FUNDAMENTALS OF SOLID STATE ENGINEERING, *2nd Edition*

FUNDAMENTALS OF SOLID STATE ENGINEERING, *2nd Edition*

Manijeh Razeghi
Northwestern University
Evanston, IL, USA

Springer

Manijeh Razeghi
Northwestern University
Evanston, IL, USA

Fundamentals of Solid State Engineering, 2^{nd} Edition

Library of Congress Control Number: 2005937004

ISBN 10: 0-387-28152-5
ISBN 13: 978-0-387-28152-0
EISBN: 0-387-28751-5 (ebook)

Printed on acid-free paper.

© 2006 Springer Science+Business Media, Inc.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, Inc., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 11530428

springeronline.com

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 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
$\Delta n, \Delta p$	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_{F_n}	Quasi-Fermi energy for electrons
E_{F_p}	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", $(=\frac{h}{2\pi})$
η	Quantum efficiency
η	Viscosity
i	$\sqrt{-1}$
i, I	Current
$J, \vec{J}$	Current density, current density vector
$J^{diff}, \overrightarrow{J^{diff}}$	Diffusion current density
$J^{drift}, \overrightarrow{J^{drift}}$	Drift current density
J_T	Thermal current
κ	Thermal conductivity coefficient
κ	Damping factor (imaginary part of the complex refractive index $\overline{N}$)
$\vec{K}$	Reciprocal lattice vector
$k, \vec{k}$	Wavenumber $(=\frac{2\pi}{\lambda}=\frac{2\pi\nu}{c})$, 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 $\overline{N}$)
$\overline{N}$	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
$\mathcal{N}_A$	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)

Foreword

It is a pleasure to write this foreword to the second edition of *Fundamentals of Solid State Engineering* by Professor Manijeh Razeghi.

Professor Razeghi is one of the world's foremost experts in the field of electronic materials crystal growth, bandgap engineering and device physics. The text combines her unique expertise in the field, both as a researcher and as a teacher. The book is all-encompassing and spans fundamental solid state physics, quantum mechanics, low-dimensional structures, crystal growth, semiconductor device processing and technology, transistors and lasers. It is excellent material for students of solid state devices in electrical engineering and materials science. The book has learning aids through exceptional illustrations and end of Chapter summaries and problems. Recent publications are often cited.

The text is a wonderful introduction to the field of solid state engineering. The breadth of subjects covered serves a very useful integrative function in combining fundamental science with application.

I have enjoyed reading the book and am delighted Professor Razeghi has put her lectures at Northwestern into a text for the benefit of a wider audience.

Venkatesh Narayanamurti
John A. and Elizabeth S. Armstrong Professor and Dean
of Engineering and Applied Sciences,
Dean of Physical Sciences and Professor of Physics
Harvard University
Cambridge, Massachusetts

Preface

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 lifeform. 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 acts to transduce 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 on to 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 biowarfare 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 100mV 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 achieve not only 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. A). Forecasts show that the current microelectronics technology is not expected to reach similar levels because of its physical limitations.

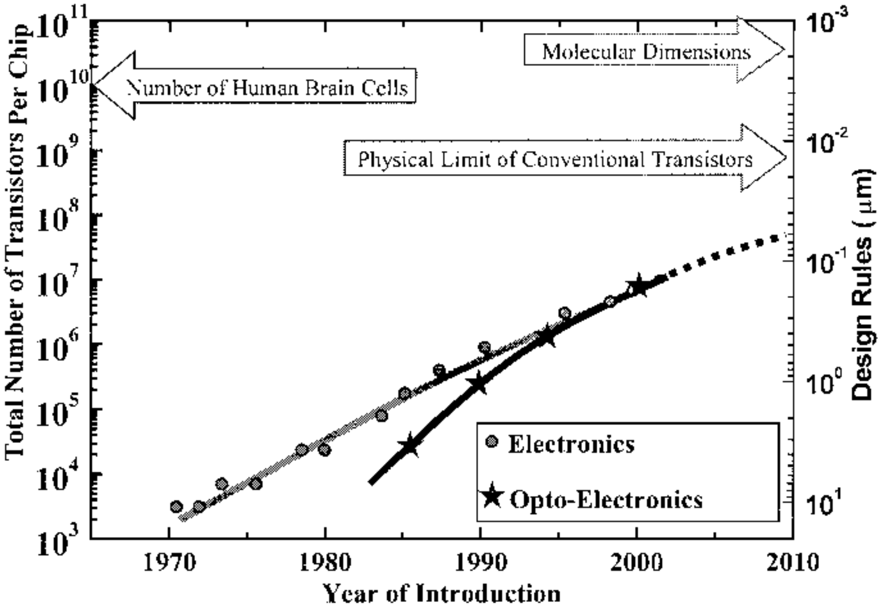


Fig. A. Evolution of the total number of transistors per computer chip and their corresponding dimensions (in an inverted scale) as a function of year. For comparison, the number of human brain cells is shown on the left scale. In addition, the physical dimension limit for conventional transistors and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

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*.

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. B below, 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 nano-sensors, nano-machines and smart materials will be made possible.

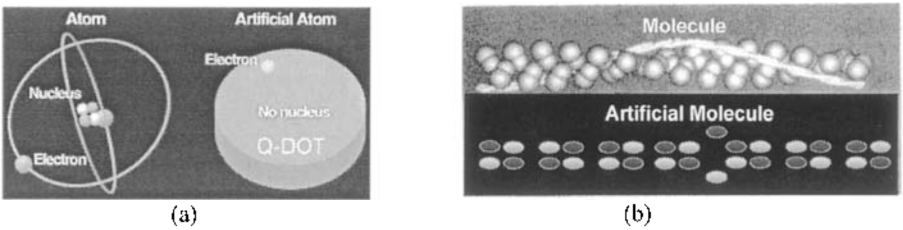


Fig. B. Schematic comparisons: (a) between a real atom and an artificial atom in the form of a quantum dot, and (b) between a real molecule and an artificial molecule.

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.

The 20th 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. C: 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.

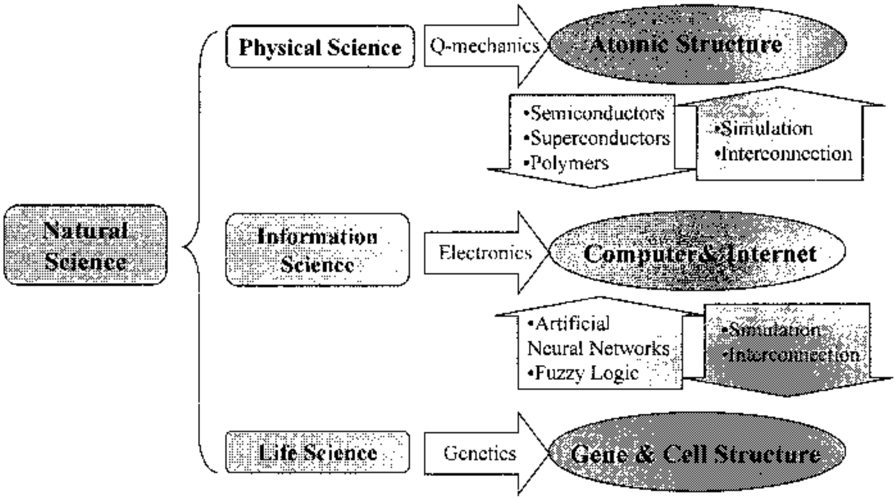


Fig. C. Three branches of Natural Science and Technology have impacted all aspects of human life in the 20th century: Physical, Information, and Life Sciences (▨). For each one, a key scientific discipline or technology has been developed: quantum mechanics, electronics, and genetics (⇨). These have 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 (▨ ▨).

The scientific and technological accomplishments of earlier centuries represent the first stage in the development of Natural Science and Technology, that of understanding (Fig. D). As the 21st 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.

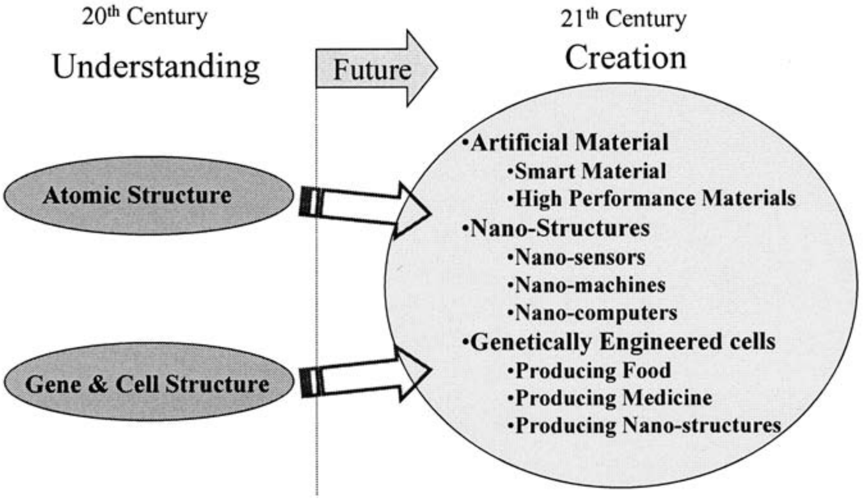


Fig. D. The scientific and technological advances of the 20^{th} century can be regarded as the understanding stage in the development of Natural Science and Technology. The 21^{st} century will be the creation stage in which novel opportunities will be discovered and carried out.

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 for example Stanford University, the Massachusetts Institute of Technology, the University of Illinois at Urbana-Champaign, the California Institute of Technology, and the University of Michigan. I reviewed numerous textbooks and reference texts in order to put together the teaching material students needed to learn nanotechnology, 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*.

The book was primarily aimed at the undergraduate level, while graduate students and researchers in the field will also find useful material in it as well. 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 second edition, I have taken into account feedback comments from students who took the courses associated with this text and from numerous colleagues in the field. The second 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 second edition is structured in two major parts. It first addresses the basic physics concepts which are at the base of solid state matter in general and semiconductors in particular. The text starts by providing an understanding of the structure of matter, real and reciprocal crystal lattices (Chapter 1), followed by a description of the structure of atoms and electrons (Chapter 2). An introduction to basic concepts in quantum mechanics (Chapter 3) and to the modeling of electrons and energy band structures in crystals (Chapter 4) is then given. A few crystal properties are then described in detail, by introducing the concept of phonons to describe vibrations of atoms in crystals (Chapter 5) and by interpreting the thermal properties of crystals (Chapter 6). The equilibrium and non-equilibrium electrical properties of semiconductors will then be reviewed, by developing the statistics (Chapter 7) as well as the transport, generation and recombination properties of these charge carriers in semiconductors (Chapter 8). These concepts will allow then to model semiconductor p-n and semiconductor-metal junctions (Chapter 9) which constitute the building blocks of modern electronics. The optical properties of semiconductors (Chapter 10) will then be described in detail. The first part of this book ends with a discussion on semiconductor heterostructures and low-dimensional quantum structures including quantum wells and superlattices, wires and dots (Chapter 11). In these Chapters, the derivation of the mathematical relations has been spelled out in thorough detail so that the reader can understand the limits of applicability of these expressions and adapt them to his or her particular situations.

The second part of this book reviews the technology associated with modern Solid State Engineering. This includes a review of compound semiconductors and crystal growth techniques (Chapter 12), including that of epitaxial thin films, followed by a brief description of the major semiconductor characterization techniques (Chapter 13) and defects in crystals (Chapter 14). Current semiconductor device processing and nanofabrication technologies will subsequently be examined (Chapters 15 and 16). A few examples of semiconductor devices, including transistors (Chapter 17), semiconductor lasers (Chapter 18), and photodetectors (Chapter 19 and 20), will then be reviewed along with a description of their theory of operation.

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

This textbook is partially based on lecture notes from the different classes, both undergraduate and graduate, which I have taught at Northwestern University. I am therefore grateful to many of my students and group for their assistance during the preparation process of this manuscript: Pierre-Yves Delaunay, Allan Evans, Aaron Gin, Darin Hoffman, Andrew Hood, Ho-Chul Lim, Ryan McClintock, Kathryn Minder, Binh Minh Nguyen, Jean Nguyen, John Szafraniec, Maho Taguchi, Stanley Tsao, Yajun Wei, Alireza Yasan, Wei Zhang, Dr. Euzi DaSilva, Dr. Patrick Kung, Dr. Erick Michel, Dr. Bijan Movaghar, Dr. Alain Quivy, Dr. Steven Slivken, and George Mach.

Finally, I would like to express my deepest appreciation to the Northwestern University Administration for their permanent support and encouragement, especially President Henry S. Bienen, Provost Lawrence B. Dumas, Vice President for Research C. Bradley Moore, and Dean of the McCormick School of Engineering Julio Ottino.

M.R.

1. Crystalline Properties of Solids

1.1. 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. 1.1). 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.

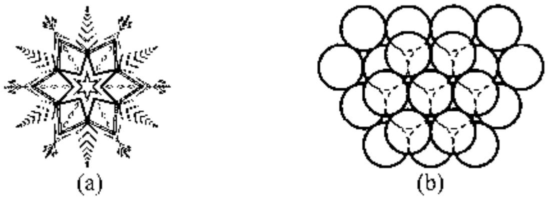
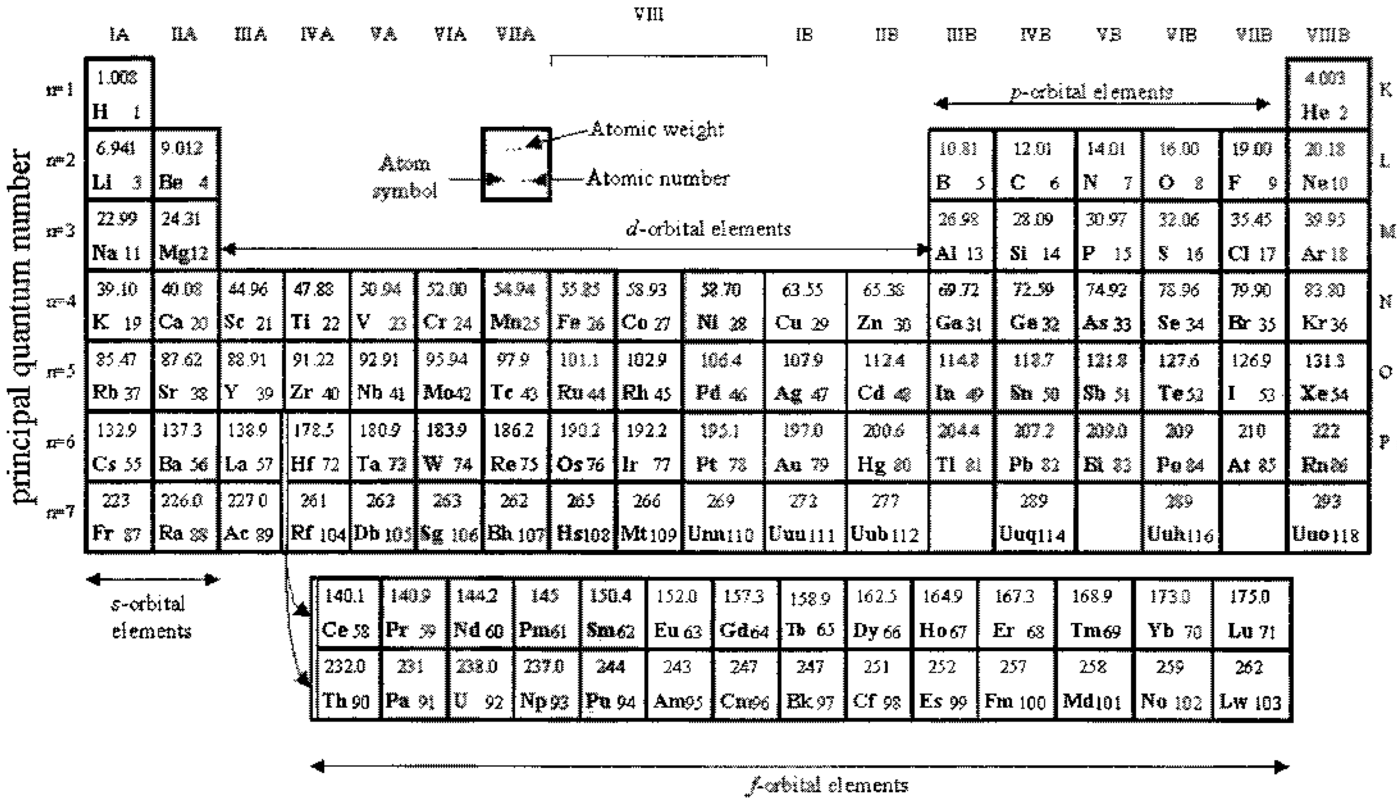


Fig. 1.1. (a) Snowflake crystal, and (b) the close-packing of spheres which gives rise to a six corner pattern. The close-packing of spheres can be thought as the way to most efficiently stack identical spheres.

All elements from the periodic table (Fig. 1.2) 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. 1.3. The natural forms of each element in the periodic table are given in Fig. A.1 in Appendix A.3.



	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	1.008 H 1																	4.003 He 2	K
n=2	6.941 Li 3	9.012 Be 4											10.81 B 5	12.01 C 6	14.01 N 7	16.00 O 8	19.00 F 9	20.18 Ne 10	L
n=3	22.99 Na 11	24.31 Mg 12											26.98 Al 13	28.09 Si 14	30.97 P 15	32.06 S 16	35.45 Cl 17	39.95 Ar 18	M
n=4	39.10 K 19	40.08 Ca 20	44.96 Sc 21	47.88 Ti 22	50.94 V 23	52.00 Cr 24	54.94 Mn 25	55.85 Fe 26	58.93 Co 27	58.70 Ni 28	63.55 Cu 29	65.38 Zn 30	69.72 Ga 31	72.59 Ge 32	74.92 As 33	78.96 Se 34	79.90 Br 35	83.80 Kr 36	N
n=5	85.47 Rb 37	87.62 Sr 38	88.91 Y 39	91.22 Zr 40	92.91 Nb 41	95.94 Mo 42	97.9 Tc 43	101.1 Ru 44	102.9 Rh 45	106.4 Pd 46	107.9 Ag 47	112.4 Cd 48	114.8 In 49	118.7 Sn 50	121.8 Sb 51	127.6 Te 52	126.9 I 53	131.3 Xe 54	O
n=6	132.9 Cs 55	137.3 Ba 56	138.9 La 57	178.5 Hf 72	180.9 Ta 73	183.9 W 74	186.2 Re 75	190.2 Os 76	192.2 Ir 77	195.1 Pt 78	197.0 Au 79	200.6 Hg 80	204.4 Tl 81	207.2 Pb 82	209.0 Bi 83	209 Po 84	210 At 85	222 Rn 86	P
n=7	223 Fr 87	226.0 Ra 88	227.0 Ac 89	261 Rf 104	262 Db 105	263 Sg 106	262 Bh 107	265 Hs 108	266 Mt 109	269 Unn 110	273 Uuu 111	277 Uub 112		289 Uuq 114		289 Uuh 116		293 Uuo 118	

140.1 Ce 58	140.9 Pr 59	144.2 Nd 60	145 Pm 61	150.4 Sm 62	152.0 Eu 63	157.3 Gd 64	158.9 Tb 65	162.5 Dy 66	164.9 Ho 67	167.3 Er 68	168.9 Tm 69	173.0 Yb 70	175.0 Lu 71
232.0 Th 90	231 Pa 91	238.0 U 92	237.0 Np 93	244 Pu 94	243 Am 95	247 Cm 96	247 Bk 97	251 Cf 98	252 Es 99	257 Fm 100	258 Md 101	259 No 102	262 Lw 103

Fig. 1.2. Periodic table of elements. For each element, its symbol, atomic number and atomic weight are shown.

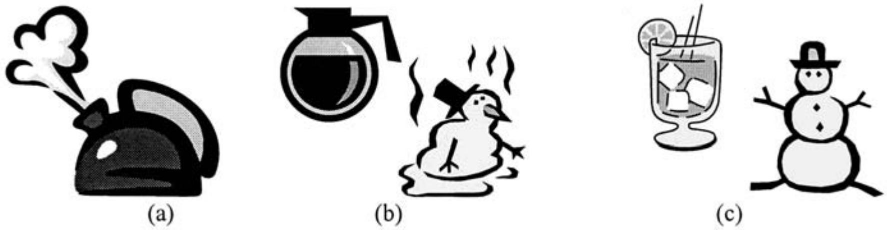


Fig. 1.3. Illustration of the physical states of water: (a) gas also known as water vapor, (b) liquid or common water, (c) solid also known as snow or ice.

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. 1.4. 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 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. 1.4.

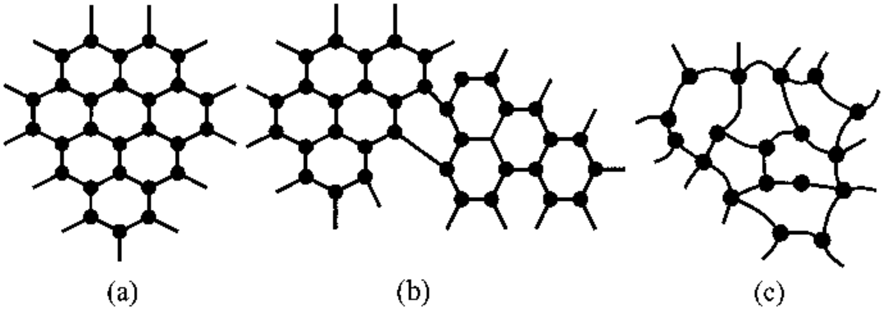


Fig. 1.4. Arrangement of atoms: (a) a single-crystalline, (b) a polycrystalline, and (c) an amorphous material.

1.2. 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 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. 1.5(a). With the pattern shown in Fig. 1.5(b), one can obtain the two-dimensional crystal in Fig. 1.5(c) which shows that a pattern associated with each lattice point.

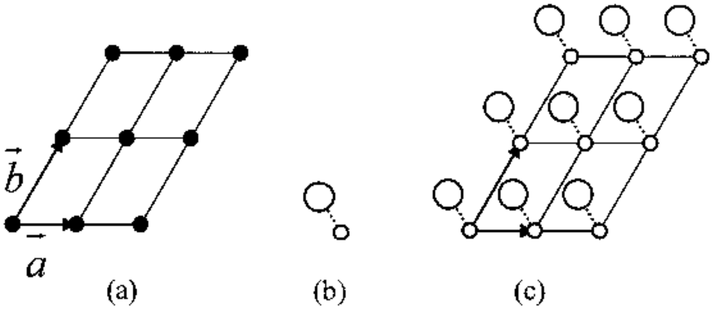


Fig. 1.5. Example of (a) two-dimensional lattice, (b) pattern, and (c) two-dimensional crystal illustrating a pattern associated with each lattice point.

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. 1.5(a) and Fig. 1.6, 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:

Eq. (1.1)
$$\vec{R} = n_1\vec{a} + n_2\vec{b} + n_3\vec{c},$$
$$n_{1,2,3} = 0, \pm 1, \pm 2, \ldots$$

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 1.1, 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 1.1.

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

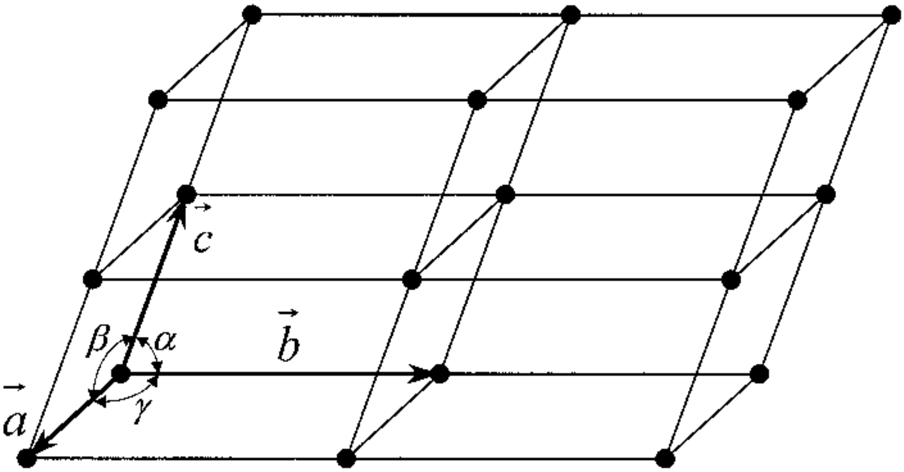


Fig. 1.6. Example of a three-dimensional lattice, with translation vectors and the angles between two vectors. By taking the origin at one lattice point, the position of any lattice point can be determined by a vector which is the sum of integer numbers of translation vectors.

Crystal systems	Axial lengths and angles
Cubic	Three equal axes at right angles a=b=c, α=β=γ=90°
Tetragonal	Three axes at right angles, two equal a=b≠c, α=β=γ =90°
Orthorhombic	Three unequal axes at right angles a≠b≠c, α=β=γ=90°
Trigonal	Three equal axes, equally inclined a=b=c, α=β=γ≠90°
Hexagonal	Two equal coplanar axes at 120°, third axis at right angles a=b≠c, α=β=90°, γ=120°
Monoclinic	Three unequal axes, one pair not at right angles a≠b≠c, α=γ=90°≠β
Triclinic	Three unequal axes, unequally inclined and none at right angles a≠b≠c, α≠β≠γ≠90°

Table 1.1. The seven crystal systems.

1.3. 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). Fig. 1.7 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.

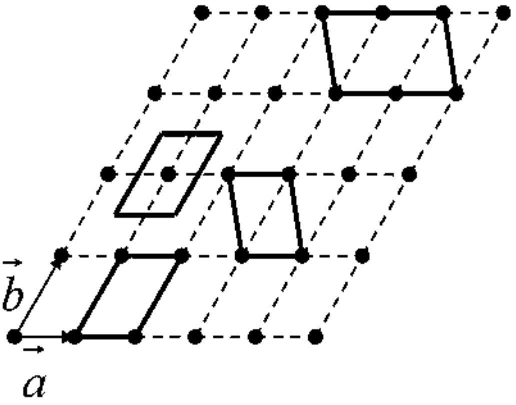


Fig. 1.7. Three examples of possible unit cells for a two-dimensional lattice. The unit cells are delimited in solid lines. The same principle can be applied for the choice of a unit cell in three dimensions.

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 Eq. (1.1). An example of primitive unit cell in a three-dimensional lattice is shown in Fig. 1.8. 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 contains only one lattice point. The volume of a primitive unit cell is found from vector algebra:

Eq. (1.2) $$V = = \left|\vec{a}.(\vec{b}\times\vec{c})\right|$$

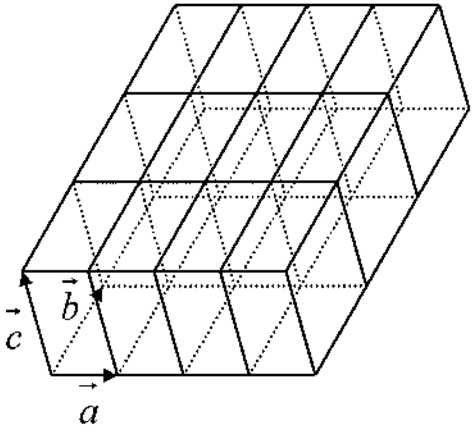


Fig. 1.8. Three-dimensional lattice and a corresponding primitive unit cell defined by the three basis vectors $\vec{a}, \vec{b}, \vec{c}$.

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:

$$primitive\ \ unit\ cell\ volume = \frac{crystal\ volume}{N}$$

A primitive unit cell is in many cases characterized by non-orthogonal lattice vectors (as in Fig. 1.6). 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. 1.9 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 (each n_c point is shared by eight corners). For example, the number of atoms per unit cell in the fcc lattice (Fig. 1.9(c)) (n_i=0, n_f=6, and n_c=8) is:

Eq. (1.3) $n_u = n_i + \frac{n_f}{2} + \frac{n_c}{8}$ =4 atoms/unit cell

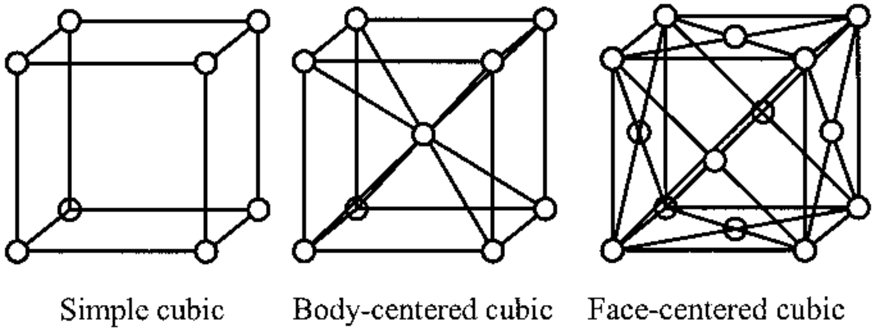


Fig. 1.9. Three-dimensional unit cells: simple cubic (left), body-centered cubic (bcc)(middle), and face-centered cubic (fcc) (right).

1.4. 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. 1.10, 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.

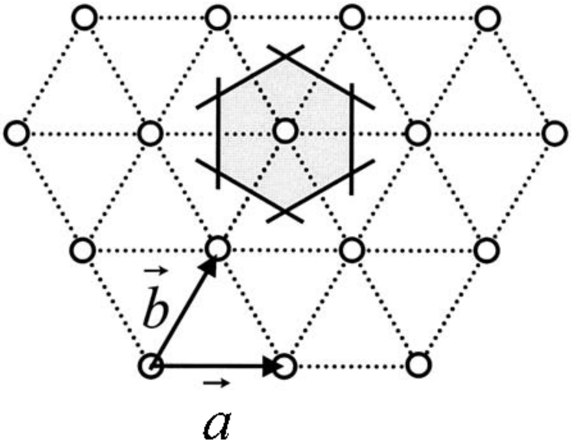
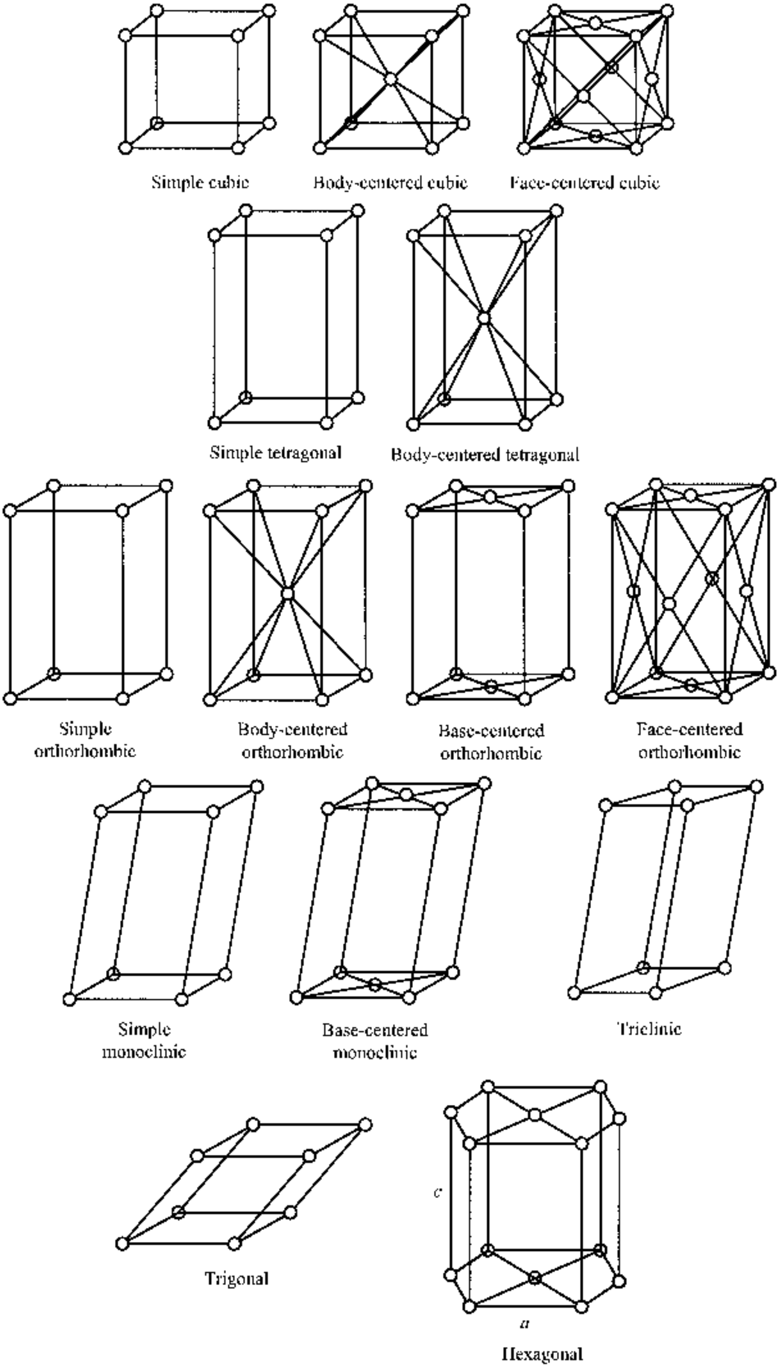


Fig. 1.10. Two-dimensional Wigner-Seitz cell and its construction method: select a lattice point, draw lines from a given lattice point to all nearby points, bisect these lines with orthogonal planes, construct the smallest polyhedron that contains the first selected lattice.

1.5. 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. 1.11 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 Table 1.1, no crystal can have a lattice other than one of those 14 Bravais lattices.



Simple cubic

Body-centered cubic

Face-centered cubic

Simple tetragonal

Body-centered tetragonal

Simple orthorhombic

Body-centered orthorhombic

Base-centered orthorhombic

Face-centered orthorhombic

Simple monoclinic

Base-centered monoclinic

Triclinic

Trigonal

Hexagonal

Fig. 1.11. The fourteen Bravais lattices, illustrating all the possible three-dimensional crystal lattices.

1.6. 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 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.

1.6.1. 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. 1.12. 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.

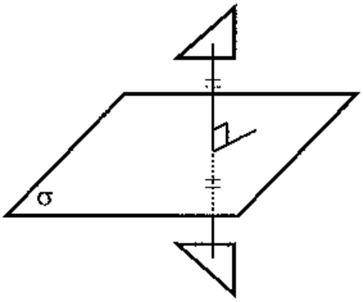


Fig. 1.12. Illustration of a plane reflection. The triangular object and its reflected image are mirror images of each other.

1.6.2. 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. 1.13. In crystallography, the angle of rotation cannot be arbitrary but can only take the following fractions of 2π: $\theta = \frac{2\pi}{1}, \frac{2\pi}{2}, \frac{2\pi}{3}, \frac{2\pi}{4}, \frac{2\pi}{6}$.

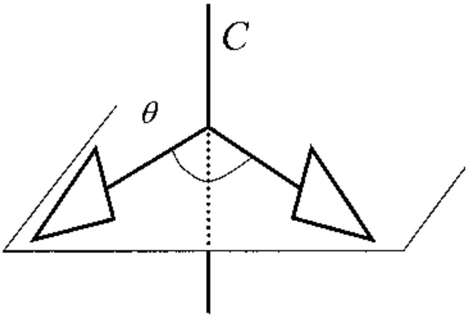


Fig. 1.13. Illustration of a rotation symmetry. The triangular object and its image are separated by an angle equal to θ.

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. Fig. 1.14 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. 1.15.

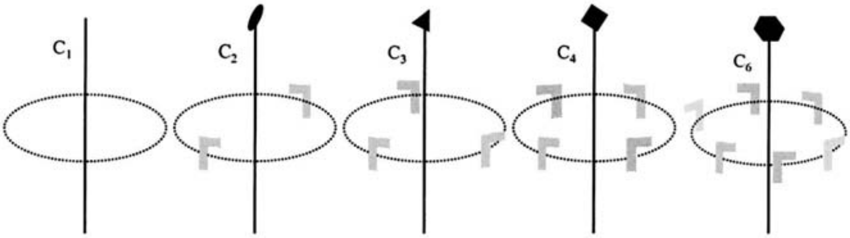


Fig. 1.14. Crystal bodies with symmetries C_1, C_2, C_3, C_4 *and* C_6. *The elbow patterns are brought into self-coincidence after a rotation around the axis shown and through an angle equal to* $2\pi/n$ *where* $n=1, 2, 3, 4,$ *or* 6.

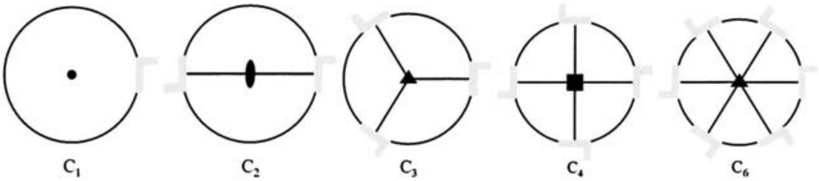


Fig. 1.15. Crystal bodies with symmetries C_1, C_2, C_3, C_4 *and* C_6 *with the rotation axes perpendicular to the plane of the figure.*

1.6.3. 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. Fig. 1.16(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. Fig. 1.16(b) shows the bodies with a symmetry C_{4v}. The number of elements is $2n$ too.

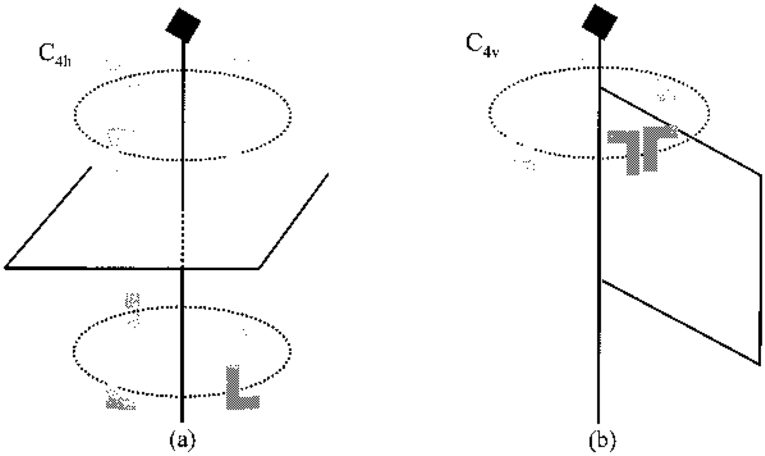


Fig. 1.16. Crystal bodies with symmetries (a) C_{4h} where the reflection plane is perpendicular to the rotation axis; and (b) C_{4v} where the reflection plane passes through the rotation axis.

1.6.4. 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. 1.17.

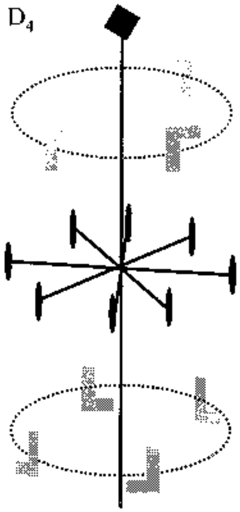


Fig. 1.17. Crystal bodies with symmetry D_4. In addition to the C_4 axis, there are 4 C_2 axes of rotation perpendicular to the C_n axis.

1.6.5. 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 the possible combinations of such symmetry operations leads 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. 1.18(a).

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 Fig. 1.18(b).

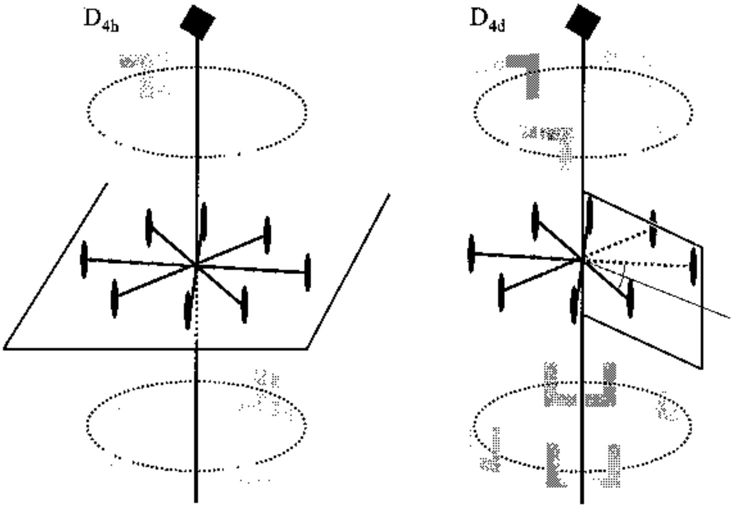


Fig. 1.18. Bodies with symmetries (a) D_{4h} and (b) D_{4d}.

1.6.6. 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. 1.19.

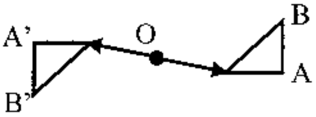


Fig. 1.19. Illustration of an inversion symmetry. Any point of the triangular object and its image are such that the inversion center is at the middle of these two points.

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.

1.6.7. 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 sub-group 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 sub-group is made by combining roto-inversion operation, is denoted S_4 and is illustrated in Fig. 1.20. Its number of elements is 4.

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

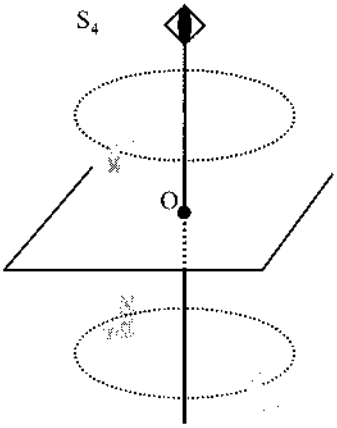


Fig. 1.20. Bodies with symmetry S_4.

1.6.8. T group

The tetrahedron axes group T is illustrated in Fig. 1.21. 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.

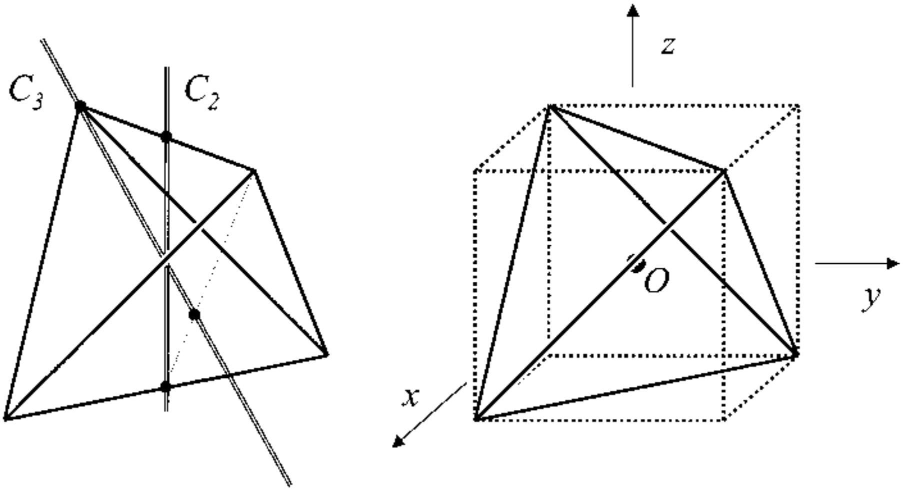


Fig. 1.21. Axes of rotation for the T group, including four C_3 and three C_2 axes. The orientation of the tetrahedron with respect to the cubic coordinate axes is shown on the right.

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 8 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 (3 elements),
- and the identity (1 element).

1.6.9. T_d group

The T_d point group contains all the symmetry elements of a regular tetrahedron (Fig. 1.22). Basically, it includes all the symmetry operations of the T group in addition to an inversion center at the center of the tetrahedron.

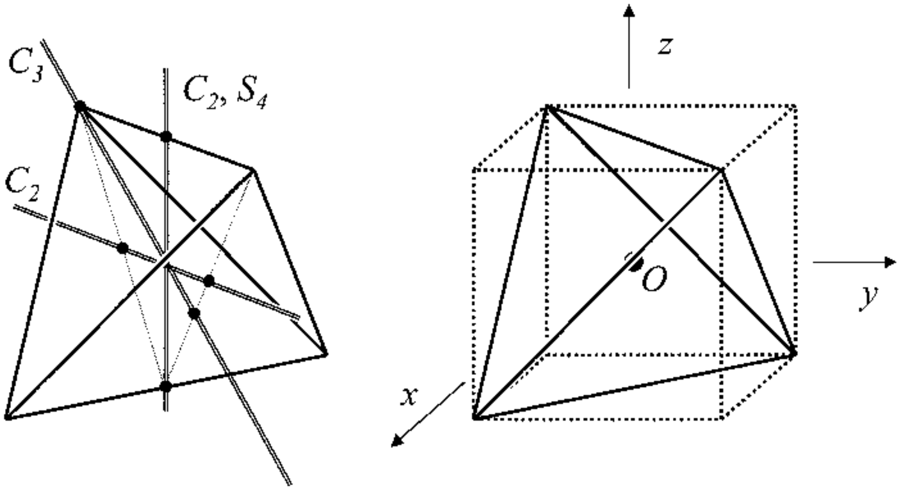


Fig. 1.22. Axes of rotation for T_d group, including four C_3, three C_2 axes passing through the center of opposite faces, three S_4 axes, and six C_2 axes passing through the centers of diagonally opposite sides.

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 8 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 (3 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, (6 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, (6 elements),
- and finally, the identity (1 element).

1.6.10. 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 9 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 (8 elements),
- rotations through an angle π, about the six C_2 axes passing through the midpoints of opposite edges (6 elements),
- and finally, the identity (1 element).

1.6.11. 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-center point of a cube (24 elements).

1.6.12. List of crystallographic point groups

The point groups previously reviewed are constructed by considering all the possible combinations basic symmetry operations (plane reflections and rotations) discussed in sub-sections 1.6.1 and 1.6.2. 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 1.2 shows the correspondence between two widely used notations.

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}	$4mm$
	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}	$6mm$
	D_{3h}	$\bar{6}m2$
	D_{6h}	$6/mmm$

Table 1.2. List of the 32 crystallographic point groups.

1.7. 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 $\vec{R}$ (Eq. (1.1)).

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.

1.8. 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. 1.23.

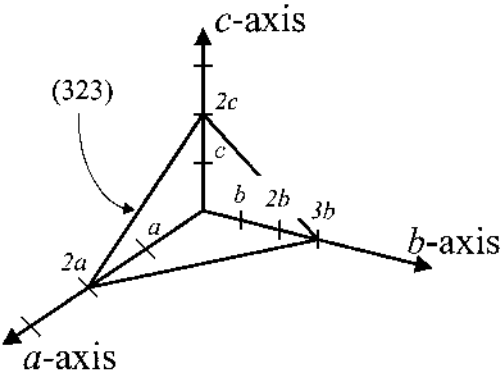


Fig. 1.23. Example of a plane which passes through lattice points. Its Miller indices are (hkl)=(323) and are used to identify this plane in the crystal. These indices are obtained as follows: note where the plane intersects the coordinate axes, it is either an integer multiple or an irreducible fraction of the axis unit length; invert the intercept values; using the appropriate multiplier, convert these inverted values into integer numbers; enclose the integer numbers in parenthesis.

Fig. 1.23 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 a few planes in a cubic lattice are shown in Fig. 1.24. 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 1.3.

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. 1.23, which coincides with vector $\vec{a}$, is [100].

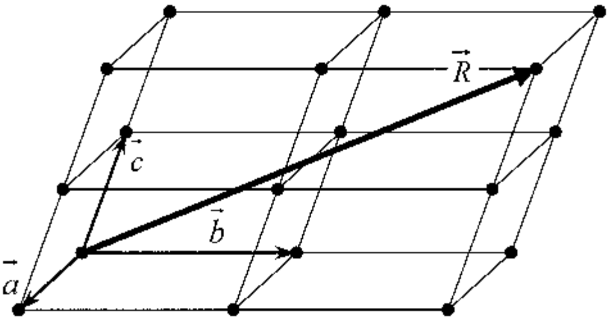
For the indices of both plane and directions, a negative value of the index is written with a bar sign above the index, such as $(\bar{h}kl)$ or $[u\bar{v}w]$.

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

Table 1.3. Conventions used to label directions and planes in crystallography.

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. 1.23 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. 1.24.

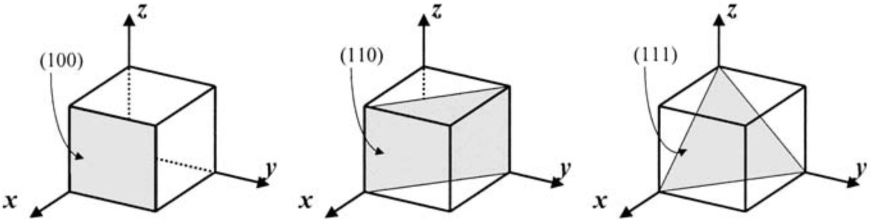


Fig. 1.24. Miller indices of the three principal planes in the cubic structure. If a plane is parallel to an axis, we consider that it "intersects" this axis at infinity and we get the Miller indices: 1,∞,∞ => 1/1:1/∞:1/∞=1:0:0 => (100).

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. 1.25:

Eq. (1.4) $$d_{hkl} = a \Big/ \sqrt{h^2 + k^2 + l^2}$$

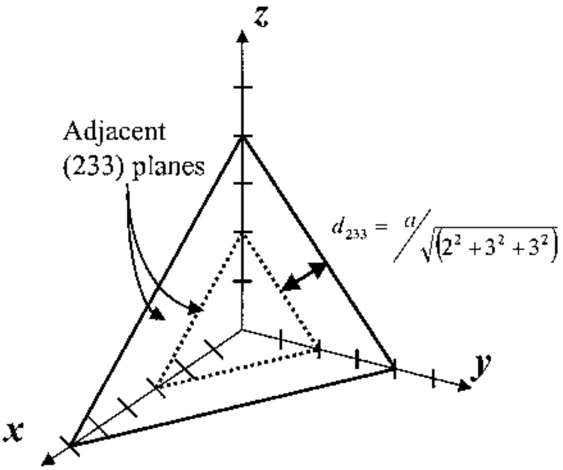


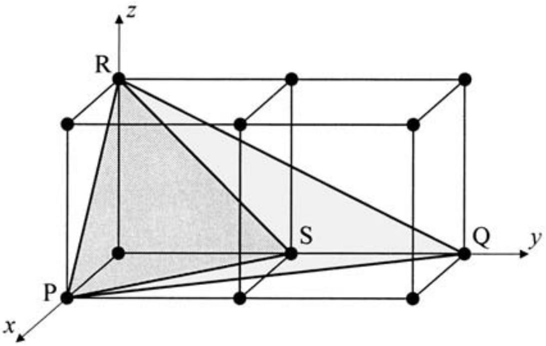
Fig. 1.25. Illustration of the interplanar spacing in a cubic lattice between two adjacent (233) planes.

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

Eq. (1.5) $$\cos(\theta) = (h_1h_2 + k_1k_2 + l_1l_2) \Big/ \sqrt{(h_1^2 + k_1^2 + l_1^2)(h_2^2 + k_2^2 + l_2^2)}$$

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: $\cos(\theta)=\dfrac{1\times 2+1\times 1+1\times 2}{\sqrt{(1^2+1^2+1^2)(2^2+1^2+2^2)}}=\dfrac{5\sqrt{3}}{9}$.

The angle between the two planes is therefore: 15.8 deg.

In hexagonal systems, the *a*- and *b*-axes of Fig. 1.23 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 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. 1.26. It is then customary to use a four-index system for planes and directions: (*hkil*) and [*uvtw*], respectively, as shown in Table 1.3. 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 $\vec{a}_3$.

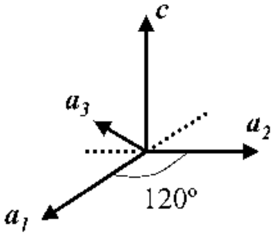


Fig. 1.26. Coordinate axes used to determine Miller indices for hexagonal systems.

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. 1.27. A primary flat is such that it is perpendicular to the [110] direction, while a smaller secondary flat is perpendicular to the $[01\bar{1}]$ direction.

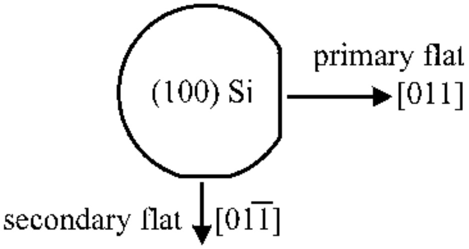


Fig. 1.27. Illustration of the use of primary and secondary flats on a (100) oriented silicon crystal wafer to indicate the in-plane crystallographic orientation of the wafer.

1.9. 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.

1.9.1. 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), (¼,¾,¾).

The tetrahedral bonding characteristic of the diamond structure is shown in Fig. 1.28(a). Each atom has 4 nearest neighbors and 12 second nearest neighbors. For example, the atom located at (¼,¼,¼) at the center of the cube in Fig. 1.28(b) has four nearest neighbors also shown in Fig. 1.28(b) 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, n_c are the numbers of points in the interior, on faces and on corners of the cubic unit cell shown in Fig. 1.28(a), 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:

$n_u = 4 + \frac{6}{2} + \frac{8}{8} = 8$ 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/(0.543×10^{-7})^3=5×10^{22} atoms/cm^3.

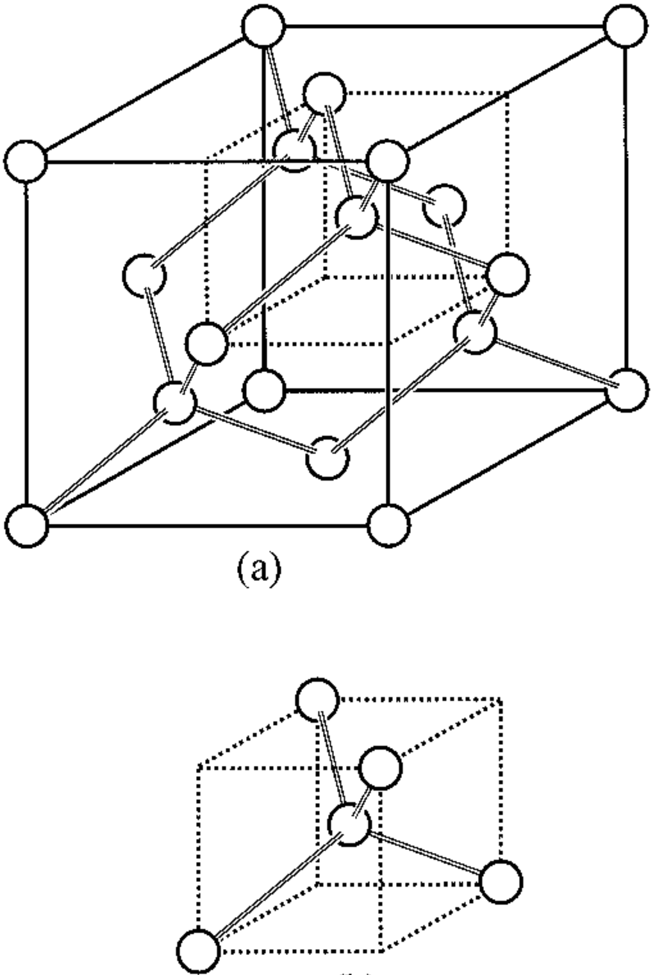


Fig. 1.28. (a) Diamond lattice. The Bravais lattice is face-centered cubic with a basis consisting of two identical atoms displaced from each other by a quarter of the cubic body diagonal. The atoms are connected by covalent bonds. The cube outlined by the dashed lines shows one tetrahedral unit. (b) Tetrahedral unit of the diamond lattice.

1.9.2. 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. 1.29. 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. 1.29 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); (¾,¾,¾).

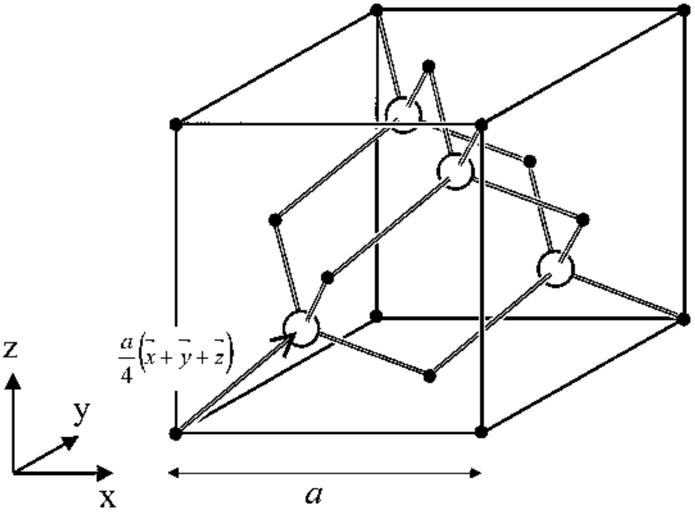


Fig. 1.29. Cubic unit cell for the zinc blende structure. The Bravais lattice is face-centered cubic with a basis of two different atoms represented by the open and solid spheres, and separated by a quarter of the cubic body diagonal. The crystal does not appear the same when viewed along a body diagonal from one direction or the other.

1.9.3. Sodium chloride structure

The structure of sodium chloride, NaCl, is shown in Fig. 1.30. 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).

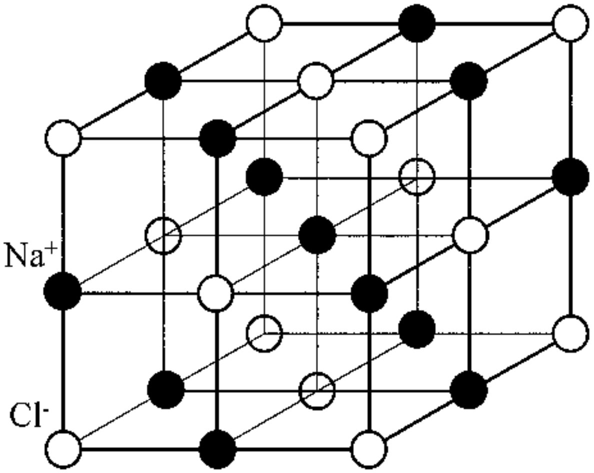


Fig. 1.30. Sodium chloride crystal. The Bravais lattice is face-centered cubic with a basis of two ions: one Cl^- ion at (0,0,0) and one Na^+ ion at (½,½,½), separated by one half of the cubic body diagonal. The figure shows one cubic unit cell.

1.9.4. Cesium chloride structure

The cesium chloride structure is shown in Fig. 1.31. 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.

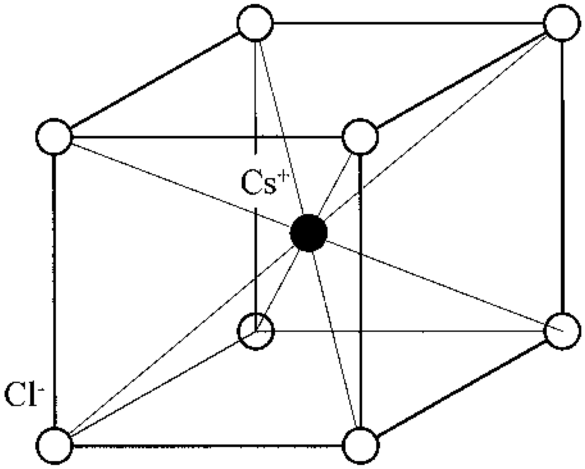


Fig. 1.31. The cesium chloride crystal structure. The Bravais lattice is cubic with a basis of two ions: one Cl$^-$ ion at (0,0,0) and one Cs$^+$ ion at (½,½,½), separated by one half the cubic body diagonal.

1.9.5. 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 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. 1.32) 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. 1.32. 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. 1.32. We obtain the hexagonal close-packed structure (Fig. 1.33) 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).

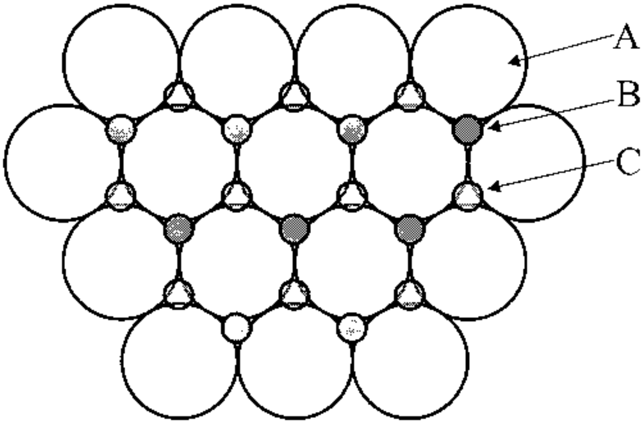


Fig. 1.32. The closed-packed array of spheres. Note the three different possible positions, A, B, and C for the successive layers. The most space efficient way to arrange identical spheres or atoms in a plane is to first place each sphere in contact with six others in that plane (positions A). The most 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.

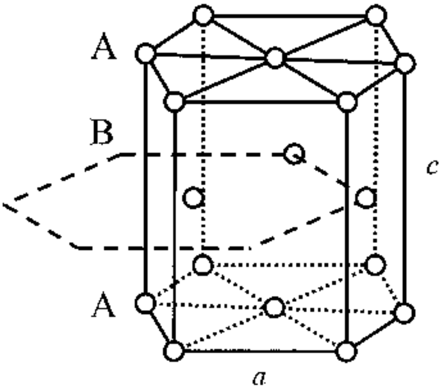


Fig. 1.33. The hexagonal close-packed (hcp) structure. This Bravais lattice of this structure is hexagonal, with a basis of two identical atoms. It is constructed by stacking layers in the ABABAB... sequence. The lattice parameters a *and* c *are indicated.*

Zinc, magnesium and low-temperature form of titanium have the hcp structure. The ratio *c/a* for ideal hexagonal close-packed structure in Fig. 1.33 is 1.633. The number of nearest-neighbor atoms is 12 for hcp structures. Table 1.4 shows the *c/a* parameter for different 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

Table 1.4. c/a parameter for various hexagonal crystals.

1.9.6. Wurtzite structure

A few III-V and several II-VI semiconductor compounds have the wurtzite structure shown in Fig. 1.34.

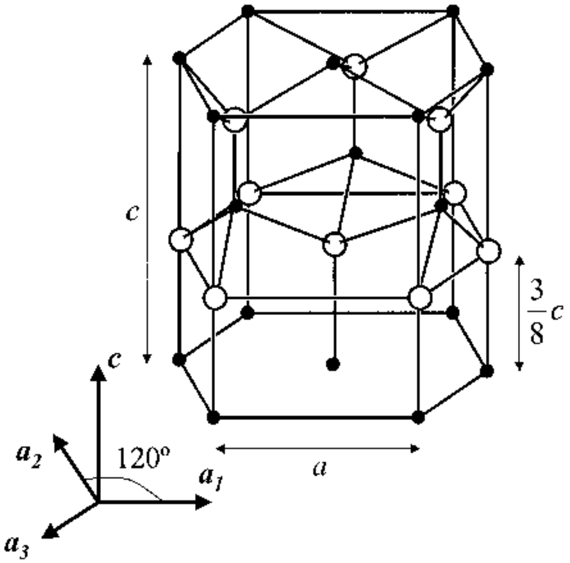


Fig. 1.34. The wurtzite structure consists of two interpenetrating hcp structures, each with a different atom, shifted along the c-direction. The bonds between atoms and the hexagonal symmetry are shown.

This structure consists of two interpenetrating hexagonal close-packed lattices, each with different atoms, ideally displaced from each other by 3/8c 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}.

1.9.7. 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:

$$\frac{\frac{4}{3}\pi(\frac{a}{2})^3}{a^3} = 0.52 .$$

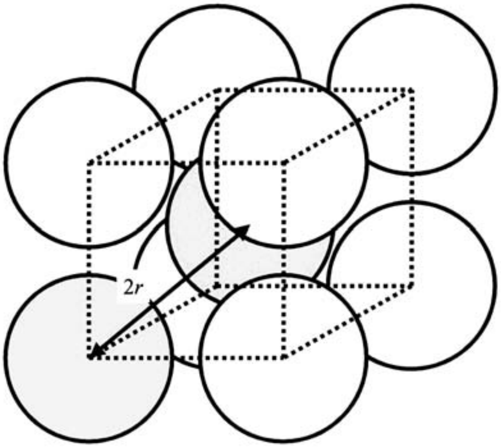
The following two examples illustrate the determination of the packing factor for the other two cubic lattices.

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:

$$\frac{2\times\frac{4\pi}{3}\left(\frac{\sqrt{3}}{4}a\right)^3}{a^3}=\frac{\pi\sqrt{3}}{8}=0.68\,.$$

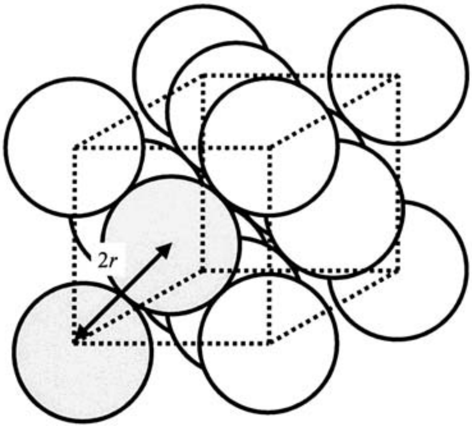


Example

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:

$$\frac{4\times\frac{4}{3}\pi(\frac{\sqrt{2}}{4}a)^3}{a^3}=0.7405\,.$$



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).

1.10. 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:

Eq. (1.6) $$\rho(\vec{r})=\sum_{\vec{K}}P(\vec{K})\exp(i\vec{K}\vec{r})$$

where the vector $\vec{K}$ is used to index the summation and the Fourier coefficients $P(\vec{K})$. This vector $\vec{K}$ has the dimension of an inverse distance and for a periodic function, can take discrete values and 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 $\vec{R}$:

Eq. (1.7) $$\rho(\vec{r}) = \rho(\vec{r} + \vec{R}) = \sum_{\vec{K}} P(\vec{K}) \exp\left[i\vec{K}.\left(\vec{r} + \vec{R}\right)\right]$$

which becomes:

Eq. (1.8) $$\sum_{\vec{K}} P(\vec{K}) \exp(i\vec{K}.\vec{r}) = \sum_{\vec{K}} P(\vec{K}) \exp\left[i\vec{K}.\left(\vec{r} + \vec{R}\right)\right]$$

Eq. (1.8) 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:

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

or:

Eq. (1.9) $$\exp\left(i\vec{K}.\vec{R}\right) = 1$$

for any lattice vector $\vec{R}$. Eq. (1.9) 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. (1.9) which is equivalent to:

Eq. (1.10) $$\vec{K}.\vec{R} = 2\pi m$$

where m=0, ±1, ±2,... is an integer. Using the expression for $\vec{R}$ from Eq. (1.1) of Chapter 1, we obtain:

Eq. (1.11) $$\left(\vec{K}.\vec{a}\right)n_1 + \left(\vec{K}.\vec{b}\right)n_2 + \left(\vec{K}.\vec{c}\right)n_3 = 2\pi m$$

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, Eq. (1.11) leads us to:

Eq. (1.12)
$$\begin{cases} \vec{K}.\vec{a} = 2\pi h_1 \\ \vec{K}.\vec{b} = 2\pi h_2 \text{ with } h_{1,2,3} = 0;\pm1;\pm2,... \\ \vec{K}.\vec{c} = 2\pi h_3 \end{cases}$$

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. (1.1) of Chapter 1. These basis vectors define what we call the reciprocal lattice. Any reciprocal lattice vector $\vec{K}$ can thus be represented as:

Eq. (1.13) $$\vec{K} = h_1\vec{A} + h_2\vec{B} + h_3\vec{C}$$

From Eq. (1.12) and Eq. (1.13) we have:

Eq. (1.14)
$$\begin{cases} (\vec{A}.\vec{a})h_1 + (\vec{B}.\vec{a})h_2 + (\vec{C}.\vec{a})h_3 = 2\pi h_1 \\ (\vec{A}.\vec{b})h_1 + (\vec{B}.\vec{b})h_2 + (\vec{C}.\vec{b})h_3 = 2\pi h_2 \\ (\vec{A}.\vec{c})h_1 + (\vec{B}.\vec{c})h_2 + (\vec{C}.\vec{c})h_3 = 2\pi h_3 \end{cases}$$

Eq. (1.14) can be satisfied only when:

Eq. (1.15)
$$\begin{cases} \vec{A}.\vec{a} = \vec{B}.\vec{b} = \vec{C}.\vec{c} = 2\pi \\ and \\ \vec{A}.\vec{b} = \vec{A}.\vec{c} = 0 \\ \vec{B}.\vec{a} = \vec{B}.\vec{c} = 0 \\ \vec{C}.\vec{b} = \vec{C}.\vec{a} = 0 \end{cases}$$

Eq. (1.15) 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 ($\vec{a},\vec{b},\vec{c}$):

Eq. (1.16) $$\begin{cases} \vec{A} = 2\pi \dfrac{\vec{b} \times \vec{c}}{\vec{a}.(\vec{b} \times \vec{c})} \\ \vec{B} = 2\pi \dfrac{\vec{c} \times \vec{a}}{\vec{a}.(\vec{b} \times \vec{c})} \\ \vec{C} = 2\pi \dfrac{\vec{a} \times \vec{b}}{\vec{a}.(\vec{b} \times \vec{c})} \end{cases}$$

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):

Eq. (1.17) $$\vec{A}.\left(\vec{B} \times \vec{C}\right) = \frac{8\pi^3}{\vec{a}.(\vec{b} \times \vec{c})}$$

We note that the vectors of reciprocal space have the same dimensions as the wavenumbers and momenta of electromagnetic waves for example. 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.

1.11. 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. 1.10. 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.

1.12. 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

Further reading

Dalven, R., *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, 1990.

Holden, A., *The Nature of Solids*, Dover, New York, 1992.

Kittel, C., *Introduction to Solid State Physics*, John Wiley & Sons, New York, 1986.

Loretto, M.H., *Electron-Beam Analysis of Materials*, Chapman & Hall, London, 1994.

Lovett, D.R., *Tensor Properties of Crystals*, Institute of Physics, Bristol, UK, 1999.

Mayer, J.W. and Lau, S., *Electronic Materials Science for Integrated Circuits in Si and GaAs*, Macmillan, New York, 1990.

McKelvey, J.P., *Solid State and Semiconductor Physics,* Harper and Row, New York, 1966.

Pierret, R.F., *Advanced Semiconductor Fundamentals,* Addison-Wesley, Reading, MA, 1989.

Rhodes, G., *Crystallography Made Crystal Clear*, Academic Press, San Diego, Calif., 1993.

Rosenberg, H.M., *The Solid State*, Oxford physics series, Clarendon Press, Oxford, 1978.

Scott, W.R., *Group Theory*, Dover Publications, New York, 1964.

Weyl, H., *The Theory of Groups and Quantum Mechanics,* Dover Publications, New York, 1950.

Wolfe, C.M., Holonyak, N., and Stillman, G.E., *Physical Properties of Semiconductors*, Prentice-Hall, Englewood Cliffs, N.J., 1989.

Yu, P.Y. and Cardona, M., *Fundamentals of Semiconductors: Physics and Materials Properties*, Springer, New York, 1999.

Ziman, J.M., *Elements of Advanced Quantum Theory,* Cambridge University Press, London, 1969.

Ziman, J.M., *Principles of the Theory of Solids,* Cambridge University Press, Cambridge, 1998.

Problems

1. Fig. 1.6 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 $\theta=\dfrac{2\pi}{5}$ 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 $\vec{c}=a\left(\vec{x}+\vec{y}+\vec{z}\right)$. 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. (1.16). Prove that the volume defined by these vectors is given by:

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

2. Electronic Structure of Atoms

2.1. 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 Chapter 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 Chapter 3.

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 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.

2.2. 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:

Eq. (2.1) $$\frac{\upsilon}{c}=\frac{1}{\lambda}=Ry\left(\frac{1}{n^2}-\frac{1}{\left(n'\right)^2}\right)$$

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 (=2.99792×10^{8} m.s^{-1}=2.99792×10^{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:

$$E = \frac{hc}{\lambda} = h\upsilon$$

where h (=6.62617×10^{-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. (2.1) 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. 2.1, 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 20th century, a suitable explanation was not available, even for the simplest case of the hydrogen atom. Prior to 1913, the explanation for 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 nuclei 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.

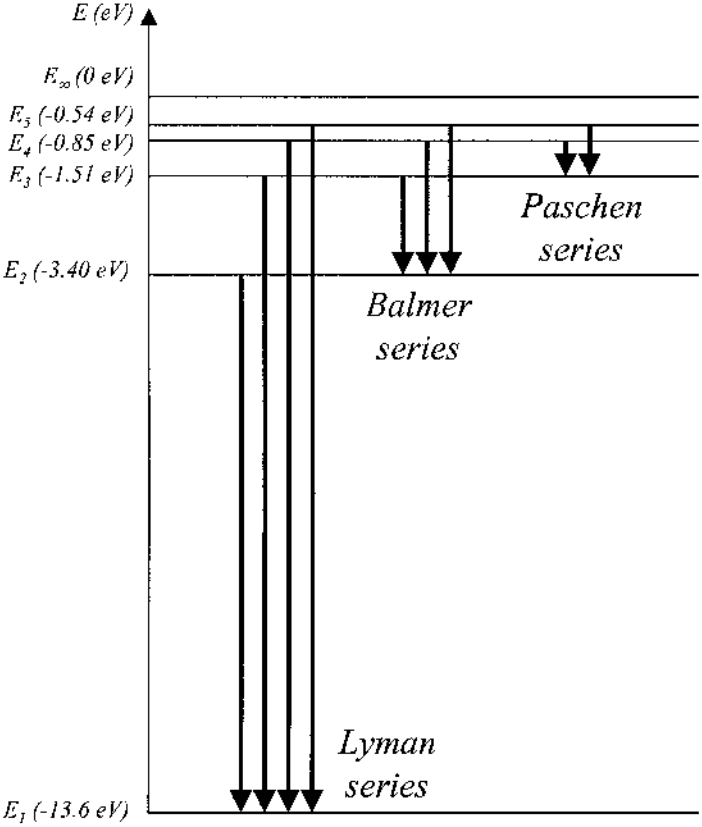


Fig. 2.1. Energies of the light emitted from the hydrogen atom (shown by arrows). The Lyman series corresponds to n=1 in Eq. (2.1), the Balmer series corresponds to n=2 and the Paschen series to n=3.

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 $\upsilon = \frac{|E_n - E_{n'}|}{h}$.

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:

Eq. (2.2) $$L_n = m v r_n = n\frac{h}{2\pi}, \quad n = 1,2,...$$

where m is the mass of the electron, v is the linear electron velocity, and n an integer expressing the quantization and used to index the electron orbits. Since the orbit is circular, the electron experiences a centripetal acceleration v^2 / r_n. The coulombic force between the electron and nucleus provides this acceleration, as illustrated in Fig. 2.2.

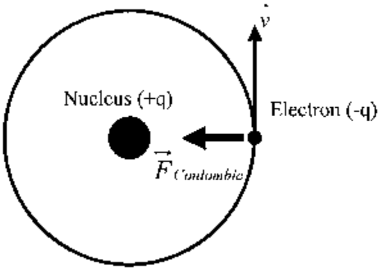


Fig. 2.2. Schematic diagram showing the electron orbit, the attractive coulombic force between the positively charged nucleus and the orbiting negatively charged electron, and the velocity of the electron which is always tangential to its circular orbit.

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

Eq. (2.3) $$\frac{q^2}{4\pi\varepsilon_0 r_n^2} = \left|\vec{F}_{coulombic}\right| = \frac{m v^2}{r_n}$$

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

Combining Eq. (2.2) and Eq. (2.3), one obtains the discrete radius of an electron orbit:

Eq. (2.4) $$r_n = \frac{\varepsilon_0 n^2 h^2}{\pi m q^2}$$

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:

Eq. (2.5) $$E_n = \frac{1}{2}\frac{q^2}{4\pi\varepsilon_0 r_n} - \frac{q^2}{4\pi\varepsilon_0 r_n} = -\frac{1}{8}\frac{q^2}{\pi\varepsilon_0 r_n}$$

With Eq. (2.4) we finally have:

Eq. (2.6) $$E_n = \frac{-m\,q^4}{8\,(\varepsilon_0 n\,h)^2} = -\frac{13.6}{n^2} \quad \text{in units of electron-volts (eV)}$$

This theory thus provided an explanation for each series of spectroscopic lines in the emission spectrum from atomic hydrogen as shown in Fig. 2.1. 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. (2.4) 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. 2.1 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. 2.1. 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. 2.3. 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 Chapter 3 for more details).

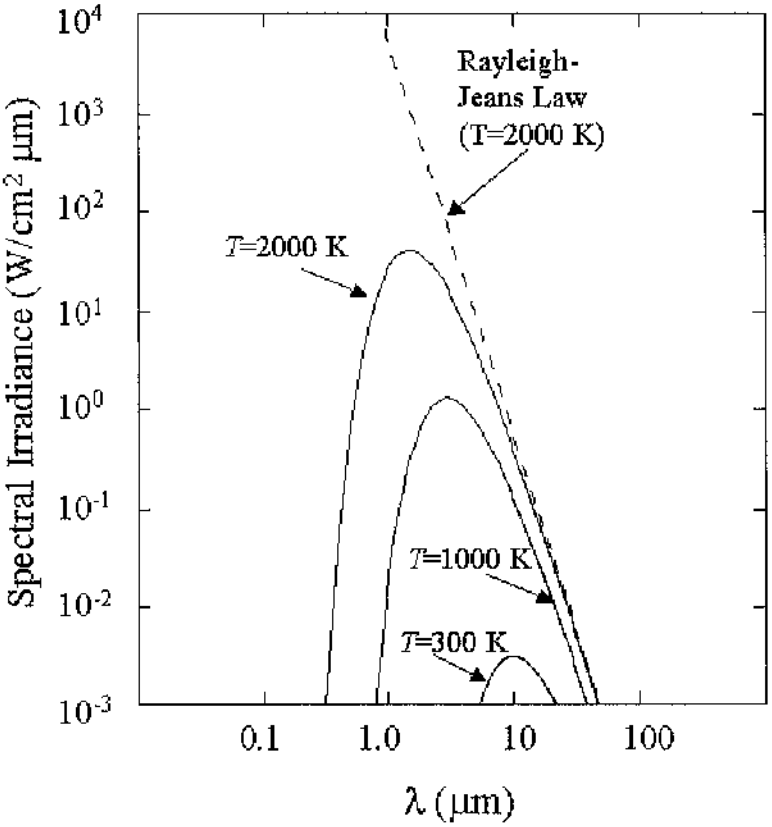


Fig. 2.3. Spectral irradiance of a blackbody at different temperatures. When the temperature is at or below room temperature, the radiation is mostly in the infrared spectral region, undetectable by the human eye. When the temperature is raised, the emission power increases and its peak shifts toward 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.

2.3. 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, 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. 2.4(a) as a spherical surface with radius a_o.

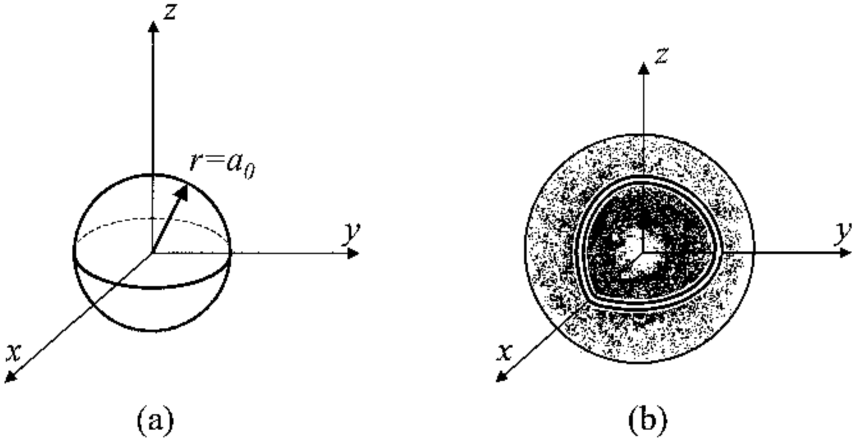


Fig. 2.4. (a) The precise spherical orbit of an electron in the first Bohr orbit, for which the radius is a_0=0.52917 Å. as calculated by Bohr's model. (b) The electron probability density pattern for the comparable atomic orbital using a quantum mechanical model. The darker areas indicate a higher probability 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.

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. 2.4(b). 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. 2.5 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.

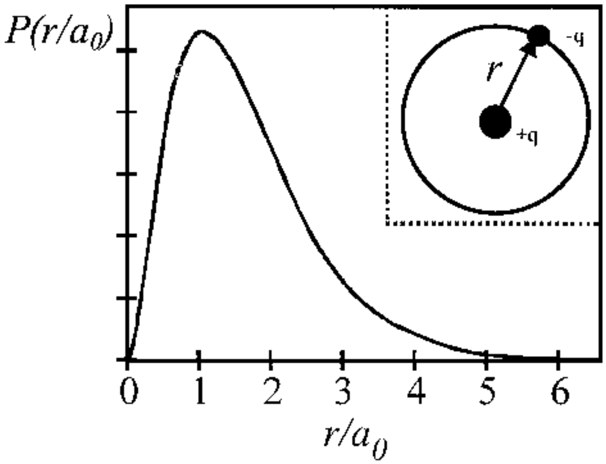


Fig. 2.5. The electron radial probability density function $P(r/a_0)$, which describes the probability of finding an electron in a spherical surface at a distance r from the nucleus in the hydrogen atom (for n=1). This probability has a maximum value when the electron is at a distance equal to the Bohr radius: $r=a_0$.

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. (2.6.). 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 2.1 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.

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 a 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.

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

Table 2.1. 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.

2.4. Structures of atoms with many electrons

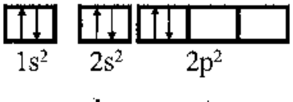
In multi-electron atoms, the energy of an electron depends on the orbital principal quantum number n and the orbital momentum quantum number l, i.e. whether the electron is in an *s*, *p*, *d*, or *f* state. The different m_l quantum numbers for a fixed set of n and l are degenerate (they have the same 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.

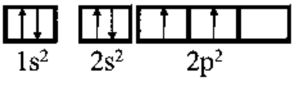
Example

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, then the total orbital momentum is maximized ($\sum l$). Determine the electronic configuration, including the spin, of the carbon atom, which has 6 electrons in its ground state.

A: Carbon has 6 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.

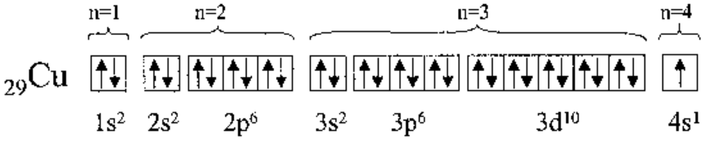
Examples of the ground state electron configurations in a number of elements are shown below. The sequence for *Z*=1 to *Z*=18 is built in a straightforward and logical manner, by filling the allowed *s*, *p*, *d*... orbitals successively (i.e. in this order). For *Z*=19, the first deviation to this procedure occurs: the 4*s* orbitals are filled with electrons *before* the 3*d* orbitals. Elements in the periodic table with partially filled 3*d* orbitals are usually transition metals and the electrons in these 3*d* 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	Li	$1s^22s^1$
Z=4	Be	$1s^22s^2$
Z=5	B	$1s^22s^22p^1$
Z=6	C	$1s^22s^22p^2$
Z=7	N	$1s^22s^22p^3$
Z=8	O	$1s^22s^22p^4$
Z=9	F	$1s^22s^22p^5$
Z=10	Ne	$1s^22s^22p^6$
...	...	...
Z=31	Ga	$1s^22s^22p^63\mathrm{s}^23\mathrm{p}^64s^23d^{10}4p^1$

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: $[\mathrm{Ar}]=1s^22s^22p^63s^23p^6$. The remaining 11 electrons must be distributed inside the 3*d* and 4*s* orbitals. Suppose that the two possible configurations are $[\mathrm{Ar}]3d^94s^2$ and $[\mathrm{Ar}]3d^{10}4s^1$. 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 has exhibits the second electronic configuration: $[\mathrm{Ar}]3d^{10}4s^1$ or $1s^22s^22p^63s^23p^63d^{10}4s^1$ 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 2.2. 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.

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

Table 2.2. The full nuclear charge Z and effective nuclear charge Z^* for the first ten elements.

Let us consider an example of electronic configuration in the multi-electron atom of Si. As shown in Fig. 2.6, ten of the fourteen Si-atom electrons (two in the 1*s* orbital, two in the 2*s* orbital, and six 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 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. 2.6, the four valence electrons occupy four of the eight allowed states belonging to the 3*s* and 3*p* 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.

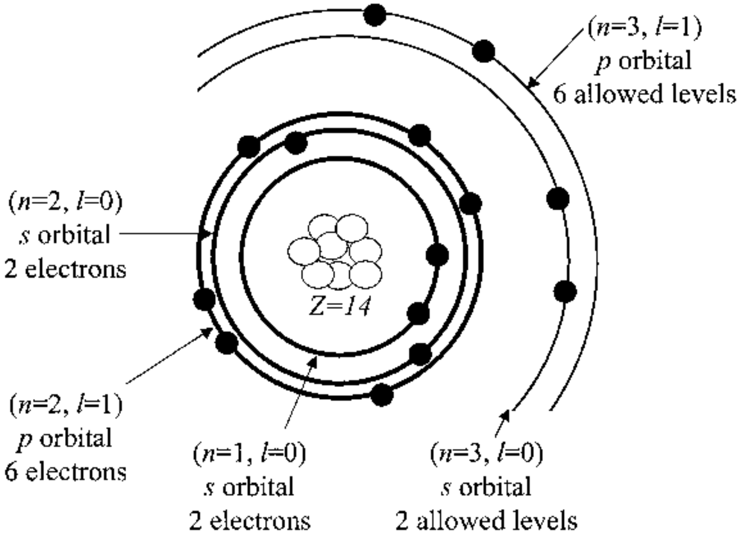


Fig. 2.6. Electron configuration for electrons in a Si atom. The ten electrons in the core orbitals, 1s (n=1), 2s (n=2, l=0), and 2p (n=2, l=1) are tightly bound to the nucleus. The remaining four electrons in the 3s (n=3, l=0) and 3p (n=3, l=1) orbitals are weakly bound.

2.5. Bonds in solids

2.5.1. 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 electron-to-electron shell repulsion, called core repulsion, that dominates. Fig. 2.7 shows the inter-atomic 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 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. 2.7).

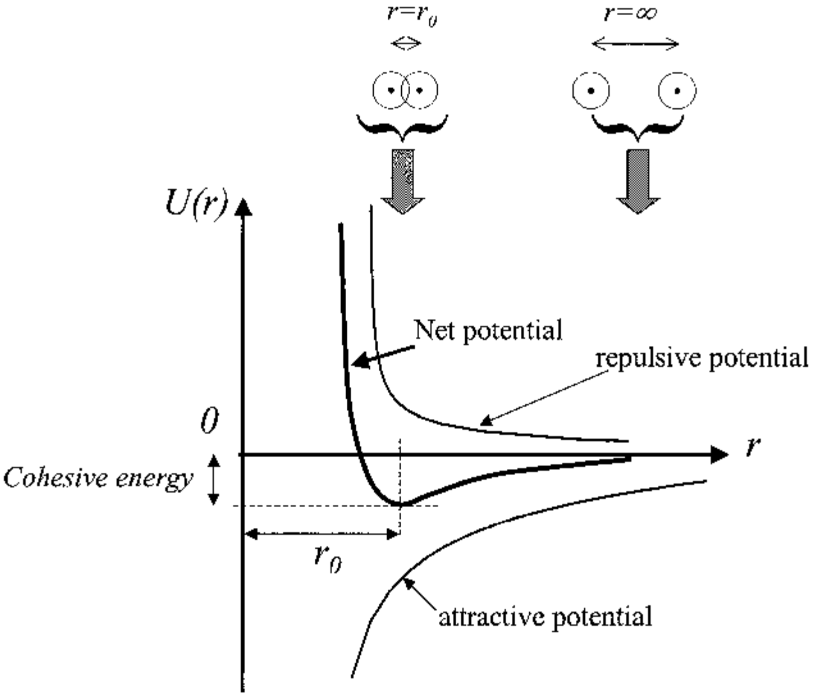


Fig. 2.7. Potential energy versus inter-atomic separation r. The net potential is the sum of repulsive and attractive components. The minimum of the net potential corresponds to the equilibrium distance r_0 between the two atoms.

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. 2.7. 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 $a=1.54\sigma$.

2.5.2. 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. 2.8 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. 2.9.

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. 2.7 can be estimated from the sum of the coulombic potential energies between the ions (see Problem 11).

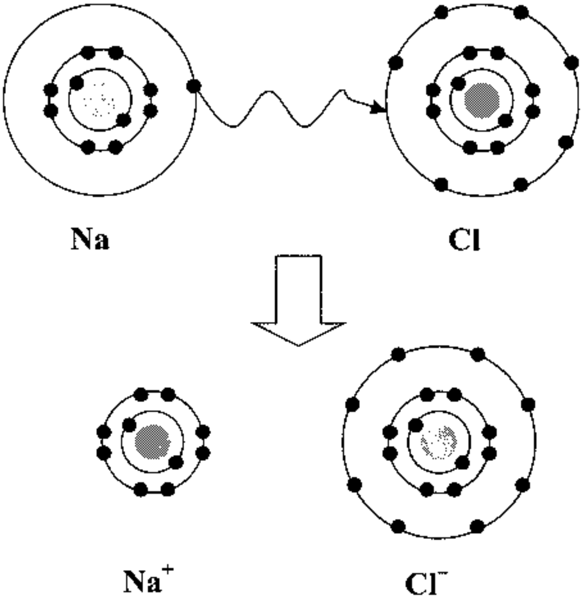


Fig. 2.8. Schematic illustration of the formation of an ionic bond in NaCl, showing the electron transfer between the two elements and their final electronic configurations.

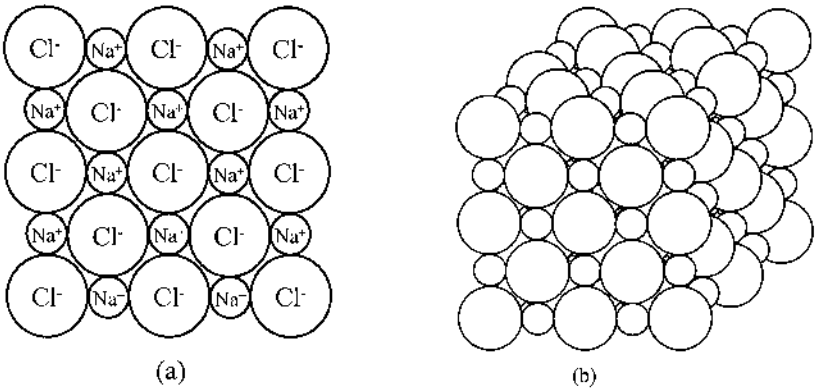

Cl^- Na^+ Cl^- Na^+ Cl^-
Na^+ Cl^- Na^+ Cl^- Na^+
Cl^- Na^+ Cl^- Na^+ Cl^-
Na^+ Cl^- Na^+ Cl^- Na^+
Cl^- Na^- Cl^- Na^- Cl^-
(a)
(b)

Fig. 2.9. (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. 1.31, one can see that one Cs^+ ion (at the center of the cube) has 8 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: $E = -8\frac{q^2}{4\pi\varepsilon_0 r_{nn}}$ =-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.

2.5.3. 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. 2.7, 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. Fig. 2.10 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 4 valence electrons with its neighbors as shown in Fig. 2.10. 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. 2.11. The coordination number (CN) is the number of nearest neighbors for a given atom in the solid. As it is seen in Fig. 2.11, the coordination number for a carbon atom in the diamond crystal structure is four, as discussed in Chapter 1.

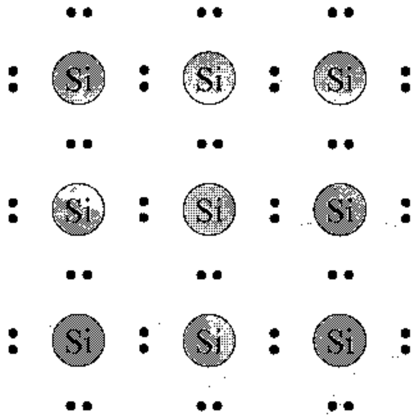


Fig. 2.10. Schematic of covalent bonds in Si. Each Si atom contributes one of its 4 outer shell electrons with each neighboring Si atom. This creates a pair of shared electrons between two Si atoms, which constitutes the covalent bond. Because the two atoms arę identical, the electrons have the highest probability of being located at equal distances between the two atoms, as illustrated here.

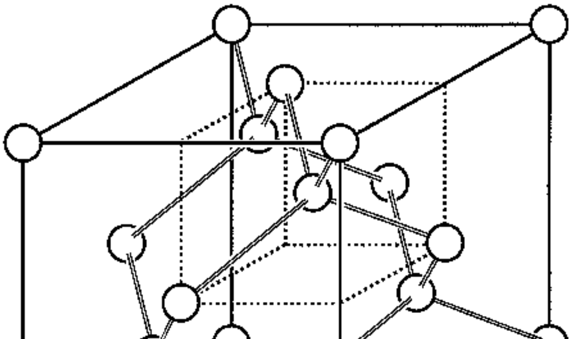


Fig. 2.11. The diamond crystal with covalent bonds. The diamond crystal is most often represented using a cubic unit cell, as shown here. Each atom in the structure is covalently bonded to four neighboring atoms.

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 (2*s* shell in C and 3*s* shell in Si for example) into the doubly occupied *p*-shell. This costs energy, but this energy is more than 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 4-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.

2.5.4. 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 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 Ill-V compounds (e.g., GaAs, InSb, and so on), have polar covalent bonds. In GaAs, for example, the electrons in a covalent bond arc closer to (i.e. more probably found near) the As ion core than the Ga ion core. This example is shown in Fig. 2.12.

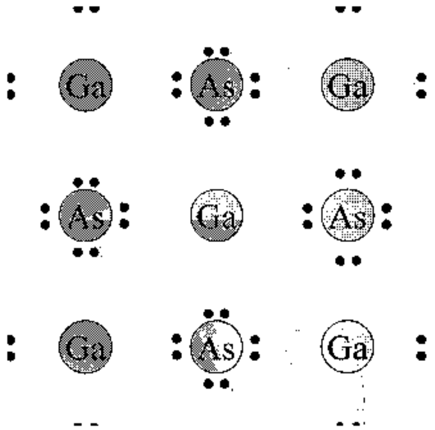


Fig. 2.12. Polar bonds in an III-V intermetallic compound. Similar to the case of Si in Fig. 2.10, a covalent bond is formed by the sharing of an electron from a Ga atom and one from a neighboring As atom. However, because a Ga atom has only 3 electrons in its outer shell while an As atom has 5, one of the four covalent bond 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.

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.

2.5.5. 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. 2.13. 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. 2.7 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 in many cases to "pull apart" metal regions, which explains why metals are usually malleable.

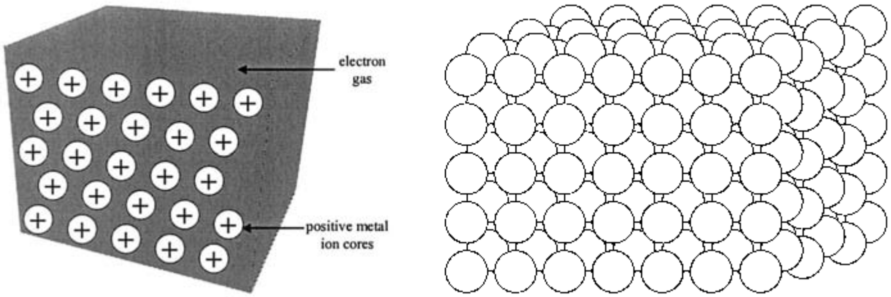


Fig. 2.13. Metallic bonding resulting from the attraction between the electron gas and the positive metal ions. The electrons are delocalized inside the volume between the atoms in the crystal.

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.

2.5.6. 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 involve 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. 2.14(a). 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 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 τ 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. 2.14(b) and (c). In a uniform field, the torque τ 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 non-uniform 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 $\vec{E'}(\vec{r})$ of its own around it as shown in Fig. 2.14(d), just as a single charge does. Therefore, a dipole can interact with another dipole as shown in Fig. 2.14(e). 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 $1/R^6$. Fig. 2.7 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 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.

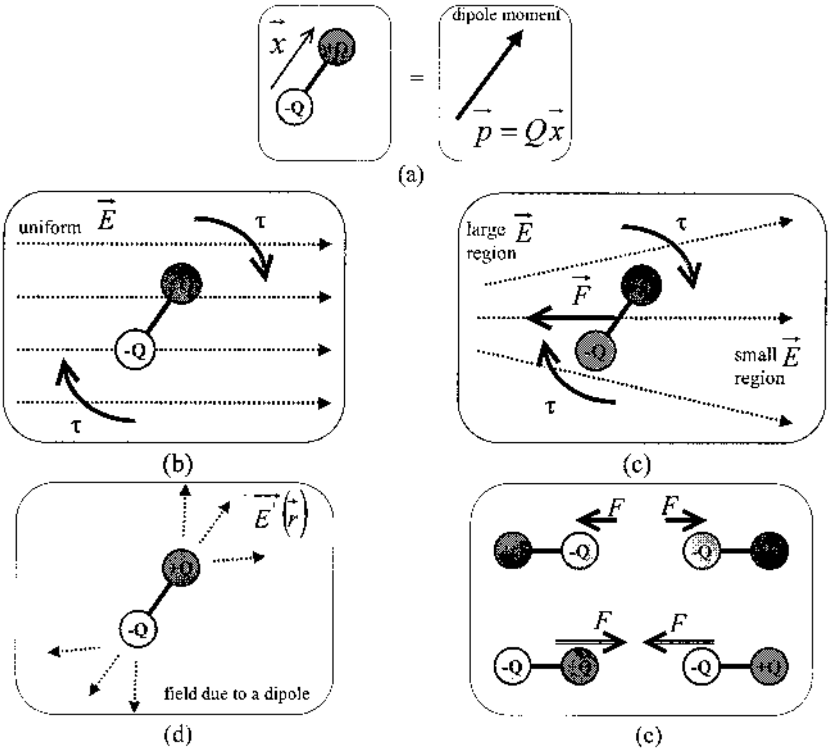


Fig. 2.14. Electric dipole moment and its properties. (a) A dipole is formed when two electrical charges with opposite signs and equal magnitude are separated by a distance. This creates a dipole moment. (b), (c) A dipole can rotate and be translated in the presence of an electric field. (d) A dipole 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.

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. Fig. 2.15 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.

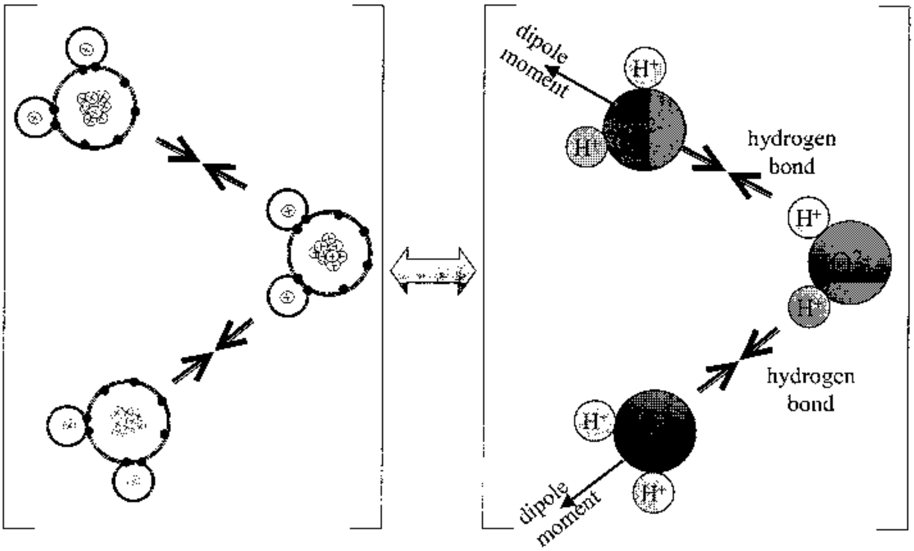


Fig. 2.15. The origin of hydrogen bonding between water molecules. A H_2O molecule has a net permanent dipole moment as a result of its lack of central symmetry. The H_2O molecules can therefore interact with one another. Attractions between the various dipole moments in water give rise to hydrogen bonding.

2.6. Atomic property trends in the periodic table

2.6.1. 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 periodic table shown in Fig. 2.16 can be divided into 3 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.

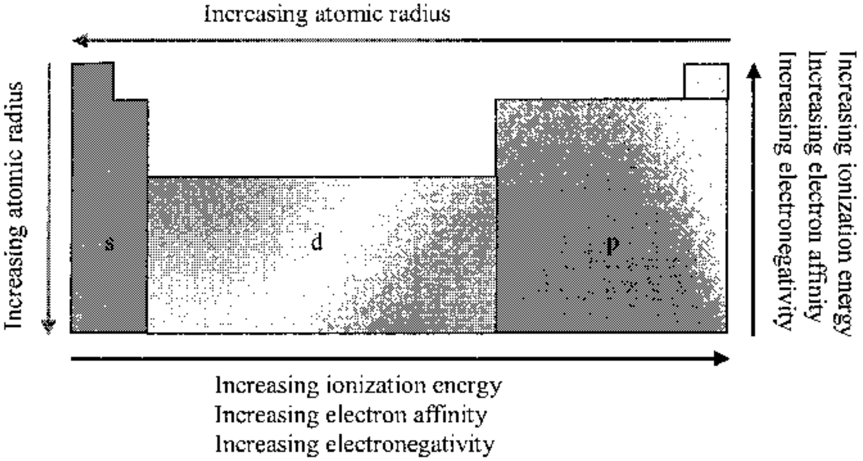

Increasing atomic radius
Increasing atomic radius
s
d
p
Increasing ionization energy
Increasing electron affinity
Increasing electronegativity
Increasing ionization energy
Increasing electron affinity
Increasing electronegativity

Fig. 2.16. Part of the periodic table with divisions indicating valence shells and a summary of atomic property trends.

2.6.2. 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.

2.6.3. 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.

2.6.4. 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 towards 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 particular 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.

2.6.5. 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.

2.6.6. Summary of trends

The different trends are summarized in Fig. 2.16. 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).

2.7. 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 Chapter 4 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 ten of its fourteen 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. 2.17(b). Exactly half of the allowed states become depressed in energy (bonding states) and half increase in energy (anti bonding 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. 2.17(b), 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.

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 pre-selected crystalline direction (x) is as shown in Fig. 2.17(a). This plot is the basic energy band model. E_C introduced in Fig. 2.17(a) 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 Chapter 4.

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.

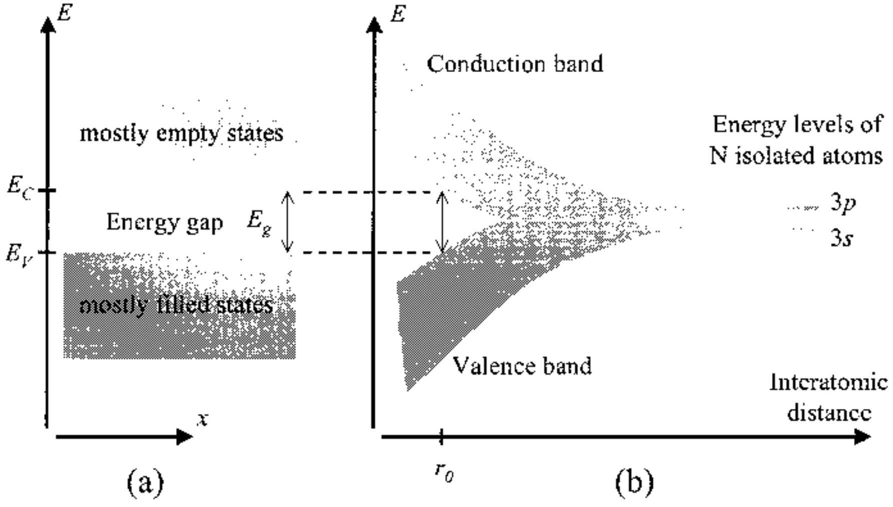


Fig. 2.17. Illustration of the formation of energy bands in a Si crystal. A system of N isolated Si atoms has discrete allowed energy levels, all located at the energies of the 3s and 3p orbitals of an isolated Si atom. When the atoms come into close proximity, the energy levels are modified as shown in the figure, as a result of the interaction between the atoms. The allowed energy levels start to form energy bands.

2.8. 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.

Further reading

Atkins, P.W., *Molecular Quantum Mechanics*, Oxford University Press, New York, 1983.

Cohen, M.M., *Introduction to the Quantum Theory of Semiconductors*, Gordon and Breach, New York, 1972.

Ferry, D.K., *Semiconductors*, Macmillan, New York, 1991.

Kasap, S.O., *Principles of Engineering Materials and Devices*, McGraw-Hill, New York, 1997.

Kittel, C., *Introduction to Solid State Physics*, John Wiley & Sons, New York, 1976.

Pierret, R.F., *Semiconductor Fundamentals*, Addison-Wesley, Reading, MA, 1989.

Pierret, R.F., *Advanced Semiconductor Fundamentals*, Addison-Wesley, Reading, MA, 1989.

Yu, P.Y. and Cardona, M., *Fundamentals of Semiconductors: Physics and Materials Properties*, Springer, New York, 1999.

Ziman, J.M., *Principles of the Theory of Solids*, Cambridge University Press, Cambridge, 1998.

Zumdahl S.S. and Zumdahl S.A., *Chemistry*, Houghton Mifflin Company, Boston, 2003.

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 Lymann 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 2 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 $I(\lambda) = \frac{2\pi hc^2}{\lambda^5}\{\frac{1}{e^{\frac{hc}{\lambda k_b T}} - 1}\}$ 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, Boltzman'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:

$$E(r) = -\frac{4.03\times10^{-28}}{r} + \frac{6.97\times10^{-96}}{r^8}$$

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 $A=1.037\times10^{-77}$ J.m^6 and $B=1.616\times10^{-134}$ 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 your answer.

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.11 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 right-most group) have electron affinity values listed in Fig. A.12 in Appendix A.3?

3. Introduction to Quantum Mechanics

3.1. The quantum concepts

In Chapter 2 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 electromagnetic field, 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.

3.1.1. Blackbody radiation

As introduced in Chapter 2, a blackbody is an ideal source of electromagnetic radiation and the radiated power dependence was depicted as a function of wavelength in Fig. 2.3 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. 2.3. Several attempts to explain this observed blackbody spectrum were made using classical mechanics in the latter half of the 19th 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. They have a *continuous* 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 $k_bT/2$, where k_b is the Boltzmann constant ($=8.614\times10^{-5}$ 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 $\dfrac{8\pi\upsilon^2}{c^3}$. Therefore, the distribution of energy radiated by a blackbody per unit volume and per unit frequency interval is $u(\upsilon,T)=\dfrac{8\pi\upsilon^2}{c^3}k_bT$. Considering that this energy is radiated at the velocity 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: $w(\lambda,T)=\dfrac{8\pi c}{\lambda^4}k_bT$. 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. 2.3 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 that the classical theorem 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 υ in a material could only radiate or absorb energy in discrete or *quantized* packages proportional to the frequency:

Eq. (3.1) $E_n = nh\upsilon = n\hbar\omega \; n=0, 1, 2, \ldots$

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 υ did not vary continuously (with an average value k_bT), but was an integral multiple of $h\upsilon$. 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/k_bT} and can be expressed as:

$$P(E_n) = \frac{e^{-E_n/k_bT}}{\sum_E e^{-E/k_bT}}$$

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. (3.1), the average energy $<E>$ associated with each frequency mode υ can thus be written as:

$$\langle E\rangle = \sum_{E_n} E_n P(E_n) = \frac{\sum_{n=0}^{\infty}(nh\nu)e^{-nh\nu/k_bT}}{\sum_{n=0}^{\infty}e^{-nh\nu/k_bT}} = \frac{h\nu}{e^{h\nu/k_bT}-1}$$

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

Eq. (3.2) $$u(\nu,T) = \frac{8\pi\upsilon^2}{c^3}\frac{h\upsilon}{e^{h\nu/k_bT}-1}$$

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:

$$e^{h\upsilon/k_bT}-1 \approx {h\upsilon}/{k_bT}$$

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.

3.1.2. 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. 3.1. 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 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.

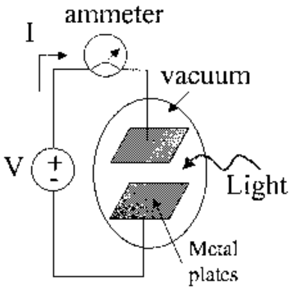


Fig. 3.1. Simplified experimental setup used by Lenard. A chamber in vacuum contains two parallel metal plates on which a voltage is applied. Light shining onto a metal plate gives enough energy to the electrons of the plate to make them leave the plate and be accelerated by the electric field.

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. 3.2. 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.

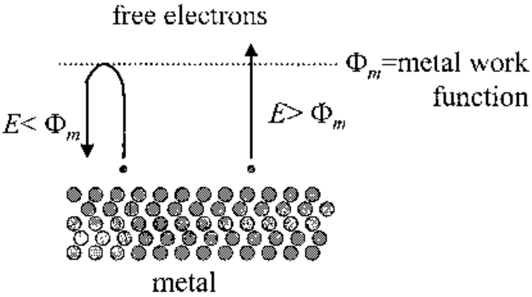


Fig. 3.2. The work function of a metal, denoted Φ_m, is the minimum amount of energy that an electron needs to acquire to leave the metal.

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:

$$\begin{aligned}\Phi_m &= \frac{hc}{\lambda} - qV_0 \\ &= \frac{(6.62617\times10^{-34})(2.99792\times10^{8})}{2263\times10^{-10}} - (1.60218\times10^{-19})(3) \\ &= 3.97\times10^{-19}\,J \\ &= 2.48\,eV\end{aligned}$$

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 thousand-fold. 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 $h\upsilon$ where h is Planck's constant and υ 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. $E=h\upsilon$. Therefore, if the frequency ν was too low, such that $E=h\upsilon$ 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. 3.3. However, if the frequency was high enough, such that $E=h\upsilon$ 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 ten 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!

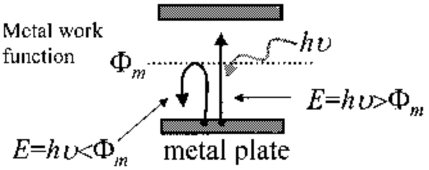


Fig. 3.3. Schematic diagram of the escape mechanism of an electron in the metal plate receiving a photon with energy hυ. If the photon energy is lower than the work function, the electron does not escape. If the photon energy is higher than the work function, the electron receives enough energy to reach the vacuum level and leave the metal.

3.1.3. Wave-particle duality

The previous discussions on the Bohr atom in Chapter 2, 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:

Eq. (3.3) $$\lambda = \frac{h}{p}$$

This postulate establishes the relationship between a particle and a wave in nature. This concept as well as other ones introduced in the previous examples clearly prove 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.

3.1.4. 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. 3.4. 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.

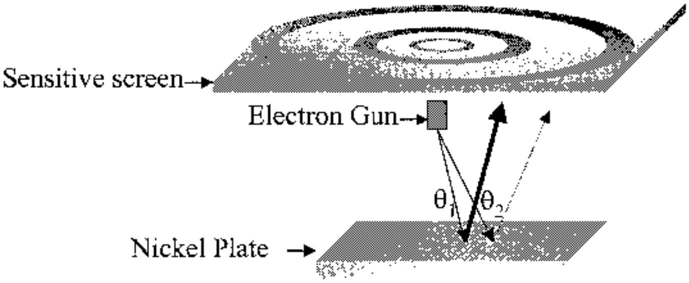


Fig. 3.4. Schematic of the experimental setup in the Davisson-Germer experiment. A beam of electrons is directed on a nickel plate from which the electrons are reflected. They then hit a sensitive screen and create a ring pattern.

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. 3.5(a). The darkest areas corresponded to the situations in which 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. 3.5(b). By quantifying the positions of the rings, Davisson and Germer were able to confirm the de Broglie relation given in Eq. (3.3).

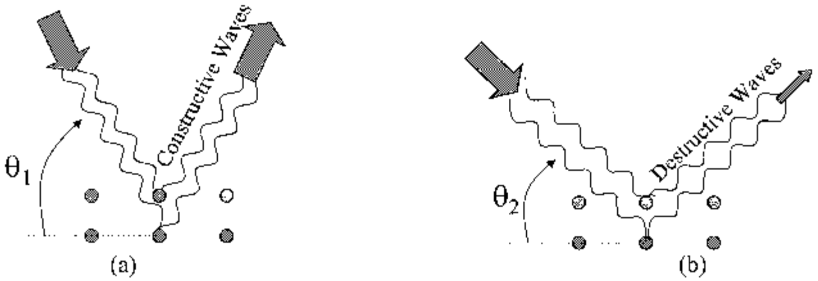


Fig. 3.5. (a) Constructive diffraction and (b) destructive diffraction condition for the waves reflected from a crystal surface. In the constructive diffraction situation, $2d\sin(\theta) = n\lambda$ *where d is the distance between two planes,* λ, θ *are wavelength and angle to the normal respectively, n is an integer, the waves are in phase whereas in the destructive diffraction configuration, the waves have opposite phases.*

3.2. 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.

3.2.1. 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 1920's 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).

Eq. (3.4) $$i\hbar\frac{\partial\Psi(x,y,z,t)}{\partial t}=H\Psi(x,y,z,t)$$

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:

Eq. (3.5) $$H=-\frac{\hbar^2}{2m}\nabla^2+U(x,y,z,t)$$

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. The Hamiltonian H describes the temporal and spatial evolution of the system (wavefunction).

The next principle of Quantum Mechanics is that one cannot know more about the system then the totality of all its wavefunction solutions of the Schrödinger equation. Having solved the SE and found the wavefunctions, we have the following properties:

(3) 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 $|\Psi(x,y,z,t)|^2 dxdydz$.

(4) The sum of the probabilities of all possible values of the spatial coordinates of the system must be, by definition, equal to unity:

Eq. (3.6) $$\int|\Psi(x,y,z,t)|^2 dxdydz=1$$

This equation is the normalization condition for the wavefunction.

(5) If $\Psi_1(x,y,z,t)$ and $\Psi_2(x,y,z,t)$ are two wavefunctions of the system, solutions of the SE, then the linear combination $c_1\Psi_1(x,y,z,t)+c_2\Psi_2(x,y,z,t)$ is also a wavefunction of the same system. This statement constitutes the so-called principle of superposition of states, the main principle of quantum mechanics.

(6) For any physical quantity, one can associate an operator f which "acts" on a wavefunction, i.e. differentiates, integrates or simply multiplies it with another function. It represents a physical observable. An operator has a set or spectrum of eigenvalues which correspond to the possible values that the associated physical quantity can take. Thus the operator f acting upon 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:

Eq. (3.7) $$f\Psi_f(x,y,z,t) = f_f\Psi_f(x,y,z,t)$$

then Ψ_f is an eigenfunction of f and f_f, its corresponding eigenvalue.

Every physical observable has a set of eigenfunctions and corresponding eigenvalues. However it is possible for different physical observables to share the same eigenfunctions. Eigenfunctions belonging to different eigenvalues are orthogonal, thus their inner product is equal to 1 when the wavefunctions belong to the same eigenvalue, and 0 otherwise, or mathematically:

Eq. (3.8) $$\int\Psi_{f_1}(x,y,z)^*\Psi_{f_{12}}(x,y,z)dxdydz = \delta_{f_1,f_2}$$

where δ_{f_1,f_2} is called the Kronecker δ function, and is defined as:

Eq. (3.9) $$\delta_{f_1,f_2} = \begin{cases} 1 & if \quad f_1 = f_2 \\ 0 & if \quad f_1 \neq f_2 \end{cases}$$

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.

Eq. (3.10) $$\chi(x,y,z,t) = \sum_f a_f\Psi_f(x,y,z,t)$$

(7) In classical mechanics, all physical quantities can have a continuous range of values. By contrast, in quantum mechanics, there exist quantities both with a continuous spectrum (for instance the coordinates) and others with a discrete spectrum of eigenvalues (for instance energy).

(8) A system need not be in a pure state, 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. This leads us to the next definition.

(9) The mean value or expectation value of a physical quantity represented by an operator f is what is measured experimentally, is denoted $< f >$ and is given by:

Eq. (3.11) $$< f >= \int \Psi(x,y,z,t)^* f \Psi(x,y,z,t) dxdydz$$

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

Eq. (3.12) $$\Psi(x,y,z,t) = c_{f_1}\Psi_{f1}(x,y,z,t) + c_{f_2}\Psi_{f_2}(x,y,z,t)$$

the expectation value $< f >$ is given by:

Eq. (3.13) $$< f >= \left|c_{f_1}\right|^2 f_1 + \left|c_{f_2}\right|^2 f_2$$

which signifies that the expectation value is the sum of the eigenvalues multiplied by the probability of being in that particular eigenstate.

Examples of physical quantities, their associated operators and expectation values are given in Table 3.1.

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)	$i\hbar\frac{\partial}{\partial t}$	$< E >= \int \Psi^* i\hbar \frac{\partial \Psi}{\partial t} dxdydz$

Table 3.1. Examples of common physical quantities and their associated operators.

(10) Operators which are related to 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:

$$f_{ij} = \int d\vec{r} \Phi_i^* f \Phi_f$$

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

(11) Let us now take a closer look at the SE and the Hamiltonian of the system. 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:

Eq. (3.14) $$\nabla^2 \Psi(x,y,z) = \frac{\partial^2 \Psi(x,y,z)}{\partial x^2} + \frac{\partial^2 \Psi(x,y,z)}{\partial y^2} + \frac{\partial^2 \Psi(x,y,z)}{\partial z^2}$$

U(x,y,z,t) is the potential energy of the system considered, $\hbar$ is the reduced Planck's constant, and i is the complex number such that $i^2 = -1$. Solving this equation fully determines the wavefunctions, and eigenstates of energy, of the physical system under consideration.

3.2.2. The time independent Schrödinger equation

A particular and important case for the Schrödinger equation, is that for a closed system in a time-independent external field. Then, the right hand side of Eq. (3.4) 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). This means that $\langle E\rangle = E$ is constant. From Eq. (3.11) using the operator in Table 3.1, we get:

$$E =< E >= \int \Psi^* i\hbar \frac{\partial \Psi}{\partial t} dxdydz$$

By identification, we find that the following relation must be satisfied:

Eq. (3.15) $i\hbar \frac{\partial \Psi(x,y,z,t)}{\partial t} = E\Psi(x,y,z,t)$

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

Eq. (3.16) $\Psi(x,y,z,t) = \varphi(x,y,z).\exp(-\frac{i}{\hbar}Et)$

Inserting this expression into the Schrödinger equation in Eq. (3.4) and eliminating the exponential term on both sides of the equation, we obtain:

Eq. (3.17) $E\varphi(x,y,z) = -\frac{\hbar^2}{2m}\nabla^2\varphi(x,y,z) + U(x,y,z)\varphi(x,y,z)$

which can be rewritten as:

Eq. (3.18) $\frac{\hbar^2}{2m}\nabla^2\varphi(x,y,z) + [E - U(x,y,z)]\varphi(x,y,z) = 0$

This last expression is called the time-independent Schrödinger equation. From now, and in the rest of the Chapter, we will limit our discussion only to such types of situation. However, we will continue to use the symbol Ψ to denote the wavefunction of the system considered.

A system in a stationary state will always have well defined eigenfunctions Ψ_n and eigenvalues of energy E_n, but these need not be simultaneous eigenstates of momentum or angular momentum for example. For example, a system will only have well defined energy and momentum, if the Hamiltonian or total energy operator is time independent (conservation of energy), and also translationally invariant over any distance in space (conservation of momentum). Thus for example in a random or disordered medium, the Hamiltonian varies in space and one can have conservation of energy, and therefore well defined energy levels, but not well defined momentum or angular momentum eigenstates and eigenvalues.

3.2.3. 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 is therefore that:

Eq. (3.19) $\Delta E \Delta t \sim \hbar$

In other words if one knows the energy E to within an accuracy ΔE, then one cannot know the time to an accuracy better than Δt as given by Eq. (3.19). The same is true the other way around. If one knows the time to within an accuracy of Δt then one cannot know the energy to a greater accuracy than ΔE as given by the above relation. Let us apply this concept to a stationary state of the system in energy, where we know the energy level of the particle with absolute accuracy. The meaning of Eq. (3.19) is that in this case, we can say nothing about the time. Indeed the time dependence of the wavefunction as shown by Eq. (3.16) only occurs in the phase of the complex exponential, which has no consequence on the probability distribution for example. Indeed, 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 Heisenberg uncertainty principle 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:

Eq. (3.20) $\Delta p_\mu \Delta r_\mu \sim \hbar$

where p_μ are x, y, z components of momentum and r_μ of 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.

3.2.4. First summary

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 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 A is given by the generalized form of the Boltzmann distribution:

$$< A >= \frac{\sum_n e^{-E_n / k_b T} A_{nn}}{\sum_n e^{-E_n / k_b T}}$$

which involves the expectation values A_{nn} of the operator A 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 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 or eigenstate, then the outcome of this operation or act of measurement, is the corresponding eigenvalue. In general however the system is in superposition of eigenstates, and the outcome of the measurement is the weighted superposition as given by Eq. (3.13). This superposition, or thermal average, involves the expectation value of the operator A, given by A_{nn} in all eigenstates labeled by n.

3.2.5. 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 interested reader is referred to more advanced quantum mechanics textbooks for further information.

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.

3.3. Simple quantum mechanical systems

3.3.1. 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. (3.18) can then be simplified to:

Eq. (3.21) $$\frac{\hbar^2}{2m}\frac{d^2\Psi(x)}{dx^2} + E\Psi(x) = 0$$

The solution of Eq. (3.21) which is an eigenstate of both energy and momentum is:

Eq. (3.22) $$\Psi(x) = Ae^{ikx}$$

where *A* is a constant, and $k = \frac{2\pi}{\lambda}$ is the wavenumber. By applying the *x*-momentum operator on the RHS of Eq. (3.22), 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. (3.21), one obtains:

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

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

Eq. (3.23) $$E=\frac{\hbar^2 k^2}{2m}=\frac{h^2 k^2}{8\pi^2 m}$$

or conversely:

Eq. (3.24) $$k=\sqrt{\frac{2mE}{\hbar^2}}$$

and is plotted in Fig. 3.6. The expectation of particle momentum, as defined by Eq. (3.11), can be expressed in quantum mechanics as:

Eq. (3.25) $$\langle p\rangle=\hbar k$$

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. Rather than dealing with an infinite system which is not normalizable, we can think of the system as being very large but of finite size [0,L], as L becomes infinite, so that the normalization constant A is given by $A=\sqrt{\frac{1}{L}}$

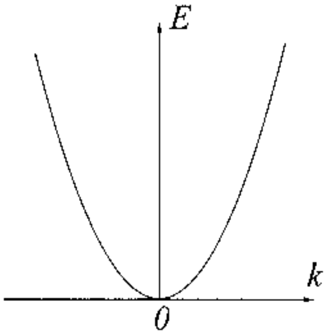


Fig. 3.6. The energy-momentum relationship for a free particle has a parabolic shape.

3.3.2. Particle in a 1D 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. 3.7.

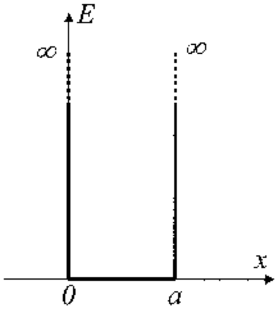


Fig. 3.7. Potential energy corresponding to the 1D box.

This potential can be mathematically expressed such that:

Eq. (3.26) $$\begin{cases} U(x) = \infty & \text{for } x < 0 \text{ and } x > a \\ U(x) = 0 & \text{for } 0 < x < a \end{cases}$$

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

Eq. (3.27) $$\begin{cases} \Psi(x) = 0 & \textit{for } x < 0 \textit{ and } x > a \\ \dfrac{\hbar^2}{2m}\dfrac{d^2\Psi(x)}{dx^2} + E\Psi(x) = 0 & \textit{for } 0 < x < a \end{cases}$$

Inside the box $\Psi(x)$ can be written as the sum of *sin* and *cos* functions so that:

Eq. (3.28) $$\Psi(x) = A\sin(kx) + B\cos(kx)$$

but with the boundary conditions:

Eq. (3.29) $$\Psi(0) = \Psi(a) = 0$$

Expressing these conditions using Eq. (3.28), we get:

Eq. (3.30) $$\begin{cases} B = 0 \\ A\sin(ka) = 0 \end{cases}$$

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

$\sin(ka) = 0$ or $k = k_n = n\dfrac{\pi}{a}$ where n is an integer equal to ±1, ±2, …

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. (3.27), we can see that the energy of a particle in a 1D box is also quantized:

Eq. (3.31) $$E_n = n^2\frac{\hbar^2\pi^2}{2ma^2}$$

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. (3.28) and Eq. (3.30), we can write the wavefunction as:

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

The value of A can be computed by substituting this expression into the normalization condition expressed by Eq. (3.6). One easily finds that:

Eq. (3.32) $$A = \sqrt{\frac{2}{a}}$$

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

Eq. (3.33) $$\Psi_n(x) = \sqrt{\frac{2}{a}} \sin\left(\frac{n\pi x}{a}\right)$$

These functions consist of standing waves as depicted in Fig. 3.8(b). One can think of the particle in a 1D box as bouncing off the walls of the box and the probability of finding a particle at x in the box is shown in Fig. 3.8(c).

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:

$$E_n = n^2 \frac{\left(1.05458\times10^{-34}\right)^2\pi^2}{2\left(0.91095\times10^{-30}\right)\left(25\times10^{-10}\right)^2}$$

$$= 9.63n^2 \times 10^{-21}\,J$$

$$= 0.060n^2\ eV$$

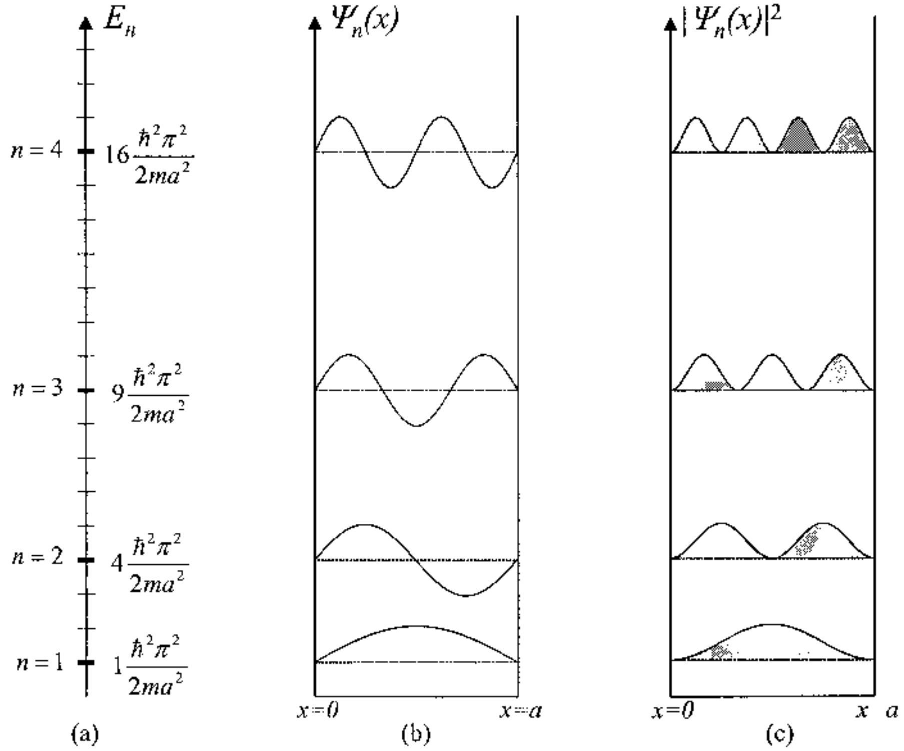

E_n
$n = 4$
$16\frac{\hbar^2\pi^2}{2ma^2}$
$n = 3$
$9\frac{\hbar^2\pi^2}{2ma^2}$
$n = 2$
$4\frac{\hbar^2\pi^2}{2ma^2}$
$n = 1$
$1\frac{\hbar^2\pi^2}{2ma^2}$
(a)
$\Psi_n(x)$
x=0
x=a
(b)
$|\Psi_n(x)|^2$
x=0
x a
(c)

Fig. 3.8. (a) Energy levels, (b) wavefunctions Ψ(x), and (c) | Ψ(x)|^2 which is proportional to the probability of finding a particle at a position x in a 1D quantum box, for the first four allowed levels.

3.3.3. 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. 3.9 and mathematically expressed as:

Eq. (3.34)
$$\begin{cases} U(x) = U_0 > 0 & \text{for } x < 0 \text{ and } x > a \\ U(x) = 0 & \text{for } 0 < x < a \end{cases}$$

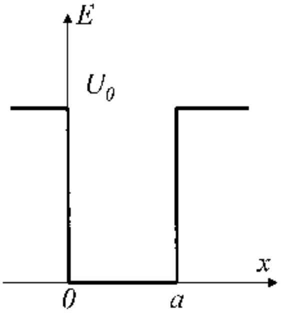


Fig. 3.9. Potential energy in a finite potential well.

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

Eq. (3.35)

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

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. (3.35) can be rewritten as:

Eq. (3.36)

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

by defining:

Eq. (3.37)

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

The general solution to Eq. (3.36) is then:

Eq. (3.38) $$\begin{cases} \Psi_-(x) = A_- e^{\alpha x} + B_- e^{-\alpha x} & \text{for } x < 0 \\ \Psi_0(x) = A_0 \sin(kx) + B_0 \cos(kx) & \text{for } 0 < x < a \\ \Psi_+(x) = A_+ e^{\alpha x} + B_+ e^{-\alpha x} & \text{for } x > a \end{cases}$$

The boundary conditions include the finite nature of $\Psi(x)$ for $x \rightarrow \infty$ and $x \rightarrow -\infty$, 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:

Eq. (3.39) $$\begin{cases} \Psi_-(-\infty) = 0 & \Psi_+(+\infty) = 0 \\ \Psi_-(0) = \Psi_0(0) & \Psi_0(a) = \Psi_+(a) \\ \frac{d\Psi_-}{dx}(0) = \frac{d\Psi_0}{dx}(0) & \frac{d\Psi_0}{dx}(a) = \frac{d\Psi_+}{dx}(a) \end{cases}$$

Utilizing Eq. (3.38), we obtain:

Eq. (3.40)
$$\begin{cases} A_+ = B_- = 0 & \\ A_- = B_0 & A_0 \sin(ka) + B_0 \cos(ka) = B_+ e^{-\alpha a} \\ \alpha A_- = kA_0 & kA_0 \cos(ka) - kB_0 \sin(ka) = -\alpha B_+ e^{-\alpha a} \end{cases}$$

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_+:

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

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

Eq. (3.41) $$\left(k^2-\alpha^2\right)\sin(ka)-2\alpha k\cos(ka)=0$$

This condition can be rewritten into:

Eq. (3.42) $$\tan(ka)=\frac{2\alpha k}{k^2-\alpha^2}$$

By introducing the constants:

Eq. (3.43) $$\begin{cases}\alpha_0=\sqrt{\dfrac{2mU_0}{\hbar^2}}\\ \zeta=\dfrac{E}{U_0} & (0<\zeta<1)\end{cases}$$

we can first rewrite Eq. (3.37) as:

Eq. (3.44) $$\begin{cases}\alpha=\alpha_0\sqrt{1-\zeta}\\ k=\alpha_0\sqrt{\zeta}\end{cases}$$

and therefore Eq. (3.42):

Eq. (3.45) $$\tan\left(a\alpha_0\sqrt{\zeta}\right)=\frac{2\sqrt{\zeta(1-\zeta)}}{2\zeta-1}$$

The only variable in Eq. (3.45) is ζ, and any value that satisfies it leads to a value of *E*, *k*, 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. (3.45) is easiest solved graphically. For example, Fig. 3.10 shows a plot of the two functions on either side of Eq. (3.45) The intersection points correspond to values of ζ which satisfy Eq. (3.45) 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. 3.10, there are two solutions.

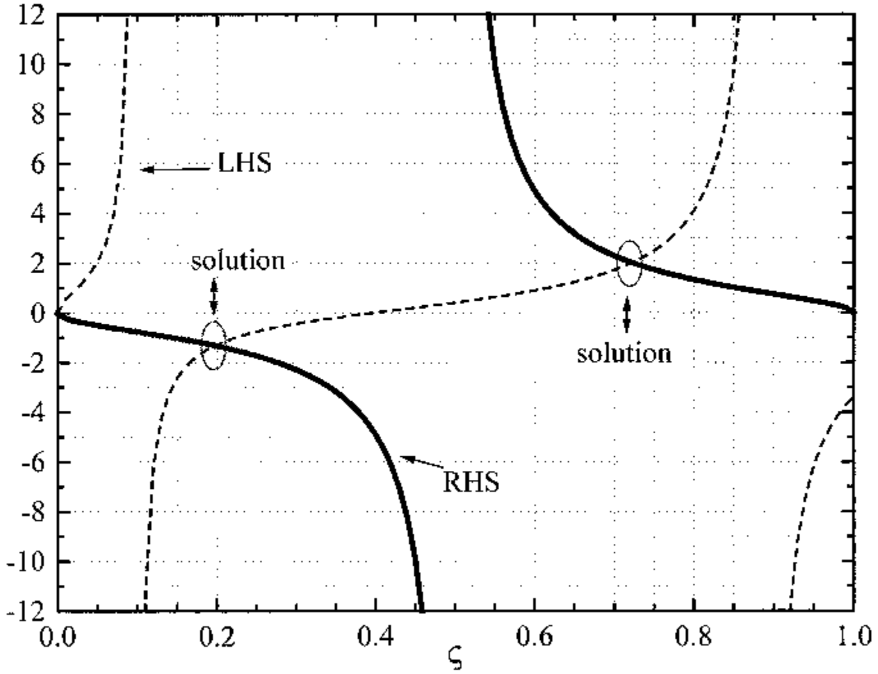


Fig. 3.10. Graphical representations of the functions on the left hand side (LHS) and right hand side (RHS) of Eq. (3.45), shown in dashed and solid lines. The intersections between these curve yield the solutions of the finite potential well problem.

As the well potential U_0 increases, α_0 increases as defined by Eq. (3.44) and thus a higher number of tangent function branches can be fitted for ζ between 0 and 1 (left hand side of Eq. (3.45)). 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. 3.11. 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 lower than those in the infinite well potential case, as shown with the dashed lines in Fig. 3.11.

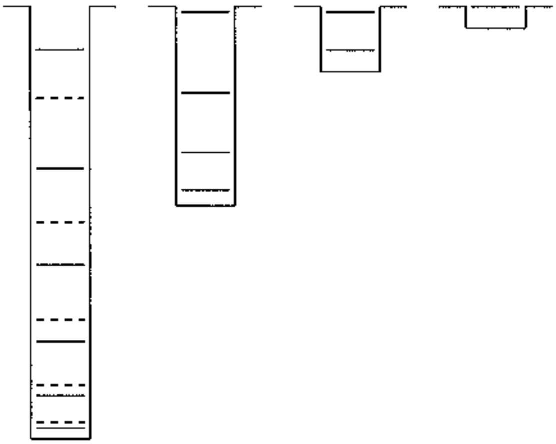


Fig. 3.11. Quantized energy levels in a finite potential well (solid lines) as a function of potential well depth. For comparison, the energy levels of the infinite well case are shown in dashed lines for the quantum well on the left.

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 non-zero wavefunction exists in the regions $x<0$ and $x>a$, which means that the probability of finding a particle there is non zero. In other words, even if a particle has an energy E lower than the potential barrier U_0, it has a non zero probability of being found beyond the barrier. This is schematically shown in Fig. 3.12.

In the case of $E>U_0$, the solution of Eq. (3.36) can again, as before, be written as a sum of a cosine and a sine term (see Eq. (3.28)), for each of the regions defined by Eq. (3.34). Another, more elegant way, is to represent the solution as a sum of a plane waves, one going to the left and one going to the right. The two plane waves have different wavenumbers 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. (3.6), one can analytically determine the wavefunction. This analysis would lead to result similar for a free particle, and in particular that there is a continuum of energy states $E>U_0$ allowed.

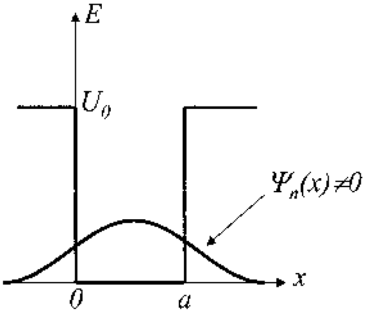


Fig. 3.12. Illustration of the tunneling effect in a finite potential well. The wavefunction is non zero outside the potential well. This means that there exists a non zero probability of finding an electron outside the potential well even when its energy E is lower than the potential barrier height U_0.

3.4. Summary

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 concepts, including an infinite and a finite potential well. Through these, some aspects of quantum mechanics have been discussed, including the quantization of energy levels and momenta, and tunneling effects.

Further reading

Bastard, G., *Wave Mechanics Applied to Semiconductor Heterostructures*, Halsted Press, New York, 1988.

Cohen-Tannoudji, C., Diu, B. and Laloë, F., *Quantum Mechanics*, John Wiley & Sons, New York, 1977.

Dalven, R., *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, 1990.

Davydov, A.S., *Quantum Mechanics*, Pergamon, New York, 1965.

Kittel, C., *Introduction to Solid State Physics*, John Wiley & Sons, New York, 1976.

Liboff, R.L., *Introductory Quantum Mechanics*, Addison-Wesley, Reading, MA, 1998.

McKelvey, J.P., *Solid State and Semiconductor Physics,* Harper and Row, New York, 1966.

Pierret, R.F., *Advanced Semiconductor Fundamentals*, Addison-Wesley, Reading, MA, 1989.

Powell, J.L. and Crasemann, B., *Quantum Mechanics,* Addison-Wesley, Reading, MA, 1961.

Ziman, J.M., *Elements of Advanced Quantum Theory*, Cambridge University Press, London, 1969.

Problems

1. According to quantum mechanics, electromagnetic radiation of frequency υ can be regarded as consisting of photons of energy $h\upsilon$ where h=6.626×10^{-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 (λ=336 nm) emit? A cell phone that emits 0.4 W of 850 MHz radiation? A microwave oven operating at 2 GHz generating a microwave power of 1000 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.kg^{-1}.K^{-1}) by 20 °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 20 % of the incident power reaches the retina, how many photons at the receptors generate the signal that the test person recognizes as a flash of light?

3. (a) The thermal energy scale is k_bT, where k_b=1.38×10^{-23} J/K is Boltzmann's 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 20% and take the photon wavelength as 600 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?
A simulation of black body radiation and related topics (Planck's law, Wien's law) can be found 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. (3.2), show 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. 2.3). Do so by first rewriting the expression in Eq. (3.2) for $u(\nu, T)$, the distribution per unit energy, as a distribution per unit wavelength, i.e. $w(\lambda, T)$.

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 de Broglie wavelength of an automobile (2000 kg) traveling at 35 miles per hour? A dust particle of radius 1 μm and density 300 kg.m^{-3} being jostled by air molecules at room temperature (T=300 K)? An ^{14}N atom that has been laser cooled to a temperature of T=77 K? 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. Prove that the normalization constant A in Eq. (3.28) is equal to $A=(2/a)^{1/2}$ as given in Eq. (3.32).

8. A particle with mass 6.10×10^{-32} kg is confined to an infinite square well of width L. The energy of the third level is 3×10^{-17} J. Calculate the value of L.

9. 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. Assume that the wavefunction is the same immediately after the change, if the change happens fast enough.
(a) What is the physical meaning of "fast enough" in quantum mechanics?
(b) Determine the probability of finding the particle 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 well before its expansion. Do not forget to properly normalize the wave-functions for your calculations.

(c) What is the probability that the particle is found in the ground state after the sudden expansion?

(d) Calculate the expectation value of the energy of the particle before and after the sudden expansion?

10. An electron is confined to a 1 micron layer of GaAs. 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 GaAs is $0.067m_0$, where m_0=9.10×10^{-31} kg is the free electron rest mass).

11. In this exercise, we will develop the material of section 3.3.3 to calculate the factor of confinement of a particle in a finite well.

Making the system symmetric for convenience, we translate the Ox axis so that the potential equals to 0 in the region: $-a/2 < x < a/2$.

(a) Rewrite Eq. (3.38) in this new coordinate. Use the appropriate boundary conditions to eliminate some trivial constants.

(b) By symmetry, we search for solutions in 2 families of functions: even and odd functions. Show that the even solutions satisfy 2 equations:

$$\begin{cases} \tan(\dfrac{ka}{2}) = \dfrac{\alpha}{k} \\ k^2 + \alpha^2 = \dfrac{2mU_0}{\hbar^2} \end{cases}$$

while the odd solutions satisfy

$$\begin{cases} -\cot(\dfrac{ka}{2}) = \dfrac{\alpha}{k} \\ k^2 + \alpha^2 = \dfrac{2mE}{\hbar^2} \end{cases}$$

How can you resolve these equations graphically?

(c) The particle is in the ground state, which is even, and has energy E. Find the relative density of the particle inside the well. This quantity is defined as the confinement factor (or coefficient of confinement).

A simulation of this Problem can be found at:

http://www.sgi.com/fun/java/john/wave-sim.html

12. Consider a particle of mass m moving in the potential:

$$V(x) = -\frac{\hbar^2 a^2}{m}\frac{1}{\cosh^2(ax)}$$

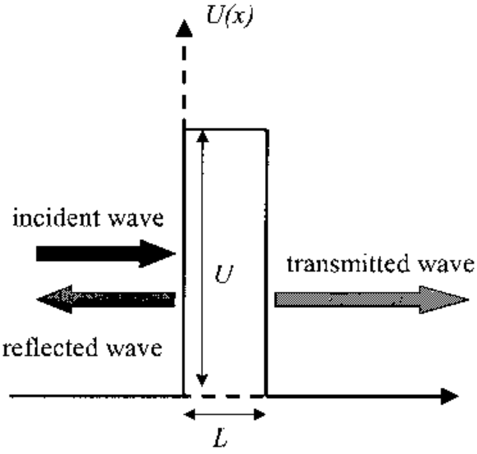
(a) Show that this potential has a bound eigenstate described by the wavefunction $\psi_0(x) = \frac{A}{\cosh(ax)}$ and find the corresponding eigenenergy. Normalize Ψ_0 and sketch it. This turns out to be the only bound state for this potential.

(b) Show that the wavefunction is $\psi_k(x) = B\left(\frac{ik - a\tanh(ax)}{ik + a}\right)e^{ikx}$

where $\hbar k = \sqrt{2mE}$. Solve the Schrödinger equation for any positive energy E. Verify that for $x \to \pm\infty$ the asymptotic behavior of $\psi_K(x)$ has the plane wave form. Determine the transmission coefficient if it's 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?

A simulation of this Problem can be found at:
http://www.kfunigraz.ac.at/imawww/thaller/visualization/vis.html

13. Consider a particle of energy E traveling from the left hits a barrier of height $U > E$ and thickness L. Calculate the transmission coefficient.



A simulation of this Problem can be found at:
http://www.kfunigraz.ac.at/imawww/thaller/visualization/vis.html
http://www.sgi.com/fun/java/john/wave-sim.html

4. Electrons and Energy Band Structures in Crystals

4.1. Introduction

In Chapter 3, 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".

4.2. 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.

4.2.1. Bloch theorem

The Bloch theorem provides a powerful mathematical simplification for the wavefunctions of particles evolving in a periodic potential. The solutions of the Schrödinger equation in such a potential are not pure plane waves as they were in the case of a free particle (Eq. (3.22)), 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:

Eq. (4.1) $$\Psi\left(\vec{k},\vec{r}\right)=\exp\left(i\vec{k}.\vec{r}\right)u\left(\vec{k},\vec{r}\right)$$

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:

Eq. (4.2) $$u\left(\vec{k},\vec{r}+\vec{R}\right)=u\left(\vec{k},\vec{r}\right)$$

The expression in Eq. (4.1) means that the Bloch wavefunction is a plane wave, given by the exponential term in Eq. (4.1), which is modulated by a function which has the periodicity of the crystal lattice. An illustration of this is shown in Fig. 4.1 in the one-dimensional case.

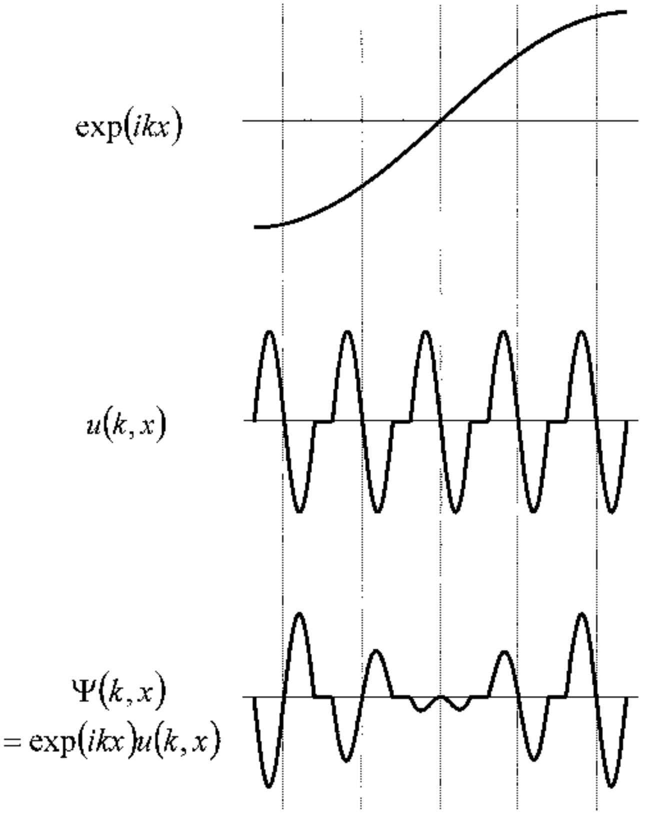


Fig. 4.1. One-dimensional illustration of a Bloch wavefunction (bottom) as a plane wave (top) modulated by a periodic function which has the period of the lattice (middle).

Combining Eq. (4.1) and Eq. (4.2) leads us to the form:

Eq. (4.3) $$\Psi\left(\vec{k},\vec{r}+\vec{R}\right)=\exp\left(i\vec{k}.\vec{R}\right)\Psi\left(\vec{k},\vec{r}\right)$$

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:

Eq. (4.4) $$\Psi(k,x+d)=\exp(ikd)\Psi(k,x)$$

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:

Eq. (4.5) $$\Psi(k,x)=\Psi(k,x+Nd)=\exp(ikNd)\Psi(k,x)$$

which means that:

Eq. (4.6) $$\exp(ikNd)=1$$

or:

Eq. (4.7) $$k=\frac{2\pi n}{Nd}$$

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*.

4.2.2. 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. 4.2(a) and is approximated by the one depicted in Fig. 4.2(b).

The solution of the Kronig-Penney model partially utilizes the results from the finite potential well problem discussed in sub-section 3.3.3 and the same notations has therefore been used in Fig. 4.2(b). The mathematical analysis will first be conducted locally, in the region $-b<x<a$, where the potential can be approximated by Eq. (3.35) except that there is a new limit for the variable *x*.

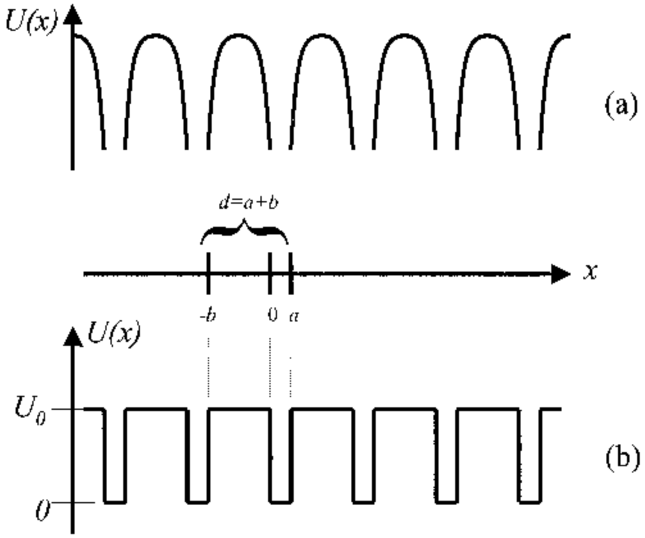


Fig. 4.2. (a) Real crystal potential experienced by electrons in a crystal and (b) simplified crystal potential used in the Kronig-Penney model.

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:

$$\text{Eq. (4.8)} \quad \begin{cases} \dfrac{d^2\Psi_1(x)}{dx^2} + \alpha^2\Psi_1(x) = 0 & \quad for -b < x < 0 \\ \dfrac{d^2\Psi_2(x)}{dx^2} + \beta^2\Psi_2(x) = 0 & \quad for\, 0 < x < a \end{cases}$$

by defining:

$$\text{Eq. (4.9)} \quad \begin{cases} \alpha = \begin{cases} i\alpha_-, with\, \alpha_- = \sqrt{\dfrac{2m(U_0 - E)}{\hbar^2}} & when\, 0 < E < U_0 \\ \alpha_+, with\, \alpha_+ = \sqrt{\dfrac{2m(E - U_0)}{\hbar^2}} & when\, U_0 < E \end{cases} \\ \beta = \sqrt{\dfrac{2mE}{\hbar^2}} \end{cases}$$

The general solution to Eq. (4.8) can be expressed as:

Eq. (4.10)
$$\begin{cases}\Psi_1(x) = A_1\sin(\alpha x) + B_1\cos(\alpha x)\\ \Psi_2(x) = A_2\ \sin(\beta x) + B_2\cos(\beta x)\end{cases}$$

with the understanding that sin(αx) and cos(αx) become -$i$sinh($\alpha\ x$) and cosh(α_x), respectively, when $\alpha=i\alpha\_$ 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. (4.4) between points $x=a$ and $x=-b$:

Eq. (4.11)
$$\begin{cases}\Psi_1(0) = \Psi_2(0)\\ \frac{d\Psi_1}{dx}(0) = \frac{d\Psi_2}{dx}(0)\end{cases}$$
$$\begin{cases}e^{ik(a+b)}\Psi_1(-b) = \Psi_2(a)\\ e^{ik(a+b)}\frac{d\Psi_1}{dx}(-b) = \frac{d\Psi_2}{dx}(a)\end{cases}$$

Utilizing Eq. (4.10), we obtain:

Eq. (4.12)
$$\begin{cases}B_1 = B_2\\ \alpha A_1 = \beta A_2\\ e^{ik(a+b)}\left[-A_1\sin(\alpha b) + B_1\cos(\alpha b)\right] = A_2\sin(\beta a) + B_2\cos(\beta a)\\ e^{ik(a+b)}\left[\alpha A_1\cos(\alpha b) + \alpha B_1\sin(\alpha b)\right] = \beta A_2\cos(\beta a) - \beta B_2\sin(\beta a)\end{cases}$$

which can be simplified by expressing A_2 and B_2 in terms of A_1 and B_1:

Eq. (4.13)
$$\begin{cases}A_1\left[e^{ik(a+b)}\sin(\alpha b) + \frac{\alpha}{\beta}\sin(\beta a)\right] + B_1\left[\cos(\beta a) - e^{ik(a+b)}\cos(\alpha b)\right] = 0\\ A_1\left[\alpha e^{ik(a+b)}\cos(\alpha b) - \alpha\cos(\beta a)\right] + B_1\left[\beta\sin(\beta a) + \alpha e^{ik(a+b)}\sin(\alpha b)\right] = 0\end{cases}$$

This system of two equations with two unknowns has a non-zero solution (i.e. A_I and B_I 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:

Eq. (.4.14)

$$\left[e^{ik(a+b)}\sin(\alpha b)+\frac{\alpha}{\beta}\sin(\beta a)\right]\left[\beta\sin(\beta a)+\alpha e^{ik(a+b)}\sin(\alpha b)\right]$$
$$-\left[\cos(\beta a)-e^{ik(a+b)}\cos(\alpha b)\right]\left[\alpha e^{ik(a+b)}\cos(\alpha b)-\alpha\cos(\beta a)\right]=0$$

or after simplification:

Eq. (4.15) $$\cos k(a+b)=-\frac{\alpha^2+\beta^2}{2\alpha\beta}\sin(\alpha b)\sin(\beta a)+\cos(\alpha b)\cos(\beta a)$$

Using the same constants as in Eq. (3.44), we can rewrite Eq. (4.9) as:

Eq. (4.16) $$\begin{cases}\alpha=\begin{cases}i\alpha_-, with\,\alpha_-=\alpha_0\sqrt{1-\zeta} & when\,0<E<U_0\\ \alpha_+, with\,\alpha_+=\alpha_0\sqrt{\zeta-1} & when\,U_0<E\end{cases}\\ \beta=\alpha_0\sqrt{\zeta}\end{cases}$$

Therefore, Eq. (4.15) can be simplified into:

Eq. (4.17)

$$\begin{cases}\cos k(a+b)=\dfrac{1-2\zeta}{2\sqrt{\zeta(1-\zeta)}}\sin\left(\alpha_0 a\sqrt{\zeta}\right)\sinh\left(\alpha_0 b\sqrt{1-\zeta}\right)+\cos\left(\alpha_0 a\sqrt{\zeta}\right)\cosh\left(\alpha_0 b\sqrt{1-\zeta}\right)\\ \qquad for\,0<\zeta<1\\ \cos k(a+b)=\dfrac{1-2\zeta}{2\sqrt{\zeta(\zeta-1)}}\sin\left(\alpha_0 a\sqrt{\zeta}\right)\sin\left(\alpha_0 b\sqrt{\zeta-1}\right)+\cos\left(\alpha_0 a\sqrt{\zeta}\right)\cos\left(\alpha_0 b\sqrt{\zeta-1}\right)\\ \qquad for\,1<\zeta\end{cases}$$

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. (4.17) allows us to determine the values of the energy as well as the wavefunctions (after normalization).

4.2.3. 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). Eq. (4.17) is most easily solved graphically. The shape of the right hand side function of Eq. (4.17), which we will call $f(\zeta)$, can be visualized in Fig. 4.3.

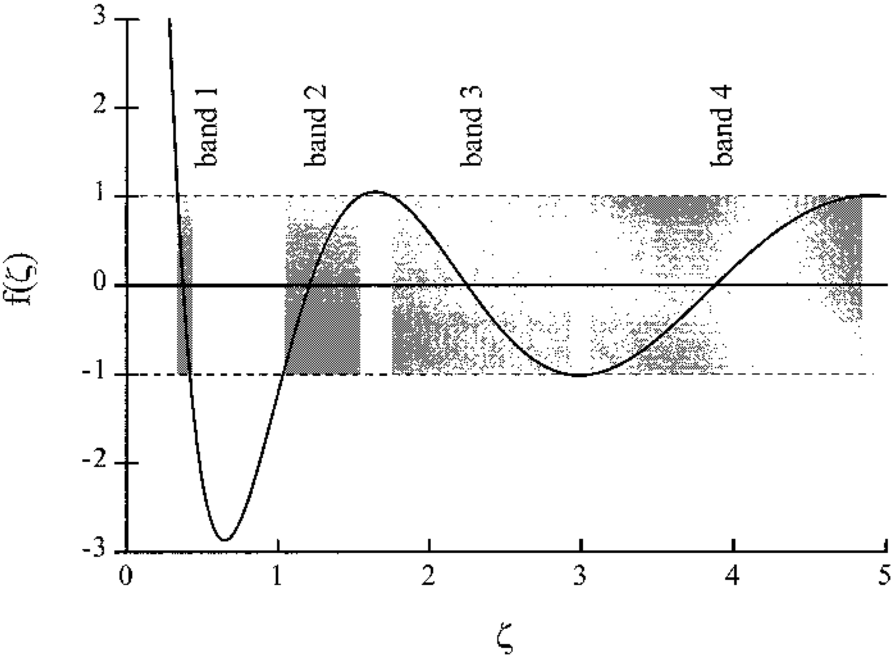


Fig. 4.3. Plot of the right hand side of Eq. (4.17), showing the graphical determination of the E-k relationship. There exist a solution to Eq. (4.17) only when the right hand side of the equation is between -1 and +1, which correspond to the shaded areas.

Because of the cosine on the LHS of Eq. (4.17), 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. 4.3. 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. 4.3, 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. 4.4.

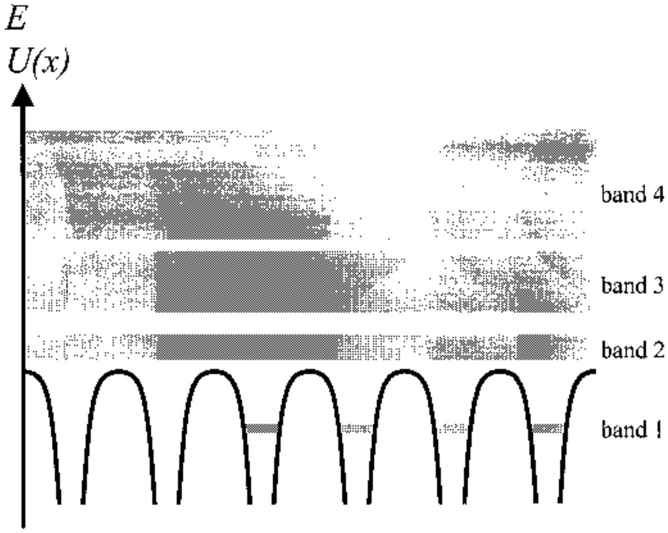


Fig. 4.4. Illustration of the concept of energy bands in the crystal.

Furthermore, as we can see from Fig. 4.3, 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. 4.5 and is called the energy spectrum, the band diagram or band structure. This type of diagram is very important in determining the 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 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.

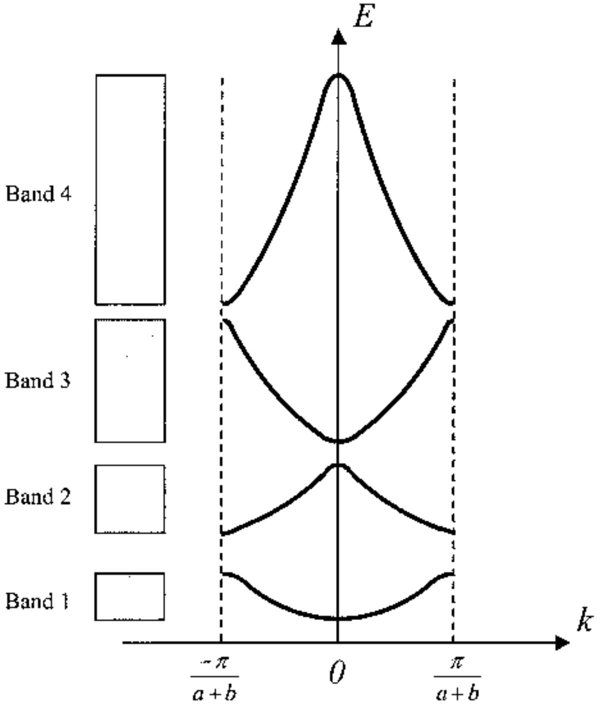


Fig. 4.5. One-dimensional E-k relationship in the reduced-zone representation in the Kronig-Penney model.

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. 4.6., 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 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.

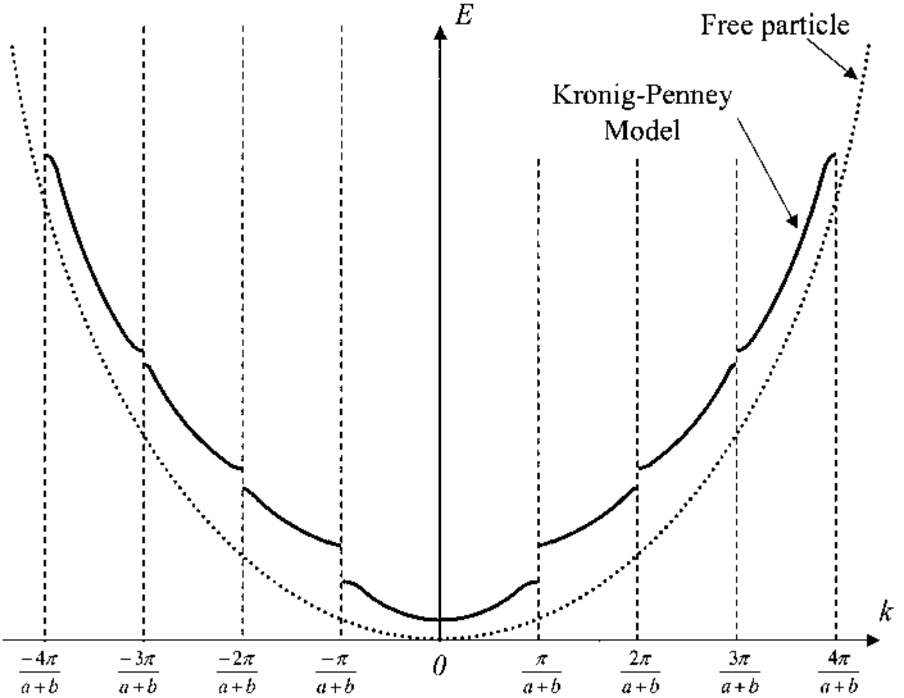


Fig. 4.6. One-dimensional E-k relationship in the extended-zone representation in the Kronig-Penney model. The parabolic relation for the free particle is shown in dotted lines for comparison. The deviation from a parabolic shape occurs mainly at the Brillouin zone boundaries.

4.2.4. 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 model. In this sub-section, 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. 4.7 (solid line) and compared with that of a free electron (dashed lines).

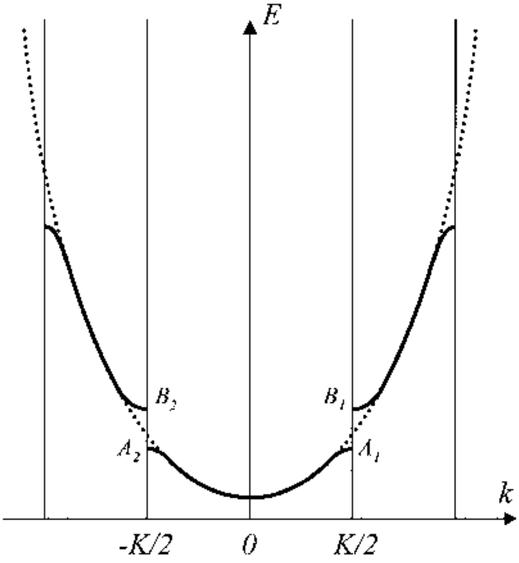


Fig. 4.7. Electron energy in a lattice (solid curve) and energy spectrum of free electrons (dashed curve). The deviation from the parabolic shape occurs at the Brillouin zone boundaries.

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 Chapter 1 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 sub-section.

The energy difference between branches at points A_l and B_l (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. 3.6.

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. 4.8.

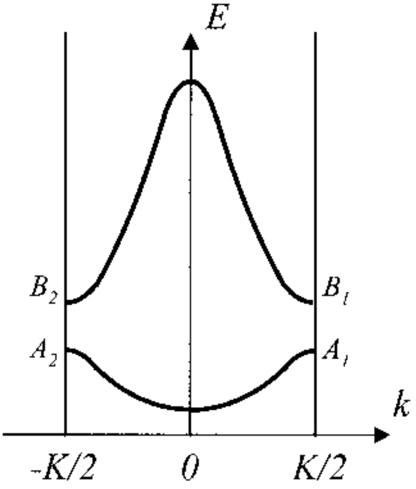


Fig. 4.8. Electron energy in the reduced zone scheme.

4.2.5. *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 presence of electrons from different atoms will be non-zero in the same position in space. These result in a non-zero 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

Eq. (4.18) $$\Psi_{\vec{k}}(\vec{r}) = \sum_{j,n} \beta_j \Phi_j\left(\vec{r} - \vec{R}_n\right)\exp\left(i\vec{k}.\vec{R}_n\right)$$

where β_j are the admixture coefficients of the j^{th} orbital, and $\Phi_j(\vec{r} - \vec{R}_n)$ is the j^{th} orbital itself on the atom located at $\vec{R}_n$ respectively. Substituting Eq. (4.18) 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:

$$\{-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r} - \vec{R}_l)\}\Phi_\alpha(\vec{r} - \vec{R}_l) = E_\alpha \Phi_\alpha(\vec{r} - \vec{R}_l)$$

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 $t_{l+1,l} = \int d\vec{r}\Phi^*(\vec{r} - \vec{R}_{l-1})V(\vec{r} - \vec{R}_l)\Phi(\vec{r} - \vec{R}_l)$. 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 $t_{ij} \sim t$, is illustrated in Fig. 4.9 and Fig. 4.10.

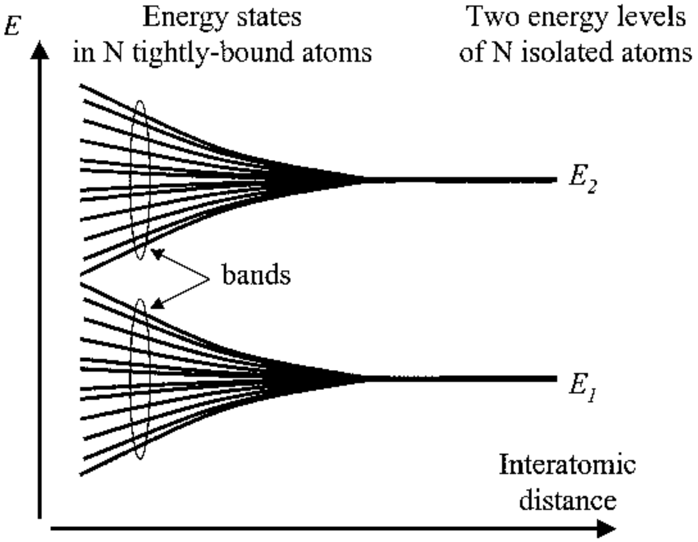


Fig. 4.9. Broadening of the atomic energy levels in a solid. When the atoms are isolated, they all have the same allowed discrete energy levels (e.g. E_1 and E_2). When the interatomic distance decreases, the atoms interact with one another and the allowed energy levels split: some increase while some others decrease.

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.

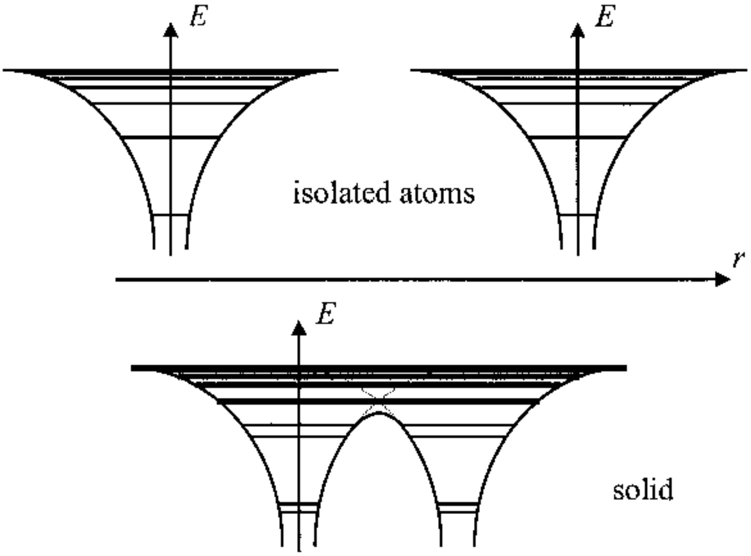


Fig. 4.10. Change in energy spectrum from single atoms to a solid. Each of the discrete energy levels in two isolated atoms split into two separate energy levels when the atoms are bound in a solid.

4.2.6. 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 sub-sections.

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:

Eq. (4.19) $$v_g = \frac{d\omega}{dk}$$

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. (4.19) thus becomes:

Eq. (4.20) $$v_g = \frac{1}{\hbar}\frac{dE}{dk}$$

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:

Eq. (4.21) $$dE = Fdx = Fv_g dt$$

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:

Eq. (4.22) $$F = \frac{1}{v_g}\frac{dE}{dt} = \frac{1}{v_g}\frac{dE}{dk}\frac{dk}{dt}$$

or:

Eq. (4.23) $$F = \hbar\frac{dk}{dt} = \frac{d(\hbar k)}{dt}$$

after using Eq. (4.20). On the other hand, differentiating Eq. (4.20) 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:

Eq. (4.24) $$\frac{dv_g}{dt} = \frac{1}{\hbar^2}\frac{d^2E}{dk^2}\frac{d(\hbar k)}{dt}$$

Eliminating $\frac{d(\hbar k)}{dt}$ in Eq. (4.23) and Eq. (4.24), we find:

Eq. (4.25) $$F = \left(\frac{1}{\frac{1}{\hbar^2}\frac{d^2E}{dk^2}}\right)\frac{dv_g}{dt}$$

This expression resembles Newton's law of motion when rewritten as:

Eq. (4.26) $$F = m^*\frac{dv_g}{dt}$$

where we have defined m^* as:

Eq. (4.27) $$m^* = \frac{\hbar^2}{d^2E/dk^2}$$

m^* is called the electron effective mass and has a very significant meaning in solid state physics. Eq. (4.26) 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 sub-sections 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. (3.23)):

$$E(k) = \frac{\hbar^2 k^2}{2m}$$

where m is the mass of the electron. Using Eq. (4.27), 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. 4.7, 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 (sub-section 4.3.3). 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. (4.26). 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: $m^* = \dfrac{1}{\dfrac{1}{\hbar^2}\dfrac{d^2E}{dk^2}} = \dfrac{1}{\dfrac{1}{\hbar^2}\dfrac{d^2\left(Ak^2\right)}{dk^2}} = \dfrac{\hbar^2}{2A}$. In units of free electron mass, we get:

$$\frac{m^*}{m_0} = \frac{\hbar^2}{2Am_0}$$

$$= \frac{\left(1.05458\times10^{-34}\right)^2}{2\times\left(84.67\times10^{-20}\times1.60218\times10^{-19}\right)\left(0.91095\times10^{-30}\right)}$$

$$= 0.045$$

4.2.7. 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

(Chapter 2), the electrons in a crystal fill the lower energy bands first, while satisfying the Pauli exclusion principle.

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 ($m>>n$): 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. 4.11(b). 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. 4.11(a). The energy gap between the valence band and the conduction band is called the bandgap and is denoted E_g.

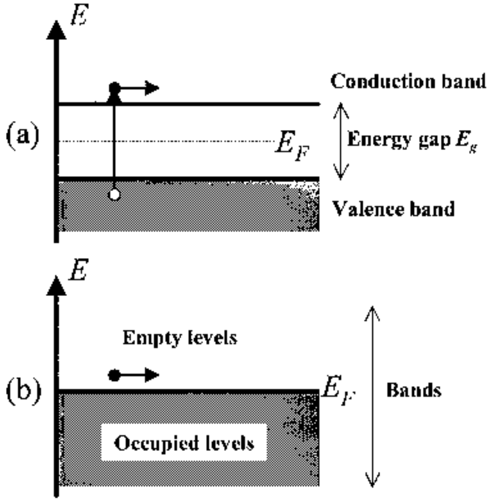


Fig. 4.11. Bands in (a) semiconductors and (b) metals. In most semiconductors E_F is in the bandgap. In semiconductors, there is an energy region that does not contain allowed energy levels and the Fermi energy is located in it. In metals, the Fermi energy is located inside an allowed energy band.

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 $k_b T$, where k_b (=1.38066×10^{-23} J.K^{-1}=0.08625 meV.K^{-1}) is the Boltzmann constant and T is the temperature in degrees K.

4.2.8. 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:

$$\text{Eq. (4.28)} \qquad f_e(E) = \frac{1}{\exp\left(\dfrac{E - E_{\mathrm{F}}}{k_b T}\right) + 1}$$

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. 4.12 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. (4.28) 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.

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.

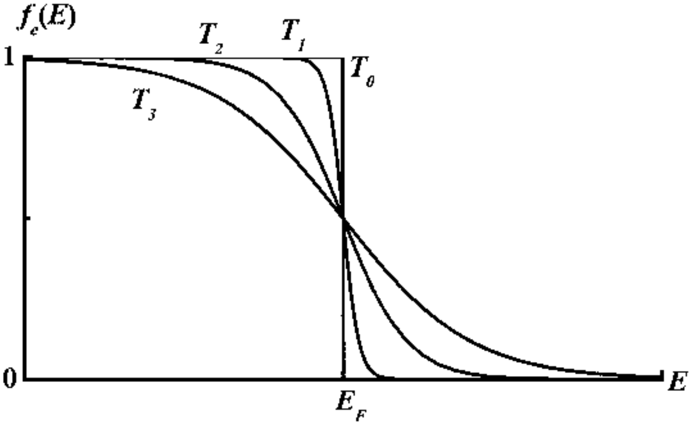


Fig. 4.12. Fermi-Dirac distribution function at different temperatures: T_3>T_2>T_1,T_0=0 K. At the absolute zero temperature, the probability of an electron to have an energy below the Fermi energy E_F is equal to 1, whereas its probability to have a higher energy is zero.

4.3. 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 motivation 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 section 4.2 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. (4.7).

The analysis in section 4.2 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 $\vec{r} = (x, y, z)$.

4.3.1. Direct calculation

Let us assume that the shape of the crystal is a rectangular parallelepiped of linear dimensions L_x, L_y, L_z and volume $V = L_x L_y L_z$. The periodic boundary conditions, similar to Eq. (4.5), require the electron quantum states to be the same at opposite surfaces of the sample:

Eq. (4.29)

$$\Psi(x+L_x, y, z) = \Psi(x, y+L_y, z) = \Psi(x, y, z+L_z) = \Psi(x, y, z)$$

Using the Bloch theorem, these conditions mean that:

Eq. (4.30) $$\exp(ik_x L_x) = \exp(ik_y L_y) = \exp(ik_z L_z) = 1$$

or:

Eq. (4.31) $$\begin{cases} k_x = \dfrac{2\pi}{L_1} n_x \\ k_y = \dfrac{2\pi}{L_2} n_y \\ k_z = \dfrac{2\pi}{L_3} n_z \end{cases}$$

where $n_x, n_y, n_z = 0, \pm 1 \ldots$ 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. Eq. (4.31) 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 Chapter 1, 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 Chapter 4, we understood 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:

Eq. (4.32) $$g(\vec{k}) = 2\frac{V}{(2\pi)^3}$$

where the extra factor 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: $g(\vec{k}) = 2\frac{V}{(2\pi)^3} = 2\times\frac{1mm^3}{8\pi^3} = 8.063\times10^{-3}$ 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: $N = V_k g(\vec{k}) = (5.492\times10^{21} mm^{-3})(8.063\times10^{-3} mm^3)$.
$\approx 4.43\times10^{19}$

The density of states *g(E)* as defined earlier is therefore related to its counterpart in *k*-space, $g(\vec{k})$, by:

Eq. (4.33) $$g(E)dE = g(\vec{k})d\vec{k}$$

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 section 4.2, the shape of the $E(\vec{k})$ relationship near the bottom of the conduction band can generally be considered parabolic:

Eq. (4.34) $$E(\vec{k}) = \frac{\hbar^2 k^2}{2m^*}$$

where *k* is the norm or length of the wavevector $\vec{k}$, and *m** is the electron effective mass as defined in sub-section 4.2.6. Using this expression, we can express successively:

Eq. (4.35) $$dE = \frac{\hbar^2}{2m^*}(2k)dk$$

When considering orthogonal coordinates, the unit volume in *k*-space is defined given by:

Eq. (4.36) $$d\vec{k} = dk_x dk_y dk_z$$

which is equal, when using spherical coordinates, to:

Eq. (4.37) $$d\vec{k} = d\left(\frac{4\pi}{3}k^3\right) = 4\pi k^2 dk$$

Therefore, by replacing into Eq. (4.35), we get:

Eq. (4.38) $$dE = \frac{\hbar^2}{2m^*}\left(\frac{1}{2\pi k}\right)d\vec{k}$$

Using Eq. (4.34) to express k in terms of E, and replacing into Eq. (4.38):

Eq. (4.39) $$\begin{aligned} dE &= \frac{\hbar^2}{2m^*}\left(\frac{1}{2\pi}\sqrt{\frac{\hbar^2}{2m^*E}}\right)d\vec{k} \\ &= \frac{1}{2\pi}\left(\frac{\hbar^2}{2m^*}\right)^{3/2}\frac{1}{\sqrt{E}}d\vec{k} \end{aligned}$$

Now, by replacing into Eq. (4.33), we obtain successively:

$$g(E) = 2\pi\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E}\,g(\vec{k})$$

Finally, using Eq. (4.32), we get:

Eq. (4.40) $$g_{3D}(E) = \frac{V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E}$$

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. 4.13.

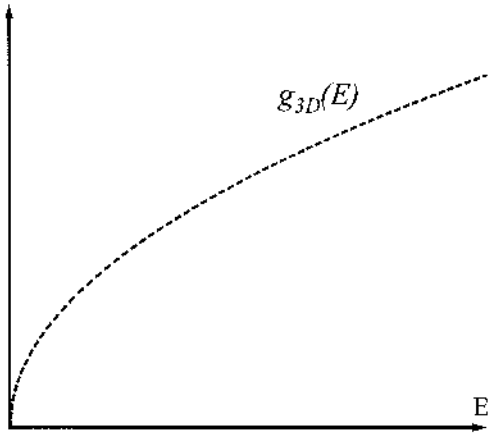


Fig. 4.13. Energy dependence of density of states for a three-dimensional semiconductor conduction band. The density of states follows a parabolic relationship.

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 $g_{3D}(E)$: $N = \int_0^{eV} g_{3D}(E)dE$. 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:

$$\begin{aligned} N &= \int_0^{1eV} g_{3D}(E)dE = \frac{V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\int_0^{1eV}\sqrt{E}dE \\ &= \frac{V}{3\pi^2}\left(\frac{2m^*\times 1eV}{\hbar^2}\right)^{3/2} \\ &= \frac{\left(10^{-3}\right)^3}{3\pi^2}\left(\frac{2\left(0.067*0.91095\times10^{-30}\right)\times\left(1.60218\times10^{-19}\right)}{\left(1.05458\times10^{-34}\right)^2}\right)^{3/2} \\ &\approx 7.88\times10^{16} \end{aligned}$$

4.3.2. 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)=2x(number of states which have an energy $E\left(\vec{k}\right)$ equal to *E*)

which can be mathematically expressed as:

Eq. (4.41) $$g(E) = 2\sum_{\vec{k}}\delta\left[E\left(\vec{k}\right) - E\right]$$

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:

Eq. (4.42) $$\begin{cases} \delta(x) = 0 & for\, x \neq 0 \\ \int_{-\infty}^{+\infty}\delta(x)dx = 1 & \end{cases}$$

Some of the most important properties of the Dirac delta function include:

Eq. (4.43)
$$\begin{cases} \int_{-\infty}^{+\infty} \delta(x) Y(x) dx = Y(0) \\ \int_{-\infty}^{+\infty} \delta(x - x_0) Y(x) dx = Y(x_0) \end{cases}$$

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 $\Upsilon(\vec{k})$ 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:

Eq. (4.44)
$$\sum_{\vec{k}} \Upsilon(\vec{k}) \equiv \frac{V}{(2\pi)^3} \iiint_{\vec{k}} \Upsilon(\vec{k}) d\vec{k} = \frac{V}{(2\pi)^3} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Upsilon(k_x, k_y, k_z) dk_x dk_y dk_z$$

The factor $\frac{V}{(2\pi)^3}$ is the volume occupied by a reciprocal lattice point in k-space. Eq. (4.41) can therefore be rewritten into:

Eq. (4.45) $$g(E) = \frac{1}{4\pi^3} \iiint_{\vec{k}} \delta\left[E(\vec{k}) - E\right] d\vec{k}$$

Now, we need to use the expression of $d\left[E(\vec{k})\right]$ as a function of $d\vec{k}$ found in Eq. (4.39):

Eq. (4.46) $$d\left[E(\vec{k})\right] = \frac{1}{2\pi} \left(\frac{\hbar^2}{2m^*} \right)^{3/2} \frac{1}{\sqrt{E(\vec{k})}} d\vec{k}$$

Eq. (4.45) therefore becomes:

Eq. (4.47) $$g(E)=\frac{2\pi V}{4\pi^3}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\int_0^\infty \delta\left[E(\vec{k})-E\right]\sqrt{E(\vec{k})}d\left[E(\vec{k})\right]$$

and after the change of variable $E(\vec{k})\rightarrow x$:

Eq. (4.48) $$g(E)=\frac{V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\int_0^\infty \delta(x-E)\sqrt{x}dx$$

Using Eq. (4.43), and because E>0:

Eq. (4.49) $$g(E)=\frac{V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E}$$

which is the same expression as Eq. (4.40) 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:

Eq. (4.50) $$n_{total}=\int_0^\infty g(E)f_e(E)dE$$

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 Chapter 7 in which we will discuss the equilibrium electronic properties of semiconductors. Here, the density of electrons n, with effective mass m_e, in the conduction band is given by:

Eq. (4.51) $$n = \frac{1}{V}\int_{E_C}^{\infty} g(E) f_e(E) dE$$

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:

Eq. (4.52) $$g(E) = \frac{V}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E - E_C)^{1/2}$$

Combining this expression with Eq. (4.28), the density of electrons becomes:

Eq. (4.53) $$n = \frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\int_{E_C}^{\infty}(E - E_C)^{1/2}\frac{1}{\exp\left(\frac{E - E_F}{k_b T}\right) + 1}dE$$

or:

Eq. (4.54) $$n = N_c F_{\frac{1}{2}}\left(\frac{E_F - E_C}{k_b T}\right)$$

where:

Eq. (4.55) $$N_e = 2\left(\frac{2\pi k_b T m_e}{h^2}\right)^{3/2}$$

is the effective density of states in the conduction band, and:

Eq. (4.56) $$F_{\frac{1}{2}}(x) = \frac{2}{\sqrt{\pi}}\int_0^{\infty}\frac{y^{1/2}}{1 + \exp(y - x)}dy$$

is the Fermi-Dirac integral. A more detailed discussion on the effective density of states and the Fermi-Dirac integral will be given in Chapter 7.

4.3.3. 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. 4.8, 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 sub-section 4.2.6 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. Fig. 4.14(a) and (b) are equivalent descriptions of the same physical phenomenon. In Fig. 4.14(a), we are showing the energy states occupied by electrons. In Fig. 4.14(b), we are showing the energy states in the valence band which are occupied by holes, i.e. vacated by electrons.

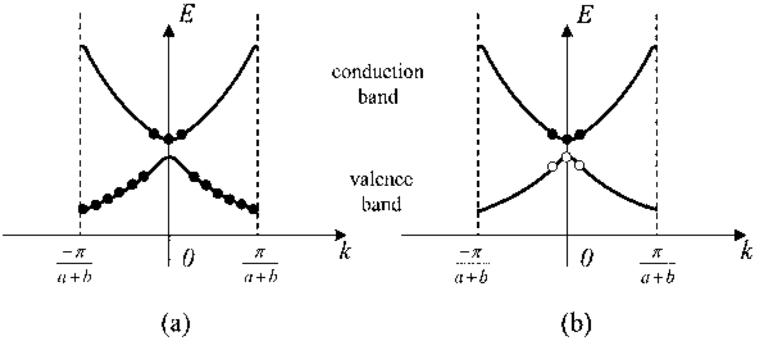


Fig. 4.14. Electron energy states in the reduced zone scheme. In (a), the solid circles show the states occupied by electrons. In (b), the closed circles show the states in the conduction band which are occupied by electrons, and the open circles the states in the valence band occupied by holes.

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. 4.8 and is computed as:

Eq. (4.57) $$m^* = -\frac{\hbar^2}{d^2E/dk^2}$$

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:

Eq. (4.58) $$f_h(k) = 1 - f_e(k)$$

The electrical current from the electrons in the band is:

Eq. (4.59) $$j = -2q\sum_k f_e(k)v_\mathrm{k}$$

where v_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:

Eq. (4.60) $$\begin{aligned} j &= -2q\sum_k f_e(k)v_\mathrm{k} = -2q\sum_k \left[1 - f_h(k)\right]v_\mathrm{k} \\ &= -2q\sum_k v_\mathrm{k} + 2q\sum_k f_h(k)v_\mathrm{k} \end{aligned}$$

We can now use the fact that the electron energy spectrum is always symmetrical, i.e. $E(k) = E(-k)$, hence $v_\mathrm{k} = -v_\mathrm{-k}$ from Eq. (4.20), and the sum of velocities over the entire first Brillouin zone is zero. The first sum in Eq. (4.60) is thus equal to zero and we obtain:

Eq. (4.61) $$j = +2q\sum_k f_h(k)v_\mathrm{k}$$

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.

4.4. 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 $\frac{2\pi}{a} \approx 6.28 * 10^{10} m^{-1} >> \frac{2\pi}{Na} \approx 6.28 * 10^{10} m^{-1}$. As a result, at the scale of the reciprocal lattice, the wavenumber can be considered to vary continuously.

4.4.1. First Brillouin zone of an fcc lattice

The first Brillouin zone of an fcc lattice is shown in Fig. 4.15. 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:

Γ	$\frac{2\pi}{a}$ (0,0,0)
X	$\frac{2\pi}{a}$ (0,0,1)
W	$\frac{2\pi}{a}$ (½,0,1)
K	$\frac{2\pi}{a}$ (¾,¾,0)

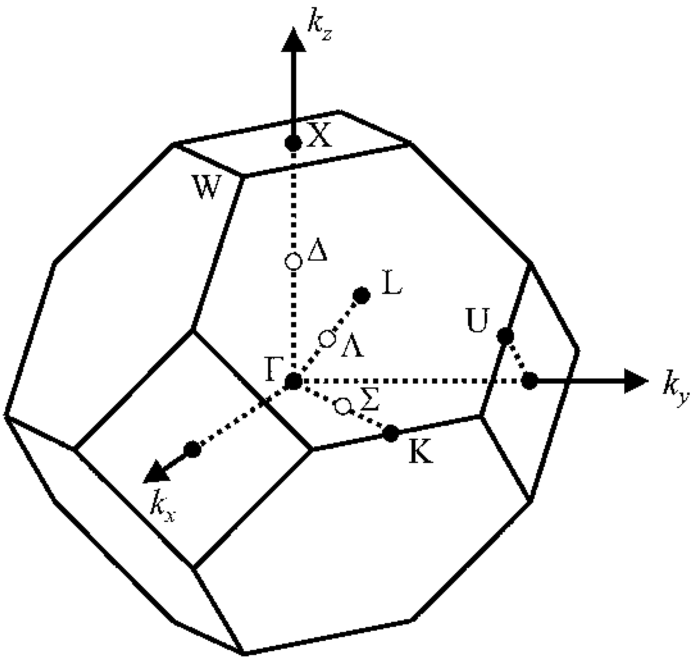


Fig. 4.15. First Brillouin zone of an fcc lattice.

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 $\frac{2\pi}{a}(0,0,\pm 1)$, $\frac{2\pi}{a}(0,\pm 1,0)$ and $\frac{2\pi}{a}(\pm 1,0,0)$.

Using Miller indices, the symmetry directions can be denoted as:

Δ :	Γ →X	(parallel to <100>)
Λ :	Γ →L	(parallel to <111>)
Σ :	Γ →K	(parallel to <110>).

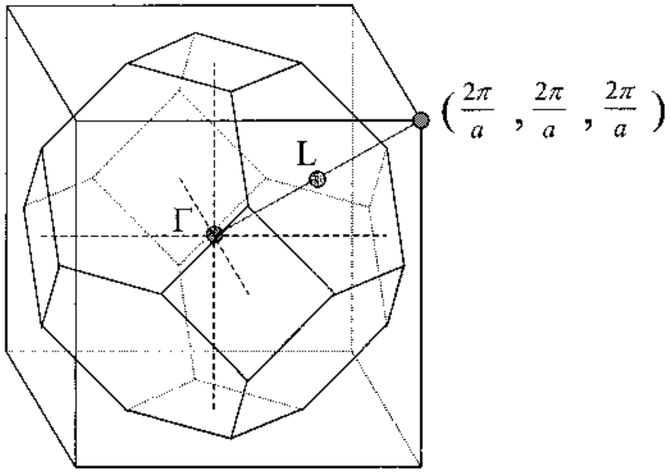
These notations come from the crystal group theory where they are used to label the symmetry operation groups at those particular high-symmetry points and directions. For example, Γ is the symmetry group at the zone center ($\vec{k}$=(0,0,0)) and is isomorphic to the lattice point group.

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: $(\frac{\pi}{a},\frac{\pi}{a},\frac{\pi}{a})$.



4.4.2. 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. 4.16.

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:

Γ	$\frac{2\pi}{a}(0,0,0)$
H	$\frac{2\pi}{a}(0,0,1)$
P	$\frac{2\pi}{a}(½,½,½)$
N	$\frac{2\pi}{a}(½,½,0)$.

Using Miller indices for the directions:

Δ :	Γ →H	(parallel to <100>)
Λ :	Γ →P	(parallel to <111>)

D:	N→P	(parallel to <100>)
Σ :	Γ →N	(parallel to <110>)
G:	N→H	(parallel to $<1\bar{1}0>$).

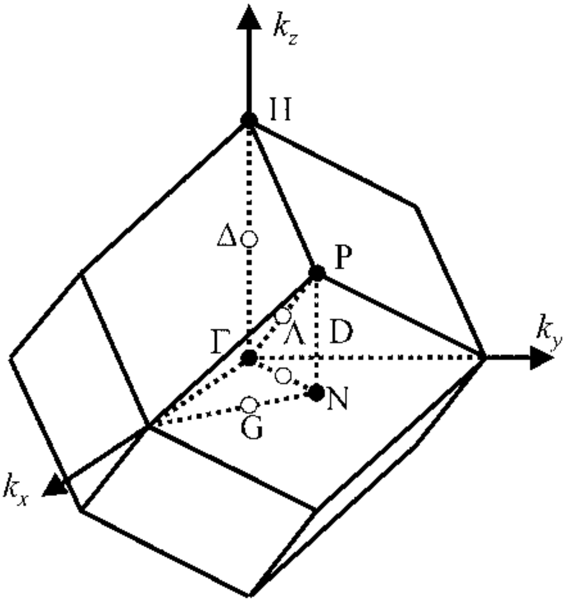


Fig. 4.16. First Brillouin zone of a bcc lattice.

4.4.3. First Brillouin zones of a few semiconductors

As discussed in Chapter 1, 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.

Fig. 4.17 depicts the *E-k* diagrams characterizing the band structures in Ge (Fig. 4.17(a)), Si (Fig. 4.17(b)), and GaAs (Fig. 4.17(c)). 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. 4.17 are two-direction composite diagrams. The <111> direction is toward point L, 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.

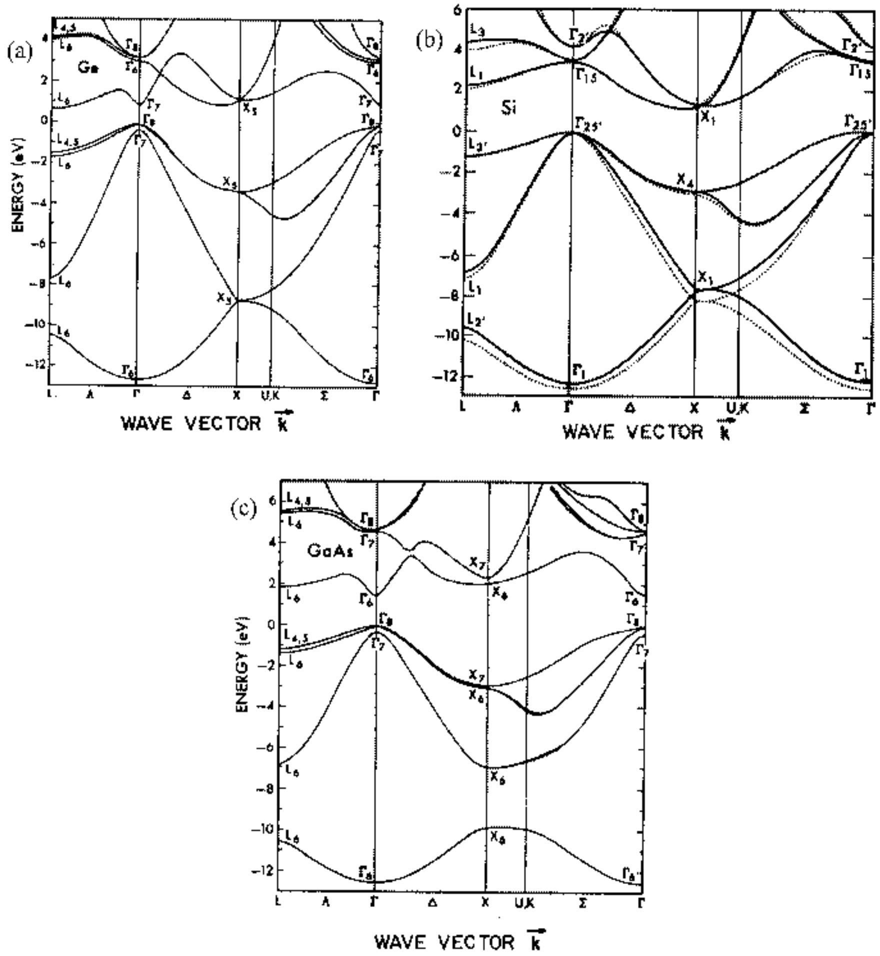


Fig. 4.17. E-k diagram of a few semiconductor crystals: (a) Ge, (b) Si, and (c) GaAs. The structures of the conduction and valence bands are plotted. The origin of the energy is chosen to be at the top of the valence band. [Reprinted figure with permission from Chelikowsky, J.R. and Cohen, M.L., Physical Review B Vol. 14, pp. 559 & 566, 1976. Copyright 1976 by the American Physical Society.]

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 sub-bands. 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. 4.17(b) and the maximum of the third band is only 0.044 eV below E_V at k=0.

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 Appendix A.8).

Conduction band

There are a number of sub-bands in each of the conduction bands shown in Fig. 4.17. These sub-bands 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 sub-bands 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. 4.17(a). 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 *four-fold degeneracy* or a *multiplicity of 4*. 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 six-fold 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. 4.17, 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 is 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.

4.5. 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. Fig. 4.18 and Fig. 4.19 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.

The band structures in the $\Gamma \rightarrow$X, $\Gamma \rightarrow$K and $\Gamma \rightarrow$L directions are nearly parabolic, and are therefore similar to the free electron case. Electrons in aluminum thus behave almost like free electrons.

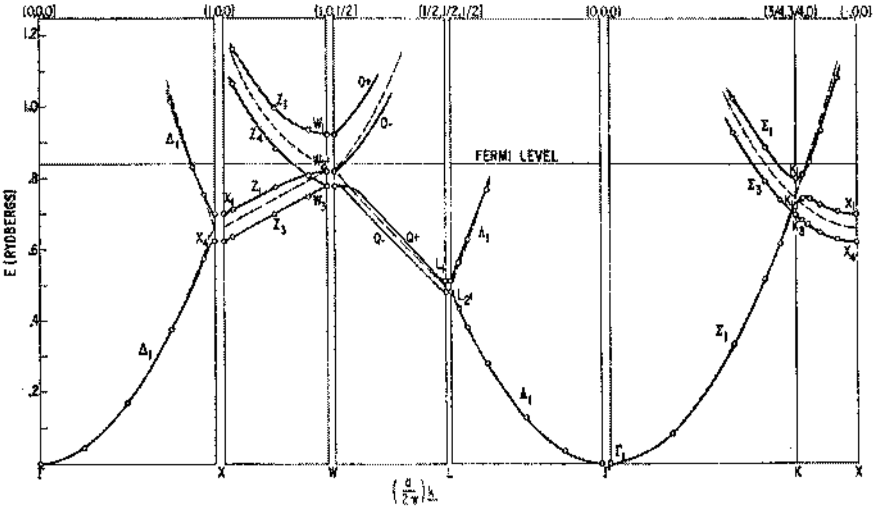


Fig. 4.18. Electron band structure diagram of aluminum. The energy is expressed in units of Rydberg. The dashed lines show the energy bands for a free electron. [Reprinted figure with permission from Segall, B., The Physical Review Vol. 124, p. 1801, 1961. Copyright 1961 by the American Physical Society.]

The dashed lines in Fig. 4.18 and Fig. 4.19 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 are 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 10). 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.

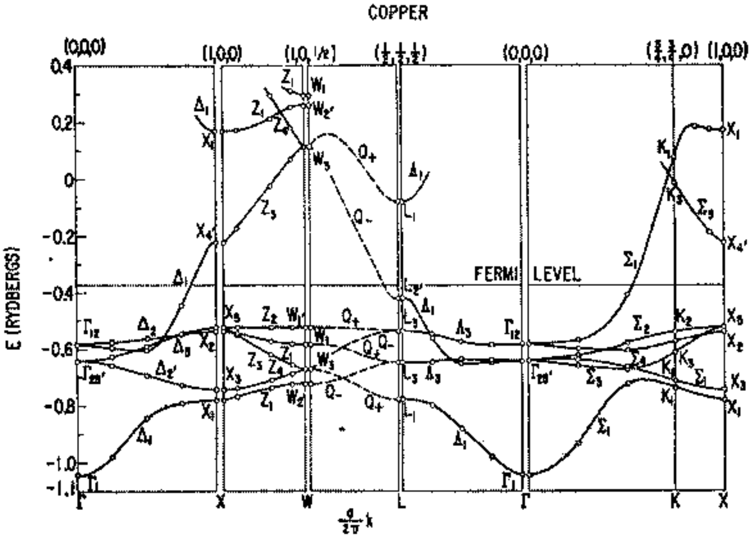


Fig. 4.19. Electron band structure diagram of copper. The energy is expressed in units of Rydberg. There are a few narrow bands located just below the Fermi energy, corresponding to the 4d-orbitals in copper. [Reprinted figure with permission from Segall, B., The Physical Review Vol. 125, p. 113, 1962. Copyright 1962 by the American Physical Society.]

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.

4.6. 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, bandgap, Bloch wavefunction, effective mass, Fermi energy, 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, has also been presented, and the main features were described and compared to those of semiconductors.

References

Chelikowsky, J.R. and Cohen, M.L., "Nonlocal pseudopotential calculations for the electronic structure of eleven diamond and zinc-blende semiconductors," *Physical Review* B 14, pp. 556-582, 1976.

Segall, B., "Energy bands of aluminum," *The Physical Review* 124, pp. 1797-1806, 1961.

Segall, B., "Fermi surface and energy bands of copper," *The Physical Review* 125, pp. 109-122, 1962.

Further reading

Altmann, S.L., *Band Theory of Solids: an Introduction from the Point of View of Symmetry*, Oxford University Press, New York, 1991.

Bastard, G., *Wave Mechanics Applied to Semiconductor Heterostructures*, Halsted Press, New York, 1988.

Christman, J.R., *Fundamentals of Solid State Physics*, John Wiley & Sons, New York, 1988.

Kittel, C., *Introduction to Solid State Physics*, John Wiley & Sons, New York, 1976.

McKelvey, J.P., *Solid State and Semiconductor Physics,* Harper and Row, New York, 1966.

Powell, J.L. and Crasemann, B., *Quantum Mechanics,* Addison-Wesley, Reading, MA, 1961.

Rosenberg, H.M., *The Solid State: an Introduction to the Physics of Crystals for Students of Physics, Materials Science, and Engineering*, Oxford University Press, New York, 1988.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

Talley, H.E. and Daugherty, D.G., *Physical Principles of Semiconductor Devices,* Iowa State University Press, Ames, 1976.

Tu, K.N, Mayer, J.W., and Feldman, L.C., *Electronic Thin Films Science for Electrical Engineers and Materials Scientists*, Macmillan Publishing, New York, 1992.

Problems

1. *Equations of motion of an electron in the presence of an electric field.*

   Assuming a dispersion relation: $\varepsilon = \varepsilon_C + \frac{\hbar^2}{ma^2}[1-\cos(ka)]$

   (a) Calculate the velocity of the electron at k=π/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=-π/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=0.2m_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 -π /a. Then, the electron is again accelerated to k=π/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.

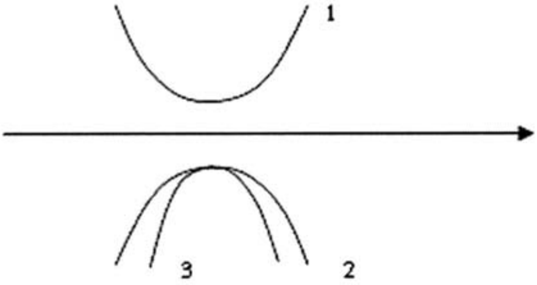
   (a) Show that the period of the Bloch oscillation is given by $\tau = \frac{2\pi\hbar}{qEa}$,

   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 $\varepsilon(\vec{k}) = -\alpha - 8\gamma\cos(\frac{k_x a}{2})$.
   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 $\varepsilon(\vec{k}) = \frac{\hbar^2}{2m} A\left(k_x^2 - \frac{a^2}{2\pi^2} k_x^4\right)$.
   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. 4.15.

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 (C_B) and in the valence band (V_B). Then, the probability that a state is filled at the conduction band edge (E_C) is equal to the probability 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.

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 ($m^*=1.08m_0$) 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.

5. Phonons

5.1. Introduction

In 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 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.

5.2. Interaction of atoms in crystals: origin and formalism

We saw in section 2.5, 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. 5.1. This figure shows a minimum energy for a specific atomic separation, which we understood was true at all time.

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 afterwards. 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.

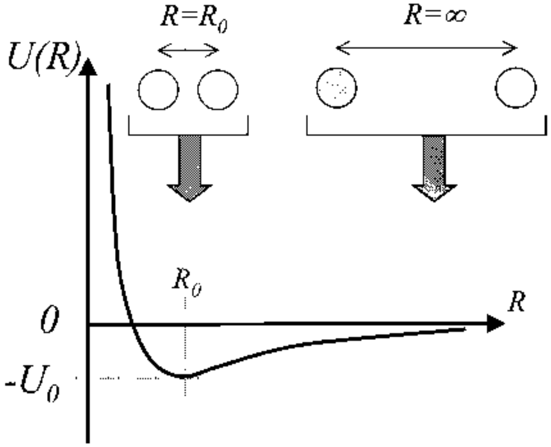


Fig. 5.1. Potential energy of two neighboring atoms in a crystal as a function of the interatomic spacing. When the two atoms are very far away from each other, they do not interact and the interaction potential energy nears zero. When they get closer to one another, they are attracted to each other 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.

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 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. 5.1 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):

Eq. (5.1)

$$U(R)=U(R_0)+\left(\frac{dU}{dR}\right)_{R_0}(R-R_0)+\frac{1}{2}\left(\frac{d^2U}{dR^2}\right)_{R_0}(R-R_0)^2+\frac{1}{6}\left(\frac{d^3U}{dR^3}\right)_{R_0}(R-R_0)^3+\ldots$$

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 $r=R_0$. $(R\text{-}R_0)$ 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. (5.1).

By denoting $U_0=-U(R_0)$ and $x=R-R_0$ the displacement, Eq. (5.1) can be rewritten:

Eq. (5.2) $$U(x)+U_0=\frac{1}{2}C_1x^2+C_2x^3+...$$

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. (5.2), $\frac{1}{2}C_1u^2$, 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. (5.2) and denote *$C=C_1$*:

Eq. (5.3) $$U(x)+U_0\approx\frac{1}{2}Cx^2$$

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.

5.3. One-dimensional monoatomic 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. 5.2. 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.

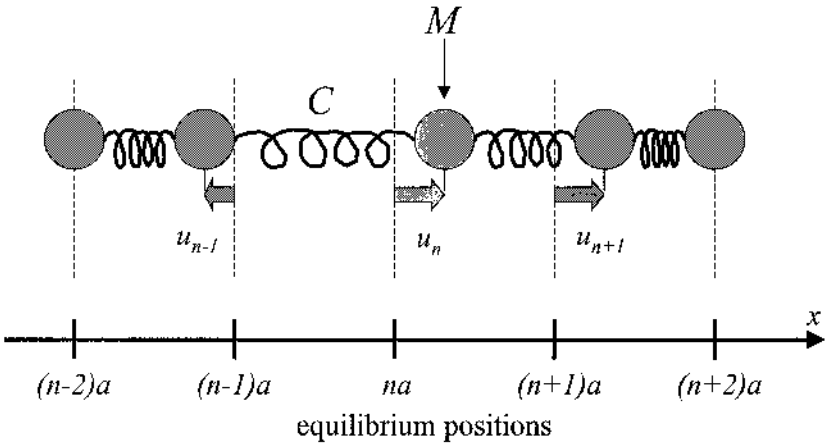


Fig. 5.2. Model for the interaction of identical atoms in a harmonic crystal. The relative movement of the atoms is modeled by a spring such that atoms displaced from their equilibrium positions are forced back by neighboring atoms. The displacement can travel like a wave throughout the lattice.

5.3.1. 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. 5.2. In other words, the force acted on the n^{th} atom:

- by the $(n\text{-}1)^{\text{th}}$ atom is $F_{n,n-1} = -C(u_n - u_{n-1})$
- and 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. (5.3) and corresponds to the first level of approximation for the interactions between atoms. The resultant force acting on the n^{th} atom is therefore:

Eq. (5.4) $$F_n = F_{n,n-1} + F_{n,n+1} = -C(2u_n - u_{n-1} - u_{n+1})$$

The equation of motion for the n^{th} atom is then expressed using classical mechanics Newton's mass action law:

Eq. (5.5) $$M\frac{d^2u_n}{dt^2} = F_n = -C(2u_n - u_{n-1} - u_{n+1})$$

where M is the mass and $\frac{d^2u_n}{dt^2}$ 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. (5.5) in the form of traveling waves such as:

Eq. (5.6) $$u_n(t) = A\exp[i(kan - \omega t)]$$

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:

Eq. (5.7) $$\begin{aligned} u_{n+1}(t) &= A\exp[i(ka(n+1) - \omega t)] \\ &= A\exp\left[i\left(kan - \omega\left(t - \frac{ka}{\omega}\right)\right)\right] = u_n\left(t - \frac{ka}{\omega}\right) \end{aligned}$$

which shows that the value of the displacement $u_{n+1}(t)$ at the $(n+1)^{th}$ 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:

Eq. (5.8) $$\frac{a}{ka/_{\omega}}=\frac{\omega}{k}$$

The wavelength λ and frequency υ of the traveling wave are related to the wavenumber or angular frequency through the definition relations:

Eq. (5.9) $$\begin{cases}\lambda=\dfrac{2\pi}{k}\\ \upsilon=\dfrac{\omega}{2\pi}\end{cases}$$

5.3.2. Boundary conditions

Before solving the equation of motion in Eq. (5.5), 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 section 4.3. 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 do not affect the motion of all the other atoms inside the sample.

Because of the general exponential expression of $u_n(t)$ (Eq. (5.6)), these conditions lead to the discretization of the wavenumber *k*, similar to what was obtained in Chapter 4:

Eq. (5.10) $$k=k_m=\frac{2\pi}{a}\frac{m}{N}$$

where $m=0,\pm1...$ 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:

Eq. (5.11) $$u_n^{'}(t) = A\exp[i(k'an - \omega t)] \\ = A\exp[i2\pi n + i(kan - \omega t)] = u_n(t)$$

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:

Eq. (5.12) $$\frac{1}{2}\left(-\frac{2\pi}{a}\right) \leq k \leq \frac{1}{2}\left(\frac{2\pi}{a}\right)$$

And all the physical properties of our one-dimensional crystal that depend 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 Chapter 1 and 4 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.

5.3.3. Phonon dispersion relation

Now we can solve the equation of motion in Eq. (5.5), by substituting Eq. (5.6) into it:

$$-M\omega^2 A\exp[i(kan - \omega t)] = -C(2 - e^{-ika} - e^{ika})A\exp[i(kan - \omega t)]$$

which successively becomes, after simplification of the exponential and the constant A:

$$-M\omega^2 = -C(2 - e^{-ika} - e^{ika})$$

or:

Eq. (5.13) $$\omega^2 = \frac{2C}{M}(1 - \cos ka) = \frac{4C}{M}\sin^2\frac{ka}{2}$$

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:

Eq. (5.14) $$\omega = \omega_{\max}\left|\sin\frac{ka}{2}\right|$$

where $\omega_{\max} = \sqrt{\frac{4C}{M}}$. This relation is called the phonon dispersion relation and is plotted in Fig. 5.3.

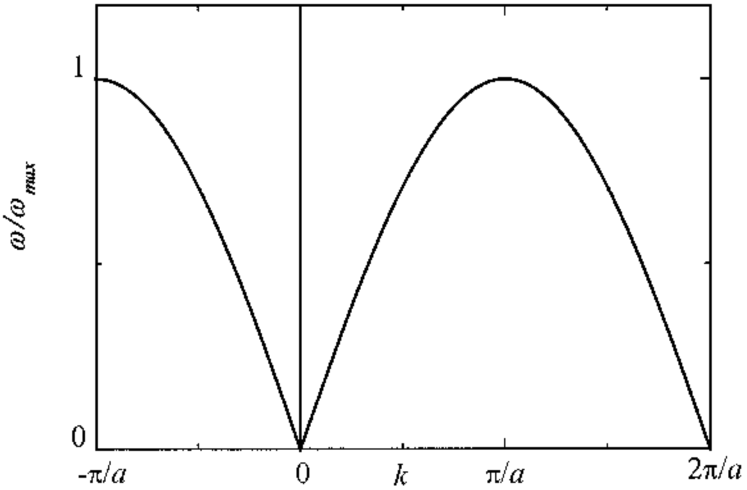


Fig. 5.3. Phonon dispersion relation in a one-dimensional monoatomic harmonic crystal, expressed through the dependence of the angular frequency as a function of the wavenumber k.

We see that the solutions of Eq. (5.5) of the traveling-wave type exist only if the relation in Eq. (5.14) is satisfied by the wavenumber k and the angular frequency ω of the traveling wave. The frequency and wavenumber of the traveling wave 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 section 5.6.

For a small wavenumber ($k\rightarrow 0$), i.e. in the long wave limit, Eq. (5.14) becomes:

Eq. (5.15) $\omega = \omega_{\max} \frac{a}{2} k$

where we have used the approximation for the sine function: $\sin(x) \approx x$ for $x \to 0$, which is in fact the Taylor expansion of the sine function near zero (see Eq. (5.1)). Eq. (5.15) 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.

5.4. One-dimensional diatomic harmonic crystal

5.4.1. Formalism

In the previous sections we have discussed the motion of atoms in a one-dimensional monoatomic 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, 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 in Eq. (5.2), 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 lattice (i.e. longitudinal vibration). All these parameters and their simplifications are summarized in Fig. 5.4.

Two coupled sets of equations of motion, similar to Eq. (5.5), 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:

Eq. (5.16)
$$\begin{cases} M_1 \dfrac{d^2u_n}{dt^2} = -C(2u_n - v_{n-1} - v_n) \\ M_2 \dfrac{d^2v_n}{dt^2} = -C(2v_n - u_n - u_{n+1}) \end{cases}$$

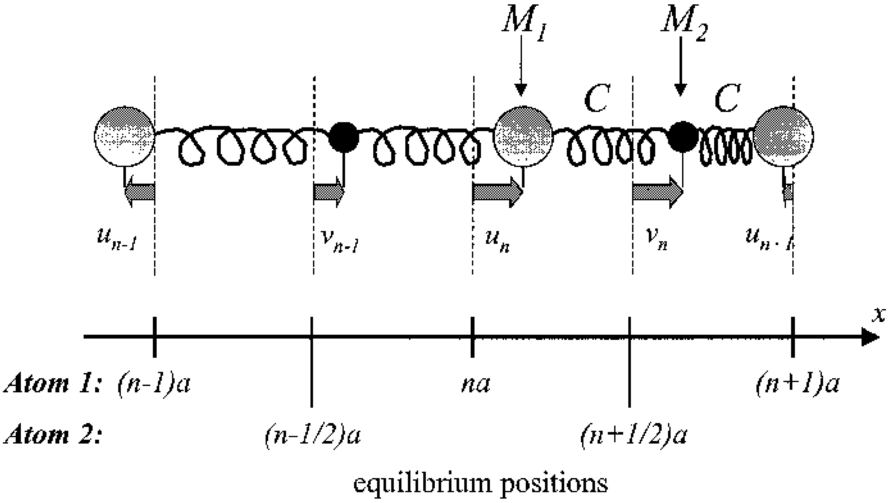


Fig. 5.4. One-dimensional model for the interaction of atoms in a diatomic harmonic crystal structure with atom masses M_1 and M_2. It is assumed here that all the springs have the same constant.

Here again, we seek solutions to the set of Eq. (5.16) in the form of traveling waves with the same wavenumber k and angular frequency ω:

Eq. (5.17)
$$\begin{cases} u_n(t) = A\exp[i(kan - \omega t)] \\ v_n(t) = B\exp\left[i\left(ka\left(n + \frac{1}{2}\right) - \omega t\right)\right] \end{cases}$$

where A and B are the amplitude of the displacements.

5.4.2. Phonon dispersion relation

Substituting these traveling wave expressions into Eq. (5.16), we obtain:

$$\begin{cases} -M_1\omega^2 A\exp[i(kan-\omega t)] \\ \quad = -C\left(2A\exp[i(kan-\omega t)] - B\exp\left[i\left(ka\left(n-\frac{1}{2}\right)-\omega t\right)\right] - B\exp\left[i\left(ka\left(n+\frac{1}{2}\right)-\omega t\right)\right]\right) \\ -M_2\omega^2 B\exp\left[i\left(ka\left(n+\frac{1}{2}\right)-\omega t\right)\right] \\ \quad = -C\left(2B\exp\left[i\left(ka\left(n+\frac{1}{2}\right)-\omega t\right)\right] - A\exp[i(kan-\omega t)] - A\exp[i(ka(n+1)-\omega t)]\right) \end{cases}$$

Simplifying by $\exp[i(kan-\omega t)]$ the first expression and $\exp\left[i\left(ka\left(n+\frac{1}{2}\right)-\omega t\right)\right]$ the second, we get:

$$\begin{cases} -M_1\omega^2 A = -C\left(2A - B\exp\left(-\frac{ika}{2}\right) - B\exp\left(+\frac{ika}{2}\right)\right) \\ -M_2\omega^2 B = -C\left(2B - A\exp\left(-\frac{ika}{2}\right) - A\exp\left(+\frac{ika}{2}\right)\right) \end{cases}$$

After re-arranging the terms with A and those with B:

$$\begin{cases} A\left[2C - M_1\omega^2\right] - BC\left[\exp\left(-\frac{ika}{2}\right) + \exp\left(+\frac{ika}{2}\right)\right] = 0 \\ -AC\left[\exp\left(-\frac{ika}{2}\right) + \exp\left(+\frac{ika}{2}\right)\right] + B\left[2C - M_2\omega^2\right] = 0 \end{cases}$$

Expressing the sum of exponentials with trigonometric functions, we get:

Eq. (5.18)
$$\begin{cases} A\left[2C - M_1\omega^2\right] - B\left[2C\cos\left(\frac{ka}{2}\right)\right] = 0 \\ -A\left[2C\cos\left(\frac{ka}{2}\right)\right] + B\left[2C - M_2\omega^2\right] = 0 \end{cases}$$

This system of equation has a non-zero solution, i.e. A and B not both equal to zero, if and only if the determinant of the system is zero:

Eq. (5.19)

$$\left[2C - M_1\omega^2\right]\left[2C - M_2\omega^2\right] - \left[2C\cos\left(\frac{ka}{2}\right)\right]\left[2C\cos\left(\frac{ka}{2}\right)\right] = 0$$

which becomes after developing the products:

$$M_1M_2\omega^4 - 2C(M_1 + M_2)\omega^2 + 4C^2 - 4C^2\cos^2\left(\frac{ka}{2}\right) = 0$$

or:

Eq. (5.20) $$M_1M_2\omega^4 - 2C(M_1 + M_2)\omega^2 + 4C^2\sin^2\left(\frac{ka}{2}\right) = 0$$

This equation is of the form $\alpha\omega^4 - 2\beta\omega^2 + \gamma = 0$, with α, β, and γ>0, and has two solutions for ω^2, denoted ω_+^2 and ω_-^2 such that:

Eq. (5.21) $$\omega_\pm^2 = \frac{\beta \pm \sqrt{\beta^2 - \alpha\gamma}}{\alpha}$$

Therefore, the solutions of Eq. (5.20) 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:

$$\omega_\pm^2(k) = C\left(\frac{M_1 + M_2}{M_1M_2}\right) \pm C\sqrt{\left(\frac{M_1 + M_2}{M_1M_2}\right)^2 - \frac{4\sin^2\left(\frac{ka}{2}\right)}{M_1M_2}}$$

Using the trigonometric identity $\cos(2x) = 1 - 2\sin^2(x)$, this equation becomes:

Eq. (5.22) $$\omega_{\pm}^{2}(k)=C\left(\frac{M_1+M_2}{M_1M_2}\right)\left[1\pm\sqrt{1-\frac{2M_1M_2}{(M_1+M_2)^2}(1-\cos(ka))}\right]$$

which constitutes the phonon dispersion relation in the model considered, similar to that obtained in Eq. (5.14). 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. 5.5. These plots represent the so-called phonon spectrum of a one-dimensional diatomic harmonic crystal.

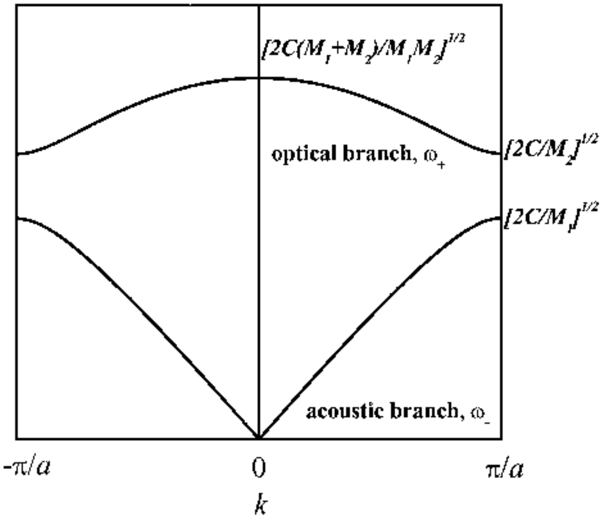


Fig. 5.5. Optical and acoustic branches in the dispersion relation.

The values for $\omega_+(k)$ and $\omega_-(k)$ at $k=0$ and $k=\pm\frac{\pi}{a}$ can be easily calculated from Eq. (5.22) (note that we have chosen $M_1>M_2$). The top curve in Fig. 5.5 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. (5.22). To do so, we use start by using an approximate expression for the cosine function in the Eq. (5.22):

$$\cos(ka) \approx 1 - \frac{1}{2}(ka)^2$$

This approximation is in fact the Taylor expansion of the cosine function near zero (see Eq. (5.1)). We therefore obtain successively:

$$1 - \cos(ka) \approx \frac{1}{2}(ka)^2$$

$$\sqrt{1 - \frac{2M_1M_2}{(M_1+M_2)^2}(1-\cos(ka))} \approx \sqrt{1 - \frac{M_1M_2}{(M_1+M_2)^2}(ka)^2}$$

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). Eq. (5.22) can then be approximated by the following expression:

Eq. (5.23) $$\omega_\pm^2(k) \approx C\left(\frac{M_1+M_2}{M_1M_2}\right)\left[1 \pm \left(1 - \frac{M_1M_2}{2(M_1+M_2)^2}(ka)^2\right)\right]$$

Consequently, in the long wave limit, the angular frequency of the acoustic phonon branch can be written as:

$$\omega_-^2(k) \approx C\left(\frac{M_1+M_2}{M_1M_2}\right)\left[\frac{M_1M_2}{2(M_1+M_2)^2}(ka)^2\right]$$

Eq. (5.24) $$\omega_-(k) \approx k\sqrt{\frac{Ca^2}{2(M_1+M_2)}}$$

which means that the angular frequency $\omega_-(k)$ in the acoustic phonon branch is proportional to the wavenumber *k*, similar the result obtained in Eq. (5.15). 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. (5.24):

Eq. (5.25) $$\omega_+^2(k) \approx C\left(\frac{M_1+M_2}{M_1M_2}\right)[1+1] = 2C\left(\frac{M_1+M_2}{M_1M_2}\right)$$

which shows that the angular frequency $\omega_+(k)$ in the optical phonon branch is a 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. (5.17) can be taken for two different values, depending on the branch chosen, calculated from either one of Eq. (5.18):

Eq. (5.26) $$\left(\frac{B}{A}\right)_\pm = \frac{2C - M_1\omega_\pm^2}{2C\cos\left(\frac{ka}{2}\right)}$$

Again, in the long wave limit (*k*→*0*) and for the acoustic phonon branch, we have $\omega_-(k)\to 0$ as seen from Eq. (5.24) and $\cos\left(\frac{ka}{2}\right)\to 1$, so that:

$$\text{Eq. (5.27)}\quad \left(\frac{B}{A}\right)_- \to \frac{2C}{2C}=1$$

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. 5.6.

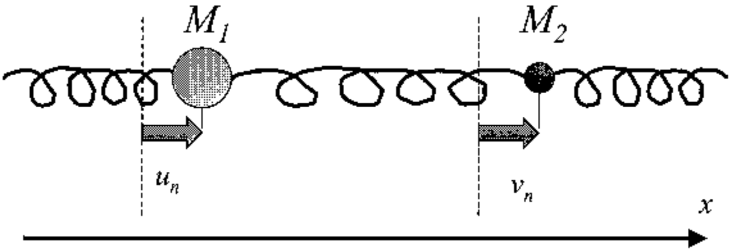


Fig. 5.6. Atomic vibrations in a one-dimensional diatomic harmonic crystal, corresponding to the acoustic phonon branch. In this configuration, the two atoms forming the unit cell move in the same direction at the same time.

In the long wave limit (*k*→*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. (5.25) and therefore, by substituting into Eq. (5.26):

$$\text{Eq. (5.28)}\quad \left(\frac{B}{A}\right)_+ \to \frac{2C-M_1\dfrac{2C}{\left(\dfrac{M_1M_2}{M_1+M_2}\right)}}{2C}=-\frac{M_1}{M_2}$$

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. 5.7. Thus, optic 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.

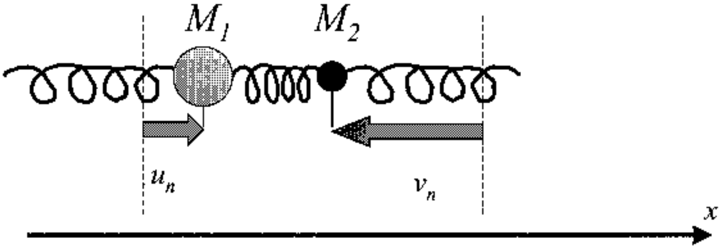


Fig. 5.7. Atomic vibrations in a one-dimensional diatomic harmonic crystal, corresponding to the optical phonon branch. In this configuration, the two atoms forming the unit cell move in opposite directions at the same time.

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 sub-section 2.5.6. 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 photons. Fig. 5.8 shows the crossing of the dispersion relation for both acoustic and optic phonons with a photon. Because the photon and optic phonon curves cross, energy and momentum can be exchanged. An optic 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.

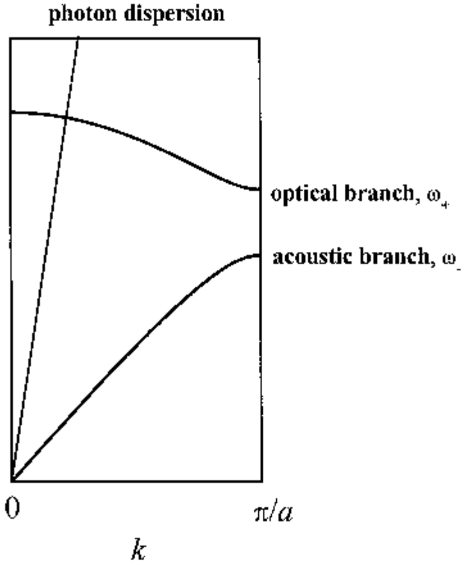


Fig. 5.8. The dispersion curves for a photon and an acoustic and optic phonon. The optic branch crosses with the photon branch, allowing for energy and momentum conservation.

5.5. Extension to three-dimensional case

5.5.1. 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 $\overrightarrow{u_{\vec{R}}}(t)$. Moreover, a wavevector $\vec{k}$ must be used, similarly to the way it was done in Chapter 4 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. (5.6), becomes in the three-dimensional case now:

Eq. (5.29) $$\overrightarrow{u_{\vec{R}}}(t)=\vec{A}\exp\left[i\left(\vec{k}.\vec{R}-\omega t\right)\right]$$

where $\vec{A}$ is the amplitude vector of the displacement, and $\vec{k}.\vec{R}$ 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. 5.9: 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 non-zero finite value in the long wave limit. Each type of phonon 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 .

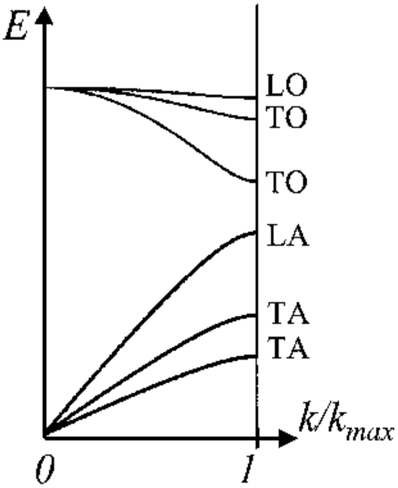


Fig. 5.9. Typical phonon dispersion spectrum for a three-dimensional diatomic lattice (s=2).

For example, in Fig. 5.9, 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. Fig. 5.9 shows a typical example for *s*=2. A monoatomic Bravais lattice (*s*=1) can only have acoustic phonon branches.

Fig. 5.10 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.

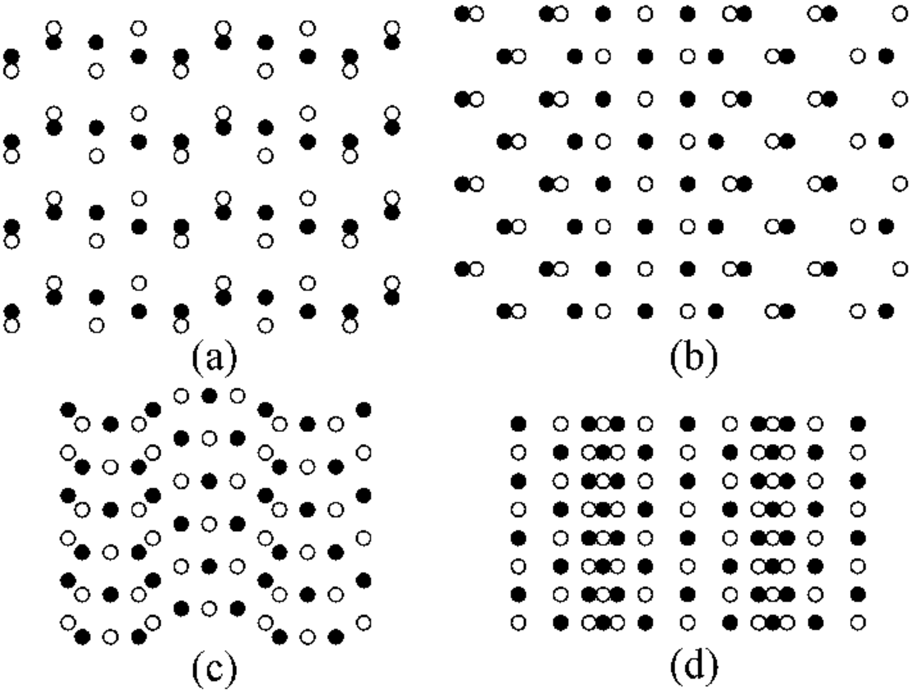


Fig. 5.10. The propagation of the four different phonon modes through a lattice: (a) transverse optic, (b) longitudinal optic, (c) transverse acoustic, and (d) longitudinal acoustic.

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 towards 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 remain fixed, while the pair as a whole is displaced relative to other atoms pairs.

For all the modes, the frequency of vibration is directly proportional to the mass of each atom, the bond length of the atomic pairs, and the electronegativity of each atom.

5.5.2. Silicon

Silicon crystals only have 2 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 not a restoring force like that in GaAs.

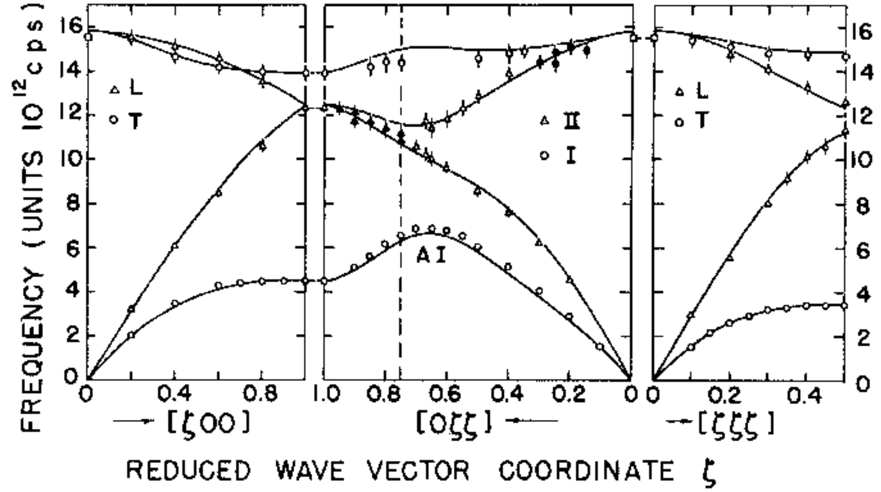


Fig. 5.11. Phonon dispersion relation for Silicon in three crystal directions. Solid lines are calculated. Data points: open circles represent transverse (T) modes, open triangles longitudinal (L) modes, and solid points undetermined polarization modes. [Reprinted with permission from Inelastic Scattering of Neutrons in Liquids and Solids Vol. 2, Dolling, G., "Lattice Vibrations in Crystals with the Diamond Structure," p. 41. Copyright 1963, International Atomic Energy Association.]

5.5.3. 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 5 electrons to the bonds compared to gallium atoms, which contribute 3. Consequently, the electrons spend on average more time near the arsenic atoms resulting in the arsenic atoms being 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.

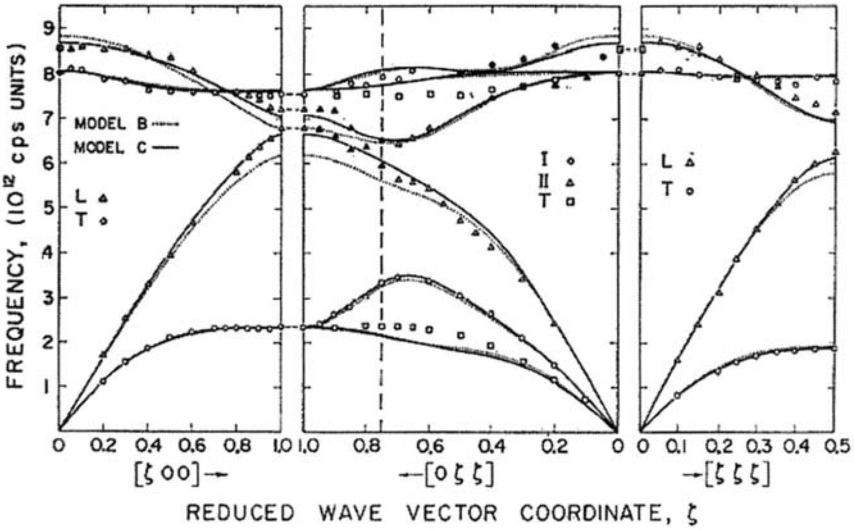


Fig. 5.12. Phonon dispersion relation for Gallium Arsenide in three crystal directions. Dotted and solid lines denote calculated values. Solid points denote undetermined polarization modes. [Reprinted figure with permission from Waugh, J.L.T. and Dolling, G., The Physical Review Vol. 132, p. 2411, 1963. Copyright 1963 by the American Physical Society.]

5.6. Phonons

In Chapter 4, 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 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 quasi-particle or phonon with a momentum and a (quantized) energy such that:

Eq. (5.30) $$\begin{cases} \vec{p} = \hbar\vec{k} \\ E = \hbar\omega(\vec{k}) \end{cases}$$

This is analogous to the quantization of the electromagnetic field discussed in sub-section 3.1.1. The energy in Eq. (5.30) 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 for example 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:

Eq. (5.31) $$\overrightarrow{v_g} = \frac{\partial\omega(\vec{k})}{\partial\vec{k}} = \nabla_{\vec{k}}\,\omega(\vec{k})$$

In Cartesian coordinates with unit vectors $(\vec{x}, \vec{y}, \vec{z})$, this relation can be written as:

Eq. (5.32) $$\overrightarrow{v_g} = \frac{\partial\omega(k_x, k_y, k_z)}{\partial k_x}\vec{x} + \frac{\partial\omega(k_x, k_y, k_z)}{\partial k_y}\vec{y} + \frac{\partial\omega(k_x, k_y, k_z)}{\partial k_z}\vec{z}$$

In this quantum picture, the propagation of harmonic lattice waves, i.e. up to the second order term in Eq. (5.2), is equivalent to the free movement of non-interacting phonon quasi-particles, 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:

Eq. (5.33) $$N_{\vec{k}} = \frac{1}{\exp\left(\frac{\hbar\omega(\vec{k})}{k_b T}\right) - 1}$$

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. (5.33) can be approximated by:

Eq. (5.34) $$\exp\left(\frac{\hbar\omega(\vec{k})}{k_b T}\right) \approx 1 + \frac{\hbar\omega(\vec{k})}{k_b T}$$

where we have used the approximation $\exp(x) \approx 1 + x$ for $x \to 0$ (again this comes from the Taylor expansion of $\exp(x)$ 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 Eq. (5.1) and Eq. (5.2).

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: $N(E) \approx \exp\left(-\frac{E}{k_b T}\right)$.

5.7. 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 section 5.3, 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. (5.8):

Eq. (5.35) $$v_{\mathrm{ph}} = \frac{\omega}{k}$$

Using Eq. (5.13) and Eq. (5.14), we obtain:

Eq. (5.36) $$v_{\mathrm{ph}} = \sqrt{\frac{4C}{M}} \left| \frac{\sin\left(\frac{ka}{2}\right)}{k} \right| = a\sqrt{\frac{C}{M}} \left| \frac{\sin\left(\frac{ka}{2}\right)}{\frac{ka}{2}} \right| = v_0 \left| \frac{\sin\left(\frac{ka}{2}\right)}{\frac{ka}{2}} \right|$$

where:

Eq. (5.37) $$v_0 = a\sqrt{\frac{C}{M}}$$

Therefore:

Eq. (5.38) $$v_{\mathrm{ph}} = v_0 \left| \frac{\sin\left(\frac{ka}{2}\right)}{\frac{ka}{2}} \right|$$

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. 5.13, 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:

Eq. (5.39) $$v_g = \left|\frac{d\omega}{dk}\right|$$

Using Eq. (5.13) and Eq. (5.14), we obtain:

Eq. (5.40) $$v_g = \sqrt{\frac{4C}{M}}\frac{a}{2}\left|\cos\left(\frac{ka}{2}\right)\right| = a\sqrt{\frac{C}{M}}\left|\cos\left(\frac{ka}{2}\right)\right|$$

and therefore:

Eq. (5.41) $$v_g = v_0\left|\cos\left(\frac{ak}{2}\right)\right|$$

The group velocity is also plotted in Fig. 5.13. We see that this quantity drops to zero when $k \to \frac{\pi}{a}$, i.e. at boundary of the first Brillouin zone.

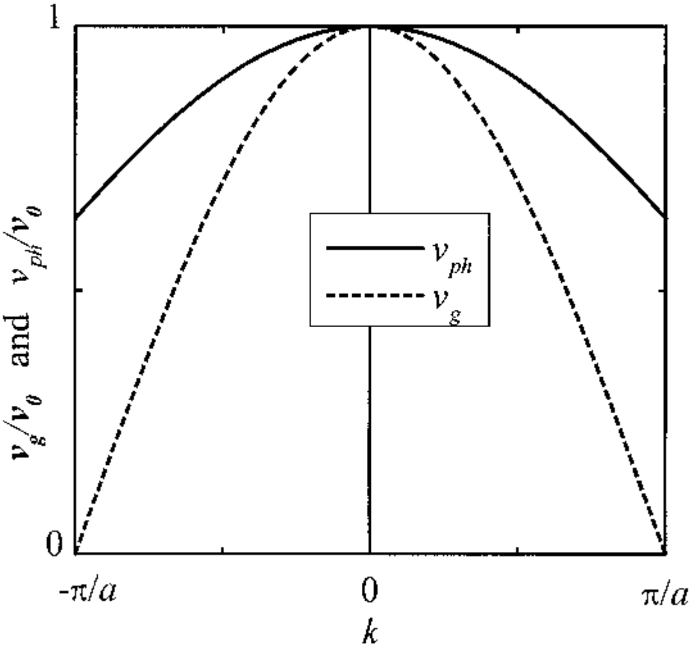


Fig. 5.13. Phase and group velocities versus wavenumber k.

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 form 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:

$$C = M\,\frac{v_0^2}{a^2} = \left(28\times1.67264\times10^{-27}\right)\times\frac{\left(2.2\times10^3\right)^2}{\left(5.43\times10^{-10}\right)^2}$$

$$\approx 0.77\,N.m^{-1}$$

From Eq. (5.37), 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:

Eq. (5.42) $$v_s = \sqrt{\frac{B}{\rho}}$$

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

Eq. (5.43) $$B = -\frac{\Delta p}{\frac{\Delta V}{V}}$$

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. (5.42) 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.

5.8. 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 monoatomic and diatomic harmonic lattices. Several important concepts have been introduced such as the lattice vibration 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 quasi-particles, or phonons, with a momentum and energy and which obey Bose-Einstein statistics.

References

Dolling, G., "Lattice Vibrations in crystals with the diamond structure," in *Inelastic Scattering of Neutrons in Liquids and Solids* 2, International Atomic Energy Agency, Vienna, Austria, p. 41, 1963.

Waugh, J.L.T. and Dolling, G., "Crystal dynamics of gallium arsenide," *The Physical Review* 132, p. 2411, 1963.

Further reading

Ashcroft, N.W. and Mermin, N.D., *Solid State Physics*, Holt, Rinehart and Winston, New York, 1976.

Born, M. and Huang, K., *Dynamical Theory of Crystal Lattices*, Clarendon Press, Oxford, 1954.

Cochran, W., *The Dynamics of Atoms in Crystals*, Edward Arnold Limited, London, 1973.

Cohen, M.M., *Introduction to the Quantum Theory of Semiconductors*, Gordon and Breach, New York, 1972.

Ibach, H. and Lüth, H., *Solid-State Physics: an Introduction to Theory and Experiment*, Springer-Verlag, New York, 1990.

Kasap, S.O., *Principles of Engineering Materials and Devices*, McGraw-Hill, New York, 1997.

Kittel, C., *Introduction to Solid State Physics*, John Wiley & Sons, New York, 1976.

Ferry, D.K., *Semiconductors*, Macmillan, New York, 1991.

Maxwell, J.C., *Matter and Motion*, Dover, New York, 1952.

Peyghambarian, N., Koch, S.W., and Mysyrowicz, A., *Introduction to Semiconductor Optics*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

Sapoval, B. and Hermann, C., *Physics of Semiconductors*, Springer-Verlag, New York, 1995.

Problems

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 optic and acoustic phonon branches. Solve Eq. (5.22) for the case when $k=\pi/a$.

3. The one-dimensional monatomic harmonic crystal (section 5.3) is in fact a particular case of the diatomic model described in section 5.4, 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. (5.22) 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 $k=0$ was 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 section 5.4, we calculated the ratio of the displacement amplitudes A and B for the long wave limit ($k\rightarrow 0$) for both the optic and acoustic phonon branches and then determined the displacement of the atoms with respect to each other. Calculate Eq. (5.26), 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 optic 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 $N(\omega)=\frac{1}{\exp\left(\frac{\hbar\omega}{k_bT}\right)-1}$ for at least five values of T to show its evolution with increasing temperatures. For each one, plot the function $F(\omega)=\frac{k_bT}{\hbar\omega}$ and show that it is a good approximation for $N(\omega)$ for high temperatures, i.e. $k_bT >> \hbar\omega$.

10. Let us model a rigid bar as a linear monatomic chain of atoms, as in section 5.3 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-section 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 section 5.7, is equal to the ratio: $\sqrt{\frac{E_Y}{M_V}}$.

11. From the speed of sound equation, $v=(B/\rho)^{1/2}$, 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?

6. Thermal Properties of Crystals

6.1. Introduction

In Chapter 5, 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 quasi-particle 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 Chapter 5 to interpret the thermal properties of crystals, in particular their heat capacity, thermal expansion and thermal conductivity.

6.2. Phonon density of states (Debye model)

6.2.1. 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 (section 6.3). The model relies on a simplification of the phonon dispersion relation (see for example Eq. (5.22), Fig. 5.5 or Fig. 5.9). 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:

Eq. (6.1) $$\omega_n(\vec{k}) = v_n\left|\vec{k}\right| = v_n k$$

where n (=l or t) is an index; k is the norm or length of the wavevector $\vec{k}$, 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. 6.1. 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.

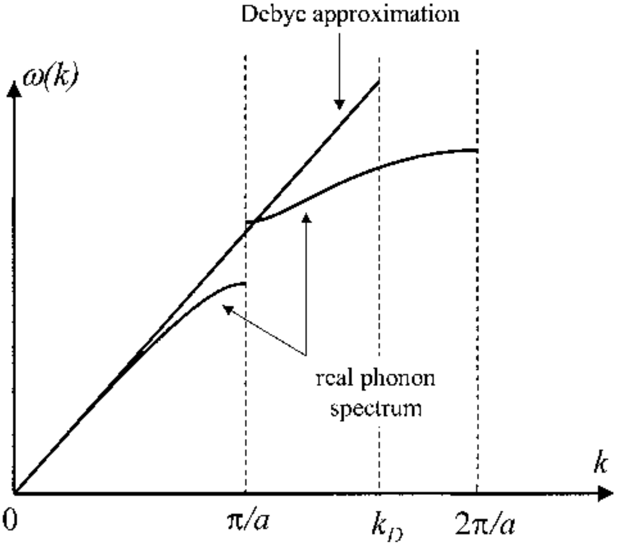


Fig. 6.1. Illustration of the Debye model in the phonon dispersion curve. In the Debye model, all the phonon branches are replaced with three acoustic branches. This corresponds to a simplification of the phonon dispersion spectrum, through a linearization of the phonon branches. In order for this model to be accurate, a Debye wavenumber needs to be introduced.

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 section 4.3 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 section 4.3 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 $0 \le k \le k_D$. We thus obtain:

Eq. (6.2) $$k_D^3 = \frac{6\pi^2 N}{V}$$

This wavenumber corresponds to a Debye frequency ω_D defined by:

Eq. (6.3) $$\hbar\omega_D = \hbar v_0 k_D$$

where v_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:

Eq. (6.4) $$k_b\Theta_D = \hbar\omega_D = \hbar v_0 k_D$$

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.

Material	Θ_D (K)	Material	Θ_D (K)
Pb	105	W	383
Au	162	Al	433
Ag	227	Fe	477
NaCl	275	Si	650
GaAs	345	BN	1900
Cu	347	C (diamond)	2250
Ge	373		

Table 6.1. Debye temperatures of a few solids. [Grigoriev and Meilikhov 1997.]

Example

Q: Calculate the Debye wavelength for GaAs, given that the density of GaAs is d=5.32×10^3 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, we have: $d = \frac{1}{V}\frac{N}{2}\left(M_{Ga} + M_{As}\right)$, 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 the atoms in the volume are Ga atoms, and the other half are As atoms.

Therefore, we can write:

$$\lambda_D = 2\pi\left(6\pi^2 \frac{2d}{\left(M_{Ga} + M_{As}\right)}\right)^{-1/3}$$

$$= 2\pi\left(6\pi^2 \frac{2\times 5.32\times 10^3}{(69.7 + 74.9)\times 1.67264\times 10^{-27}}\right)^{-1/3}$$

or $\lambda_D = 4.57$ Å.

6.2.2. 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 section 4.3.

Eq. (6.5) $$g(\omega) = \sum_{\vec{k},n} \delta\left\lfloor \omega_n(\vec{k}) - \omega \right\rfloor$$

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 Chapter 4 and the discrete summation can be replaced by an integral ((Eq. (4.44)).

Eq. (6.6)

$$\sum_{\vec{k}} \mathrm{Y}\left(\vec{k}\right) \equiv \frac{V}{(2\pi)^3} \iiint_{\vec{k}} \mathrm{Y}\left(\vec{k}\right) d\vec{k} = \frac{V}{(2\pi)^3} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{Y}\left(k_x, k_y, k_z\right) dk_x dk_y dk_z$$

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. Eq. (6.5) then becomes:

Eq. (6.7) $$g(\omega) = \frac{V}{(2\pi)^3} \sum_n \iiint_{\vec{k}} \delta\left[\omega_n(\vec{k}) - \omega\right] d\vec{k}$$

We now make use of Eq. (4.37):

$$d\vec{k} = d\left(\frac{4\pi}{3}k^3\right) = 4\pi k^2 dk$$

where k is the norm or length of the wavevector $\vec{k}$. Therefore, Eq. (6.7) becomes:

Eq. (6.8) $$g(\omega) = \frac{4\pi V}{(2\pi)^3} \sum_n \int_0^{k_D} \delta\left[\omega_n(\vec{k}) - \omega\right] k^2 dk$$

where the integration is now from 0 to the Debye wavenumber k_D, in agreement with the Debye model described earlier. Substituting Eq. (6.1), we get successively:

Eq. (6.9) $$g(\omega) = \frac{V}{2\pi^2} \sum_n \int_0^{k_D} \delta\left[v_n k - \omega\right] k^2 dk$$

or:

Eq. (6.10) $$g(\omega)=\frac{V}{2\pi^2}\sum_n\int_0^{k_D}\delta[x-\omega]\frac{x^2}{v_n^3}dx$$

after the change of variable $x=v_nk$ (and thus $dx=v_ndk$). There is a non zero solution only if there is a wavenumber *k* between *0* and k_D such that $x=v_nk=\omega$, and:

Eq. (6.11) $$\begin{cases} g(\omega)=\dfrac{V}{2\pi^2}\displaystyle\sum_n\frac{\omega^2}{v_n^3} & for\, 0\le\omega\le\omega_D \\ g(\omega)=0 & for\,\omega_D\le\omega \end{cases}$$

Remembering that the Debye model takes into account one longitudinal (l) and two transversal (t) modes, we obtain:

Eq. (6.12) $$\begin{cases} g(\omega)=\dfrac{V}{2\pi^2}\left(\dfrac{\omega^2}{v_l^3}+2\dfrac{\omega^2}{v_t^3}\right) & for\, 0\le\omega\le\omega_D \\ g(\omega)=0 & for\,\omega_D\le\omega \end{cases}$$

which can also be rewritten as:

Eq. (6.13) $$\begin{cases} g(\omega)=\dfrac{3V\omega^2}{2\pi^2v_0^3} & for\, 0\le\omega\le\omega_D \\ g(\omega)=0 & for\,\omega_D\le\omega \end{cases}$$

where:

Eq. (6.14) $$\frac{1}{v_0^3}=\frac{1}{3}(\frac{1}{v_l^3}+\frac{2}{v_t^3})$$

is the inverse average sound velocity. This phonon density of states is illustrated in Fig. 6.2 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(ω)* 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.

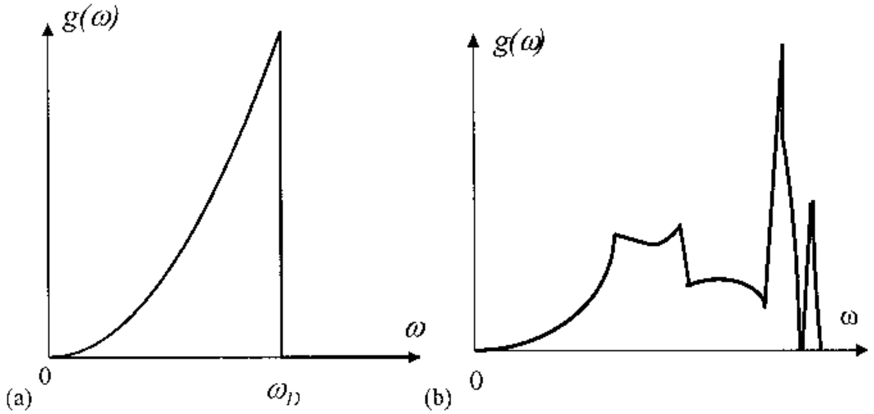


Fig. 6.2. (a) Illustration of the phonon density of states in the Debye model, where the relationship is parabolic until the Debye frequency is reached, after which the density of states is equal to zero. (b) Illustration of a typical phonon spectrum of a real crystal with discontinuities due to singularities in the spectrum. The singularities are due to zeroes in the group velocity.

6.3. Heat capacity

6.3.1. 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 Joule. 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:

Eq. (6.15) $$C_v = \left(\frac{dE}{dT}\right)_v$$

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.7 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 vibration 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. (5.33)), the phonon density of states $g(\omega)$ and that a phonon with frequency ω has an energy $\hbar\omega$ (Eq. (5.30)):

Eq. (6.16) $$E = \int_0^\infty N(\omega)g(\omega)\hbar\omega d\omega$$

In the Debye model, we can use Eq. (6.13) for $g(\omega)$ and rewrite Eq. (6.15) as:

$$E = \int_0^{\omega_D} \frac{1}{\exp\left(\frac{\hbar\omega}{k_bT}\right)-1}\frac{3V\omega^2}{2\pi^2v_0^3}\hbar\omega d\omega$$

or:

Eq. (6.17) $$E = \frac{3V\hbar}{2\pi^2v_0^3}\int_0^{\omega_D}\frac{\omega^3}{\exp\left(\frac{\hbar\omega}{k_bT}\right)-1}d\omega$$

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 $x = \frac{\hbar\omega}{k_bT}$ (and thus $dx = \frac{\hbar}{k_bT}d\omega$), this equation becomes:

Eq. (6.18) $$E = \frac{3V\hbar}{2\pi^2 v_0^3}\left(\frac{k_b T}{\hbar}\right)^4 \int_0^{\frac{\hbar\omega_D}{k_b T}} \frac{x^3}{e^x - 1} dx$$

Let us now make use of the Debye temperature Θ_D defined in Eq. (6.4) and the Debye wavenumber k_D in Eq. (6.2) to express:

$$\frac{1}{(\Theta_D)^3} = \frac{1}{k_D^3}\left(\frac{k_b}{\hbar v_0}\right)^3 = \frac{V}{6\pi^2 N}\left(\frac{k_b}{\hbar v_0}\right)^3$$

Using Eq. (6.4) for the boundary of the integral, Eq. (6.18) can then be rewritten as:

Eq. (6.19) $$E = 9Nk_b \frac{1}{(\Theta_D)^3} T^4 \int_0^{\frac{\Theta_D}{T}} \frac{x^3}{e^x - 1} dx$$

For *high temperatures*, where $k_b T >> \hbar\omega_D$ or simply $T >> \Theta_D$, the integral in Eq. (6.19) 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. (6.19) 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:

Eq. (6.20) $E \approx 3Nk_bT$

The heat capacity is thus obtained after differentiating this expression with respect to the temperature as in Eq. (6.15):

Eq. (6.21) $C_v = \left(\frac{dE}{dT}\right)_v = 3Nk_b$

This relation shows that, for high temperatures, i.e. $T >> \Theta_D$, 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 $k_b T$. Therefore, for a solid with N atoms, each having 3 vibrational degrees of freedom, we get *3N* modes, the total thermal energy is then $3Nk_b T$, as derived in Eq. (6.20), and the heat capacity is found to be equal to Eq. (6.21). 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 one mole of atoms, is calculated for N equal to the Avogadro number $\mathcal{N}_A$=6.02204×10^{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 $T >> \Theta_D$, the Debye model fits the classical model.

For *low temperatures* however, where $k_bT << \hbar\omega_D$ or simply $T << \Theta_D$, 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. (6.19) can be extended up to infinity without much error. Moreover, the exponential fraction in the integral can be expressed as:

Eq. (6.22)
$$\begin{aligned}\frac{1}{e^x - 1} &= \left(\frac{1}{e^x}\right)\left(\frac{1}{1-e^{-x}}\right) \\ &= \frac{1}{e^x}\sum_{n=0}^{\infty}\left(e^{-x}\right)^n = \sum_{n=1}^{\infty}\left(e^{-x}\right)^n = \sum_{n=1}^{\infty} e^{-nx}\end{aligned}$$

because $x>0$ and $e^{-x} < 1$. Therefore, the integral in Eq. (6.19) becomes:

Eq. (6.23)
$$\begin{aligned}\int_0^{\frac{\Theta_D}{T}} \frac{x^3}{e^x - 1}dx &\approx \int_0^{\infty}\frac{x^3}{e^x - 1}dx \\ &= \int_0^{\infty}\left(\sum_{n=1}^{\infty} x^3 e^{-nx}\right)dx \\ &= \sum_{n=1}^{\infty}\left(\int_0^{\infty} x^3 e^{-nx}dx\right) \\ &= \sum_{n=1}^{\infty} I_n\end{aligned}$$

where the integral I_n can be simplified after the following successive integration by parts:

$$
\begin{aligned}
I_n &= \int_0^\infty x^3 e^{-nx} dx \\
&= \left[-x^3 \frac{e^{-nx}}{n} \right]_0^\infty + \int_0^\infty 3x^2 \frac{e^{-nx}}{n} dx \qquad = 0 + \frac{3}{n} \int_0^\infty x^2 e^{-nx} dx \\
&= \frac{3}{n} \left[-x^2 \frac{e^{-nx}}{n} \right]_0^\infty + \frac{3}{n} \int_0^\infty 2x \frac{e^{-nx}}{n} dx \quad = 0 + \frac{6}{n^2} \int_0^\infty x e^{-nx} dx \\
&= \frac{6}{n^2} \left[-x \frac{e^{-nx}}{n} \right]_0^\infty + \frac{6}{n^2} \int_0^\infty \frac{e^{-nx}}{n} dx \qquad = 0 + \frac{6}{n^3} \int_0^\infty e^{-nx} dx \\
&= \frac{6}{n^3} \left[-\frac{e^{-nx}}{n} \right]_0^\infty \\
&= \frac{6}{n^4}
\end{aligned}
$$

Thus, Eq. (6.23) can be rewritten as:

Eq. (6.24) $$\int_0^{\frac{\Theta_D}{T}} \frac{x^3}{e^x - 1} dx \approx 6 \sum_{n=1}^{\infty} \frac{1}{n^4}$$

The sum in this expression corresponds to $\zeta(4)$, which is called the Riemann zeta function evaluated at 4, and is equal to:

Eq. (6.25) $$\zeta(4) = \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}$$

And Eq. (6.19) becomes:

$$E = 9Nk_b \frac{1}{(\Theta_D)^3} T^4 \frac{6\pi^4}{90}$$

or:

Eq. (6.26) $E = \frac{3\pi^4}{5} Nk_b \frac{T^4}{(\Theta_D)^3}$

To determine the heat capacity, we must differentiate this expression with respect to the temperature as in Eq. (6.15):

$$C_v = \left(\frac{dE}{dT}\right)_v = \frac{d}{dT}\left(\frac{3\pi^4}{5} Nk_b \frac{T^4}{(\Theta_D)^3}\right)$$

or:

Eq. (6.27) $C_v = \frac{12\pi^4}{5} Nk_b \left(\frac{T}{\Theta_D}\right)^3$

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. 6.3 for a few solids as a function of temperature.

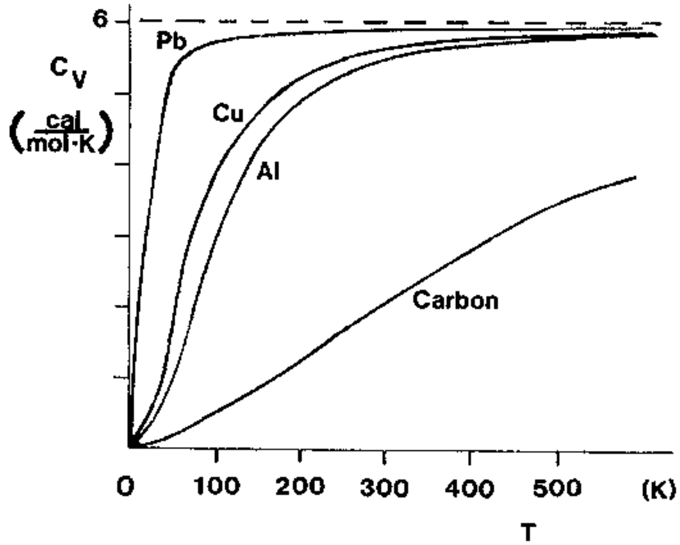


Fig. 6.3. Temperature dependence of the molar heat capacity C_v of some materials. At low temperatures, the heat capacity follows a T^3 relation. [Electronic Properties of Materials, 1993, p. 335, Hummel, R.E., Fig. 19.1. © 1985, 1993 by Springer-Verlag Berlin Heidelberg. With kind permission of Springer Science and Business Media.]

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=4.594×10^3 m.s^{-1}, v_t=3.085×10^3 m.s^{-1}, and the mass density of InP is d=4.81×10^3 kg.m^{-3}.

A: We make use of the expression giving the Debye temperature: $\Theta_D = \dfrac{\hbar\omega_D}{k_b}$, where the Debye frequency $\omega_D = v_0 k_D$ is calculated knowing $\dfrac{1}{v_0^3} = \dfrac{1}{3}\left(\dfrac{1}{v_l^3} + \dfrac{2}{v_t^3}\right)$ and $k_D^3 = \dfrac{6\pi^2 N}{V} = 6\pi^2\left(\dfrac{2d}{M_{In} + M_P}\right)$ similarly to the previous example. Numerically, we successively obtain:

$$v_0 = \left[\frac{1}{3}\left(\frac{1}{v_l^3} + \frac{2}{v_t^3}\right)\right]^{-\frac{1}{3}}$$

$$= \left[\frac{1}{3}\left(\frac{1}{\left(4.594\times10^3\right)^3} + \frac{2}{\left(3.085\times10^3\right)^3}\right)\right]^{-\frac{1}{3}} \text{ or}$$

$v_0 = 3.37\times10^3$ m.s^{-1}.

In addition, we have:

$$k_D = \left(6\pi^2\left(\frac{2\times4.81\times10^3}{(114.8+31)\times1.67264\times10^{-27}}\right)\right)^{1/3} \text{ or}$$

$k_D = 1.33\times10^{10}$ m^{-1}

which leads to:

$\omega_D = 4.47\times10^{13}$ Hz 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. 6.3 to Eq. (6.27).

6.3.2. 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 increase the overall heat capacity of the metal. The contribution of electrons to the total heat capacity, denoted C_v^{el}, can be found as:

Eq. (6.28) $$C_v^{el} = \frac{\pi^2}{2}\frac{N_f k_b{}^2}{E_F}T$$

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:

Eq. (6.29) $$C_v^{el} = \left(\frac{dE}{dT}\right)_{N_f}$$

where E is the energy of all the electrons in the crystal and is given by:

Eq. (6.30) $$E = \int_0^{\infty} \varepsilon g_{3D}(\varepsilon) f_e(\varepsilon) d\varepsilon$$

where $f_e(\varepsilon)$ is the Fermi-Dirac distribution defined in Eq. (4.28) and $g_{3D}(\varepsilon)$ is the three-dimensional electronic density of states of free electrons given by:

Eq. (6.31) $$g_{3D}(\varepsilon)=\frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{\varepsilon}$$

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. (6.28) 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. (6.27)) at low temperatures. It is interesting to consider the ratio of C_v^{el} to C_v^{ph}:

Eq. (6.32) $$\frac{C_v^{el}}{C_v^{ph}}=\frac{\frac{\pi^2}{2}\frac{N_f k_b{}^2}{E_F}T}{\frac{12\pi^4}{5}Nk_b\left(\frac{T}{\Theta_D}\right)^3}=\frac{5}{24\pi^2}\frac{N_f}{N}\frac{k_b}{E_F}\frac{\Theta_D^3}{T^2}$$

where Θ_D is the Debye temperature. By introducing the Fermi temperature T_F such that:

Eq. (6.33) $$E_F=k_bT_F$$

Eq. (6.32) becomes:

Eq. (6.34) $$\frac{C_v^{el}}{C_v^{ph}}=\frac{5}{24\pi^2}\frac{N_f}{N}\frac{\Theta_D^3}{T^2T_F}$$

The ratio $\frac{N_f}{N}$ expresses the average number of free electrons that each atom contributes to the crystal. Eq. (6.34) 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:

Eq. (6.35) $$T_0=\sqrt{\frac{5}{24\pi^2}\frac{N_f}{N}\frac{\Theta_D^3}{T_F}}$$

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 $C_v^{el}\big/C_v^{ph}$ 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: $\dfrac{C_v^{el}}{C_v^{ph}} = \dfrac{5}{24\pi^2}\dfrac{N_f}{N}\dfrac{k_b}{E_F}\dfrac{\Theta_D^3}{T^2}$. Since Cu has two free electrons per atom, we can write $\dfrac{N_f}{N} = 2$. 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: $\dfrac{C_v^{el}}{C_v^{ph}} = 1.16$ (4.2K), 0.023 (30K), 0.034 (77K), 0.00023 (296K).

6.4. 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:

Eq. (6.36) $$\frac{\Delta L}{L} = \alpha_L \Delta T$$

The linear expansion coefficients of a few solids are shown in Table 6.2:

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

Table 6.2. Thermal expansion coefficients of a few solids. [Chemical Rubber Company 1997] [Grigoriev and Meilikhov 1997].

As Eq. (6.36) 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. (6.36). 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. 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. 6.4. Some examples of advanced packaging processes that rely on optimizing the coefficient of thermal expansion can be found in Chapter 16.

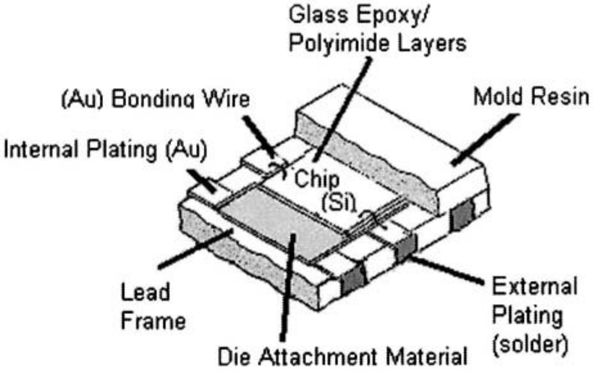
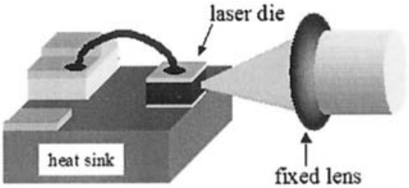


Fig. 6.4. Cut away illustration of an advanced semiconductor device package. To avoid cracking, stresses, and for devices where alignment is critical packaging materials must be chosen with compatible thermal expansion coefficients.

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 is 48 °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 Eq. (6.36) 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 1.67×10^{-5} 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:
$\Delta L = (1.67x10^{-5} K^{-1})(28K)(1 mm) = 4.68\times10^{-4}\,mm$
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 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 section 5.2 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:

Eq. (6.37) $$\langle R\rangle = \frac{\int_{-\infty}^{\infty} R\, e^{-\frac{U(R)}{k_bT}}\, dR}{\int_{-\infty}^{\infty} e^{-\frac{U(R)}{k_bT}}\, dR}$$

By introducing the displacement $x = R - R_0$ and expressing *U(R)* as a function of *x* as was done in section 5.2 (Eq. (5.2)), we can rewrite this equation as:

$$\langle R\rangle = \frac{\int_{-\infty}^{\infty} \left(R_0 + \left(R - R_0\right)\right) e^{-\frac{U(R)}{k_bT}}\, dR}{\int_{-\infty}^{\infty} e^{-\frac{U(R)}{k_bT}}\, dR} = R_0 \frac{\int_{-\infty}^{\infty} e^{-\frac{U(R)}{k_bT}}\, dR}{\int_{-\infty}^{\infty} e^{-\frac{U(R)}{k_bT}}\, dR} + \frac{\int_{-\infty}^{\infty} x\, e^{-\frac{U(x)}{k_bT}}\, dx}{\int_{-\infty}^{\infty} e^{-\frac{U(x)}{k_bT}}\, dx}$$

or:

Eq. (6.38) $$\langle R\rangle = R_0 + \frac{\int_{-\infty}^{\infty} x\, e^{-\frac{U(x)}{k_bT}}\, dx}{\int_{-\infty}^{\infty} e^{-\frac{U(x)}{k_bT}}\, dx}$$

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. (5.3)). 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: $\int_{-\infty}^{\infty} x\, e^{-\frac{U(x)}{k_bT}}\, dx = 0$ 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. (5.2) need to be included (e.g. x^3), causing anharmonicity effects. In this case, the exponential $e^{-\frac{U(x)}{k_B T}}$ 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 anharmonicity effects in the atomic interaction potential.

6.5. 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 as well as 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 κ (read "kappa") which is defined as:

Eq. (6.39) $$J_T = -\kappa \frac{dT}{dx}$$

where J_T is the thermal current density, i.e. the thermal energy transported across an unit area per unit time. This is expressed in units of J.cm^{-2}.s^{-1} or W.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 W.cm^{-1}.K^{-1} (or W.m^{-1}.K^{-1}). Values of the thermal conductivity of a few materials are given below in Table 6.3 and Fig. A.5 in Appendix A.3.

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

Table 6.3. Thermal conductivities of a few solids. [Chemical Rubber Company 1997] [Adachi 2004].

Eq. (6.39) 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 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 (quasi-particle) 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 quasi-particle is modeled as a free moving particle in space with a momentum and an energy, (ii) which 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*, (iv) and 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:

Eq. (6.40) $$\kappa_{ph} = \frac{1}{3}\left(\frac{C_v^{ph}}{V}\right) v_0 \Lambda$$

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: $\overrightarrow{k_1}+\overrightarrow{k_2}+\overrightarrow{k_3}=0$, but not phonon number (phonons are bosons and are not subject to particle number conservation) where $\overrightarrow{k_1}$, $\overrightarrow{k_2}$ and $\overrightarrow{k_3}$ are the momenta of three interacting phonons. The second type is called umklapp processes and is such that: $\overrightarrow{k_1}+\overrightarrow{k_2}+\overrightarrow{k_3}=n\overrightarrow{K}$, where n=1, 2, 3… is an integer, and $\overrightarrow{K}$ is a reciprocal lattice vector. We recall from Chapter 4 and Chapter 5 that electron and lattice momentum in a crystal is only conserved give or take a reciprocal lattice vector. Eq. (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. (5.33) 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. (6.27)). In particular, $\kappa_{ph} \to 0$ when $T \to 0$. These are shown in Fig. 6.5(a) for Λ and Fig. 6.5(b) for κ_{ph}.

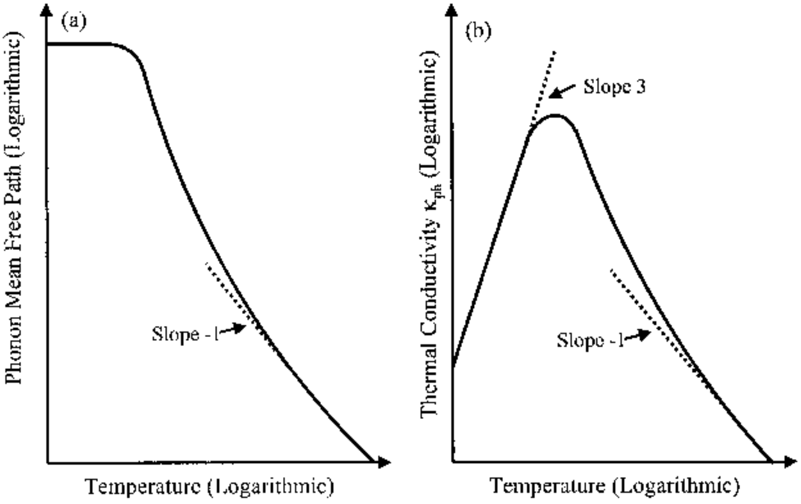


Fig. 6.5. Variation of (a) phonon mean free path and (b) lattice thermal conductivity as a function of temperature. At low temperatures, as the phonon-phonon interaction and scattering decrease, the phonon mean free path is determined by crystal imperfections which are independent of temperature, and 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}.

For higher temperatures, i.e. $T \gg \Theta_D$, we saw in section 5.6 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. 6.5(a). At the same time, we saw that in the heat capacity C_v^{ph} saturates at high temperatures (Eq. (6.21)). The thermal conductivity κ_{ph} therefore has a *1/T* dependence in this regime, as shown in Fig. 6.5(b).

The other contribution to the thermal conductivity arises from electrons and mainly concern 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. (6.40):

Eq. (6.41) $$\kappa_{el} = \frac{1}{3}\left(\frac{C_v^{el}}{V}\right)v_e\Lambda_e$$

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 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 detail in Chapter 8. Nevertheless, 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.

6.6. 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.

References

Adachi, S., *Handbook on Physical Properties of Semiconductors Volume 2-III-V Compound Semiconductors*, Kluwer Academic, Boston, 2004.

Chemical Rubber Company, *CRC Handbook of Chemistry and Physics*, CRC Press, Cleveland, 1997.

Grigoriev, I.S. and Meilikhov, E.Z., *CRC Handbook of Physical Quantities*, CRC Press, Boca Raton, FL, 1997.

Hummel, R.E., *Electronic Properties of Materials*, Springer-Verlag, New York, p. 335, 1993.

Further reading

Ashcroft, N.W. and Mermin, N.D., *Solid State Physics*, Holt, Rinehart and Winston, New York, 1976.

Cochran, W., *The Dynamics of Atoms in Crystals*, Edward Arnold Limited, London, 1973.

Cohen, M.M., *Introduction to the Quantum Theory of Semiconductors*, Gordon and Breach, New York, 1972.

Ferry, D.K., *Semiconductors*, Macmillan, New York, 1991.

Ibach, H. and Lüth, H., *Solid-State Physics: an Introduction to Theory and Experiment*, Springer-Verlag, New York, 1990.

Kasap, S.O., *Principles of Engineering Materials and Devices*, McGraw-Hill, New York, 1997.

Kittel, C., *Introduction to Solid State Physics*, John Wiley & Sons, New York, 1976.

Maxwell, J.C., *Matter and Motion*, Dover, New York, 1952.

Peyghambarian, N., Koch, S.W., and Mysyrowicz, A., *Introduction to Semiconductor Optics*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

Reissland, J.A., *Physics of Phonons*, John Wiley & Sons, London, 1973.

Sapoval, B. and Hermann, C., *Physics of Semiconductors*, Springer-Verlag, New York, 1995.

Problems

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 the lattice constant for this material is 3.56 Å, the density of diamond is 3.52×10^3 kg.m^{-3}, and that 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?

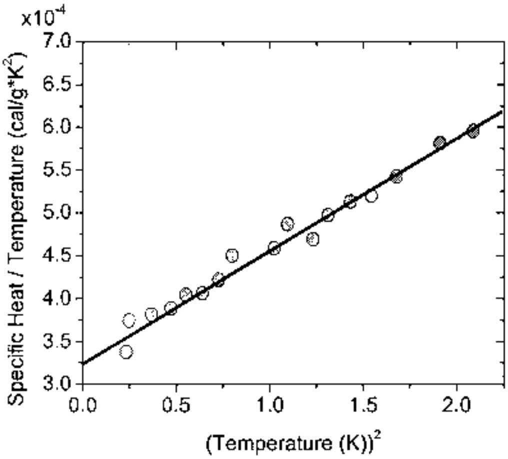
5. Starting from the expression of the total energy carried by the lattice vibrations in Eq. (6.19), show that the heat capacity $C_v = \left(\frac{dE}{dT}\right)_v$ can be written as:

$$C_v = 9Nk_b\left(\frac{T}{\Theta_D}\right)^3 \int_0^{\frac{\Theta_D}{T}} \frac{x^4 e^x}{\left(e^x - 1\right)^2} dx$$

6. It takes 450 cal to raise the temperature of a metallic sample from 20 °C 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 γ=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. IIow can this be explained?

7. Equilibrium Charge Carrier Statistics in Semiconductors

7.1. Introduction

In Chapter 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 Chapter 3 and Chapter 4. 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 Chapter 4, we calculated the density of states of electrons of the conduction band in a three-dimensional semiconductor to be:

Eq. (7.1) $$g_c(E)=\frac{V}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_C)^{1/2}$$

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. However, in the case of many others semiconductors, one has to take into account the degeneracy or number g_d of equivalent conduction-band minima in the first Brillouin zone.

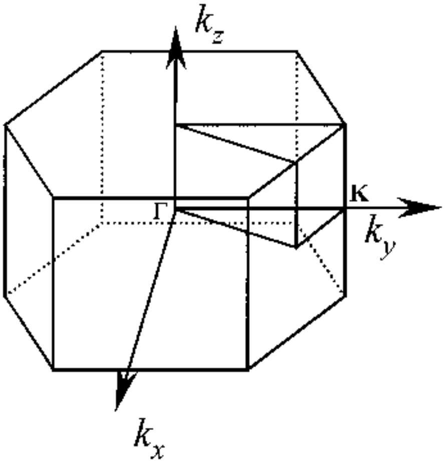
For example, we saw in Fig. 4.17(a), that the conduction band minimum in Ge occurred along the <111> direction. As there are eight equivalent <111> directions, there are eight equivalent conduction band minima in Ge. However, because the minima occur exactly at the boundary of the first Brillouin zone, 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. (7.1) needs to be multiplied by a factor 4. In addition, we also saw in Fig. 4.17(b) 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 six-fold 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. 4.17(c), 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):

Eq. (7.2) $$g_c(E)=\frac{V}{2\pi^2}g_d\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_C)^{1/2}$$

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 six-fold 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. (4.27), 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:

Eq. (7.3) $$E(k) \approx \frac{\hbar^2}{2m_e} k^2$$

In the more general case of a three-dimensional crystal, the effective mass is a 3x3 matrix and each element is 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:

Eq. (7.4) $$E(\vec{k}) \approx \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_{xx}} + \frac{k_y^2}{m_{yy}} + \frac{k_z^2}{m_{zz}}\right)$$

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. (7.2) is the average of these three masses and is given by:

Eq. (7.5) $$m_e = \left(m_{xx} m_{yy} m_{zz}\right)^{1/3}$$

In the particular case when the energy spectrum can be approximated as:

Eq. (7.6) $$E(\vec{k}) \approx \frac{\hbar^2}{2}\left(\frac{\left(k_x^2 + k_y^2\right)}{m_t} + \frac{k_z^2}{m_l}\right)$$

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. (7.2) is the average of these three masses and is given by:

Eq. (7.7) $$m_e = \left(m_t^2 m_l\right)^{1/3}$$

A similar relation can be obtained for the electronic density of states in the valence band (E_V<E):

Eq. (7.8) $$g_v(E) = \frac{V}{2\pi^2}\left(\frac{2m_h}{\hbar^2}\right)^{3/2}\left(E_V - E\right)^{1/2}$$

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 section 4.4 that the valence band of a semiconductor is composed of two main sub-bands, 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. (7.8) is the following average of these two masses:

Eq. (7.9) $$m_h = \left(m_{hh}^{3/2} + m_{lh}^{3/2} \right)^{2/3}$$

7.3. Effective density of states (conduction band)

As discussed in sub-section 4.2.8, 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. (4.28)) which gives the probability of occupation of an energy state:

Eq. (7.10) $$n = \frac{1}{V} \int_{E_C}^{\infty} g_c(E) f_e(E) dE$$

Expanding this expression using Eq. (7.2) and Eq. (4.28), we get:

Eq. (7.11) $$n = \frac{g_d}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \int_{E_C}^{\infty} \frac{\left(E - E_C\right)^{1/2}}{\exp\left(\frac{E - E_F}{k_b T} \right) + 1} dE$$

Making the change of variable $y = \frac{E - E_C}{k_b T}$, and thus $dy = \frac{1}{k_b T} dE$, the previous integral becomes:

Eq. (7.12)
$$\int_{E_C}^{\infty} \frac{\left(E - E_C\right)^{1/2}}{\exp\left(\frac{E - E_F}{k_b T} \right) + 1} dE = \left(k_b T\right)^{3/2} \int_0^{\infty} \frac{y^{1/2}}{\exp\left(y - \frac{E_F - E_C}{k_b T} \right) + 1} dy$$

We can define the Fermi-Dirac integral as in Eq. (4.56):

Eq. (7.13) $$F_{\frac{1}{2}}(x)=\frac{2}{\sqrt{\pi}}\int_0^\infty\frac{y^{1/2}}{1+\exp(y-x)}dy$$

using:

Eq. (7.14) $$x=\frac{E_F-E_C}{k_bT}$$

Eq. (7.12) can be rewritten as:

Eq. (7.15) $$\int_{E_C}^{\infty}\frac{(E-E_C)^{1/2}}{\exp\left(\frac{E-E_F}{k_bT}\right)+1}dE=(k_bT)^{3/2}\frac{\sqrt{\pi}}{2}F_{\frac{1}{2}}\left(\frac{E_F-E_C}{k_bT}\right)$$

and therefore Eq. (7.11) becomes:

Eq. (7.16) $$n=\frac{g_d}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(k_bT)^{3/2}\frac{\sqrt{\pi}}{2}F_{\frac{1}{2}}\left(\frac{E_F-E_C}{k_bT}\right)$$

Remembering that $\hbar=\frac{h}{2\pi}$, this can be simplified as:

Eq. (7.17) $$n=2g_d\left(\frac{2\pi k_bTm_e}{h^2}\right)^{3/2}F_{\frac{1}{2}}\left(\frac{E_F-E_C}{k_bT}\right)$$

or:

Eq. (7.18) $$n=N_cF_{\frac{1}{2}}\left(\frac{E_F-E_C}{k_bT}\right)$$

with:

Eq. (7.19) $$N_c = 2g_d\left(\frac{2\pi k_b T m_e}{h^2}\right)^{3/2}$$

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_bT=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. 7.1(a). For most of the practical calculations, a distance of $3k_bT$ or more, i.e. $E_C - E_F \geq 3k_bT$, is sufficient.

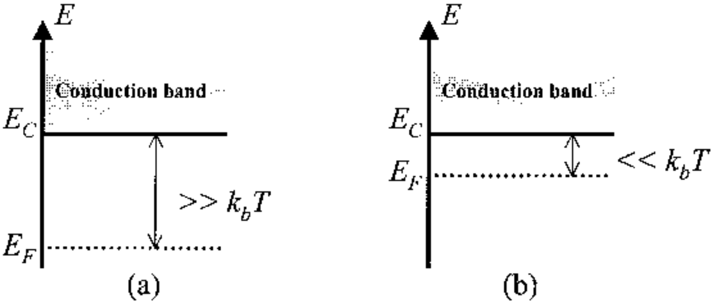


Fig. 7.1. Illustration of the position of the Fermi level with respect to the conduction band (a) in a non-degenerate n-type semiconductor: the Fermi energy is far from the edge of the conduction band. (b) In a degenerate semiconductor n-type semiconductor, the Fermi energy is close to the edge of the conduction band.

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:

Eq. (7.20) $$F_{\frac{1}{2}}(x) \approx \frac{2}{\sqrt{\pi}}\int_0^\infty \frac{y^{1/2}dy}{\exp(y-x)} = \frac{2}{\sqrt{\pi}}e^x\int_0^\infty y^{1/2}e^{-y}dy$$

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. (7.18), we obtain for a non-degenerate semiconductor:

$$F_{\frac{1}{2}}(x) \approx e^x$$

and from Eq. (7.18) and Eq. (7.14):

Eq. (7.21) $$n \approx N_c \exp\left(\frac{E_F - E_C}{k_b T}\right)$$

This expression is much simpler than Eq. (7.18) and is more amenable for calculations. However, when the Fermi energy is close to or even higher than the 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 $E_F - E_C >> k_b T$, 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:

$$F_{\frac{1}{2}}(x) \approx x^{3/2}$$

Fig. 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~0, the Fermi-Dirac integral has to be used.

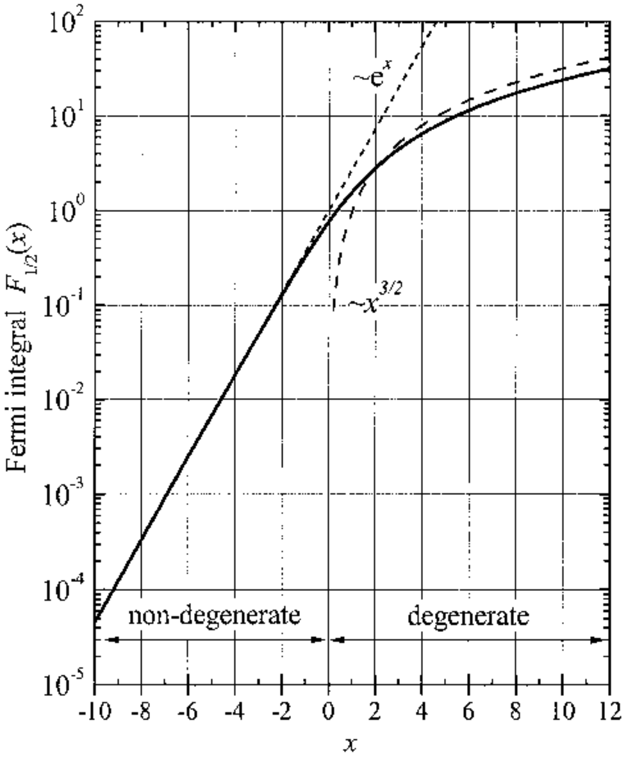


Fig. 7.2. The Fermi integral of order ½ and its approximations.

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 section 7.6), the Fermi energy is about in the middle of the bandgap, $E_C - E_F \approx E_g / 2$, which is about 85 meV at 300K. 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:

Eq. (7.22) $$p = \frac{1}{V}\int_{-\infty}^{E_V} g_v(E) f_h(E) dE$$

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. (4.58)):

Eq. (7.23) $$f_h(E) = [1 - f_e(E)] = \frac{1}{\exp\left(\frac{E_F - E}{k_b T}\right) + 1}$$

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 Eq. (7.8) and Eq. (7.23), we get:

Eq. (7.24) $$p = \frac{1}{2\pi^2}\left(\frac{2m_h}{\hbar^2}\right)^{3/2}\int_{-\infty}^{E_V}\frac{(E_V - E)^{1/2}}{\exp\left(\frac{E_F - E}{k_b T}\right) + 1} dE$$

Using the change of variable $y = \frac{E_V - E}{k_b T}$, thus $dy = -\frac{1}{k_b T} dE$, and:

Eq. (7.25) $$x = \frac{E_V - E_F}{k_b T}$$

in the previous integral and identifying it with the Fermi-Dirac integral, we obtain a relation similar to Eq. (7.17) for p:

Eq. (7.26) $$p = 2\left(\frac{2\pi k_b T m_h}{h^2}\right)^{3/2} F_{\frac{1}{2}}\left(\frac{E_V - E_F}{k_b T}\right)$$

or:

Eq. (7.27) $$p = N_v F_{\frac{1}{2}}\left(\frac{E_V - E_F}{k_b T}\right)$$

where:

Eq. (7.28) $$N_v = 2\left(\frac{2\pi k_b T m_h}{h^2}\right)^{3/2}$$

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 section 7.3, the general expression in Eq. (7.18) 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. 7.3(a). 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. (7.21):

Eq. (7.29) $$p \approx N_v \exp\left(\frac{E_V - E_F}{k_b T}\right)$$

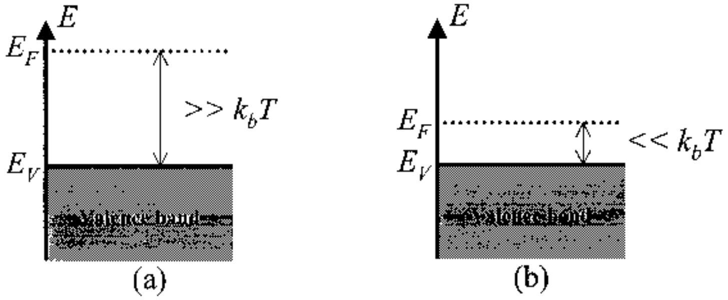


Fig. 7.3. Illustration of the position of the Fermi level with respect to the valence band (a) in a non-degenerate p-type semiconductor: the Fermi energy is far from the edge of the valence band. (b) In a degenerate p-type semiconductor , the Fermi energy is close to the edge of the valence band.

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_bT (25.9 meV at room temperature). This situation is more often encountered in practice that 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 Eq. (7.21) and Eq. (7.29) given by:

$$\begin{aligned} np &= N_c \exp\left(\frac{E_F - E_c}{k_b T}\right) N_v \exp\left(\frac{E_V - E_F}{k_b T}\right) \\ &= N_c N_v \exp\left(\frac{E_V - E_c}{k_b T}\right) \end{aligned}$$

or:

Eq. (7.30) $$np = N_c N_v \exp\left(-\frac{E_g}{k_b T}\right)$$

where $E_g=E_C-E_V$ 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:

Eq. (7.31) $$n_i^2 = np = N_c N_v \exp\left(-\frac{E_g}{k_b T}\right)$$

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:

Eq. (7.32) $$n = p = n_i = \sqrt{N_c N_v} \exp\left(-\frac{E_g}{2k_b T}\right)$$

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:

$$\begin{aligned} n_i &= \sqrt{4g_d\left(\frac{2\pi k_b T m_e}{h^2}\right)^{3/2}\left(\frac{2\pi k_b T m_h}{h^2}\right)^{3/2}} \exp\left(-\frac{E_g}{2k_b T}\right) \\ &= 2\left(\frac{2\pi k_b T m_0}{h^2}\right)^{3/2}\sqrt{g_d\left(\frac{m_e}{m_0}\frac{m_h}{m_0}\right)^{3/2}} \exp\left(-\frac{E_g}{2k_b T}\right) \end{aligned},$$

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: $m_h^{3/2} = m_{hh}^{3/2} + m_{lh}^{3/2}$. We therefore get:

$$
\begin{aligned}
n_i &= 2\left(\frac{2\pi\times\left(1.38066\times10^{-23}\right)\times300\times\left(0.91095\times10^{-30}\right)}{\left(6.62617\times10^{-34}\right)^2}\right)^{3/2} \\
&\quad\times\sqrt{(0.067)^{3/2}\left(0.45^{3/2}+0.082^{3/2}\right)} \\
&\quad\times\exp\left(-\frac{1.424\times1.60218\times10^{-19}}{2\times\left(1.38066\times10^{-23}\right)\times300}\right) \\
&= 2.06\times10^{12}\ m^{-3} \\
&= 2.06\times10^{6}\ cm^{-3}
\end{aligned}
$$

7.6. Doping: intrinsic vs. 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-section 4.2.7). Since the Fermi-Dirac distribution function has an exact step shape at T=0 K (Fig. 4.12), there is no electron with an energy $E>E_F$, including the conduction band, and all the electrons are located at an energy $E<E_F$.

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 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.

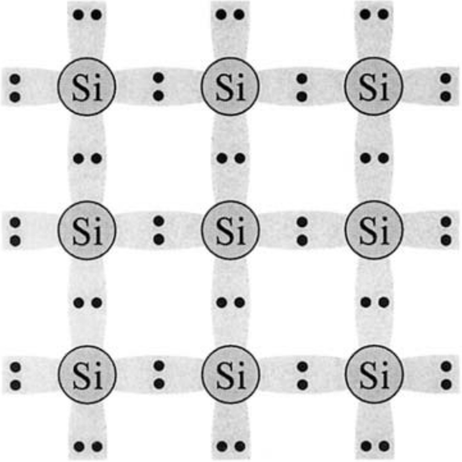


Fig. 7.4. Schematic of a Si semiconductor crystal showing the distribution of electrons in the outer shell of each Si atom. Each Si atom has eight electrons in this shell: four from its own outer shell and one from each of the four nearest Si atoms to which it is covalently bonded to.

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} to 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.

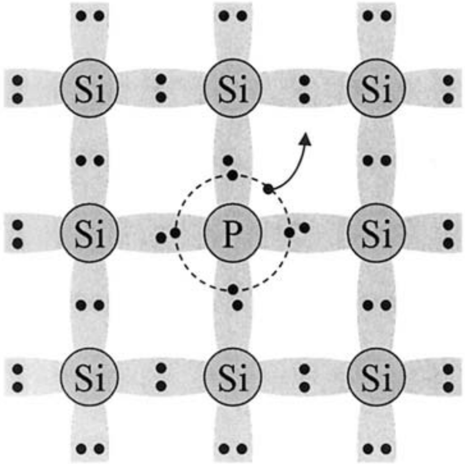


Fig. 7.5. Schematic of a Si semiconductor crystal with one Si atom replaced by a P atom to achieve n-type doping. The dotted circle symbolizes the outer shell of the P atom which contains 5 electrons. Because the fifth electron does not contribute to the bonding, it can be free (ionized) to move inside the crystal. P is thus called a donor.

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 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. 7.6(a). Moreover, because the extra electron around the P atom is easily ionized, i.e. 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.

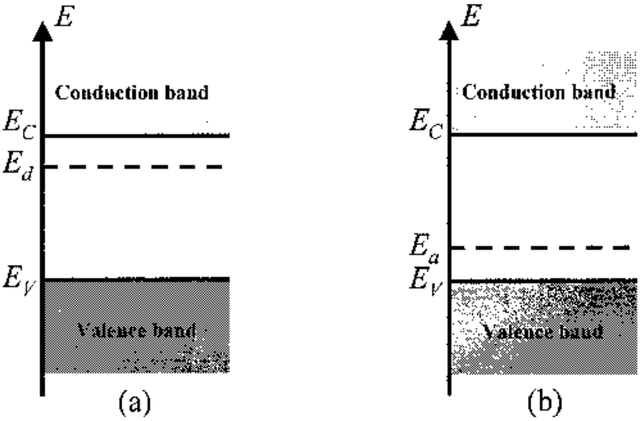


Fig. 7.6. Schematic of the energy levels introduced by: (a) a donor or a (b) an acceptor dopant in a semiconductor crystal. The energy level of a donor is closer to the edge of the conduction band, whereas that of an acceptor is closer to the edge of the valence band.

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.

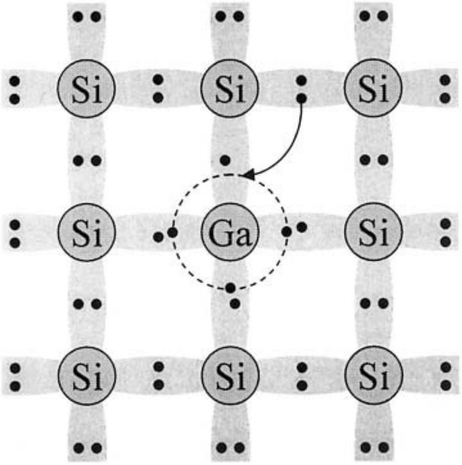


Fig. 7.7. Schematic of a Si semiconductor crystal with one Si atom replaced by a Ga atom to achieve p-type doping. The dotted circle symbolizes the outer shell of the Ga atom which contains 3 electrons. The Ga atom can accept one more electron from a neighboring bond. Ga is thus called an acceptor.

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 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-section 4.3.3, 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 Ea, which is called acceptor energy level, as shown Fig. 7.6(b). 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.

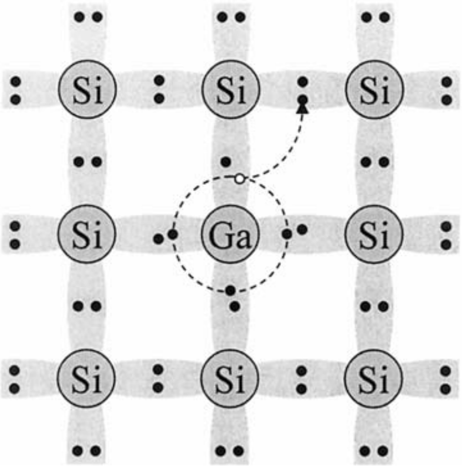


Fig. 7.8. Schematic showing the movement of a hole in a Si semiconductor crystal doped p-type using a Ga atom. The hole is represented by an open circle. When the Ga atom accepts an electron, the process can be equivalently viewed as the Ga atom releasing a hole inside the crystal.

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 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. Additional data on other impurities and their ionization energies will be given in Table 14.2.

(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

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.]

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:

Eq. (7.33) $n + N_A^- = p + N_D^+$

For a given semiconductor crystal, the concentrations n and p solely depend on the Fermi energy E_F through Eq. (7.21) and Eq. (7.29) in the non-degenerate case, or Eq. (7.18) and Eq. (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:

Eq. (7.34) $$\frac{N_D^+}{N_D} = \frac{1}{2\exp\left(\frac{E_F - E_d}{k_b T}\right) + 1}$$

Eq. (7.35) $$\frac{N_A^-}{N_A} = \frac{1}{4\exp\left(\frac{E_a - E_F}{k_b T}\right) + 1}$$

where E_F is the Fermi energy, E_d and E_a are the donor and acceptor energy levels in the bandgap respectively, 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 sub-bands in the valence band at the center of the Brillouin zone: the heavy-hole band and the light-hole band (sub-section 4.4.3). Similar to the Fermi-Dirac distribution, the Eq. (7.34) and Eq. (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 Eq. (7.21) and Eq. (7.29) in the non-degenerate case, or Eq. (7.18) and Eq. (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 Eq. (7.21) and Eq. (7.32), we can write an expression for the Fermi energy:

$$n = \sqrt{N_c N_v}\exp\left(-\frac{E_g}{2k_b T}\right) = N_c \exp\left(\frac{E_F - E_C}{k_b T}\right)$$

which becomes, knowing that $E_g=E_C-E_V$:

Eq. (7.36) $$\exp\left(\frac{E_F}{k_bT}\right)=\sqrt{\frac{N_v}{N_c}}\exp\left(\frac{E_C+E_V}{2k_bT}\right)$$

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:

Eq. (7.37) $$E_F=\frac{E_C+E_V}{2}+\frac{1}{2}k_bT\ln\left(\frac{N_v}{N_c}\right)$$

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:

$$E_F - \frac{E_C + E_V}{2} = \frac{1}{2} k_b T \ln\left(\frac{m_{hh}^{3/2} + m_{lh}^{3/2}}{m_e^{3/2}} \right)$$

$$= \frac{1}{2} \times 1.38066 \times 10^{-23} \times 296 \times \ln\left(\frac{0.45^{3/2} + 0.082^{3/2}}{0.067^{3/2}} \right)$$

$$= 6.00 \times 10^{-21}\, J$$

$$= 37.4\, meV$$

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. 7.9(a), 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. 4.12, 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. 7.9(b), which means that the Fermi energy is much closer to the top of the valence band than in the case of an intrinsic 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. 4.12, at room temperature. And these holes can easily participate in electrical conduction phenomena.

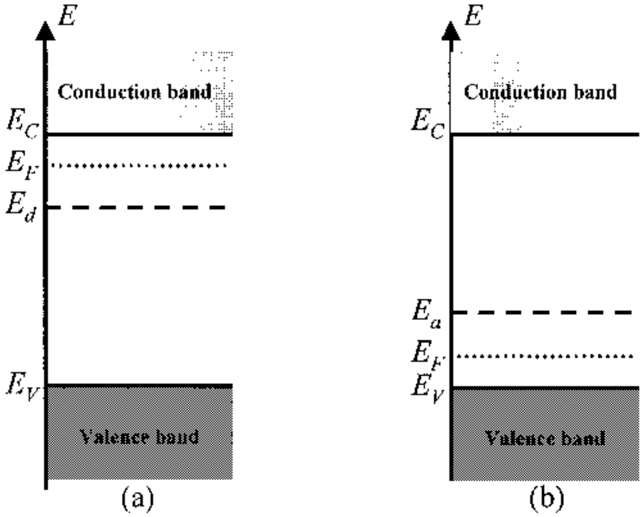


Fig. 7.9. Position of the Fermi energy at T=0 K in (a) an n-type semiconductor: it is located between the donor energy level and the bottom of the conduction band; and (b) in a p-type semiconductor: it is located between the acceptor energy level and the top of the valence band.

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→ ∞):

Eq. (7.38) $$n = p = n_i = \sqrt{N_c N_v}\exp\left(-\frac{E_g}{2k_b T}\right) \approx \sqrt{N_c N_v}$$

Moreover, from Eq. (7.19) and Eq. (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:

Eq. (7.39) $$n \approx p$$

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.

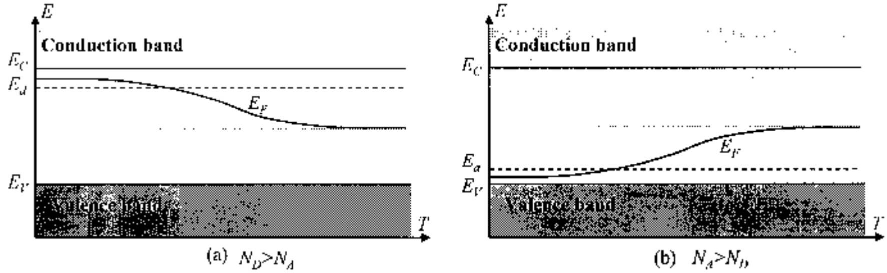


Fig. 7.10. Evolution of the Fermi energy as a function of temperature in a: (a) n-type or (b) p-type semiconductor crystal. As the temperature is raised, the position of the Fermi energy shifts from its position in Fig. 7.9 to the position for an intrinsic semiconductor.

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:

Eq. (7.40) $n = p + N_D^+$

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 sub-section, all the donors are ionized ($N_D^+ = N_D$). 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.:

Eq. (7.41) $n \approx p \approx n_i = \sqrt{N_c N_v}\exp\left(-\frac{E_g}{2k_b T}\right)$

As the temperature is lowered, while the donors remain ionized ($N_D^+ = N_D$), 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 ($p << N_D^+$) because the only contribution to p is the intrinsic contribution. Therefore, Eq. (7.40) becomes:

Eq. (7.42) $n \approx N_D$

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 $p << N_D^+$. In this case, Eq. (7.40) becomes:

Eq. (7.43) $n \approx N_D^+$

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:

Eq. (7.44) $$N_D^+ \approx \frac{N_D}{2}\exp\left(-\frac{E_F - E_d}{k_b T}\right)$$

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:

Eq. (7.45) $$N_c \exp\left(\frac{E_F - E_C}{k_b T}\right)\frac{N_D}{2}\exp\left(-\frac{E_F - E_d}{k_b T}\right)$$

after using Eq. (7.21) and Eq. (7.44). We then obtain:

Eq. (7.46) $$n^2 \approx \frac{N_c N_D}{2}\exp\left(-\frac{E_C - E_d}{k_b T}\right)$$

which yields:

Eq. (7.47) $$n \approx \sqrt{\frac{N_c N_D}{2}}\exp\left(-\frac{E_C - E_d}{2k_b T}\right)$$

The three expressions of n in Eq. (7.41), Eq. (7.42) and Eq. (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. (7.41) 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.

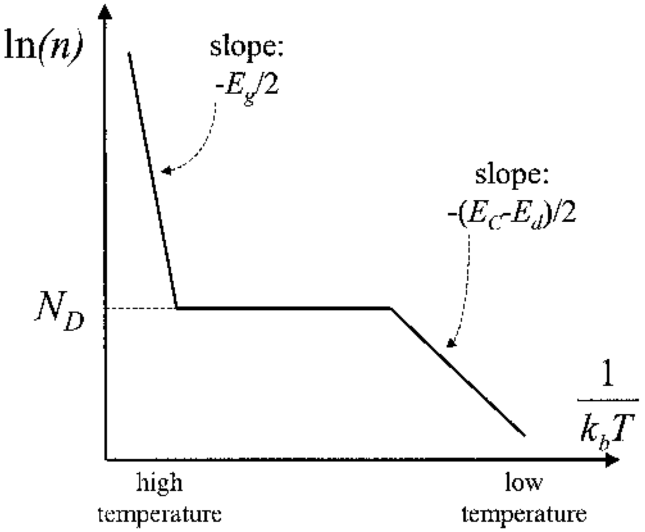


Fig. 7.11. Simple schematic diagram of the dependence of the electron concentration in the conduction band as a function of temperature in a typical n-type semiconductor crystal. At low temperatures, the carrier concentration follows a relation dependent on the donor energy inside the bandgap. At moderate temperatures, the electron concentration is nearly constant equal to the donor concentration. At high temperatures, the carrier concentration approaches that of an intrinsic semiconductor.

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):

Eq. (7.48) $$n \approx p \approx n_i = \sqrt{N_c N_v}\exp\left(-\frac{E_g}{2k_bT}\right)$$

In the second regime, Eq. (7.42) can be transformed for a p-type semiconductor into:

Eq. (7.49) $$p \approx N_A$$

In the third regime, as the temperature is further lowered, Eq. (7.43) can also be transformed for a p-type semiconductor into:

Eq. (7.50) $p \approx N_A^-$

From Eq. (7.35), using the same derivation as between Eq. (7.44) and Eq. (7.47), we get:

Eq. (7.51) $$p \approx \sqrt{\frac{N_v N_D}{4}} \exp\left(-\frac{E_a - E_V}{2k_b T}\right)$$

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.

References

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

Wolfe, C.M., Holonyak, Jr., N., and Stillman, G.E., *Physical Properties of Semiconductors*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

Further reading

Anselm, A., *Introduction to Semiconductor Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

Ashcroft, N.W. and Mermin, N.D., *Solid State Physics*, Holt, Rinehart and Winston, New York, 1976.

Cohen, M.M., *Introduction to the Quantum Theory of Semiconductors*, Gordon and Breach, New York, 1972.

Ferry, D.K., *Semiconductors*, Macmillan, New York, 1991.

Hummel, R.E., *Electronic Properties of Materials*, Springer-Verlag, New York, 1986.

Pierret, R.F., *Advanced Semiconductor Fundamentals*, Addison-Wesley, Reading, MA, 1989.

Sapoval, B. and Hermann, C., *Physics of Semiconductors*, Springer-Verlag, New York, 1995.

Streetman, B.G., *Solid State Electronic Devices*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

Wang, S., *Fundamentals of Semiconductor Theory and Device Physics*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

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: $m_h = \left(m_{hh}^{3/2} + m_{lh}^{3/2} \right)^{2/3}$.

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 carriers concentration when the Maxwell-Boltzmann statistics is applied for semiconductors with Fermi energy within the energy gap, if the Fermi level is $3k_bT$, $2k_bT$, k_bT 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 section 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 $\Delta E_a = E_a - E_V = 125\,meV$. 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 $\Delta E_d = E_C - E_d = 25\,meV$. 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=$1.38\cdot10^{10}$ 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=$1.38\cdot10^{10}$ 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, and
(b) 175 °C.

8. Non-Equilibrium Electrical Properties of Semiconductors

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 section 6.5.

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.9).

Let us start by conceptually considering an electron with an electrical charge -q in an 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:

Eq. (8.1) $$m\frac{d\vec{v}}{dt}=\vec{F}=-q\vec{E}$$

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. 8.1(a). The current density is expressed in units of A.cm^{-2}.

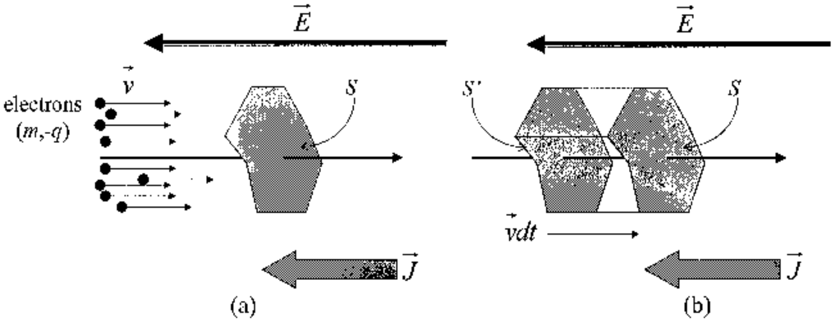


Fig. 8.1. Schematic diagrams showing: (a) the flow of electrons and current density vector in a uniform electric field, (b) the displacement of the surface area A after a time dt at a velocity equal to that of the flowing electrons.

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. 8.1(b). This volume is equal to $A\left|\vec{v}\right|dt$, where A is the area of the surface S.

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:

Eq. (8.2) $$nqA\left|\vec{v}\right|dt$$

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:

Eq. (8.3) $$\vec{J} = -nq\vec{v}$$

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.9).

In order to calculate the average velocity of the electron gas that results from the electric field, we have to introduce, as was done in earlier in Chapter 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 $\overrightarrow{v^{drift}}$. This quantity can be estimated by integrating Eq. (8.1) over time from $t = 0$ and $t = \tau$:

Eq. (8.4) $$m\overrightarrow{v^{drift}} = -q\tau\vec{E} \text{ or } \overrightarrow{v^{drift}} = -\frac{q\tau}{m}\vec{E}$$

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:

Eq. (8.5) $$\begin{cases} \overrightarrow{v^{drift}} = -\mu\vec{E} \\ \mu = \dfrac{q\tau}{m} \end{cases}$$

This quantity is expressed in units of $cm^2.V^{-1}.s^{-1}$, and it represents the velocity that an electron gains per unit electric field strength (velocity ($cm.s^{-1}$) divided by electric field strength ($V.cm^{-1}$)). 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):

Eq. (8.6) $$\overrightarrow{J^{drift}} = -nq\overrightarrow{v^{drift}} = nq\mu\vec{E} = \frac{nq^2\tau}{m}\vec{E}$$

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 S.cm^{-1} (Siemens per cm) or inverse (Ω.cm):

Eq. (8.7) $$\begin{cases} \overrightarrow{J^{drift}} = \sigma\vec{E} \\ \sigma = nq\mu = \dfrac{nq^2\tau}{m} \end{cases}$$

It is also common practice to consider the inverse of the conductivity which is called the resistivity of the material:

Eq. (8.8) $$\rho = \frac{1}{\sigma} = \frac{1}{nq\mu}$$

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: $\mu = \dfrac{1}{nq\rho}$, where n is the electron concentration participating in the conduction. Since there are two electrons in valence shell of copper, this concentration can be determined by the concentration of Cu atoms or the density of Cu (d=8.92 g.cm^{-3}): $n = 2 \times \dfrac{d}{m_{Cu}}$, where m_{Cu} is the mass of a Cu atom. Assuming the resistivity of Cu is about ρ=1.7×10^{-6} Ω.cm, we get the mobility:

$$\begin{aligned}\mu &= \frac{m_{Cu}}{2dq\rho} \\ &= \frac{63.55\times 1.67264\times 10^{-27}}{2\times\left(8.92\times 10^{-3}\right)\times\left(1.60218\times 10^{-19}\right)\times\left(1.7\times 10^{-6}\right)} \\ &= 21.9\, cm^2 / Vs\end{aligned}$$

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.

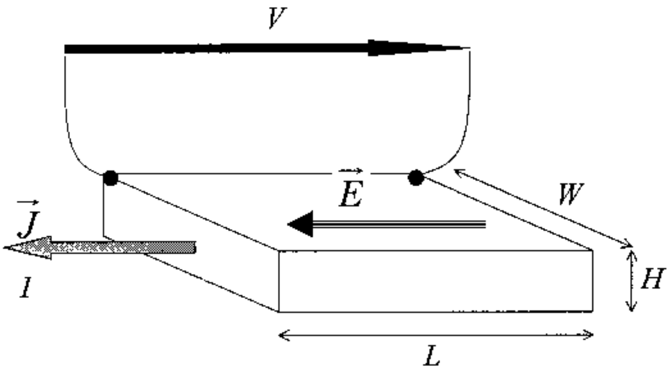


Fig. 8.2. Schematic diagram illustrating the geometry used to illustrate Ohm's law. A voltage is applied across two opposite faces of a rectangular solid and separated by a distance L. This results in an electric field and a current density perpendicular to these two faces.

In this configuration, the electrical current I is equal to the magnitude of the current density multiplied by the area WH, i.e. $I = WHJ_{drift}$. 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:

Eq. (8.9)
$$\begin{aligned} I &= WHJ^{drift} = WH\sigma\left|\vec{E}\right| \\ &= \frac{WII}{L}\sigma L\left|\vec{E}\right| = \frac{WH}{L}\sigma V \end{aligned}$$

We thus recognize the relation:

Eq. (8.10) $$I = \frac{V}{R}$$

where:

Eq. (8.11) $$R = \frac{L}{WII}\frac{1}{\sigma} \quad \text{or} \qquad R = \frac{L}{A}\rho$$

and $A=WII$ 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. 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:

$$\overrightarrow{J^{drift}} = \overrightarrow{J_e^{drift}} + \overrightarrow{J_h^{drift}}$$

where each of the $\overrightarrow{J_e^{drift}}$ and $\overrightarrow{J_h^{drift}}$ 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:

Eq. (8.12) $$\begin{cases} \overrightarrow{v_e^{drift}} = -\mu_e \vec{E} \\ \overrightarrow{v_h^{drift}} = +\mu_h \vec{E} \end{cases} \text{ and } \begin{cases} \overrightarrow{J_e^{drift}} = -nq\overrightarrow{v_e^{drift}} \\ \overrightarrow{J_h^{drift}} = +pq\overrightarrow{v_h^{drift}} \end{cases}$$

The total drift current density can then be written as:

Eq. (8.13) $$\begin{cases} \overrightarrow{J^{drift}} = \sigma \vec{E} \\ \sigma = q(n\mu_e + p\mu_h) \end{cases}$$

The typical room temperature conductivity in metals is $(0.1\sim3)\times10^4$ S.cm^{-1}, while the conductivity in semiconductors depends on the carrier concentrations and therefore the doping level, as discussed in Chapter 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. 4.11) 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. 4.12). 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. 4.11), 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 Chapter 4, 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), 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:

$$\vec{k}' = \vec{k} \pm \vec{q}$$

$$E(\vec{k}') = E(\vec{k}) \pm \hbar\omega(\vec{q})$$

where $E(\vec{k}')$ is the energy of the particle after the scattering process and $\hbar\omega(\vec{q})$ is the energy of the phonon absorbed or emitted.

We have seen in Chapter 5 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-optic 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-optic 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 $\frac{1}{\tau} = \frac{1}{\tau_{el}} + \frac{1}{\tau_{op}} + \frac{1}{\tau_{ac}}$ where the terms denote the inverse of the elastic, optic, an 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 section 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 A.4) 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; 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 ~20cm^2/Vs. This is typical for good metals and interestingly lower than for many semiconductors.

8.4. Hall effect

At the end of the 19th 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.

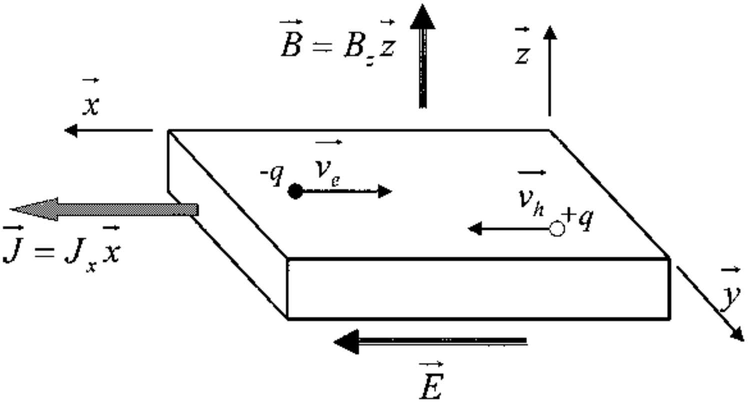


Fig. 8.3. Geometry used for a Hall effect experiment. A uniform electric field strength is applied inside a solid in the x-direction (by applying a voltage across the solid, for example), which results in an electric current in the same direction. The movement of holes and electrons in the solid are shown. The solid is immersed in a magnetic induction which is directed in the z-direction, perpendicularly to this electric field.

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:

Eq. (8.14) $$\overrightarrow{F_{Lorentz}} = q\overrightarrow{v_h} \times \vec{B}$$

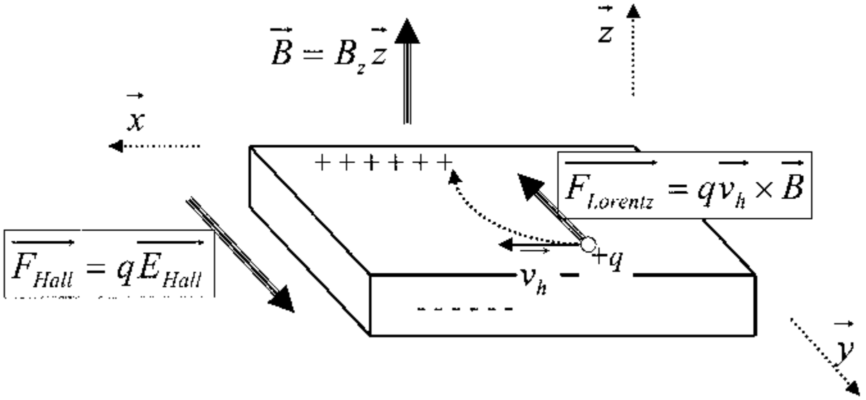
which is in the y-direction. If the sample was without limits, the hole would exhibit a cyclotron (circular) motion around an axis parallel to $\vec{B}$. In the case of a finite size solid as the one shown in Fig. 8.4, holes would accumulate on one of its sides to create a surplus of positive charges. 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 $\overrightarrow{E_{Hall}}$, 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:

Eq. (8.15) $$\vec{0} = \overrightarrow{F_{h,Lorentz}} + \overrightarrow{F_{Hall}} = q\overrightarrow{v_h} \times \vec{B} + q\overrightarrow{E_{Hall}}$$

The Hall electric field strength is thus:

Eq. (8.16) $\overrightarrow{E_{Hall}} = -\overrightarrow{v_h} \times \overrightarrow{B}$



$\overrightarrow{B} = B_z \overrightarrow{z}$
$\overrightarrow{z}$
$\overrightarrow{x}$
$\overrightarrow{y}$
+ + + + + +
- - - - - -
$\overrightarrow{F_{Lorentz}} = q\overrightarrow{v_h} \times \overrightarrow{B}$
$\overrightarrow{F_{Hall}} = q\overrightarrow{E_{Hall}}$
$\overrightarrow{v_h}$
$+q$

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.

The component of the Hall electric field strength in the *y*-direction (i.e. $\overrightarrow{E_{Hall}} = \left(E_{Hall}\right)_y \overrightarrow{y}$), in the geometry shown in Fig. 8.4 is:

Eq. (8.17) $\left(E_{Hall}\right)_y = \left(v_h\right)_x B_z \;\; > 0$

From Eq. (8.12), we get:

$$J_x = pq\left(v_h\right)_x$$

where *p* is the hole concentration in the solid and we can rewrite Eq. (8.17) as:

Eq. (8.18) $\left(E_{Hall}\right)_y = \frac{J_x}{pq} B_z$

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 common practice to introduce the Hall constant given by:

Eq. (8.19) $$R_H = \frac{\left(E_{Hall}\right)_y}{J_x B_z} = \frac{1}{pq} \quad > 0$$

The Hall constant therefore yields a direct measure of the hole concentration in the solid. We can define a hole Hall mobility as:

Eq. (8.20) $$\mu_{H,h} = \sigma R_H$$

This Hall mobility has the same units as the drift mobility encountered in Eq. (8.12) in section 8.2, i.e. cm^2.V^{-1}.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. 8.4.

However, because electrons have a negative charge, the Hall electric field is in the opposite direction in comparison to the one from holes:

Eq. (8.21) $$\vec{0} = \overrightarrow{F_{e,Lorentz}} + \overrightarrow{F_{Hall}} = -q\overrightarrow{v_e} \times \vec{B} - q\overrightarrow{E_{Hall}}$$

The Hall electric field strength is thus:

Eq. (8.22) $$\overrightarrow{E_{Hall}} = -\overrightarrow{v_e} \times \vec{B}$$

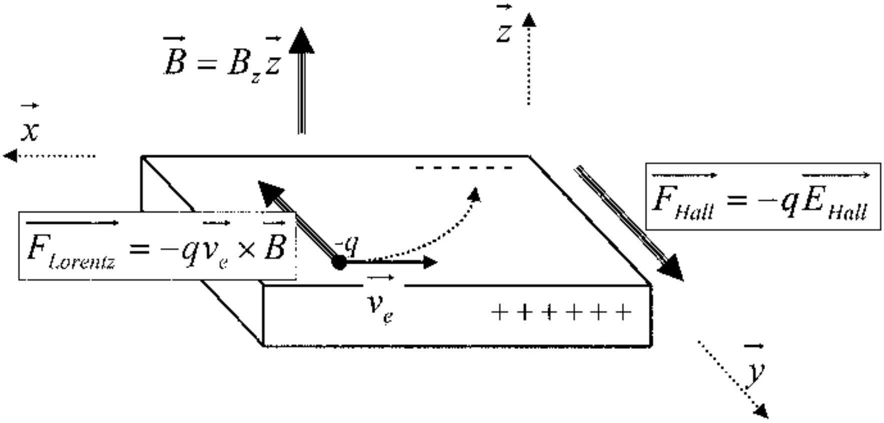

$\vec{B} = B_z\vec{z}$
$\vec{z}$
$\vec{x}$
- - - - - -
$\overrightarrow{F_{Lorentz}} = -q\vec{v_e} \times \vec{B}$
$-q$
$\overrightarrow{F_{Hall}} = -q\overrightarrow{E_{Hall}}$
$\vec{v_e}$
+ + + + + + +
$\vec{y}$

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.

The component of the Hall electric field strength in the *y*-direction (i.e. $\overrightarrow{E_{Hall}} = (E_{Hall})_y \vec{y}$), in the geometry shown in Fig. 8.5 is:

Eq. (8.23) $$(E_{Hall})_y = (v_e)_x B_z \quad < 0$$

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:

Eq. (8.24) $$(E_{Hall})_y = -\frac{J_x}{nq} B_z$$

This expression is similar to Eq. (8.18) and the Hall constant defined in Eq. (8.19) becomes:

Eq. (8.25) $$R_H = \frac{(E_{Hall})_y}{J_x B_z} = -\frac{1}{nq} \quad < 0$$

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:

Eq. (8.26) $$\mu_{H,e} = \sigma\left|R_H\right| \quad > 0$$

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 be focused to compensate the local doping, and thus produce submicron devices (see also Chapter 13).

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 photoexciting 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:

Eq. (8.27) $$\begin{aligned} m_e \frac{dv_x}{dt} + m_e v_x \frac{1}{\tau} &= -qE_x - qv_x B \\ m_e \frac{dv_y}{dt} + m_e v_y \frac{1}{\tau} &= -qE_y + qv_x B \end{aligned}$$

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 $q \to -q$. 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. Then, we can write Eq. (8.27) as:

Eq. (8.28)
$$v_y^e = -\mu_e E_y - \mu_e^2 BE_x$$
$$v_y^h = \mu_h E_y - \mu_h^2 BE_x$$

The above equations can be related to the total current J_y giving:

Eq. (8.29)
$$J_y = nq\mu_e(E_y + \mu_e E_x B) + pq\mu_h(E_y - \mu_h BE_x)$$

Under equilibrium condition, i.e. when the current $J_y = 0$, the ratio of the components of the electric field is such that:

Eq. (8.30)
$$\frac{E_y}{E_x} = \{\frac{p\mu_h^2 - n\mu_e^2}{n\mu_e + p\mu_h}\}B$$

and the Hall constant is now given by:

Eq. (8.31)
$$R_H = \frac{1}{q}\frac{p\mu_h^2 - n\mu_e^2}{(p\mu_h + n\mu_e)^2}$$

where p and n are the hole and electron concentrations, 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:

Eq. (8.32)
$$\mu_H = \sigma|R_H| = \left|\frac{p\mu_h^2 - n\mu_e^2}{p\mu_h + n\mu_e}\right|$$

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.

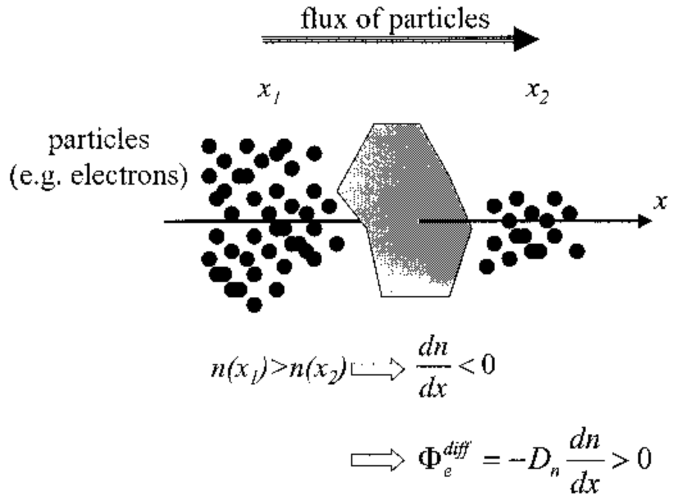

flux of particles
x_1
x_2
particles
(e.g. electrons)
x
$n(x_1)>n(x_2) \Rightarrow \frac{dn}{dx}<0$
$\Rightarrow \Phi_e^{diff} = -D_n \frac{dn}{dx} > 0$

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.

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:

Eq. (8.33) $$\Phi_e^{diff} = -D_n \frac{dn}{dx}$$

where D_n is called the diffusion coefficient or diffusivity and has the units of cm^2.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 section 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_h^{diff} :

Eq. (8.34) $\Phi_e^{diff} = nv_e^{diff}$

which leads to the relation:

Eq. (8.35) $v_e^{diff} = -D_n \dfrac{1}{n}\dfrac{dn}{dx}$

The movement of these electrons creates an electrical current. The diffusion current density of electrons is then determined from Eq. (8.12):

Eq. (8.36) $J_e^{diff} = -nqv_e^{diff} = +qD_n \dfrac{dn}{dx}$

Similar relations to Eq. (8.35) and Eq. (8.36) can be obtained for the diffusion of holes:

Eq. (8.37) $v_h^{diff} = -D_p \dfrac{1}{p}\dfrac{dp}{dx}$

Eq. (8.38) $J_h^{diff} = -pqv_h^{diff} = -qD_p \dfrac{dp}{dx}$

where p is the concentration of holes. Note that there is a sign change from Eq. (8.36) to Eq. (8.38) which is due to the positive charge of the hole. There is no such sign change from Eq. (8.35) to Eq. (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 an 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): $J_e^{drift} = nq\mu_e E$; while the electrical diffusion current density is given by Eq. (8.36): $J_e^{diff} = +qD_n \frac{dn}{dx}$. At the thermal equilibrium of this system, the sum of these two current densities:

Eq. (8.39) $$J_e^{drift} + J_e^{diff} = nq\mu_e E + qD_n \frac{dn}{dx}$$

must be equal to zero, i.e.:

Eq. (8.40) $$nq\mu_e E + qD_n \frac{dn}{dx} = 0$$

This first order differential equation can be rewritten as:

Eq. (8.41) $$\frac{dn}{dx} + \frac{\mu_e E}{D_n} n = 0$$

which leads to the solution:

Eq. (8.42) $$n(x) = n(0)\exp\left(-\frac{\mu_e Ex}{D_n}\right)$$

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

Chapter 3. For a non-degenerate semiconductor, the electron concentration according to Boltzmann statistics should be given by:

Eq. (8.43) $$n(x) = n(0)\exp\left(-\frac{qEx}{k_b T}\right)$$

because qEx is the potential energy of the electron in an electric field strength of magnitude E. Comparing Eq. (8.42) and Eq. (8.43), we obtain the relation:

$$\frac{\mu_e E}{D_n} = \frac{qE}{k_b T}$$

or:

Eq. (8.44) $$\frac{D_n}{\mu_e} = \frac{k_b T}{q}$$

A similar relation can be obtained for holes:

Eq. (8.45) $$\frac{D_p}{\mu_h} = \frac{k_b T}{q}$$

Eq. (8.44) and Eq. (8.45) are called the Einstein relations and are valid only for non-degenerate 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 $x \rightarrow \infty$. 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:

Eq. (8.46) $\Delta n(x) = n(x) - n_0$

the excess electron concentration which is a function of position. A possible shape for $\Delta n(x)$ is shown in Fig. 8.7.

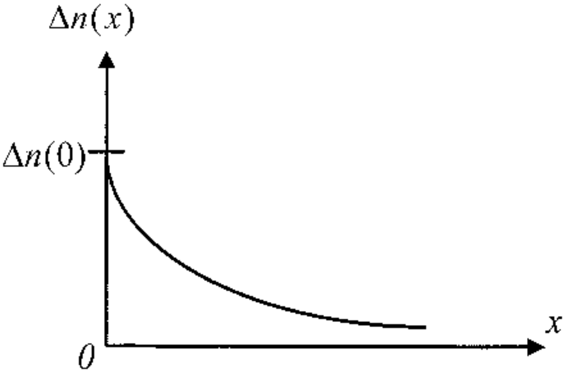


Fig. 8.7. Excess electron concentration in a one-dimensional model. The excess concentration decreases, as it gets deeper into the material as a result of recombination. The decrease has an exponential dependence.

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 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:

Eq. (8.47) $$R(x)=\frac{\Delta n(x)}{\tau_n}$$

This quantity has the units of cm^{-3}.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 in-flux and an out-flux of electrons, denoted respectively $\left(\Phi_e^{diff}\right)_{in}$ and $\left(\Phi_e^{diff}\right)_{out}$ and shown in Fig. 8.8.

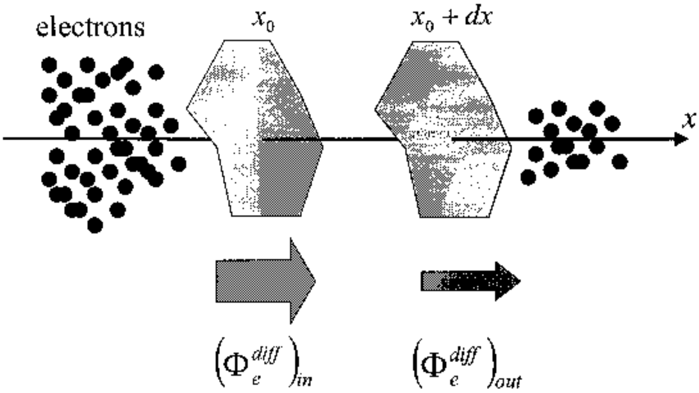

electrons
x_0
x_0+dx
x
$\left(\Phi_e^{diff}\right)_{in}$
$\left(\Phi_e^{diff}\right)_{out}$

Fig. 8.8. Schematic of the in-flux and out-flux 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.

If $\left(\Phi_e^{diff}\right)_{in}>\left(\Phi_e^{diff}\right)_{out}$, there is a net in-flux or accumulation of electrons, but if $\left(\Phi_e^{diff}\right)_{out}>\left(\Phi_e^{diff}\right)_{in}$, there is a net out-flux or depletion of electrons in this region. Under steady-state conditions, there must not be a never-ending accumulation or depletion of electrons. The in-flux of electrons must therefore be equal to the sum of the out-flux 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:

Eq. (8.48) $$\left(\Phi_e^{diff}\right)_{in}=\left(\Phi_e^{diff}\right)_{out}+R(x_0)dx$$

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:

Eq. (8.49) $$\begin{cases} \left(\varPhi_e^{diff}\right)_{in} = -D_n\left(\dfrac{d(\varDelta n)}{dx}\right)_{x=x_0} \\ \left(\varPhi_e^{diff}\right)_{out} = -D_n\left(\dfrac{d(\varDelta n)}{dx}\right)_{x=x_0+dx} \end{cases}$$

Eq. (8.48) becomes then:

$$-D_n\left(\frac{d(\Delta n)}{dx}\right)_{x=x_0} = -D_n\left(\frac{d(\Delta n)}{dx}\right)_{x=x_0+dx} + R(x_0)dx$$

which can be rewritten as:

$$D_n\frac{\left(\dfrac{d(\Delta n)}{dx}\right)_{x=x_0+dx} - \left(\dfrac{d(\Delta n)}{dx}\right)_{x=x_0}}{dx} = R(x_0)$$

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.:

Eq. (8.50) $$D_n\left(\frac{d^2(\Delta n)}{dx^2}\right)_{x=x_0} = R(x_0)$$

This relation is valid for any arbitrarily chosen position x_0, which means that the following equation must be satisfied:

$$D_n\frac{d^2(\Delta n)}{dx^2} = R(x)$$

Equating to Eq. (8.47), we get the differential equation that governs the shape of the excess electron concentration $\Delta n(x)$:

Eq. (8.51) $$D_n \frac{d^2(\Delta n)}{dx^2} = \frac{\Delta n}{\tau_n}$$

This equation can be rewritten as:

Eq. (8.52) $$\frac{d^2(\Delta n)}{dx^2} - \frac{\Delta n}{D_n \tau_n} = 0$$

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:

Eq. (8.53) $$L_n = \sqrt{D_n \tau_n}$$

The solution to Eq. (8.52) then has the general form:

Eq. (8.54) $$\Delta n(x) = Ae^{\frac{x}{L_n}} + Be^{-\frac{x}{L_n}}$$

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:

Eq. (8.55) $$\Delta n(x) = \Delta n(0) e^{-\frac{x}{L_n}}$$

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:

Eq. (8.56) $$L_p = \sqrt{D_p \tau_p}$$

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 τ_p=2×10^{-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's relations, we can determine the diffusion coefficient: $D_p = \dfrac{k_b T \mu_h}{q}$. The hole mobility in silicon being about μ_h=450 cm^2/Vs, we get:

$$L_p = \sqrt{\frac{k_b T \mu_h}{q} \tau_p}$$

$$= \sqrt{\frac{\left(1.38066 \times 10^{-23}\right) \times 300 \times \left(450 \times 10^{-4}\right)}{1.60218 \times 10^{-19}} \times 2 \times 10^{-10}}$$

$$= 0.48\,\mu m$$

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:

Eq. (8.57) $$\begin{cases}\Delta n(t) = n(t) - n_0 \\ \Delta p(t) = p(t) - p_0\end{cases}$$

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 $\text{cm}^{-3}.\text{s}^{-1}$.

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.

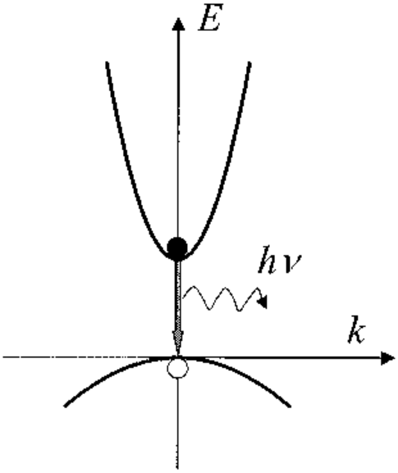


Fig. 8.9. Schematic E-k diagram of a direct band-to-band recombination process. The recombining electron and hole have the same wavevector.

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:

Eq. (8.58) $$R = r(T)n(t)p(t)$$

where $r(T)$ is the recombination coefficient, which is expressed in units of cm^3.s, and T is the temperature.

In a non-equilibrium situation when the excess generation rate is non zero, the net change in the electron and hole densities is given by:

Eq. (8.59) $$-\frac{dn}{dt} = -\frac{d(\Delta n)}{dt} = R - G - G_t$$

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.):

Eq. (8.60) $G_t = r(T)n_0 p_0$ or simply $G_t = r(T)n_i^2$

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 Eq. (8.58) and Eq. (8.60), Eq. (8.59) becomes:

Eq. (8.61) $$-\frac{d(\Delta n)}{dt} = r\left[np - n_i^2\right]$$

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.:

Eq. (8.62) $$-\frac{d(\Delta n)}{dt} = r\left[n_0 p_0 + n_0 \Delta p + p_0 \Delta n + \Delta n \Delta p - n_i^2\right]$$

One obvious simplification can be immediately made in the previous expression as $n_0 p_0 = n_i^2$ from Eq. (7.31). For further simplicity, we can assume the $\Delta n = \Delta p$, 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. Eq. (8.62) then becomes:

Eq. (8.63) $$-\frac{d(\Delta p)}{dt} = -\frac{d(\Delta n)}{dt} = r\left[(n_0 + p_0) + \Delta n\right]\Delta n$$

We can successively transform Eq. (8.63) into:

$$-\frac{d(\Delta n)\big/dt}{\left[(n_0 + p_0) + \Delta n\right]\Delta n} = r$$

$$\frac{1}{(n_0+p_0)}\left(\frac{d(\Delta n)/dt}{(n_0+p_0)+\Delta n}-\frac{d(\Delta n)/dt}{\Delta n}\right)=r$$

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:

$$\frac{1}{(n_0+p_0)}\left[\ln((n_0+p_0)+\Delta n)-\ln(\Delta n)\right]_0^t=rt$$

$$\frac{1}{(n_0+p_0)}\left[\ln\left(\frac{(n_0+p_0)+\Delta n}{\Delta n}\right)\right]_0^t=rt$$

$$\ln\left(\frac{(n_0+p_0)+\Delta n(t)}{\Delta n(t)}\right)-\ln\left(\frac{(n_0+p_0)+\Delta n(0)}{\Delta n(0)}\right)=r(n_0+p_0)t$$

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:

Eq. (8.64)

$$\Delta p(t)=\Delta n(t)=\frac{(n_0+p_0)\Delta n(0)}{[(n_0+p_0)+\Delta n(0)]\exp[r(n_0+p_0)t]-\Delta n(0)}$$

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:

$$\Delta n(t) \approx \frac{(n_0 + p_0)\Delta n(0)}{(n_0 + p_0)\exp[r(n_0 + p_0)t] - \Delta n(0)}$$
$$\approx \frac{(n_0 + p_0)\Delta n(0)}{(n_0 + p_0)\exp[r(n_0 + p_0)t]}$$

or simply:

Eq. (8.65) $\Delta n(t) \approx \Delta n(0)\exp[-r(n_0 + p_0)t]$

and similarly for $\Delta p(t)$:

Eq. (8.66) $\Delta p(t) \approx \Delta p(0)\exp[-r(n_0 + p_0)t]$

By defining a direct band-to-band recombination lifetime for electrons and holes as:

Eq. (8.67) $\tau_p = \tau_n = \dfrac{1}{r(n_0 + p_0)}$

we obtain:

Eq. (8.68) $\begin{cases} \Delta n(t) \approx \Delta n(0)e^{-\frac{t}{\tau_n}} \\ \Delta p(t) \approx \Delta p(0)e^{-\frac{t}{\tau_p}} \end{cases}$

This is the same lifetime introduced in Eq. (8.47). Indeed, in the current conditions, we have by using Eq. (8.59)and Eq. (8.68):

$$R - G_i = -\frac{d(\Delta n)}{dt} = \frac{1}{\tau_n}\Delta n(0)e^{-\frac{t}{\tau_n}}$$

or:

Eq. (8.69) $R - G_i = \dfrac{\Delta n(t)}{\tau_n}$

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 recombinations (SRH) 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 sub-section 8.6.1.

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.

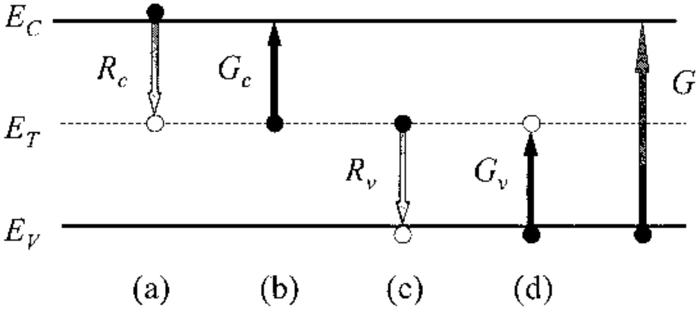


Fig. 8.10. The four possible transitions for an electron and involving a recombination center in the bandgap: (a) capture of an electron from the conduction band by the center, (b) emission of an electron from the center into the conduction band, (c) emission of an electron from the center into a vacant state in the valence band, and (d) capture of an electron from the valence band by the center.

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:

Eq. (8.70) $$N_T f_e(E_T) = \frac{N_T}{\exp\left(\frac{E_T - E_F}{k_b T}\right) + 1}$$

where f_e is the Fermi-Dirac distribution given by Eq. (4.28). The density of the recombination center states which are empty of electrons at equilibrium is given by:

Eq. (8.71) $$N_T[1 - f_e(E_T)] = \frac{N_T}{1 + \exp\left(-\frac{E_T - E_F}{k_b T}\right)}$$

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.

(1) 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}.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 $v_{th}\sigma_n$ of two quantities: the electron thermal velocity v_{th} (in units of cm.s^{-1}) and the capture cross-section σ_n of electrons for this particular recombination center (in units of cm^2). These two parameters can be better understood by considering the illustration in Fig. 8.11. It shows that the electrons which have a velocity v_{th} and which will reach a surface of area σ_n are located in a volume equal to the product $v_{th}\sigma_n$ during a unit time.

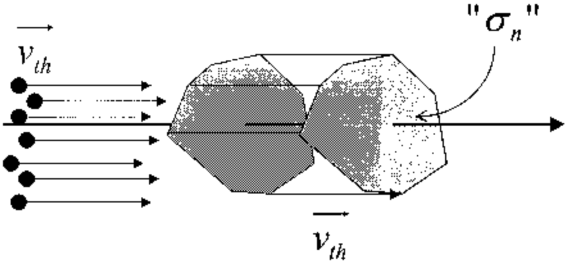


Fig. 8.11. Schematic illustration of the concepts of electron thermal velocity and capture cross-section. Using ballistic terminology, the electrons moving with the thermal velocity which would collide with an object having a cross-section equal to σ_n are located in the volume delimited by the two shaded surfaces in this figure.

The electron thermal velocity in a non-degenerate semiconductor is given by:

Eq. (8.72) $$v_{th} = \sqrt{\frac{3k_bT}{m}}$$

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:

Eq. (8.73) $R_c = v_{th}\sigma_n n N_T(1-f)$

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:

Eq. (8.74) $G_c = e_n N_T f$

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 Eq. (8.73) and Eq. (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:

Eq. (8.75) $R_v = v_{th}\sigma_p p N_T f$

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:

Eq. (8.76) $G_v = e_p N_T(1-f)$

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:

Eq. (8.77) $\begin{cases} R_c = G_c + G \\ R_v = G_v + G \end{cases}$

(2) 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 Eq. (8.73) to Eq. (8.76) we get:

$$\begin{cases} v_{th}\sigma_n n_0 N_T\left(1-f_e\right)=e_n N_T f_e \\ v_{th}\sigma_p p_0 N_T f_e = e_p N_T\left(1-f_e\right) \end{cases}$$

which allow us to extract the electron and hole emission probabilities:

Eq. (8.78)
$$\begin{cases} e_n = v_{th}\sigma_n n_0 \dfrac{1-f_e}{f_e} \\ e_p = v_{th}\sigma_p p_0 \dfrac{f_e}{1-f_e} \end{cases}$$

This last set of equations can be simplified by using the expression for the Fermi-Dirac distribution in Eq. (4.28) to obtain:

Eq. (8.79)
$$\frac{1-f_e}{f_e}=\exp\left(\frac{E_T-E_F}{k_bT}\right)$$

and by using the expressions of n_0 and p_0 given in Eq. (7.21) and Eq. (7.29) for a non-degenerate semiconductor:

$$\begin{aligned} n_0\frac{1-f_e}{f_e} &= N_c\exp\left(\frac{E_F-E_C}{k_bT}\right)\exp\left(\frac{E_T-E_F}{k_bT}\right) \\ &= N_c\exp\left(\frac{E_T-E_C}{k_bT}\right) \end{aligned}$$

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 ($E_F=E_T$):

Eq. (8.80)
$$n_T=n_0\frac{1-f_e}{f_e}=N_c\exp\left(\frac{E_T-E_C}{k_bT}\right)$$

A similar expression can be derived for:

Eq. (8.81) $$p_T = p_0 \frac{f_e}{1-f_e} = N_v \exp\left(\frac{E_V - E_T}{k_b T}\right)$$

Therefore, Eq. (8.78) is simplified into:

Eq. (8.82) $$\begin{cases} e_n = v_{th}\sigma_n n_T \\ e_p = v_{th}\sigma_p p_T \end{cases}$$

(3) The non-equilibrium distribution function f

The non-equilibrium distribution function, included in the expressions of the transition rates in Eq. (8.73) to Eq. (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:

$$v_{th}\sigma_n n N_T(1-f) - v_{th}\sigma_p p N_T f = e_n N_T f - e_p N_T(1-f)$$

Using Eq. (8.82), we obtain:

$$v_{th}\sigma_n n N_T(1-f) - v_{th}\sigma_p p N_T f = v_{th}\sigma_n n_T N_T f - v_{th}\sigma_p p_T N_T(1-f)$$

and, after simplifying by v_{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:

Eq. (8.83) $$f = \frac{\sigma_n n + \sigma_p p_T}{\sigma_n(n+n_T) + \sigma_p(p+p_T)}$$

(4) 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.:

Eq. (8.84) $$-\frac{d(\Delta n)}{dt} = R_c - G_c$$

This quantity is also equal to the net recombination rate of holes from the valence band in view of Eq. (8.77):

Eq. (8.85) $$-\frac{d(\Delta p)}{dt} = R_v - G_v$$

Using the non-equilibrium distribution function (Eq. (8.83))and the expressions for R_c, G_c and e_n in Eq. (8.73), Eq. (8.74) and Eq. (8.82), we can calculate successively:

$$\begin{aligned}
R_c - G_c &= v_{th}\sigma_n n N_T(1-f) - e_n N_T f \\
&= v_{th}\sigma_n N_T\left[n - (n+n_T)f\right] \\
&= v_{th}\sigma_n N_T\left[n - (n+n_T)\frac{\sigma_n n + \sigma_p p_T}{\sigma_n(n+n_T)+\sigma_p(p+p_T)}\right] \\
&= \frac{v_{th}\sigma_n N_T}{\sigma_n(n+n_T)+\sigma_p(p+p_T)}\left[n\sigma_p(p+p_T) - (n+n_T)\sigma_p p_T\right] \\
&= \frac{v_{th}\sigma_n N_T}{\sigma_n(n+n_T)+\sigma_p(p+p_T)}\sigma_p\left[np - n_T p_T\right]
\end{aligned}$$

From the definitions of n_T and p_T in Eq. (8.80) and Eq. (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:

Eq. (8.86) $$R_c - G_c = v_{th}\sigma_n\sigma_p N_T \frac{\left(np - n_i^2\right)}{\sigma_n(n+n_T)+\sigma_p(p+p_T)}$$

Introducing the excess carriers Δn and Δp as in Eq. (8.57), and still assuming $\Delta n = \Delta p$, we get:

Eq. (8.87)

$$R_c - G_c = v_{th}\sigma_n\sigma_p N_T \frac{(n_0 + p_0 + \Delta n)\Delta n}{\sigma_n(n_0 + n_T + \Delta n) + \sigma_p(p_0 + p_T + \Delta n)}$$

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):

Eq. (8.88) $$-\frac{d(\Delta n)}{dt} = R_c - G_c \approx v_{th}\sigma_p N_T \Delta n$$

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.:

Eq. (8.89) $$\tau_{p_0} = \frac{1}{v_{th}\sigma_p N_T}$$

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:

Eq. (8.90) $$\tau_{n_0} = \frac{1}{v_{th}\sigma_n N_T}$$

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 Eq. (8.89) and Eq. (8.90), we can simplify Eq. (8.87):

Eq. (8.91) $$R_c - G_c = \frac{(n_0 + p_0 + \Delta n)\Delta n}{\tau_{p_0}(n_0 + n_T + \Delta n) + \tau_{n_0}(p_0 + p_T + \Delta n)}$$

From Eq. (8.84) and Eq. (8.85), we can write:

Eq. (8.92)
$$-\frac{d(\Delta n)}{dt} = -\frac{d(\Delta p)}{dt} = \frac{(n_0 + p_0 + \Delta n)\Delta n}{\tau_{p_0}(n_0 + n_T + \Delta n) + \tau_{n_0}(p_0 + p_T + \Delta n)}$$

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.:

Eq. (8.93) $$\tau_n(t) = \tau_p(t) = \frac{\tau_{p_0}(n_0 + n_T + \Delta n) + \tau_{n_0}(p_0 + p_T + \Delta n)}{(n_0 + p_0 + \Delta n)}$$

which becomes independent of time for weak excitation levels $\Delta n << n_0, p_0$:

Eq. (8.94) $$\tau_n = \tau_p = \frac{\tau_{p_0}(n_0 + n_T) + \tau_{n_0}(p_0 + p_T)}{(n_0 + p_0)}$$

From this relation, we can easily find the two previous particular cases, i.e. that 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. 8.12(a), 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 band (CC). In the case of an Auger hole, the valence band structure is more complex than the conduction band, as we saw in sub-section 4.4.3. We must then distinguish whether this hole is excited into the light-hole band (CHLH process, Fig. 8.12(b) or the spin-orbit split-off band (CHSH process, Fig. 8.12(c)).

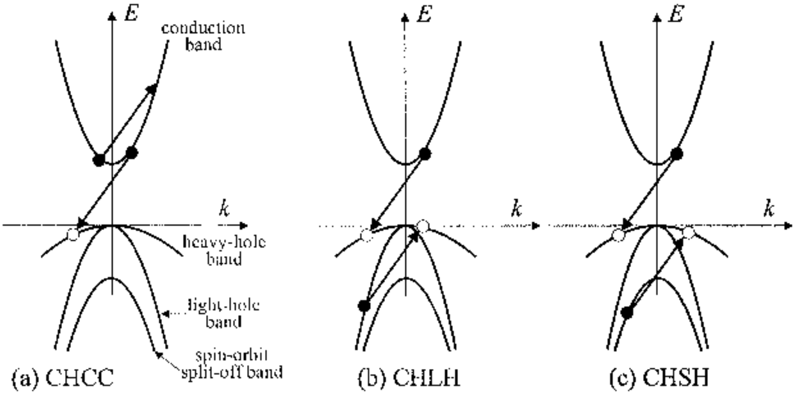


Fig. 8.12. Auger recombination processes semiconductors. The energy released through the recombination of an electron in the conduction band and a hole in the valence band is yielded to: (a) another electron in the conduction band which is then excited to a higher state in the band, (b) an electron in 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.

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 cm^{-3}.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:

Eq. (8.95) $$R_{CHCC} = r_1 n^2 p$$

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:

Eq. (8.96) $$R_{CHLH+CHSH} = r_2 n p^2$$

where r_2 is the Auger recombination coefficient when Auger holes are excited.

The total Auger recombination rate is therefore:

Eq. (8.97) $R = R_{CHCC} + R_{CHLH+CHSH} = r_1 n^2 p + r_2 n p^2$

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:

Eq. (8.98) $G_t = R = r_1 n_0{}^2 p_0 + r_2 n_0 p_0{}^2$

Let us now consider the relaxation process, which occurs after the external source of generation is removed $(G = 0)$. Taking into account Eq. (8.97) and Eq. (8.98), Eq. (8.59) becomes:

Eq. (8.99) $-\frac{d(\Delta n)}{dt} = R - G_t = r_1 (n^2 p - n_0^2 p_0) + r_2 (n p^2 - n_0 p_0{}^2)$

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 $\tau_n = \tau_p$ such that:

$$-\frac{d(\Delta n)}{dt} = -\frac{d(\Delta p)}{dt} = \frac{\Delta p}{\tau_p} = \frac{\Delta n}{\tau_n}$$

Eq. (8.100)

$$\tau_n(t) = \tau_p(t) = \frac{1}{\begin{array}{r} r_1\left[\left(n_0^2 + 2n_0 p_0\right) + (2n_0 + p_0)\Delta n + (\Delta n)^2\right] \\ + r_2\left[\left(p_0^2 + 2n_0 p_0\right) + (2p_0 + n_0)\Delta n + (\Delta n)^2\right] \end{array}}$$

which becomes independent of time for weak excitation levels $\Delta n << n_0, p_0$:

Eq. (8.101) $$\tau_n = \tau_p = \frac{1}{r_1\left(n_0^2 + 2n_0 p_0\right) + r_2\left(p_0^2 + 2n_0 p_0\right)}$$

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 sub-section 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}.s^{-1} and can be obtained by analogy with the SRH recombination in Eq. (8.87):

Eq. (8.102)
$$\left(R - G_t\right)_s = v_{th}\sigma_n\sigma_p\left(N_T\right)_s \frac{\left(n_0 + p_0 + \Delta n\right)\Delta n}{\sigma_n\left(n_0 + n_T + \Delta n\right) + \sigma_p\left(p_0 + p_T + \Delta n\right)}$$

Here, Δn is the excess electron concentration near the surface considered. We can rewrite this relation as:

Eq. (8.103) $$-\frac{d(\Delta n)}{dt} = \left(R - G_t\right)_s = S_n \Delta n$$

where:

Eq. (8.104) $$S_n = v_{th}\sigma_n\sigma_p\left(N_T\right)_s \frac{\left(n_0 + p_0 + \Delta n\right)}{\sigma_n\left(n_0 + n_T + \Delta n\right) + \sigma_p\left(p_0 + p_T + \Delta n\right)}$$

This quantity is expressed in units of cm.s^{-1}, and has thus the same dimension as a velocity. It is called the surface recombination velocity.

8.7. Quasi-Fermi energy

In section 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 Eq. (7.18) and Eq. (7.27) in the general case, and Eq. (7.21) and Eq. (7.29) in the non-degenerate. 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:

Eq. (8.105) $$\begin{cases} n = n_0 + \Delta n \\ p = p_0 + \Delta p \end{cases}$$

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 non-degenerate semiconductor only.

Therefore, by analogy with Eq. (7.21), the non-equilibrium electron concentration in the conduction band is given by:

Eq. (8.106) $$n = N_c \exp\left(\frac{E_{F_n} - E_C}{k_b T}\right)$$

where the quantity E_{F_n} is used instead of the Fermi energy E_F. This quantity is called the electron quasi-Fermi energy. Using this expression, Eq. (7.21) and Eq. (8.105), we can write:

Eq. (8.107) $$\frac{\Delta n}{n_0} = \frac{n}{n_0} - 1 = \exp\left(\frac{E_{F_n} - E_F}{k_b T}\right) - 1$$

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. (4.28) with E_F replaced by E_{F_n} :

Eq. (8.108) $$f_{e_n}(E)=\frac{1}{\exp\left(\frac{E-E_{F_n}}{k_bT}\right)+1}$$

A similar concept can be introduced for holes in the valence band. The hole quasi-Fermi energy E_{F_p} is defined such that:

Eq. (8.109) $$p=N_v\exp\left(\frac{E_V-E_{F_p}}{k_bT}\right)$$

A quasi-Fermi-Dirac distribution for holes can also be defined by analogy with Eq. (7.23):

Eq. (8.110) $$f_{h_p}(E)=\frac{1}{\exp\left(\frac{E_{F_p}-E}{k_bT}\right)+1}$$

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. $E_{F_n}=E_{F_p}=E_F$.

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} is related to the excess carrier concentration through the expression:

$$E_{Fn} - E_F = k_b T \ln\left(\frac{\Delta n}{n_0}\right) \text{ and } E_F - E_{Fp} = k_b T \ln\left(\frac{\Delta p}{p_0}\right),$$

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: $E_{Fn} - E_F = E_F - E_{Fp} = 0.0095 k_b T$.

8.8. Summary

In this Chapter, we have covered a few important non-equilibrium transport phenomena involving charge carriers. First 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.

Finally, 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.

Further reading

Anselm, A., *Introduction to Semiconductor Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

Ashcroft, N.W. and Mermin N.D., *Solid State Physics*, Holt, Rinehart and Winston, New York, 1976.

Cohen, M.M., *Introduction to the Quantum Theory of Semiconductors*, Gordon and Breach, New York, 1972.

Highlights in Condensed Matter Physics and Future Concepts, ed. L. Esaki, NATO Science Forum Series, Vol. 285, Plenum Press, New York, 1991.

Ferry, D.K., *Semiconductors*, Macmillan, New York, 1991.

Hummel, R.E., *Electronic Properties of Materials*, Springer-Verlag, New York, 1986.

Orton, J.W. and Blood, P., *The Electrical Characterization of Semiconductors: Measurement of Minority Carrier Properties*, Academic Press, San Diego, 1990.

Pankove, J.I., *Optical Processes in Semiconductors*, Dover, New York, 1975.

Peyghambarian, N., Koch, S.W., and Mysyrowicz, A., *Introduction to Semiconductor Optics*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

Pierret, R.F., *Advanced Semiconductor Fundamentals*, Addison-Wesley, Reading, MA, 1989.

Pollock, C.R., *Fundamentals of Optoelectronics*, Irwin, Burr Ridge, IL, 1995.

Ridley, B.K., *Quantum Processes in Semiconductors*, Oxford University Press, New York, 1999.

Rogalski, A., *Infrared Photon Detectors*, Bellingham, Washington, 1995.

Streetman, B.G., *Solid State Electronic Devices*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

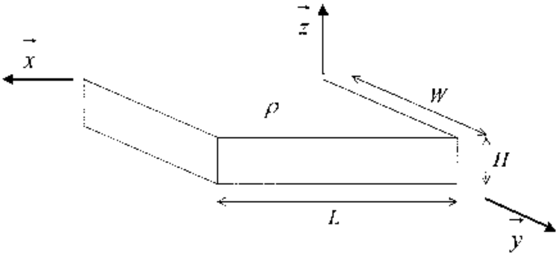
Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

Wang, S., *Fundamentals of Semiconductor Theory and Device Physics*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

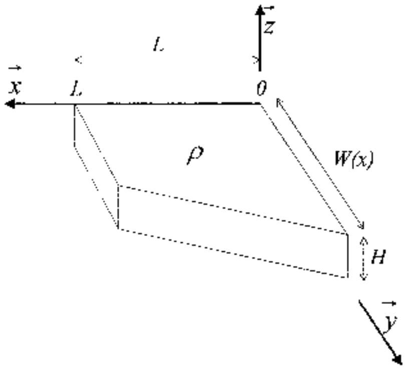
Wolfe, C.M., Holonyak, Jr., N., and Stillman, G.E., *Physical Properties of Semiconductors*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

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 and 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.

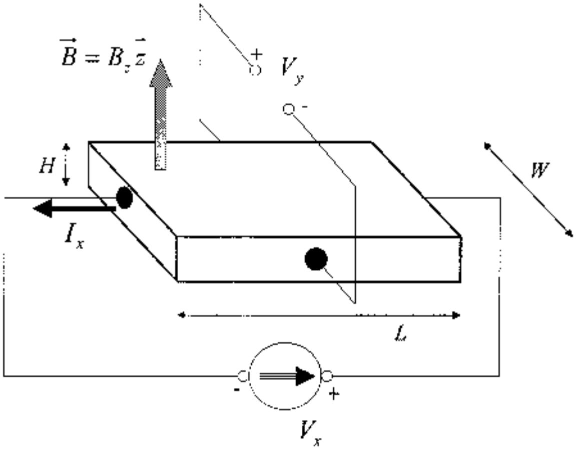


3. Do the same as in Problem 2, but in the y-direction.

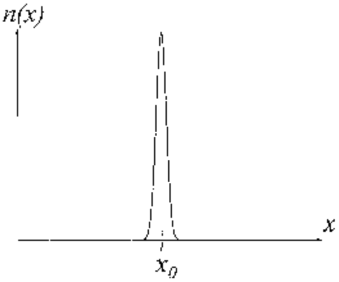
4. Consider the Hall effect measurement experiment depicted in the figure below. The dimensions of the semiconductor slab are L=2 mm, W=1

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, the resistivity of the semiconductor (assumed uniform).

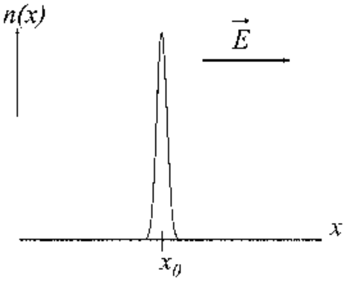


5. Consider an experiment where excess electrons are generated in a "burst" at t=0 at x=x_0 in a semiconductor, resulting in the concentration profile $n(x)$ shown in the figure below.



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 an 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.

9. Semiconductor p-n and Metal-Semiconductor Junctions

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 non-homogeneous 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. 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.

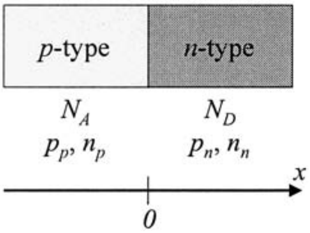


Fig. 9.1. Ideal p-n junction model, in which one side of the junction is a purely p-type semiconductor and the other a purely n-type semiconductor. Both materials are uniformly doped.

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).

Eq. (9.1) $$p_p n_p = p_n n_n = n_i^2$$

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:

Eq. (9.2) $$\begin{cases} p_p = N_A \left(10^{16}\, cm^{-3}\right) \\ n_p = \dfrac{n_i^2}{N_A} \left(10^5\, cm^{-3}\right) \end{cases} \text{and} \begin{cases} n_n = N_D \left(10^{17}\, cm^{-3}\right) \\ p_n = \dfrac{n_i^2}{N_D} \left(10^4\, cm^{-3}\right) \end{cases}$$

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. 9.2. 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 a 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.

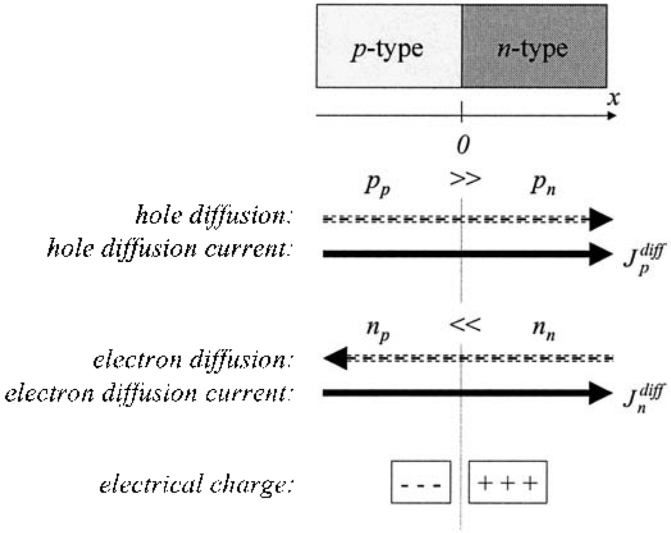


Fig. 9.2. Hole and electron diffusion across a p-n junction. The holes diffuse from the left to the right, which leads to a diffusion electrical current from the left to the right as well. By contrast, the electrons diffuse from the right to the left, but this leads to a diffusion electrical current 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 result 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 Chapter 8, this electric field will generate the drift of the positively charged holes and the negatively charged electrons. By comparing Fig. 9.2 and Fig. 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 $J^{diffusion}$ and drift currents J^{drift} are exactly balanced for each type of carrier, i.e. holes and electrons taken independently:

Eq. (9.3) $$\begin{cases} J_h^{diff} + J_h^{drift} = 0 \\ J_e^{diff} + J_e^{drift} = 0 \end{cases}$$

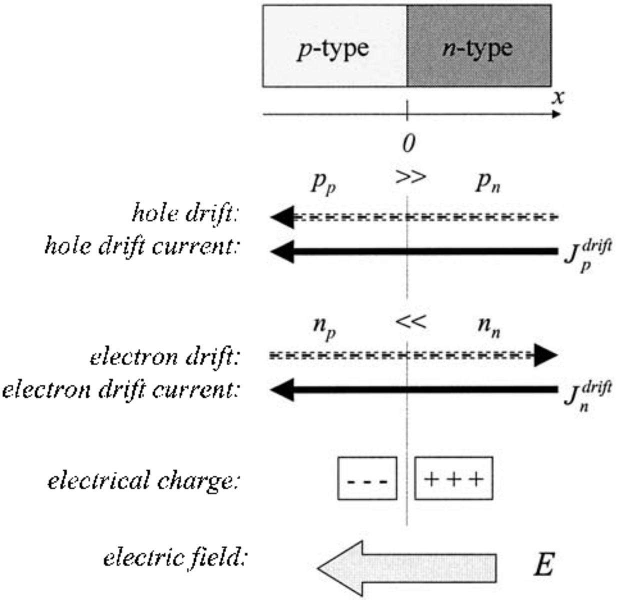


Fig. 9.3. Hole and electron drift across a p-n junction. Under the influence of the built-in electric field, the holes drift from the right to the left, which leads to a drift electrical current from the right to the left as well. By contrast, the electrons drift from the left to the right, but this 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.

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. 9.4(a). The charge distribution within this region is modeled as follows: we consider that there is a uniform concentration of negative charges for $-x_{p0}<x<0$ equal to $Q(x)=-qN_A$ (where N_A is the total concentration of acceptors in the *p*-type region), and a uniform concentration of positive charges for $0<x<x_{n0}$ and equal to $Q(x)=+qN_D$ (where N_D is the total concentration of donors 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. 9.4(b). The width of the space charge region, also called depletion width, is then given by:

Eq. (9.4) $W_0 = x_{n0} + x_{p0}$

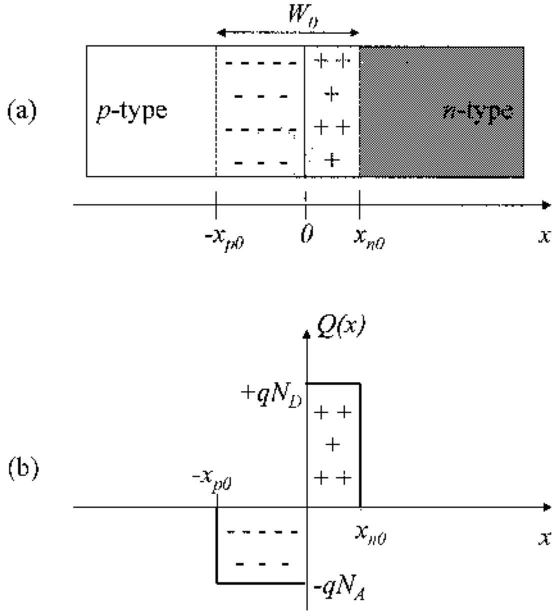


Fig. 9.4. (a) Space charge region in a p-n junction. Near the junction area, the p-type region is negatively charged as a result of the diffusion of charge carriers. (b) Electrical charge density in a p-n junction. To keep the overall charge neutrality, the total number of negative charges in the p-type 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}.

Outside of this space charge region, we assume that the semiconductor is at thermal equilibrium, i.e. 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. 9.5. In particular, we have:

Eq. (9.5)
$$\begin{cases} p(-x_{p0}) = p_p & \text{and} \quad p(x_{n0}) = p_n \\ n(-x_{p0}) = n_p & \text{and} \quad n(x_{n0}) = n_n \end{cases}$$

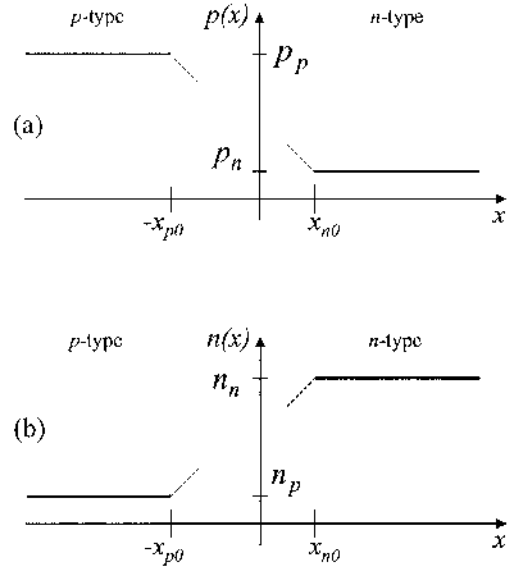


Fig. 9.5. (a) Hole and (b) electron concentrations in a p-n junction. In the depletion approximation, the hole and electron concentrations are assumed to be constant and equal to their equilibrium values outside of the depletion region.

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.:

$$qAN_A x_{p0} = qAN_D x_{n0}$$

where A is the cross-section area of the junction, and after simplification:

Eq. (9.6) $N_A x_{p0} = N_D x_{n0}$

Combining Eq. (9.4) and Eq. (9.6), we can express the quantities x_{p0} and x_{n0} as a function of the depletion width W_0:

Eq. (9.7)
$$\begin{cases} x_{p0}=\dfrac{N_D}{N_A+N_D}W_0 \\ x_{n0}=\dfrac{N_A}{N_A+N_D}W_0 \end{cases}$$

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:

$$\frac{x_{p0}}{x_{n0}}=\frac{N_D}{N_A}=\frac{10^{17}}{10^{18}}=0.1\,.$$

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:

Eq. (9.8)
$$\frac{dE(x)}{dx}=\frac{Q(x)}{\varepsilon}$$

where ε is the permittivity of the semiconductor material. This relation can be rewritten for either sides of the junction:

Eq. (9.9)
$$\begin{cases} \dfrac{dE(x)}{dx}=-\dfrac{qN_A}{\varepsilon} & \mathit{for}\ -x_{p0}<x<0 \\ \dfrac{dE(x)}{dx}=\dfrac{qN_D}{\varepsilon} & \mathit{for}\ 0<x<x_{n0} \end{cases}$$

From these relations we see that the electric field strength varies linearly on either side of the junction. By integrating Eq. (9.9) using the boundary conditions assumed in the depletion approximation:

Eq. (9.10) $E(-x_{p0}) = E(x_{n0}) = 0$

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:

$$\begin{cases} E(x) = \int\limits_{-x_{p0}}^{x} dEdx = \int\limits_{-x_{p0}}^{x} -\frac{qN_A}{\varepsilon}dx & for\ -x_{p0} < x < 0 \\ E(x) = \int\limits_{x_{n0}}^{x} dEdx = \int\limits_{x_{n0}}^{x} \frac{qN_D}{\varepsilon}dx & for\ 0 < x < x_{n0} \end{cases}$$

Eq. (9.11) $$\begin{cases} E(x) = -\frac{qN_A}{\varepsilon}(x + x_{p0}) & for\ -x_{p0} < x < 0 \\ E(x) = \frac{qN_D}{\varepsilon}(x - x_{n0}) & for\ 0 < x < x_{n0} \end{cases}$$

For x=0, we obtain two expressions for the electric field strength from the two previous expressions for $E(x)$:

Eq. (9.12) $$\begin{cases} E(0) = -\frac{qN_A}{\varepsilon}(x_{p0}) \\ E(0) = \frac{qN_D}{\varepsilon}(-x_{n0}) \end{cases}$$

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.

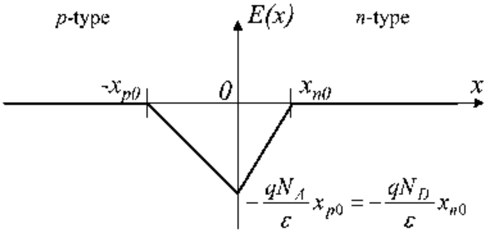


Fig. 9.6. Built-in electric field strength profile across a p-n junction. In the depletion approximation, the electric field strength is zero outside the depletion region because there is no net electrical charge. Within the depletion region, the electric field strength varies linearly with distance.

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:

Eq. (9.13) $$E(x) = -\frac{dV(x)}{dx}$$

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):

Eq. (9.14) $$\begin{cases} V(x) = \frac{qN_A}{\varepsilon}\left(\frac{x^2}{2} + x_{p0}x\right) & for -x_{p0} < x < 0 \\ V(x) = -\frac{qN_D}{\varepsilon}\left(\frac{x^2}{2} - x_{n0}x\right) & for\ 0 < x < x_{n0} \end{cases}$$

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 Fig. 9.7.

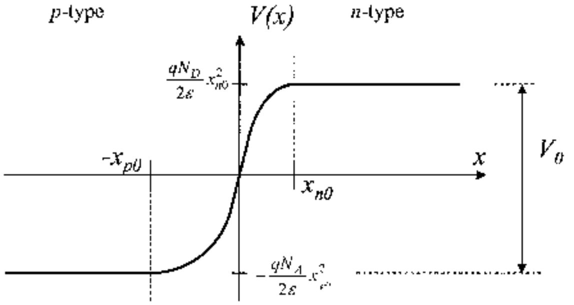


Fig. 9.7. Built-in potential profile across a p-n junction. In the depletion approximation, there is no variation of the potential outside the depletion region.

The 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_{p0} and x=x_{n0}:

Eq. (9.15) $$V_0 = V(x_{n0}) - V(-x_{p0})$$

This can be rewritten as:

Eq. (9.16) $$V_0 = \frac{qN_D}{\varepsilon}\frac{x_{n0}^2}{2} + \frac{qN_A}{\varepsilon}\frac{x_{p0}^2}{2}$$

Expressing *-x_{p0}* and *x_{n0}* as a function of the depletion width given in Eq. (9.7), we obtain:

Eq. (9.17) $$V_0 = \frac{q}{2\varepsilon}\frac{N_A N_D}{(N_A + N_D)}W_0^2$$

Another independent expression of the built-in potential can be obtained by expressing the balancing of the diffusion and drift currents. In Chapter 8 we determined analytical expressions for these currents in Eq. (8.12) and Eq. (8.38) for holes, and Eq. (8.12) and Eq. (8.36) for electrons. The total current from the motion of holes and that from the motion of electrons are given by:

Eq. (9.18) $$\begin{cases} J_h^{diff} + J_h^{drift} = -qD_p \dfrac{dp(x)}{dx} + q\mu_h p(x)E(x) \\ J_e^{diff} + J_e^{drift} = qD_n \dfrac{dn(x)}{dx} + q\mu_e n(x)E(x) \end{cases}$$

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:

Eq. (9.19) $$\begin{cases} D_p \dfrac{dp(x)}{dx} = \mu_h p(x)E(x) \\ D_n \dfrac{dn(x)}{dx} = -\mu_e n(x)E(x) \end{cases}$$

which can be rewritten using Eq. (9.13) as:

$$\begin{cases} \dfrac{D_p}{\mu_h}\dfrac{1}{p(x)}\dfrac{dp(x)}{dx} = -\dfrac{dV(x)}{dx} \\ \dfrac{D_n}{\mu_e}\dfrac{1}{n(x)}\dfrac{dn(x)}{dx} = \dfrac{dV(x)}{dx} \end{cases}$$

By integrating these equations, we get successively:

$$\begin{cases} \dfrac{D_p}{\mu_h} \displaystyle\int_{-x_{p0}}^{x_{n0}} \dfrac{1}{p(x)} \dfrac{dp(x)}{dx} dx = - \displaystyle\int_{-x_{p0}}^{x_{n0}} \dfrac{dV(x)}{dx} dx \\ \dfrac{D_n}{\mu_e} \displaystyle\int_{-x_{p0}}^{x_{n0}} \dfrac{1}{n(x)} \dfrac{dn(x)}{dx} dx = \displaystyle\int_{-x_{p0}}^{x_{n0}} \dfrac{dV(x)}{dx} dx \end{cases}$$

Using Eq. (9.5) and Eq. (9.15), and by taking into account the Einstein relations $\frac{D_p}{\mu_h} = \frac{D_n}{\mu_e} = \frac{k_b T}{q}$ obtained from Eq. (8.44) and Eq. (8.45), we get:

$$\begin{cases} \dfrac{k_b T}{q} \displaystyle\int_{p_p}^{p_n} \dfrac{dp}{p} = - \displaystyle\int_{0}^{V_0} dV \\ \dfrac{k_b T}{q} \displaystyle\int_{n_p}^{n_n} \dfrac{dn}{n} = \displaystyle\int_{0}^{V_0} dV \end{cases}$$

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.:

Eq. (9.20) $$V_0 = \frac{k_b T}{q} \ln\left(\frac{p_p}{p_n}\right) = \frac{k_b T}{q} \ln\left(\frac{n_n}{n_p}\right)$$

This can be rewritten into the form:

Eq. (9.21) $$\frac{p_p}{p_n} = \frac{n_n}{n_p} = \exp\left(\frac{qV_0}{k_b T}\right)$$

Using the expressions in Eq. (9.2), we can write the built-in potential as a function of the doping concentrations:

Eq. (9.22) $$V_0 = \frac{k_b T}{q} \ln\left(\frac{N_A N_D}{n_i^2}\right)$$

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, as soon as the probes are in contact with the material, contact potentials are created at the probes which cancel the built-in potential in the measurement.

9.2.5. Depletion width

It is now possible to relate the width *W₀* 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. (9.17), we can express the depletion width as:

Eq. (9.23) $$W_0 = \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A + N_D}{N_A N_D}\right)V_0}$$

which becomes, after considering Eq. (9.22):

Eq. (9.24) $$W_0 = \sqrt{\frac{2\varepsilon k_b T}{q^2}\left(\frac{N_A + N_D}{N_A N_D}\right)\ln\left(\frac{N_A N_D}{n_i^2}\right)}$$

The extent of the depletion width into each side of the p-n junction can then be determined by replacing *W₀* from Eq. (9.23) into Eq. (9.7):

Eq. (9.25) $$\begin{cases} x_{p0} = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_D}{N_A(N_A + N_D)}\right)V_0} \\ x_{n0} = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_A}{N_D(N_A + N_D)}\right)V_0} \end{cases}$$

These last two expressions show that the space charge region extends more into the region of lower doping, in accordance with sub-section 9.2.2.

Example

Q: Consider a GaAs abrupt p-n junction with a doping level on the *p*-type side of N_A=2×10^{17} cm^{-3} and a doping level on the *n*-type side of N_D=1×10^{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 (ε=13.1ε_0) and V_0 is the built-in potential. The latter is calculated from:

$$\begin{aligned} V_0 &= \frac{k_b T}{q}\ln\left(\frac{N_A N_D}{n_i^2}\right) \\ &= \frac{(1.38066\times 10^{-23})\times 300}{1.60218\times 10^{-19}}\ln\left(\frac{(2\times 10^{17})(1\times 10^{17})}{(1.79\times 10^{6})^2}\right) \\ &= 1.297\,V \end{aligned}$$

because the intrinsic carrier concentration in GaAs at 300 K is n_i=1.79×10^{6} cm^{-3}. The widths can then be calculated as:

$$x_{p0} = \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_D}{N_A(N_A+N_D)}\right)V_0}$$

$$= \sqrt{\frac{2\times(13.1\times 8.85418\times 10^{-14})}{1.60218\times 10^{-19}} \times\left(\frac{1\times 10^{17}}{(2\times 10^{17})(2\times 10^{17}+1\times 10^{17})}\right)\times 1.297}$$

$$x_{p0} = 5.6\times 10^{-6}\ cm$$

$$x_{p0} = 56\,nm$$

and

$$x_{n0}=\sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A}{N_D\left(N_A+N_D\right)}\right)V_0}$$
$$=\frac{N_A}{N_D}x_{p0}$$
$$=\frac{2\times10^{17}}{1\times10^{17}}5.6\times10^{-6}$$
$$=11.2\times10^{-6}\ cm$$
$$x_{n0}=112\,nm$$

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. 9.8, 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.

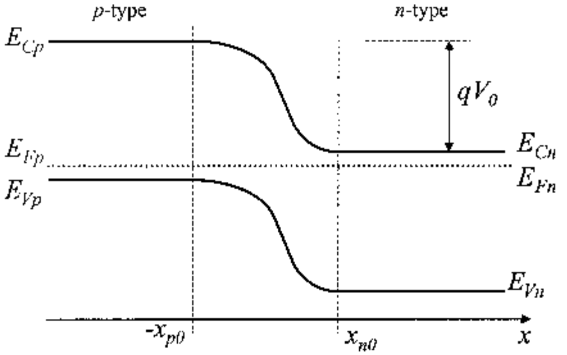


Fig. 9.8. Energy band profile across a p-n junction. This profile is obtained by multiplying the potential in Fig. 9.7 by -q, the electrical charge of electrons.

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:

Eq. (9.26) $$E_{Vp} - E_{Vn} = E_{Cp} - E_{Cn} = qV_0$$

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=2×10^{17} cm^{-3} and a doping level on the *n*-type side of N_D=1×10^{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 energy 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 non-degenerate case:

Eq. (9.27) $$\begin{cases} p_p = N_v \exp\left(\dfrac{E_{Vp} - E_{Fp}}{k_bT}\right) \\ p_n = N_v \exp\left(\dfrac{E_{Vn} - E_{Fn}}{k_bT}\right) \end{cases}$$

Utilizing Eq. (9.21), we get:

$$\exp\left(\frac{qV_0}{k_bT}\right) = \frac{p_p}{p_n} = \frac{\exp\left(\dfrac{E_{Vp} - E_{Fp}}{k_bT}\right)}{\exp\left(\dfrac{E_{Vn} - E_{Fn}}{k_bT}\right)}$$

Eq. (9.28) $$\exp\left(\frac{qV_0}{k_bT}\right) = \exp\left(\frac{E_{Vp} - E_{Vn}}{k_bT}\right)\exp\left(\frac{E_{Fn} - E_{Fp}}{k_bT}\right)$$

In addition, by using Eq. (9.26) in this expression, we get:

$$1 = \exp\left(\frac{E_{Fn} - E_{Fp}}{k_b T}\right)$$

which means that E_{Fn}=E_{Fp}, i.e. the Fermi energy 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 transfer 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 non-symmetrical nature, a p-n junction will exhibit different properties depending on the polarity of the external voltage or bias applied. 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.

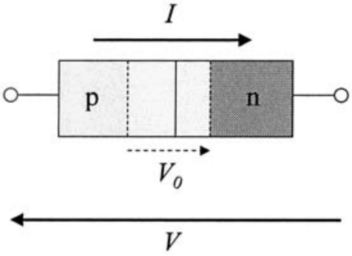


Fig. 9.9. Convention for the polarity of the external voltage and current.

When an external bias is applied, the diffusion and drift currents do not balance each other any more. 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. 9.9, 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 (section 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. 9.9, 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.

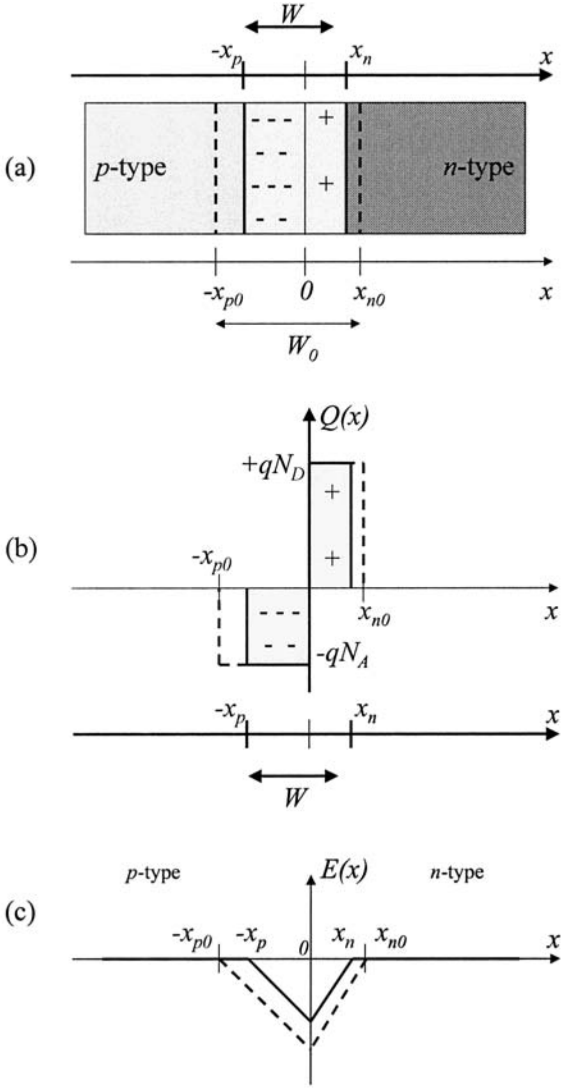
In the forward bias regime, corresponding to $V>0$, the total voltage or potential barrier across the transition region is actually reduced from V_0 to V_0-V, which has a number of consequences. First, the strength of the internal electric field associated with the lower potential barrier is reduced as well, as shown in Fig. 9.10(c). 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. 9.10(b). 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. 9.10(a). As the internal voltage is reduced from its 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. 9.10(e). This means that:

Eq. (9.29) $$E_{Vp} - E_{Vn} = E_{Cp} - E_{Cn} = q(V_0 - V)$$

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. 9.10(e), we must have:

Eq. (9.30) $$E_{Fp} - E_{Fn} = -qV$$

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.



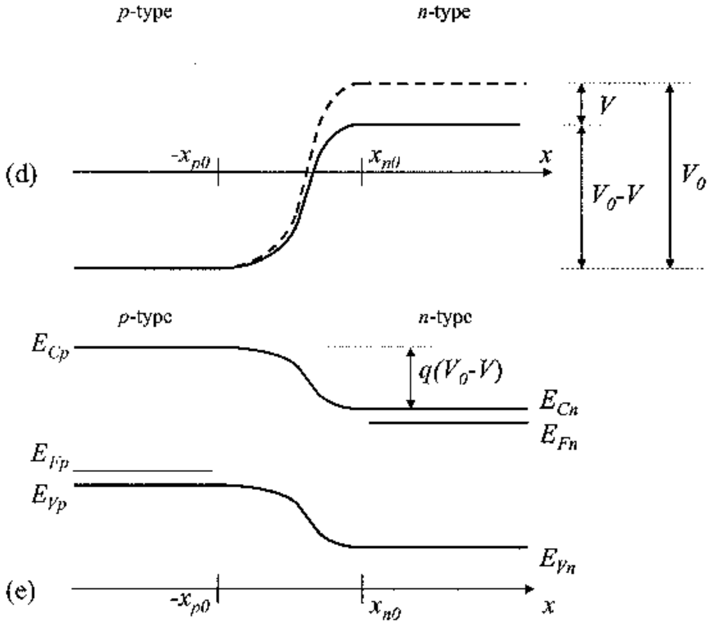


Fig. 9.10. (a) Space charge region width, (b) electrical charge density, (c) electric field strength, (d) potential profile, and (e) energy band profile of a p-n junction under forward bias (V>0). The thick dashed curves represent the equilibrium case for comparison.

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 density 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. 9.10(c), 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.

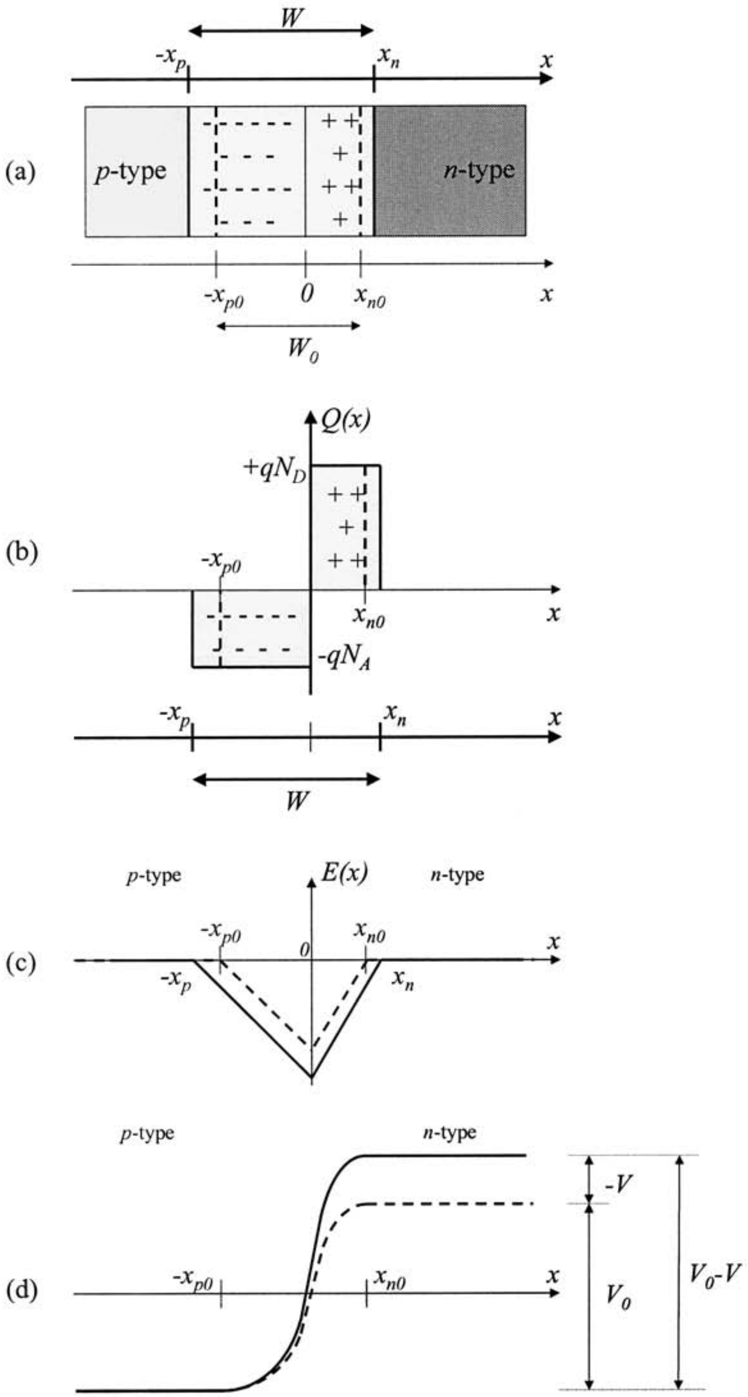
This can also be understood by examining the energy band profile. For example: when the electrons in the *n*-type region, on the right hand side of

Fig. 9.10(e) where they are more concentrated, diffuse towards 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. 9.10(e), and more electrons can thus participate in the diffusion towards 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 section 10.2 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. 9.11(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. 9.11(a)) because more electrical charges are needed to maintain this electric field, as shown in Fig. 9.11(b). 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. 9.11(e). 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).


(a)
W
$-x_p$
x_n
x
p-type
n-type
$-x_{p0}$
0
x_{n0}
x
W_0
(b)
$Q(x)$
$+qN_D$
$-x_{p0}$
x_{n0}
x
$-qN_A$
$-x_p$
x_n
x
W
(c)
p-type
$E(x)$
n-type
$-x_{p0}$
x_{n0}
0
x
$-x_p$
x_n
(d)
p-type
n-type
$-x_{p0}$
x_{n0}
x
$-V$
V_0
V_0-V

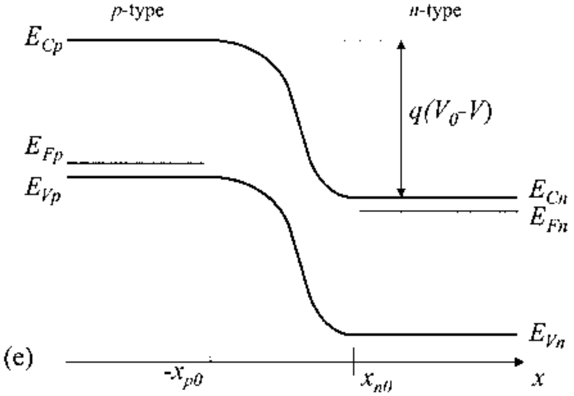


Fig. 9.11. (a) Space charge region width, (b) electrical charge density, (c) electric field strength, (d) potential profile, and (e) energy band profile of a p-n junction under reverse bias (V<0). The thick dashed curves represent the equilibrium case for comparison.

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 sub-sections, 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 (Eq. (9.29) and Eq. (9.30) respectively).

In fact, most of the relations that were derived in section 9.2 for the equilibrium case are valid when an external bias voltage V is applied, provided we make the following transformations:

Eq. (9.31)
$$\begin{cases} W_0 & \rightarrow & W \\ x_{p0} & \rightarrow & x_p \\ x_{n0} & \rightarrow & x_n \\ V_0 & \rightarrow & V_0 - V \end{cases}$$

This statement is justified by the fact that most of the expressions in section 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):

Eq. (9.32) $$W = \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A + N_D}{N_A N_D}\right)(V_0 - V)}$$

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 sub-section.

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 $N_A=2\times10^{17}$ cm^{-3} and a doping level on the n-type side of $N_D=1\times10^{17}$ 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:

$$\begin{aligned} \frac{W}{W_0} &= \sqrt{\frac{(V_0 - V)}{V_0}} \\ &= \sqrt{\frac{(1.297 - 0.3)}{1.297}} \\ &= 0.877 \end{aligned}$$

The depletion width is then:

$$\begin{aligned} W &= 0.877 W_0 = 0.877(x_{p0} + x_{n0}) \\ &= 0.877(56 + 112) \\ &= 147\, nm \end{aligned}$$

The extent of the space charge region inside the p-type and n-type regions, as shown in Fig. 9.10(a) and Fig. 9.11(a), can be obtained from Eq. (9.25):

Eq. (9.33)
$$\begin{cases} x_p = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_D}{N_A(N_A+N_D)}\right)(V_0-V)} \\ x_n = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_A}{N_D(N_A+N_D)}\right)(V_0-V)} \end{cases}$$

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):

Eq. (9.34)
$$\frac{p(-x_p)}{p(x_n)} = \frac{n(x_n)}{n(-x_p)} = \exp\left(\frac{q(V_0-V)}{k_bT}\right)$$

In addition, following our previous discussion, we realize that the majority carrier concentrations is 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:

$$\frac{p_p}{p(x_n)} = \frac{n_n}{n(-x_p)} = \exp\left(\frac{q(V_0-V)}{k_bT}\right)$$

and by using Eq. (9.21) to eliminate p_p and n_n from this latest equation:

Eq. (9.35)
$$\frac{p(x_n)}{p_n} = \frac{n(-x_p)}{n_p} = \exp\left(\frac{qV}{k_bT}\right)$$

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 sub-section.

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 $\frac{p(x_n)}{p_n}=\frac{n(-x_p)}{n_p}=\exp\left(\frac{qV}{k_bT}\right)$, 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:

$$p_n=\frac{n_i^2}{N_D}=\frac{\left(1.79\times10^6\right)^2}{1\times10^{17}}=3.20\times10^{-5}\ cm^{-3}$$

and

$$n_p=\frac{n_i^2}{N_A}=\frac{\left(1.79\times10^6\right)^2}{2\times10^{17}}=1.60\times10^{-5}\ cm^{-3}.$$

In addition, the exponential is numerically equal to:

$$\exp\left(\frac{qV}{k_bT}\right)=\exp\left(\frac{\left(1.60218\times10^{-19}\right)\times0.3}{\left(1.38066\times10^{-23}\right)\times300}\right)=1.1\times10^5.$$

Thus, we get:

$$p(x_n)=\left(3.20\times10^{-5}\right)\left(1.1\times10^5\right)$$
$$\approx3.5\,cm^{-3}$$

and

$$n(x_p)=\left(1.60\times10^{-5}\right)\left(1.1\times10^5\right)$$
$$\approx1.76\,cm^{-3}$$

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:

Eq. (9.36) $$C_{dep}=\frac{\varepsilon A}{W}$$

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.

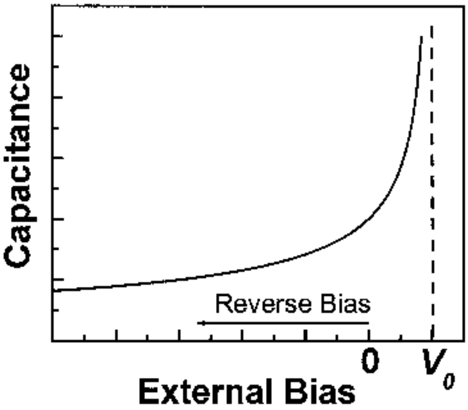


Fig. 9.12. Depletion layer capacitance as a function of bias voltage, showing the increase in capacitance with forward bias and the decrease with reverse bias.

More formally, the capacitance of the p-n junction can be derived starting from the definition of capacitance:

Eq. (9.37) $$C_{dep} = \left|\frac{dQ}{dV}\right|$$

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

Eq. (9.38) $$Q_{dep} = qAN_A x_n = qAN_D x_p$$

where x_n and x_p are given by Eq. (9.33). Substituting in Eq. (9.38) for either term gives the equation:

$$Q_{dep} = A\sqrt{2q\varepsilon \frac{N_A N_D}{(N_A + N_D)}(V_0 - V)}$$

which can then be differentiated with respect to *V* to yield:

Eq. (9.39) $$C_{dep} = \frac{A}{\sqrt{(V_0 - V)}}\sqrt{\frac{q\varepsilon}{2}\frac{N_A N_D}{(N_A + N_D)}}$$

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 $1/C_{dep}$ vs. applied voltage:

Eq. (9.40)
$$V = A^2\left[\frac{q\varepsilon(N_A N_D)}{2(N_A+N_D)}\right]\frac{1}{C_{dep}^2} - V_o$$

In the case of an abrupt one sided junction (such as a p^+n^- or a metal-semiconductor Schottky diode (see section 9.5)), this equation reduces further, and the carrier concentrations can be extracted more directly:

Eq. (9.41)
$$V = \frac{A^2 q\varepsilon}{2}N_A\frac{1}{C_{dep}^2} - V_o, \qquad (N_D >> N_A)$$
$$V = \frac{A^2 q\varepsilon}{2}N_D\frac{1}{C_{dep}^2} - V_o, \qquad (N_A >> N_D)$$

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):

Eq. (9.42)
$$\begin{cases} \Delta p_n = p_n\left(e^{\frac{qV}{k_bT}} - 1 \right) \\ \Delta n_p = n_p\left(e^{\frac{qV}{k_bT}} - 1 \right) \end{cases}$$

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

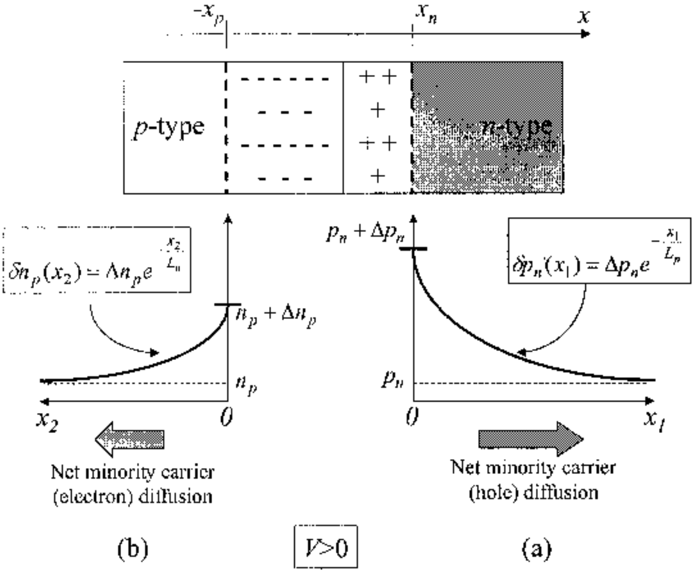

$-x_p$
x_n
x
p-type
n-type
$\delta n_p(x_2) = \Delta n_p e^{\frac{x_2}{L_n}}$
$n_p + \Delta n_p$
n_p
x_2
0
$p_n + \Delta p_n$
$\delta p_n(x_1) = \Delta p_n e^{-\frac{x_1}{L_p}}$
p_n
0
x_1
Net minority carrier (electron) diffusion
Net minority carrier (hole) diffusion
(b)
V>0
(a)

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.

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 towards the edges of the space charge region. This is illustrated in Fig. 9.14.

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 Chapter 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 Chapter 8 and the analytical expression for $\delta p_n(x_1)$, the excess hole concentration at a position x_1, is obtained from Eq. (8.55):

Eq. (9.43) $$\delta p_n(x_1) = \Delta p_n e^{-\frac{x_1}{L_p}}$$

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 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. 9.13(a) for the forward bias case and Fig. 9.14(a) for the reverse bias case.

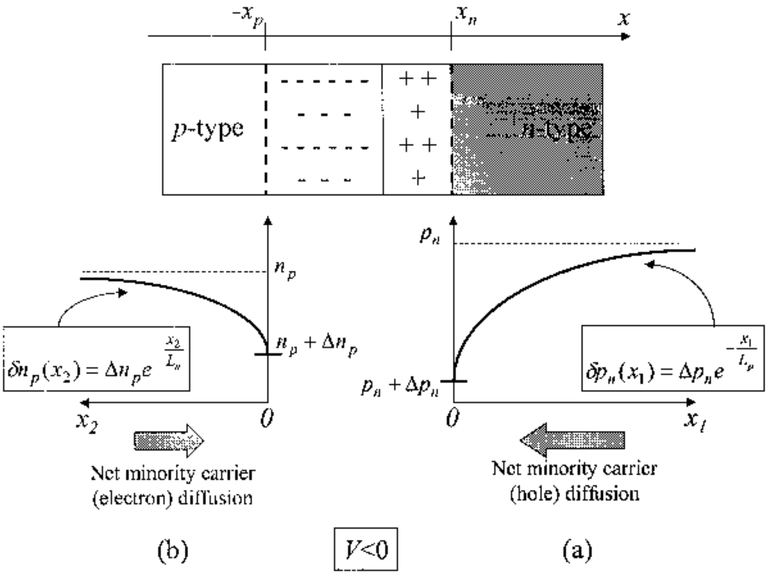

$-x_p$
x_n
x
p-type
n-type
n_p
$n_p + \Delta n_p$
p_n
$p_n + \Delta p_n$
$\delta n_p(x_2) = \Delta n_p e^{\frac{x_2}{L_n}}$
$\delta p_n(x_1) = \Delta p_n e^{-\frac{x_1}{L_p}}$
x_2
0
0
x_1
Net minority carrier (electron) diffusion
Net minority carrier (hole) diffusion
(b)
$V<0$
(a)

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.

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. 9.13(b) for the forward bias case and Fig. 9.14(b) for the reverse bias case, and it is analytically given by:

Eq. (9.44) $$\delta n_p(x_2) = \Delta n_p e^{-\frac{x_2}{L_n}}$$

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 sub-section 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, 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. 9.9.

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 Eq. (8.36) and Eq. (8.38), and are given by:

Eq. (9.45)
$$\begin{cases} J_h^{diff}(x_1) = -qD_p \dfrac{d(\delta p_n(x_1))}{dx_1} \\ J_e^{diff}(x_2) = qD_n \dfrac{d(\delta n_p(x_2))}{dx_2} \end{cases}$$

Using the expressions of the excess carrier concentrations in Eq. (9.43) and Eq. (9.44), we get:

Eq. (9.46)
$$\begin{cases} J_h^{diff}(x_1) = +q\dfrac{D_p}{L_p}\Delta p_n e^{-\frac{x_1}{L_p}} \\ J_e^{diff}(x_2) = -q\dfrac{D_n}{L_n}\Delta n_p e^{-\frac{x_2}{L_n}} \end{cases}$$

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 $x_1=x_2=0$:

Eq. (9.47)
$$\begin{cases} J_h^{diff}(0) = +q\dfrac{D_p}{L_p}\Delta p_n \\ J_e^{diff}(0) = -q\dfrac{D_n}{L_n}\Delta n_p \end{cases}$$

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:

Eq. (9.48) $$J_{total} = J_h^{diff}(0) - J_e^{diff}(0)$$

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:

Eq. (9.49) $$J_{total} = q\left(\frac{D_p}{L_p}\Delta p_n + \frac{D_n}{L_n}\Delta n_p\right)$$

and using Eq. (9.42), we finally obtain:

Eq. (9.50) $$J_{total} = q\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)\left(e^{\frac{qV}{k_bT}} - 1\right)$$

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:

Eq. (9.51) $$I_{total} = AJ_{total} = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)\left(e^{\frac{qV}{k_bT}} - 1\right)$$

By introducing a new term I_0, this can be rewritten as:

Eq. (9.52) $$I_{total} = I_0\left(e^{\frac{qV}{k_bT}} - 1\right)$$

with:

Eq. (9.53) $$I_0 = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)$$

Eq. (9.52) and Eq. (9.53) represent the diode equation for an ideal p-n junction. This function is plotted in Fig. 9.15.

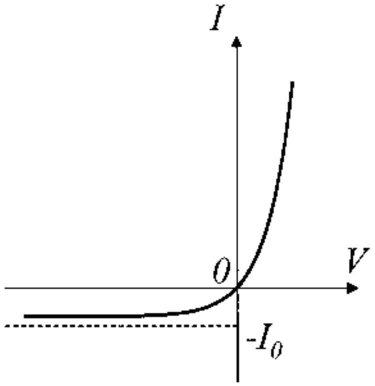


Fig. 9.15. Current-voltage characteristic for an ideal p-n junction diode. The dependence of the current on the voltage follows an exponential expression. The current is zero when the voltage is zero, without external excitation.

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 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 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.:

Eq. (9.54) $Q_p = qAL_p\Delta p_n$

where A is the area of the p-n junction. This excess charge is illustrated in Fig. 9.16(a), 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 sub-section 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 $I_p = \frac{Q_p}{\tau_p}$.

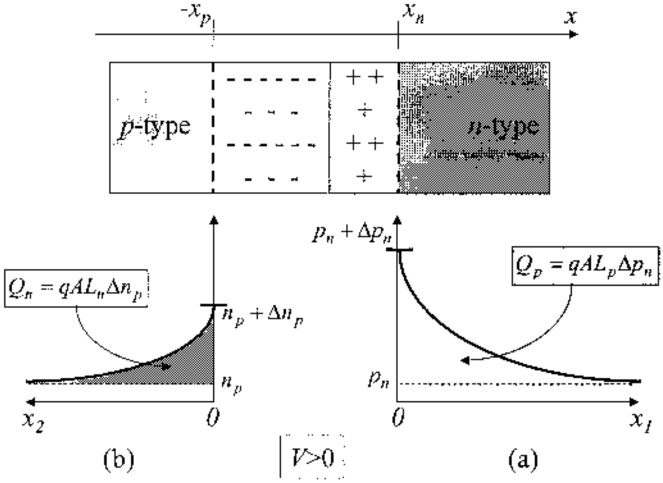


Fig. 9.16. (a) Excess positive charge in the n-type region and (b) excess negative charge in the p-type region, under a forward bias. The total excess charges are calculated by integrating the excess carrier concentrations over the volume of the regions outside the depletion region.

Similarly, the excess negative charge in the p-type region is given by:

Eq. (9.55) $Q_n = qAL_n\Delta n_p$

and is shown in Fig. 9.16(b). 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:

Eq. (9.56) $$I_{total} = qA\left(\frac{L_p}{\tau_p}\Delta p_n + \frac{L_n}{\tau_n}\Delta n_p\right)$$

Using the definition of the diffusion lengths given in Eq. (8.53) and Eq. (8.56), and using Eq. (9.42), we can transform this last expression into:

$$I_{total} = AJ_{total} = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)\left(e^{\frac{qV}{k_bT}} - 1\right)$$

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 Eq. (9.43) and Eq. (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 carriers "consume" majority carriers (e.g. electrons). There thus must be a flow of majority carriers (e.g. electrons) in the opposite direction to re-supply 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 sub-section 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.

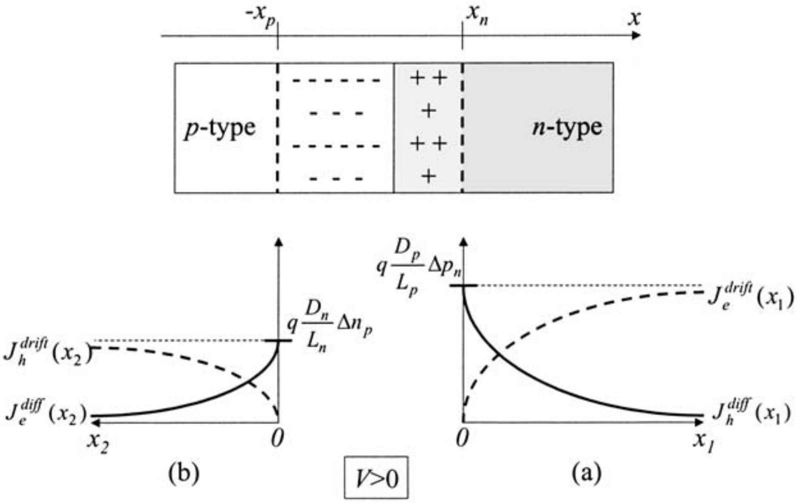


Fig. 9.17. Diffusion current of minority carriers and drift current of majority carriers in the (a) n-type region and (b) p-type region, under a forward bias. As the minority carriers diffuse further away from the edges of the depletion region, they recombine with majority carriers. The diffusion current 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.

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 $J_h^{diff}(x_1)$ and $J_e^{diff}(x_2)$ from Eq. (9.46), the drift current densities will be the difference:

Eq. (9.57)
$$\begin{cases} J_h^{drift}(x_2) = J_e^{diff}(0) - J_e^{diff}(x_2) \\ J_e^{drift}(x_1) = J_h^{diff}(0) - J_h^{diff}(x_1) \end{cases}$$

Recalling Eq. (9.46) and Eq. (9.49), we get successively:

$$\begin{cases} J_h^{drift}(x_2) = -q\dfrac{D_n}{L_n}\Delta n_p + q\dfrac{D_n}{L_n}\Delta n_p e^{-\frac{x_2}{L_n}} \\ J_e^{drift}(x_1) = q\dfrac{D_p}{L_p}\Delta p_n - q\dfrac{D_p}{L_p}\Delta p_n e^{-\frac{x_1}{L_p}} \end{cases}$$

Eq. (9.58)
$$\begin{cases} J_h^{drift}(x_2) = q\dfrac{D_n}{L_n}\Delta n_p\left(e^{-\frac{x_2}{L_n}} - 1\right) \\ J_e^{drift}(x_1) = q\dfrac{D_p}{L_p}\Delta p_n\left(1 - e^{-\frac{x_1}{L_p}}\right) \end{cases}$$

It is important to remember that the sign convention chosen for $J_h^{drift}(x_2)$ 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 sub-section 9.4.3.

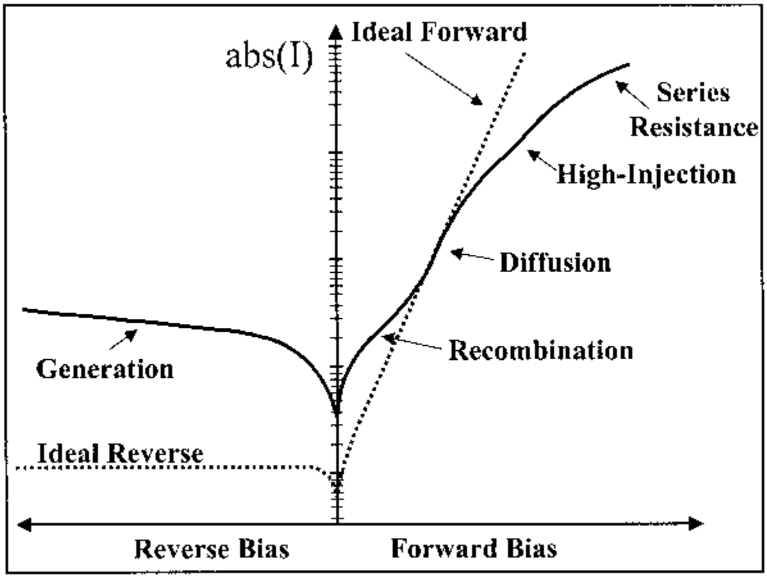


Fig. 9.18. The current-voltage characteristic for a real Si p-n junction diode (solid) do not exactly match the behavior of a Si junction diode predicted by the ideal diode model (dotted), both shown above in semi-log scale. A real Si diode shows the following deviations from the ideal (diffusion limited) case: reverse leakage current due to thermal generation and surface leakage effects, recombination in the depletion region, high-injection deviation, and series resistance effects.

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 towards 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 $G_t(T)$, expressed in cm^{-3}.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:

Eq. (9.59) $$I_{gen} = qWAG_t(T)$$

Under reverse bias the current can then be expressed as the sum of the diffusion and generation components:

Eq. (9.60) $$I_{rev} = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right) + qWAG_t(T).$$

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 (W) 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, and will be discussed in more detail in Chapter 20.

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 though 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):

Eq. (9.61) $$I_R \approx I_{R0} \exp\left(\frac{qV}{2k_bT}\right)$$

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:

Eq. (9.62) $$I = I_0 \exp\left(\frac{qV}{k_bT}\right) + I_{R0} \exp\left(\frac{qV}{2k_bT}\right)$$

In working with real diodes this equation is generally represented in an empirical form by introducing a new factor *n* called the ideality factor:

Eq. (9.63) $$I \approx I_0 \exp\left(\frac{qV}{nk_bT}\right)$$

In this combined equation, the ideality factor *n* tends towards 2 when recombination current dominates, and tends towards 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 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. 9.18.

Under higher reverse bias, the contact potentials and the potential drop across the bulk regions of the semiconductor ceases 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:

$$\text{Eq. (9.64)} \quad I \approx I_0 \exp\left(\frac{q(V - IR_S)}{nk_bT}\right)$$

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 a 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).

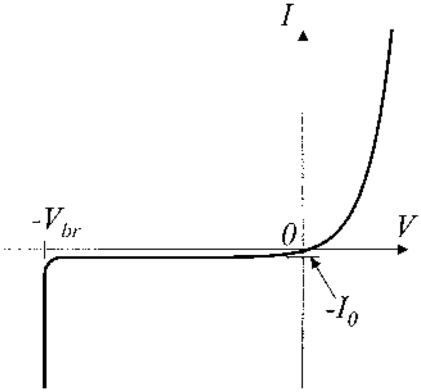


Fig. 9.19. Current-voltage characteristic for an ideal p-n junction diode showing a reverse breakdown. When the voltage across the p-n junction is equal to the reverse breakdown voltage, the current increases dramatically. If it is not limited, this current can damage the diode through heating.

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:

Eq. (9.65) $$M = \frac{1}{1-\left(\frac{V_r}{V_{br}}\right)^n}$$

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.

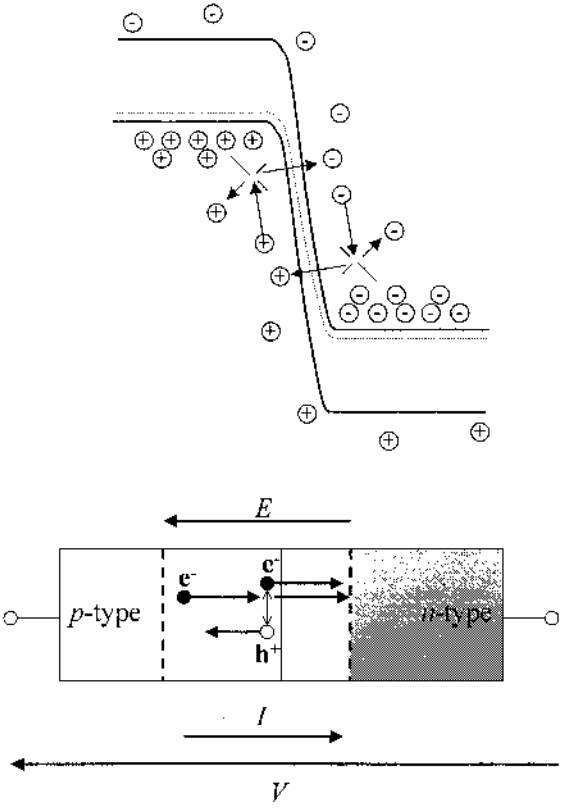


Fig. 9.20. Impact ionization process: under strong reverse bias, electrons and holes are injected into the depletion region; when they gain enough kinetic energy they impact on the semiconductor lattice to create electron-hole pairs. These newly created carriers can then lead to the same impact ionization process if they can gain enough kinetic energy within the space charge region.

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.

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_{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: $M = \dfrac{1}{1-\left(\dfrac{V_r}{V_{br}}\right)^n}$.

For the two reverse biases mentioned, we get the ratio of the multiplication factor:

$$\frac{M_1}{M_2} = \frac{1-\left(\dfrac{V_2}{V_{br}}\right)^n}{1-\left(\dfrac{V_1}{V_{br}}\right)^n}$$

$$= \frac{1-\left(\dfrac{13.990}{14}\right)^6}{1-\left(\dfrac{13.995}{14}\right)^6}$$

$$= 2$$

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:

Eq. (9.66) $$T_Z = \exp\{-\frac{4\sqrt{2m^*}}{3qV\hbar} E_g^{3/2}\}$$

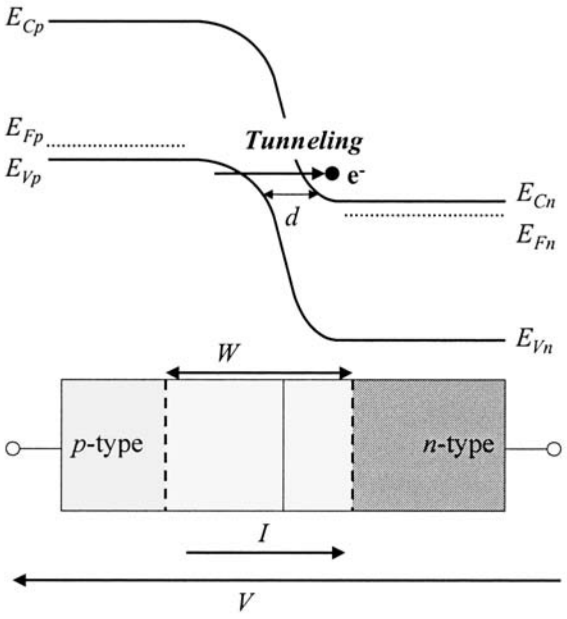


Fig. 9.21. Zener breakdown mechanism involving electrons tunneling from the valence band of the p-type side to the conduction band of the n-type side.

9.5. Metal-semiconductor junctions

As we have already mentioned in sub-section 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. 4.11 in sub-section 4.2.7, 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 Chapter 3, 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 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.

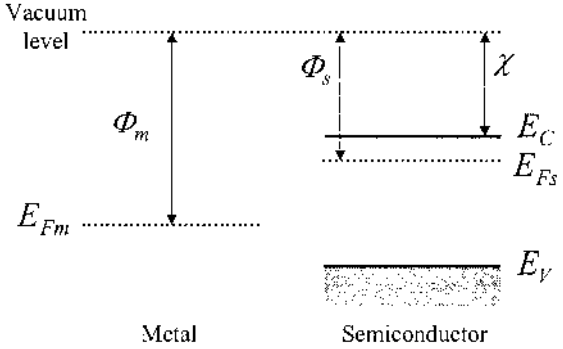


Fig. 9.22. Fermi energies, work functions in a metal and a semiconductor, when considered isolated from each other. The vacuum level is the same for both materials, but the Fermi energies are generally different.

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 Φ_m-Φ_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.

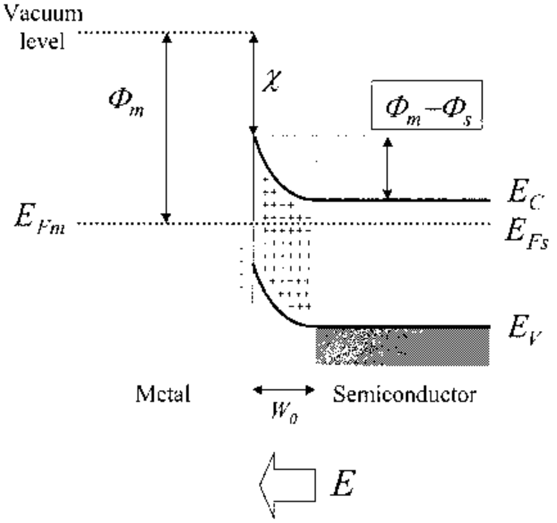


Fig. 9.23. Energy levels, accumulated charge carriers and built-in electric field in a metal-semiconductor junction. When the metal and the semiconductor are brought into contact, at equilibrium, the energy band profile of the semiconductor near the junction is modified so that the Fermi energies become equal in both materials.

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 external bias is applied. This is shown in Fig. 9.24(a). 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. 9.25(a). 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 Fig. 9.24 and Fig. 9.25, and summarized in Table 9.1.

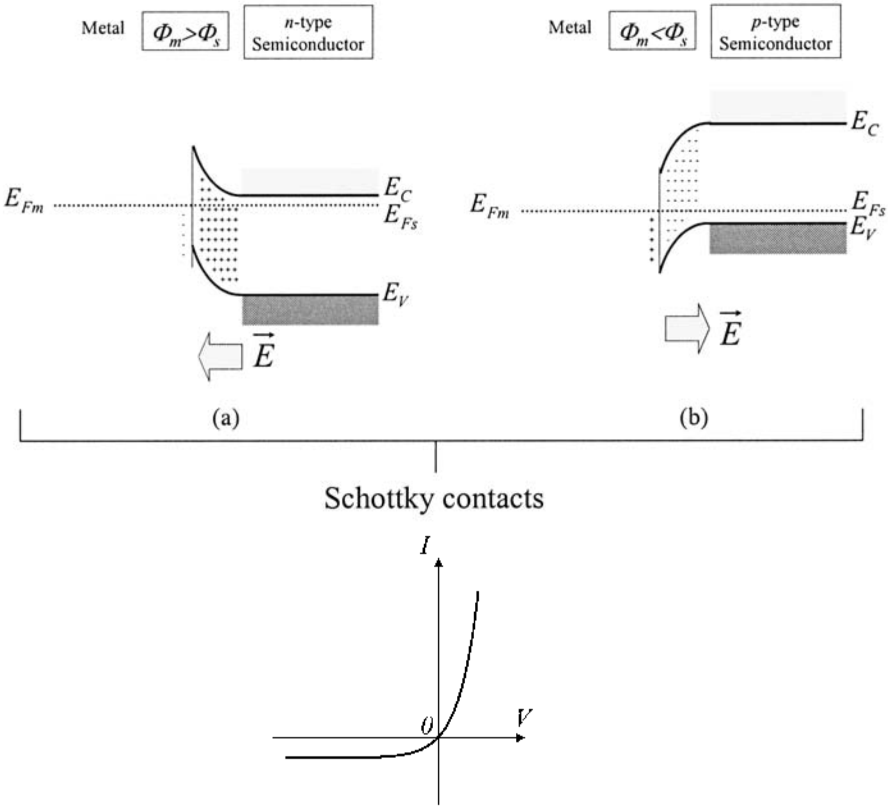


Fig. 9.24. These two of the four possible metal-semiconductor junction configurations lead to a Schottky contact: (a) $\Phi_m > \Phi_s$ and n-type, (b) $\Phi_m < \Phi_s$ and p-type. A Schottky contact is obtained in each case because the majority carriers in the semiconductor experience a potential barrier 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.

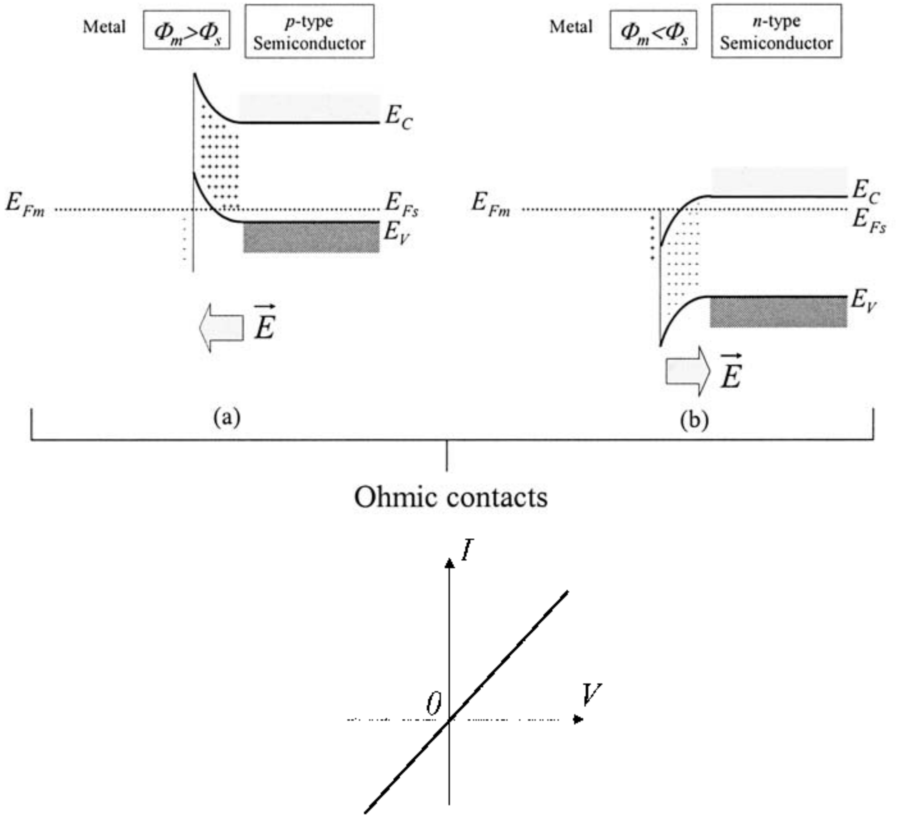


Fig. 9.25. These two of the four possible metal-semiconductor junction configurations lead to an ohmic contact: (a) $\Phi_m > \Phi_s$ and p-type, (b) $\Phi_m < \Phi_s$ and n-type. Unlike the configurations shown in Fig. 9.24, the energy band profiles here are such that the majority carriers in the semiconductor 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.

	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

Table 9.1. Four possible metal-semiconductor junction configurations and the resulting contact types.

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.

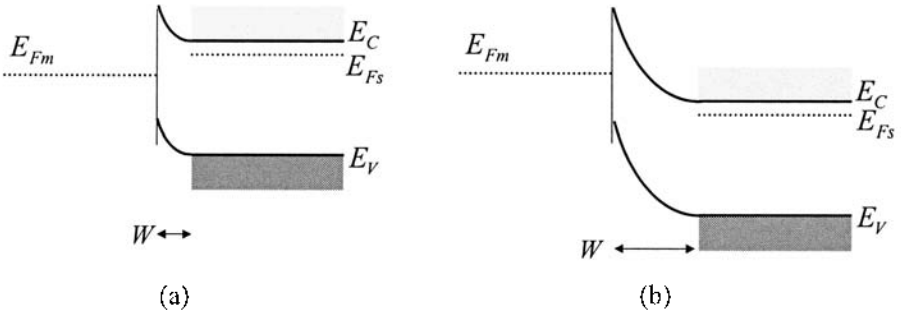


Fig. 9.26. Band alignment in a Schottky metal-(n-type) semiconductor contact under (a) forward bias where the potential barrier is reduced, and under (b) reverse bias where the potential barrier is increased, thus reducing the tunneling of carriers.

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 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:

Eq. (9.67) $$I = I_0\left(e^{\frac{qV}{k_bT}} - 1\right)$$

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:

Eq. (9.68) $$I_0 = AB_eT^2e^{\left(-\frac{(\Phi_m-\chi)}{k_bT}\right)}$$

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 A.10), as well as impurity and interface states. In this case, the current-voltage relation is given by:

Eq. (9.69) $$I = I_0\left(e^{\frac{qV}{nk_bT}} - 1\right)$$

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.

Further reading

Ashcroft, N.W. and Mermin, N.D., *Solid State Physics*, Holt, Rinehart and Winston, New York, 1976.

Neudeck, G.W., *The PN Junction Diode*, Addison-Wesley, Reading, MA, 1989.

Pierret, R.F., *Advanced Semiconductor Fundamentals*, Addison-Wesley, Reading, MA, 1989.

Sapoval, B. and Hermann, C., *Physics of Semiconductors*, Springer-Verlag, New York, 1995.

Streetman, B.G., *Solid State Electronic Devices*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

Wang, S., *Fundamentals of Semiconductor Theory and Device Physics*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

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=5×10^{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).
   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.

4. 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.

5. 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 τ_n=τ_p=1 μs, and diffusion coefficients equal to D_n=35 $\text{cm}^2.\text{s}^{-1}$ and D_p=12.5 $\text{cm}^2.\text{s}^{-1}$.

6. Consider a Si p-n step junction with N_A=10^{17} cm^{-3} and N_D=10^{16} cm^{-3}, with recombination lifetimes τ_p=0.1 μs and τ_n=0.01 μs, and carrier mobilities μ_h=450 cm^2/Vs and μ_e=800 cm^2/Vs at 300 K.
   Determine the total reverse saturation current density, the reverse saturation current density due to holes and that due to electrons.

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.

Assume a reverse bias equal to $-V_0/2$ is applied. Calculate the minority carrier currents at the edges of the space charge region.

7. A Si p-n junction is doped with an acceptor concentration $N_A=5\times10^{18}$ cm^{-3}, a donor concentration $N_D=5\times10^{15}$ 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 $N_D=5\times10^{17}$ cm^{-3}.

8. Consider an ideal metal-semiconductor junction between *p*-type silicon and polycrystalline aluminum. The Si is doped with $N_A=5\times10^{16}$ 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 barriers height ϕ_B.

9. 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.

10. Optical Properties of Semiconductors

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 Chapter that band structure depends on the dimensionality of the system, and we have already seen in Chapters 7 and 8 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 Chapter 11.

Maxwell showed many year 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.

Quantum theorists, as we have seen in Chapter 3, 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. As for phonons, the quantum of lattice vibrations, it turns out in practice that for most purposes the classical description of light is quite adequate, and we shall therefore continue our study of optical properties in terms of 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.

We now study how a given medium modifies the electromagnetic wave and in effect determines the way this wave propagates. Later in Chapters 19 and 20 we shall see how the frequency and power of electromagnetic waves can be measured to great accuracy using photodetectors.

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. The Maxwell's equations can be written in a form which from the very beginning distinguishes a conducting medium from a non conducting medium, by writing:

Eq. (10.1) $$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

Eq. (10.2) $$\nabla \times \vec{H} = \sigma(\omega)\vec{E} + \frac{\partial \vec{D}}{\partial t}$$

Eq. (10.3) $$\nabla \times \vec{D} = \rho$$

Eq. (10.4) $$\nabla \cdot \vec{B} = 0$$

where $\sigma(\omega)$ is the complex frequency dependent conductivity of the medium with a density of ρ mobile charges, $\vec{E}$, $\vec{D}$, and $\vec{H}$, 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 $\vec{B} = \mu_0 \vec{H}$, $\mu = \mu_0$, 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 Eq. (10.1) and Eq. (10.2) by taking the "curl" (or "rot") of Eq. (10.1) to give the wave equation for an EM wave as:

Eq. (10.5) $$\nabla^2 \vec{E} = \frac{1}{c^2}\left(\varepsilon_b \frac{\partial^2 \vec{E}}{\partial t^2} + \frac{\sigma}{\varepsilon_0}\frac{\partial \vec{E}}{\partial t} \right)$$

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:

Eq. (10.6) $$\vec{E} = \vec{E}_0 \exp\left[i\left(\vec{k} \cdot \vec{r} - \omega t \right) \right]$$

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:

Eq. (10.7) $$k = \frac{\omega}{c}(\varepsilon_b + \frac{i\sigma}{\varepsilon_0 \omega})^{1/2}$$

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 $\overline{N}$:

Eq. (10.8) $$\overline{N} = (\varepsilon_b + \frac{i\sigma}{\varepsilon_0 \omega})^{1/2}$$

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:

Eq. (10.9) $$\overline{N}^2 = \varepsilon_r + i(\varepsilon_i + \frac{\sigma}{\varepsilon_0 \omega})$$

The free electron permittivity is now by definition:

Eq. (10.10) $$\varepsilon_f(\omega) = i\frac{\sigma(\omega)}{\varepsilon_0 \omega}$$

We can now rewrite the complex refractive index and complex wavevector as:

Eq. (10.11) $$\overline{N} = \overline{n} + i\kappa$$

Eq. (10.12) $$k = \frac{\overline{n}\omega}{c} + \frac{i\kappa\omega}{c} = \overline{N}k_0$$

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:

Eq. (10.13) $$\vec{E} = \vec{E}_0 \exp\{i\omega(\frac{\overline{n}z}{c} - t)\}\exp(-\frac{\omega\kappa z}{c})$$

For $\vec{E}_0 = E_0^x \vec{x}$ the corresponding H_0^y is given by $H_0^y = N\sqrt{\frac{\varepsilon_0}{\mu_0}}E_0^x$

where we also have from Eq. (10.9) and $\sigma = \sigma_r + i\sigma_i$

Eq. (10.14) $$\overline{n}^2 - \kappa^2 = \varepsilon_r - \frac{\sigma_i}{\omega\varepsilon_0}$$
$$2\overline{n}\kappa = \varepsilon_i + \frac{\sigma_r}{\varepsilon_0\omega}$$

The medium has modified the electromagnetic wave or photon, in two ways. It has changed the velocity of propagation from c to $c/\overline{n}$, 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 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:

Eq. (10.15)
$$|E|^2 = |E_0|e^{-\alpha z}$$
$$\alpha = 2\kappa\frac{\omega}{c}$$

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 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:

$$\vec{S} = 1/2\,\mathrm{Re}(\vec{E}\times\vec{H}^*) = \frac{\overline{n}c}{2}\varepsilon_0(E_0^x)^2 e^{-\alpha z}\vec{z}$$

and is discussed in more detail in Chapter 18.

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.

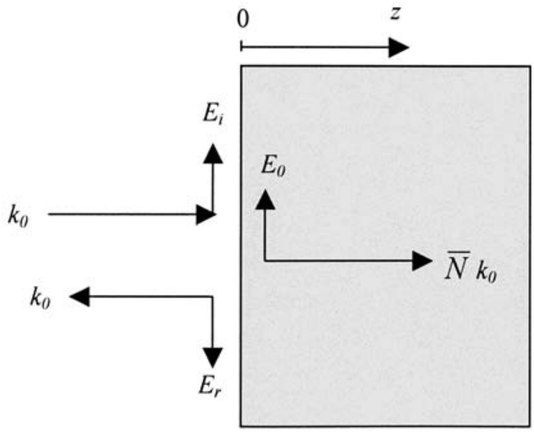


Fig. 10.1. The reflection and transmission process expressed in terms of a diagram.

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:

Eq. (10.16) $$E_x(t,z) = E_0 \exp\left(i\omega(\frac{\overline{N}z}{c} - t) \right)$$

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:

Eq. (10.17) $$E_x(t,z) = E_i \exp\left[i\omega\left(\frac{z}{c} - t\right)\right] + E_r \exp\left[-i\omega\left(\frac{z}{c} + t\right)\right]$$

The continuity requirement of the electric field at the boundary *z=0* gives us:

Eq. (10.18) $$E_0 = E_i + E_r$$

Knowing the electric field allows us to deduce the magnetic field using Maxwell's equation so that, for *z*>0:

Eq. (10.19) $$H_y = \frac{-1}{\omega\mu_0}(\overline{N}k_0)E_x$$

and then use the continuity condition for H at the boundary, which gives

Eq. (10.20) $$\overline{N}E_0 = E_i - E_r$$

Note the magnetic field at z=0 depends on the direction of propagation. From this pair of equations we can deduce the relation:

Eq. (10.21) $$\frac{E_r}{E_i} = \frac{1-\overline{N}}{1+\overline{N}}$$

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. (10.21) giving:

Eq. (10.22) $$R = \left|\frac{1-\overline{N}}{1+\overline{N}}\right|^2 = \frac{(\overline{n}-1)^2+\kappa^2}{(\overline{n}+1)^2+\kappa^2}$$

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. (10.22) 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 2 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 $e^{\pm idk_0N}$. 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 4 equations for 4 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:

Eq. (10.23)
$$T = \frac{(1-|r_{01}|^2)^2 e^{-\alpha d}}{(1-|r_{01}|^2 e^{-\alpha d})^2}$$
$$|r_{01}|^2 = \left|\frac{1-\overline{N}}{1+\overline{N}}\right|^2$$

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. (10.22) 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:

Eq. (10.24) $$R = \frac{|r_{01}|^2(1-e^{-\alpha d})^2}{(1-|r_{01}|^2 e^{-\alpha d})^2}$$

From Eq. (10.23) and Eq. (10.24) one can now deduce the absorbance $A = 1 - R - T$. In the limit of a very thick slab, $e^{-\alpha d} \rightarrow 0$ 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 Chapter 8 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.

The 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 (Chapter 8) can be written as:

Eq. (10.25) $$m^* \frac{d^2 x}{dt^2} + m^* \frac{dx}{dt} \frac{1}{\tau} = -qE(t)$$

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. (10.25) and differentiate in time. The condition that this be a solution to Eq. (10.25) is that:

Eq. (10.26) $$-m^* \omega^2 x_0 - m^* \frac{i\omega}{\tau} x_0 = -qE_0$$

which immediately allows us to extract the amplitude x_0 as:

Eq. (10.27) $$x_0 = \frac{q\tau}{m^* i\omega} \left(\frac{1}{1 - i\omega\tau} \right) E_0$$

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:

Eq. (10.28) $$P = -n_c q x(t) = -\frac{n_c q^2 \tau}{m^* i\omega} \left(\frac{1}{1 - i\omega\tau} \right) E_0 e^{-i\omega t}$$

from which we can now also deduce the polarizability or optical susceptibility as the ratio:

Eq. (10.29) $$\alpha_p(\omega) = \frac{P_c(t)}{E_0 e^{-i\omega t}}$$

and write:

Eq. (10.30) $$\alpha_p(\omega) = -\frac{n_c q^2 \tau}{m^* i\omega}\left(\frac{1}{1 - i\omega\tau}\right)$$

And for the complex conductivity we have, from the current:

Eq. (10.31) $$-\frac{n_c q \dfrac{dx}{dt}}{E_0 e^{-i\omega t}} = \sigma(\omega) = \frac{n_c q^2 \tau}{m^*}\left(\frac{1}{1 - i\omega\tau}\right)$$

From the polarizability, we can deduce the relative permittivity produced by nearly free electrons, in the usual electrodynamic way ($\varepsilon_f = (1 + \frac{\alpha_p}{\varepsilon_0})$):

Eq. (10.32) $$\varepsilon_f = 1 - \frac{n_c q^2 \tau}{\varepsilon_0 m^* i\omega}\left(\frac{1}{1 - i\omega\tau}\right)$$

It is convenient and useful to rewrite the relative permittivity in a form which involves the plasma frequency ω_p and rewrite it as:

Eq. (10.33) $$\varepsilon_f = 1 - \frac{\omega_p^2}{\omega^2}\left(\frac{\omega\tau(\omega\tau - i)}{1 + (\omega\tau)^2}\right)$$

Eq. (10.34) $$\omega_p^2 = \frac{n_c q^2}{m^* \varepsilon_0}$$

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 $P = n_c qx$, 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 RHS of Eq. (10.33) dominates and $\varepsilon_f(\omega)$ reduces to:

Eq. (10.35) $$\varepsilon_f(\omega) \sim i\frac{n_c q^2 \tau}{\varepsilon_0 m^* \omega}$$

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:

Eq. (10.36) $$n(\omega) = \{\frac{n_c q^2 \tau}{2\varepsilon_0 m^* \omega}\}^{1/2}$$

and which via Eq. (10.22) gives rise to a high reflection coefficient for small frequencies.

Secondly, in the limit that $\omega\tau >> 1$, the relative permittivity is dominated by the real part and reduces to the form:

Eq. (10.37) $$\varepsilon_f(\omega) \sim \{1 - \frac{\omega_p^2}{\omega^2}\}$$

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. (10.22) 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].

When the frequency is above the plasma frequency, the permittivity is real and $\varepsilon<1$, it vanishes at the plasma frequency. The refractive index in this limit becomes:

Eq. (10.38) $$\overline{n}(\omega)=\sqrt{(1-\frac{\omega_p^2}{\omega^2})}$$

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. (10.32) into Eq. (10.22). The result is shown in Fig. 10.2.

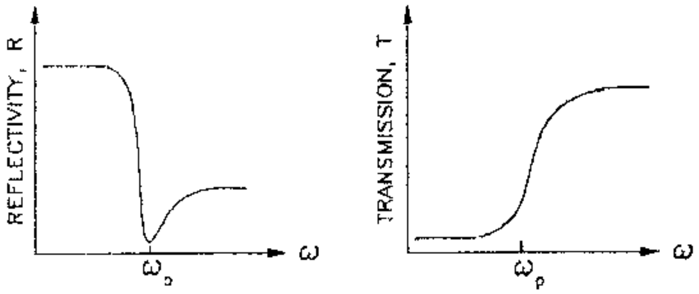


Fig. 10.2. The reflectivity and transmissivity of an electron gas (thin film). [Introduction to Semiconductor Optics, 1993, p. 62, Peyghambarian, N., Koch, S.W., and Mysyrowicz, A., Fig. 3.1. Reprinted with permission.]

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... (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:

Eq. (10.39)
$$\begin{aligned} n_c &\rightarrow \frac{1}{2} g(E_f) m^* \left| v_f \right|^2 \\ \sigma(0) &= g(E_f) E_F q^2 \tau / m^* \end{aligned}$$

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 RHS of Eq. (10.39) 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 it's energy dependence and it's 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 a an additional new term in the total energy or Hamiltonian of the system:

Eq. (10.40) $V(t) = -q\vec{E}.\vec{r}$

The next step is to solve the time dependent Schrödinger equation Chapter 3 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 Chapter 3 namely as the set:

Eq. (10.41) $\Psi_n(\vec{r},t) = \Phi_n(\vec{r})e^{-iE_n t/\hbar}$

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, 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. (10.40). We emphasize that the principle of superposition is rigorously true, and part of the principles of quantum mechanics we discussed in Chapter 3. So we can write for *t*>0:

Eq. (10.42) $\Psi(\vec{r},t)=\Phi_g e^{-iE_g t/\hbar}+\sum_{n\neq g} c_n(t)\Phi_n e^{-iE_n t/\hbar}$

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. (10.42) into the time dependent Schrödinger equation (see Eq. (3.4)). On one side we take the derivative with respect to time to obtain:

Eq. (10.43)

$$i\hbar\frac{\partial\Psi}{\partial t}=E_g\Phi_g e^{\ iE_g t/\hbar}+\sum_{n\neq g}E_n c_n(t)\Phi_n e^{-iE_n t/\hbar}+\sum_{n\neq g}i\hbar\frac{\partial c_n}{\partial t}\Phi_n e^{-iE_n t/\hbar}$$

on the other side of the Schrödinger equation we have

Eq. (10.44)

$$\{H_0+V(t)\}\Psi(\vec{r},t)=E_g\Phi_g e^{-iE_g t/\hbar}+\sum_{n\neq g}E_n c_n(t)\Phi_n e^{-iE_n t/\hbar}-q\vec{r}\cdot\vec{E}_0(e^{i\omega t}+e^{-i\omega t})\Psi(\vec{r},t)$$

We now equate Eq. (10.43) and Eq. (10.44), and cancel the common terms. This leaves the last terms of the RHS of Eq. (10.43) and Eq. (10.44) as equal to each other. Now we multiply the new equation on both sides with $\Phi_j^* e^{iE_j t/\hbar}$ 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. (3.8)). 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 RHS of Eq. (10.44) which now gives:

Eq. (10.45)

$$i\hbar\frac{\partial c_j}{\partial t}=-\int d\vec{r}\Phi*_j(\vec{r})q\vec{r}\vec{E}_0(e^{i\omega t}+e^{-i\omega t})e^{i(E_j-E_g)t/\hbar)}\Phi_g(\vec{r})$$

this can be integrated to give:

Eq. (10.46)

$$c_j(t)=-q\vec{E}_0\cdot\vec{r}_{jg}\left[\frac{1-e^{i(\hbar\omega+E_j-E_g)t/\hbar}}{\hbar\omega+(E_j-E_g)}-\frac{1-e^{i(-\hbar\omega+E_j-E_g)t/\hbar}}{\hbar\omega-(E_j-E_g)}\right]$$

where the position matrix element is:

Eq. (10.47) $$\vec{r}_{jg} = \int d\vec{r}\Phi^*{}_j(\vec{r})\vec{r}\Phi_g(\vec{r})$$

For simplicity we assume that the wave is polarized in the *x*-direction so the first factor reduces to $qE_0^x x_{jg}$. Eq. (10.47) 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. (10.47), in this case the displacement, $\vec{r}_{\alpha\beta}$ is sometimes also written in the Dirac notation $\langle\alpha|\vec{r}|\beta\rangle$.

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:

Eq. (10.48) $$P_x(t) = -q\int d\vec{r}\Psi^*(\vec{r},t)x\Psi(\vec{r},t)$$

Substitute the solution from the wave function and keep only the linear terms in the coefficients immediately gives us:

Eq. (10.49) $$P_x(t) = -\sum_j q(x_{gj}c_j(t)e^{-i\omega_j t} + x_{jg}c_j{}^*(t)e^{i\omega_j t})$$

Eq. (10.50) $$P_x(t) = \sum_j q^2\left|x_{gj}\right|^2\left(\frac{1}{E_{j0}-\hbar\omega}+\frac{1}{E_{j0}+\hbar\omega}\right)E_0^x(e^{i\omega t}+e^{-i\omega t})$$

From the dipole moment induced by the field we can now deduce the polarizability in the usual way:

Eq. (10.51) $$\alpha_p(\omega) = \sum_j q^2\left|x_{gj}\right|^2\frac{2E_{j0}}{E_{j0}^2-(\hbar\omega)^2}$$

and by introducing the oscillator strength F_j:

Eq. (10.52) $F_j = \frac{2m_0}{\hbar^2} E_{jg} \left| x_{gj} \right|^2$

We can rewrite the ground state polarizability in an elegant form:

Eq. (10.53) $\alpha_p(\omega) = \frac{q^2}{m_0} \sum_j \frac{F_j}{\omega_{jg}^2 - \omega^2}$

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:

Eq. (10.54) $\sum_j F_j = 1$

This sum rule is important. It is a check of consistency and follows from two quantum mechanical identities. The momentum position commutation relation:

Eq. (10.55) $xp_x - p_x x = i\hbar$

and taking the expectation value of this equation and expanding over a complete set of intermediate states:

Eq. (10.56) $i\hbar = \sum_l (x_{il} p_{li,x} - p_{il,x} x_{li})$

and using an identity from Heisenberg's equation motion which reads:

Eq. (10.57) $p_{ij,x} = x_{ij}(E_j - E_i) m_0 / i\hbar$

Substituting Eq. (10.57) into Eq. (10.56) gives the sum rule. Now we know the bound electron polarizability, we can compute the relative permittivity by considering the polarizability of N_b such atoms or molecules per unit volume.

Eq. (10.58) $\varepsilon(\omega) = 1 + \frac{N_b q^2}{\varepsilon_0 m_0} \sum_j \frac{F_j}{\omega_j^2 - \omega^2}$

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

Eq. (10.59) $$\varepsilon(0) = 1 + \sum_j \frac{F_j \omega_p^2}{\omega_j^2}$$

and in the high frequency limit when the light energy exceeds all bound to bound transitions, we recover the corresponding Drude result:

Eq. (10.60) $$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2}$$

which also implies that close to the plasma frequency, the permittivity can be negative, and the refractive index purely imaginary implying from Eq. (10.22) 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. (10.58), 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. (10.46) 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. (10.46) 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:

Eq. (10.61) $$\left|c_j(t)\right|^2 \sim \left|\frac{q x_{gj}}{\hbar} E_0^x\right|^2 \frac{\sin^2(\omega_j - \omega)t/2}{(\omega_j - \omega)^2}$$

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 divide by time t to study $W_{gj} = \left|c_j(t)\right|^2 / t$. Dividing the RHS of

Eq. (10.61) by *t*, and letting time go to infinity gives us a function which we recognize to be the well known Dirac delta function:

Eq. (10.62)

$$t \to \infty \Rightarrow \frac{\sin^2[(\hbar\omega_j - \hbar\omega)t/2\hbar]}{t\ \hbar^2(\omega_j - \omega)^2/4} = \frac{2\pi}{\hbar}\delta(\hbar\omega_j - \hbar\omega)$$

the Dirac delta function $\delta(x)$ has the property that:

Eq. (10.63) $$\int_{-\infty}^{\infty} dx\delta(x) = 1$$

And also as the imaginary part of the fraction:

Eq. (10.64) $$\mathrm{Im}\left(\frac{1}{x - i\eta}\right) = \pi\delta(x)$$

with infinitesimal η. So basically Eq. (10.61) 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, 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 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:

Eq. (10.65) $$W_{gj} = \frac{2\pi}{\hbar}\left|\int d\vec{r}\Phi_j{}^* V(\vec{r})\Phi_g\right|^2 \delta(\hbar\omega - E_j + E_g)$$

Now we can understand the meaning of the resonances in the permittivity expression Eq. (10.58). 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. (10.46) so that the relative permittivity becomes the complex function (T=0 K):

$$\text{Eq. (10.66)} \quad \varepsilon_b(\omega) = 1 + \frac{N_b q^2}{\varepsilon_0 m_0} \sum_j \frac{F_j}{\omega_j^2 - \omega^2 - i\omega\Gamma_j}$$

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:

$$\text{Eq. (10.67)} \quad \varepsilon(\omega) = 1 + \frac{N_b q^2}{\varepsilon_0 m_0} \sum_j \frac{F_j}{\omega_j^2 - \omega^2 - i\omega\Gamma_j} - \frac{n_c q^2}{\varepsilon_o m^*} \left(\frac{\omega\tau - i}{\omega^2\tau^2 + 1} \right)$$

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):

$$\text{Eq. (10.68)} \quad \varepsilon_b(\omega) = 1 + \frac{N_b q^2}{\hbar^2 \varepsilon_0} \sum_{i \neq j} \frac{\hbar |x_{ij}|^2 (f_i - f_j)(\omega_j - \omega_i)}{(\omega_j - \omega_i)^2 - \omega^2 - i\omega\Gamma_{ij}}$$

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. 4.17. The indirect bandgap systems will be discussed later on in this Chapter. In direct bandgap materials, or for sufficiently high photon energy, Eq. (10.66) 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 empty so that we have (T=0 K):

Eq. (10.69) $$\varepsilon_s(\omega) \sim 1 + \frac{q^2}{\varepsilon_0 m_0} \sum_{\vec{k}} F_{\vec{k}} \frac{1}{(\omega_{\vec{k},c} - \omega_{\vec{k},v})^2 - \omega^2}$$

where the Bloch sum over the occupied states is normalized by the volume and defined as:

Eq. (10.70) $$\sum_{k_v} = N_b$$

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:

Eq. (10.71) $$\frac{q^2}{m_0 \varepsilon_0} \sum_{k} F_{\vec{k}} \sim \frac{N_b q^2}{\varepsilon_0 m_0} = (\omega_p^b)^2$$

Eq. (10.72) $$\varepsilon_s(0) \sim 1 + \left(\frac{\hbar \omega_p^b}{E_g} \right)^2$$

Where ω_p^b is the effective bound electron plasma frequency and can be obtained by comparison with 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 ~ 3×10^{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:

Eq. (10.73)

$$\varepsilon_s(\omega) \sim 1 + \frac{q^2}{\hbar\varepsilon_0}\sum_{\vec{k}} | \; x_{\vec{k}c,\vec{k}v} \; |^2 \frac{(\omega - \omega_{\vec{k}c} + \omega_{\vec{k}v}) - i\gamma}{(\omega - \omega_{\vec{k},c} + \omega_{\vec{k},v})^2 + \gamma^2}[f(E_{\vec{k}v}) - f(E_{\vec{k}c})]$$

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]:

Eq. (10.74) $$\left|x_{kvkc}\right|^2 = \left|\int d\vec{r}\Psi^*_c(\vec{k})x\Psi_v(\vec{k})\right|^2 = \frac{1}{3}\frac{\hbar^2}{E_g^2}\frac{E_P}{m_0}$$

where E_P is the Kane parameter and a number which varies only slightly between 20 and 25 eV in most semiconductors (see also Appendix A.8). This powerful last equation now allows us to compute the permittivity for most situations of interest in semiconductor physics. All we need for Eq. (10.73) 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 $\omega\tau >> 1$ 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

Eq. (10.75) $$\bar{n}(\omega) = \left((1+\varepsilon_b)\left(1 - \frac{\omega_p^2}{\omega^2(1+\varepsilon_b)}\right)\right)^{1/2}$$

or as is the notation of some other authors one can also replace:

Eq. (10.76) $\varepsilon(\infty) = 1 + \varepsilon_b$

One can think of Eq. (10.76) as a renormalization of the plasma frequency of the electrons to $\omega_p^2 \to \omega_p^2 /(1+\varepsilon_b)$. 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 A.4. For example GaAs: $\varepsilon_b = 13.1$, Si: $\varepsilon_b = 11.9$, C: $\varepsilon_b = 5.7$. From Eq. (10.72), 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. (10.73).

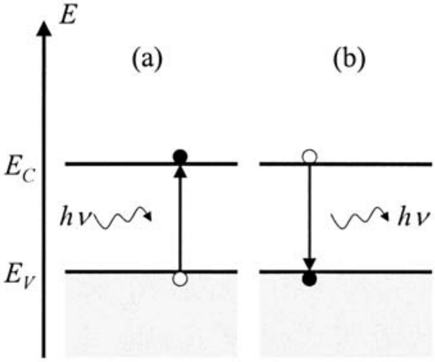


Fig. 10.3. Electronic transition, (a) from the valence band to the conduction band resulting from the absorption of a photon, (b) from the conduction band to the valence band resulting into the emission of a photon.

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 the optical absorption process: (i) the total energy (electron+hole+photon) must be conserved, (ii) the total momentum or wavevector must also be conserved. Assuming that in Eq. (10.69), the oscillator strength $F_{\vec{k}}$ is only a weak function of $\vec{k}$ allows us to take the imaginary part of the permittivity as the delta function sum to obtain for absorption, with:

Eq. (10.77)

$$\operatorname{Im}\left(\frac{1}{\hbar\omega - E_C\left(\vec{k}\right) + E_V\left(\vec{k}\right) - i\eta}\right) = \pi\delta\left[\hbar\omega - E_C\left(\vec{k}\right) + E_V\left(\vec{k}\right)\right]$$

and therefore having split up the expression Eq. (10.69) we have:

Eq. (10.78) $$2\overline{n}\kappa = \frac{\hbar^2 q^2}{m_0\varepsilon_0} F_{vc} \int \frac{1}{2\hbar\omega} d\vec{k}\,\delta(\hbar\omega - E_C\left(\vec{k}\right) + E_V\left(\vec{k}\right))$$

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):

Eq. (10.79) $$\hbar\omega = \frac{\hbar^2 k^2}{2m_e} + \frac{\hbar^2 k^2}{2m_h} = \frac{\hbar^2 k^2}{2}\left(\frac{1}{m_r}\right)$$

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:

Eq. (10.80) $$m_r^* = \frac{m_e m_h}{m_e + m_h}$$

For example, in a three-dimensional bulk semiconductor structure with direct bandgap:

Eq. (10.81) $$2\overline{n}\kappa = \frac{\hbar q^2}{2\omega m_0 \varepsilon_0} F_{vc} \left\{ \frac{1}{2\pi^2} \left(\frac{2m_r^*}{\hbar^2} \right)^{3/2} \sqrt{\hbar\omega - E_g} \right\}$$

where by definition, the absorption coefficient is $\alpha = \frac{\omega}{\overline{n}c} 2\overline{n}\kappa$ 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. (10.74) (see also Appendix A.8).

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. (10.52) 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. (10.52) should be replaced by the Kane m^* (see Appendix A.8). The expression in the curly bracket on the RHS of Eq. (10.81) 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 2.2 for the hydrogen atom, we obtain the energy associated with the relative motion of an exciton:

Eq. (10.82) $$E_n = -\frac{E_{Ry}}{n^2}$$

where n=1,2,... 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 $l = 0,1,...,n-1$ and magnetic quantum numbers $m = -l,-l+1,...,l$. The Rydberg energy is given by:

Eq. (10.83) $$E_{Ry} = \frac{m_r^* q^4}{8(\varepsilon_r \varepsilon_0 h)^2}$$

with ε_r is the real part of the zero frequency relative permittivity or the dielectric constant of the material, ε_0 is the permittivity of free space and h is Plank's constant. Furthermore, by defining an exciton Bohr radius, a_B, derived from Eq. (2.4) such that:

Eq. (10.84) $$a_B = \frac{\varepsilon_r \varepsilon_0 h^2}{\pi m_r^* q^2}$$

we can rewrite the Rydberg energy as:

Eq. (10.85) $$E_{Ry}=\frac{q^2}{8\pi\varepsilon_r\varepsilon_0 a_B}=\frac{\hbar^2}{2m_r^* a_B{}^2}$$

The first fraction is similar to Eq. (2.5) 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 eV 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.

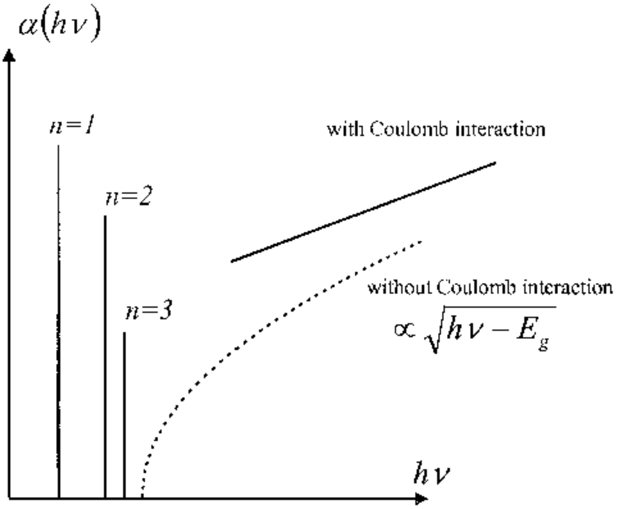


Fig. 10.4. Excitonic absorption peaks (n=1,2,3) in the optical absorption spectra of a bulk semiconductor (3D). These peaks are located inside the energy gap. In addition, the effect of coulombic interaction between electrons and holes on the absorption coefficient is shown.

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 Chapter 11.

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. 4.17 in Chapter 4 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. (10.73).

From Fig. 4.17 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 sources. 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:

Eq. (10.86) $E_c(\vec{k}+\vec{Q}) = E_v(\vec{k}) \pm \hbar\Omega + \hbar\omega$

the required momentum $\vec{Q}$ is fixed by the band structure. The process can be one of emission in which case the photon needs more energy than the indirect bandgap. The emission process is weakly dependent on temperature and involves the factor $1/(e^{\hbar\Omega/k_bT}-1)+1 = N(\omega)+1$. 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 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):

Eq. (10.87)

$$\alpha_{ep} = A_e \frac{(\hbar\omega - E_{ind}^{g} - \hbar\Omega)^2}{1 - e^{-\hbar\Omega / k_b T}} \theta(\hbar\omega - E_{ind}^{g} - \hbar\Omega)$$

$$\alpha_{ep} = A_a \frac{(\hbar\omega - E_{ind}^{g} + \hbar\Omega)^2}{-1 + e^{\hbar\Omega / k_b T}} \theta(\hbar\omega - E_{ind}^{g} + \hbar\Omega)$$

The $A's$ are constants, the theta function θ is zero when the argument is less than zero and one otherwise [Peyghambarian *et al.* 1993]. Note the different (squared) behavior of the band edge absorption Eq. (10.87) with photon energy when compared with the direct bandgap case Eq. (10.81) (square root).

Fig. 10.5 and Fig. 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.

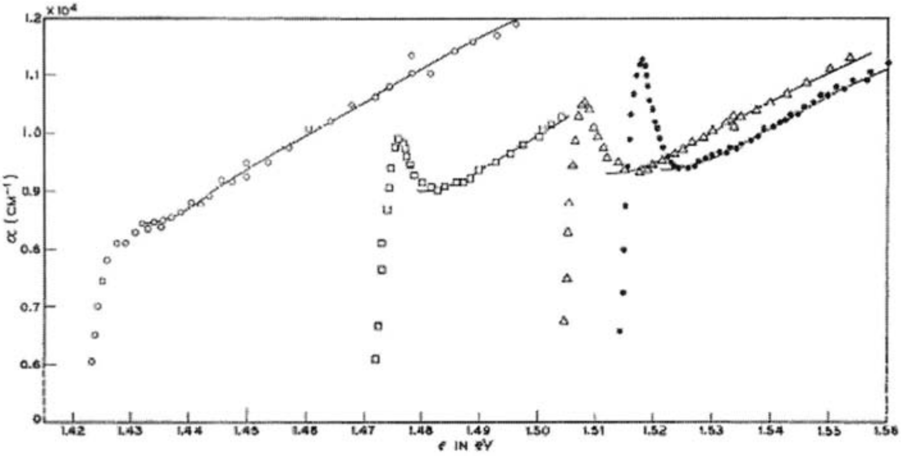


Fig. 10.5. Band edge absorption of GaAs showing also the evolution of the exciton absorption for different temperatures left to right: 294 K;186 K; 90 K; 21 K. [Reprinted figure with permission from Sturge M.D., Physical Review Vol. 127, p. 768, 1962. Copyright 1962 by the American Physical Society.]

When a phonon is needed, the transition is more complex, involves 3 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.

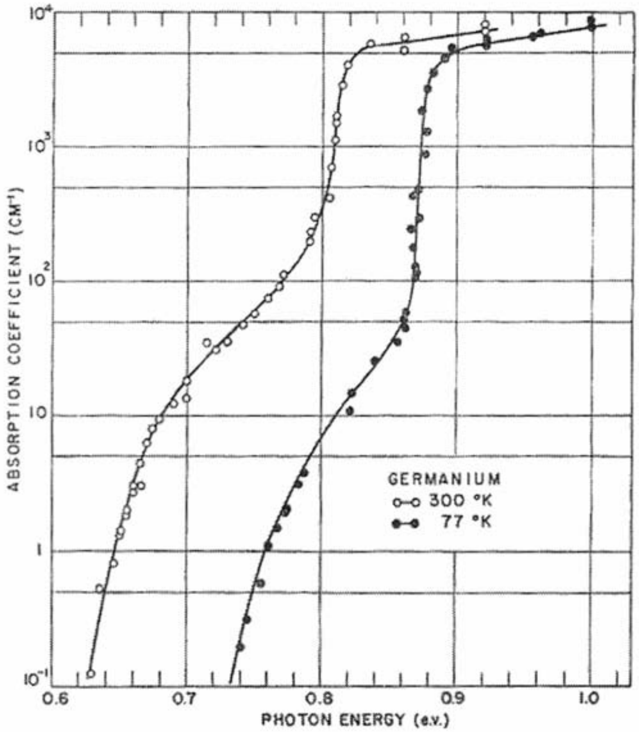


Fig. 10.6. The band edge absorption of Ge on a logarithmic scale. Note the change of behavior at 102 cm^{-1} from indirect to direct band to band transitions. [Reprinted from Solid State Physics Vol. 8, Newman, R. and Tyler, W.W., "Photoconductivity in Germanium," p. 58, Copyright (1959), with permission from Elsevier.]

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 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 may phonons effects is beyond the scope of this textbook. But the application of photon-phonon coupling as a characterization tool in semiconductors, is developed in Chapter 13. 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 in Chapter 5. 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. (5.5) now becomes:

Eq. (10.88)

$$M_1 \frac{d^2 u_n}{dt^2} + M_1 \gamma \frac{du_n}{dt} - C(v_{n+1} + v_{n-1} - 2u_n) = -qE_0 e^{-i\omega t}$$

$$M_2 \frac{d^2 v_n}{dt^2} + M_2 \gamma \frac{dv_n}{dt} - C(u_{n+1} + u_{n-1} - 2v_n) = qE_0 e^{-i\omega t}$$

where we have assumed that the M_1 mass is negatively charged and M_2 positive. 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 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. (10.88), 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:

Eq. (10.89) $$\begin{aligned} u_n(t) &= A_1 e^{-i\omega t} \\ \upsilon_n(t) &= A_2 e^{-i\omega t} \end{aligned}$$

and find from Eq. (10.88):

Eq. (10.90) $$\begin{aligned} [2C - M_1(\omega^2 + i\gamma\omega)]A_1 - 2CA_2 &= -qE_0 \\ [2C - M_2(\omega^2 + i\gamma\omega)]A_2 - 2CA_2 &= qE_0 \end{aligned}$$

the solution is

Eq. (10.91) $$\begin{aligned} A_1 &= \frac{-qE_0}{M_1[\Omega_+^2 - (\omega^2 + i\gamma\omega)]} \\ A_2 &= \frac{qE_0}{M_2[\Omega_+^2 - (\omega^2 + i\gamma\omega)]} \end{aligned}$$

Eq. (10.92) $$\Omega_+^2(k=0) = \frac{2C(M_1 + M_2)}{M_1 M_2}$$

Using this result we can now go back and compute the polarization induced by the light wave. Given N_I ion pairs per unit volume we have the volume dipole moment:

Eq. (10.93) $$P_I = -qN_I(u_n - v_n)$$

or:

Eq. (10.94) $$P_I = qN_I(A_2 - A_1)e^{-i\omega t}$$

which then using Eq. (10.91) reduces in the limit of the pure ionic permittivity to:

Eq. (10.95) $$\varepsilon_I(\omega) = 1 + \frac{qN_I}{E_0\varepsilon_0}(A_2(\omega) - A_1(\omega))$$

Eq. (10.96) $\varepsilon_l(\omega) = 1 + \frac{q^2 N_l}{\varepsilon_0 M_r} \frac{1}{\Omega_+^2 - (\omega^2 + i\gamma\omega)}$

Eq. (10.97) $M_r = \frac{M_1 M_2}{M_1 + M_2}$

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:

Eq. (10.98) $\varepsilon(\omega) = 1 + (\varepsilon_f(\omega) - 1) + (\varepsilon_b(\omega) - 1) + (\varepsilon_l(\omega) - 1)$

where the polarizability contributions are added as given by Eq. (10.96), Eq. (10.66) and Eq. (10.33), and where it is understood that in a semiconductor, the bound contribution is the same as the formula Eq. (10.68). 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 than the 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

Eq. (10.99)
$$\overline{n}^2 - \kappa^2 = \varepsilon(\infty) + \frac{q^2 N_l}{\varepsilon_0 M_r} \frac{\Omega_T^2 - \omega^2}{(\Omega_T^2 - \omega^2)^2 + \omega^2\gamma^2}$$
$$2\overline{n}\kappa = \frac{q^2 N_l}{\varepsilon_0 M_r} \frac{\gamma\omega}{(\Omega_T^2 - \omega^2)^2 + \omega^2\gamma^2}$$

Fig. 10.7 shows the reflectivity R of an ionic insulator. The effect of the resonance on the reflectivity is to produce a sharp cross over from high to low reflectance as the photon-energy is changed.

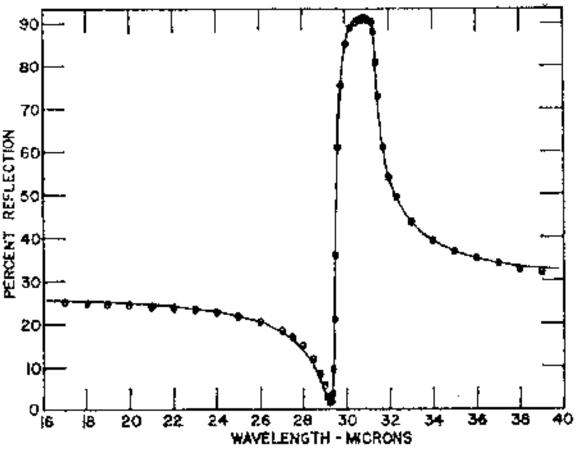


Fig. 10.7. Lattice reflection spectrum of AlSb. Points are experimental data, line is fit using the single oscillator model. [Reprinted figure with permission from Turner, W.J. and Reese, W.E., Physical Review Vol. 127, p. 126, 1962. Copyright 1962 by the American Physical Society.]

10.6.3. The phonon-polariton

The real part of the refractive index duc to the coupling with ions has a strong frequency dependence as can be seen in the previous figure and strongly modulates photons with frequencics in the neighborhood of the optical modes. Indeed the photon dispersion relation relating photon frequency and momentum k is:

Eq. (10.100) $$\omega^2(k) = \left[\frac{ck}{\overline{n}(\omega)}\right]^2$$

where $\overline{n}(\omega)$ is given by the pair of Eq. (10.99). 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. (10.99) and one has $\overline{n}(\omega)$ only so Eq. (10.100) becomes:

Eq. (10.101) $$\omega^2(k)\left\{\varepsilon(\infty) + \frac{q^2 N_l}{\varepsilon_0 M_r}\frac{1}{(\Omega_T^2 - \omega^2(k))}\right\} = c^2 k^2$$

which is a quadratic equation in ω^2 with two branches.

The frequency versus momentum of the physical roots are 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.

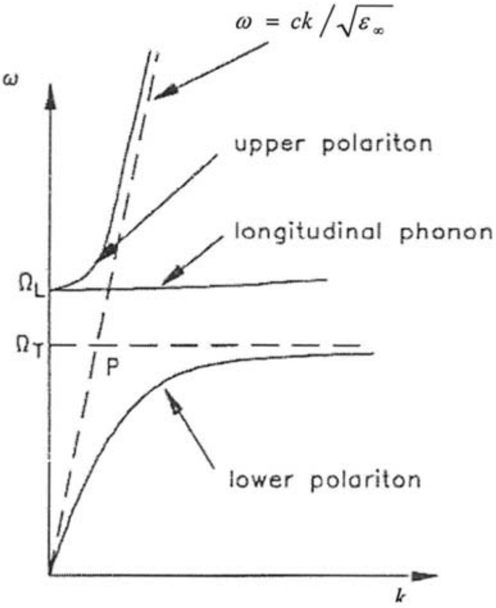


Fig. 10.8. The dispersion curve for a phonon-polariton. [Introduction to Semiconductor Optics, 1993, p. 98, Peyghambarian, N., Koch, S.W., and Mysyrowicz, A., Fig. 4.11. Modified with permission.]

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. Here, the photon becomes a mixed state, part phonon and part photon, it gets slowed down in the process because the phonon is slow 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-photon coupling and absorption is more subtle and involves 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. Opto-electronics 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 semi-classical 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 wave functions. 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 are nearly free electron like, so that the total wavefunction and energy of the charge is separable:

Eq. (10.102) $$\Psi_{\vec{K}}(x,y,z) = e^{ik_x x}e^{ik_y y}\Phi(z)$$

Eq. (10.103) $$E = \frac{\hbar^2}{2m^*}(k_x^2 + k_y^2) + E_z$$

The Schrödinger equation in the field direction becomes:

Eq. (10.104) $$-\frac{\hbar^2}{2m^*}\frac{\partial^2\Phi(z)}{\partial z^2} - qE_0^z z\Phi(z) = E_z\Phi(z)$$

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 this equations are called the Airy functions $Ai_\nu(z)$ with energy E_ν and given by an integral representation:

Eq. (10.105) $$Ai_\nu(z_\nu) = \frac{1}{\pi}\int_0^\infty \cos\left(\frac{s^3}{3} + sz_\nu\right)ds$$

Eq. (10.106) $$z_\nu = \left(\frac{2m^* qE_0^z}{\hbar^2}\right)^{1/3}\left(z - \frac{E_\nu}{qE_0^z}\right)$$

The normalized eigenstates of Eq. (10.104) labeled with their energies are:

Eq. (10.107)
$$\Phi_{E_\nu} = \left(\frac{2m*}{\hbar^2}\right)^{1/3}\left(\frac{1}{qE_0^z}\right)^{1/6} Ai\left[\left(\frac{2m^* qE_0^z}{\hbar^2}\right)^{1/3}\left(z - \frac{E_\nu}{qE_0^z}\right)\right]$$

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 $E_\nu / qE_0^z = a\nu$ where ν is an integer in the range $\{\infty, -\infty\}$, and $a\nu$ defines the origin of the ν_{th} Airy state. The Airy function decay asymptotically as $e^{-z_\nu^{3/2}}$ 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:

Eq. (10.108) $$Ai(z) = \frac{1}{\sqrt{\pi}} \frac{1}{(-z_\nu)^{1/4}} \sin\left(\frac{2}{3}(-z_\nu)^{3/2} + \frac{\pi}{4} \right)$$

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. (10.108) 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 the Fermi golden rule:

Eq. (10.109)
$$W_{\nu\nu'} = \frac{2\pi}{\hbar} \left| \int \Phi_{\nu,v} * (z) qE_0^z z \Phi_{\nu',c}(z) dz \right|^2 \delta(E_{\nu',c} - E_{\nu,v} - \hbar\omega)$$

Here the momentum rule reappears as a reduction in the overlap integral Eq. (10.109) between levels which are not vertically above each other, i.e. differ by the ν index of the valence to the ν' index of the conduction band Airy states. So we see that non-diagonal transitions $\nu \neq \nu'$ 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:

Eq. (10.110) $n(E,z)=\sum_{n}\left|\Phi_n(z)\right|^2\delta(E-E_n)$

where here, Φ_n are the energy eigenstates Eq. (10.107) 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:

Eq. (10.111)

$$\int\sum_{n}\left|\Phi_n(z)\right|^2\delta(E-E_n)dz=\sum_{n}\delta(E-E_n)=g(E)$$

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. (10.107) and eigenvalues into Eq. (10.110) 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. Fig. 10.10 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 Chapter 11.

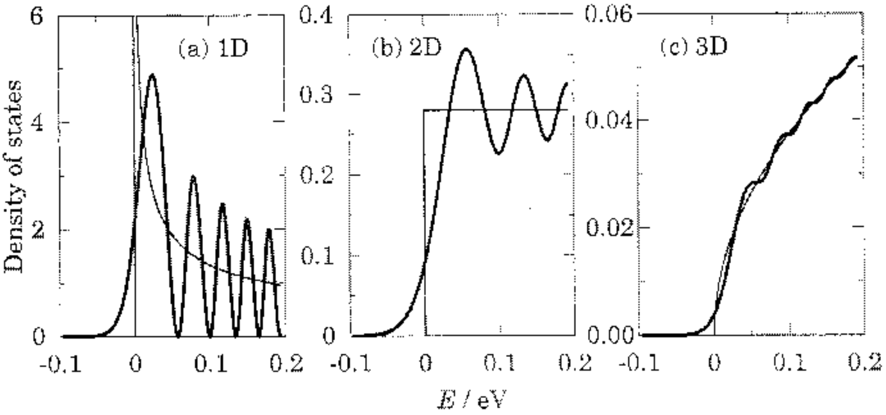


Fig. 10.9. The density of states from left to right for a one-, two- and three-dimensional free electron system in the presence of an electric field. [Davies, J.H., The Physics of Low Dimensional Semiconductors: an Introduction, p. 211, Fig. 6.2. © Cambridge University Press 1998. Reprinted with the permission of Cambridge University Press.]

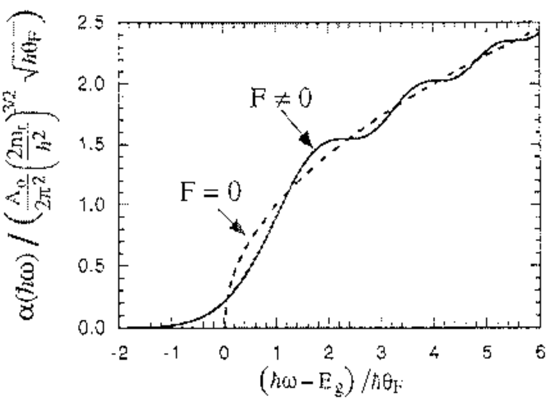


Fig. 10.10. Franz-Keldysh oscillations in the absorption of bulk semiconductors. The dashed line is the spectrum without a field. [Physics of Optoelectronic Devices, Chuang, S.L. Copyright © 1995 by John Wiley & Sons, Inc. Reprinted with permission of Wiley-Liss Inc., a subsidiary of John Wiley & Sons, 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:

Eq. (10.112) $\vec{B}=\vec{\nabla}\times\vec{A}$

and consequently the time independent Schrödinger becomes:

Eq. (10.113)

$$\frac{1}{2m^*}\left[-\hbar^2\frac{\partial^2}{\partial x^2}+(-i\hbar\frac{\partial}{\partial y}+qBx)^2-\hbar^2\frac{\partial^2}{\partial z^2}\right]\Psi(x,y,z)=E\Psi(x,y,z)$$

From Eq. (10.113) it follows that in the z-direction, the Hamiltonian is that of the free particle, in the y-direction the interaction is an x-y product term so we try the solution:

Eq. (10.114) $\Psi(x,y,z)=u(x)e^{ik_y y}e^{ik_z z}$

with

Eq. (10.115) $E=\frac{\hbar^2k_z^2}{2m^*}+E_{n//}$

Including the spin degree of freedom $s=\pm 1/2$ we also have the Zeeman splitting in a magnetic field:

Eq. (10.116) $E_{nk_z s}=E_{n//}+\frac{\hbar^2k_z^2}{2m^*}-sg\mu_B B$

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. (10.113) into Eq. (10.112) we find that the function $u(x)$ must satisfy:

Eq. (10.117) $\left[-\frac{\hbar^2}{2m^*}\frac{d^2}{dx^2}+\frac{m^*\omega_c^2}{2}\left(x+\frac{\hbar k_y}{qB}\right)^2\right]u(x)=\varepsilon u(x)$

This equation is similar to the one of the harmonic oscillator where:

Eq. (10.118) $\omega_c = \dfrac{qB}{m^*}$

is the cyclotron frequency and $\dfrac{\hbar k_y}{qB}$ is a length which we shall denote with $\{-x_{k_y}\}$. The equation Eq. (10.117) is a standard differential equation of mathematical physics which has the Hermite polynomials H_n as solutions. We can therefore now write the complete wave function as:

Eq. (10.119) $\Psi(x,y,z) = Ae^{ik_y y}e^{ik_z z}H_{nk_y}\left(\dfrac{x-x_{k_y}}{l_B}\right)\exp\left[-\dfrac{(x-x_{k_y})^2}{2l_B^2}\right]$

where n are integers, A is the normalization constant, and $l_B = \sqrt{\dfrac{\hbar}{qB}}$ is called the cyclotron radius that is typically ~25 nm for B=1 T. The first few normalized Hermite polynomials are tabulated and given by:

Eq. (10.120) $$\begin{aligned} H_0(s) &= 1 \\ H_1(s) &= 2s \\ H_2(s) &= 4s^2 - 2 \end{aligned}$$

The corresponding *x-y* energy levels are independent of the index k_y, and given by (n is an integer including 0):

Eq. (10.121) $E_{//} = \varepsilon_n = (n + 1/2)\hbar\omega_c$

These levels are called the Landau levels. Each Landau level is highly degenerate because there are many k_y levels in each Landau level. In fact there are exactly $\dfrac{L_x L_y qB}{h}$ k_y 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 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 satisfies the condition that $\frac{h}{\tau} < \hbar\omega_c$. This condition is difficult to satisfy in practice because in a metal $\tau \sim 10^{-13} - 10^{-14}$ s, 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 Chapter 11 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 ~70K. We will come back to this topic when we discuss the low-dimensional semiconducting systems in Chapter 11. 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 wave functions and energies into Eq. (10.110) 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 (Chapter 4 replace $4\pi k^2 dk \rightarrow 2\pi k dk$ in Eq. (4.37)). 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. Fig. 10.11 illustrates how the two-dimensional constant density of states collapses into Landau levels which are not ultra-sharp delta functions, but broadened by disorder or phonon scattering processes. Thus in a two dimensional system, the electrons would fill the Landau levels up the Fermi energy which can the be in the Landau band or in the gap according to the electron concentration for a given field. Such quasi two-dimensional systems can be made using multilayers and quantum wells as we shall see in detail in Chapter 11. 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. (10.39) 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 non conducting 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 Chapter 11.

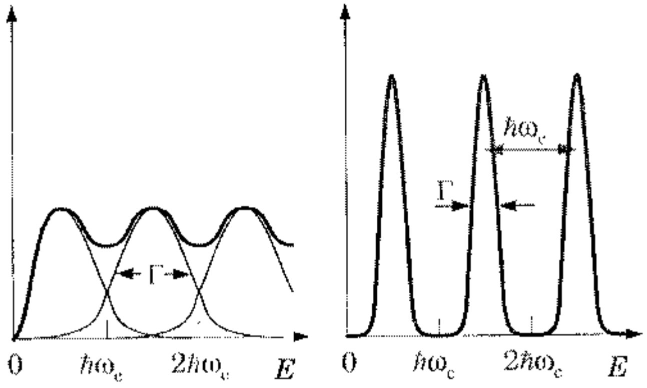


Fig. 10.11. The density of states of a two-dimensional electron gas in a magnetic field for two different values of broadening. As the broadening is reduced, the Landau levels become delta function like peaks. With increased broadening, the trend is to a constant density of states as in the B=0 limit. [Davies, J.H., The Physics of Low Dimensional Semiconductors: an Introduction, p. 225, Fig. 6.7b and 6.7c. © Cambridge University Press 1998. Reprinted with the permission of Cambridge University Press.]

In a three-dimensional system, the k_z degree of freedom broadens the Landau bands and we have (spinless case):

Eq. (10.122) $$g(E) = \frac{qBL_xL_y}{h}\sum_{n,k_z}\delta(E-\varepsilon_n-\frac{\hbar^2k_z^2}{2m^*})$$

Eq. (10.123) $$g(E)=\frac{qVB\sqrt{2m^*}}{(2\pi\hbar)^2}\sum_{n=0}^{n_{\max}}[E-(n+1/2)\hbar\omega_c]^{-1/2}$$

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, 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 opto-electronics.

The permittivity of free electrons in a magnetic field can be computed using the wavefunctions we obtained Eq. (10.119) and substituting them into Eq. (10.68), however it is often adequate to compute the optical spectra of materials within a semi-classical treatment. This can be done by adding the Lorentz force $F_L=-q\frac{d\vec{r}}{dt}\times\vec{B}$ to the RHS of Eq. (10.25), 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. (10.17), 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. (10.25):

Eq. (10.124) $$\begin{aligned} m^*\frac{d^2x}{dt^2}+m^*\frac{dx}{dt}\frac{1}{\tau}&=-qE(t)-q\frac{dy}{dt}B \\ m^*\frac{d^2y}{dt^2}+m^*\frac{dy}{dt}\frac{1}{\tau}&=q\frac{dx}{dt}B \end{aligned}$$

These equations are solved by making the same assumption as before for the displacements $x(t)=x_0e^{-i\omega t}$ and $y(t)=y_0e^{-i\omega t}$. We find the new B field dependent free carrier relative permittivity contribution and add it to the bound relative permittivity to obtain:

Eq. (10.125) $\varepsilon(\omega)=\varepsilon_b(\omega)+\dfrac{i}{\omega\varepsilon_0}\sigma(B,\omega)$

where the complex conductivity now is dependent on the B field via the cyclotron frequency:

Eq. (10.126) $\sigma(B,\omega)=\dfrac{n_c q^2\tau}{m^*}\left(\dfrac{1}{\tau}\right)\left\{\dfrac{1/\tau-i\omega}{(i\omega-1/\tau)^2+\omega_c^2}\right\}$

Eq. (10.126) reduces to the usual result Eq. (10.29) 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 Eq. (10.126) and Eq. (10.15):

Eq. (10.127) $\alpha(\omega)=\dfrac{\omega}{c}\left\{\dfrac{((1/\tau)^2+\omega_c^2+\omega^2)}{[(1/\tau)^2+\omega_c^2-\omega^2]^2+4\omega^2/\tau^2}\right\}\dfrac{n_c q^2}{\varepsilon_0 m^*}\left(\dfrac{1}{\omega\tau}\right)$

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 semi-classical 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-optic 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. Fig. 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 Chapter 11.

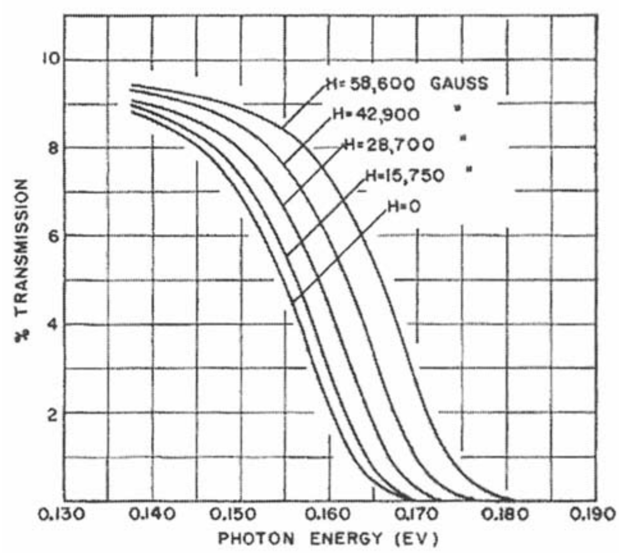


Fig. 10.12. the band edge absorption of InSb with magnetic field at room temperature. [Reprinted figure with permission from Burstein, E., Picus, G.S., Gebbie, H.A., and Blatt, F., Physical Review Vol. 103, p. 827, 1956. Copyright 1956 by the American Physical Society.]

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:

Eq. (10.128) $\vec{P} = \chi^{(1)}\vec{E} + \chi^{(2)}\vec{E}\cdot\vec{E} + \chi^{(3)}\vec{E}\cdot\vec{E}\cdot\vec{E} + \ldots$

or equivalently:

Eq. (10.129) $P_i = \sum_l \chi_{il}^{(1)} E_l + \sum_{l,k} \chi_{ilk}^{(2)} E_l E_k + \sum_{lks} \chi_{ilks}^{(3)} E_l E_k E_s + \ldots$

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. (10.49)) 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. (10.128) 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 processes 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 3-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 respond 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 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 the Fermi 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 a 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 strong impact on technology. This is specially 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 it forces us to use the exact wavefunctions derived above, however weak 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 semi-classical 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 wave-functions 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 Chapter 11.

References

Burstein, E., Picus, G.S., Gebbie, H.A., and Blatt, F. "Magnetic optical bandgap effects in InSb," *Physical Review* 103, pp. 826-828, 1956.

Chuang, S.L., *Physics of Optoelectronic Devices*, John Wiley & Sons, New York, 1995.

Davies, J.H., *The Physics of Low Dimensional Semiconductors: an Introduction*, Cambridge University Press, 1998.

Newman, R. and Tyler, W.W., "Photoconductivity in Germanium,"in *Solid State Physics* 8, eds. F. Seitz and D. Turnbull, pp. 49-103, Academic Press, New York, 1959.

Peyghambarian, N., Koch, S.W., and Mysyrowicz, A., *Introduction to Semiconductor Optics*, Prentice-Hall, Englewood cliff, New Jersey, 1993.

Rosencher, E. and Vinter, B., *Optoelectronics*, Cambridge University Press, 2002.

Seeger, K., *Semiconductor Physics: an Introduction*, Springer, 1997.

Sturge, M.D., "Optical absorption of Gallium Arsenide between 0.6 and 2.75 eV," *Physical Review* 127, pp. 768-773, 1962.

Turner, W.J. and Reese, W.E., "Infrared lattice absorption of AlSb," *Physical Review* 127, pp. 126-131, 1962.

Further reading

Bockrath, M., Cobden, D.H., Lu, J., Rinzler, A.G., Smalley, R.E., Balents, L., and McEuen, P.L., "Luttinger liquid behaviour in carbon nanotubes," *Nature* 397, pp. 598-601, 1999.

Bastard, G., *Wave Mechanics Applied to Semiconductor Heterostructures*, Halsted Press, New York, 1988.

Cohen-Tannoudji, C., Diu, B. and Laloë, F., *Quantum Mechanics*, John Wiley & Sons, New York, 1977.

Davydov, A.S., *Quantum Mechanics*, Pergamon, New York, 1965.

Kittel, C., *Introduction to Solid State Physics*, John Wiley & Sons, New York, 1976.

Liboff, R.L., *Introductory Quantum Mechanics*, Addison-Wesley, Reading, MA, 1998.

Madelung, O., *Introduction to Solid State Theory*, Springer, New York, 1978.

Powell, J.L. and Crasemann, B., *Quantum Mechanics*, Addison-Wesley, Reading, MA, 1961.

Ziman, J.M., *Principles of the Theory of Solids*, Cambridge University Press, 1964.

Ziman, J.M., *Elements of Advanced Quantum Theory*, Cambridge University Press, London, 1969.

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=1nm, effective mass m*=0.023m_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*=0.045m_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 by $L_x L_y qB/h$ for an area of size $L_x L_z$ 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). Shows that this leads to a quadratic equation from which the real and imaginary part of the complex refractive index $\overline{n}$ 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 photon polariton equation using Eq. (10.101). 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 $m_e = 0.067m_0$ and the effective mass of the hole is $m_h = 0.0.82m_0$. The relative static permittivity ε_r is 13.1. Using Eq. (10.84) and Eq. (10.85), 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. (10.124), derive the magnetic field dependent complex conductivity of an electron gas as given by Eq. (10.126): $\sigma(B,\omega) = \frac{nq^2\tau}{m^*}\left(\frac{1}{\tau}\right)\left\{\frac{1/\tau - i\omega}{(i\omega - 1/\tau)^2 + \omega_c^2}\right\}$. 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?

11. Semiconductor Heterostructures and Low-Dimensional Quantum Structures

11.1. Introduction

In Chapter 3, we have introduced the basic concepts and formalism of quantum mechanics. In Chapter 4, 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), metalorganic chemical vapor deposition (MOCVD), has made it possible to control with atomic scale precision 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 heterostructures including semiconductor lasers, light-emitting diodes and 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 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.

11.2. 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.

11.2.1. 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. 11.1(a). 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.

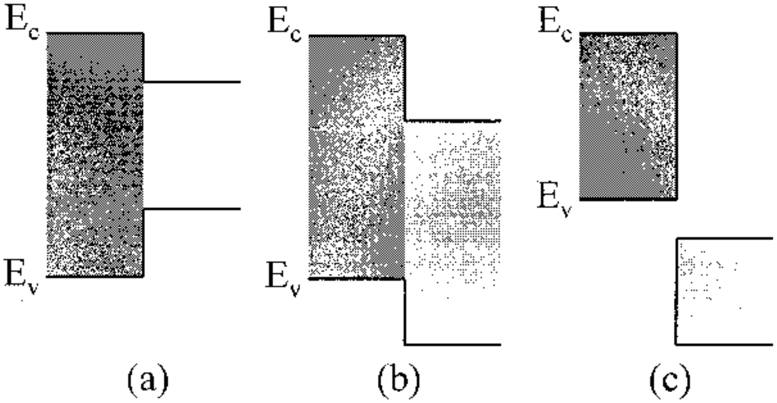


Fig. 11.1. Heterojunction band line ups for isolated, but adjacent, semiconductors: (a) type I, (b) type II staggered, and (c) type II broken gap alignments.

11.2.2. 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. 11.1(b). 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. 11.1(c).

11.3. 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$_x$Ga$_{1-x}$As ($0<x<0.4$) junction system.

We assume A and B represent two III-V semiconductors that have the same lattice constant. The valence band position can be calculated as:

Eq. (11.1) $$E_V = E_{V,av} + \frac{\Delta}{3}$$

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:

Eq. (11.2) $$\Delta E_V = \left(E^A_{V,av} - E^B_{V,av}\right) + \frac{1}{3}\left(\Delta_A - \Delta_B\right)$$

The conduction band edge is obtained by adding the bandgap value to the valence band position:

Eq. (11.3) $$E_C = E_V + E_g$$

Therefore the conduction band offset can be calculated as:

Eq. (11.4) $$\Delta E_C = \left(E^A_{V,av} - E^B_{V,av}\right) + \frac{1}{3}\left(\Delta_A - \Delta_B\right) + \left(E^A_g - E^B_g\right)$$

All these quantities are summarized in Fig. 11.2.

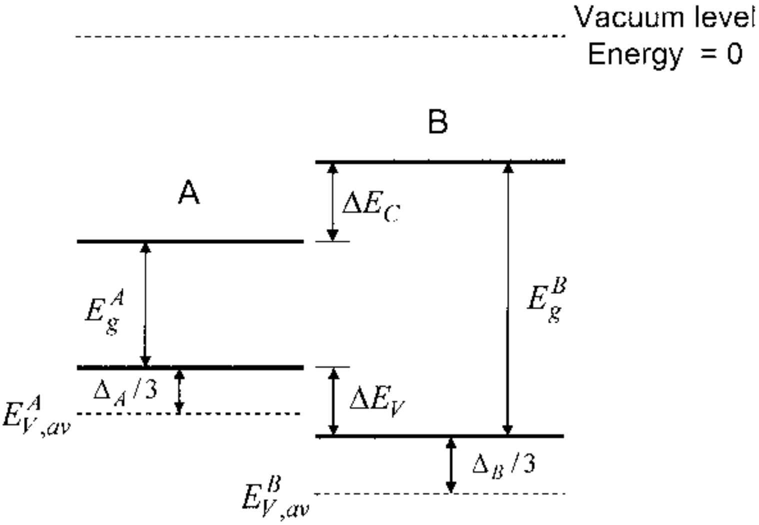


Fig. 11.2. Band alignment diagram for calculation of band offset.

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 11.1.

	$E_{V,av}$ (eV)	Δ (eV)	E_g (eV)
GaAs	-6.92	0.34	1.52
AlAs	-7.49	0.28	3.13

Table 11.1. Material parameters for GaAs and AlAs.

A: For GaAs, we have:

$$E_V^{GaAs} = E_{V,av}^{GaAs} + \frac{\Delta_{GaAs}}{3} = -6.807eV$$

For $Al_{0.2}Ga_{0.8}As$ we use the arithmetic average of 20 % AlAs and 80 % of GaAs:

$$E_V^{Al_{0.2}Ga_{0.8}As} = 0.2\times\left(E_{V,av}^{AlAs}+\frac{\Delta_{AlAs}}{3}\right)+0.8\times\left(E_{V,av}^{GaAs}+\frac{\Delta_{GaAs}}{3}\right)$$
$$= -6.925eV$$

$$E_g^{Al_{0.2}Ga_{0.8}As} = 0.2\times E_g^{AlAs} + 0.8\times E_g^{GaAs} = 1.842eV$$

Therefore, we obtain the band offset as following,

$$\begin{cases} \Delta E_V = E_V^{GaAs} - E_V^{Al_{0.2}Ga_{0.8}As} = (-6.807)-(-6.925) \\ \quad = 0.118eV \\ \Delta E_C = \left(E_V^{Al_{0.2}Ga_{0.8}As} + E_g^{Al_{0.2}Ga_{0.8}As}\right) - \left(E_V^{GaAs} + E_g^{GaAs}\right) \\ \quad = (-5.287)-(-5.555) \\ \quad = 0.268eV \end{cases}$$

11.4. 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_{F_A} and E_{F_B}. We use V_0 to represent the potential difference due to the energy difference between E_{F_A} and E_{F_B}, as shown in Fig. 11.2(a). According to Fig. 11.2(a) we have:

Eq. (11.5) $$V_0 = E_g^A + \Delta E_C - \left(E_{F_A} - E_V^A\right) - \left(E_C^B - E_{F_B}\right)$$

For non-degenerate semiconductors, we have:

Eq. (11.6)
$$\begin{cases} E_{F_A} - E_V^A = -k_b T \ln\left(\dfrac{N_A}{N_v^A}\right) \\ E_V^B - E_{F_B} = -k_b T \ln\left(\dfrac{N_d}{N_c^B}\right) \end{cases}$$

where N_v^A and N_c^B are the valence band and conduction band density of states for semiconductor A and B respectively. Substituting Eq. (11.6) into Eq. (11.5) we obtain the expression for V_0:

Eq. (11.7)
$$V_0 = E_g^A + \Delta E_C + k_b T \ln\left(\frac{N_A \cdot N_D}{N_v^A \cdot N_c^B}\right)$$

After we bring semiconductor A and B together into contact, there will be a net electron transfer from B to A (see Fig. 11.3(c)) until the Fermi levels on both sides reach the same value, as shown in Fig. 11.3(b).

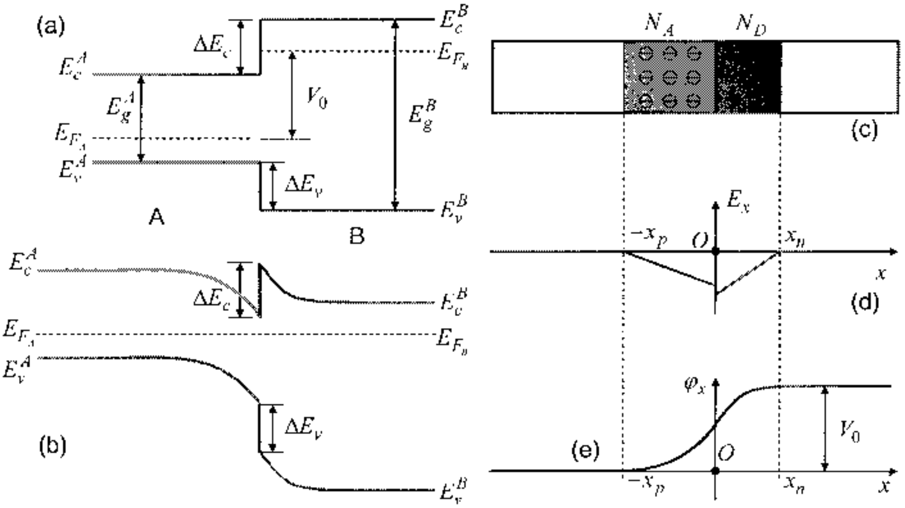


Fig. 11.3. Illustrations for (a) band diagram for the heterojunction before charge transfer, (b) band diagram after charge transfer, (c) depletion approximation, (d) electric field distribution, and (e) electric potential distribution.

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:

Eq. (11.8) $$N_A x_p = N_D x_n$$

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 electric 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:

Eq. (11.9) $$\begin{cases} E_x = -\dfrac{qN_A\left(x+x_p\right)}{\varepsilon_A\varepsilon_0}, & -x_p \le x < 0 \\ E_x = -\dfrac{qN_D\left(x_n-x\right)}{\varepsilon_B\varepsilon_0}, & 0 < x \le x_n \end{cases}$$

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 towards the potential zero point to obtain the electric potential profile:

Eq. (11.10) $$\varphi_x = \int_x^{-x_p} E_x dx$$

Substituting Eq. (11.9) into Eq. (11.10) we have:

Eq. (11.11) $$\begin{cases} \varphi_x = 0, \quad x < -x_p \\ \varphi_x = \dfrac{qN_A\left(x+x_p\right)^2}{2\varepsilon_A\varepsilon_0}, & -x_p \le x < 0 \\ \varphi_x = \dfrac{qN_A x_p^{\;2}}{2\varepsilon_A\varepsilon_0} + \dfrac{qN_D\left(2x_n x - x^2\right)}{2\varepsilon_B\varepsilon_0}, & 0 \le x \le x_n \\ \varphi_x = \dfrac{qN_A x_p^{\;2}}{2\varepsilon_A\varepsilon_0} + \dfrac{qN_D x_n^{\;2}}{2\varepsilon_B\varepsilon_0}, & x > x_n \end{cases}$$

We recall that the total potential drop is V_0 as calculated before, i.e.:

Eq. (11.12) $$\frac{qN_A x_p{}^2}{2\varepsilon_A\varepsilon_0}+\frac{qN_D x_n{}^2}{2\varepsilon_B\varepsilon_0}=V_0$$

Combining Eq. (11.8) and Eq. (11.12), we obtain the values of x_n and x_p in terms of V_0:

Eq. (11.13) $$\begin{cases} x_n=\sqrt{\dfrac{N_A}{N_D}\dfrac{2\varepsilon_0 V_0}{q}\dfrac{\varepsilon_A\varepsilon_B}{N_A\varepsilon_A+N_D\varepsilon_B}} \\ x_p=\sqrt{\dfrac{N_D}{N_A}\dfrac{2\varepsilon_0 V_0}{q}\dfrac{\varepsilon_A\varepsilon_B}{N_A\varepsilon_A+N_D\varepsilon_B}} \end{cases}$$

We define the junction depletion width as $x_w=x_n+x_p$. Taking into account Eq. (11.13) we can obtain:

Eq. (11.14) $$x_w=\sqrt{\frac{2\varepsilon_0 V_0}{qN_D N_A}\frac{\varepsilon_A\varepsilon_B}{N_A\varepsilon_A+N_D\varepsilon_B}\cdot(N_D+N_A)}$$

Substituting Eq. (11.13) into Eq. (11.11), 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. 11.3(a) we will obtain a calculated electron energy profile for the heterojunction under equilibrium as illustrated in Fig. 11.3(b).

11.5. 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. 11.4.

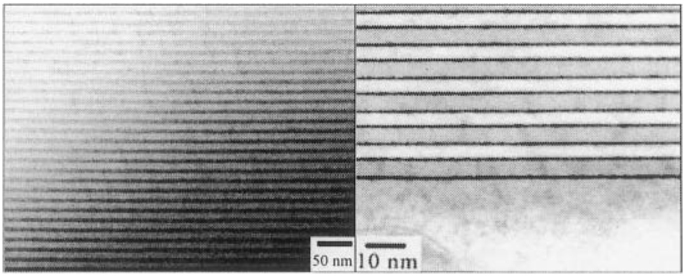


Fig. 11.4. Transmission electron microscope images of type II InAs/GaSb superlattice. The dark regions correspond to the InSb interface between InAs and GaSb layers.

11.6. Two-dimensional structures: quantum wells

11.6.1. 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. 11.5. 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 Chapter 4, 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 sub-section 3.3.3. 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 sub-section 3.3.1. The total potential can therefore be expressed as:

Eq. (11.15) $$U(x,y,z)=\begin{cases}0 & for\, 0<x<a\\ U_0>0 & for\, x<0\, and\; x>a\end{cases}$$

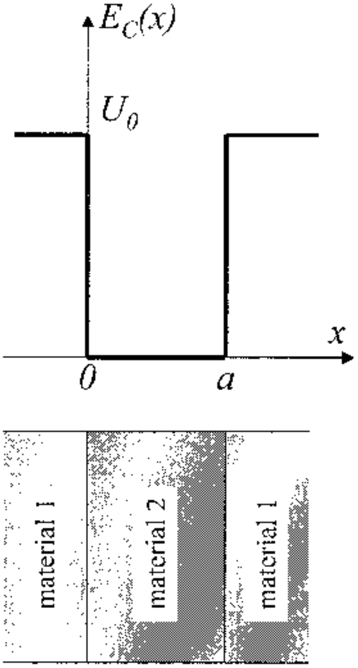


Fig. 11.5. Potential energy profile of a quantum well. This profile can be obtained by sandwiching a thin and flat semiconductor film of material 2 between two semiconductor crystals of another material 1.

and the time independent Schrödinger equation becomes:

Eq. (11.16) $$-\frac{\hbar^2}{2m^*}\nabla^2\Psi(x,y,z)-[E-U(x,y,z)]\Psi(x,y,z)=0$$

where m^* is the electron effective mass. The shape of the potential in Eq. (11.15) implies that the motion in the x-direction and that in the (y,z)-plane are independent. It is 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, $\vec{r}_\perp$ 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, $\Psi_{total}(x,y,z)=\Psi_{//}(\vec{r}_{//})\Psi_\perp(\vec{r}_\perp)$, 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.

(1) In-plane motion

In the (y,z)-plane, the motion of the electron is similar to that of a free particle discussed in sub-section 3.3.1. The wavefunction $\Psi_{//}\left(\vec{r}_{//}\right)$ can therefore be considered to be a plane wave similar to Eq. (3.22) and can be expressed as:

Eq. (11.17) $$\Psi_{//}(\vec{r}_{//}) = A\exp\left(i\vec{k}_{//}\cdot\vec{r}_{//}\right)$$

where A is a normalization constant. The energy spectrum in the (y,z) plane is given by:

Eq. (11.18) $$E_{//}\left(\vec{k}_{//}\right)=\frac{\hbar^2\vec{k}_{//}^2}{2m^*}=\frac{\hbar^2\left(k_y{}^2+k_z{}^2\right)}{2m^*}$$

Note that these expressions are correct only for small values of the momentum such that $\left|\vec{k}_{//}\right| << \left|\vec{K}\right|$, 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 Chapter 4.

(2) 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 sub-section 3.3.3 Although no analytical solution was derived, the main results can be summarized as follows.

The set of equations from Eq. (3.43) to Eq. (3.45) 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,… and these quantities must satisfy Eq. (3.37):

Eq. (11.19) $$\begin{cases} E_{\perp n}=\dfrac{\hbar^2\left(k_{\perp n}\right)^2}{2m^*} \\ \alpha_n=\sqrt{\dfrac{2m^*\left(U_0-E_{\perp n}\right)}{\hbar^2}} \end{cases}$$

Note that we are now using the effective mass of the electron, m^*. The spacing between consecutive energy levels is on the order of $\frac{\hbar^2\pi^2}{m^* a^2}$ from Eq. (3.31). For $E_n<U_0$, the wavefunction $\Psi'_\perp(\vec{r}_\perp)=\Psi'_\perp(x)$ 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. (3.38), Eq. (3.40) and the values of $E_{\perp_n}$, $k_{\perp n}$ and α_n as illustrated in Fig. 11.6

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. (3.3)).

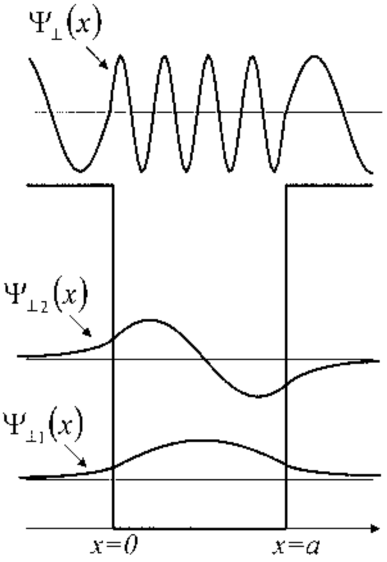


Fig. 11.6. Shapes of the wavefunctions $\Psi'_\perp(x)$ for the allowed energy levels of a quantum well. In this example, there are only two allowed confined states. The wavefunctions of these have an oscillatory behavior inside the well ($0<x<a$) but vanish rapidly when outside the quantum well. A third 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.

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 $\frac{\hbar}{\tau}$ where τ is a characteristic time between electron collisions, or electron lifetime, which can be 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 ($\frac{\hbar^2\pi^2}{m^*a^2}$) is larger than the broadening ($\frac{\hbar}{\tau}$). In other words, the inequality $\frac{\hbar^2\pi^2}{m^*a^2} >> \frac{\hbar}{\tau}$ ensures that the quantized behavior can be observed.

11.6.2. Density of states

The total energy spectrum for an electron in a quantum well is given by considering Eq. (11.18) and Eq. (11.19):

Eq. (11.20) $$E(\vec{k}_{//},n) = E_{//}\left(\vec{k}_{//}\right) + E_{\perp n} = \frac{\hbar^2\vec{k}_{//}^{\,2}}{2m^*} + \frac{\hbar^2\left(k_{\perp n}\right)^2}{2m^*}$$

where the values of $\vec{k}_{//}$ are continuous, while $k_{\perp n}$ is quantized and indexed by an integer *n*. Similar to Eq. (4.41) in sub-section 4.3.2, 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:

Eq. (11.21) $$g_{2D}(E) = 2\sum_{n,\vec{k}_{//}}\delta\left[E_{//}\left(\vec{k}_{//}\right) + E_{\perp n} - E\right]$$

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. (4.44) is performed on two coordinates only:

Eq. (11.22)
$$\sum_{\vec{k}_{//}}\Upsilon\left(\vec{k}_{//}\right) \equiv \frac{S}{(2\pi)^2}\iint_{\vec{k}_{//}}\Upsilon\left(\vec{k}_{//}\right)d\vec{k}_{//} = \frac{S}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Upsilon\left(k_x,k_y,k_z\right)dk_y dk_z$$

where *S* is the cross-section area of the crystal, in the (*y*,*z*)-plane. Eq. (11.21) then becomes:

Eq. (11.23) $$g_{2D}(E)=\frac{2S}{(2\pi)^2}\sum_n \iint_{\vec{k}_{//}}\delta\left[E_{//}(\vec{k}_{//})+E_{\perp n}-E\right]d\vec{k}_{//}$$

Now, we must determine the relation between $d\left[E_{//}(\vec{k}_{//})\right]$ as a function of $d\overrightarrow{k_{//}}$ in order to perform the integration in Eq. (11.23). For this, we follow the same analysis conducted in Eq. (4.35) to Eq. (4.39). Eq. (11.18) yields:

Eq. (11.24) $$d\left[E_{//}(\vec{k}_{//})\right]=\frac{\hbar^2}{2m^*}.(2k_{//})dk_{//}$$

where $k_{//}$ is the norm or length of the vector $\vec{k}_{//}$. On the other hand, in two dimensions, Eq. (4.37) becomes:

Eq. (11.25) $$d\vec{k}_{//}=d\left(\pi k_{//}{}^2\right)=2\pi k_{//}dk_{//}$$

Thus:

Eq. (11.26) $$d\left[E_{//}(\vec{k}_{//})\right]=\frac{\hbar^2}{2m^*}\frac{1}{\pi}d\vec{k}_{//}$$

and Eq. (11.23) becomes:

Eq. (11.27)
$$\begin{cases} g_{2D}(E)=\dfrac{S}{2\pi^2}\left(\dfrac{2m^*\pi}{\hbar^2}\right)\displaystyle\sum_n\int_0^\infty\delta\left[E_{//}(\vec{k}_{//})+E_{\perp n}-E\right]d\left[E_{//}(\vec{k}_{//})\right] \\ or \\ g_{2D}(E)=\dfrac{Sm^*}{\pi\hbar^2}\displaystyle\sum_n\int_0^\infty\delta\left[x+E_{\perp n}-E\right]dx \end{cases}$$

The integral will be zero if the argument of the Dirac function, i.e. $\left[x+E_{\perp n}-E\right]$, never reaches zero when the variable x is varied from 0 to $+\infty$. In other words:

Eq. (11.28)
$$\begin{cases} \int_0^\infty \delta\left[x+E_{\perp n}-E\right]dx = 0 \ \ if \left[E_{\perp n}-E\right]>0 \\ \int_0^\infty \delta\left[x+E_{\perp n}-E\right]dx = 1 \ \ if \left[E_{\perp n}-E\right]<0 \end{cases}$$

This can be best expressed by considering the step function which is defined as:

Eq. (11.29)
$$\begin{cases} \Theta(X)=0 & for\, x<0 \\ \Theta(X)=1 & for\, x>0 \end{cases}$$

Therefore, we can write:

Eq. (11.30)
$$\int_0^\infty \delta\left[x+E_{\perp n}-E\right]dx = \Theta\left[E-E_{\perp n}\right]$$

and Eq. (11.27) becomes:

Eq. (11.31)
$$g_{2D}(E)=\frac{Sm^*}{\pi\hbar^2}\sum_n \Theta\left[E-E_{\perp n}\right]$$

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 $\frac{Sm^*}{\pi\hbar^2}$ each time the energy E crosses an allowed value of $E_{\perp n}$, as shown in Fig. 11.7. 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. 11.7.

The modification of the density of states in a quantum well (2D) from that in the bulk case (3D), shown in Fig. 11.7, reflects the change in the motion of an electron. The in-plane motion is two-dimensional, which makes the density of states independent of energy in a subband. For the motion perpendicular to the well plane, we have a new quantum number n, introduced in Eq. (11.19), 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. 11.7.

It can be mathematically demonstrated that the density of states for two-dimensional and three-dimensional electrons do coincide at values of $E = E_{\perp n}$, as illustrated in Fig. 11.7, 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.

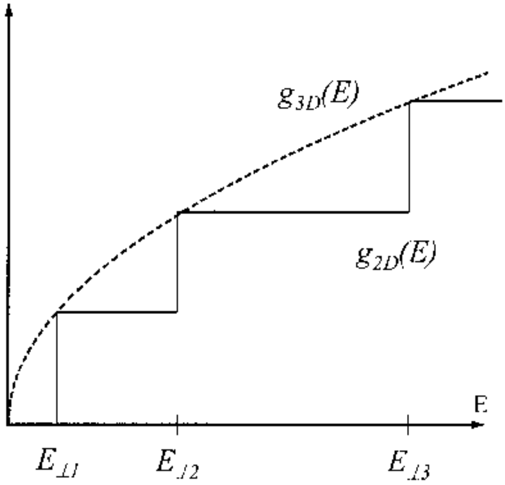


Fig. 11.7. Density of states in the conduction band in a quantum well (2D). The density of states is constant for values of energy between two consecutive quantized energy levels. For comparison, the density of states of a bulk material (3D) is shown in dashed lines.

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, 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 $\vec{k}_{//} = \vec{0}$):

$$g_{2D}(E) = \frac{Sm^*}{\pi\hbar^2}\sum_n \Theta\left[E - E_{\perp n}\right],$$ we obtain:

$$N = \int_{E_1}^{E_2} g_{2D}(E)dE = \frac{Sm^*}{\pi\hbar^2}\sum_n \Theta\left[E - E_{\perp n}\right]$$
$$= \frac{Sm^*}{\pi\hbar^2}(E_2 - E_1)$$
$$= \frac{\left(10^{-3}\right)^2\left(0.067 * 0.91095 \times 10^{-30}\right)}{\pi\left(1.05458 \times 10^{-34}\right)^2}\left(0.3 \times 1.60218 \times 10^{-19}\right)$$
$$\approx 8.40 \times 10^{10}$$

11.6.3. 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. 11.5. The effective mass of the electron traveling across the structure thus depends on position, $m^*(x)$. Two Schrödinger equations must then be considered:

Eq. (11.32)
$$\begin{cases} -\dfrac{\hbar^2}{2m_1*}\nabla^2\Psi(x,y,z) \quad - \quad [E - U(x,y,z)]\Psi(x,y,z) = 0 \\ \qquad\qquad for\, x < 0\, and\, x > a \\ -\dfrac{\hbar^2}{2m_2*}\nabla^2\Psi(x,y,z) \quad - \quad [E - U(x,y,z)]\Psi(x,y,z) = 0 \\ \qquad\qquad for\, 0 < x < a \end{cases}$$

The other important change concerns the boundary conditions outlined in Eq. (3.39). The continuity of the first derivative of the wavefunction $\dfrac{\partial\Psi(x,y,z)}{\partial x}$ is no longer valid, but must be replaced by the continuity of the product $\dfrac{1}{m*(x)}\dfrac{\partial\Psi(x,y,z)}{\partial x}$, which takes into account the spatial dependence of the electron effective mass. As a result, the boundary conditions in Eq. (3.39) must be replaced by:

Eq. (11.33)
$$\begin{cases} \dfrac{1}{m_1^*}\dfrac{\partial\Psi_-}{\partial x}(0)=\dfrac{1}{m_2^*}\dfrac{\partial\Psi_0}{\partial x}(0) \\ \qquad and \\ \dfrac{1}{m_2^*}\dfrac{\partial\Psi_0}{\partial x}(a)=\dfrac{1}{m_1^*}\dfrac{\partial\Psi_+}{\partial x}(a) \end{cases}$$

11.7. One-dimensional structures: quantum wires

11.7.1. 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:

Eq. (11.34) $$\Psi_{total}(x,y,z)=\Psi_z(z)\Psi_y(y)\Psi_x(x)$$

Only the wavefunction in the z-direction can be easily expressed as a plane wave:

Eq. (11.35) $$\Psi_z(z)=A\exp(ik_z.z)$$

where A is a normalization constant. The total energy is the sum of three components:

Eq. (11.36)
$$E(k_z, n, m) = E_z(k_z) + (E_x)_n + (E_y)_m = \frac{\hbar^2 k_z^{\ 2}}{2m^*} + \frac{\hbar^2 (k_x)_n^{\ 2}}{2m^*} + \frac{\hbar^2 (k_y)_m^{\ 2}}{2m^*}$$

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 sub-section 3.3.3.

The most important characteristic of a quantum wire is its electron density of states in the conduction band which is given by:

Eq. (11.37) $$g_{1D}(E) = 2 \sum_{n,m,k_z} \delta\left[E_z(k_z) + (E_x)_n + (E_y)_m - E\right]$$

In this one-dimensional case, we can make use of the quasi-continuous nature of k_z to write the identity:

Eq. (11.38) $$\sum_{k_z} \Upsilon(k_z) \equiv \frac{L}{2\pi} \int_{-\infty}^{+\infty} \Upsilon(k_z) dk_z$$

which allows us to simplify Eq. (11.37) into:

Eq. (11.39) $$g_{1D}(E) = \frac{2L}{2\pi} \sum_{n,m} \int_{-\infty}^{+\infty} \delta\left[E_z(k_z) + (E_x)_n + (E_y)_m - E\right] dk_z$$

where L is the length of the quantum wire. Moreover, in the one-dimensional case, we have:

Eq. (11.40) $$d\left[E_z(k_z)\right] = \frac{\hbar^2}{m^*} k_z dk_z = \frac{\hbar^2}{m^*} \sqrt{\frac{2m^* E_z(k_z)}{\hbar^2}} dk_z$$

Therefore, Eq. (11.39) becomes:

Eq. (11.41)

$$\begin{cases} g_{1D}(E)=\dfrac{L}{\pi}\sqrt{\dfrac{m^*}{2\hbar^2}}\displaystyle\sum_{n,m}\int_0^{\infty}\delta\left[E_z(k_z)+(E_x)_n+\left(E_y\right)_m-E\right]\frac{1}{\sqrt{E_z(k_z)}}dE_z \\ or \\ g_{1D}(E)=\dfrac{L\sqrt{m^*}}{\hbar\pi\sqrt{2}}\displaystyle\sum_{n,m}\int_0^{\infty}\delta\left[x+(E_x)_n+\left(E_y\right)_m-E\right]\frac{1}{\sqrt{x}}dx \end{cases}$$

Using Eq. (4.43) and the same argument as Eq. (11.31), we obtain:

Eq. (11.42) $$g_{1D}(E)=\frac{L}{\pi\hbar}\sqrt{\frac{m^*}{2}}\sum_{m,n}\frac{\Theta(E-\left[(E_x)_n+\left(E_y\right)_m\right])}{\sqrt{E-\left[(E_x)_n+\left(E_y\right)_m\right]}}$$

This expression means that, in a quantum wire, the density of states depends on the energy like $\frac{1}{\sqrt{E}}$ in each of the subband defined by two consecutive energy levels $(E_x)_n+\left(E_y\right)_m$, as shown in Fig. 11.8.

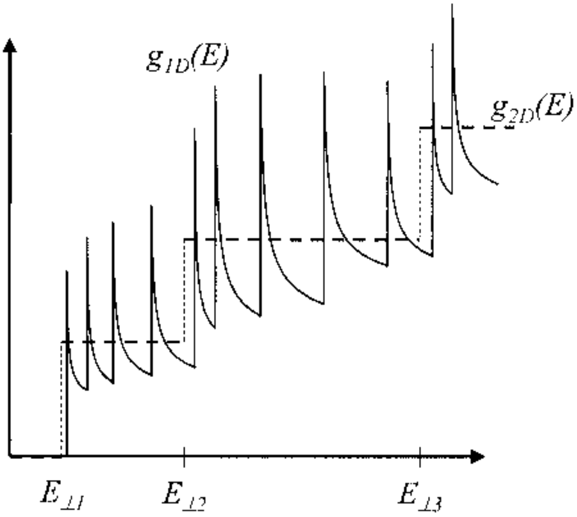


Fig. 11.8. Density of states in the conduction band for a quantum wire (1D). For comparison, the density of states of a quantum well (2D) is shown in dashed lines.

Eq. (11.42) also reveals infinite divergences at points where the energy E coincides with the bottoms of quasi-one-dimensional subbands at $(E_x)_n+\left(E_y\right)_m$. These discontinuities take place in an idealized model. In real structures, they are smeared out by the electron collisions mentioned earlier, in sub-section 11.6.1. The maximum values of g_{1D} in Fig. 11.8 are not infinite, but correspond to the value of Eq. (11.42) when the denominator is equal to $E-\left[\left(E_x\right)_n+\left(E_y\right)_m\right]\approx\frac{\hbar}{\tau}$, where τ is the electron lifetime discussed earlier.

11.7.2. 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. 11.9 and can be considered to be the two-dimensional analogy to the one-dimensional confinement potential of the standard infinitely deep quantum well.

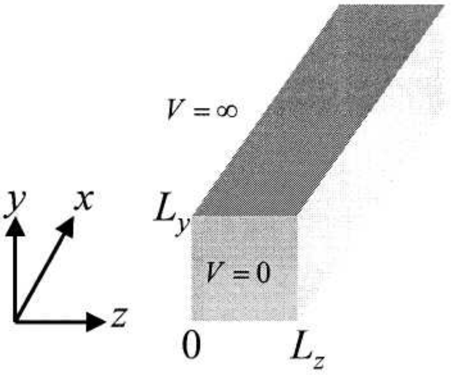


Fig. 11.9. The infinitely deep rectangular cross-section quantum wire.

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:

$$\text{Eq. (11.43)}\quad -\frac{\hbar^2}{2m^*}\left[\frac{\partial^2\Psi(y,z)}{\partial y^2}+\frac{\partial^2\Psi(y,z)}{\partial z^2}\right]=E_{y,z}\Psi(y,z)$$

The separation of the coordinates in the Schrödinger equation allows the motion to be decoupled further, and leads to:

Eq. (11.44) $\Psi(y,z)=\Psi(y)\Psi(z)$

and then the Schrödinger equation can be written as:

Eq. (11.45)
$$-\frac{\hbar^2}{2m^*}\Psi(z)\frac{\partial^2\Psi(y)}{\partial y^2}-\frac{\hbar^2}{2m^*}\Psi(y)\frac{\partial^2\Psi(z)}{\partial z^2}=(E_y+E_z)\Psi(y)\Psi(z)$$

Here the energy components can also be separated into $E_{y,z}=E_y+E_z$. The decoupling is completed with the following equations:

Eq. (11.46) $-\frac{\hbar^2}{2m^*}\frac{\partial^2\Psi(y)}{\partial y^2}=E_y\Psi(y)$

Eq. (11.47) $-\frac{\hbar^2}{2m^*}\frac{\partial^2\Psi(z)}{\partial z^2}=E_z\Psi(z)$

The above equations are exactly the same as the infinite quantum well problems (see sub-section 3.3.2). The wavefunction solutions are:

Eq. (11.48) $\Psi(y)=\sqrt{\frac{2}{L_y}}\sin\left(\frac{\pi n_y y}{L_y}\right)$

and

Eq. (11.49) $\Psi(z)=\sqrt{\frac{2}{L_z}}\sin\left(\frac{\pi n_z z}{L_z}\right)$

which give the components of the energy as:

Eq. (11.50) $E_y=\frac{\hbar^2\pi^2 n_y^2}{2m^* L_y^2}$

Eq. (11.51) $E_z=\frac{\hbar^2\pi^2 n_z^2}{2m^* L_z^2}$

Thus, the total energy of the particle due to the confinement is given by the sum of the two discrete components:

Eq. (11.52) $$E_{y,z} = \frac{\hbar^2\pi^2}{2m^*}\left(\frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2}\right)$$

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 Chapter 3.

11.8. Zero-dimensional structures: quantum dots

11.8.1. 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:

Eq. (11.53) $$\begin{aligned} E(n,m,l) &= (E_x)_n + (E_y)_m + (E_z)_l \\ &= \frac{\hbar^2 (k_x)_n^2}{2m^*} + \frac{\hbar^2 (k_y)_m^2}{2m^*} + \frac{\hbar^2 (k_z)_l^2}{2m^*} \end{aligned}$$

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 sub-section 3.3.3 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:

Eq. (11.54)
$$g_{0D}(E) = 2\sum_{n,m,l} \delta\left[E(n,m,l) - E\right] = 2\sum_{n,m,l} \delta\left[\left(E_x\right)_n + \left(E_y\right)_m + \left(E_z\right)_l - E\right]$$

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. 11.10.

Again, the divergences in the density of states shown in Fig. 11.10 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.

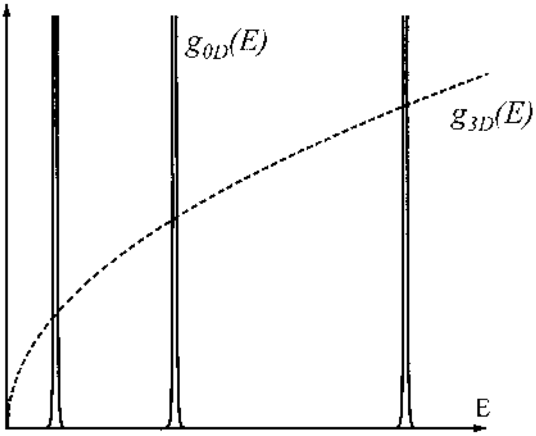


Fig. 11.10. Density of states in the conduction band for a quantum dot (0D). For comparison, the density of states of a bulk crystal (3D) is shown.

11.8.2. 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:

Eq. (11.55) $$\begin{cases} V(\vec{r}) = 0 & \text{if } r < R \\ V(\vec{r}) = \infty & \text{otherwise} \end{cases}$$

Due to the spherical symmetry of the potential, the Schrödinger-like equation for the envelope function $\Psi(\vec{r})$ in spherical coordinates is given as:

Eq. (11.56) $$\left[-\frac{\hbar^2}{2m^*}\left(\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)-\frac{\vec{L}^2}{r^2}\right)+V(\vec{r})\right]\Psi(\vec{r}) = E\Psi(\vec{r})$$

where $\vec{L}^2$ is the orbital momentum operator which commutes with the Hamiltonian. The solution to Eq. (11.56) 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:

Eq. (11.57) $$E_{nl} = \frac{\hbar^2}{2m^*}\left(\frac{\alpha_{nl}}{R}\right)^2, \quad n = 1,2,3..., \quad l = 0,1,2...$$
$$\Psi(r,\theta,\varphi) = Aj_l\left(\frac{\alpha_{nl}r}{R}\right)Y_{lm}(\theta,\varphi)$$

where A is a constant and 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 11.2 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 $-l$.

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

Table 11.2. Values of α_{nl} for the lowest states in a spherical well.

11.9. Optical properties of low-dimensional structures

Fig. 11.11 illustrates the band diagram in a GaAs-AlGaAs quantum well with several electron and hole subbands and the notations used in this section.

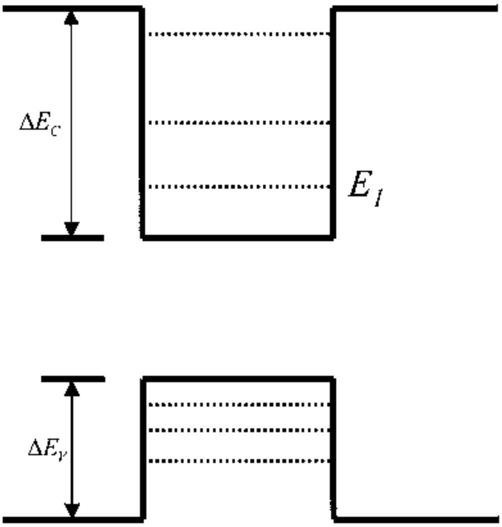


Fig. 11.11. Schematic of band diagram of GaAs-AlGaAs QW with electron and hole subbands.

11.9.1. Interband absorption coefficients of quantum wells

The absorption coefficient for a transition from a valence band state of energy E_l to a conduction band state of energy E_l has been given earlier in Chapter 10 and can be written as:

$$\text{Eq. (11.58)} \quad \alpha(\hbar\omega) = \frac{\pi q^2}{c m_0^2 \bar{n} \varepsilon_0 \omega V} \sum_{1,2} p_{12}(f_1 - f_2)\delta(E_{12} - \hbar\omega)$$

where:

$$\text{Eq. (11.59)} \quad p_{12} = \left\langle 1 \middle| \exp(-i\vec{k}_\lambda \cdot \vec{r})\vec{e}_\lambda \cdot \vec{p} \middle| 2 \right\rangle$$

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 m^{th} heavy hole subband with 2D wavevector $\vec{k}_h$ absorbs a photon and enters a state with wavevector $\vec{k}_e$ in the nth subband of the conduction band. In terms of the 2D vector $\vec{\rho}$ and the coordinate z normal to the QW layer plane, the wavefunctions are then written as:

$$\text{Eq. (11.60)} \quad \begin{aligned} |1\rangle &= \left|h, m, \vec{k}_h\right\rangle = U_h(\vec{\rho}, z)\exp(i\vec{k}_h \cdot \vec{\rho})\phi_{hm}(z), \\ |2\rangle &= \left|c, n, \vec{k}_e\right\rangle = U_c(\vec{\rho}, z)\exp(i\vec{k}_e \cdot \vec{\rho})\phi_{cn}(z), \end{aligned}$$

where U_h and U_c are the cell-periodic parts of the Bloch function and the $\phi's$ are envelope functions. We decompose the photon wavevector as $\vec{k}_\lambda = \left(\vec{k}_{\lambda //}, k_{\lambda z}\right)$ and write Eq. (11.59) as:

$$\text{Eq. (11.61)} \quad p_{12} = \left|\left\langle c, n, \vec{k}_e \middle| \exp(i\vec{k}_{\lambda //} \cdot \vec{\rho} + i k_{\lambda z} z)\vec{e}_\lambda \cdot \vec{p} \middle| h, m, \vec{k}_h \right\rangle\right|$$

The matrix element can be evaluated by using Eq. (11.60) 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 $\vec{k}_e = \vec{k}_h + \vec{k}_{\lambda //} \approx \vec{k}_h$, the squared matrix element can be written as:

Eq. (11.62) $$|p_{12}|^2 = \left\langle |p_{cv}|^2 \right\rangle_{QW} \delta_{\vec{k}_e,\vec{k}_h} C_{mn}$$

with

Eq. (11.63) $$C_{mn} = \left|\left\langle \phi_{hm} \middle| \phi_{cn} \right\rangle\right|^2 = \left| \int \phi^*_{hm} \phi_{cn} dz \right|^2$$

In the present case, $\left\langle |p_{cv}|^2 \right\rangle_{QW}$ is the polarization dependent momentum matrix element for transitions between conduction and valence subbands in a QW. It is different from the momentum matrix element in bulk semiconductors. The factor $\left\langle \phi_{hm} \middle| \phi_{cn} \right\rangle$ 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 $C_{mn} = \delta_{mn}$. 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:

Eq. (11.64)
$$\alpha(\hbar\omega) = \frac{\pi q^2}{c m_0^2 \bar{n} \varepsilon_0 \omega V} C_{mn} \sum_{\vec{k}_e,\vec{k}_h} \left\langle |p_{cv}|^2 \right\rangle_{QW} \delta_{\vec{k}_e,\vec{k}_h} (f_e - f_h) \delta(E_e(\vec{k}_e) - E_h(\vec{k}_h) - \hbar\omega)$$

Using the parabolic $E(\vec{k})$ relation, the energies are expressed as:

Eq. (11.65) $$\begin{cases} E_e(\vec{k}_e) = E_{cn} + \dfrac{\hbar^2 k_e^2}{2m_e} \\ E_h(\vec{k}_h) = -E_g - E_{hm} - \dfrac{\hbar^2 k_h^2}{2m_h} \end{cases}$$

The Fermi occupational probability can be written as:

Eq. (11.66) $$f(E)=\frac{1}{1+\exp\left[(E-E_f)/k_bT\right]}$$

where E_f is the quasi-Fermi level. Using $\vec{k}$ -conservation, the double summation in Eq. (11.64) is reduced to a single summation over $\vec{k}_h$. The argument of the energy-conserving δ -function then becomes:

Eq. (11.67) $$E_e(\vec{k}_e)-E_h(\vec{k}_h)-\hbar\omega=\left(E_g+E_{cn}+E_{hm}\right)-\frac{\hbar^2k_h^2}{2m_r}-\hbar\omega,$$

where m_r is the reduced mass (Eq. (10.80)). The remaining sum in Eq. (11.64) becomes:

Eq. (11.68)
$$\sum_{k_h}\rightarrow\frac{2S}{(2\pi)^2}\int\delta\left(E_g+E_{cn}+E_{hm}-\frac{\hbar^2k_h^2}{2m_r}-\hbar\omega\right)2\pi k_h dk_h\left[f_h(k_h)-f_e(k_h)\right]$$

where S is the area of the QW and the factor 2 is from spin degeneracy. The integration in Eq. (11.68) is performed easily due to the presence of the δ-function, so that we obtain:

Eq. (11.69)
$$\alpha(\hbar\omega)=\frac{m_rq^2C_{mn}\left\langle\left|p_{cv}\right|^2\right\rangle_{QW}}{\varepsilon_0\hbar^2cm_0^2\bar{n}\omega L}(f_e-f_h)H(\hbar\omega-E_g-E_{cn}-E_{hm})$$

where L is the thickness of QW and $H(x)$ is the Heaviside step function. Eq. (11.69) may be compared with the expression for bulk. Remember from Chapter 10 that $\left|p_{vc}\right|^2$ can be expressed and estimated in terms of the Kane-matrix elements (Eq. (10.74) and Appendix A.8). 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. 11.7. The experimental measurement of absorption coefficient in GaAs/AlGaAs quantum wells and thick GaAs layer are compared in Fig. 11.12.

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.

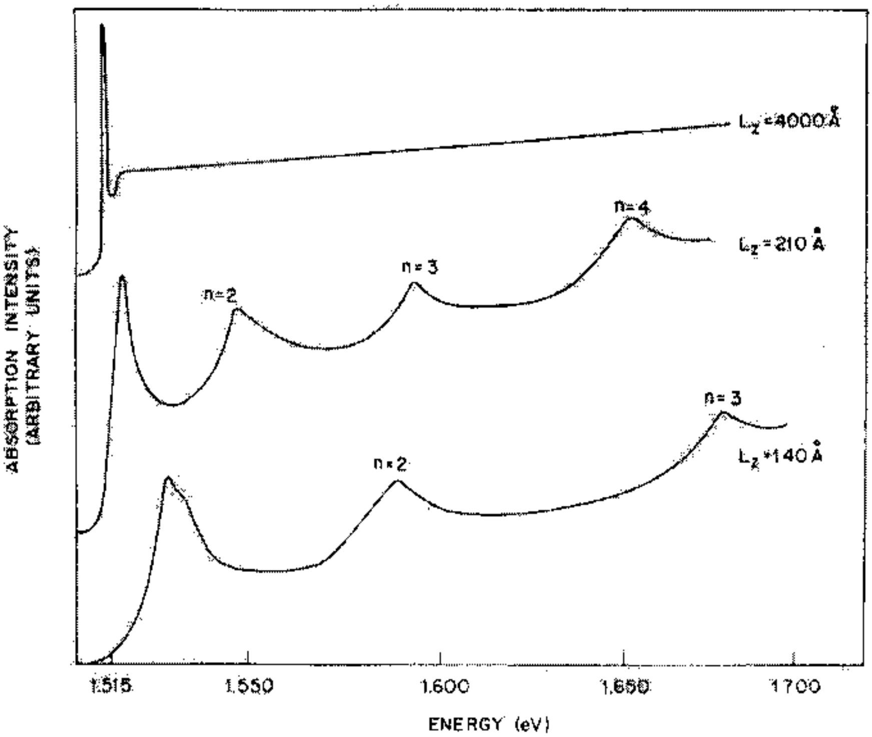


Fig. 11.12. Absorption coefficient in GaAs/AlGaAs quantum wells and thick GaAs layers (upper curve). The peaks correspond to quantum confined subbands n. [Reprinted figure with permission from Dingle, R., Wiegmann, W., and Henry, C.H., Physical Review Letters Vol. 33, p. 829, 1974. Copyright 1974 by the American Physical Society.]

11.9.2. 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:

Eq. (11.70) $$\alpha(\hbar\omega) = -B_1 \sum_{\vec{k}} (f_e - f_h)\delta\left(E_g + \hbar^2 k^2/2m_r - \hbar\omega\right)$$

where B_l is a constant, 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:

Eq. (11.71) $$\alpha(\hbar\omega, a) = -\frac{q^2 C_{1D} \left\langle |p_{cv}|^2 \right\rangle_{QWR} (2m_r)^{1/2}}{2m_0^2 \varepsilon_0 \bar{n} \hbar \omega c S} (\hbar\omega - E_g)$$

where the coefficient C_{1D} is the overlap integral of a quantum wire and $\left\langle |p_{cv}|^2 \right\rangle_{QWR}$ 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.

Eq. (11.71) 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. 11.8.

11.9.3. Absorption coefficient of quantum dots

The absorption coefficient of a cubic QD system of side length a may be written as:

Eq. (11.72) $$\alpha(\hbar\omega) = \frac{2\pi q^2 \left\langle |p_{cv}|^2 \right\rangle}{m_0^2 \bar{n} \varepsilon_0 c \omega a^3} \sum_m g(m^2) \delta(\hbar\omega - E_g - \pi^2 \hbar^2 m^2 / 2m_r a^2)$$

where $g(m^2)$ is the degeneracy of the energy level determined by m^2. Only $\Delta m = 0$ transitions are allowed. Eq. (11.72) 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:

Eq. (11.73) $$\hbar\omega = E_g + \pi^2 \hbar^2 m^2 / 2m_r a^2$$

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:

Eq. (11.74) $P(a)=\frac{1}{(2\pi)^{1/2}D}\exp\left[-(a-a_0)^2/2D^2\right]$

where a_0 is the average value and $D^2=\left\langle (a-a_0)^2\right\rangle$ is the standard deviation.

Using Eq. (11.72) and Eq. (11.74) the absorption coefficient for a non-uniform quantum dot system can be calculated as:

Eq. (11.75) $\alpha = \int_0^{\infty} P(a)\alpha(\hbar\omega,a)da$

The line broadening also occurs due to phonon scattering processes in addition to the size distribution of QDs.

11.10. 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 which will be described in more detail in Chapter 18. Such low-dimensional structures are fabricated in practice using a succession of processes involving epitaxy, lithography, and etching, which will all be discussed in Chapters 13 and 16. An illustration of the principle of quantum wells, wires, and dots is shown in Fig. 11.13.

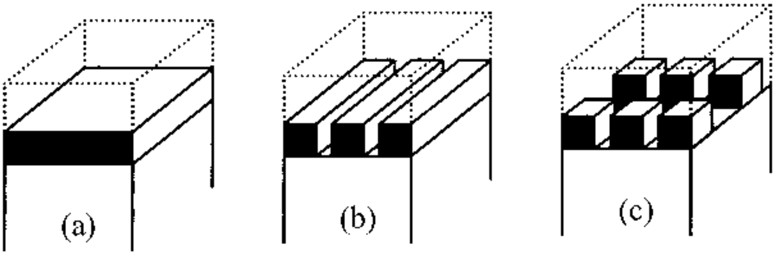


Fig. 11.13. Illustration of a: (a) 2D structure (quantum well), (b) 1D structure (quantum wires), and (c) 0D structure (quantum dots), showing the various levels of spatial confinement.

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 dependence devices (see Chapter 18 for a detailed discussion of these concepts). These 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. 11.14 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.

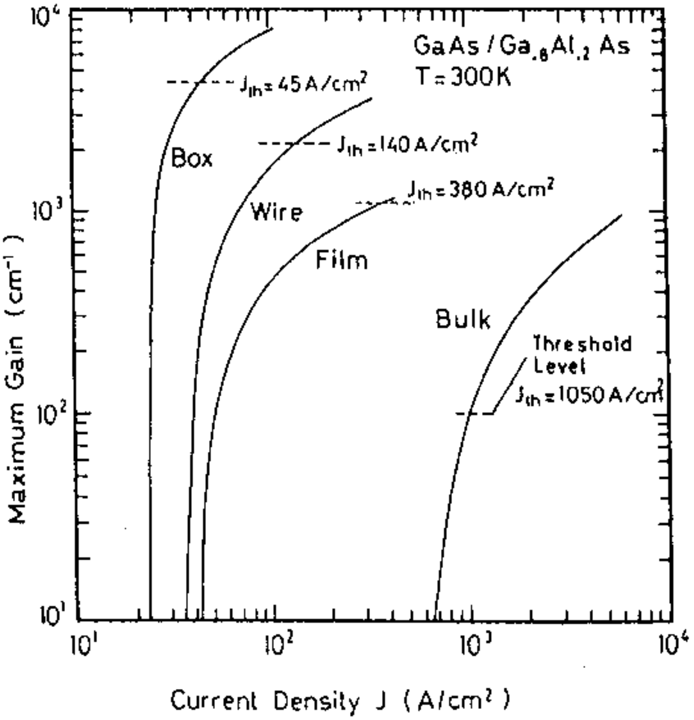


Fig. 11.14. Coefficient of light amplification (gain) for different structures. The dashed lines show the threshold current density above which laser emission starts. [Reprinted with permission from IEEE Journal of Quantum Electronics Vol. 22, Asada, M., Miyamoto, Y., and Suematsu, Y., "Gain and the threshold of 3-dimensional quantum-box lasers," p. 1918. Copyright 1986, IEEE.]

11.10.1. Quantum wires

Fig. 11.15 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.

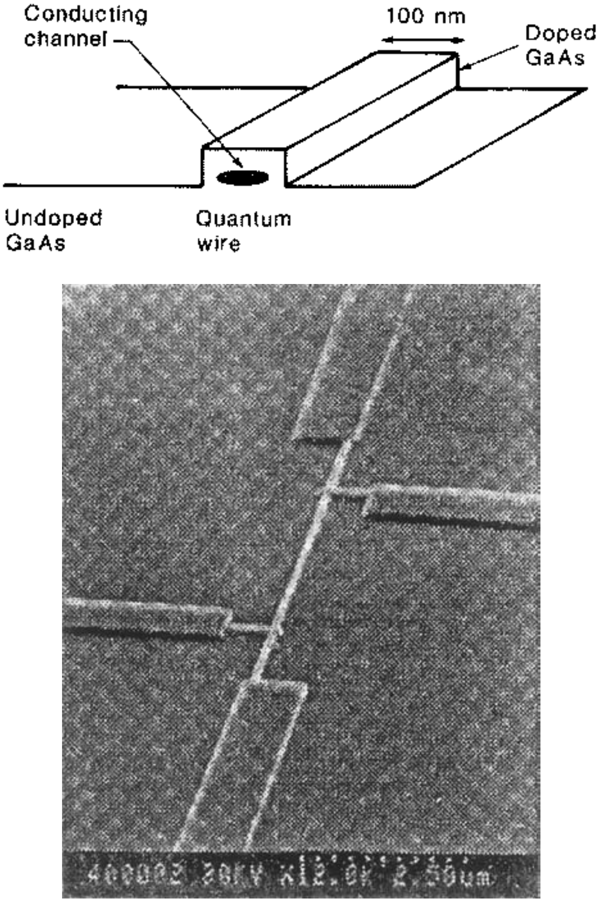


Fig. 11.15. Quantum wire formed by etching away all but a thin strip of doped semiconductor on an undoped substrate: (a) schematic diagram; (b) practical example. ["Figure 11.1", from LOW-DIMENSIONAL SEMICONDUCTORS: MATERIALS, PHYSICS, TECHNOLOGY, DEVICES by M.J. Kelly; taken after Physica Scripta Vol. T45, Beaumont, S.P., "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.]

Another example of quantum wire is shown in Fig. 11.16. 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.

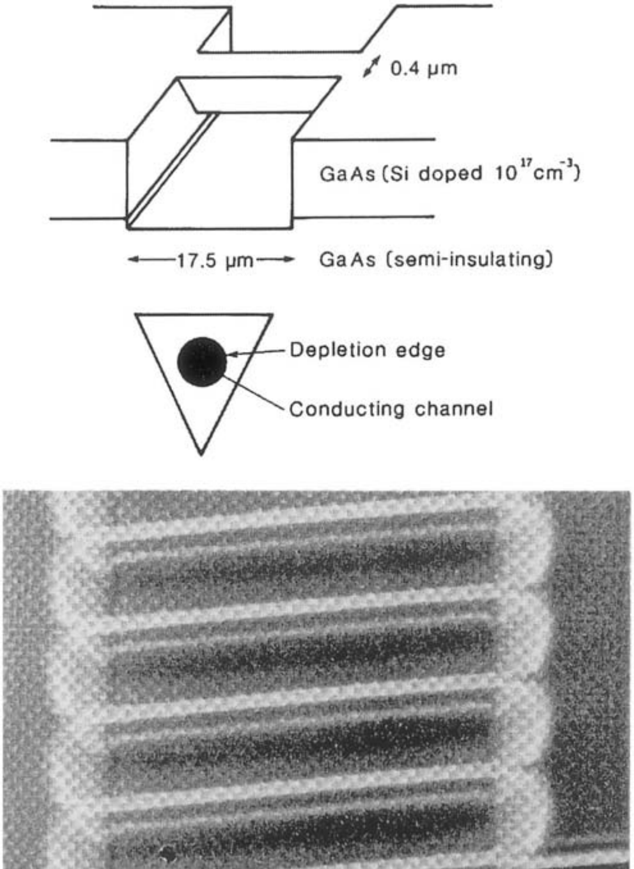


Fig. 11.16. Schematic diagram and image of quantum wires of doped GaAs on an insulating substrate. ["Figure 11.2", from LOW-DIMENSIONAL SEMICONDUCTORS: MATERIALS, PHYSICS, TECHNOLOGY, DEVICES by M.J. Kelly; taken after Journal of Vacuum Science and Technology B Vol. 6, Hasko, D.G., Potts, A., Cleaver, J.R.A., Smith, C.G., 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. Reprinted with permission of Oxford University Press, Inc. and American Institute of Physics.]

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 Chapter 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 1999], and exhibits exciting new science which has been studied only very recently. When moving along a "line", carriers are more likely to be affected by each others' 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. 11.17, in doped (multilayer) semiconductors is more complicated. The processing steps are shown in Fig. 11.17(c) and (d) which result in the structure shown in Fig. 11.17(b). A sub-micrometer diameter metal dot is laid down onto the film (step 2 in Fig. 11.17(c)), 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 & 2 in Fig. 11.17(d)). The whole surface can then be coated with metal (steps 3 & 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.

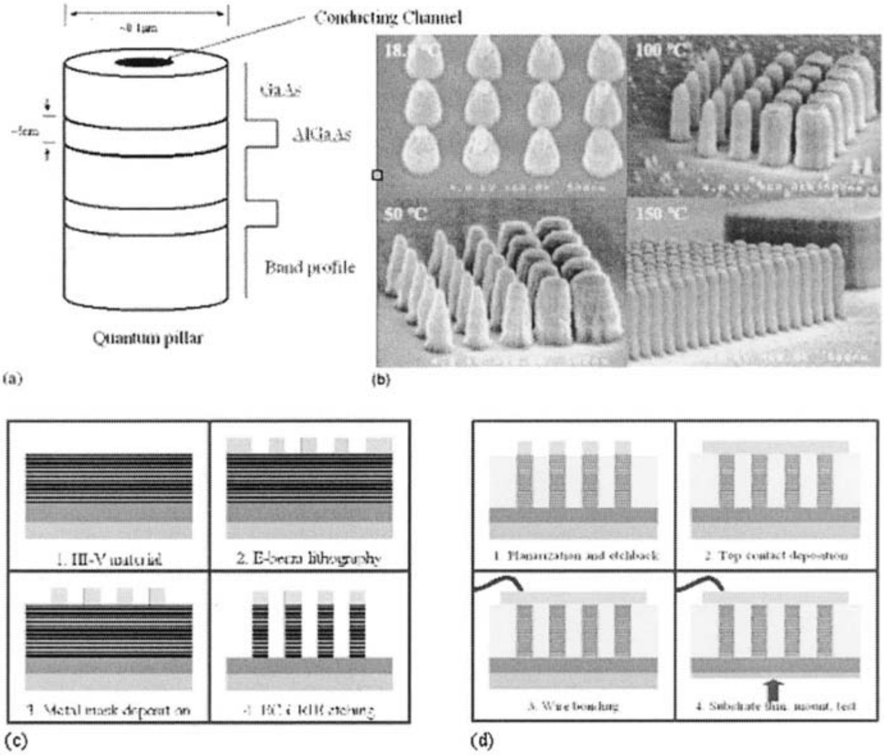


Fig. 11.17. A quantum pillar formed from resonant tunneling semiconductor multilayers showing (a) a schematic diagram of the pillar, (b) the partially processed structure after the first etch, and (c) and (d) the full processing route.

11.10.2. Quantum dots

The structure shown in Fig. 11.17 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. 11.17(a)) 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 atomic layer, and after a certain critical thickness is reached, nanometer size islands spontaneously forms. This is known as the Stranski-Krastanow 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 ultra sharp 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.

11.10.3. 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 Chapter 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. We write for the new ground state, i.e. for the wavefunction and energy of the perturbed system the perturbation expansions:

Eq. (11.76)
$$\begin{aligned}\Phi_g &= \Phi_g^0 + \Phi_g^1 + \Phi_g^{(2)}\\ E_g &= E_g^0 + E_g^{(1)} + E_g^{(2)}\end{aligned}$$

The admixtures are as discussed before in section 3.2.1, i.e. linear combinations of excited states so that

Eq. (11.77)
$$\Phi_g^{(n)} = \sum_{l\neq g} a_{lg}^{(n)} \Phi_l^0$$

and where the time independent Schrödinger equation with perturbation V is given by:

Eq. (11.78) $(H_0 + V)\Phi_g = E_g \Phi_g$

Substituting Eq. (11.76) in Eq. (11.78), and comparing the coefficients of the same order then gives to zeroth order in the potential:

Eq. (11.79) $H_0 \Phi_g^0 = E_g^0 \Phi_g^0$

First order in the perturbing potential:

Eq. (11.80) $H_0 \Phi_g^1 + V\Phi_g^0 = E_g^0 \Phi_g^1 + E_g^1 \Phi_g^0$

Second order in the perturbing potential:

Eq. (11.81) $H_0 \Phi_g^{(2)} + V\Phi_g^1 = E_g^0 \Phi_g^{(2)} + E_g^1 \Phi_g^1 + E_g^{(2)} \Phi_g^0$

In order to obtain the zeroth order solution we substitute the expansion Eq. (11.77) into the first order equation Eq. (11.80), multiply the left hand on both side by $(\Phi_g^0)^*$, integrate over all space, and use the orthogonality condition:

Eq. (11.82) $\int d\vec{r} \Phi^*{}_n \Phi_m = \delta_{mn}$

to obtain the first order term of the energy expansion of the perturbed system (see Eq. (11.76)):

Eq. (11.83) $E_g^1 = \int d\vec{r} (\Phi_g^0)^* V(\vec{r}) \Phi_g^0$

We carry on the procedure to calculate the first order change in the wavefunction by multiplying Eq. (11.80) this time on both sides with $(\Phi_l^0)^*$ and integrating while using the orthogonality again and the relation:

Eq. (11.84) $H_0 \Phi_m^0 = E_m^0 \Phi_m^0$

to find the first coefficient of the wavefunction expansion:

Eq. (11.85) $a_{lg}^{(1)} = \dfrac{V_{lg}}{E_g - E_l}$

After multiplying and integrating again with $(\Phi_g^0)^*$ on the second order equation given by Eq. (11.81) and after doing some algebra, we find the second order energy term:

$$\text{Eq. (11.86)}\quad E_g^{(2)} = \sum_{l\neq g} \frac{V_{gl}V_{\mathrm{lg}}}{E_g - E_l}$$

The wavefunction is given to first order by:

$$\text{Eq. (11.87)}\quad \Phi_g = \Phi_g^0 + \sum_{l\neq g} \frac{V_{\mathrm{lg}}}{E_g - E_l}\Phi_l^0$$

where as before the matrix element of the potential is defined by:

$$\text{Eq. (11.88)}\quad V_{ls} = \int \Phi *_s^0 V(\vec{r})\Phi_l^0 d\vec{r}$$

Knowing the unperturbed wavefunctions and energy levels, allows us to compute the perturbed ones. This also gives a straightforward rule to obtain the new permittivities of the perturbed system. We do this by substituting the new wavefunctions and energies into Eq. (10.53). Thus if the perturbation is due to an applied field in the *z*-direction, i.e. in the quantum well growth direction, then $V = -qzE_0^z$, and we can compute the result to a good approximation with the box wavefunctions of Chapter 3. In this case, by symmetry it follows that $V_{gg} = 0$, and we have only the second order term which can be related to a sum of terms involving the oscillator strength (Chapter 10) and which is proportional to the squared of the applied field. The usual representation of the second order energy term is in the form:

$$\text{Eq. (11.89)}\quad E_g^{(2)} = \sum_{l} (qE_{0z})^2 \frac{\left|z_{gl}\right|^2}{E_g - E_l}$$

In a confined system, the electric field induced shift of the energy of the free subband eigenstates is called the Stark shift and is a lowering of energy when we start with box eigenstates. Fig. 11.18 shows the absorption spectra of a quantum well in 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, specially when the exciton is broken up by the field, and then the simple method might not suffice.

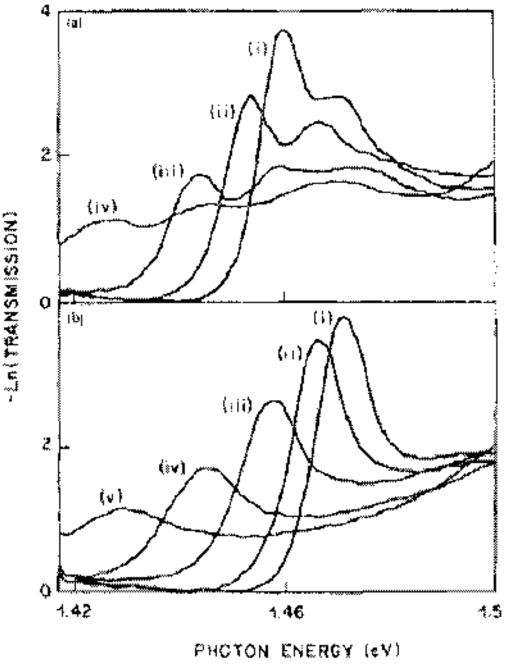


Fig. 11.18. Electroabsorption spectra of GaAs quantum well waveguide device as a function of electric field applied field perpendicular to the plane of the layers. (i)=1.6×10^4 V.cm^{-1}; (ii)=10^5 V.cm^{-1}; (iii)=1.4×10^5 V.cm^{-1}; (iv)=1.8×10^5 V.cm^{-1}; (v)=2.2×10^5 V.cm^{-1}. [Reprinted with permission from Applied Physics Letters Vol. 47, Weiner J.S., Miller D.A.B., Chemla D.J., Damen T.C., Burrus C.A., Wood T. H., Gossard A.C., Wiegmann W., "Strong polarization sensitive electroabsorption in GaAs/AlGaAs quantum well waveguides," p. 1149. Copyright 1985, American Institute of Physics.]

Fig. 11.19 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.

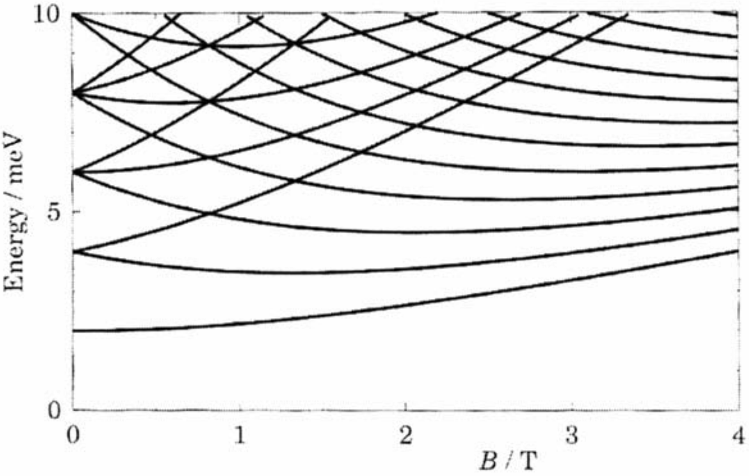


Fig. 11.19. The energy levels of a parabolically confined quantum dot with intrinsic energy level splitting $\hbar\omega_0 = 2meV$ in a magnetic field. [Davies, J.H., The Physics of Low Dimensional Semiconductors: an Introduction, p. 237, Fig. 6.16. © Cambridge University Press 1998. Reprinted with the permission of Cambridge University Press.]

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 Chapter 10, and from Eq. (10.113) the perturbation is of the form (m^* is the effective mass):

Eq. (11.90) $$V = [(qBx)^2 - 2qBx(i\hbar\frac{\partial}{\partial y})]\frac{1}{2m^*}$$

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. 11.20 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 Chapter 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 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. 11.20. 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 orbital radii of the localized states are smaller than the cyclotron radius, and their energies are not sensitive to the *B* field. 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 semi classical description, the Hall voltage exists because the Lorentz force creates an asymmetric charge redistribution for drifting carriers.

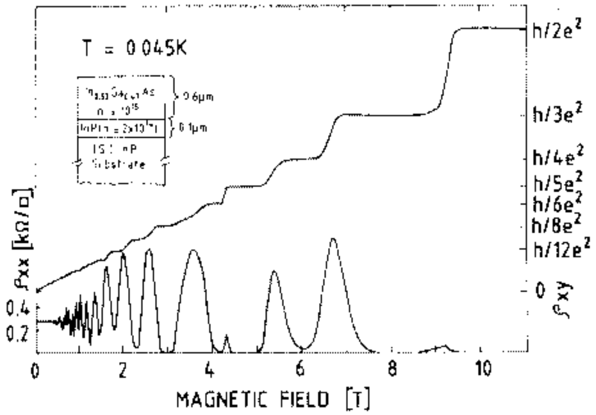


Fig. 11.20. Shubnikov de-Haas trace (ρ_{xx}) and quantum Hall effect (ρ_{xy}) as a function of magnetic field normal to the plane at T= 0.045 K in $Ga_{0.47}In_{0.53}As$-InP heterostructures. [Reprinted with permission from Applied Physics Letters Vol. 48, Razeghi, M., Duchemin, J.P., Portal, J.C., Dmowski, L., Remeni, G., Nicolas, R.J., 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. 713. Copyright 1986, American Institute of Physics.]

11.11. 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 $\Delta E_{n,n+1} \sim k_b T$, 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 Chapter 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 Chapter 10 combined with the perturbation expansion given here.

References

Asada, M., Miyamoto, Y., and Suematsu, Y., "Gain and the threshold of 3-dimensional quantum-box lasers," *IEEE Journal of Quantum Electronics* 22, pp. 1915-1921, 1986.

Bockrath, M, Cobden, D.H., Lu, J., Rinzler, A.G., Smalley, R.E., Balents, L., and McEuen, P.L, "Luttinger liquid behaviour in carbon nanotubes," *Letters to Nature* 397, pp. 598-601, 1999.

Kelly, M.J., *Low-Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Oxford University Press, New York, 1995.

Dingle, R., Wiegmann, W., and Henry, C.H., "Quantum states of confined carriers in very thin $Al_xGa_{1-x}As$-GaAs- $Al_xGa_{1-x}As$ heterostructures," *Physical Review Letters* 33, pp. 827-830, 1974.

Mohseni, H., *Type-II InAs/GaSb superlattices for infrared detectors*, Thesis, Ph.D. dissertation, Northwestern University, 2001.

Razeghi, M., Duchemin, J.P., Portal, J.C., Dmowski, L., Remeni, G., Nicolas, R.J., 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," *Applied Physics Letters* 48, pp. 712-715, 1986.

Shockley, W., U.S. Patent 2,569,347, 1951.

Weiner, J.S., Miller, D.A.B., Chemla, D.J., Damen, T.C., Burrus, C.A., Wood, T.H., Gossard, A.C., and Wiegmann, W., "Strong polarization sensitive electroabsorption in GaAs/AlGaAs quantum well waveguides," *Applied Physics Letters* 47, pp. 1148-1151, 1985.

Further reading

Ahmed, H., "An integration microfabrication system for low dimensional structures and devices," in *The Physics and Fabrication of Microstructures and Microdevices*, eds. M.J. Kelly and C. Weisbuch, Springer-Verlag, Berlin, pp. 435-442, 1986.

Ashcroft, N.W. and Mermin, N.D., *Solid State Physics*, Holt, Rinehart, Winston, New York, 1976.

Bassani, F. and Pastori Parravicini, G., *Electronic States and Optical Transitions in Solids*, Chap. 6, Pergamon, New York, 1975.

Bastard, G., *Wave Mechanics Applied to Semiconductor Heterostructures*, Halsted Press, New York, 1988.

Beaumont, S.P., "Quantum wires and dots-defect related effects," *Physica Scripta* T45, pp. 196-199, 1992.

Chuang, S.L., *Physics of Optoelectronic Devices*, John Wiley & Sons, New York, 1995.

Davies, J.H., *The Physics of Low Dimensional Semiconductors: an Introduction*, Cambridge University Press, New York, 1998

Dingle, R., "Confined carrier quantum states in ultrathin semiconductor heterostructures," in *Feskorperproblem* XV, ed. H.J. Queisser, pp. 21-48.

Einspruch, N.G., and Frensley, W.R., *Heterostructures and Quantum Devices*, Academic Press Limited, London, 1994.

Hasko, D.G., Potts, A., Cleaver, J.R., Smith, C., and Ahmed, H., "*Fabrication of sub-micrometer free standing single crystal GaAs and Si structures for quantum transport studies*," *Journal of Vacuum Science and Technology* B. 6, pp. 1849-1851, 1988.

Rosencher, E., Vinter, B., *Optoelectronics*, Cambridge University Press, Cambridge, 2002.

Scherer, A., Jewell, J., Lee, Y.H., Harbison, J., and Florez, L.T., "*Fabrication of microlasers and microresonator optical switches*," Applied Physics Letters 55, pp. 2724-2726, 1989.

Sze, S., *Physics of Semiconductor Devices*, 2nd Edition, John Wiley & Sons, New York, 1981.

Tewordt, M., Law, V., Kelly, M., Newbury, R., Pepper, M., and Peacock, C., "*Direct experimental determination of the tunneling time and transmission probability of electrons through a resonant tunneling system*," Journal of Physics-Condensed Matter 2, pp. 896-899, 1990.

Vasko, F.T. and Kuznetsov, A.V., *Electronic States and Optical Transitions in Semiconductor Heterostructures*, Springer, New York, 1999.

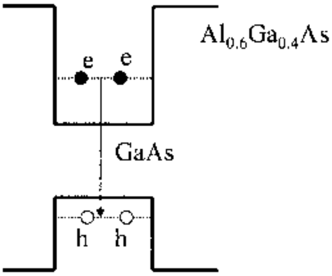
Weisbuch, C. and Vinter, B., *Quantum Semiconductor Structures*, Academic Press, New York, 1991.

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 Chapter 3.

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 $E_1 = \dfrac{\hbar^2\pi^2}{2m^*a^2}$, 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: $E_{nl} = \dfrac{\hbar^2}{2m^*}\left[\dfrac{\alpha_{nl}}{R}\right]^2$, where m^* is the effective mass of the electron or

hole, and the value of α_{nl} is given by α_{10}=π, α_{11}=4.49, α_{12}=5.76, α_{20}=6.28, α_{21}=7.72, α_{22}=9.09. Now consider very small GaAs quantum dots of radius 10nm. 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, $E_F = 3E_1^A$ 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.

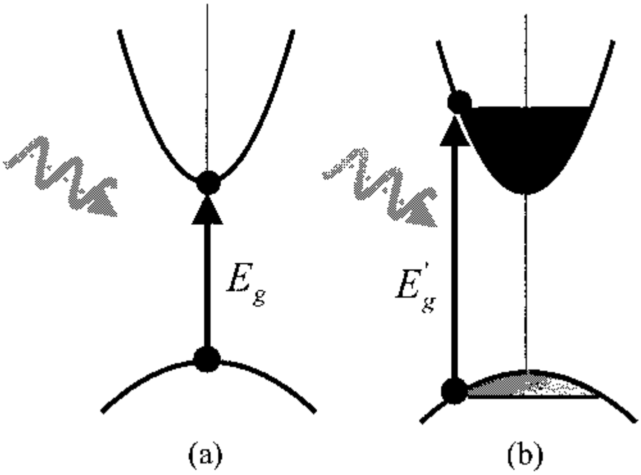
7. *The graphic of the two-dimensional density of states.*
Fig. 11.7 shows the density of states of a quantum well. The confinement energy of the lowest level (E_1) is 17 meV and the first excited state (E_2) has a confinement energy of 30 meV. The Fermi level is located 50 meV above the bottom of the conduction band. Determine the number of electrons contained in the well.

8. *The 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 $m_e^* = 0.067m_0$ and $m_h^* = 0.45m_0$ 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.

9. 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.

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), 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. (11.56) 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.

12. Compound Semiconductors and Crystal Growth Techniques

12.1. 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 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.

12.2. III-V semiconductor alloys

12.2.1. 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 12.1 lists some of the fundamental physical parameters of common binary III-V compounds.

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
AlSb	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	-	-
AlAs	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	-
AlP	14	5.451	2.45	3.027	-	0.63	0.20	-	-
AlN	10	a=3.112 c=4.982	6.2	2.2	-	-	-	9.14	-

Table 12.1. Physical constants of some III-V binary compounds at 300 K.

These binary compounds are the simplest III-V compounds, and constitute the basis for more complex ternary or quaternary compounds.

12.2.2. 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:

Eq. (12.1) $$E_g(x) = E_g(0) + bx + cx^2$$

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 12.2 [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}	$E_g(x) = 0.36 + 0.891x + 0.101x^2$
$InAs_xSb_{1-x}$	$E_g(x) = 0.18 - 0.41x + 0.58x^2$

Table 12.2. Compositional dependence of the bandgap energy in some III-V ternary compound semiconductors at 300 K. [Casey and Panish 1978.]

The bowing parameter *c* can be theoretically determined [Van Vechten *et al.* 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:

Eq. (12.2) $$a_{A_xB_{1-x}C} = xa_{AC} + (1-x)a_{BC}$$

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.

12.2.3. 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 *et al.* [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 *et al.* [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:

Eq. (12.3) $$a_{A_xB_{1-x}C_yD_{1-y}} = xya_{AC} + x(1-y)a_{AD} + (1-x)ya_{BC} + (1-x)(1-y)a_{BD}$$

The quaternary III-V alloys which can be used for multilayer heterostructures are listed in Table 12.3 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:

Eq. (12.4) $$E_g = xyE_{AC} + x(1-y)E_{AD} + (1-x)yE_{BC} + (1-x)(1-y)E_{BD}$$

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

Table 12.3. Binary to quaternary III-V lattice-matched systems of multilayer heterostructures. [Casey and Panish 1978.]

By using advanced epitaxial growth techniques, such as the ones discussed in section 12.5, multilayer structures of compounds with different bandgap associated wavelengths can be synthesized.

Fig. 12.1 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 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. 12.1. 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.

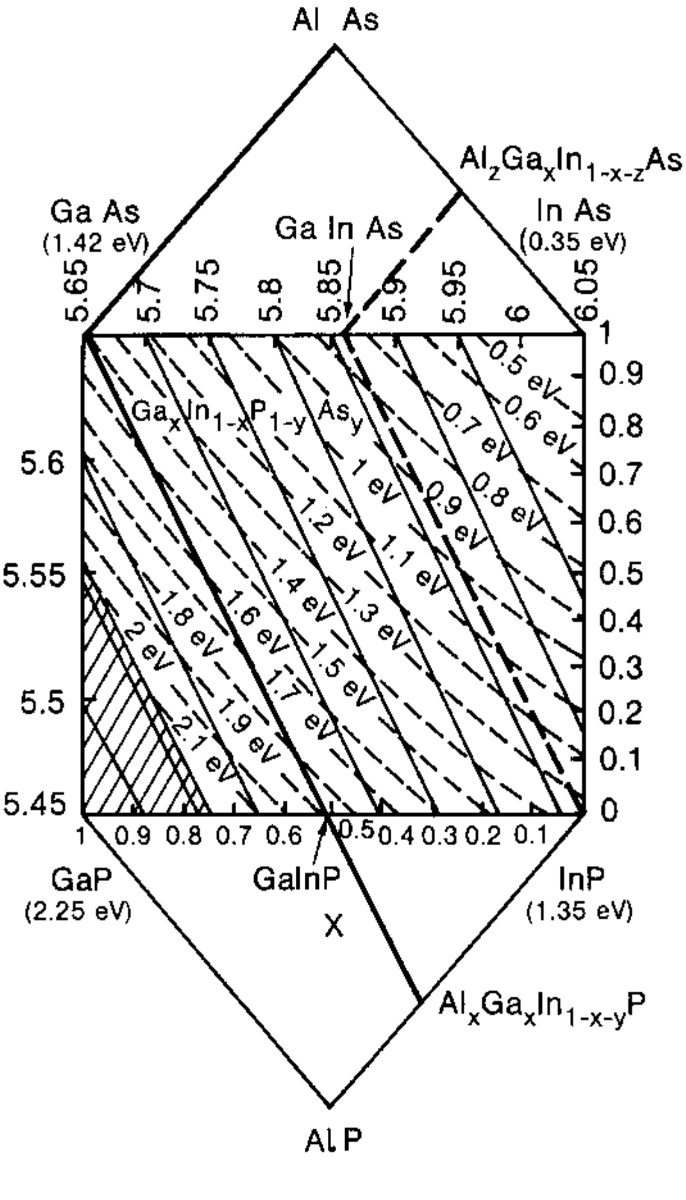


Fig. 12.1. The x-y-z compositional plane for quaternaries III-V alloys at 300 K. The solid lines in the square center region represent the x-y coordinates for which the quaternary alloy has a constant lattice parameter, while the curved dashed lines represent the x-y coordinates for which the alloy has 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 Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

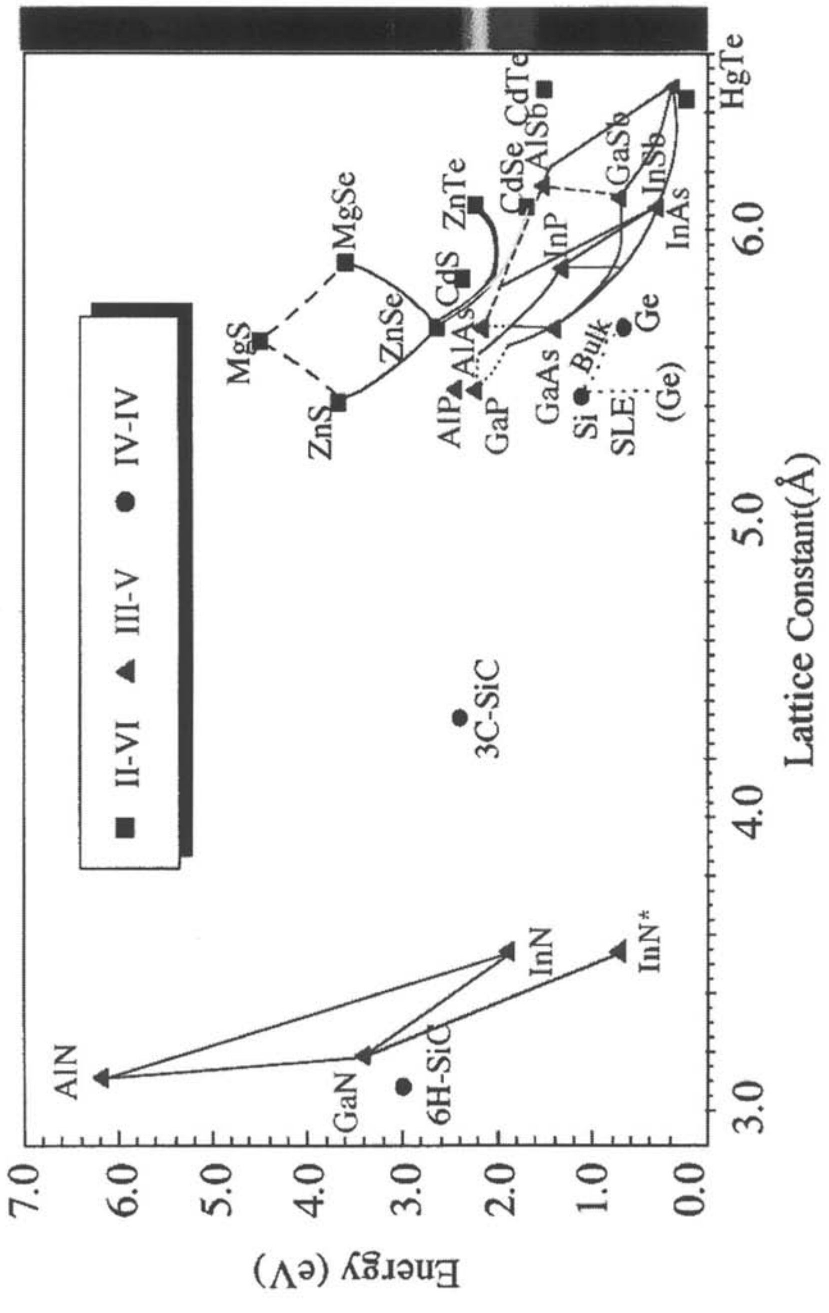


Fig. 12.2. Bandgap energy vs. lattice constant diagram of common semiconductors. A dashed line indicates an indirect bandgap material.
**(The bandgap energy of pure InN has been found to be 0.7 eV which is much smaller than the previously reported value of 1.9 eV).*

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. 12.2. The 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.

12.3. 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. lists some of the II-VI compound semiconductors and their respective bandgap energies, crystalline structures, and the rate of change of their bandgap energy as a function of temperature [Ray 1969].

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	-

Table 12.4. 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]

12.4. 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 sub-sections.

12.4.1. 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. 12.3 is 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. 12.3). 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.

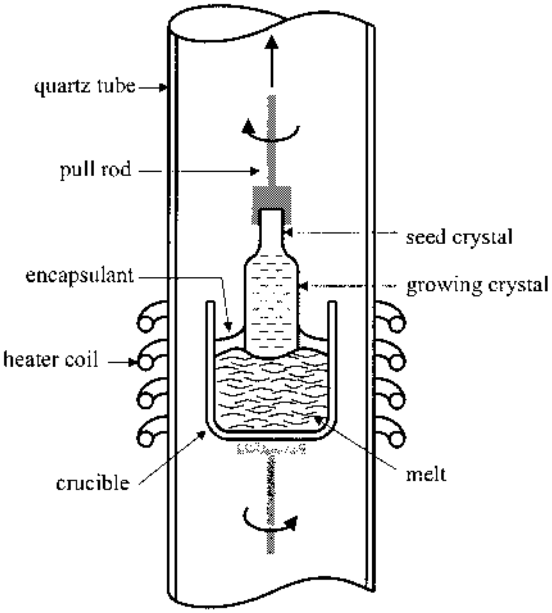


Fig. 12.3. Cross-section of a furnace used for the growth of single-crystal semiconductor boules by the Czochralski process, in which a tiny single crystal is suspended in a pool of hot molten material, and is slowly drawn upward as the crystal grows from the melt. The resulting boule can have a diameter over 30 cm and a length up to 2 m.

Since both the molten semiconductor and the solid are at the same pressure and have approximately the same composition, crystallization results due 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:

Eq. (12.5) $$\left(-k_l A \frac{dT}{dx}\Big|_l\right) - \left(-k_s A \frac{dT}{dx}\Big|_s\right) = L\frac{dm}{dt}$$

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:

Eq. (12.6) $$v_{\max} = \frac{kA}{L}\frac{dT}{dm} = \frac{k}{M_V L}\frac{dT}{dx}\Big|_s$$

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, 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:

Eq. (12.7) $$k = \frac{C_s}{C_l}$$

where C_l and C_s are the impurity concentrations in the liquid and solid sides of the liquid/solid interface, respectively. Table 12.5 lists the values of the segregation constant for some common impurities in silicon.

Impurity	*k*
Al	0.002
As	0.3
B	0.8
O	0.25
P	0.35
Sb	0.023

Table 12.5. Segregation constants for a few 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:

Eq. (12.8) $$C_s = kC_o(1-X)^{k\ 1}$$

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 10000 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.

12.4.2. 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. 12.4.

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.

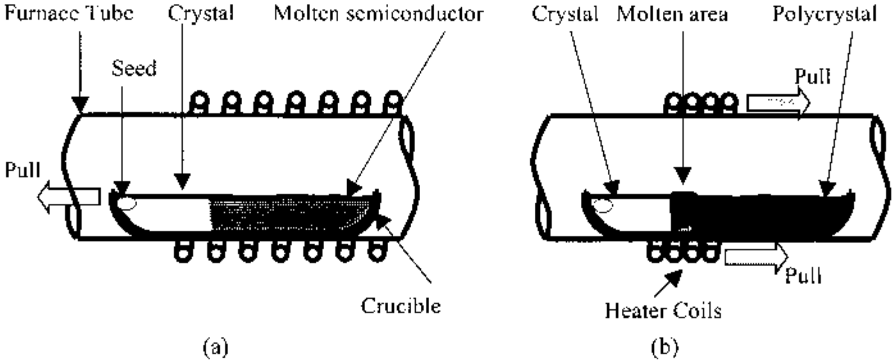


Fig. 12.4. Bridgman growth method in a crucible (a) solidification from one end of the melt (b) melting and solidification in a moving heated zone.

In the case of compound semiconductors, the process is slightly different from that for silicon. The basic process is shown in Fig. 12.5 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.

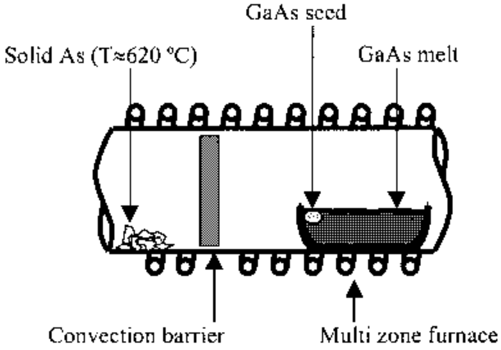


Fig. 12.5. Schematic diagram of the Bridgman growth method for a compound semiconductor such as gallium arsenide.

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.

12.4.3. 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 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. 12.6. 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.

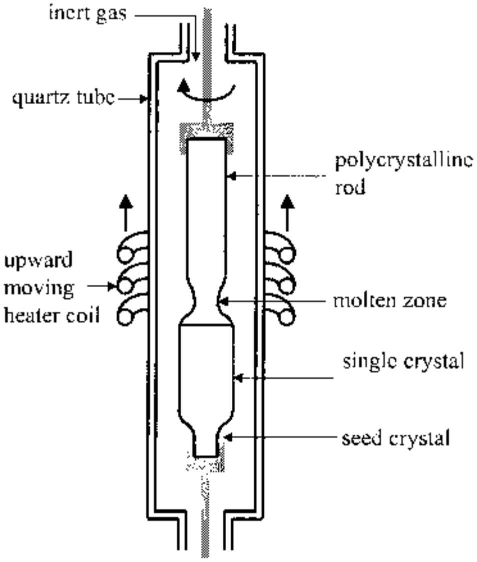


Fig. 12.6. Cross-section of a furnace used for the growth of single-crystal semiconductor boules by the float-zone process.

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:

Eq. (12.9) $$C(z) = C_c \left[\frac{r_d}{r_f} \right] \left[1 - (1-k) e^{-kz/l} \right]$$

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 12.6.

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.

Impurity	k
B	0.9
P	0.5
Sb	0.07

Table 12.6. Distribution coefficients for float-zone growth.

12.4.4. 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 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. 12.7. 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.

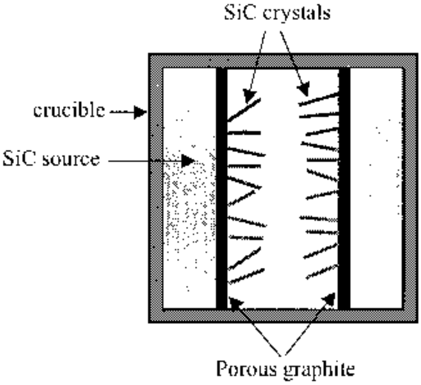


Fig. 12.7. Schematic cross-section diagram of a cylindrical crucible used for the Lely growth of SiC.

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. 12.8. 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.

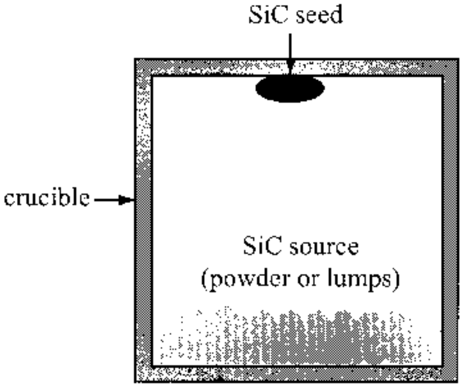


Fig. 12.8. Schematic cross-section diagram of a cylindrical crucible used for the Lely growth of SiC.

Although the Modified Lely method is more than twenty 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 non-uniform, and the resulting crystals still have high density of defects, such as micropipes and dislocations.

12.4.5. 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. Fig. 12.9 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.

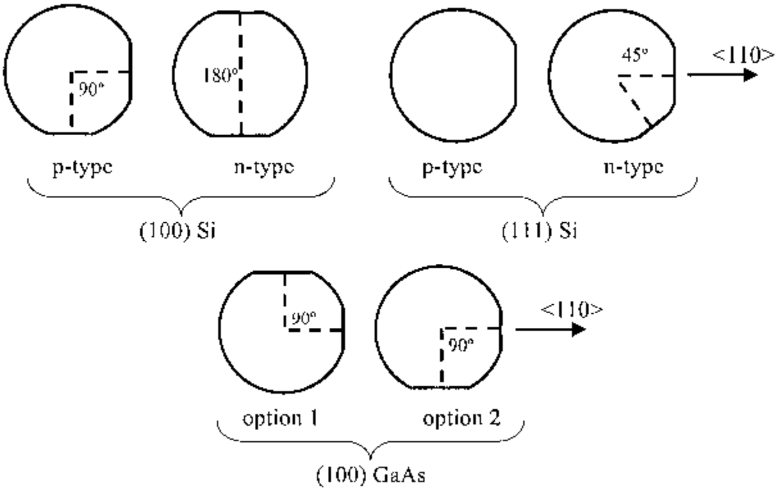


Fig. 12.9. Standard flat orientations for various types of semiconductor wafers. The longer flat is called the primary flat, whereas the shorter one is referred to as the secondary flat.

12.5. 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 sub-sections.

12.5.1. Liquid phase epitaxy

The LPE growth technique uses a system shown in Fig. 12.10 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. 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 effect 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.

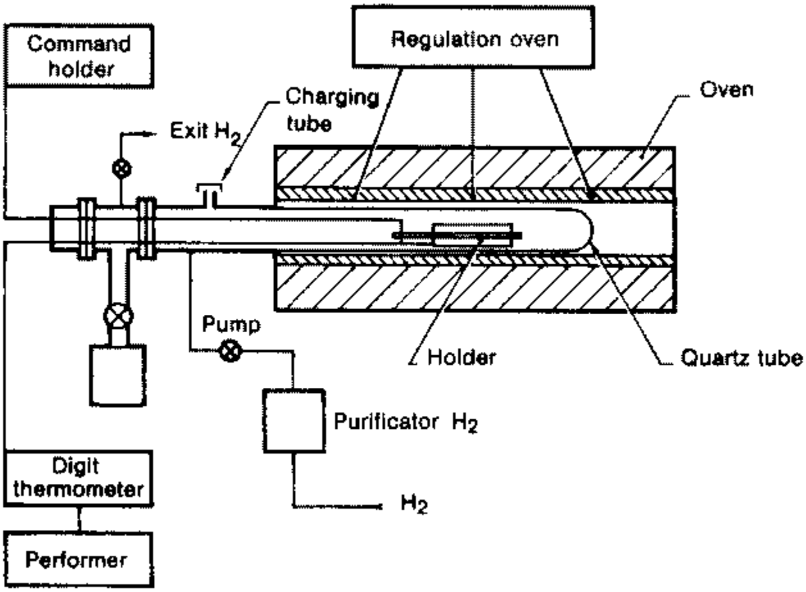


Fig. 12.10. Cross-section of a liquid phase epitaxy system. Inside the horizontal furnace, there is a sliding graphite boat upon which a substrate is held. [Copyright © 1989 From The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

12.5.2. 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. 12.11.

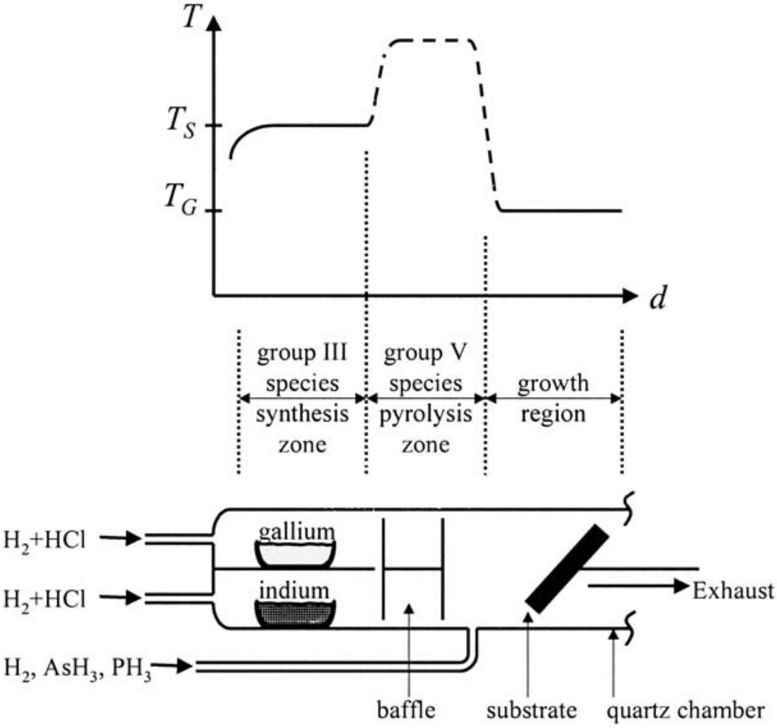


Fig. 12.11. Cross-section schematics of a typical VPE reactor, showing the group III species synthesis, group V species pyrolysis, and the growth zones with their respective temperature profiles for the growth of a few selected semiconductors.

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:

$$Ga_{liq} + HCl_g \rightarrow GaCl_g + \frac{1}{2}H_{2_g}$$

$$In_{liq} + HCl_g \rightarrow InCl_g + \frac{1}{2}H_{2_g}$$

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:

$$AsH_3 \rightarrow \frac{u}{4}As_4 + \frac{1-u}{2}As_2 + \frac{3}{2}H_2$$
$$PH_3 \rightarrow \frac{v}{4}P_4 + \frac{1-v}{2}P_2 + \frac{3}{2}H_2$$

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. 12.12: 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.

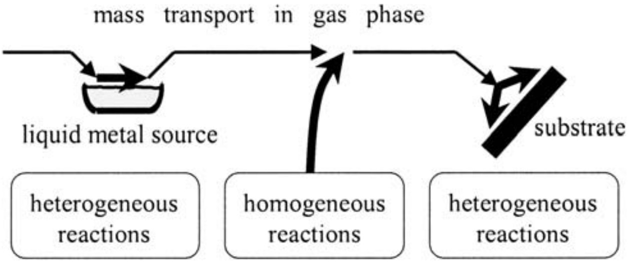


Fig. 12.12. Location of heterogeneous and homogeneous chemical reactions taking place during the vapor phase epitaxy growth process.

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. 12.13. The concentrations of reactants in the bulk gas phase is denoted C_G, while that at the surface of the substrate is C_S. Two fluxes are considered.

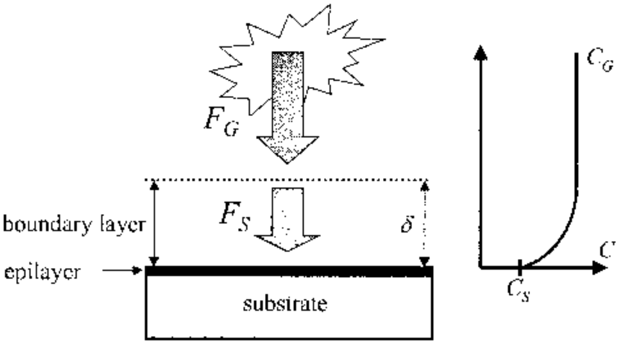


Fig. 12.13. Schematic diagram of the boundary layer near the epilayer/substrate surface in vapor phase epitaxy. A plot of the concentration of reactants in the bulk gas phase and at the surface as a function of the distance to the substrate is shown on the right.

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:

Eq. (12.10) $$F_G = \frac{D}{\delta}\left(C_G - C_S\right) = h_G\left(C_G - C_S\right)$$

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:

Eq. (12.11) $$F_S = k_S C_S$$

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:

Eq. (12.12) $$C_S = \frac{h_G}{h_G + k_S} C_G$$

The growth rate can be calculated as:

Eq. (12.13) $$\frac{dX}{dt} = \frac{F_S}{C} = \frac{h_G k_S}{h_G + k_S} \frac{C_G}{C}$$

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 $h_G >> k_S$, the growth rate can be approximated by:

Eq. (12.14) $$\frac{dX}{dt} \approx k_S \frac{C_G}{C}$$

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 $h_G << k_S$, the growth rate can be approximated by:

Eq. (12.15) $$\frac{dX}{dt} \approx h_G \frac{C_G}{C}$$

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.

12.5.3. 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. 12.14. 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.

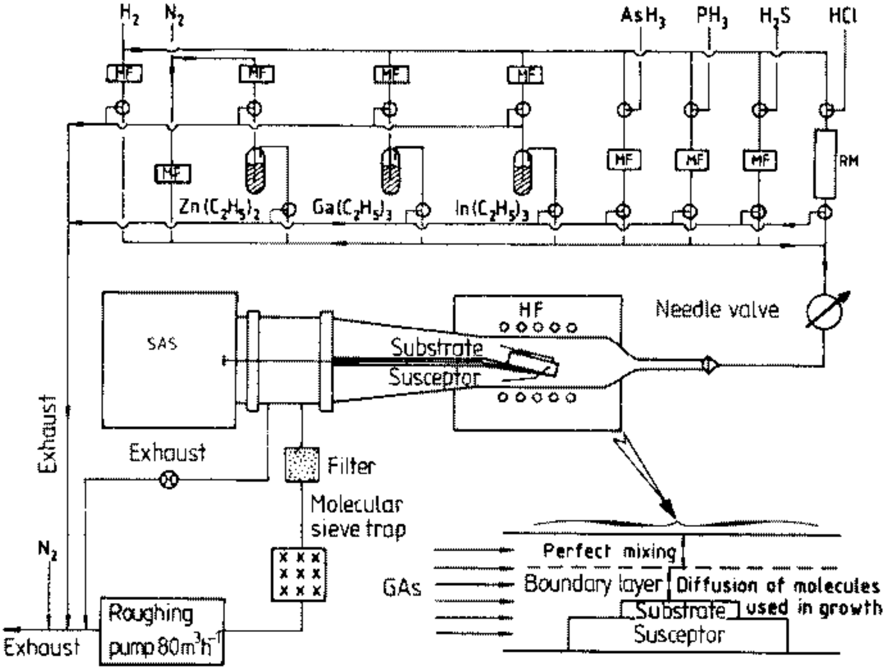


Fig. 12.14. Schematic diagram of a typical low-pressure MOCVD reactor. [Copyright © 1989 From The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

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.

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, 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 byproducts 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.11. 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 12.7.

For a thorough discussion of these source materials, the interested reader is referred to other books [Razeghi 1989].

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 sulphide	H_2S	VI element

Table 12.7. 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.

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 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 1989].

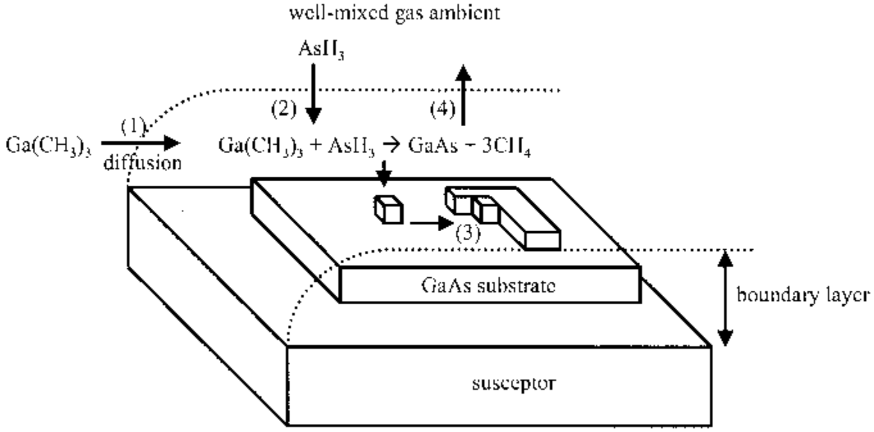


Fig. 12.15. A simplified schematic illustration of the GaAs growth process involving different steps.

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. 12.15. 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. 12.16 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. 12.16(a)), chemical reactions at the solid/vapor interface limit the growth rate as they follow an Arrhenius relation of the form $\exp\left(-\frac{E_A}{k_b T}\right)$ 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. 12.16(b)), 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. 12.16(c)), 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.

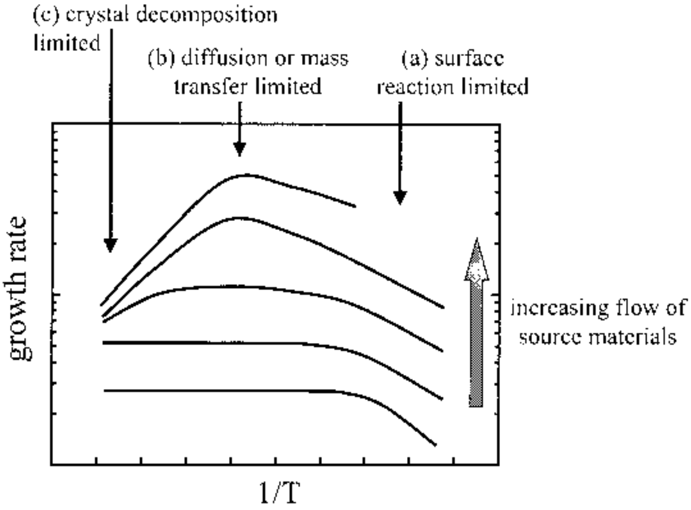


Fig. 12.16. Typical growth rate profile as a function of temperature.

In-situ characterization techniques. Although the MOCVD growth technique cannot accommodate as many in-situ characterization techniques as molecular beam epitaxy (sub-section 12.5.4), 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.

12.5.4. 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. 12.17.

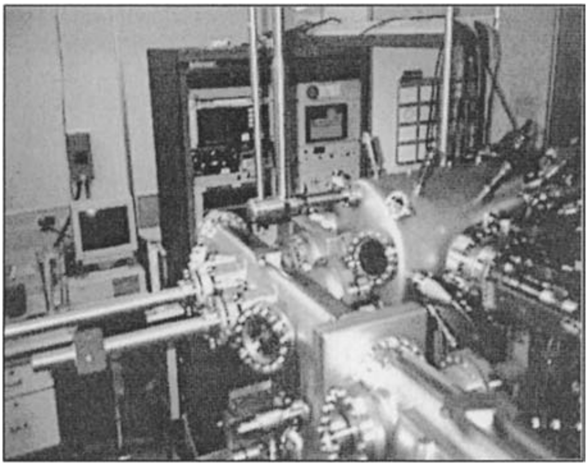


Fig. 12.17. Photograph of a molecular beam epitaxy reactor.

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 ultra high vacuum conditions, as illustrated in Fig. 12.18. 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.

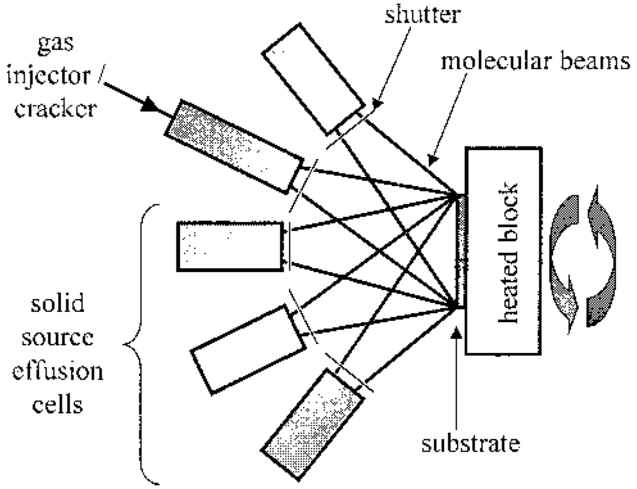


Fig. 12.18. Schematic diagram of an MBE growth system showing a few solid effusion cells, a gas injector/cracker, shutters controlling which sources are used at one time, the path of the beams, and a substrate mounted on a heated block that can be rotated.

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 ultra high 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:

$$\text{Eq. (12.16)} \quad \Lambda = \frac{1}{\sqrt{2}\pi n d^2}$$

in which d is the diameter of the atom or molecule, and n is its concentration in the growth chamber given by:

$$\text{Eq. (12.17)} \quad n = \frac{P}{k_b T}$$

where k_b is the Boltzmann constant, 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 stem from the fact that the growth is carried out in an ultra high 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 into four fundamental steps illustrated in Fig. 12.19. (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.

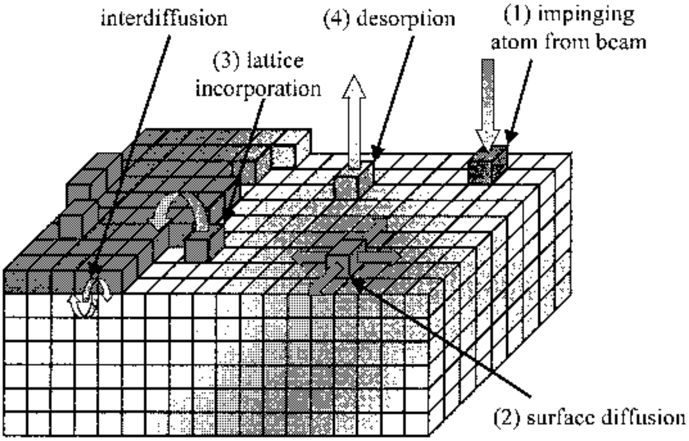


Fig. 12.19. Schematic illustration of the surface processes during MBE epitaxial growth, including: (1) the adsorption of the constituent atoms or molecules impinging on the substrate surface, (2) the surface migration and dissociation of the absorbed species, (3) the incorporation of the constituent 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.

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 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. 12.20. 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.

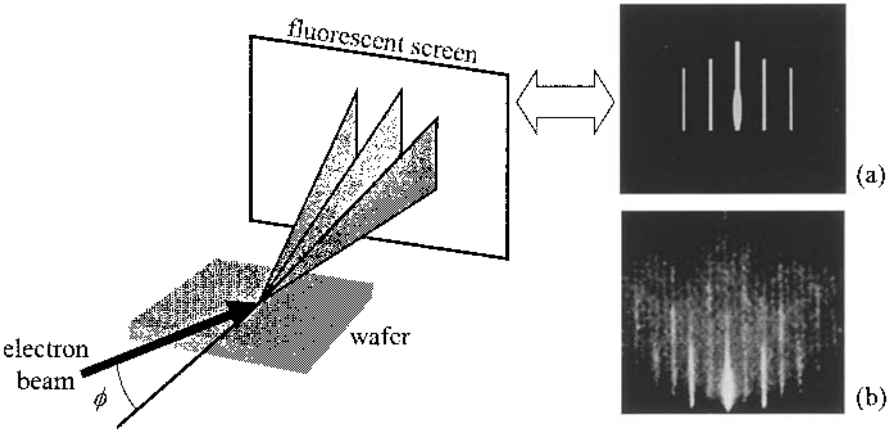


Fig. 12.20. Geometry of RHEED technique. A beam of electrons with an energy in the range 5~50 keV is directed on the substrate surface at an angle ϕ. The electrons are then partially reflected and diffracted by the wafer surface, which leads to the appearance of a bright spot and intensity-modulated streaks on a fluorescent screen, as is schematically shown in (a). An actual RHEED pattern is shown in (b).

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. 12.21. Indeed, during the epitaxial growth process, starting from an atomically flat surface (i.e. coverage: θ=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 % (θ=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 (θ=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.

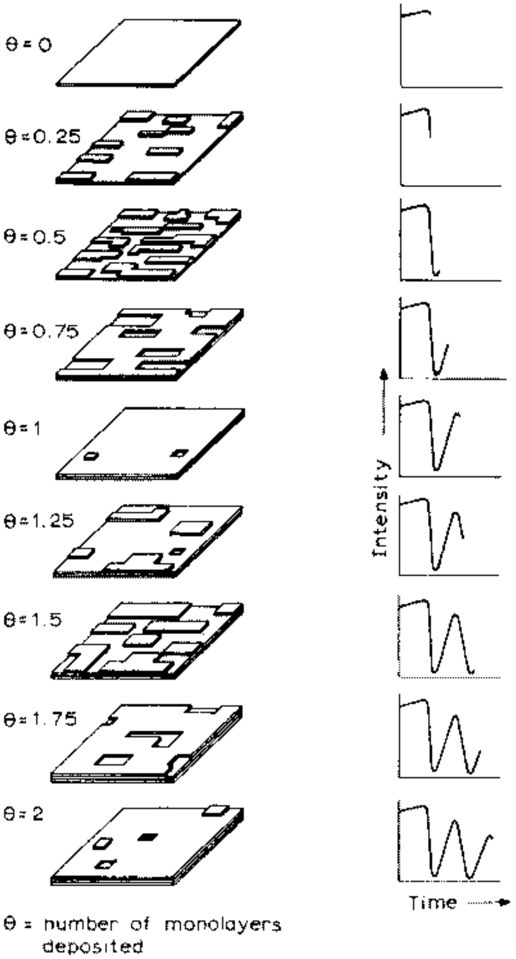


Fig. 12.21. Schematic diagram illustrating the dynamic RHEED process. The sketches on the left show the various stages of the surface morphology during epitaxial growth, while the right plots show the intensity of the RHEED signal from the specular beam as a function of time. [Reprinted from Surface Science Vol. 168, Joyce, B.A., Dobson, P.J., Neave, J.H., Woodbridge, K., Zhang, J., Larsen, P.K., and Bölger, B., "RHEED studies of heterojunction and quantum well formation during MBE growth-from multiple scattering to band offsets," p. 426, Copyright 1986, with permission from Elsevier.]

In spite of its technological advantages over other epitaxial growth techniques, MBE suffers from the high cost to maintain the ultra high 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.

12.5.5. 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 re-arranging 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).

12.5.6. 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 Chapter 13.

12.6. Thermodynamics and kinetics of growth

In section 12.5.3, 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 sub-sections, 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.

12.6.1. Thermodynamics

In this sub-section 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:

Eq. (12.18) $G = H - TS$

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:

Eq. (12.19) $H = E + PV$

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:

Eq. (12.20) $\Delta G = G_f - G_i = \Delta H - T\Delta S$

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 (ΔG<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, ΔG>0. Therefore, the possibility of occurrence of a particular reaction can be determined via the sign of ΔG.

12.6.2. 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:

Eq. (12.21) $xX + yY \rightarrow zZ$

The free-energy change of the reaction is given by:

Eq. (12.22) $\Delta G = zG_Z - xG_X - yG_Y$

The free energy of individual reactants is often written as:

Eq. (12.23) $G_i = G_i^0 + RT \ln a_i$

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 ºC. A pure liquid or solid is the standard state of the relevant substance. Table 12.8 lists the standard values of the change of enthalpy and entropy for the formation of various substances. Substitution of Eq. (12.23) into Eq. (12.22) and letting $\Delta G = 0$ yields:

Eq. (12.24) $-\Delta G^0 = RT \ln K$

where:

Eq. (12.25) $K = \dfrac{a^{z}_{Z(eq)}}{a^{x}_{X(eq)} a^{y}_{Y(eq)}}$

Species	State	ΔH_f (kJ.mol^{-1})	S (J.mol^{-1}.K^{-1})
H_2O	*l*	-286	70
H_2O	*g*	-242	190
CO_2	*g*	-394	214
O_2	*g*	0	205
HCl	*g*	-92	190
HCl	*aq*	-167	57
H	*g*	218	115
Cl_2	*g*	0	220
NaCl	*s*	-411	72

Table 12.8. 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).

Let us see how thermodynamics can help us find out about feasibility of a chemical reaction. Table 12.9 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.

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

Table 12.9. Free-energy change and classification of some select reactions.

12.6.3. 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:

Eq. (12.26) $f = C - P + 2$

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. Typical 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. 12.22).

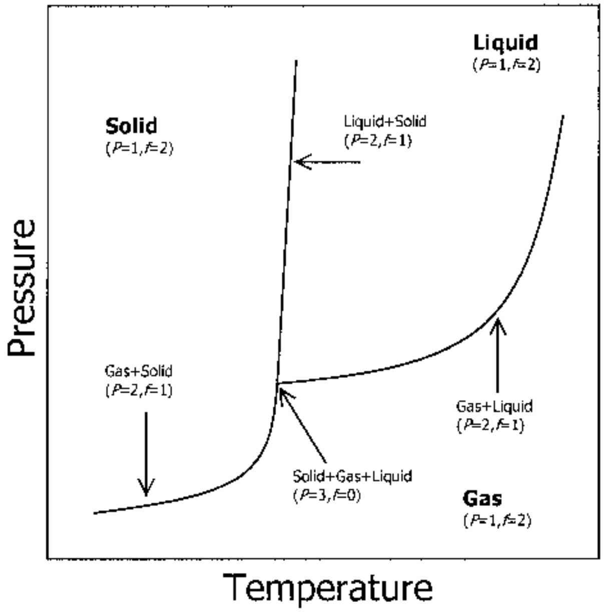


Fig. 12.22. P-T diagram of a one-component system showing degrees of freedom for different number of phases.

12.6.4. 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. 12.23). The rates of the forward and reverse reactions can be described as:

Eq. (12.27) $Rate = nk$

where n is the concentration of reactants/products and k is the rate constant usually expressed in terms of the Arrhenius equation:

Eq. (12.28) $k = Ae^{-E^*/RT}$

In this equation, A is a pre-exponential factor and E^* is the activation energy of the process. R is the gas constant.

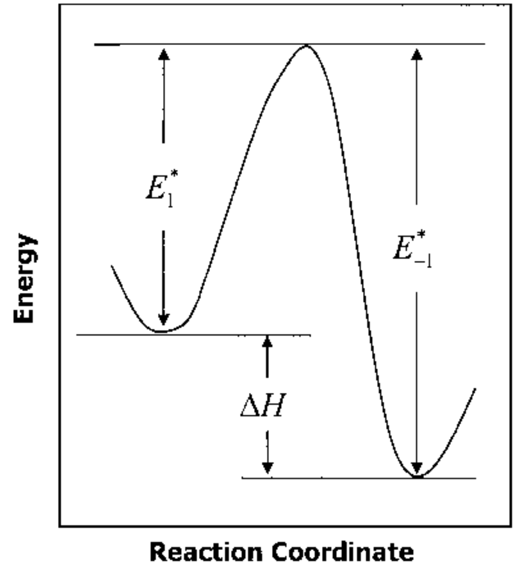


Fig. 12.23. Schematic diagram of energy vs. reaction coordinate. E_1^* *and* E_{-1}^* *are the activation energies of the forward and reverse reactions, respectively.*

From Fig. 12.23, we find the thermodynamic enthalpy difference from the initial to the final state, *ΔH*, to be:

Eq. (12.29) $\Delta H = E_1^* - E_{-1}^*$

At equilibrium, the rates of the forward and reverse reactions are equal:

Eq. (12.30) $n_i k_1 = n_f k_{-1}$

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:

Eq. (12.31) $\dfrac{n_f}{n_i} = \dfrac{k_1}{k_{-1}} = K_1 = \exp\left(\dfrac{-\Delta G_1^0}{RT}\right)$

In Eq. (12.31), K_I 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 of 25 °C (298 K), reactants and products at concentration of 1 mole.

12.7. 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-Krastanow 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. 12.24(a)). 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. 12.24(b)). In this growth mode, the deposited atoms are more 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-Krastanow transition (Fig. 12.24(c)). 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.

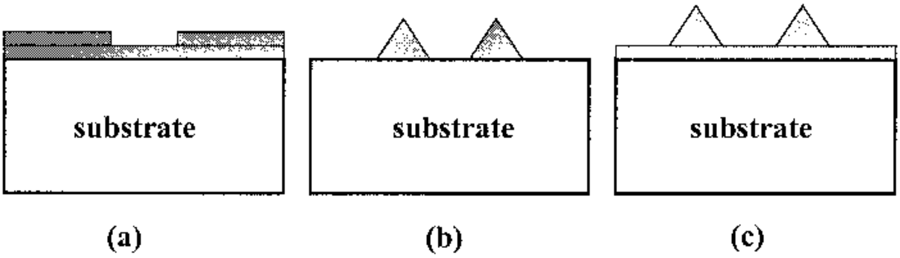


Fig. 12.24. Schematic presentation of the (a) Frank-van der Merwe, (b) Volmer-Weber, and (c) Stranski-Krastanow growth modes.

12.8. 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 section 12.6. Finally, the various growth modes were discussed, covering both lattice-matched and lattice-mismatched systems.

References

Casey Jr., H.C. and Panish, M.B., *Heterostructure Lasers, parts A & B*, Academic Press, New York, 1978.

Cho, A.Y., *The Technology and Physics of Molecular Beam Epitaxy*, Plenum Press, New York, 1985.

Glasstone, S., Leidler, K.J., and Eyring, H. *The theory of rate processes*, McGraw-Hill, New York, 1941.

Horikoshi, Y., "Migration-enhanced epitaxy of GaAs and AlGaAs," *Semiconductor Science Technology* 8, pp. 1032-1051, 1993.

Ilegems, M. and Panish, M.B., "Phase equilibria in III-V quaternary systems-Application to Al-Ga-P-As," *Journal of Physics and Chemistry of Solids* 35, pp. 409-420, 1974

Jordan, A.S. and Ilegems, M., "Solid-liquid equilibria for quaternary solid solutions involving compound semiconductors in the regular solution approximation," *Journal of Physics and Chemistry of Solids* 36, pp. 329-342, 1974

Joyce, B.A., Dobson, P.J., Neave, J.H., Woodbridge, K., Zhang, J., Larsen, P.K., and Bôlger, B., "RHEED studies of heterojunction and quantum well formation during MBE growth-from multiple scattering to band offsets," *Surface Science* 168, p. 426, 1986.

Ray, B., *II-VI compounds*, Pergamon Press, Oxford, UK, 1969.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, 1989.

Razeghi, M., *The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for Photonic and Electronic Device Applications*, Institute of Physics, Bristol, UK, pp. 21-29, 1995.

Roberts, G.G. and Zallen R., "Quenching of photoconductivity and luminescence in natural crystals of mercury sulphide", *Journal of Physical Chemistry: Solid State Physics* 4, 1890-1897, 1971.

Stringfellow, G.B., *Organometallic Vapor-Phase Epitaxy: Theory and Practice*, Academic Press, Boston, MA, 1989.

Van Vechten, J.A., and Bergstresser, T.K., "Electronic structures of semiconductor alloys," *Physical Review B* 1, pp. 3351-3358, 1970.

Further reading

Gise, P.E. and Blanchard, R., *Semiconductor and Integrated Circuit Fabrication Techniques*, Reston Publishing, Reston, VA, 1979.

Middleman, S. and Hochberg, A.K., *Process Engineering Analysis in Semiconductor Device Fabrication*, McGraw-Hill, New York, pp. 230-257, 1993.

Nakajima, K., "Liquid-phase epitaxy," in *GaInAsP Alloy Semiconductor*, ed. T.P. Pearsall, John Wiley & Sons Limited, Chichester, UK, pp. 43-59, 1982.

Ohring, M., *The Materials Science of Thin Films*, Academic Press, San Diego, US, 1992.

Razeghi, M., "LP-MOCVD growth, characterization, and application of InP material," *Semiconductors and Semimetals* 31, pp. 256-257, 1990.

Stringfellow, G.B., *Organometallic Vapor-Phase Epitaxy: Theory and Practice*, Academic Press, Boston, MA, 1989.

Tu, K.N, Mayer, J.W., and Feldman, L.C., *Electronic Thin Films Science for Electrical Engineers and Materials Scientists*, Macmillan Publishing, New York, 1992.

Problems

1. From the expressions of the bandgap energy of ternary alloys given in Table 12.2 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:
   $Al_xGa_{1-x}As$, $Al_xIn_{1-x}As$, $Ga_xIn_{1-x}P$, $Ga_xIn_{1-x}As$, $Ga_xP_{1-x}As$, and $In_xP_{1-x}As$.

2. Derive the relation given in Eq. (12.3):
   $$a_{A_xB_{1-x}C_yD_{1-y}} = xya_{AC} + x(1-y)a_{AD} + (1-x)ya_{BC} + (1-x)(1-y)a_{BD}$$

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
   $$E_g(Al_xIn_yGa_{1-x-y}N) = (1-x-y)E_g(GaN) + xE_g(AlN) + yE_g(InN)$$

4. Using the diagram in Fig. 12.1, 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.

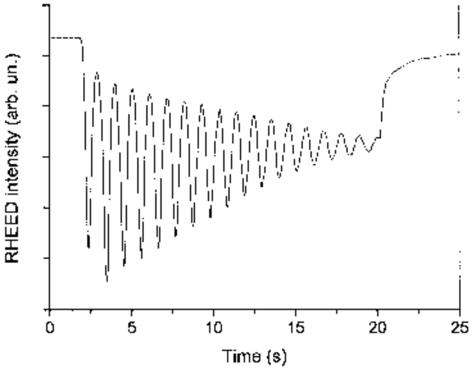
6. Derive Eq. (12.8): $C_s = kC_0(1-X)^{k-1}$, where:
   C_s = impurity concentration in the solid,
   C_0 = original impurity concentration in the melt,

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 4 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 a 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?

13. Semiconductor Characterization Techniques

13.1. 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, their structures and their 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.

13.2. Structural characterization techniques

13.2.1. 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. 13.1) 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 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.

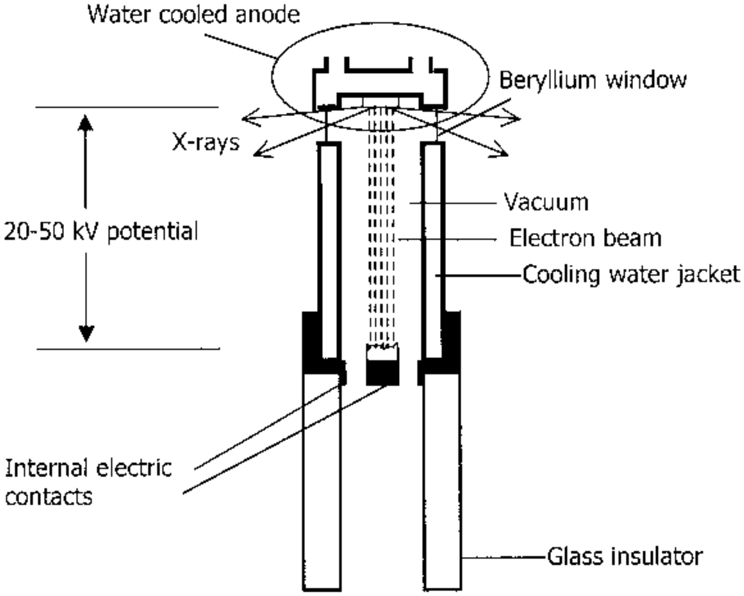


Fig. 13.1. Schematic diagram of an x-ray tube.

The peaks in an x-ray diffraction pattern are directly related to the atomic distances. For a given set of lattice planes with an inter-plane distance d, the condition for a diffraction (peak) to occur can be found using Bragg's law:

Eq. (13.1) $$2d\sin\theta = n\lambda$$

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. 13.2.

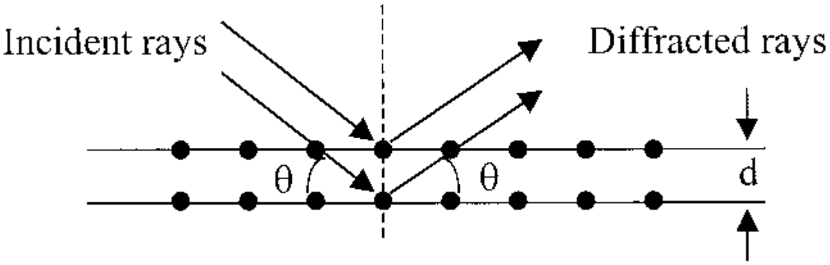


Fig. 13.2. Schematic of diffraction of x-rays by a crystal.

Fig. 13.3 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 multi-layer 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 is a measure of the film quality.

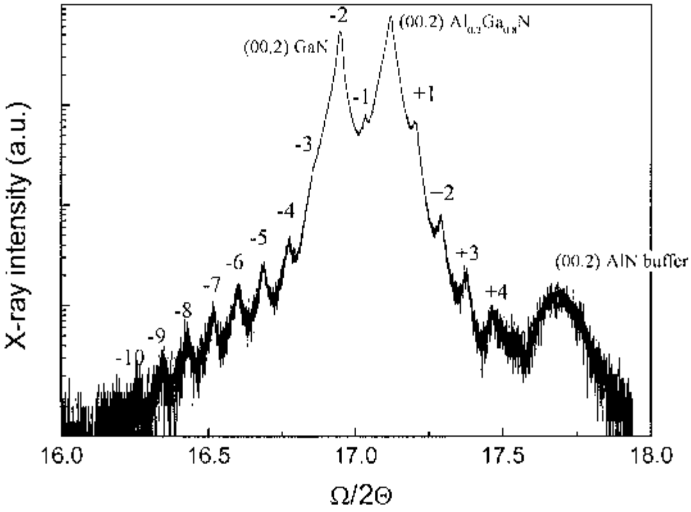


Fig. 13.3. X-ray curve of an $Al_{0.2}Ga_{0.8}N$/GaN superlattice grown on GaN/AlN buffer layer. The individual $Al_{0.2}Ga_{0.8}N$, GaN, and AlN peaks as well as the superlattice satellite peaks are clearly discernible on the graph.

13.2.2. 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. 13.4. 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.

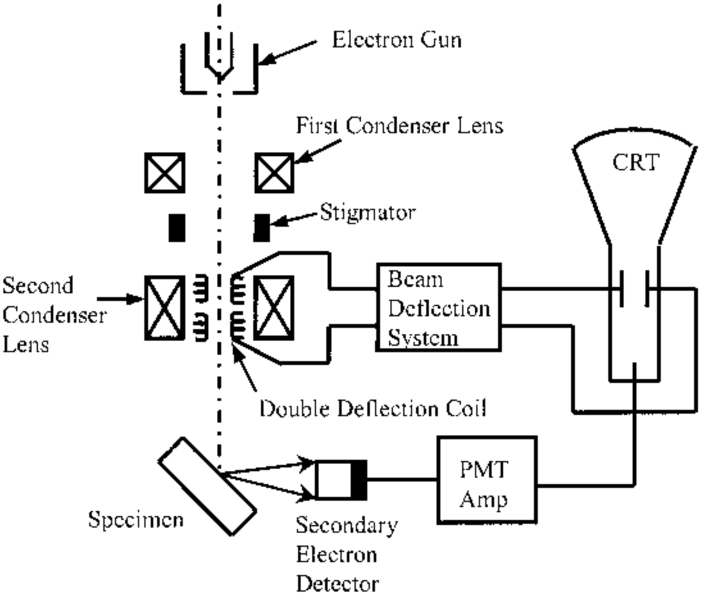


Fig. 13.4. Schematic of a scanning electron microscope.

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). Fig. 13.5(a) illustrates a bird's eye view image of a surface of a "nanopillar" sample while Fig. 13.5(b) displays the cross-section of a multi-layer semiconductor structure.

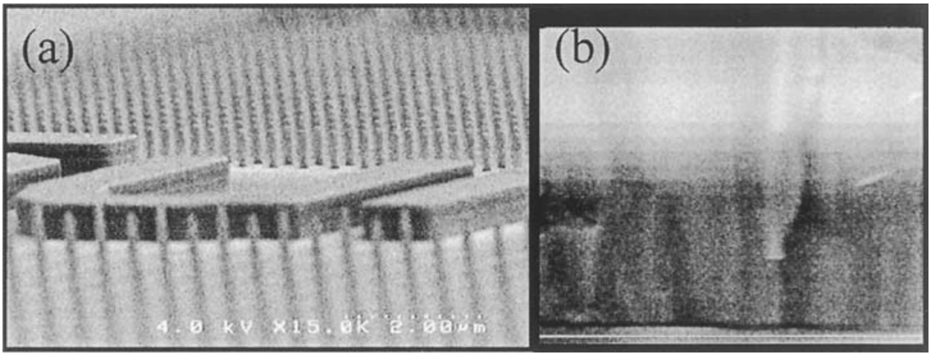


Fig. 13.5. (a) Bird's eye view of the surface of a nanopillar sample and (b) Cross-sectional SEM image of a multi-layer semiconductor structure.

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.

Fig. 13.6 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. Fig. 13.7 shows a high-resolution lattice image of the AlN/Al_2O_3 interface. Dislocations can be identified when any of the atomic planes terminates.

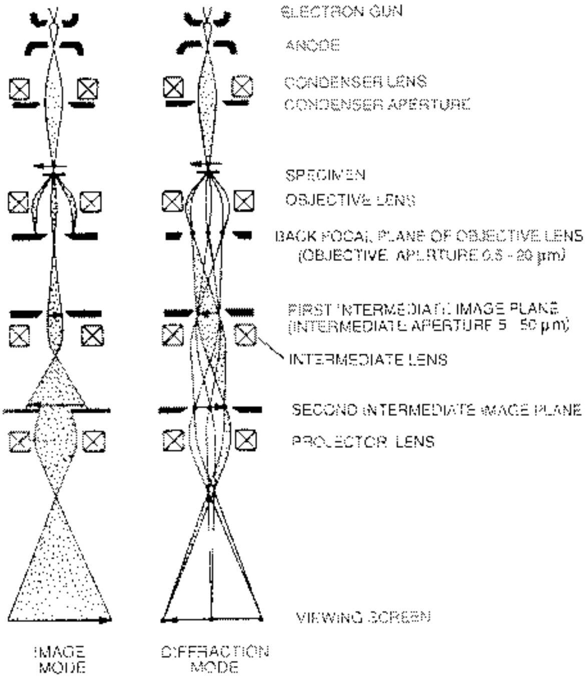

ELECTRON GUN
ANODE
CONDENSER LENS
CONDENSER APERTURE
SPECIMEN
OBJECTIVE LENS
BACK FOCAL PLANE OF OBJECTIVE LENS
(OBJECTIVE APERTURE 0.8 - 20 μm)
FIRST INTERMEDIATE IMAGE PLANE
(INTERMEDIATE APERTURE 5 - 50 μm)
INTERMEDIATE LENS
SECOND INTERMEDIATE IMAGE PLANE
PROJECTOR LENS
VIEWING SCREEN
IMAGE MODE
DIFFRACTION MODE

Fig. 13.6. Schematic of the TEM in imaging and diffraction modes. [Thomas, G. and Goringe, M.J., Transmission Electron Microscopy of Materials. Copyright © 1979 by John Wiley and Sons. Reprinted with permission of CBLS.]

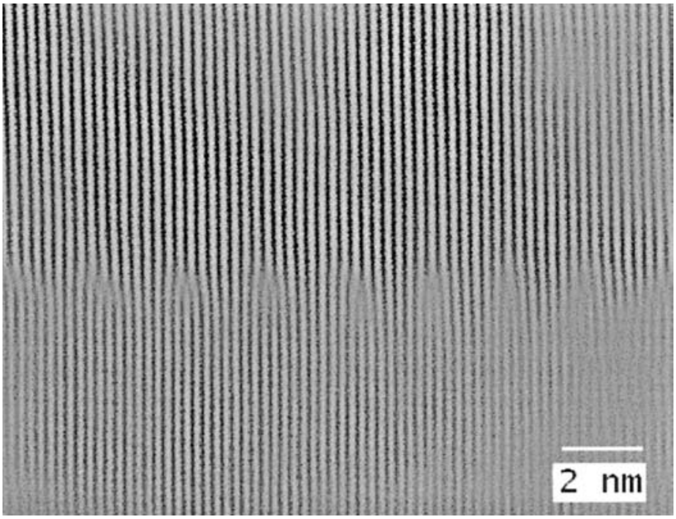


Fig. 13.7. High resolution TEM image of the interface of AlN and sapphire (Al_2O_3). One misfit dislocation generates when an atomic plane ends.

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)):

Eq. (13.2) $$\lambda = \frac{h}{p}$$

As mentioned above, electrons are accelerated to very high energies. If we let this potential energy, eV, equal the kinetic energy of the electrons:

Eq. (13.3) $$eV = \frac{m_0 v^2}{2}$$

the momentum of an electron can be written as:

Eq. (13.4) $$p = m_0 v$$

Therefore, the wavelength of the electrons, from the above three equations, can be expressed as:

Eq. (13.5) $$\lambda = \frac{h}{\sqrt{2m_0 eV}}$$

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:

Eq. (13.6) $$\lambda = \frac{h}{\sqrt{2m_0 eV\left(1 + \frac{eV}{2m_0 c^2}\right)}}$$

For example, with an acceleration voltage of 1 MV, the non relativistic wavelength is 0.0122 Å while the relativistic value is only 0.0087 Å [Williams and Carter 1996].

13.2.3. 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-ray 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. Fig. 13.8 shows an example of an EDX plot.

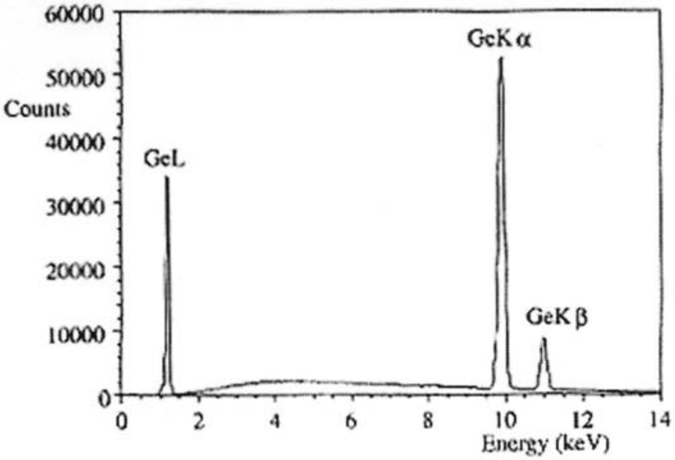


Fig. 13.8. An example of an EDX measurement. Multiple lines of Ge emission correspond to the various electron energy transitions. [Transmission Electron Microscopy, 1996, p. 557, Williams, D.B. and Carter, C.B., Fig. 32.2. © 1996 Plenum Press, New York. With kind permission of Springer Science and Business Media.]

13.2.4. Auger electron spectroscopy (AES)

The AES technique takes advantage of the Auger transitions that were introduced in Chapter 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.

A typical Auger spectrometer is kept under ultra-high 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. 13.9.

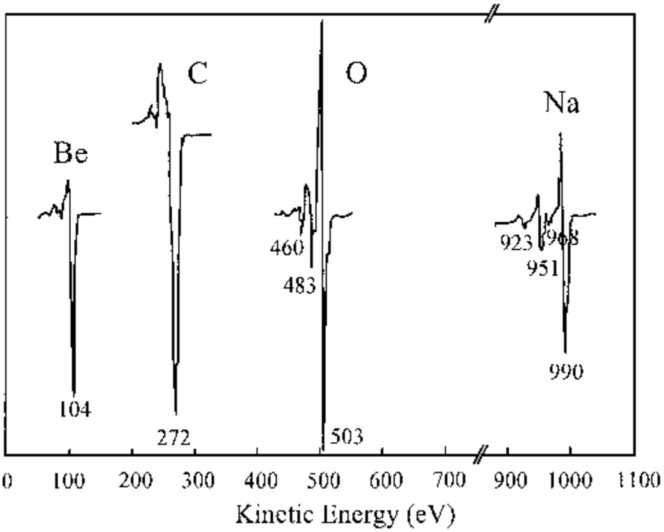


Fig. 13.9. Auger electron spectra of various elements.

13.2.5. 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:

Eq. (13.7) $$E_{KE} = h\upsilon - E_{BE}$$

where E_{KE} is the energy of the ejected electron, $h\upsilon$ 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 Ag) is shown in Fig. 13.10. It should be noted that for multi-component samples the intensities of the peaks are proportional to the concentration of the element within the sampled region.

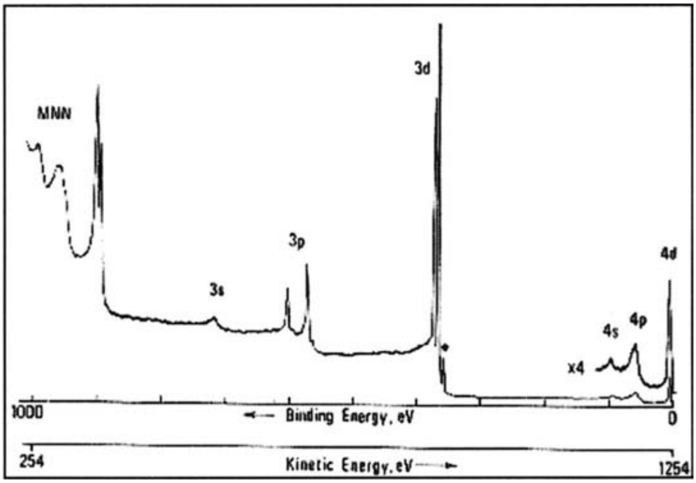


Fig. 13.10. An XPS spectrum from a silver sample. [Practical Surface Analysis: by Auger and X-ray Photoelectron Spectroscopy, Briggs, D. and Seah, M.P. Copyright 1983. ©John Wiley & Sons Limited. Reproduced with permission.]

13.2.6. 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. Fig. 13.11 shows a SIMS depth profile of a GaN sample showing its concentration of impurities (oxygen, carbon, silicon) using Cs^+ bombardment.

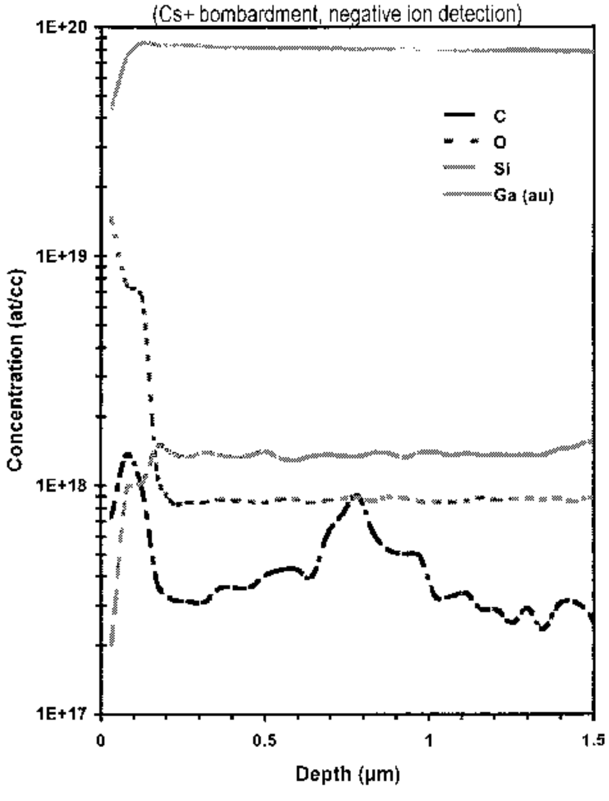


Fig. 13.11. A SIMS depth profile showing the concentration of impurities in a GaN sample. The impact energy was 15.5 keV at oblique incidence and the detected area was 33 μm in diameter.

13.2.7. 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 afterwards given by Ohring [1992]:

$$\text{Eq. (13.8)} \quad E_1 = \left\{ \frac{\left(M^2 - M_0^2 \sin^2\theta\right)^{1/2} + M_0 \cos\theta}{M_0 + M} \right\}^2 E_0$$

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 multi-layer 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.

13.2.8. 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. 13.12). 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.

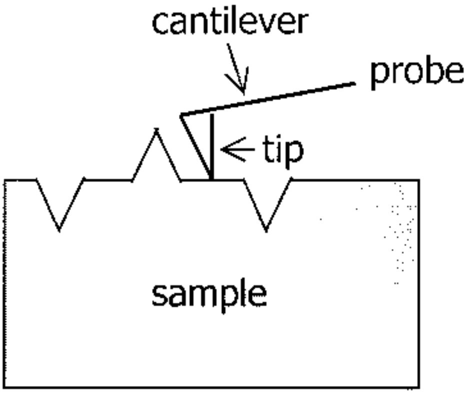


Fig. 13.12. Schematic of an AFM tip scanning over the surface of a sample

Essential to the system is a piezoelectric tube (Fig. 13.13). 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 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.

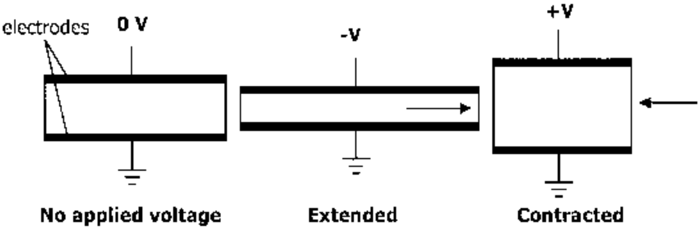


Fig. 13.13. Reaction of a piezo material to applied bias.

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. 13.14.

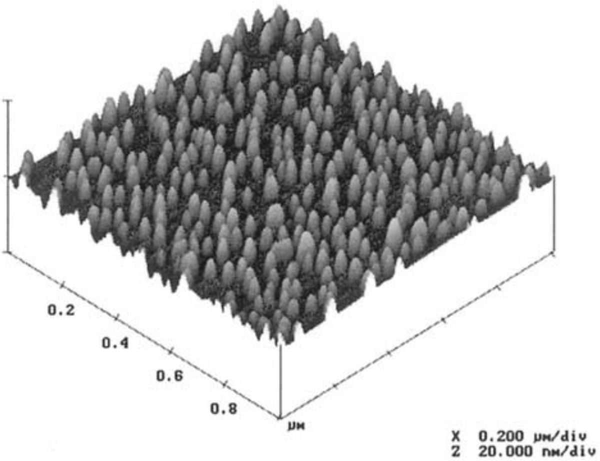


Fig. 13.14. A 3D AFM image of the surface of a sample consisting of InAs quantum dots grown on top of a GaAs/InP substrate.

13.3. Optical characterization techniques

13.3.1. Photoluminescence spectroscopy

Photoluminescence (PL) spectroscopy is a non-destructive method of probing the electrical properties of materials. Light is focused onto the sample where it is absorbed in a process called "photo-excitation". As a result of the excess energy caused by photo-excitation, 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. 13.15. 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 non-radiative processes can occur. The intensity of the PL peak and its dependence on the level of photo-excitation and temperature is directly related to the dominant recombination process.
- Material quality: The intensity and the line width (FWHM) of a PL spectrum are representative of the quality of the material. Additionally, presence of defect-related peaks is indicative of imperfections in the epitaxial layer.

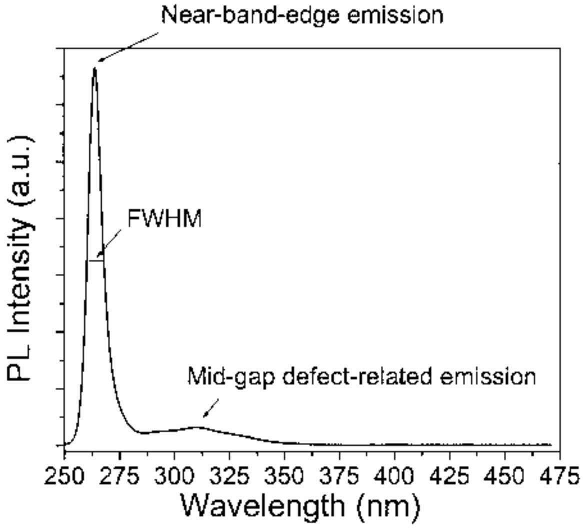


Fig. 13.15. Photoluminescence spectrum of an AlGaN sample. Shown on the graph are the near-band-edge emission peak and a defect-related emission peak.

13.3.2. 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 sub-section.

13.3.3. 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:

Eq. (13.9) $$R = \left|\frac{E_r}{E_i}\right|^2 = \left(\frac{\overline{n}-1}{\overline{n}+1}\right)^2$$

where E_r and E_i are the energy of the reflected and incident light, respectively and $\overline{n}$ is the refractive index of the medium.

13.3.4. 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 Chapter 10) gives the following relationship between the absorption coefficient α (see Eq. (10.81)), the light energy E, and the bandgap energy E_g :

Eq. (13.10) $$\alpha \propto \sqrt{E - E_g}$$

13.3.5. 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:

$$\text{Eq. (13.11)} \quad \tan(\Psi)e^{i\Delta} = \frac{R_p}{R_s}$$

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. 13.16, 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.

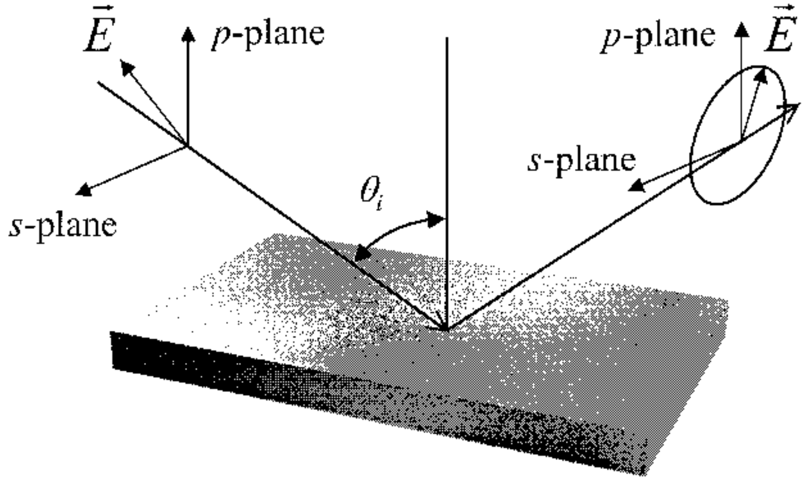


Fig. 13.16. Schematic of the geometry of an ellipsometry measurement. The coordinate system used to describe the ellipse of polarization is the p-s coordinate system. The s-direction is taken to be perpendicular to the direction of propagation and parallel to the sample surface. The p-direction is taken to be perpendicular to the direction of propagation and contained in the plane of incidence.

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:

Eq. (13.12) $T = \frac{I_t}{I_i}$ and $R = \frac{I_r}{I_i}$

where I_i, I_t, and I_r are the intensities of the incident, transmitted, and reflected lights, respectively.

13.3.6. 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. 13.17, 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.

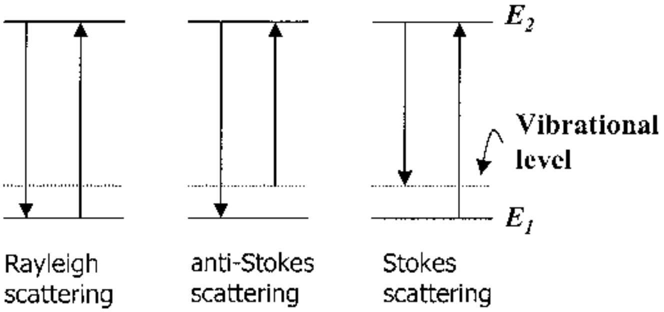


Fig. 13.17. Schematic depiction of various scattering processes within a medium. The incident photon energies are marked by the right-hand-side arrows.

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. 13.18.

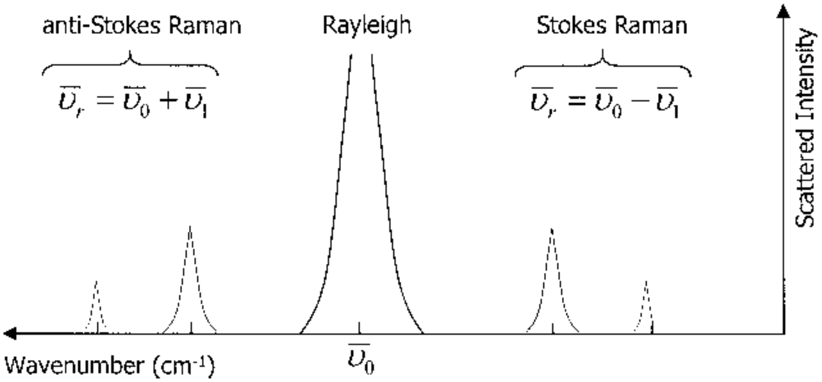


Fig. 13.18. An example of a Raman spectrum representing Rayleigh, Stokes, and anti-Stokes Raman peaks.

13.3.7. 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 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. 13.19. 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).

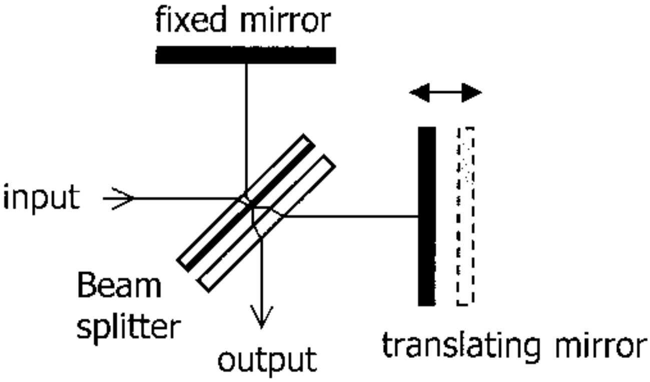


Fig. 13.19. Schematic cross-section of a Michelson interferometer.

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. 13.20. 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.

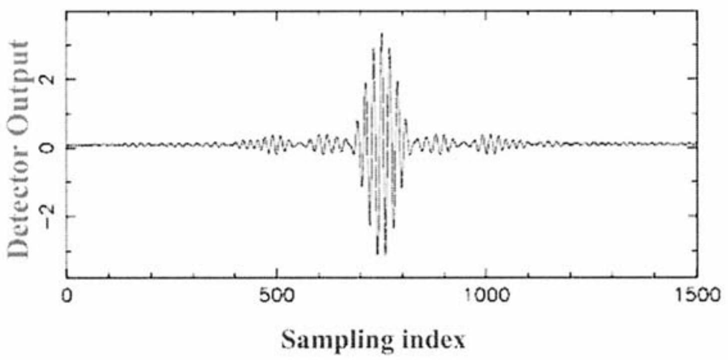


Fig. 13.20. A typical interferogram.

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 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. 13.21 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:

Eq. (13.13) $$\upsilon(cm^{-1}) = \frac{10000}{\lambda(\mu m)}$$

Therefore, it would be easy to convert wavenumber to other useful units such as wavelength or energy, as is the case in Fig. 13.21.

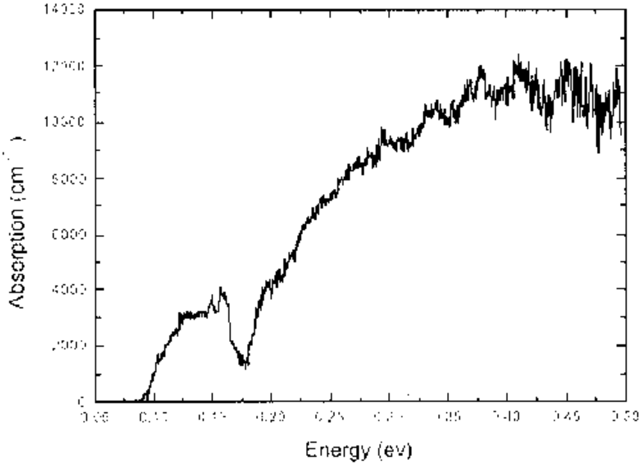


Fig. 13.21. Absorption spectrum for a semiconductor photodetector structure taken by a Fourier transform infrared (FTIR) system.

13.4. Electrical characterization techniques

13.4.1. 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:

Eq. (13.14) $\rho(z) = \left[N(z)e\mu(N)\right]^{-1}$

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:

Eq. (13.15) $\rho_s = \left[\int_0^t N(z)e\mu(N)dz\right]^{-1}$

where *t* is the thickness of the layer.

13.4.2. 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 Chapter 8 for a complete discussion on the Hall effect.

13.4.3. 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 Chapter 9 for more discussion on junction capacitance):

Eq. (13.16) $N = \dfrac{2}{\varepsilon\varepsilon_0 A^2}\left(\dfrac{1}{d(C^{-2})/dV_r}\right)$

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. Fig. 13.22 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.

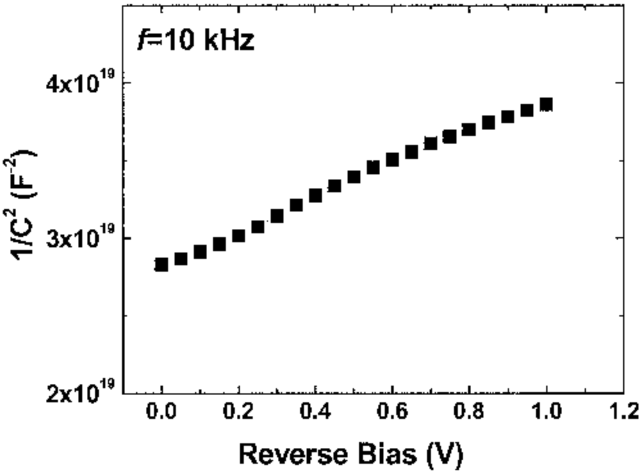


Fig. 13.22. Plot of C^{-2} vs. reverse bias for a p-type GaN sample. The measurements were taken at a frequency of 10 kHz.

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.

13.4.4. 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.

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. 13.23.

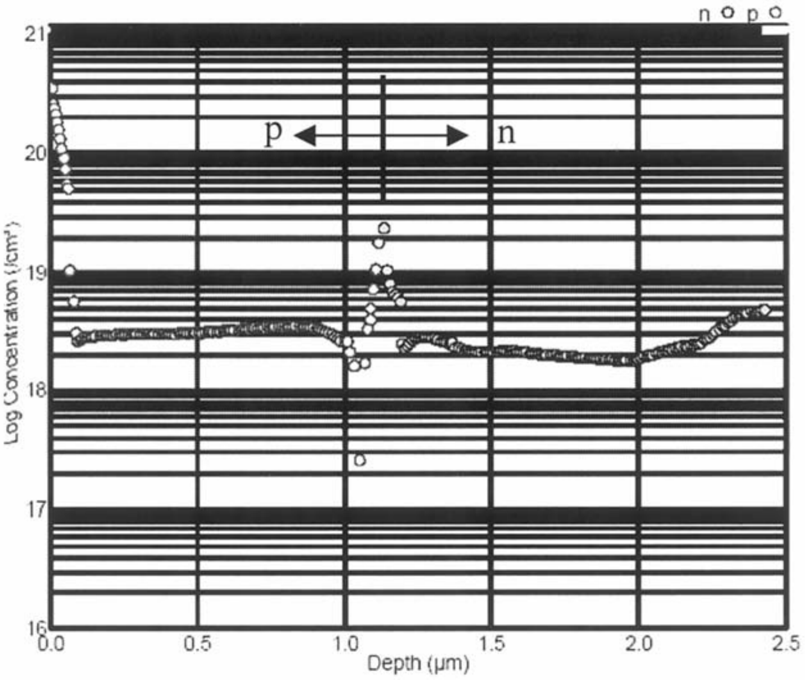


Fig. 13.23. A representative ECV profile showing the concentration of n-type and p-type dopants as a function of depth for a 980 nm laser diode structure.

13.5. 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.

References

Briggs, D. and Seah, M.P., *Practical Surface Analysis: by Auger and X-ray Photoelectron Spectroscopy*, John Wiley and Sons, Chichester, UK, 1983.

Davies, L.E., MacDonald, N.C., Palmberg, P.W., Riach, G.E., and Weber, R.E., *Handbook of Auger Electron Spectroscopy*, Physical Electronics Division, Perkin-Elmer Corporation, USA, 1978.

Ohring, M., *The Materials Science of Thin Films*, Academic Press, San Diego, CA, 1992.

Thomas, G. and Goringe, M.J., *Transmission Electron Microscopy of Materials*, John Wiley and Sons, New York, 1979.

Williams, D.B. and Carter, C.B., *Transmission Electron Microscopy*, Plenum Press, New York, 1996.

Further reading

Long, D.A., *Raman Spectroscopy*, McGraw-Hill, New York, 1977.

Perkowitz, S., *Optical Characterization of Semiconductors: Infrared, Raman, and Photoluminescence Spectroscopy*, Academic Press, London, UK, 1993.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, 1989.

Razeghi, M., *The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for Photonic and Electronic Device Applications*, Institute of Physics, Bristol, UK, pp. 21-29, 1995.

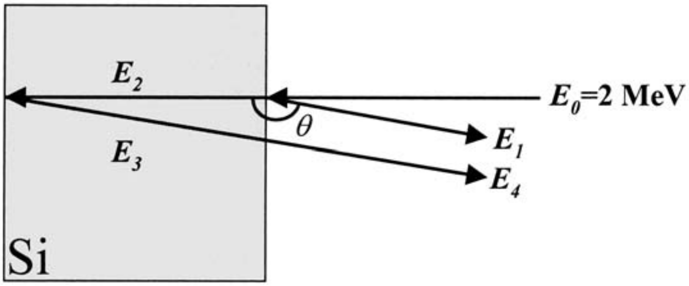
Stradling, R.A. and Klipstein, P.C., *Growth and Characterization of Semiconductors*, Adam Hilger, New York, 1990.

Warren, B.E., *X-ray Diffraction*, Dover Publications, New York, 1990.

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 ($\Delta E = E_4 - E_1$).

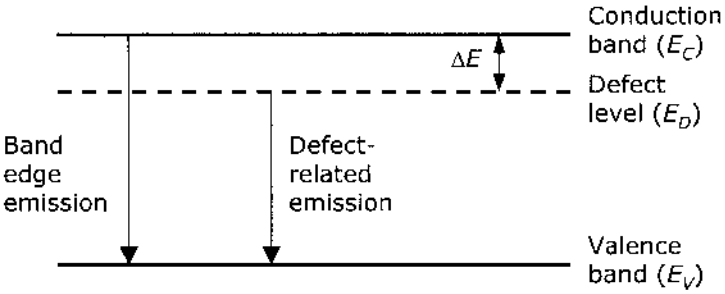


3. Estimate the acceptor concentration of the p-type GaN of Fig. 13.22 assuming a diode area of 400 μm × 150 μm and a dielectric constant of ε=10 ε_0.

4. Based on the SIMS spectrum of Fig. 13.11:
(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. 13.15:
(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 resisitivity (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. Do you think SEM and AFM are competing techniques or complementary techniques? Explain why.

8. 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.

9. Based on the TEM image provided in Fig. 13.7 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?

14. Defects

14.1. Introduction

An ideal crystalline solid has a periodic structure that is based on the chemical properties of its constituent atoms (see Chapter 1). 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 (Chapter 4). 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, 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:

(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 Chapter 12 for details)

Intrinsic defects can be introduced during crystal growth. For example, typical concentrations of intrinsic defects in Si is 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-

Krastanow growth mode (Chapter 12). 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 14.1 displays examples of these four types of defects.

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

Table 14.1. Table of dislocation dimension classifications.

14.2. Point defects

Point defects, or 0-dimensional defects, refer to missing, additional, or misplaced atoms within the crystalline lattice. Fig. 14.1 shows examples of substitutional, interstitial and vacancy point defects, each of which will be discussed in more detail in the following sections.

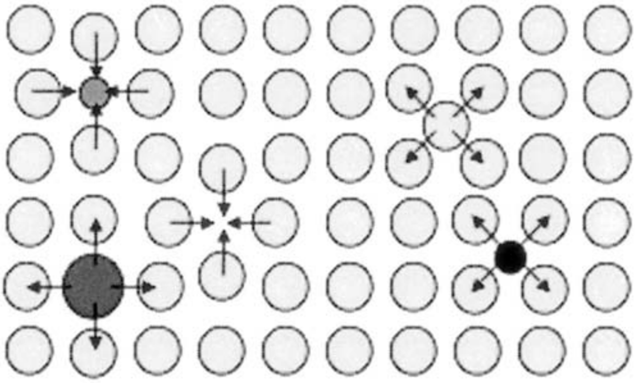


Fig. 14.1. Examples of point defects.

14.2.1. 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. 14.2(a), while a Frenkel defect is shown schematically in Fig. 14.2(b).

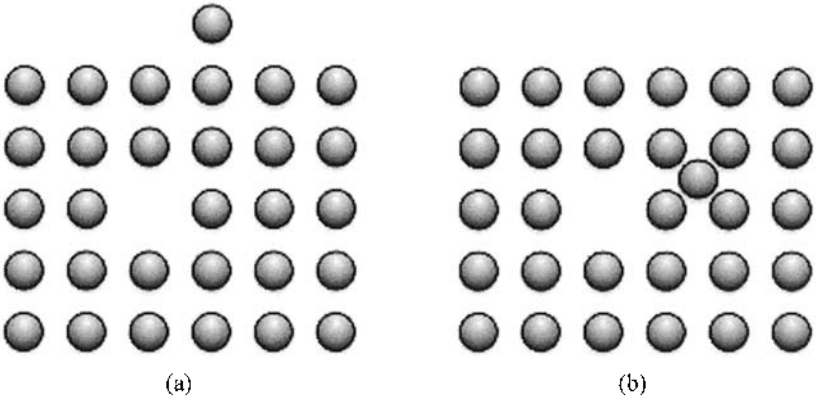


Fig. 14.2. Schematic diagrams of a: (a) Schottky defect and (b) Frenkel defect.

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. (14.1), where T is absolute temperature:

Eq. (14.1) $\Delta G = \Delta H - T\Delta S$

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 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. (14.2), 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 Boltzmann's constant. By dividing n_e by N, the equilibrium concentration of defects, n_e, may be found.

Eq. (14.2) $$n_e = NA\exp\left(\frac{-E_D}{k_b T}\right)$$

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 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. 14.3.

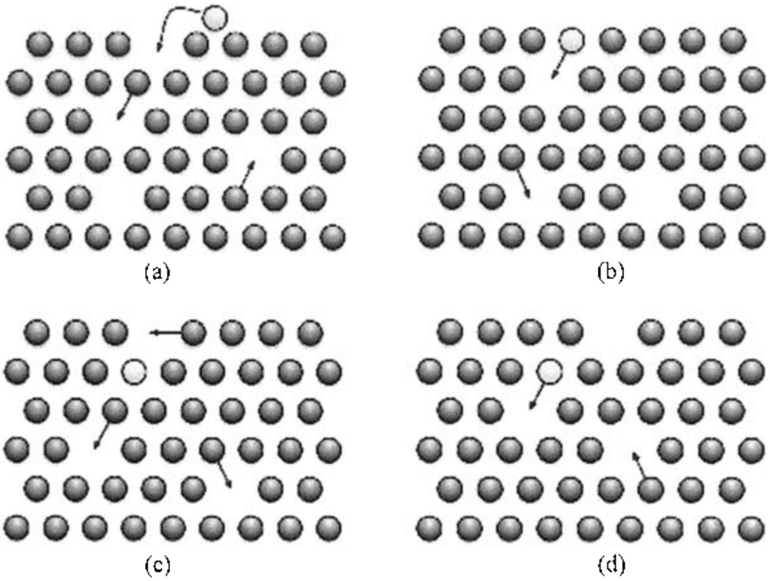


Fig. 14.3. Schematic of chemical diffusion showing how a foreign atom may diffuse into a crystal with time assisted by the presence of voids (increasing time from (a) to (d)).

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. For more details on chemical diffusion, see Chapter 15.

Another type of intrinsic point defect is an anti-site defect, shown Fig. 14.4. 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.

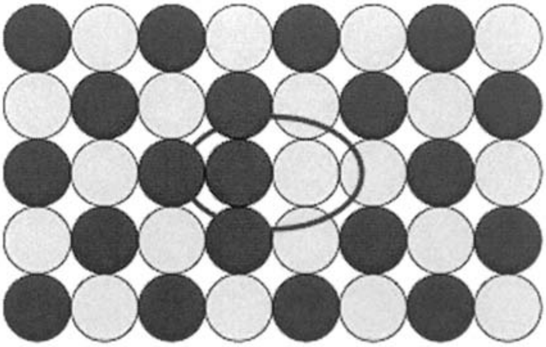


Fig. 14.4. Schematic diagram of an anti-site defect.

14.2.2. Extrinsic point defects

Extrinsic point defects, shown schematically in Fig. 14.5, 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 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.

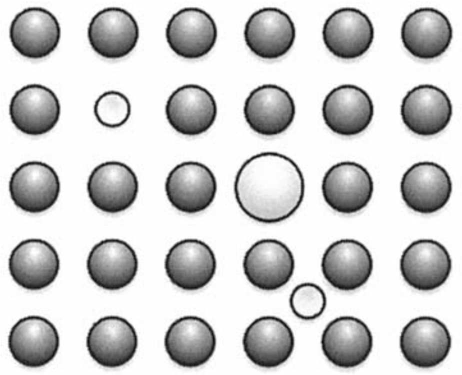


Fig. 14.5. Diagram of extrinsic point defects of substitutional impurities and an interstitial impurity.

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 Chapter 7.

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 14.2 (note: data about the most common dopants in Si, Ge, and GaAs was already listed in Table 7.1).

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		

Table 14.2. Impurity ionization energy (in meV) for several semiconductors. [Wolfe et al. 1989]

14.3. 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. 14.6.

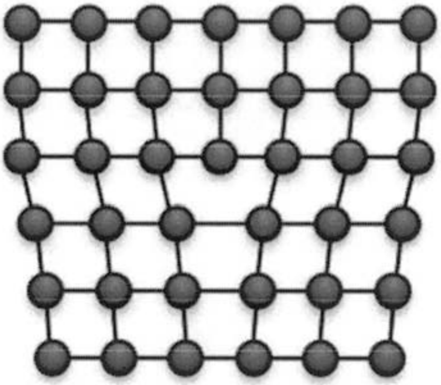


Fig. 14.6. Illustration of an edge dislocation.

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. 14.7.

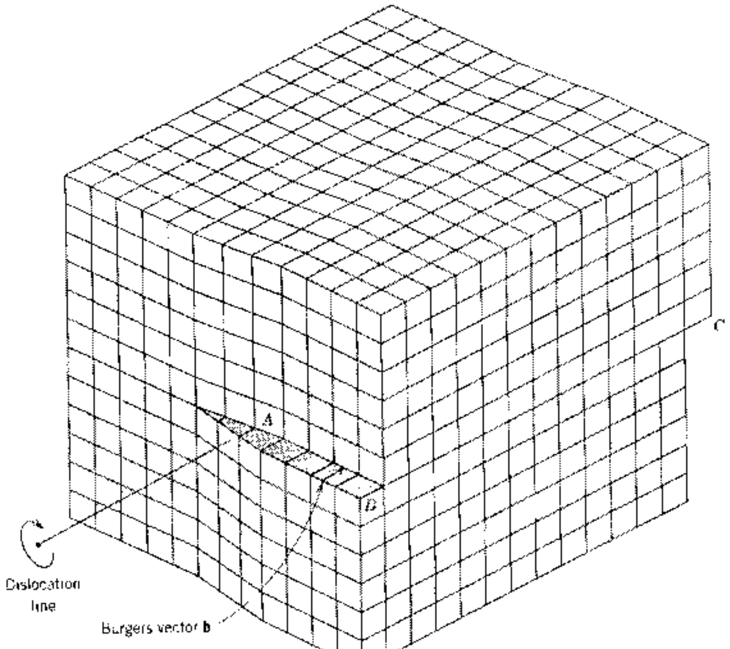


Fig. 14.7. Illustration of a screw dislocation. [Materials Science and Engineering: An Introduction, Callister, W.D. Copyright © 2000 by John Wiley & Sons, Inc. Reprinted with permission of Wiley-Liss Inc., a subsidiary of John Wiley & Sons, Inc.]

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. 14.8.

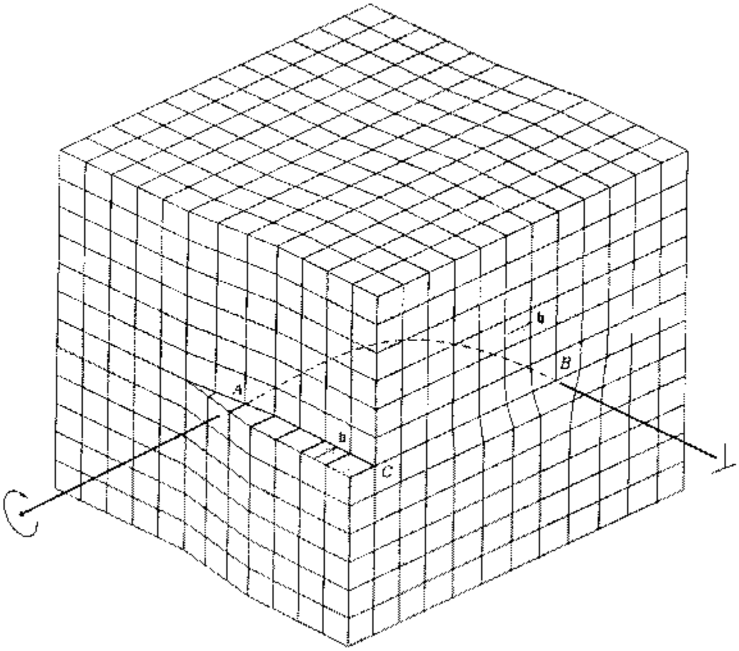


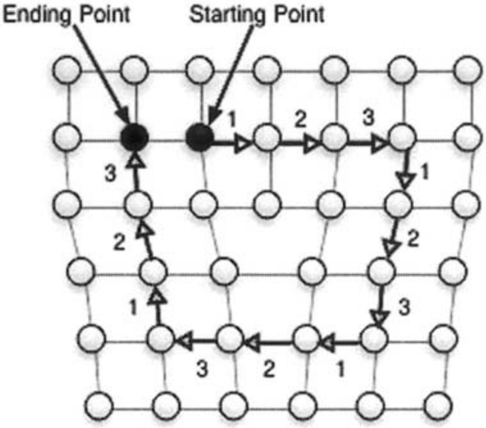
Fig. 14.8. Illustration of a mixed dislocation comprised of one edge dislocation and one screw dislocation. [Materials Science and Engineering: An Introduction, Callister, W.D. Copyright © 2000 by John Wiley & Sons, Inc. Reprinted with permission of Wiley-Liss Inc., a subsidiary of John Wiley & Sons, Inc.]

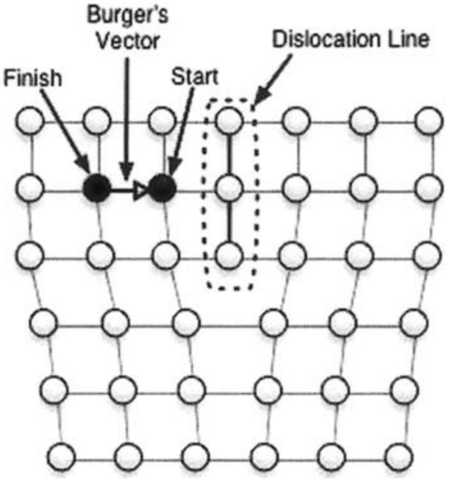
Burger's vectors are used to classify and describe dislocations. In order to construct a Burger's vector, a closed loop should be drawn around the dislocation by traveling the same amount of lattice points in all directions. If the loop does not close, it is surrounding a dislocation, and the vector that would close the circuit is the Burger's vector. The starting point, the circuit direction, and the size of the loop are arbitrary. Independent of these factors, the Burger's vector will always be perpendicular to the line of an edge dislocation and parallel to the line of a screw dislocation. It is often very complicated to find the Burger's vector for a mixed dislocation.

Example

Q: Draw the Burger's circuit to show that the Burger's 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 Burger's vector.





14.4. 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. Fig. 14.9 shows an example of a stacking fault.

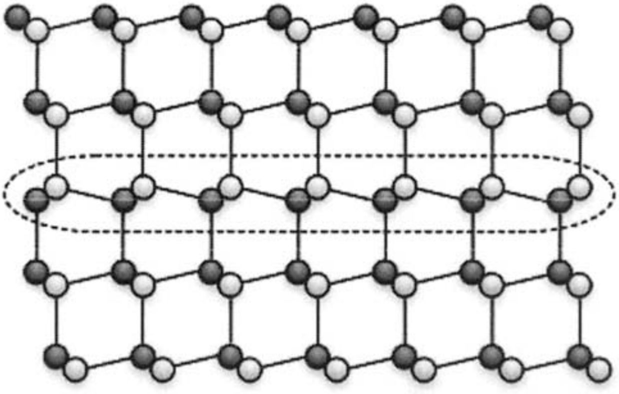


Fig. 14.9. Schematic diagram of a stacking fault.

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. 14.10.

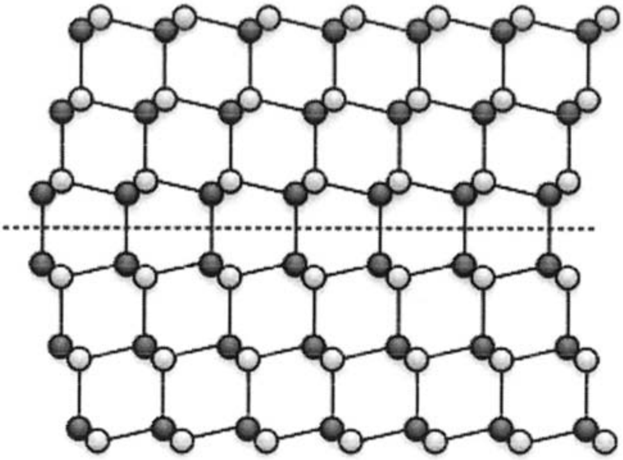


Fig. 14.10. Schematic diagram of a twin boundary.

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. 14.11.

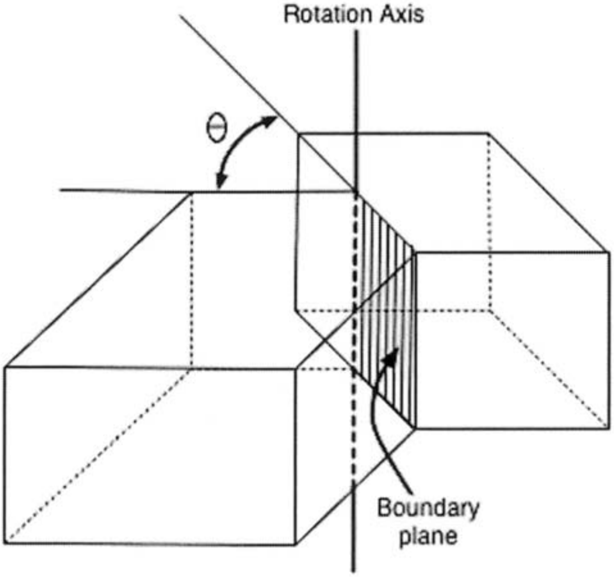


Fig. 14.11. Schematic diagram of a tilt boundary.

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 Fig. 14.12.

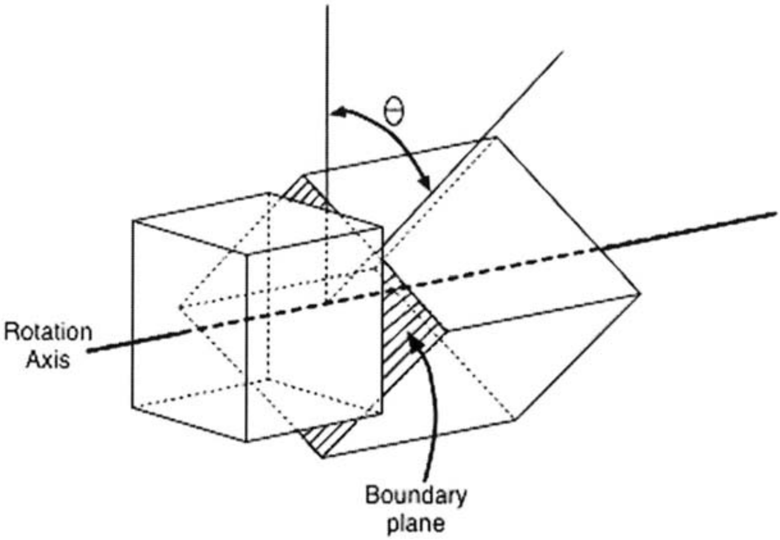


Fig. 14.12. Schematic diagram of a twist boundary.

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. (14.3), 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.

Eq. (14.3) $$D = \frac{b}{\sin\theta} \approx \frac{b}{\theta}$$

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 has 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. (14.4) where b is the magnitude of the Burger's 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):

Eq. (14.4) $$d_{crit} = \frac{b}{10 \cdot f}$$

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. 14.13.

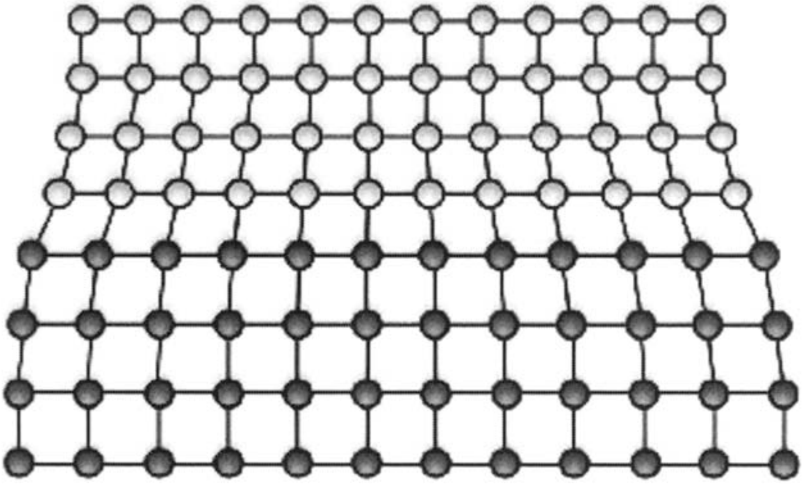


Fig. 14.13. Schematic of a coherent interphase boundary.

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. 14.14.

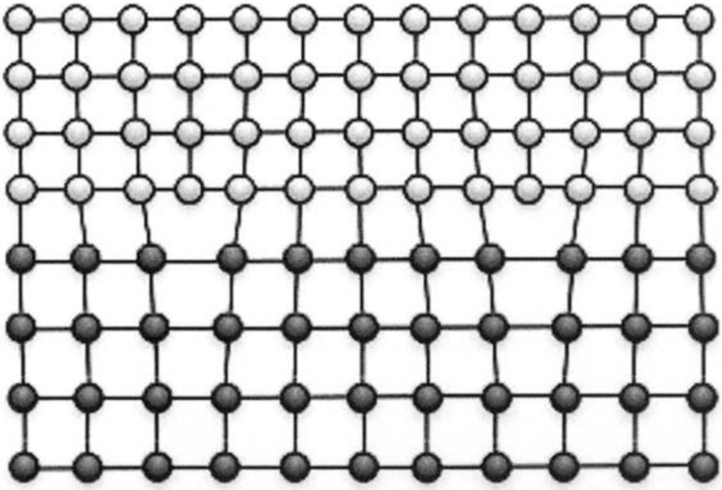


Fig. 14.14. Schematic of a semi-coherent interphase boundary.

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.

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.

14.5. 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. 14.15 shows an SEM image of voids in AlGaN.

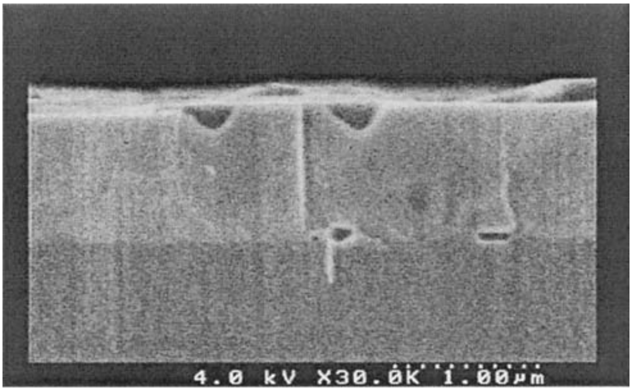


Fig. 14.15. SEM image of voids in AlGaN

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} 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.

14.6. Defect characterization

Characterization and analysis of defects is one of the biggest experimental challenges. There are conventional characterization methods to examine the over all quality or electrical features of the material such as Hall measurement and x-ray measurement (see Chapter 13). 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.

14.7. 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 become more significant. In particular, the aggregation of vacancies which results in the formation of microvoids are 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 Chapter 12.

14.8. Summary

In this Chapter we discussed defects as imperfections that disturb the periodic structure of the crystal. The defects were classified into 4 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.

References

Callister, W.D., *Materials Science and Engineering: An Introduction*, 5th ed., John Wiley and Sons, New York, 2000.

Wolfe, C.M., Holonyak, Jr., N., and Stillman, G.E., *Physical properties of Semiconductors*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

Further reading

Adachi, S., *Physical Properties of III-V Semiconductor Compounds: InP, InAs, GaAs, GaP, InGaAs, and InGaAsP*, Wiley, New York, pp. 263-286, 1992.

Anderson, J.C., Leaver, K.D., Leevers, P., and Rawlings, R.D., *Materials Science for Engineers*, Nelson Thornes Ltd, Cheltenham, UK, 2003.

Bongiorno, A., Colombo, L., and Diaz de la Rubia, T., "Structural and binding properties of vacancy clusters in silicon," *Europhysics Letters* 43, pp. 695-700, 1998.

Hayes, W. and Stoneham, A.M., *Defects and Defect Processes in Nonmetallic Solids*, Dover Publications, New York, 2004.

Hurle, D.T.J. and Rudolph, P., "A brief history of defect formation, segregation, faceting, and twinning in melt-grown semiconductors," *Journal of crystal growth* 264, pp. 550-564, 2004.

Hurle, D.T.J., "Point defects in compound semiconductors," in *Crystal Growth-From Fundamentals to Technology*, eds. G. Muller, J.J. Metois, and P. Rudolph, Elsevier, Oxford, pp. 323-343, 2004.

Kittel, C., *Introduction to Solid State Physics*, John Wiley & Sons, New York, 1986.

Murr, L.E., *Solid-State Electronics*, Marcel Dekker, New York, 1978.

Nalawa, H.S., Bloembergen, N., and Laureate, N., *Handbook of Advanced Electronic and Photonic Materials and Devices*, Academic Press, San Diego, 2000.

Shaffner, T.J., "Characterization challenges for the ULSI era," in *Proceedings of the Electrochemical Society Symposium on Diagnostic Techniques for Semiconductor Materials and Devices*, eds. P. Rai-Choudhury, J.L. Benton, D.K. Schroder, and T.J. Shaffner, Electrochemical Society, Pennington, NJ, pp. 1-13, 1997.

Swaminathan, V. and Macrander, A.T., *Materials Aspects of GaAs and InP Based Structures*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

Problems

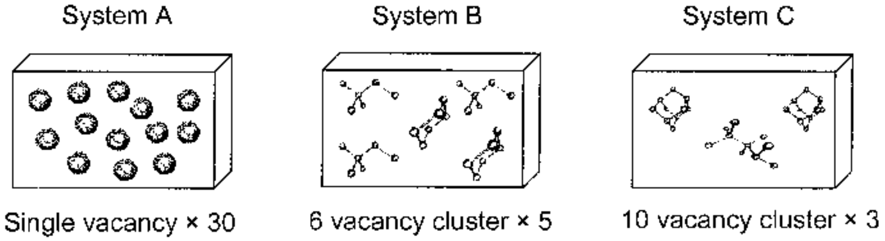
1. Give some examples of physical properties that defects can change.

2. Identify the types of point defects shown in Fig. 14.1. 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 $g.cm^{-3}$ and 55.85 $g.mol^{-1}$, 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 Europhysics Letters Vol. 43, 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 Burger's 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 $b = \frac{a_{InAs}}{\sqrt{2}}$).

9. What is preferential etching?

10. What have been the goals of the semiconductor industry in silicon crystal growth technology? Why?

15. Semiconductor Device Technology

15.1. Introduction

In the previous Chapters, we have reviewed the various techniques used to synthesize semiconductor crystals and thin films. This represented only the first step in the fabrication of semiconductor devices. Several additional steps are necessary before a final product can be obtained, which will be described in this and the following Chapter.

In this Chapter, the discussion will be inspired from the silicon device technology because of its technological predominance and maturity in modern semiconductor industry. We will first describe and model the oxidation process used to realize a silicon oxide film. We will then discuss the diffusion and ion implantation of dopant impurities in silicon to achieve controlled doping, and review the methods used to characterize their electrical properties. Although this Chapter discusses silicon, the methods can be equally applied to all types of semiconducting materials.

15.2. Oxidation

The ability to form a chemically stable protective layer of silicon dioxide (SiO_2) at the surface of silicon is one of the main reasons that makes silicon the most widely used semiconductor material. This silicon oxide layer is a high quality electrically insulating layer on the silicon surface, serving as a dielectric in numerous devices, that can also be a preferential masking layer in many steps during device fabrication. In this section, we will first review the experimental process of the formation of a silicon oxide. Then we will develop a mathematical model for it and determine the factors influencing the oxidation. We will then end this section by providing details on how to characterize the thickness of the formed oxide.

15.2.1. Oxidation process

A silicon dioxide layer is often thermally formed in the presence of oxygen compounds at a temperature in the range of 900 to 1300 °C. There exists two basic means of supplying the necessary oxygen into the reaction chamber. The first is in gaseous pure oxygen form (dry oxidation) through the reaction: $Si + O_2 \rightarrow SiO_2$. The second is in the form of water vapor (wet oxidation) through the reaction: $Si + 2H_2O \rightarrow SiO_2 + 2H_2$. For both means of oxidation, the high temperature allows the oxygen to diffuse easily through the silicon dioxide. The silicon is consumed as the oxide grows, and with a total oxide thickness of X, about $0.45X$ lies below the original surface of the Si wafer and $0.55X$ lies above it, as shown in Fig. 15.1. A typical oxidation growth cycle consists of dry-wet-dry oxidations, where most of the oxide is grown in the wet oxidation phase. Dry oxidation is slower and results in more dense, higher quality oxides. This type of oxidation method is used mostly for metal-oxide-semiconductor (MOS) gate oxides. Wet oxidation results in much more rapid growth and is used mostly for thicker masking layers.

Before thermal oxidation, the silicon is usually preceded by a cleaning sequence designed to remove all contaminants. Special care must be taken during this step to guarantee that the wafers do not contact any source of contamination, particularly inadvertent contact with a human person. Humans are a potential source of sodium, the element most often responsible for the failure of devices due to surface leakage. Sodium contamination can be reduced by incorporating a small percentage of chlorine into the oxidizing gas. Next, the cleaned wafers are dried and loaded into a quartz wafer holder called a boat.

The thermal oxidation process is performed with the wafers sitting in the boat loaded into a furnace where the temperature is carefully controlled. Generally, three or four separate furnaces are used in a stack manner, each with its own set of controls and quartzware. The quartz tube inside each furnace is enclosed around heating coils which are controlled by the amount of electrical current running through. A cross-section of a typical oxidation furnace is shown in Fig. 15.1.

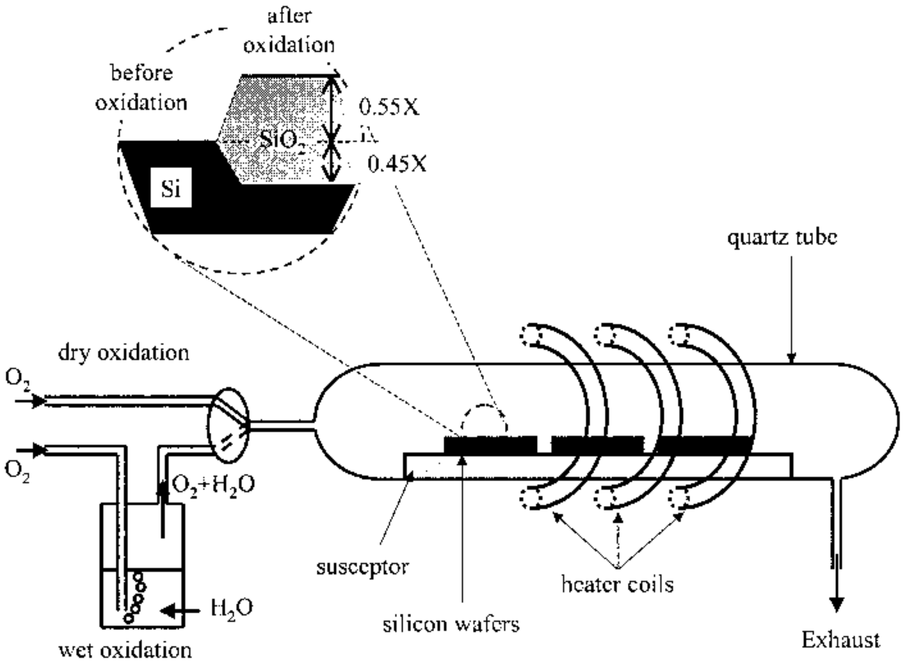


Fig. 15.1. Cross-section of an oxidation furnace: a quartz tube, heated by coils surrounding it, contains the silicon wafers in which either dry oxygen gas or water vapor can be introduced to provide the oxidizing gas. On the top left, the cross-section at the surface of a silicon wafer before and after oxidation is shown.

The furnace is suitable for either dry or wet oxidation film growth by turning a control valve. In the dry oxidation method, oxygen gas is sent into the quartz tube. High-purity gas is used to ensure that no unwanted impurities are incorporated in the layer of oxide as it forms. The oxygen gas can also be mixed with pure nitrogen gas in order to decrease the total cost of running the oxidation process, as nitrogen gas is less expensive than oxygen. In the wet oxidation method, the water vapor introduced into the furnace system is created by flowing a carrier gas into a container or bubbler filled with ultra pure water and maintained at a constant temperature below its boiling point (100 °C). The carrier gas can be either nitrogen or oxygen, and both result in equivalent oxide thickness growth rates. As the gas bubbles through the water, it becomes saturated with the water vapor. The distance to the quartz oxidation tube must be short enough to prevent condensation of the water vapor. The bubblers used in the wet oxidation process are simple and quite reproducible, but they have two disadvantages associated with the fact that they must be refilled when the water level falls too low: an improper handling of the container can result in the contamination of the water prior or during filling, and the bubbler cannot be filled during an oxidation cycle.

15.2.2. Modeling of oxidation

Using radioactive tracer experiments, the oxygen or water molecules in a dry oxidation process were found to move through the oxide film and react with the silicon atoms at the interface between the oxide film and silicon. As the oxide grows, the growth rate of the oxide layer decreases because the oxygen must pass through more oxide to reach and combine with the silicon. This is schematically illustrated in Fig. 15.2. The movement of these molecules through the forming oxide layer can be mathematically modeled using the same Fick's first law of diffusion as introduced in section 8.5.

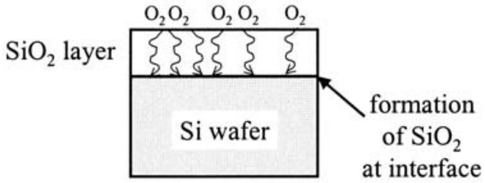

O_2 O_2 O_2 O_2 O_2
SiO_2 layer
Si wafer
formation of SiO_2 at interface

Fig. 15.2. Formation of SiO_2 in a dry oxidation process. The oxygen molecules diffuse through the existing oxide film until they reach the oxide-silicon interface where they react with silicon atoms to continue to form an oxide.

The objective of the following mathematical model is to determine the growth rate of the oxide layer, that is how fast the oxide layer forms. We will follow a similar approach to the one taken for vapor phase epitaxy in sub-section 12.5.2. In this model, we consider that there is a flow of a gas containing oxygen, called the oxidant, onto the sample surface, which we assume diffuses through the existing oxide layer and reacts with the underlying silicon. We will consider three different fluxes (units of particles per $cm^2.s^{-1}$) of oxidant, each governed by a different physical mechanism. These fluxes are shown in Fig. 15.3.

The first one is the flux of oxidant from the bulk gas phase onto the sample surface, denoted F_1. This flux is proportional to the difference in concentration of oxidant between the bulk gas phase and at the surface of the forming oxide:

Eq. (15.1) $$F_1 = h_G\left[C_G - C_s\right]$$

where h_G is the vapor phase mass transfer coefficient, C_G denotes the oxidant concentration in the bulk gas phase, and C_S denotes that at the surface of the forming oxide. These concentrations are generally different because some oxidant is consumed in the oxidation process. These concentrations are directly related to the partial pressures of the oxidant gas in the bulk gas phase, P_G, and at the oxide surface, P_S, through the ideal gas law:

Eq. (15.2) $$\begin{cases} C_G = \dfrac{P_G}{k_b T} \\ C_S = \dfrac{P_S}{k_b T} \end{cases}$$

where k_b is the Boltzmann constant and T is the absolute temperature.

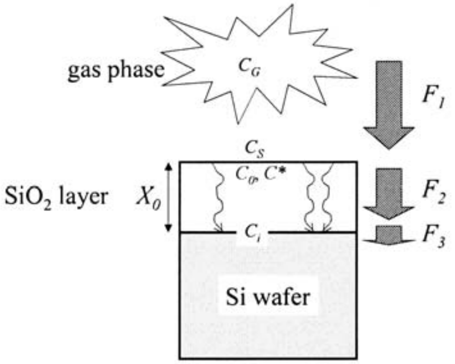


Fig. 15.3. Model for the thermal oxidation of silicon. F_1 represents the flux of oxidant from the bulk gas phase onto the sample surface, F_2 represents the flux of oxidant diffusing through the existing oxide, and F_3 represents the flux of oxidant which reaches the oxide-silicon interface and is consumed through chemical reaction with the silicon.

We can relate the oxidant concentration in the gas with the oxidant concentration in the solid phase, i.e. the oxide layer, near the surface through Henry's law:

Eq. (15.3) $C_0 = K_H P_S$

where C_0 is the oxidant concentration inside the oxide layer just below its surface, K_H is Henry's law constant and P_S is the partial pressure of the oxidant in the gas phase at the oxide surface.

It will be convenient to introduce the equilibrium value of C_0, which will be denoted C^*. This concentration is related to the partial pressure in the bulk of the gas P_G through:

Eq. (15.4) $C^* = K_H P_G$

Combining Eq. (15.2), Eq. (15.3) and Eq. (15.4), Eq. (15.1) can then be successively written as:

Eq. (15.5) $$F_1 = h_G\left[\frac{P_G}{k_b T} - \frac{P_S}{k_b T}\right] = \frac{h_G}{k_b T}\left[\frac{C^*}{K_H} - \frac{C_0}{K_H}\right] = h\left[C^* - C_0\right]$$

where we have defined:

Eq. (15.6) $$h = \frac{h_G}{k_b K_H T}$$

The second flux, denoted F_2, to consider is that of the oxidant diffusing through the oxide layer already present which can be expressed as:

Eq. (15.7) $$F_2 = \frac{D}{X_o}\left(C_0 - C_i\right)$$

where D is the diffusion coefficient of the oxidant through the oxide, C_i is the oxidant concentration at the oxide-silicon interface, and X_0 is the thickness of the oxide.

The third flux, denoted F_3, corresponds to the incorporation of oxidant molecules which reach the oxide-silicon interface and react chemically to expand the oxide. This can be expressed as:

Eq. (15.8) $F_3 = k_s C_i$

where k_S is the chemical reaction constant for the formation of oxide.

Under steady-state conditions, these three fluxes must be equal:

Eq. (15.9) $F_1=F_2=F_3=F$

This gives us three equations, for the three unknowns: C_0, C^* and C_i. Using Eq. (15.7) and Eq. (15.8) to equate F_2 and F_3, we get:

Eq. (15.10) $$C_0 = \left(1+\frac{k_S X_0}{D}\right)C_i$$

Now, using Eq. (15.5) and Eq. (15.8) to equate F_1 and F_3, we get:

$$C^* = C_0 + \frac{k_s}{h}C_i$$

which, after considering Eq. (15.10), becomes:

Eq. (15.11) $$C^* = \left(1+\frac{k_S X_0}{D}+\frac{k_S}{h}\right)C_i$$

It is convenient to rearrange these relations to express C_0 and C_i as a function of C^*:

Eq. (15.12) $$C_i = \frac{1}{\left(1+\frac{k_S X_0}{D}+\frac{k_S}{h}\right)}C^*$$

Eq. (15.13) $$C_0 = \frac{\left(1+\frac{k_S X_0}{D}\right)}{\left(1+\frac{k_S X_0}{D}+\frac{k_S}{h}\right)}C^*$$

We can now consider a particular case. If we assume that $h>>k_s$, i.e. the oxidation reaction at the oxide-silicon interface is much slower than the arrival of oxidant at the oxide surface, the oxidation process is then said to be interfacial reaction controlled. The Eq. (15.12) and Eq. (15.13) can then be simplified into:

Eq. (15.14) $$\begin{cases} C_i \approx \dfrac{1}{\left(1+\dfrac{k_s X_0}{D}\right)} C^* \\ C_0 \approx C^* \end{cases}$$

Combining Eq. (15.8) and Eq. (15.14) to eliminate C_i, we can express the flux F as a function of C_0:

Eq. (15.15) $$F = \frac{k_s}{\left(1+\dfrac{k_s X_0}{D}\right)} C_0$$

The rate at which the oxide layer grows is then given by the flux divided by the number N of oxidant molecules that can be incorporated into a unit volume of oxide:

Eq. (15.16) $$\frac{dX_o}{dt} = \frac{F}{N} = \frac{1}{N} \frac{k_s C_0}{\left(1+\dfrac{k_s X_0}{D}\right)}$$

For dry oxidation, N=2.2×10^{22} molecules per cm^3, while for wet oxidation N=4.4×10^{22} molecules per cm^3. Integrating Eq. (15.16) and using the boundary condition $X_0(t=0)=X_i$, yields the following equation for X_0:

Eq. (15.17) $$X_o^2 + AX_o = B(t+\tau)$$

where τ is an integration constant and where we have denoted:

Eq. (15.18) $$\begin{cases} A = \dfrac{2D}{k_s} \\ B = \dfrac{2DC_0}{N} \\ \tau = \dfrac{X_i^{\ 2}}{B} + \dfrac{X_i}{B/A} \end{cases}$$

where X_i is the initial thickness of the oxide. For dry oxidation, an initial oxide thickness of 250 Å must be accounted for by letting X_i=25 nm, in order to make Eq. (15.17) universal to both oxidation methods.

Solving for the oxide thickness in Eq. (15.17) as a function of oxidation time t, one obtains the following positive expression for X_0:

Eq. (15.19) $$X_0 = \frac{A}{2}\left\{\sqrt{1+\frac{(t+\tau)}{A^2/4B}}-1\right\}$$

The growth time t is given directly by Eq. (15.17):

Eq. (15.20) $$t = \frac{A^2}{4B}\left[\left(\frac{2X_0}{A}+1\right)^2-1\right]-\tau$$

For the limiting case of "short oxidation time", where $(t+\tau)<<A^2/4B$, we can simplify the expression in Eq. (15.19):

$$X_0 \approx \frac{A}{2}\left\{1+\frac{1}{2}\frac{(t+\tau)}{A^2/4B}-1\right\}$$

which is obtained after considering the Taylor expansion of the square root. We then obtain the so-called *linear oxidation law*:

Eq. (15.21) $$X_0 = \frac{B}{A}(t+\tau)$$

where B/A is the linear growth rate constant, and can be calculated using Eq. (15.18) and Table 15.1.

For the other limiting case of "long oxidation time", when $t>>A^2/4B$, one obtains the *parabolic oxidation law*:

Eq. (15.22) $$X_o^2 = Bt$$

where B is the parabolic growth rate constant, and can be calculated using Eq. (15.18) and Table 15.1.

15.2.3. Factors influencing oxidation rate

Numerous factors can influence the oxidation rate by governing each of the mechanisms discussed in the previous model. For example, one of them is the diffusion coefficient in Eq. (15.7). This parameter generally follows an Arrhenius relationship as given by:

$$\text{Eq. (15.23)} \quad D = D_0 \exp\left(-\frac{E_A}{k_b T} \right)$$

where k_b is the Boltzmann constant, E_A is the activation energy, and T is the temperature. Values for activation energy and D_0 coefficient can be found in Table 15.1. This relation indicates the strong dependence of oxide growth rate on temperature as the diffusion rate of the oxidant increases exponentially with temperature.

There exist four other factors which are commonly known to affect the oxidation rate of silicon: type of oxidation, orientation of the silicon wafer, pressure and impurity effects. For the type of oxidation, wet oxidation has a higher growth rate due to the higher solubility of the water vapor. The orientation dependence of the oxidation rate can be easily understood because the oxidation process depends on the total number of available Si atoms per unit area for oxidation at the oxide-silicon interface. Only the linear oxidation rate is expected to significantly change as a function of orientation, i.e. for short oxidation durations. For example, the oxidation rate for (111) oriented Si is faster than that for (100) oriented Si initially, in the linear region, as shown in Fig. 15.4(a) and (b). As the oxidation kinetics change from the linear rate to the parabolic rate, i.e. for longer oxidation durations, the difference between the two orientations diminishes. The pressure is proportional to the number of oxidants, and is directly proportional to both linear and parabolic growth rate constants. As can be seen in Eq. (15.17), an increase in pressure results in a slower growth rate.

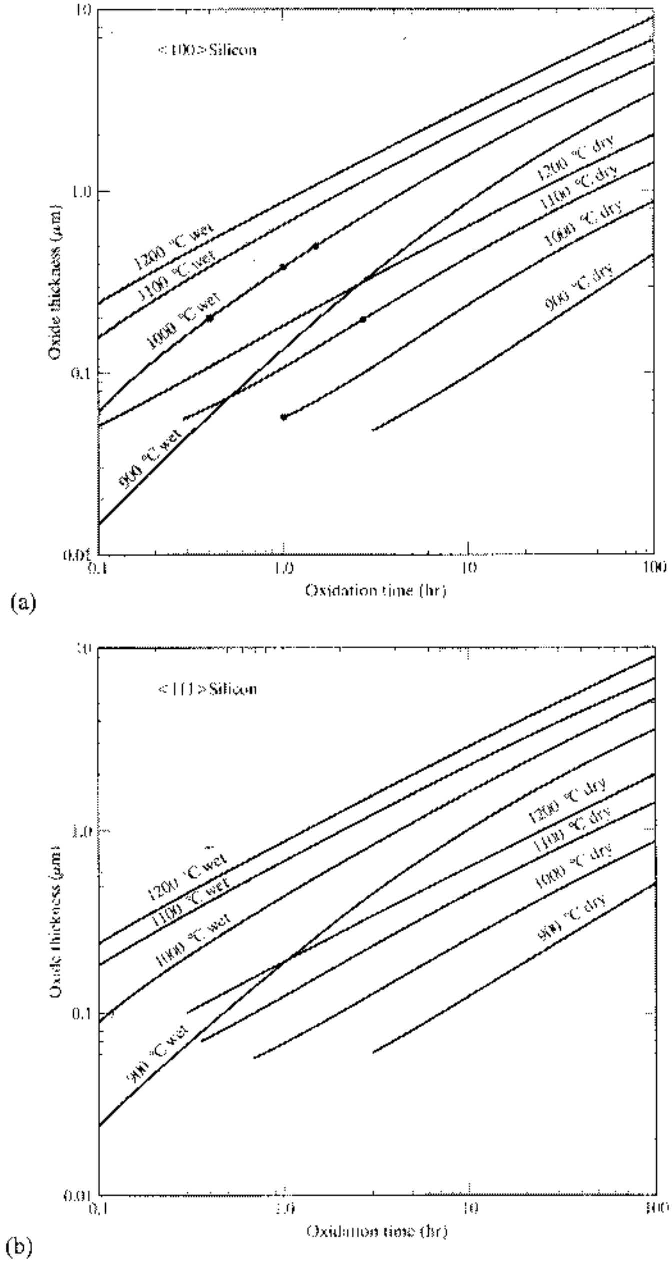


Fig. 15.4. Oxide thickness as a function of oxidation time under various conditions: (a) wet and dry oxidation of (100) silicon at several temperatures, (b) wet and dry oxidation of (111) silicon at various temperatures. [JAEGER, RICHARD C., INTRODUCTION TO MICROELECTRONICS FABRICATION: VOLUME 5 OF MODULAR SERIES ON SOLID STATE DEVICES, 2^{nd} Edition, © 2002. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

	Wet O_2 (X_i=0 nm)		Dry O_2 (X_i=25 nm)	
	D_0	E_A (eV)	D_0	E_A (eV)
<100> Si Linear	9.7×10^7 μm.hr^{-1}	2.05	3.71×10^6 μm.hr^{-1}	2.00
<100> Si Parabolic	386 μm^2.hr^{-1}	0.78	772 μm^2.hr^{-1}	1.23
<111> Si Linear	1.63×10^8 μm.hr^{-1}	2.05	6.23×10^6 μm.hr^{-1}	2.00
<111> Si Parabolic	386 μm^2.hr^{-1}	0.78	772 μm^2.hr^{-1}	1.23

Table 15.1. D_0 coefficient values and activation energy E_A for wet and dry oxygen for different types of silicon. [Jaeger 1988.]

15.2.4. Oxide thickness characterization

The accurate measurement of the thickness of a dielectric film such as silicon dioxide is very important in the fabrication of optoelectronic devices. Various techniques are available for measuring this oxide thickness, including optical interference, ellipsometry, capacitance, and the use of a color chart.

The optical interference method is a simple and nondestructive technique, which can be used to routinely measure thermal oxide thickness from less than 100 Å to more than 1 μm. The method is based on characterizing the interference pattern created by light reflected from the air/SiO_2 interface and that from the Si/SiO_2 interface, as illustrated in Fig. 15.5.

The equation governing this interference is:

Eq. (15.24) $$X_0 = \frac{\lambda(g-\Delta\varphi)}{2n^*}$$

where X_0 is the thickness of the oxide, λ is the wavelength of the incident radiation, g the order of the interference, and $\Delta\varphi$ is the net phase shift and is equal to φ_s-φ_o where φ_o is the phase shift at the air/SiO_2 interface and φ_s is the phase shift at the Si/SiO_2 interface. The parameter n^* is given by:

Eq. (15.25) $$n^* = \sqrt{\overline{n}_i^2 - \sin^2\theta}$$

where $\overline{n}_i$ is the refractive index of the oxide film and θ is the angle of incidence of the light relative to the wafer. All these parameters are illustrated in Fig. 15.5.

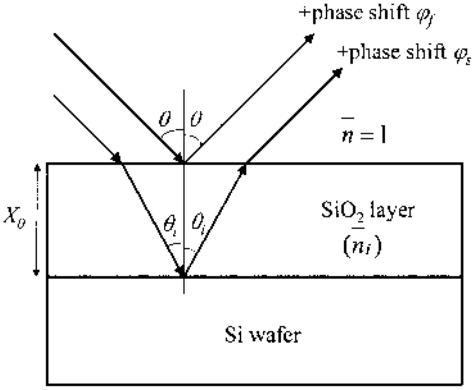


Fig. 15.5. Optical interference method for the measurement of oxide film thickness. Two rays of light with the same wavelength are shown incident on the wafer. One of them is reflected from the oxide-air interface. The other enters the oxide layer which has a different refractive index than air and is reflected at the oxide-silicon interface. A difference in optical path occurs between these two rays of light and a phase shift difference results. If the phase shift difference is an integer multiple of 2π, these two reflected rays of light interfere constructively, whereas if the phase shift difference is a half integer multiple of 2π, these rays interfere destructively.

The second method for the measurement of the oxide film thickness is ellipsometry. Ellipsometry is the most popular technique used to assess the properties of silicon dioxide films. Ellipsometry provides a non-destructive technique for accurately determining the oxide thickness, as well as the refractive index at the measuring wavelength. An illustration of an ellipsometry system is shown in Fig. 15.6. It is the most widely used tool to measure the refractive index of a wide variety of materials on any substrate, in particular SiO_2, Si_3N_4, photoresist, and aluminum oxide (Al_2O_3) on silicon substrates. Such systems can measure film thickness in the range of 20 Å to 60,000 Å with an accuracy of ±2 %. An ellipsometer operates by shining polarized monochromatic light onto the wafer surface at an angle. The light is then reflected from both the oxide and the silicon surface. A phase modulation unit, numerical data acquisition and processing system work together to measure the difference in polarization. The result is then used to calculate the oxide thickness.

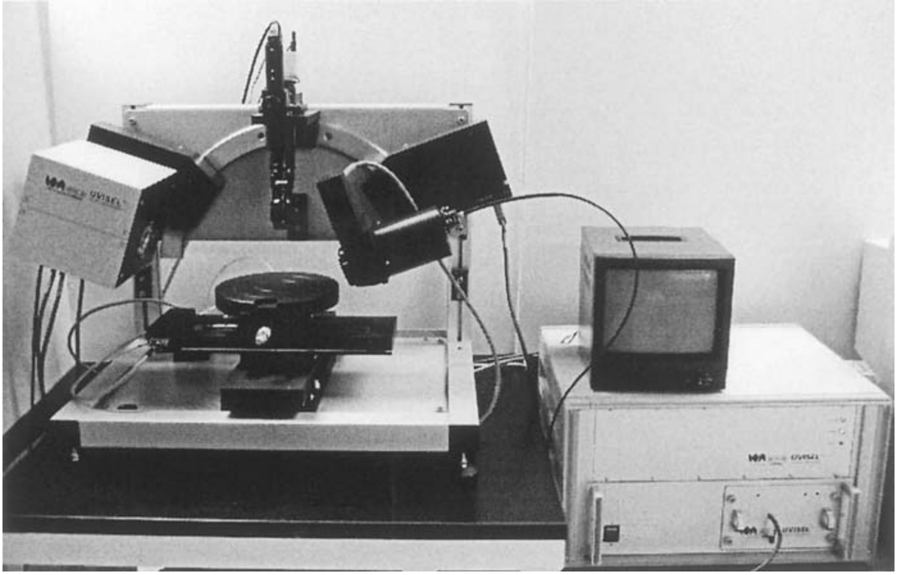


Fig. 15.6. A typical ellipsometer system, including a light source and its power supply, a sample stage, a detector and the analyzing circuits.

The third oxide film thickness measurement technique is the capacitance method, which requires the fabrication of a metal-oxide-semiconductor (MOS) capacitor. The oxide thickness is given by the following equation:

$$X_0 = \frac{A_g \varepsilon_{ox} \varepsilon_0}{C_{ox}}$$

where C_{ox} is the experimentally measured oxide capacitance, A_g is the area of the capacitor, ε_{ox} is the dielectric constant of the oxide film, and ε_0 the permittivity in vacuum.

Finally, the fourth and simplest method used to measure an oxide film thickness is by comparing the film color with a calibrated chart as shown in Table 15.2 for SiO_2. Each oxide thickness has a specific color when it is viewed under white light perpendicular to its surface. The colors are cyclically repeated for different orders of reflection.

Film thickness (µm)	Color
0.05	Tan
0.07	Brown
0.10	Dark violet to red violet
0.12	Royal blue
0.15	Light blue to metallic blue
0.17	Metallic to very light yellow green
0.20	Light gold to yellow, metallic
0.22	Gold with slight yellow orange
0.25	Orange to melon
0.27	Red violet
0.30	Blue to violet blue
0.31	Blue
0.32	Blue to blue green
0.34	Light green
0.35	Green to yellow green
0.36	Yellow green
0.37	Green yellow
0.39	Yellow
0.41	Light orange
0.42	Carnation pink
0.44	Violet red
0.46	Red violet
0.47	Violet
0.48	Blue violet
0.49	Blue
0.50	Blue green
0.52	Green (borad)
0.54	Yellow green
0.56	Green yellow
0.57	Yellow to "yellowish"
0.58	Light orange or yellow to pink borderline
0.60	Carnation pink
0.63	Violet red
0.68	Bluish
0.72	Blue green to green
0.77	"Yellowish"
0.80	Orange
0.82	Salmon
0.85	Dull, light red violet
0.86	Violet
0.87	Blue violet
0.89	Blue
0.92	Blue green
0.95	Dull yellow green
0.97	Yellow to "yellowish"
0.99	Orange
1.00	Carnation pink
1.02	Violet red
1.05	Red violet
1.06	Violet
1.07	Blue violet
1.10	Green
1.11	Yellow green
1.12	Green
1.18	Violet
1.19	Red violet
1.21	Violet red
1.24	Carnation pink to salmon
1.25	Orange
1.28	"Yellowish"
1.32	Sky blue to green blue
1.40	Orange
1.45	Violet
1.46	Blue violet
1.50	Blue
1.54	Dull yellow green

Table 15.2. SiO_2 oxide film color chart.

Example

Q: Using the Deal-Grove oxidation model, calculate the time needed to grow a 150 nm thick oxide on top of (100) silicon by wet oxidation at a temperature of 1000 °C.

A: T=1000 °C=1273 K

From Table 15.1 we obtain the value of the pre-exponential factor D_0=3.71×10^6 μm.hr^{-1} (linear oxidation) and E_A=2.00 eV. Using Eq. (15.23) we determine the diffusion coefficient D and then the ratio:

$$B/A = 3.71\times10^6 \exp[-2.00/(8.617\times10^{-5})(1273)]$$

$$= 0.04478\,\mu m.hr^{-1}$$

From Table 15.1 we obtain D_0=772 μm^2.hr^{-1} (parabolic oxidation) and E_A=1.23 eV. Using Eq. (15.23) we calculate D and then the value for B:

$$B = 772\exp[-1.23/(8.617\times10^{-5})(1273)]$$

$$= 0.01042\,\mu m^2.hr^{-1}$$

We can find the necessary oxidation time by using Eq. (15.17), which when rearranged becomes:

$$t = \frac{X_0{}^2}{B} + \frac{X_0}{B/A} - \tau$$

Since we are using wet oxidation, X_i=0 so τ=0,

$$\Rightarrow \quad t = \frac{(0.15)^2}{0.04478} + \frac{0.15}{0.10420} = 5.5\,.$$

Therefore, 5.5 hours is needed to grow a 150 nm thick oxide layer on (100) silicon using wet oxidation at 1000 °C.

15.3. Diffusion of dopants

In section 7.6, we discussed doping as a means to control the electrical properties in semiconductors. Doping is achieved by replacing the constituting atoms of the semiconductor with atoms which contain fewer or more electrons. Through doping, the crystal composition is thus slightly altered so that it contains either a higher concentration of electrons or holes, which makes the semiconductor *n*-type or *p*-type, respectively.

The doping of semiconductors can be performed during the bulk crystal or epitaxial film growth (Chapter 12) by introducing the dopant along with the precursor chemicals. This way, the entire crystal or film is uniformly doped with the same concentration of dopants. Another method consists of carrying out the doping after the film deposition by performing the diffusion or the implantation of dopants. These have the advantage that the doping can be localized to certain regions only, by using an adequate mask to prevent the doping in undesired areas. In this section, we will focus on the diffusion of dopants and we will illustrate our discussion with the doping of silicon.

The diffusion of dopants in compound semiconductor epitaxial films generally follows a similar model. However, the effects of diffusion doping in compound semiconductor heterostructures are subtler and have been discussed in detail in specialized texts [Razeghi 1989].

15.3.1. Diffusion process

The concept of diffusion has been briefly introduced in section 8.5. Diffusion is the process whereby a particle moves from regions of higher concentrations to regions of lower concentrations. The process could be visualized by thinking of a drop of black ink dropped into a glass of clean water. Initially, the ink stays in a localized area, appearing as a dark region in the clean water. Gradually, some of the ink moves away from the region of high concentration, and instead of there being a dark region and a clean region, there is a graduation of colors. As time passes, the ink spreads out until it is possible to see through it. Finally, after a very long time, a steady state is reached and the ink is uniformly distributed in the water. The movement of the ink from the region of high concentration (ink drop) to the region of low concentration (the rest of the glass of water) is an illustration of the process of diffusion.

In the doping of silicon by diffusion, the silicon wafer is placed in an atmosphere containing the impurity or dopant to incorporate. Because the silicon does not initially contain the dopant in its lattice, we are in the presence of two regions with different concentrations of impurities. At high temperatures (900 to 1200 °C), the impurity atoms can move into the crystal and diffusion can therefore occur, as schematically illustrated in Fig. 15.7.

The wafers are loaded vertically into a quartz boat and put into a furnace similar to the furnace used for oxidation. There are three types of sources to be used for the dopant atoms: solid, liquid, and gas, as shown in Fig. 15.8.

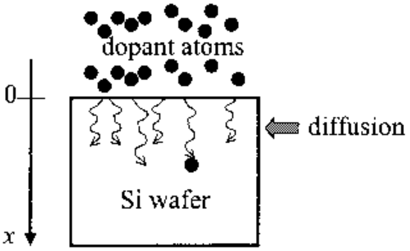


Fig. 15.7. Diffusion of dopants in a silicon wafer. The wafer is placed in an atmosphere containing the dopant. The gradient of dopant concentration between the atmosphere and the silicon crystal leads to their diffusion into the silicon.

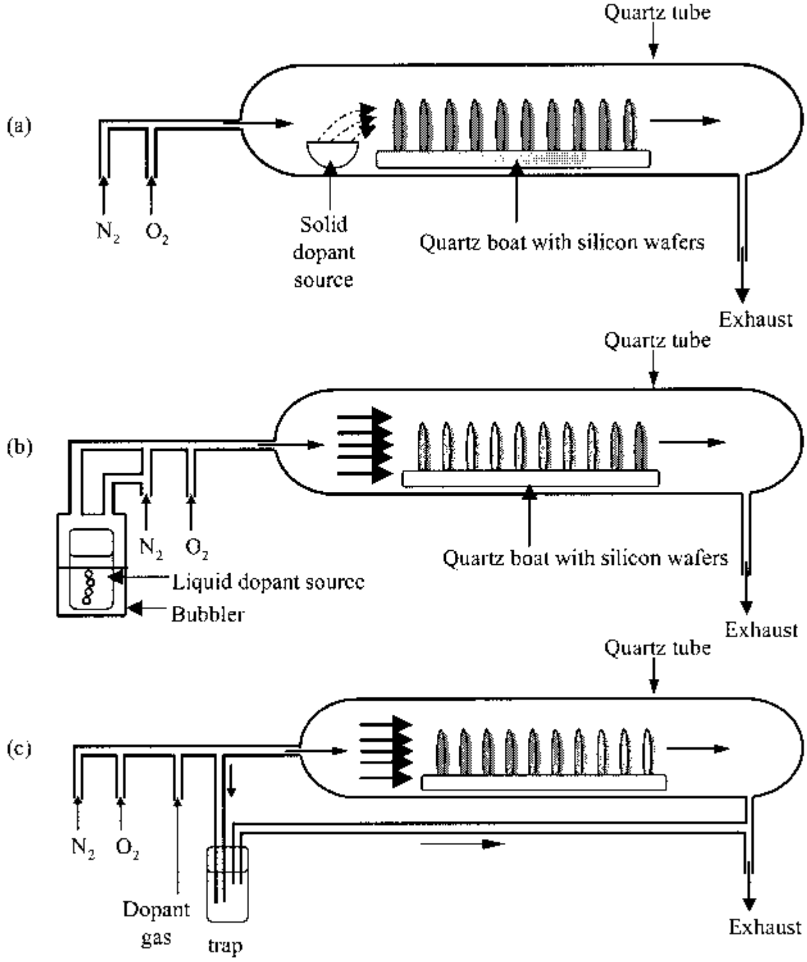


Fig. 15.8. Diffusion furnaces. (a) solid source diffusion with the source in a platinum source boat, (b) liquid source diffusion with the carrier gas passing through the bath, and (c) gas source diffusion with gaseous impurity sources.

There exist several types of diffusion mechanisms. An impurity can diffuse into an interstitial site in the lattice and can move from there to another interstitial site, as shown in Fig. 15.9(a). We then talk about interstitial diffusion Sometimes a silicon atom can be knocked into an interstitial site, leaving a vacancy in the lattice where a diffusing dopant atom can fit, as shown in Fig. 15.9(b). A third possible mechanism consists of a dopant directly diffusing into a lattice vacancy (Fig. 15.9 (c)). We then talk about substitutional diffusion. It is only in the cases that an impurity occupies a vacated lattice site that *n*-type or *p*-type doping can occur. The presence of such vacancies in the lattice can be due to defects or to heat which increases atomic vibrations (Chapter 5), thus giving enough energy to the silicon atoms to move out of their equilibrium positions into interstitial sites.

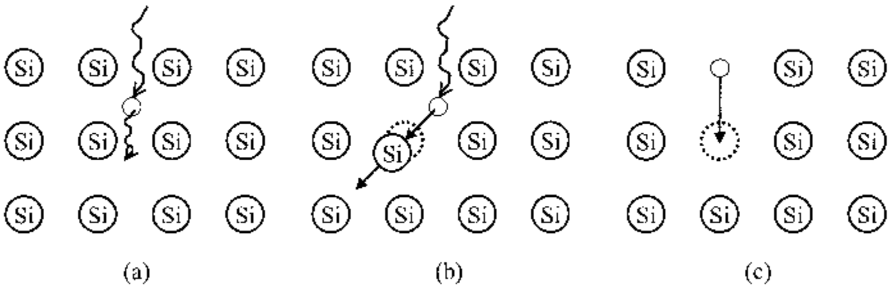


Fig. 15.9. Three possible diffusion mechanisms in a silicon wafer: (a) an impurity moves from one interstitial site to another, (b) a silicon atom is knocked into an interstitial site, thus leaving a vacancy which can be occupied by a diffusing impurity, (c) an impurity diffuses directly into a vacancy.

There are many different types of impurities that can be used for diffusion, the most common being boron, phosphorus, arsenic, and antimony. Table 15.3 lists the reactions for the materials for the three different types of diffusion sources.

The rate at which the diffusion of impurities takes place depends on how fast they are moving through the lattice. This phenomenon is quantitatively characterized by the diffusion coefficient of the impurity in silicon. Table 15.4 lists diffusion coefficient values for common impurities in silicon. We can then model the diffusion process by combining Fick's first and second law of diffusion:

$$\partial N / \partial t = D \partial^2 N / \partial x^2$$

Impurity	Type	Reaction
Boron	Solid	$2(CH_3O_3)B + 9O_2 \rightarrow B_2O_3 + 6CO_2 + 9H_2O$
	Liquid	$4BBr_3 + 3O_2 \rightarrow 2B_2O_3 + 6Br_2$
	Gas	$2B_2O_3 + 3Si \rightarrow 4B + 3SiO_2$
Phosphorus	Solid	$2P_2O_5 + 5Si \rightarrow 4P + 5SiO_2$
	Liquid	$4POCl_3 + 3O_2 \rightarrow 2P_2O_5 + 6Cl_2$
	Gas	$2PH_3 + 4O_2 \rightarrow P_2O_5 + 3H_2O$
Arsenic	Solid	$2As_2O_3 + 3Si \rightarrow 3SiO_2 + 4As$
Antimony	Solid	$2Sb_2O_3 + 3Si \rightarrow 3SiO_2 + 4Sb$

Table 15.3. Diffusion reactions for common impurity types.

Element	D_0 (cm^2.s^{-1})	E_A (eV)
B	10.50	3.69
Al	8.00	3.47
Ga	3.60	3.51
In	16.5	3.9
P	10.5	3.69
As	0.32	3.56
Sb	5.6	3.95

Table 15.4. Diffusion coefficient and activation energy values for common impurities in silicon.

The technology of diffusion in semiconductor processing consists of introducing a controlled amount of chosen impurities into selected regions of the semiconductor crystal. To prevent the diffusion of dopants in undesired areas, it is common to use a dielectric mask such as SiO_2 to selectively block the diffusion as show in Fig. 15.10. Fig. 15.11 shows a plot of the minimum mask thickness needed for a given diffusion time for boron and phosphorus diffusion.

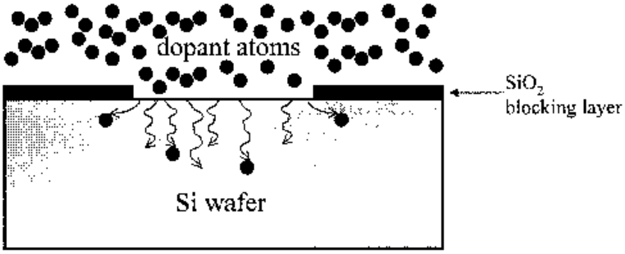


Fig. 15.10. Schematic illustration of the selective diffusion in a silicon wafer. The SiO_2 layer acts as a blocking layer for the diffusion of dopant atoms. Some dopant atoms can diffuse laterally under the blocking layer to some extent.

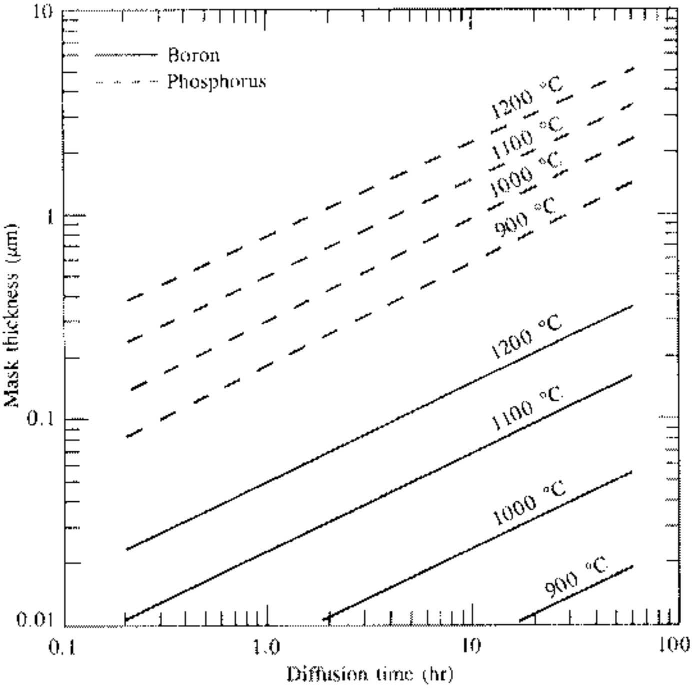


Fig. 15.11. Minimum SiO_2 mask thickness needed for successful diffusion of boron and phosphorus in silicon for a given temperature and time. [JAEGER, RICHARD C., INTRODUCTION TO MICROELECTRONICS FABRICATION: VOLUME 5 OF MODULAR SERIES ON SOLID STATE DEVICES, 2nd Edition, © 2002, p.53. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

There are two major techniques for conducting diffusion, depending on the state of the dopant on the surface of the wafer: (1) constant-source diffusion, also called predeposition or thermal predeposition, in which the concentration of the desired impurity at the surface of the semiconductor is kept constant; and (2) limited-source diffusion, or drive-in, in which a fixed total quantity of impurity is diffused and redistributed into the semiconductor to obtain the final profile.

15.3.2. Constant-source diffusion: predeposition

During predeposition, the silicon wafer is heated to a specific temperature, and an excess of the desired dopant is maintained above the wafer. The dopants diffuse into the crystal until their concentration near the surface is in equilibrium with the concentration in the surrounding ambient above it. At a given temperature, the maximum concentration that can be diffused into a solid is called the solid solubility. Having more dopants available outside the solid than can enter the solid guarantees that solid solubility will be maintained during the predeposition. For example, the solid solubility of phosphorus in silicon at 1000 °C is 9×10^{20} atoms/cm^3, while for boron in silicon at the same temperature it is only 2×10^{20} atoms/cm^3. These values only depend on the temperature, for a given dopant in a given semiconductor. As a result, the substrate temperature also determines the concentration of the dopant at the surface of the crystal wafer during diffusion.

Under predeposition conditions, let us denote N_0 the dopant concentration in the wafer near the surface. N_0 would be equal to the solid solubility of the dopant at the predeposition temperature if the excess dopant in the ambient above the wafer is sufficient. The concentration of dopant in the crystal at a depth x below the surface and after a diffusion time t can be known and is equal to:

Eq. (15.26) $$N(x,t)=N_0 erfc\left[\frac{x}{2\sqrt{Dt}}\right]$$

where D is the diffusion coefficient of the dopant at the predeposition temperature and *erfc* refers to the complementary error function. The complementary error function is found by complementing the integral of the normalized Gaussian function, and is shown in Fig. 15.12:

$$erfc(\overline{x})=1-erf(\overline{x})=\frac{2}{\sqrt{\pi}}\int_{\overline{x}}^{\infty}e^{-t^2}dt$$

The shape of the dopant concentration function is shown in Fig. 15.13 for several values of the product Dt. We see that, as the diffusion coefficient increases or, equivalently, as the diffusion time increases, the dopant reaches deeper into the crystal. The surface concentration remains the same at N_0.

The concentration N_B represents the background carrier concentration and refers to the concentration of majority carriers in the semiconductor before diffusion. The value of x for which $N(x,t)$ is equal to N_B is conventionally termed the junction depth.

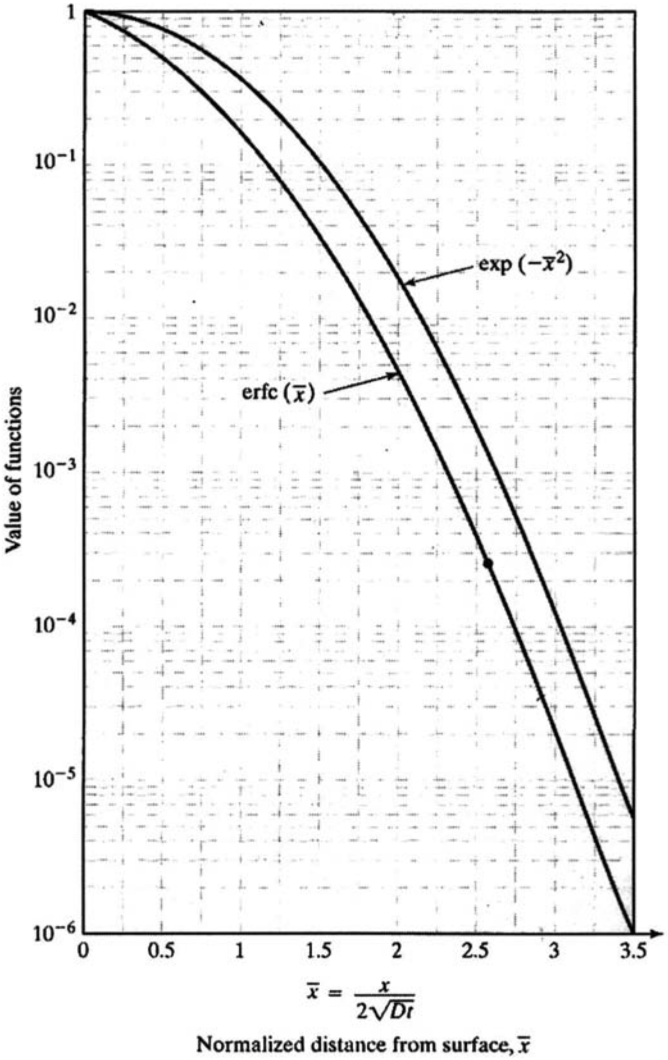


Fig. 15.12. Complementary error function, used in the calculation of the dopant concentration. [JAEGER, RICHARD C., INTRODUCTION TO MICROELECTRONICS FABRICATION: VOLUME 5 OF MODULAR SERIES ON SOLID STATE DEVICES, 2nd Edition, © 2002, p. 71. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

The total amount of impurities Q introduced per unit area, also called the dose, after a diffusion duration t in a predeposition process is found by integrating the function in Eq. (15.26) for values of x>0, which leads to the following expression:

Eq. (15.27) $$Q(t) = N_0\sqrt{\frac{4Dt}{\pi}}$$

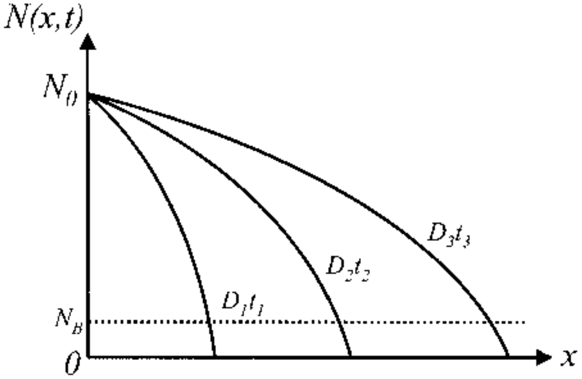


Fig. 15.13. Graph of the complementary error function, representing the dopant concentration in the crystal during predeposition, where the surface concentration is kept constant, for several values of Dt: D_3t_3>D_2t_2>D_1t_1. As the diffusion coefficient and/or the diffusion time is increased, the dopant reaches deeper into the crystal.

15.3.3. Limited-source diffusion: drive-in

Unlike predeposition, the drive-in diffusion process is carried out with a fixed total amount of impurity. This method allows us to better control the resulting doping profile and depth, which are important parameters in the fabrication of semiconductor devices.

During drive-in, the parameters which can be controlled include the duration of diffusion, the temperature, and the ambient gases. The dopant concentration profile of the drive-in has the shape of a Gaussian function, as shown in Fig. 15.14. In this type of diffusion, the dose remains constant causing the surface concentration to decrease. This relationship explains the shape of the curve, which can be expressed by solving Eq. (15.28) and using the boundary condition that the impurity concentration at the surface is equal to the dose:

Eq. (15.28) $$N(x,t) = \frac{Q}{\sqrt{\pi Dt}}\exp\left[\frac{-x^2}{4Dt}\right]$$

which is expressed in units of atoms per unit volume. D is the diffusion coefficient of the impurity at the drive-in temperature and t is the drive-in time. The drive-in can be performed after a predeposition step, in a high temperature diffusion furnace once the excess dopant remaining on the surface of the wafer from the predeposition step has been removed. In this case, Q is the total dose introduced into the crystal during a predeposition step.

The limited-source diffusion process is ideally suited when a relatively low value of surface concentration is needed in conjunction with a high diffusion depth. Typically, a short period of constant-source diffusion is followed by a period of limited-source diffusion. Predeposition is used to establish the dose into a shallow layer of the surface creating a diffusion front. Then the drive-in step moves this diffusion front to the desired depth.

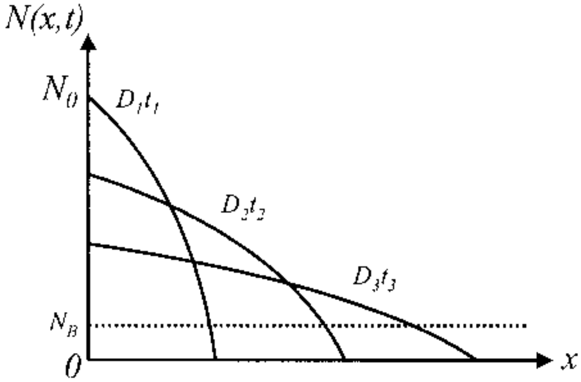


Fig. 15.14. Dopant concentration in the crystal during drive-in for several values of Dt: $D_3t_3>D_2t_2>D_1t_1$. As the diffusion coefficient and/or the diffusion time is increased, the dopant reaches deeper into the crystal. At the same time, the concentration at the surface is reduced because the drive-in is a limited-source diffusion process.

15.3.4. Junction formation

When diffusing p-type impurity dopants in an originally n-type doped semiconductor, a p-n junction can be formed, as shown in Fig. 15.15. In fact, the purpose of most diffusion processes is to form a p-n junction by changing a region of an n-type semiconductor into a p-type or vice versa.

Let us consider the example of an n-type doped silicon wafer which exhibits a background concentration N_B, and a p-type diffusing impurity with surface concentration N_0. Where the diffusing impurity profile concentration intersects the background concentration N_B, a metallurgical junction depth, x_j, is formed as shown in Fig. 15.15.

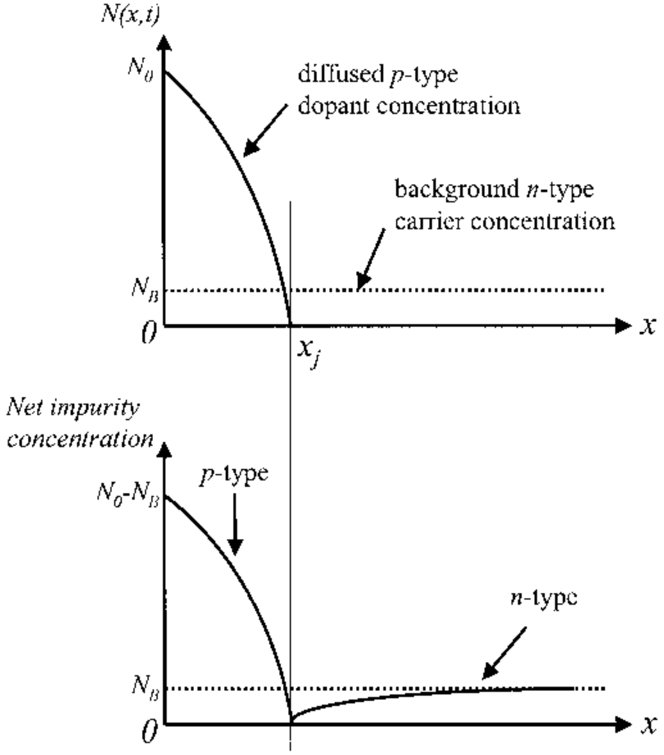

N(x,t)
N_0
diffused p-type dopant concentration
background n-type carrier concentration
N_B
0
x
x_j
Net impurity concentration
N_0-N_B
p-type
n-type
N_B
0
x

Fig. 15.15. Illustration of the formation of a p-n junction through diffusion. A p-type dopant is diffused into an n-type semiconductor which has a background concentration of N_B. The p-type dopant concentration profile after diffusion is shown in the top graph. The p-n junction will occur where the p-type dopant concentration is equal to the n-type background concentration as shown on the bottom graph.

At the metallurgical junction depth, the background concentration is equal to the surface concentration, so the net impurity concentration is zero. In the predeposition process with a complementary diffusion profile, the junction depth is found by solving Eq. (15.26) and using the boundary condition that $N(x_j,t)=N_B$:

Eq. (15.29) $$x_j = \left(2\sqrt{Dt}\right)erfc^{-1}\left(\frac{N_B}{N_0}\right)$$

where $erfc^{-1}$ refers to the reciprocal function of the complementary error function. In the drive-in process with a Gaussian diffusion profile, the junction depth is found by solving Eq. (15.28) and using the boundary condition that $N(x_j,t)=N_B$:

Eq. (15.30) $$x_j = 2\sqrt{Dt\ln\left(\frac{N_0}{N_B}\right)}$$

By successively diffusing two impurities of different types into an originally doped wafer, more complex structures can be achieved, such as for example an n-p-n transistor structure as illustrated in Fig. 15.16. The starting wafer would be an *n*-type (with a background concentration N_C in this example), the first diffusion process would introduce *p*-type dopants (N_B in this example) and the second diffusion would introduce *n*-type impurities (N_E) such that $N_E >> N_B >> N_C$.

Example

Q: Calculate the dose for a boron diffusion process at 1000 °C for 30 minutes using an *n*-type silicon substrate with a concentration of 10^{19} cm^{-3}.

A: T=1000 °C=1273 K
t=30 min=1800 s
From Table 15.4, boron has diffusion coefficient value D_0=10.5 cm^2.s^{-1} and activation energy E_A=3.69 eV. Using Eq. (15.23), the diffusion coefficient becomes

$$D = 10.5\exp[-3.69/(8.617\times10^{-5}eV)(1273K)]$$
$$= 2.5822\times10^{-14}\,cm^2.s^{-1}$$

From Eq. (15.27),

$$Q = (10^{19})\sqrt{(4\times2.5822\times10^{-14}\times1800)/\pi}$$
$$= 7.6928\times10^{13}\,cm^{-2}$$

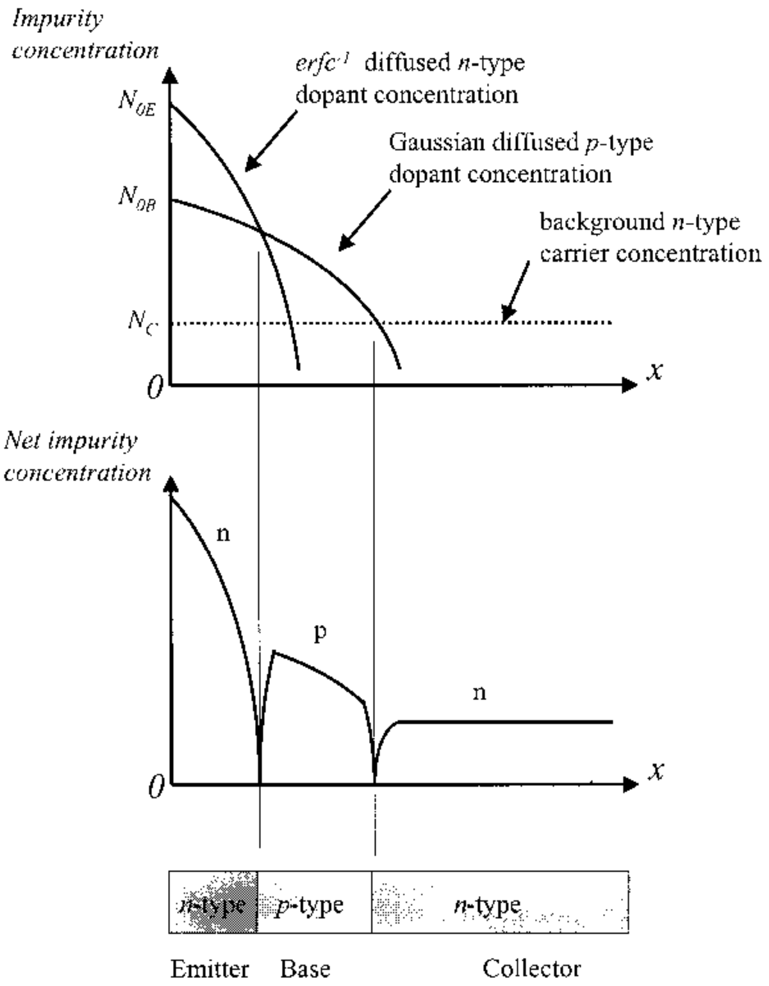

Impurity concentration
erfc^{-1} diffused n-type dopant concentration
N_{0E}
Gaussian diffused p-type dopant concentration
N_{0B}
background n-type carrier concentration
N_C
0
x
Net impurity concentration
n
p
n
0
x
n-type
p-type
n-type
Emitter
Base
Collector

Fig. 15.16. Illustration of the formation of an n-p-n transistor through diffusion. P-type dopants are first diffused into an n-type semiconductor to form the first junction. N-type dopants are subsequently diffused to form the second junction.

15.4. Ion implantation of dopants

Another technique to introduce dopants into a semiconductor wafer is through ion implantation. This technique is actually a direct alternative to the thermal predeposition described previously, and can be followed by a drive-in diffusion step.

The ion implantation process selects ions of a desired dopant, accelerates them using an electric field to form a beam of ions, and scans them across a wafer to obtain a uniform predeposition dopant profile inside the crystal. The energy imparted to the dopant ion determines the ion implantation depth. Using this technique, a controlled dose of dopant impurities can be introduced deep inside the semiconductor. This is in contrast to diffusion, where the dose of dopant is introduced only at the wafer surface. In addition, like diffusion, it is possible to conduct the ion implantation in only certain well-defined areas of the wafer by using an appropriate mask. This method yields reproducible and controlled dopant concentration for semiconductor devices.

We can choose to perform selective implantation, in which regions are selectively implanted with accelerated ions by using a patterned layer of material such as silicon dioxide or photoresist, as shown in Fig. 15.17.

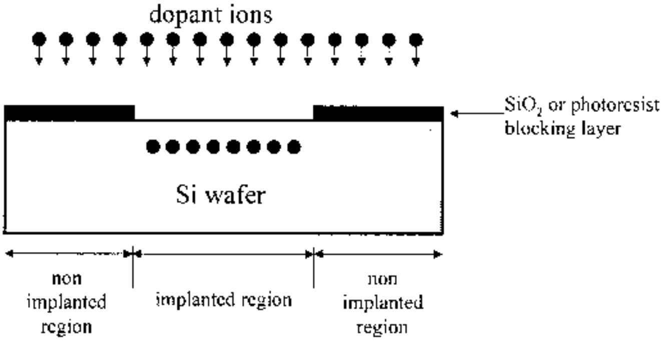


Fig. 15.17. Method of masking during ion implantation. The SiO_2 or photoresist layer acts as a blocking layer for the implantation of dopant atoms. In this process, no dopant atom can be found under the blocking layer if it is thick enough.

15.4.1. Ion generation

The first requirement of an ion implantation system is to generate ions of the desired species, accelerate them and direct them onto the wafer. A schematic of a typical ion implantation system is shown in Fig. 15.18. The dopant often comes in a gaseous form, and their ions are generated by heating them with a hot filament. These ions are then accelerated through an electric field. A magnetic field then curves the beams of ions and separates the ions, according to their atomic masses and charges, through a preset angle and output aperture. The selected ions are then further accelerated using an electric field. The beam is collimated and focused before striking the target wafer and penetrating the crystal lattice. An *x*-*y* rastering mechanism ensures that a large area of the sample is scanned by the ion implantation beam.

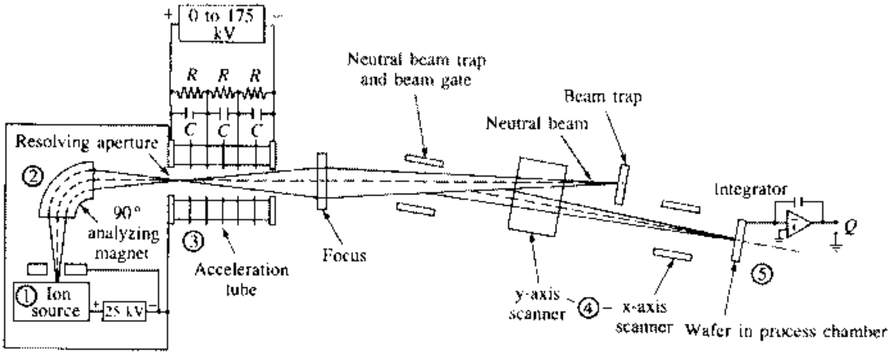


Fig. 15.18. Schematic of a typical ion implantation equipment, including an ion source (1), a primary acceleration and an analyzing magnet (2), an acceleration tube (3), a collimating and rastering magnets (4), and the sample to be implanted. [JAEGER, RICHARD C., INTRODUCTION TO MICROELECTRONICS FABRICATION: VOLUME 5 OF MODULAR SERIES ON SOLID STATE DEVICES, 2nd Edition, © 2002. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

15.4.2. Parameters of ion implantation

There are four major parameters to be controlled during ion implantation: the *energy* of the ions that reach the wafer, the dose Q of the dopant (the total number of ions that reach the wafer per unit of area), and the *depth* and *width* of the resulting implanted dopant profile.

The energy of the ions is directly controlled by the voltage used to accelerate them. It is easily understood that more energetic ions will penetrate deeper into the crystal, and potentially cause more physical damage than less energetic ones.

Because the selected ions all carry the same electrical charge, by measuring the electrical current carried by the ion beam (amount of electrical charge flowing per unit time), we can directly determine the dose. Mathematically, the latter is related to the ion beam current I through:

Eq. (15.31) $$Q = \frac{I}{qA} t$$

where q is the elementary charge, A is the implanted area and t is the duration of the ion implantation. For example, a 100 μA beam current of single ionized ions swept across a 200 cm^2 area for 60 seconds yields a dose equal to:

$$Q = \frac{\left(100\times10^{-6}\right)\times(60)}{\left(1.6\times10^{-19}\right)\times(200)} = 1.875\times10^{14}\ \text{dopants per cm}^2$$

By controlling the beam current and the implantation time, values of Q between 5×10^{9} and 5×10^{15} cm^{-2} can typically be achieved. This range of available doses is wider than that obtainable with thermal predeposition. It is therefore possible to reach doping profiles unobtainable by any other means. If the dopants were distributed uniformly over a depth of 50 nm, the dopant concentration could be controlled between values of 10^{15} and 10^{21} dopants per cm^3.

The depth and width of the resulting implanted doping profile can be represented by the projected range and straggle, as will be discussed in the next sub-section.

15.4.3. Ion range distribution

The dopant concentration profile after implantation follows a Gaussian distribution as illustrated in Fig. 15.19. As seen in the figure, the peak concentration N_p is found at a certain depth called the projected range R_p. The projected range measures the average penetration depth of the ions.

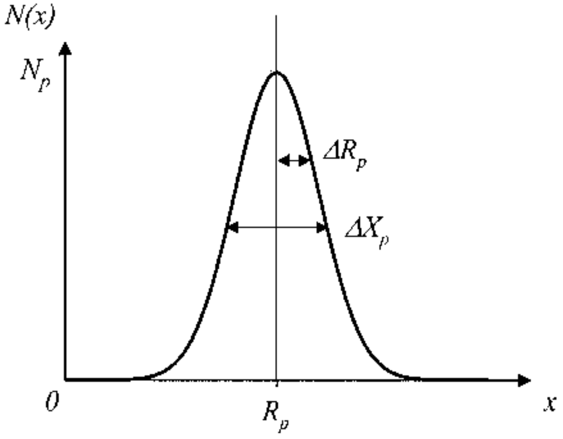


Fig. 15.19. Gaussian distribution for the concentration profile of implanted ions. The distribution is determined by its projected range, denoted R_p, corresponding to the peak concentration, and its straggle denoted by ΔR_p.

The depth at which the ions are implanted is mainly determined by the energy and the atomic number of the ions, as well as the atomic number of the substrate material. This can be easily understood because, as an impinging ion penetrates the semiconductor, it undergoes collisions with atoms and electrical repulsion with the surrounding electrons. The distance traveled between collisions and the amount of energy lost per collision are determined by a random process. Hence, even though all the ions are of the same type and have the same incident energy, they do not necessarily yield the same implantation depth. Instead, there is a distribution of depths represented by a standard deviation, called the straggle ΔR_p. Using Fig. 15.20, the impurity concentration at a given depth x can be found if the acceleration energy E_A is known:

Eq. (15.32) $$N(x) = N_p \exp\left[-\frac{(x-R_p)^2}{2\Delta R_p^2}\right]$$

For a Gaussian distribution shown in Fig. 15.19, the full width at half-maximum, denoted ΔX_p, is given by:

Eq. (15.33) $$\Delta X_p = \left(2\sqrt{2\ln 2}\right)\Delta R_p = 2.35\Delta R_p$$

The implanted dose can be determined by integrating Eq. (15.32) over all the possible depths inside the crystal:

Eq. (15.34) $$Q = \int_0^\infty N(x)dx = \sqrt{2\pi}N_p \Delta R_p$$

Example

Q: Find the full width at half-maximum for the ion-implantation using boron with an acceleration energy of 100 keV.

A: From Fig. 15.20, we obtain the normal straggle ΔR_p=0.07 μm. Using Eq. (15.33),

$$\Delta X_p = (2\sqrt{2\ln 2})(0.07) = 0.1648\,\mu m$$

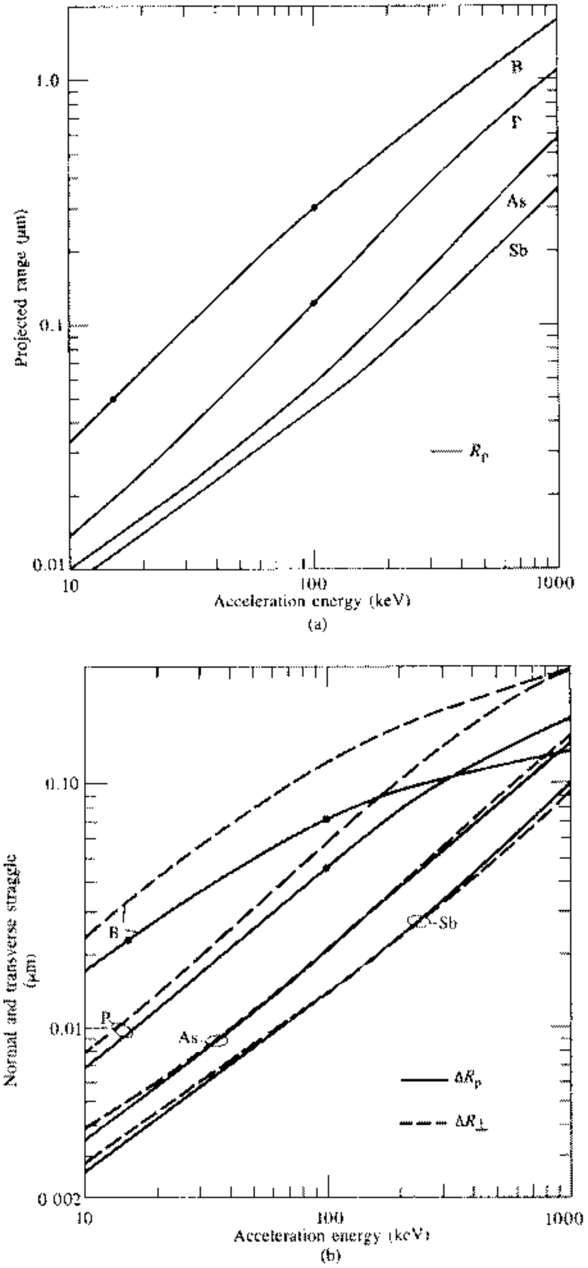


Fig. 15.20. (a) Projected range and (b) normal and transverse straggle for the ion-implantation of boron, phosphorus, arsenic, and antimony impurities for a given acceleration energy. [JAEGER, RICHARD C., INTRODUCTION TO MICROELECTRONICS FABRICATION: VOLUME 5 OF MODULAR SERIES ON SOLID STATE DEVICES, 2nd Edition, © 2002, p. 113. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

15.5. Characterization of diffused and implanted layers

Two parameters are of interest in assessing the properties of diffused or implanted layers and junctions: their electrical resistivity, junction depth, and impurity concentration. A number of techniques are available for evaluating each of these parameters.

15.5.1. Sheet resistivity

In sub-section 8.2.1, we introduced the resistivity of a crystal as being a local property which is related to the concentration and mobility of the majority carriers. In diffused or implanted layers, we are not interested in the values of the resistivity, because the carrier concentration is not uniform in space as shown in the profiles in Fig. 15.13 and Fig. 15.14. Rather, we are interested in the sheet resistivity, a quantity which can be directly measured. In order to visualize the physical meaning of this parameter, let us consider a parallelepiped semiconductor bar with a length L, a width W and a thickness H as shown in Fig. 15.21.

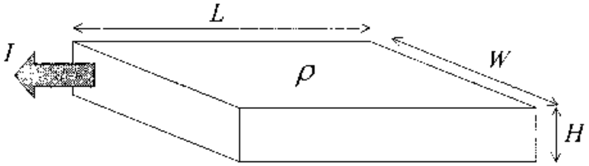


Fig. 15.21. Geometry used in determining the resistance of a block of material having uniform resistivity. When the current is flowing in the direction shown, the resistance in this direction is proportional to the length L and inversely proportional to the cross-section area WH.

We know from Eq. (8.11) that the resistance of this block for a current flowing in the direction of the shown arrow is given by:

Eq. (15.35) $$R = \frac{L}{WH}\rho = \frac{L}{W}R_s$$

where ρ is the resistivity of the material, assumed to be uniform in this case. The sheet resistivity is a quantity which does not take into account the thickness of the layer, and is defined as the resistivity divided by the thickness:

Eq. (15.36) $R_s = \dfrac{\rho}{H}$

and expressed in units of "ohm per square" or Ω/ . In practice, because the thickness of the conducting layer is not always known during experiments, the sheet resistivity is the quantity that is actually measured, and the bulk resistivity is calculated subsequently.

One of the measurement methods for the sheet resistivity is the linear four-point probe method as shown in Fig. 15.22. It consists of placing four equally spaced probes on the surface of the wafer in a linear manner. The probe spacing s is typically on the order of either 1000 or 1250 μm. By sending a fixed current I through the two outer probes and measuring the voltage V across the two inner probes, we can determine the resistivity, in units of Ω.m, given by the following expression:

Eq. (15.37) $\rho = 2\pi s \dfrac{V}{I} = \left[\dfrac{\pi t}{\ln 2}\right]\dfrac{V}{I}$

Sheet resistivity can then be determined from Eq. (15.37) as:

Eq. (15.38) $R_x = \dfrac{\rho}{t} = \left[\dfrac{\pi}{\ln 2}\right]\dfrac{V}{I}$

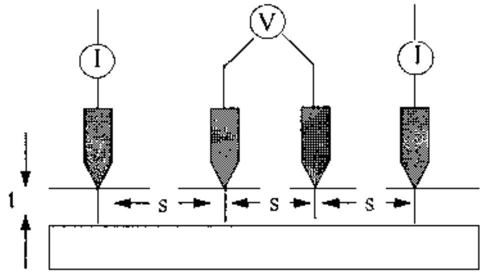


Fig. 15.22. An in-line four-point probe. Four equally spaced probes are placed on the surface of the wafer in a linear manner. A fixed current is sent through two of the outer probes and the voltage measured between the two inner probes gives a value for the sheet resistivity of the material.

Another method to measure the sheet resistivity of doped layers is van der Pauw method which can be used for any arbitrarily shaped sample of material by placing four contacts on its periphery as shown in Fig. 15.23(a). Square shaped test areas with contact regions at the four corners are usually preferred and are prepared by lithographic techniques as shown in Fig. 15.23(b).

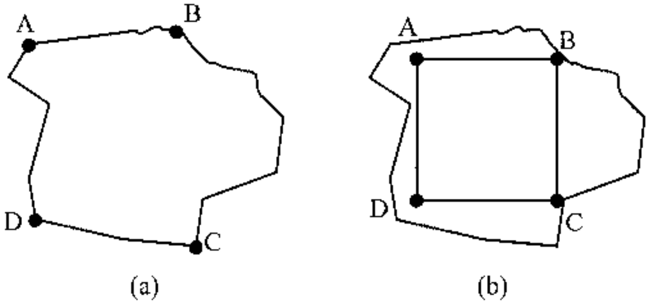


Fig. 15.23. A simple van der Pauw test structure. (a) Four contacts are placed at the periphery of any arbitrarily shaped sample and can be used to determine the resistivity of the material. (b) It is generally preferred to use a square shaped sample, which can be obtained by etching away undesired areas off the original sample.

In this configuration, a current is injected through one pair of the contacts and the voltage is measured across the remaining pair of contacts. Repeating these measurements for another pair, we are then able to define two resistances such as for example:

Eq. (15.39) $$R_{AB,CD} = \frac{V_{CD}}{I_{AB}} \text{ and } R_{BC,DA} = \frac{V_{DA}}{I_{BC}}$$

These resistances are related by the following equation relation:

Eq. (15.40) $$\exp\left[\frac{-\pi R_{AB,CD}}{R_S}\right] + \exp\left[\frac{-\pi R_{BC,DA}}{R_S}\right] = 1$$

where R_s is the sheet resistivity of the semiconductor layer. This expression allows us to implicitly determine the sheet resistivity of the sample, and thus the bulk resistivity if the layer thickness is known. For a symmetrical measurement geometry, the two resistances in Eq. (15.39) are equal and Eq. (15.40) yields a simple expression:

Eq. (15.41) $$R_s = \frac{\pi}{\ln 2} \frac{V_{CD}}{I_{AB}}$$

15.5.2. Junction depth

The junction depth x_j is defined as the distance from the top surface within the diffused or implanted layer at which the dopant concentration equals the background concentration. There exist two methods to measure the junction depth: the groove and stain method and the angle-lap method.

In the groove and stain method, a cylindrical groove is mechanically ground into the surface of the wafer as shown in the cross-section schematic in Fig. 15.24. A chemical stain creates a color contrast between the differently doped layers, thus revealing the junction.

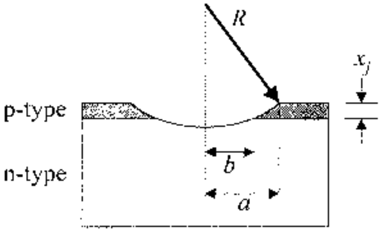


Fig. 15.24. Cross-section illustration of the junction depth measurement by the groove and stain technique. A cylindrical groove is mechanically ground into the surface of the wafer. A chemical stain creates a color contrast between the differently doped layers, thus revealing the junction.

Through purely geometrical arguments, the junction depth can be found to be equal to:

Eq. (15.42) $x_j = \sqrt{\left(R^2 - b^2\right)} - \sqrt{\left(R^2 - a^2\right)}$

For $R >> a$ and b, this expression can be simplified into:

Eq. (15.43) $x_j \approx \frac{\left(a^2 - b^2\right)}{2R}$

In the angle-lap method, a piece of the wafer is mounted on a special fixture which permits the edge of the wafer to be lapped at an angle between 1 and 5° as shown in Fig. 15.25. The sample is then stained with, for example, a 100 ml hydrofluoric acid/nitric acid solution. Once stained, the sample is observed under a collimated monochromatic light at normal incidence.

An interference pattern can be observed through a cover glass, and the junction depth may be calculated by counting the interference fringes and then applying the equation:

Eq. (15.44) $x_j \approx d\tan\theta = \frac{N\lambda}{2}$

where θ is the angle of the etched lap, λ is the wavelength of the monochromatic light and N is the number of fringes.

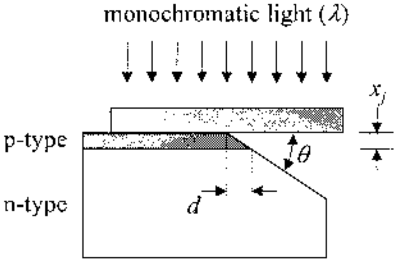


Fig. 15.25. Junction depth measured by the angle-lap method. A piece of the wafer is mounted on a special fixture which permits the edge of the wafer to be lapped at an angle. The sample is then stained and then observed under a collimated monochromatic light at normal incidence. An interference pattern can be observed through a cover glass and the junction depth can be calculated by counting the number of interference fringes.

15.5.3. Impurity concentration

There are many ways to measure the impurity concentration of a sample, but one of the most common techniques is Secondary Ion Mass Spectroscopy (SIMS). A schematic representation of the experimental setup is shown in Fig. 15.26. SIMS is a destructive characterization technique that operates with a highly energetic beam of ions hitting the sample, causing the sputtering or ejection of atoms from the sample material. Some of these ejected atoms are charged ions, or secondary ions. A mass spectrometer then separates and collects the secondary ions. The number of collected secondary ions then allows a detector to determine the material composition. An example plot of the impurity concentration information obtained from SIMS is shown in Fig. 15.27.

SIMS is an excellent technique for identifying all types of elements, unlike other measurement resources. One disadvantage to the SIMS measurement is the sensitivity of the technique. The sensitivity can be affected by the built-up charge from the sputtering process, type of ion beam. Another disadvantage to SIMS is the limitation of the beam area that hits the sample.

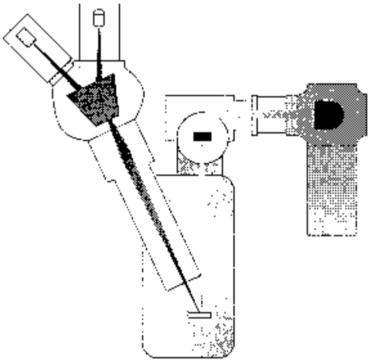


Fig. 15.26. Secondary Ion Mass Spectrometer. Low energy ion-beam hits the sample surface and atoms sputter off of it. The atoms are then ionized and sent through a mass spectrometer, where secondary ions are collected. The mass spectrometer identifies the atomic species and then the detector uses these secondary ions to determine the profile as a function of depth.

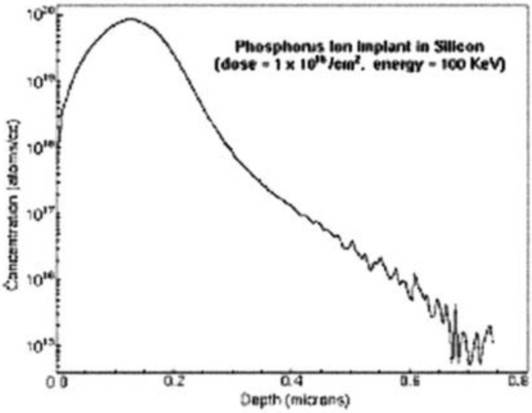


Fig. 15.27. Secondary Ion Mass Spectroscopy plot measuring the impurity concentration as a function of depth for ion implantation of phosphorus into silicon. [From http://www.me.ust.hk/~mejswu/MECH343/343SIMS.pdf.]

15.6. Summary

In this Chapter, we have reviewed a few of the steps involved in the fabrication of semiconductor devices, including oxidation, diffusion and ion implantation. Although the discussion was primarily based on silicon, the concepts introduced are applicable for the entire semiconductor industry.

We described the oxidation experimental process, mathematically modeled the formation of a silicon oxide film, discussed the factors influencing the oxidation and reviewed the methods used to characterize the oxide film. The diffusion and ion implantation of impurity dopants in silicon to achieve controlled doping in selected areas of a wafer was described, along with the resulting dopant concentration profiles inside the semiconductor. The predeposition and drive-in conditions of diffusion were discussed. Methods used to assess the electrical properties of the diffused or implanted layers were reviewed.

References

Jaeger, R.C., *Introduction to Microelectronic Fabrication*, 2^{nd} Edition, Prentice-Hall, Upper Saddle River, NJ, 2002.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, pp. 188-193, 1989.

Further reading

Campbell, S.A., *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, New York, 1996.

Diaz, J.E., *Fabrication of High Power Aluminum-Free 0.8 μm to 1.0 μm InGaP/InGaAsP/GaAs Lasers for Optical Pumping*, Ph.D. dissertation, Northwestern University, 1997.

Fogiel, M., *Microelectronics-Principle, Design Techniques, and Fabrication Processes*, Research and Education Association, New York, 1968.

Ghandi, S., *VLSI Fabrication Principles*, John Wiley & Sons, New York, 1983.

Soclof, S., *Design and Application of Analog Integrated Circuits*, Prentice-Hall, Englewood Cliffs, NJ, pp. 8-23, 1991.

Streetman, B.G., *Solid State Electronic Devices*, Prentice-Hall, Englewood Cliffs, NJ, pp. 65-70, 1980.

Trumbore, F.A., "Solid solubilities of impurity elements in germanium and silicon," *Bell System Technical Journal* 39, pp. 205-233, 1960.

Problems

1. A (100) Si wafer undergoes the following sequence of oxidation steps: one-hour dry oxidation at 1100 °C, two-hour wet oxidation at 1000 °C, and one-hour dry oxidation at 1100 °C. Calculate the thickness after each oxidation step.

2. Compare the thickness of silicon dioxide film grown on (100) Si for wet and dry oxidation at 1100 °C. Compare the two different orientations (100) and (111) Si using the same conditions for wet oxidation.

3. Calculate the time required to grow 2 μm silicon dioxide on (111) silicon wafer for wet oxidation at 1050 °C. How long would it take to grow an additional 5 μm?

4. A silicon oxide layer is grown for 65 minutes at 1100 °C on (111) silicon by passing oxygen through a 95 °C water bath. How thick an oxide layer is grown on the silicon surface?

5. A (100) oriented silicon wafer is already covered with a 0.5 μm thick oxide film. How long would it take to grow an additional 0.1 μm oxide using wet oxidation at 1373 K? Compare the result with the linear oxidation law using a rate of B/A=3 μm.hr^{-1}.

6. An npn transistor is formed by boron diffusion on an n-type silicon wafer with impurity concentration of 10^{20} cm^{-3} and doping concentration 10^{16} cm^{-3}. Constant source diffusion is performed for 30 minutes followed by limited-source diffusion for 2 hours, both at 1000 °C. Find the junction depth after each step.

7. What is the required thickness for a SiO_2 mask used for selective phosphorus diffusion. The diffusion was performed at 1000 °C for 3 hours.

8. An impurity is diffused into silicon in the constant-source diffusion case with a surface concentration N_0=10^{19} cm^{-3}. The diffusion coefficient is known to be equal to 2×10^{-12} cm^2.s^{-1} at 1100 °C with an activation energy of 4 eV. A diffusion length of 1 μm is aimed at. (a) After diffusion at 1000 °C, what is the total dose diffused in the layer? (b) How long must the diffusion last?

9. An impurity is diffused into silicon at 1100 °C during 20 minutes. A diffusion length of 1 μm is measured and the dose diffused in the sample is 2×10^{15} atoms.cm^{-2}. Determine the diffusion coefficient. Determine the surface concentration, assuming constant-source diffusion.

10. Find the dose in silicon for phosphorus that is implanted with an energy of 100 keV with a 0.1 μm SiO_2 layer and peak concentration of 10^{17} cm^{-3}. Find the time required for this implantation onto a 2" wafer using 2 μA of current.

11. Implantation using phosphorus is done such that the implantation peak is located at the Si-SiO_2 interface with an energy of 40 keV. The dose is 5.24×10^{14} cm^{-2}, and background concentration is 3×10^{16} cm^{-3}. Find the minimum oxide thickness required for a masking layer.

12. Compare the time required for implantation of phosphorus and boron with an energy of 75 keV, with a desired peak concentration of 10^{18} cm^{-3}, into a 140 mm silicon wafer with 1 μA.

13. A constant voltage of 5 V is applied to each of the five contact pads on a given sample. The location of the pads are at x_1=1, x_2=5, x_3=10, x_4=16, and x_5= 23 μm. The width is the same for each contact and is given as 1 μm. The thickness is given as 200 μm. Find the contact resistance using transmission line measurement if the current is measured as I_{12}=20 mA, I_{23}=17 mA, I_{34}=10 mA, and I_{45}=3 mA.

14. Calculate the resistivity using the van der Pauw method with measured current I_{AB}=1 mA and V_{CD}=2V. The length L=1 mm and width W=2 mm.

15. Assume that an As implant leads to a uniform electron concentration of 10^{19} cm^{-3} down to a depth of 0.1 μm and a mobility of 100 cm^2.V^{-1}.s^{-1}. Determine the resistivity and the sheet resistivity of the implanted layer. If a square van der Pauw pattern with 1 cm side length is used with a 10 V applied at two adjacent contacts, what current would be measured through the other two contacts?

16. Semiconductor Device Processing

16.1. Introduction

In Chapter 13, we reviewed a few important steps in the process of fabricating semiconductor devices, such as oxidation, diffusion and ion implantation of dopants. The resulting semiconductor wafer then undergoes a series of additional steps before the final device is obtained, which are shown in the flowchart in Fig. 16.1. The major ones include lithography, etching, metallization and packaging, each of which will be discussed in the present Chapter.

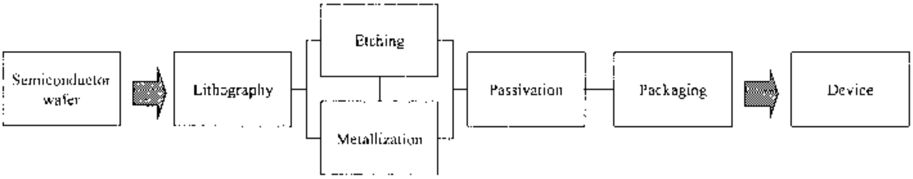


Fig. 16.1. Basic fabrication process flowchart, including the important steps: lithography, etching, metallization, passivation and packaging, which transform a semiconductor wafer into a device that can be used in electronic systems.

16.2. Photolithography

Lithography consists of preparing the surface of a semiconductor wafer in order to allow the subsequent transfer of a specific pattern. To do so, the surface of the semiconductor must be carefully prepared and a film called resist is conformally applied onto it. Parts of this resist film will be selectively "activated" through a number of processes, while others will be left untouched in order to transfer the desired pattern. This is generally achieved through what is called mask alignment and exposure. A mask is a template which contains the desired pattern to be transferred. Finally, the resist is developed to reveal the desired pattern before proceeding to the subsequent processing steps.

There are several types of lithography techniques depending on the method used to activate the resist film. The most common form of lithography uses ultraviolet light and is called photolithography. This is currently the most widely used technique in microelectronics industry and is routinely employed to achieve features as small as 0.18 μm. In this section, we will describe in detail the photolithography method.

16.2.1. Wafer preparation

In the previous Chapters, thin film epitaxy and mechanisms of wafer generation were discussed in great detail; as was the process of growing a barrier layer, i.e. an oxide layer. Once these difficult steps are completed the process of creating a pattern on the surface of the semiconductor wafer can now be started.

The most important step in the lithographic process is the minimization of defects caused by particles either falling on the surface prior to epitaxy or during the steps of lithography. This is because even the smallest dust particulate on a chip would destroy dozens if not hundreds of transistors on a state of the art microprocessor. This is the chief reason why semiconductor processing is performed in ultra clean facilities, know as a cleanrooms. A cleanroom uses sophisticated filtration to remove airborne particulates and are rated by the maximum number of particles per volume of air.

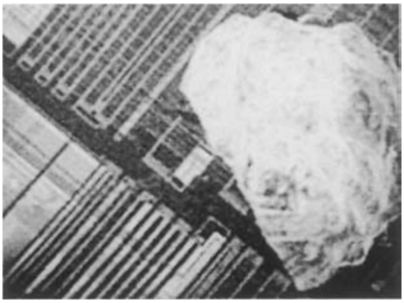


Fig. 16.2. This enlarged image of a grain of salt on a piece of a microprocessor should give you an idea of how small and complex a microprocessor really is. [From http://www.intel.com/education/images/manufacturing/salt.jpg. Reprinted with permission from Intel Corporation.]

Cleanrooms having dust is only half the problem, the other major cause of contamination are the workers themselves. The workers in the cleanroom must wear special uniforms to minimize the introduction of additional contaminants, such as hair, skin flakes, or worse outside world dirt. This protective clothing is made from a non-linting, anti-static fabric and is worn over street clothes. The final step in minimizing particulate based fabrication problems, wafers are chemically cleaned to remove any particles that may have adhered to the surface. This is to promote adhesion of the photoresist to the surface.

16.2.2. Positive and negative photoresists

In photolithography, the resist is called photoresist and it can be of either of the two types: positive or negative photoresist, depending on its chemistry which determines its property when exposed to light. The photoresist is a photosensitive material used to transfer the image from the mask to the wafer surface. The quality of the resist plays an important role in the image transfer to a semiconductor wafer. Resists are generally required to maintain a usable adhesion, uniformity, etch resistance, thermal stability, and long shelf life. Both positive and negative resists are made of complex organic molecules containing carbon, oxygen, nitrogen, and hydrogen. They can be more or less easily dissolved using a developing solution, depending on the amount of light they have been exposed to.

Let us illustrate the difference between positive and negative photoresists by considering the example of a silicon wafer with a thin silicon oxide layer and coated with a layer of photoresist, as shown in Fig. 16.3.

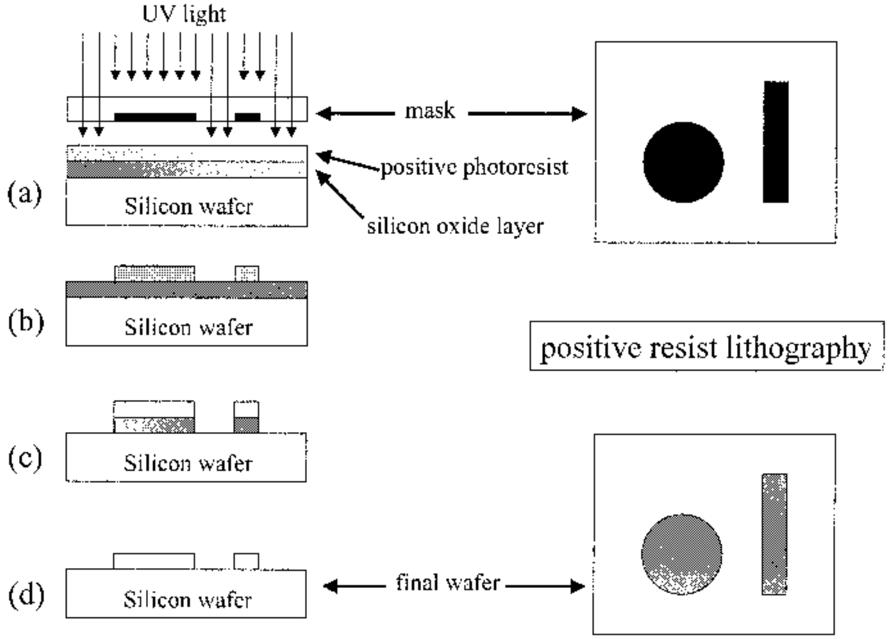


Fig. 16.3. Positive resist photolithography process sequence. When using positive resist, the exposed regions are dissolved in the developing solution, while the unexposed areas remain intact. (a) The positive photoresist is exposed using a source of intense ultraviolet light. (b) The wafer is removed from the alignment station and areas exposed are dissolved in a solution. In the steps illustrated in (c) through (d), the etch-resistant property of the resist is used in the etching of silicon dioxide in the regions which are not protected by the remaining photoresist.

In Fig. 16.3(a), the positive photoresist is exposed using a source of intense ultraviolet light such as a mercury arc lamp which alters its chemical bonding to make it more soluble where it has been exposed. The wafer is removed from the alignment station and developed (Fig. 16.3(b)). The exposed regions of the positive photoresist layer are dissolved in the developing solution, leaving the unexposed areas intact. In other words, for a positive photoresist, the light from the exposure step increases the solubility of the resist in the developing solution by depolymerizing the resist material. Following this image-transfer step to the photoresist, the image needs to be transferred to the underlying layers. In the steps illustrated in Fig. 16.3(c) through (d), the etch-resistant property of the resist is used in the etching of silicon dioxide in regions which are not protected by the remaining photoresist. If properly selected, the etchant will remove the layer of silicon dioxide but will not etch the underlying silicon or the layer of photoresist, as shown in the figure. The result of the photolithographic process is shown in Fig. 16.3(d) where after the layer of resist has been removed, only the patterned layer of silicon dioxide is left.

By contrast, when using a negative photoresist, it is the unexposed regions which are dissolved in the developing solution, leaving the exposed areas intact. In other words, in this case, the light from the exposure step causes polymerization to occur in the resist, reducing its solubility in the developing solution. This is illustrated in Fig. 16.4(a) and (b). The remaining sequence of steps is similar to the previous one and is illustrated in Fig. 16.4(c) through (d).

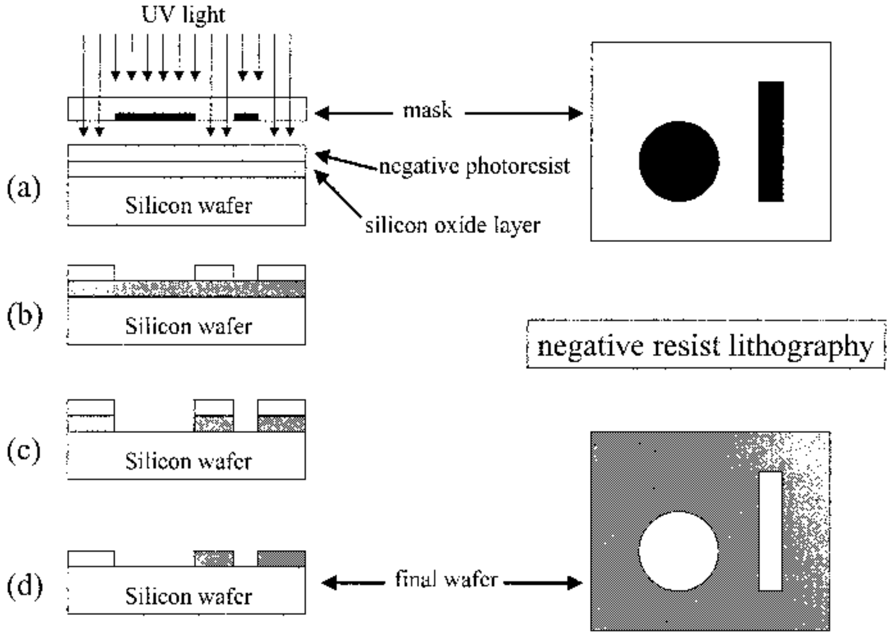


Fig. 16.4. Negative resist photolithography process sequence. When using negative photoresist, the unexposed regions are dissolved in the developing solution, while the exposed areas remain intact. The sequence of steps is similar to that of Fig. 16.3.

The choice of a positive or negative photoresist is determined by the subsequent sequence of processing steps which need to be performed.

For either type of photoresist, the following requirements must be met: it must adhere well to the wafer surface, its thickness must be uniform across the wafer and must be predictable from wafer to wafer, it must be sensitive to light so that it can be patterned, and it must not be attacked by the etchant for removing the substrate material. The photoresist film is first applied to the wafer surface in a yellow light cleanroom environment. In spinning the photoresist, a small puddle of resist is first dispensed onto the center of the substrate, which is attached to a spindle using a vacuum chuck as shown in Fig. 16.5.

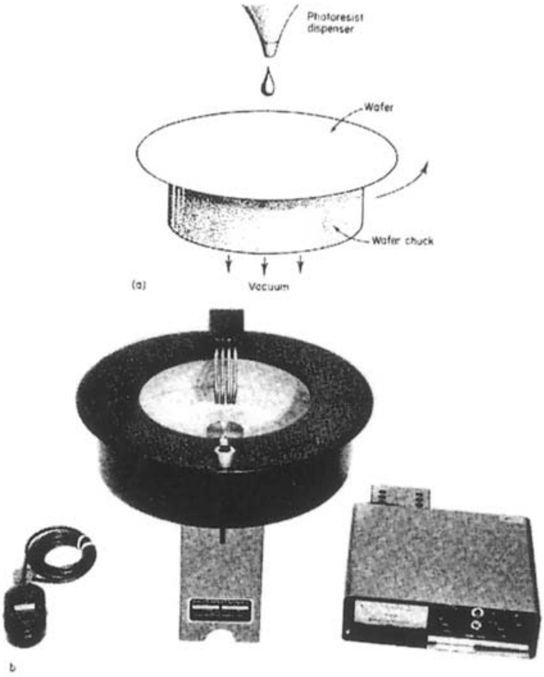


Fig. 16.5. An example of photoresist spinner system. The spinner and its control equipment are shown in (b). The top diagram (a) illustrates the spinning process where a wafer is firmly maintained onto a wafer chuck by pulling a vacuum between them. The wafer/chuck block is then spun and a drop of photoresist is dispensed at the center of the wafer from where it coats the wafer with a thin resist film. [Reprinted with permission from Headway Research, Inc.]

The spindle is then spun rapidly, rotating the substrate at several thousand revolutions per minute for a certain period of time. The formula commonly relating spin speed to final thickness is:

Eq. (16.1) $z = \frac{\kappa_r P}{\sqrt{\omega}}$

where z is the final resist thickness, P the percentage of solids in the resist, ω the angular velocity of the spinner, and κ_r is a constant that is different for each resist.

For example, for a given spinner and photoresist formulation, we know that a rotation speed of 4000 rpm gives a resist thickness of 0.7 μm. In order to increase the thickness to 0.8 μm, the proper rotation speed ω must satisfy the relation: $\sqrt{\frac{\omega}{4000rpm}} = \frac{0.7}{0.8}$, so that ω=3062 rpm.

Once the photoresist is applied a prebake, or a softbake, step is performed. Softbaking is used to remove the solvents present in the resist and to improve adhesion. Solvents left in the resist film from poor softbaking will cause a degraded image to be transferred to the wafer surface, because such solvents will be attacked by the developer and cause portions of the resist that are to remain to be removed. Properly baked wafers will have resist that has the proper amount of resins and photosensitizers (positive resist) or inhibitors (negative resist) as determined by the manufacturer. Once the prebake is completed the photoresist-covered wafer is then ready for mask alignment and exposure.

16.2.3. Mask alignment and fabrication

The word photolithography may be loosely defined as "printing with light", which is an accurate description of the heart of this processing step. The manufacture of semiconductor devices and integrated circuits consists of multiple passes through photolithography steps. Each time, it defines the region where the subsequent processing step, e.g. doping introduction, oxidation, metallization, will have its effect. These multiple passes must be aligned using simple marks to help either a computerized aligner or fabricator align the new mask with previous mask step. If masks are not aligned the whole wafer will be dead because the layers of the multiple processes will not be aligned and contacts will not work.

In photolithography, it is first necessary to produce a mask or transparency of the pattern required. Mask making begins with a large-scale layout or artwork which is then photographed by large camera to reduce it down typically more than a thousand times to the exact size on a master plate. Fig. 16.6 shows the sequential steps necessary to create an integrated mask. The master plate is used in a precision step-and-repeat printer to produce multiple sequential images of the layout on a high-resolution photosensitized emulsion glass plate which is later used as the mask in the photolithography process to transfer the layout pattern onto the wafer surface.

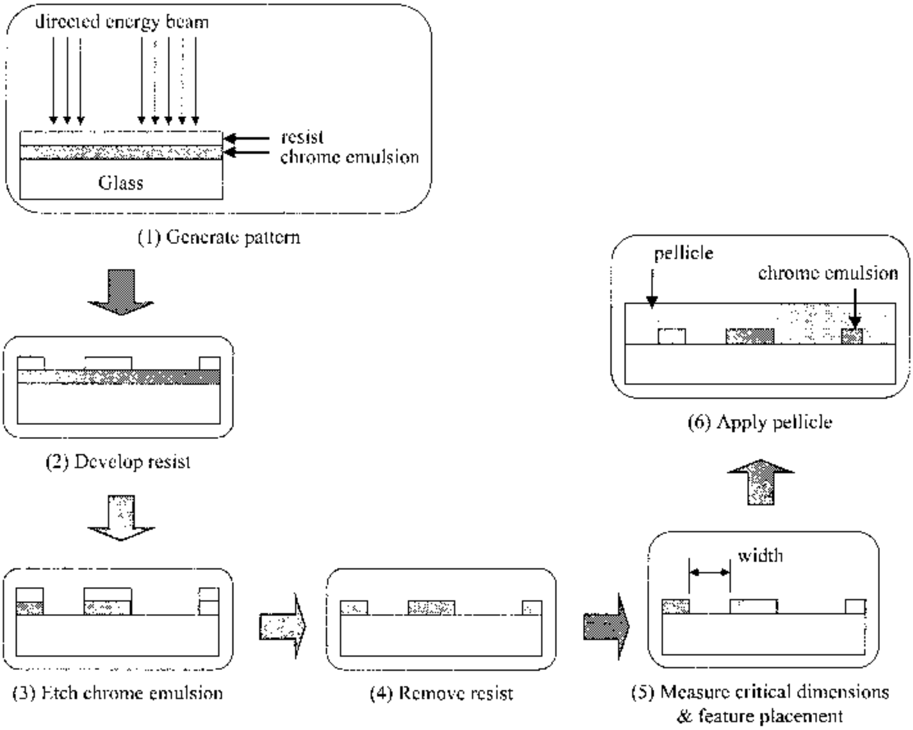


Fig. 16.6. Flow diagram illustrating the realization of a mask which would be later used for photolithography. It begins with a large-scale layout or artwork which is then photographed by large camera to reduce it down typically more than a thousand times to the exact size on a master plate. The master plate is used in a precision step-and-repeat printer to produce multiple sequential images of the layout on a high-resolution glass plate covered with an emulsion and a resist layer. A pellicle barrier layer is provided in order to ensure the integrity of the pattern from particulate.

The emulsion used on the glass plate is susceptible to scratches, wear and tear damage during usage. Alternative materials which withstand wear better than emulsion but are also considerably more expensive, such as chrome and iron oxide, are sometimes substituted for emulsion. Iron oxide masks have the additional advantage of being transparent to the yellow light used to visually align the masks, while being opaque to the intense ultraviolet light used for the exposure of the resist. Visually, each type of mask is a plate of glass with alternate clear and relatively opaque regions as shown in Fig. 16.7.

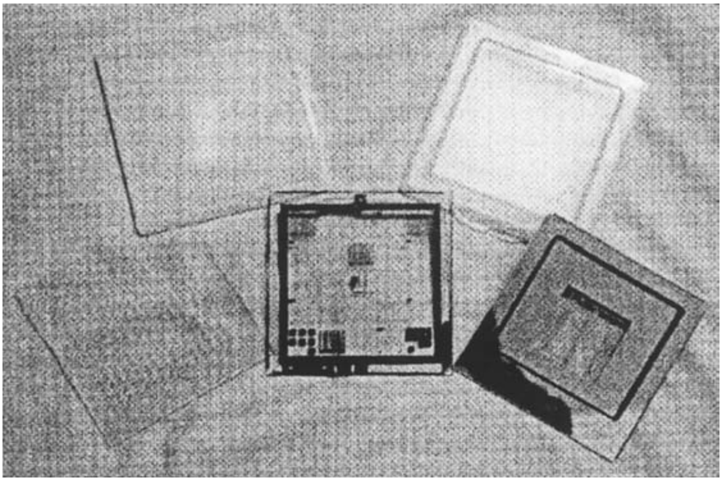


Fig. 16.7. Example of an integrated mask. Visually, a mask is a plate of glass with alternate clear and relatively opaque regions. ["Figure 7.3", from THE SCIENCE AND ENGINEERING OF MICROELECTRONIC FABRICATION, 2ND EDITION by Stephen A. Campbell, copyright © 2001 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.]

16.2.4. Exposure

As mentioned earlier, the exposure to ultraviolet light determines the property of the photoresist. It is important to know the sensitivity of the resist used and control the amount of light that it receives. In addition it is important to know how the light is passing the mask to hit the surface of the wafer and the photoresist. This affects not only the size of the image made on the wafer but also how long the mask will actually last; an important fact to know due to the length of time and expenses required to make the masks.

There are three main types of printing that are generally used in lithography: contact, proximity, and projection; a schematic diagram of these methods is shown in Fig. 16.8.

In contact printing (Fig. 16.8(a)), the mask is placed in direct contact with the photoresist. This limits diffraction and is the simplest of the techniques to use, but the mask gets worn, the photoresist can be more easily contaminated by residue of the mask, and the mask limits the size of the images produced. Contact printing has been largely replaced by proximity printing (Fig. 16.8(b)); here the mask is held slightly above the wafer. This increases the lifetime of the mask and reduces the potential for contamination but the air gap increases diffraction as well, the image is still determined by the mask dimension size. The third option for optical lithography, projection printing (Fig. 16.8(c)) is now the standard used in fabrication. This is because in this way one can make images smaller than that which appears on the mask. It is made possible because of the use of focus lenses that can shrink the image. This system also allows for larger wafers to have multiple projected images placed upon the wafer by scanning over the entire wafer. This is done using a method known as step-scan processing. While projecting printing allows for reduced contamination like proximity printing, it also able to make smaller images than the mask, and is only limited by the light's diffraction.

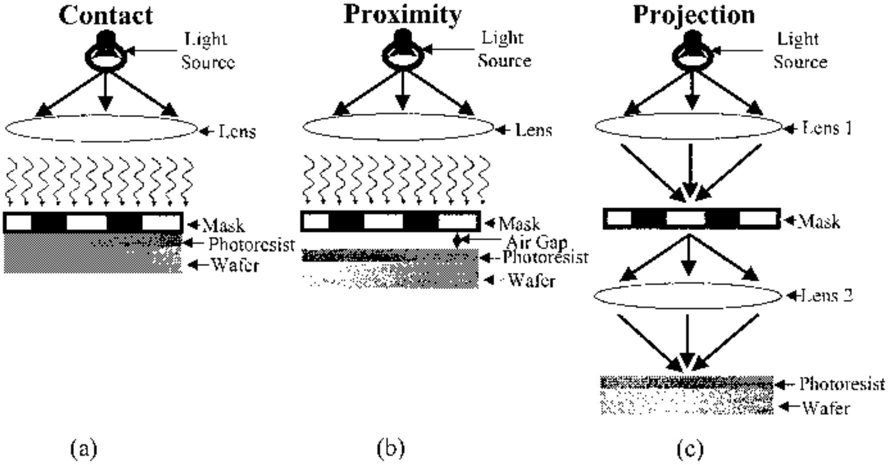


Fig. 16.8. Conceptual drawing of the three main lithographic printing techniques: contact, proximity, and projection.

16.2.5. Development

Indeed, if a photoresist film is underexposed, there is a tendency for the pattern formation to be incomplete and, in the extreme case, to cause a total loss of pattern. For a positive resist which is underexposed, the resist film remains intact after development as shown in Fig. 16.9, whereas for a negative resist it is completely removed. If the film is overexposed, and a positive photoresist is used, then the window openings in the resist are slightly larger than the mask dimension as shown below in Fig. 16.9.

The effects of overexposure also strongly depend on the nature of the photoresist used. For a positive resist, the openings are slightly larger than the mask dimensions as shown below in Fig. 16.10(a). This is due to scattered light penetrating under the mask edges and exposing a small region of film not directly irradiated by the light source. Subsequent etching of the underlying silicon oxide film accentuates this enlargement, as is shown in Fig. 16.10(a). With a negative resist, the window tends to be smaller than the mask dimensions (Fig. 16.10(b)). This can be partially compensated later by the undercut during the etching of the oxide film.

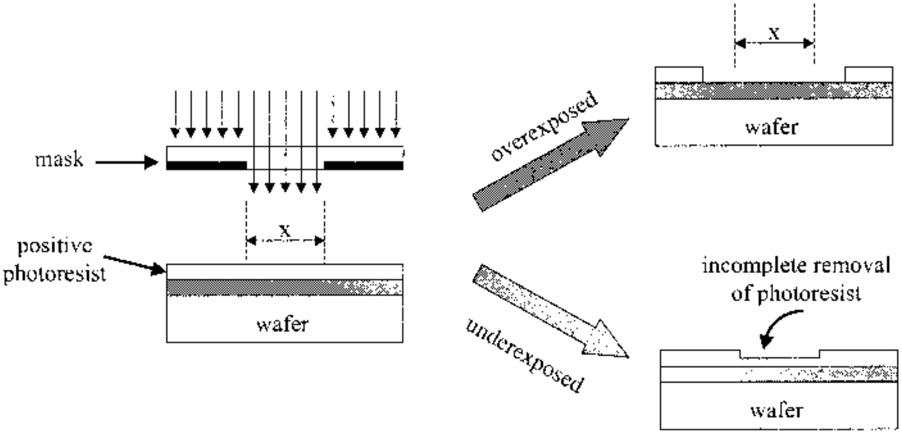


Fig. 16.9. Effect of overexposure and underexposure on the profile of a positive photoresist layer after development.

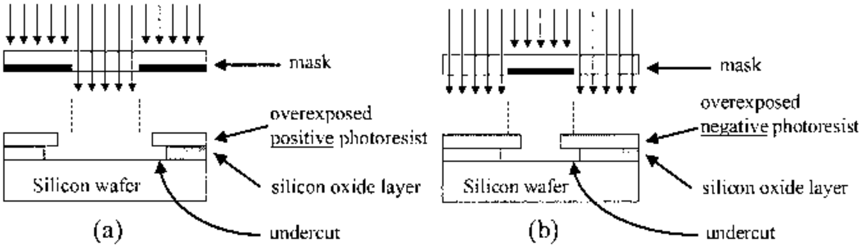


Fig. 16.10. (a) Effect of overexposure on the dimensions of a positive photoresist layer: the opening in the resist layer is larger than the size specified by the mask. When using (b) a negative resist, the opening in the resist is smaller than the size specified by the mask. In this latter case, this effect can be balanced by the undercutting which occurs when conducting the wet etching of the silicon oxide layer.

16.2.6. Direct patterning and lift-off techniques

So far, we have described the process of transferring the pattern from the photoresist film to the underlying (oxide) layer through etching, i.e. the areas which were uncovered were etched away. This method is called direct patterning.

In addition to this traditional method, one can use the lift-off technique for depositing and forming patterned metal or dielectric films onto a wafer surface. In this method, the photoresist is patterned first, before an additional (metal or dielectric) film is applied. Subsequently, by removing or "lifting-off" the photoresist, a pattern is achieved in the later deposited film.

To use lift-off process, the order of photolithography is changed in comparison with direct patterning as shown in Fig. 16.11 and the sequence of steps is summarized as follows:

direct patterning sequence

- Deposit photoresist
- Softbake
- Alignment/exposure
- Develop
- Hardbake
- Develop

lift-off sequence

- Deposit photoresist
- Softbake
- Alignment/exposure
- Postbake
- Flood exposure

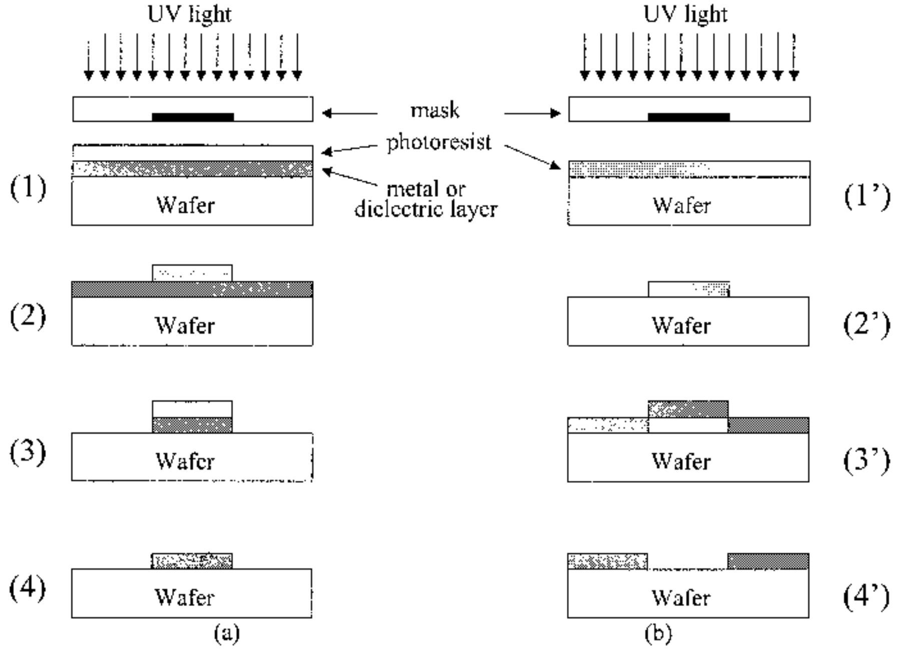


Fig. 16.11. Illustration of the (a) direct pattern vs. (b) lift-off techniques using a positive photoresist. In the direct pattern method, the photoresist is applied after the metal or dielectric layer and is exposed (1), developed (2), and the pattern in the photoresist is transferred to the underlying layer through etching (3), before removing the photoresist (4). In the lift-off technique, the photoresist layer is applied, exposed (1') and developed (2') before the metal or dielectric layer is deposited (3'). The photoresist is then lifted off the wafer, taking away with it parts of the metal or dielectric layer (4').

Fig. 16.12(a) and (b) represent cross-section images showing the photoresist profile used in direct pattern and lift-off technique, respectively, following development. The shape of the resist in Fig. 16.12(b) makes it easier to lift-off the subsequently deposited metal or dielectric layer, without peeling it off completely from the surface.

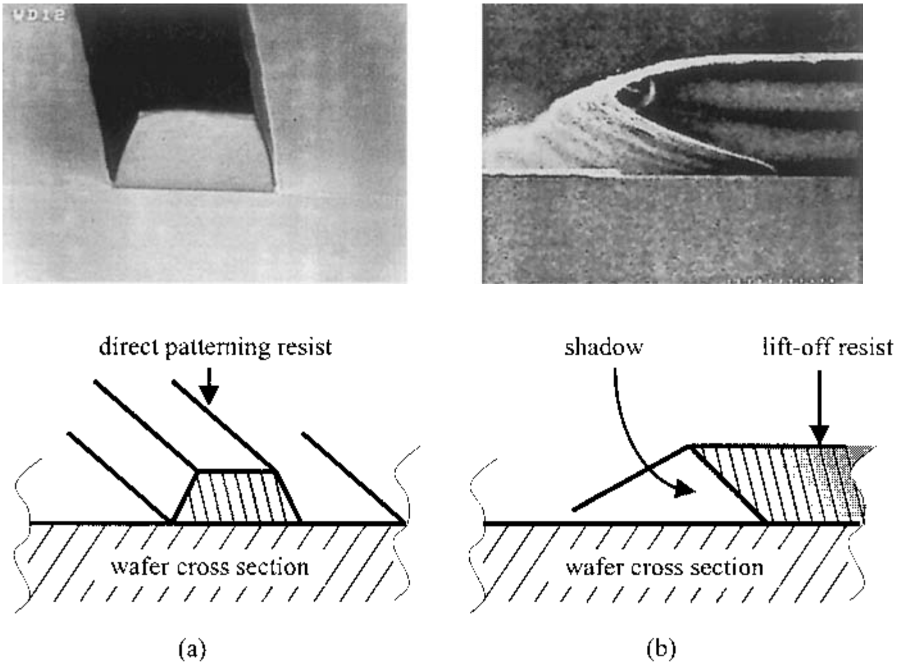


Fig. 16.12. Cross-section views of a (a) direct patterning resist profile following development, and (b) a lift-off resist profile following development. The bottom diagrams show a schematic diagram illustrating the dissimilar cross-sectional shapes of the direct patterning and the lift-off resists.

16.2.7. Alternative lithographic techniques

As microelectronic devices shrink in size, alternative approaches have been investigated due to the fact that optical lithography is fundamentally limited by the phenomenon of diffraction. Work is ongoing to develop alternative lithograghic techniques that can support Moore's law past this diffraction limit. These next generation lithographics (NGLs) include electron-beam lithography where an electron beam is used as the radiation source. This technique is currently growing in importance. It is currently used for the fabrication of the mask, as well as to define nanoscale features, as small as 30 nm. This technique will be the object of a section of the next section.

Other existing NGLs methods include x-ray lithography which is capable of achieving a few tens of nanometer size features thanks to the use of radiation with a wavelength on the order of 4 up to 50 Å. However, this method requires a complex absorber mask or a thin film support structure.

Ion-beam lithography offers patterned doping capability and a very high resolution (~10 nm). All these techniques are schematically illustrated in Fig. 16.13.

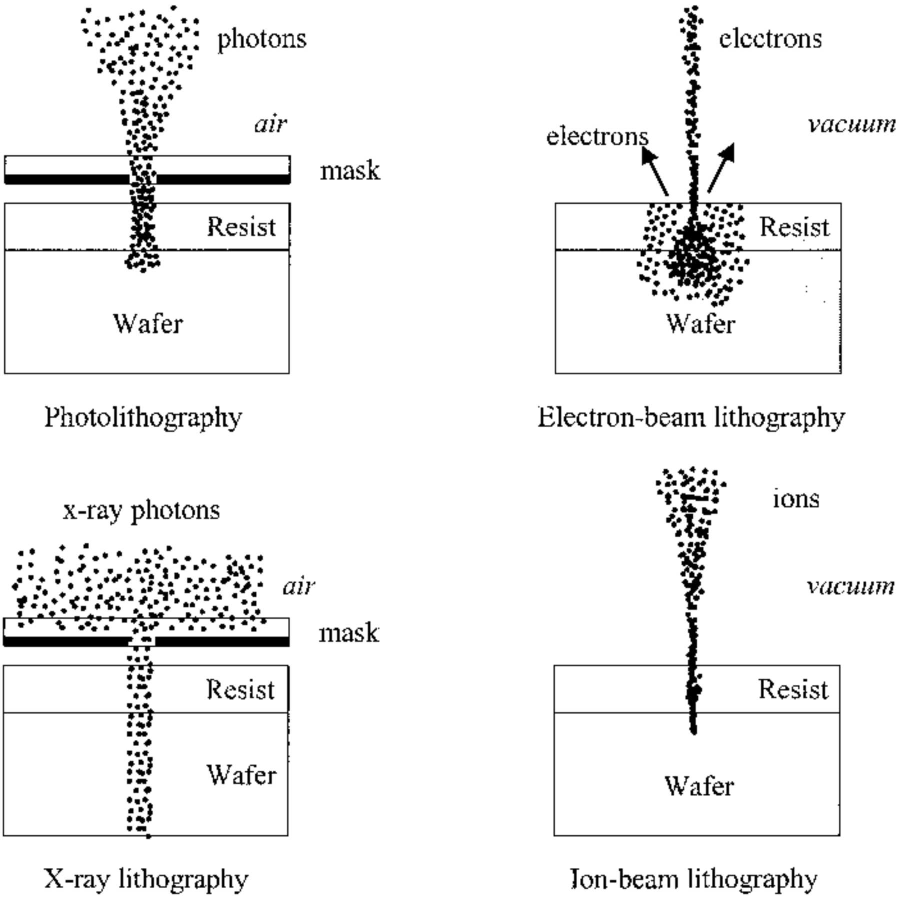


Fig. 16.13. Illustration of different types of lithography. From left to right, and top to bottom: photo-, electron-beam, x-ray and ion-beam lithography. The difference between these techniques resides in the nature of the source used to activate the resist.

While these techniques may seem different from one another, they are fundamentally similar in the fact that they use a photoresist that is activated by a specific energy source. So just like photolithography (Fig. 16.13), they use similar steps as described in this section. The major differences in these types of lithography arise in how they actually interact with the resist. In both photo- and x-ray lithography, the source energy goes through a mask, and the image is projected onto the wafer; this type of lithography is known as an indirect-write pattern. As for ion-beam, or electron beam lithography, the source energy is focused into a fine tip so that the image is directly written onto the surface of the wafer; this known as a direct write pattern.

Two extreme examples of this form of lithography are known as nano-imprint lithography in which a nano-sized stamp is used to impress an image onto the surface or proximal probe lithography (also called Dip-Pen Lithography) where an atomic force microscope tip is used to move material at the atomic scale on the wafer surface. Schematic images of both these techniques are shown below (Fig. 16.14 and Fig. 16.15).

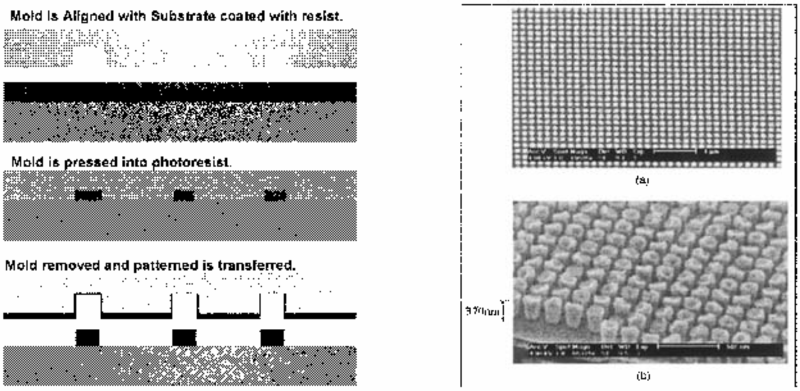

Mold is Aligned with Substrate coated with resist.
Mold is pressed into photoresist.
Mold removed and patterned is transferred.
(a)
370nm
(b)

Fig. 16.14. In nanoimprint lithography, a mold imprints its image directly onto the wafer surface that is covered by the resist. The resist is developed and the image is then fully transposed. Here, an imprint mask used to make nanopillars is shown with the SEM images of the pillars they produce. [SEM images on the right are reprinted with permission from Journal of Vacuum Science and Technology B Vol. 16, Wu, W., Cui, B., Sun, X.Y., Zhang, W., Zhuang, L., Kong, L., and Chou, S.Y., "Large area high density quantized magnetic disks fabricated using nanoimprint lithography," p. 3826. Copyright 1998, American Institute of Physics.]

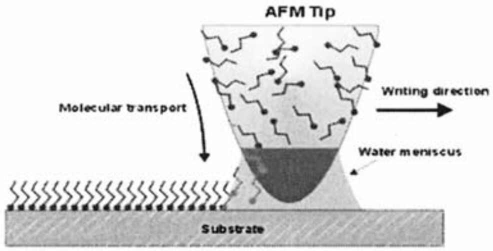


Fig. 16.15. Schematic diagram of dip-pen lithography: either a photoresist is placed down in controlled manner by pacing the AFM tip where it is desired or a molecule is directly placed. [Reprinted with permission from Piner, R.D., Zhu, J., Xu, F., Hong, S., and Mirkin, C.A., "Dip-pen nanolithography," Science Vol. 283, p. 661, Fig. 1. Copyright 1999, AAAS.]

16.3. Electron-beam lithography

The resolution of the photolithography process described in the previous section is limited by the diffraction limit of the ultraviolet light source used during the photoresist exposure. As microelectronic devices shrink in sizes to reach the nanometer scale, novel techniques such as electron-beam lithography have been developed and will be the object of this section.

Electron-beam lithography, or EBL, is a special technique for creating extremely fine features of approximately 20 nm in linewidth. This technique was developed in the 1960's and was inspired by the same technology used in scanning electron microscopy where electrons, rather than light, are used to generate an image. In EBL, the electrons strike a thin layer of resist which has been previously applied on the semiconductor wafer. The properties of the resist material are changed by the presence of the electrons that it encounters. Therefore, instead of using a mask to prevent light from passing, a fine electron beam is scanned across the sample in order to form the desired pattern on the resist.

16.3.1. Electron-beam lithography system

The major component in an electron-beam lithography system is the column, a cross-section of which is illustrated in Fig. 16.16. It contains an electron source, two or more lenses, a blanker that can switch the beam entrance on and off, a beam deflector, a stigmator for correcting any astigmatism in the beam, an aperture for helping to define the beam, alignment systems for centering the beam in the column, and an electron detector for assisting with focusing and locating marks on the sample.

The electron source consists of a heated filament which generates a beam of electrons through thermionic emission. The size of the virtual source, its brightness, and energy spread are the three most important characteristics of the electron beam. The source size dictates the amount of demagnification that the beam must undergo in order to affect the target in a small area. The brightness, measured in amperes per square centimeter through the column, must be high enough to sufficiently affect the resist. The energy spread determines the tendency of the electrons to move outward from the direction of the main beam. There is always some energy spread due to the electric field interaction between the electrons in the beam but this effect can be corrected using apertures. A change in the electron energy will cause the virtual emitter source to change its position slightly too. Provided vacuum is maintained at the specified levels, the electron source may have a lifetime of more than 2000 hours.

The lenses in the column help aim the beam toward the appropriate area of the target. They do not, however, scan the beam across the sample in order to etch patterns on the resist; as that task is left to the deflectors. There are two types of lenses that are commonly used. Electrostatic lenses employ electric fields in order to have an effect on the beam. However, they suffer from both spherical and chromatic aberrations. Spherical aberrations occur when the outside edges of the lens cause the beam to focus more strongly than the inside areas do. Chromatic aberrations are observed when the lens affects differently electrons that have different energies. Both of these effects can be reduced by drastically reducing the size of the beam so that it only passes through the center area of the lens, but this technique reduces the beam current. Magnetic lenses are more commonly used because they cause less aberration.

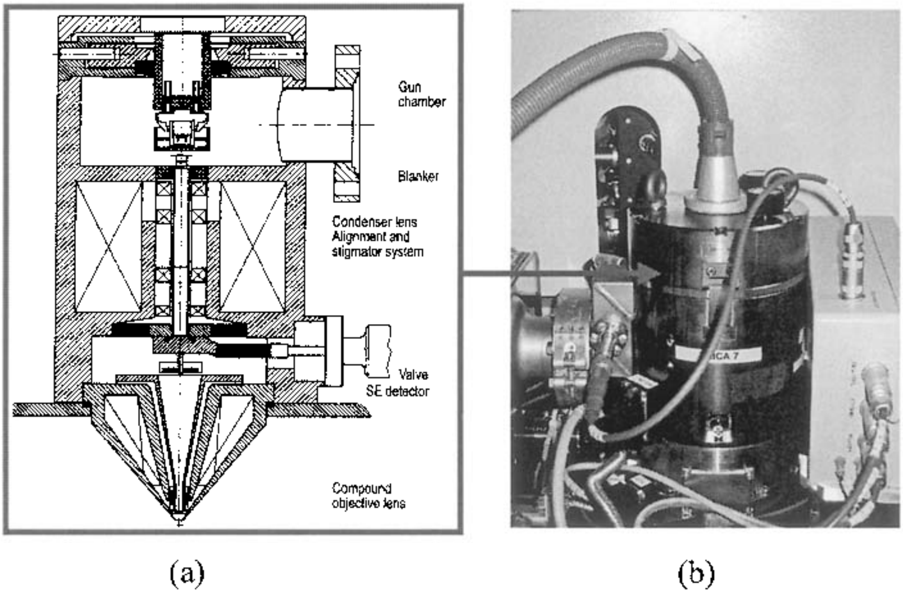


Fig. 16.16. (a) Schematic diagram of a low-voltage column, containing the condenser and objective lens systems, as well as the beam blanker, the alignment and deflector systems, and an electron detector. (b) Photograph of the low-voltage column inside an actual electron beam lithography system. [Reprinted with permission from Leica Microsystems LithographyGmbH.]

Apertures are small holes through which the beam passes as it travels down the column. There are three types of apertures, depending on their diameters. Let us assume that the diameter of the main portion of the beam (excluding stray electrons) is D_B. A spray aperture, which stops stray electrons but does not affect most of the beam, would have a diameter $D_A > D_B$. A blanking aperture, which is used to turn the beam on and off, has a diameter of D_A=0. A beam limiting aperture, which is reduces the beam diameter in order to improve the resolution, would have a diameter D_A<D_B.

The method of appropriately scanning the electron beam in a precise pattern is known as deflection. Similar to lenses, deflectors generate electric or magnetic fields that are used to affect the direction of the electron beam. Electrostatic lenses are more commonly used than magnetic lenses for this purpose because they react faster to the system's demands to scan the beam across the target surface. This difference in speed arises from the existence of inductive magnetic coils necessary to create a magnetic field. Difficulties are encountered when the electron beam interacts with its solid target. Although the beam may be extremely small when it first hits the resist material, the interaction of the beam's electrons with the material causes what is known as forward scattering. For example, an electron that penetrates 1 μm deep into the resist layer when impinging at a 90° angle would travel not only 1 μm down, but also 1 μm laterally when deflected at a 45° angle. This phenomenon results in an error in the actual area of the material which is hit by electron. Using the thinnest possible resist layers can reduce forward scattering. These errors can also be minimized with a technique called dose modulation in which small, isolated areas receive a less intense electron beam current dose than large areas.

Underneath the column, a chamber contains a stage used to load and unload the sample. In addition, a vacuum system is used to maintain an appropriate vacuum level throughout the machine and during the load and unload cycles. A computer controls the EBL system and handles such diverse functions as setting up an exposure job, loading and unloading the sample, aligning and focusing the electron beam, and sending pattern data to the pattern generator.

Electron-beam lithography is becoming an increasingly common alternative to photolithography because of its near-atomic resolution (~20 nm) capability, its flexibility in that it works well for a variety of semiconductor materials and an almost unlimited number of patterns. However, electron-beam lithography is more than ten times slower than photolithography. EBL systems are also expensive and are complex pieces of equipment which require frequent maintenance and adjustment. These limitations keep EBL from becoming the semiconductor industry's lithography standard.

16.3.2. Electron-beam lithography process

Electron-beam lithography uses resists known as polymethyl methacrylate or PMMA which are some of the highest resolution resists available. The PMMA is purchased in two high molecular weights forms (496K or 950K) in a casting solvent such as chlorobenzene or anisole.

Fig. 16.17 briefly illustrates the electron-beam lithography process, which is similar to that of conventional photolithography, including the steps such as resist spinning, exposure, development, and pattern transfer.

To avoid charging effects during electron-beam exposure, the wafer is often first coated with a thin (3~30 nm) indium-tin-oxide (ITO) or chrome layer. The PMMA resist is spun onto the wafer and baked at 170~200 °C for 30 minutes. The electron-beam exposure breaks the resist polymer into fragments that can then be dissolved by a developer solution. There are several types of developer solutions with different strengths, such as MIBK which is a 1:2 solution of (4-methyl-2-pentanone):(2-propanol) or IPA which is simply 2-propanol. MIBK alone is a strong developer and dissolves some of the unexposed resist too, while IPA is a weaker developer. Therefore, mixing them with the appropriate proportions results in a higher contrast or a higher sensitivity for the resist. These two parameters will be defined shortly below. For example, a mixture of 1 part MIBK to 2 parts IPA produces very high contrast but low sensitivity whereas, for a mixture of 1 part MIBK to 1 part IPA, the sensitivity is improved significantly with a small loss of contrast.

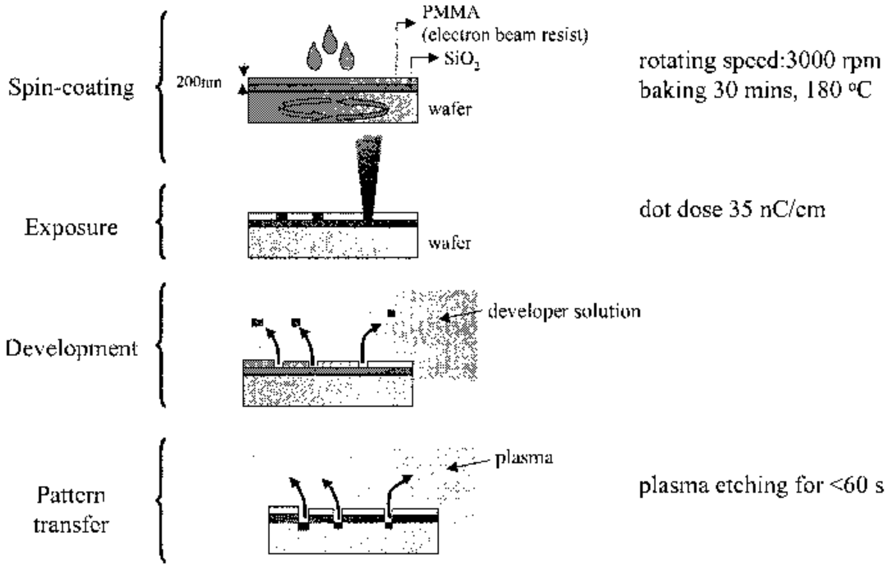


Fig. 16.17. Outline of a typical electron-beam lithography procedure: a PMMA resist film is spun onto a wafer, the resist is exposed to the electron beam following the specified design, the exposed resist is then developed and areas which have been exposed are dissolved, The pattern is transferred on the wafer through plasma etching. The resulting features can be as small as a few nanometers in width.

16.3.3. Parameters of electron-beam lithography

Several parameters need to be understood and determined for electron-beam lithography. For a given set of resist conditions, including the nature of the resist, its thickness, the nominal dose D_{nom} corresponding to a given electron energy level can be defined as the minimum dose required to ensure full dissolution (when using positive resist) or total non-solubility (when working with negative resist) of the resist in all places that were exposed to that electron beam.

If we chose to expose a uniform positive resist to a range of doses, develop the pattern and then plot the remaining average resist thickness (expressed in terms of film retention as a percentage) versus dose, we would obtain the graph shown in Fig. 16.18(a).

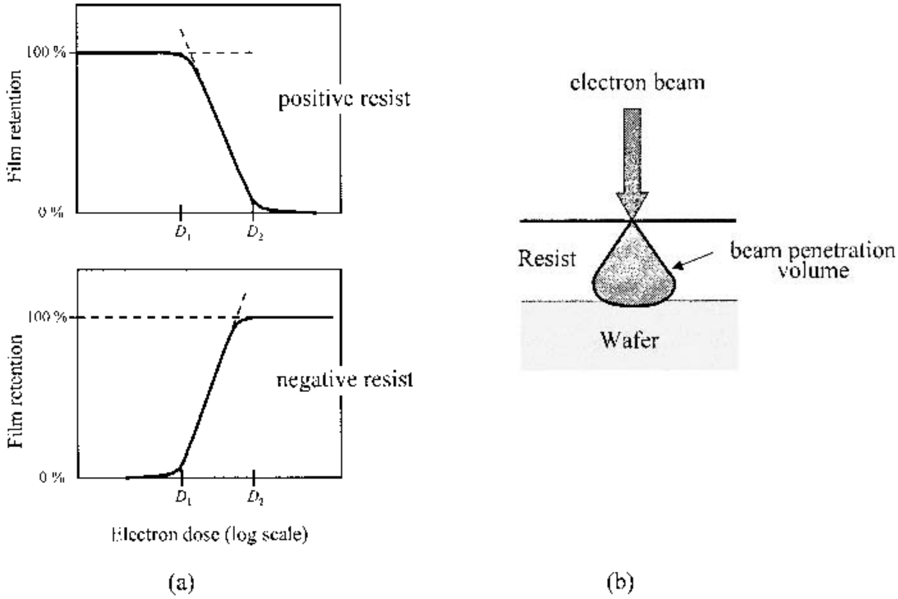


Fig. 16.18. (a) Graphical plots of the resist film thickness as a function of the electron exposure dose for a positive and for a negative resist. For a positive resist, D_1 is the largest dose at which the film remains intact while D_2 is the smallest dose at which the entire film is dissolved. For a negative resist, the meaning of these parameters is reversed. These quantities are used to determine the sensitivity and the contrast of the resist. (b) Illustration of the penetration of the electron beam into the resist film.

In the case of a positive resist, D_1 is the largest dose at which the film remains intact while D_2 is the smallest dose at which the entire film is dissolved. The sensitivity of the resist is defined as the point at which the entire resist is removed. We can also define the contrast γ of the resist as:

Eq. (16.2) $$\gamma \equiv \left| \log\left(\frac{D_2}{D_1} \right) \right|^{-1}$$

The same expression is valid for the contrast of a negative resist, but the meaning of D_1 and D_2 are swapped as shown Fig. 16.18(a). To obtain such graphs, it is essential that the electron beam penetrates the resist completely, as shown in Fig. 16.18(b). To ensure this, the resist must be thinner than the penetration depth of the electrons into the resist material, which is determined by the electron beam energy.

To determine the nominal dose for a given set of resist conditions, especially a given resist thickness, it is recommended to use the "Taxi checker" pattern as illustrated in Fig. 16.19. The inner part of the pattern contains two lines having twenty square sections with an edge length of 5.0 μm. Each square receives a set dose. The applied dose is raised incrementally toward the right, as shown in Fig. 16.19. The dose variation is achieved by varying the exposures time or dwell time, and begins with an initial dose D, hopefully close to the desired nominal dose. This selected dose corresponds to a time t_{dwell} calculated according to:

Eq. (16.3) $$D = \frac{I_{probe} \times t_{dwell}}{(SSZ)^2}$$

where D is the applied dose expressed in C.cm^{-2}, I_{probe} is the probe current at the target level (pA), t_{dwell} is the exposure time per image spot (s) and SSZ is the step size in nm.

The first square in the upper lines is exposed for one tenth of the nominal time, which corresponds to ten percent of the nominal dose. Proceeding toward the right, every further square is given a 10 % higher dose. This is accomplished by incrementing the exposure time by 10 % of t_{dwell}. Exposure of the lower line begins with the first square being subjected to a dose of 110 % of the initially selected dose. The two checker lines thus cover a range between 10 and 200 % of the originally selected dose.

The following sequence of steps is conducted to determine the nominal dose: (i) coat the wafer with PMMA resist with the desired appropriate thickness; (ii) prepare for exposure by optimizing the probe current, calibrating the main deflector unit and the beam tracking unit; (iii) expose the taxi checker patterns, if there is absolutely no initial information about the sensitivity of the resist system to be investigated, then several taxi checkers patterns should be exposed; and this should be done with doses selected such that the largest possible dose area is covered; repeat the same exposure procedure on several other wafers if necessary; (iv) develop the resist by selecting a development time slightly shorter than would be needed for complete development; (v) inspect the resist image under an optical microscope and determine the dose value; (vi) develop the same resist again if possible, for example for 20 additional seconds; (vii) inspect resist image under an optical microscope and determine appropriate nominal dose; (viii) repeat the development and dose determination until the completely developed square no longer shifts toward higher doses, i.e. toward the right, and an additional development and determination step only yield a general change in contrast.

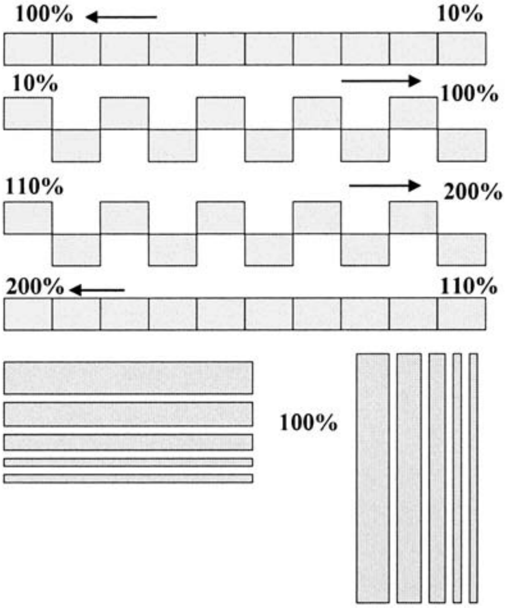


Fig. 16.19. Taxi checker pattern, commonly used to determine the nominal dose for a given set of resist conditions. The inner part of the pattern contains two lines having twenty square sections with an edge length of 5.0 μm. Each square receives a set dose. The applied dose is raised incrementally toward the right.

16.3.4. Multilayer resist systems

In electron-beam lithography, it is often necessary to employ simultaneously two or more different types of resists to achieve a specific lithographic objective when an enhanced undercut is needed for lifting off a metal layer, when a rough surface structure requires planarization, and when a thin imaging top layer is needed for high resolution. A few examples will be given in this sub-section to illustrate this process.

The first example consists of utilizing both a low and a high molecular weight PMMA resist, in a simple bilayer technique as shown in Fig. 16.20. This technique was patented in 1976 by Moreau and Ting and was later improved by Macki and Beaumont by the use of a weak solvent (xylene) for the top layer of PMMA. The low weight resist is more sensitive than the high weight one, so that it develops more easily and results in an enhanced undercut. This feature is useful when lift-off is required from densely packed layers.

The second example uses PMMA with a copolymer resist and was developed by Hatzakis. This method is often employed when it is necessary to deposit a very thick layer of metal (>1 μm in thickness). A high sensitivity copolymer of methyl methacrylate and methacrylic acid (PMMA-MAA) is spun on top of a PMMA resist layer. The exposed copolymer is soluble in solvents such as alcohol and ethers but insoluble in nonpolar solvents such as chlorobenzene. A developer such as ethoxyethanol/isopropanol is used for the top layer, stopping at the PMMA. Next, a strong solvent such as chlorobenzene or toluene is used for the bottom layer. Through this technique, a larger undercut resist profile is achieved which helps the lift-off for the thick metal layer and has been successfully used in the fabrication of memory arrays.

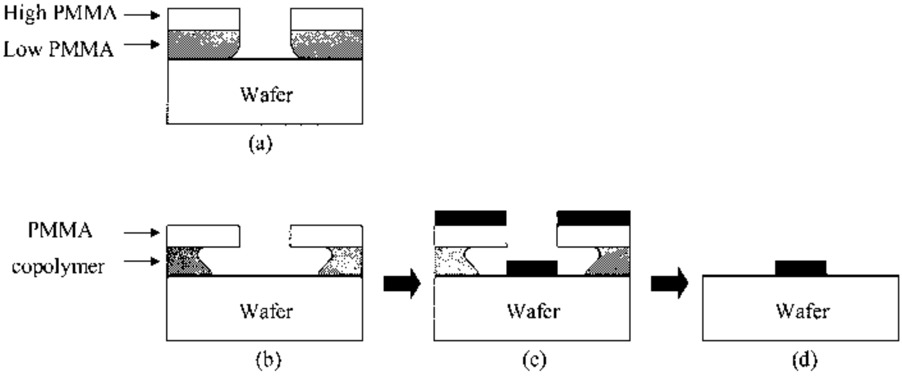


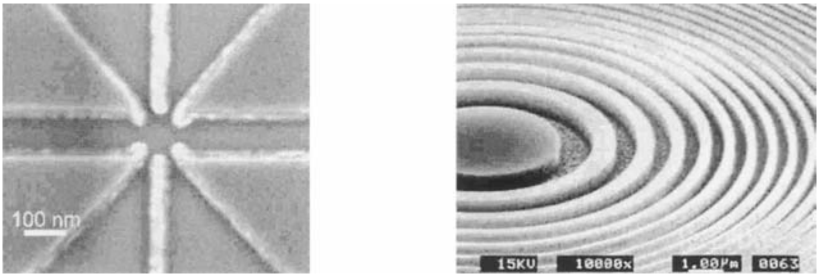
Fig. 16.20. Bilayer electron-beam resist structure: (a) a high molecular weight PMMA is spun on top of a low molecular weight PMMA. The resist is then developed in MIBK:IPA giving a slight undercut. (b) PMMA is spun on top of the copolymer. The resist is developed in MIBK:IPA 1:1 giving a larger undercut. (c) Metal is deposited on top of the resist and (d) is then removed through lift-off.

The third example consists of a trilayer system in which an interlayer is inserted between the two films of the previous bilayer system. Almost any two polymers can be combined in a multilayer system if they are separated by a barrier such as Ti, SiO_2, Al, or Ge. Once the top layer is exposed and developed, the pattern is transferred to the interlayer through dry etching methods (section 16.4) to achieve highly vertical etch profiles. This interlayer would then serve as an excellent mask for the subsequent fabrication of density packed, high aspect ratio resist profiles.

16.3.5. Examples of structures

Electron-beam lithography is by far the most widespread lithography tool for the realization of nanoscale structures. Fig. 16.21 illustrates the resolution and the uniqueness of structures that can be done using this technology. The patterns have linewidths smaller than 100 nm.

The uniqueness of most EBL systems resides in their ability to produce different types of patterns ranging from circular to linear gratings on any type of semiconductor material. This capability is currently used considerably in University research for the development of distributed feedback lasers operating at various wavelengths ranging from 300 nm up to 10 μm.



(a) (b)

Fig. 16.21. (a) Ti/Al gate structure for a SET device generated by electron-beam lithography and lift-off, and (b) a Bragg-Fresnel lens for x-rays exposed in continuous path control mode and etched into Si.

Fig. 16.22 shows an example of linear grating fabricated on top of a ridge quantum cascade laser emitting at 9.0 μm. The grating consists of a first order Bragg grating with a pitch of Λ=1.42 μm exposed by electron-beam lithography and etched 0.5 μm into the surface of the 1 μm-thick top cladding layer by reactive ion etching.

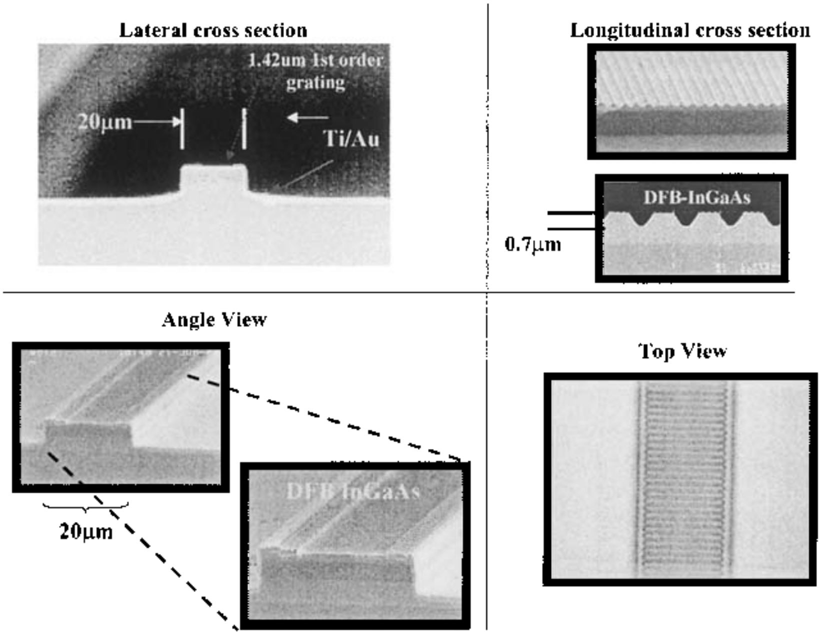


Fig. 16.22. Scanning electron microscopy images of the first order distributed feedback grating for a 9.0 μm quantum cascade laser. The grating is fabricated on the top of the ridge structure using electron-beam lithography.

16.4. Etching

In the previous sections, we have illustrated our discussion with the etching of a layer which had first been covered with a patterned photoresist film, leaving certain selected areas open and others protected. The layer to be etched is generally a dielectric material such as silicon oxide, or a metal used in providing metal contact to the semiconductor. The etching step itself is a complex process which is a function of numerous parameters. For example, the etch can be isotropic, such that the material is etched in equal proportions in all directions, or anisotropic such that one direction is etched more rapidly than any other, or a mixture of both. Several etching techniques can be used and will be described in this section, including wet chemical etching, and dry etching techniques such as plasma etching, reactive ion etching, sputter etching and ion milling.

16.4.1. Wet chemical etching

Wet chemical etching is a mostly isotropic process that etches away in all directions. The process is accomplished by immersing the wafers in an etching solution at a predetermined temperature. An example of solution can be a mixture of hydrofluoric acid and ammonium fluoride. A great variety of wet etch chemistries are available.

One can usually find a solution which is highly selective for the etching of a particular layer while leaving essentially unaffected the adjacent or underlying materials. This material selectivity is an important issue when etching a semiconductor device as it usually contains numerous layers, or when etching a metal contact layer without affecting the underlying semiconductor material.

For most wet etch processes, the material to be etched is not directly soluble in the etching solution, but rather undergoes a reaction with the chemicals present in the solution. The products of this chemical reaction can then be soluble in the solution or can be gaseous. In the later case, the gas can form bubbles which can then prevent the arrival of fresh etching chemical species from reaching the wafer surface to sustain the chemical reaction. This can be a serious problem since the occurrence of these bubbles cannot be predicted. The problem is most pronounced near the edges of the pattern. Mechanical agitation of the wet etching solution can reduce the ability of the bubbles to adhere to the wafer, as well as help sustain the supply of fresh etching reactant.

The advantages of wet etching include its lower cost and the greater versatility of the etching equipment available. Several factors however may affect the quality of wet chemical including: the fact that the photoresist often loses its adhesion to the underlying material when exposed to hot acids, the etching proceeds downward and laterally, thus producing undercutting and broadening lines, and it is difficult to control the etching for submicron geometries. Table 16.1 and Table 16.2 list a few semiconductors, dielectric materials, and metals which are commonly etched in modern microelectronic device fabrication, together with a few wet chemical etching solutions typically used.

Material	Wet etching solution
	$HNO_3 + HF + H_2O$ (5:3:3)
Si	HF
SiO_2	$HF + H_2O$
	BHF (H_2O+HF+NH_4F)
GaAs	$H_2SO_4 + H_2O_2 + H_2O$
	$NH_4OH + H_2O_2$ (pH=7)
InAs	$H_2SO_4 + H_2O_2 + H_2O$
	Br_2 + Methanol
	$HF + HNO_3 + CH_3COOH$
GaSb	Br_2 + Methanol
	$HF + HNO_3 + CH_3COOH$
	$CH_3COOH + HCl + HNO_3 + Br_2$
InGaP	HCl
Si_3N_4	H_3PO_4 (T=160~180 °C)

Table 16.1. A few semiconductors and dielectric materials commonly encountered in microelectronic device fabrication and the common wet etching solutions used.

Metal	Wet etching solution
Au (gold)	$KI + I_2 + H_2O$
Pb (lead)	$H_2O_2 + CH_3COOH$
	$HNO_3 + H_2O$
Ti (titanium)	$H_2SO_4 + HNO_3 + H_2O$ (polishing etch)
	HF
AuZn	$KI + I_2 + H_2O$
Ni (nickel)	$HNO_3 + CH_3COOH + H_2SO_4$
	HCl
	$H_2SO_4 + H_2O$ (Electrochemical polish)
Al (aluminum)	$H_3PO_4 + HNO_3 + CH_3COOH + H_2O$

Table 16.2. A few metals and their wet etching solutions.

Wet etching is also dependent on the crystallographic orientations of the semiconductor crystal, which determines the atomic packing density of the different planes exposed to the etching chemicals. Fig. 16.23 shows the etch planes and profiles when the protective resist is oriented along various directions on a (001) GaAs wafer.

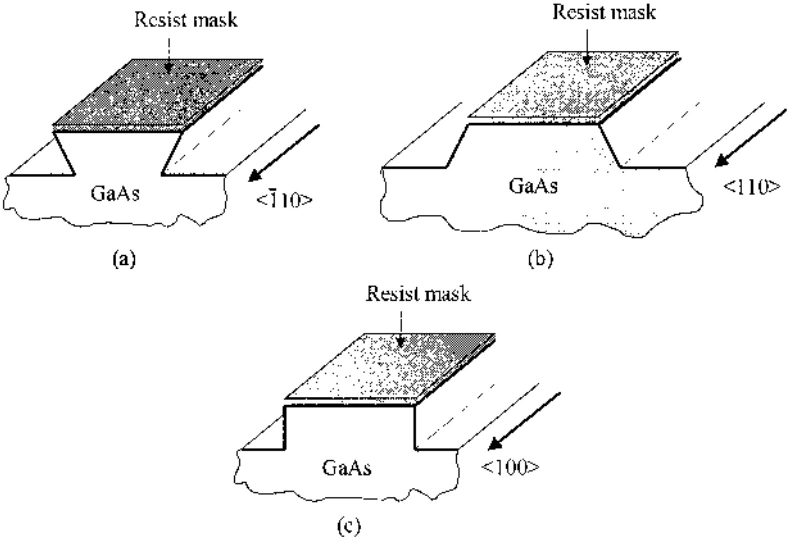


Fig. 16.23. Etch profiles of a (001) oriented GaAs crystal, obtained with most GaAs etchants as a function of the in-plane crystal orientation. When the side of the protective resist mask is oriented in the (a) $<\overline{1}10>$, (b) <110>, and (c) <100> directions.

16.4.2. Plasma etching

By contrast to wet etching, dry etching is not performed in a solution but rather in a gaseous environment. It either consists of plasma driven chemical reactions and/or energetic ion beams aimed at removing the material. Dry etching is commonly used to obtain highly anisotropic etch profiles as the one shown in Fig. 16.24. Some of the advantages of dry etching over wet etching are its greater control at a reduced cost, its substantial directionality i.e. high anisotropy, its effectiveness to reduce the undercutting of masking patterns, and the possibility to precisely etch smaller geometry features. There exist a variety of dry etching techniques including: plasma etching, reactive ion etching, sputtering etching, and ion milling. In this sub-section, we will describe the plasma etching method.

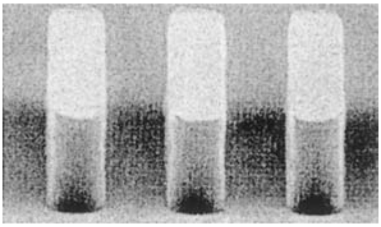


Fig. 16.24. Scanning electron image of a highly anisotropic etch profile obtained using dry etching techniques.

Plasma etching refers to any process in which a plasma, or a gas of charged particles, generates reactive enough species that they serve to chemically etch or physically remove material in the immediate proximity of the plasma. The wafers masked with a photoresist are placed in a vacuum chamber system. A small amount of reactive gas plasma, for example of oxygen, chlorine, or fluorine, is allowed into the chamber. An electromagnetic field is then applied to obtain a directional beam of excited ions and the material that is not protected by the photoresist is etched away by the excited ions. The key to plasma etching is the ability to couple the electromagnetic energy into the reactive species while not heating the rest of the gases in the chamber. Fig. 16.25 and Fig. 16.26 show two popular types of plasma etching reactors: a barrel reactor or a planar reactor.

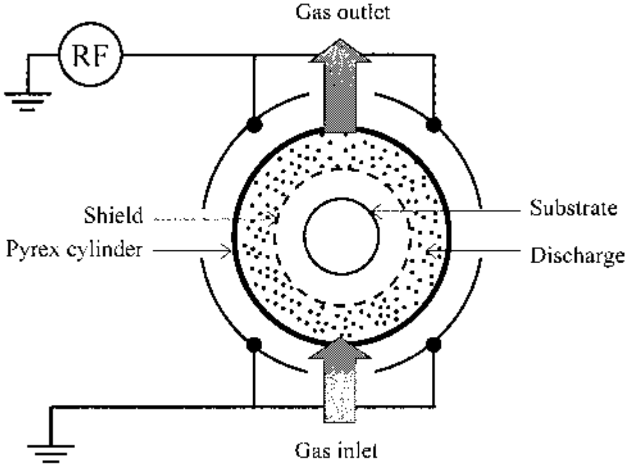


Fig. 16.25. A typical barrel reactor. The plasma is excited using inductive or capacitive electrodes outside of the quartz chamber. The substrates are held in the vertical position by a slice holder and are immersed in the plasma with no electrical bias applied.

In barrel systems, the plasma is excited using inductive or capacitive electrodes outside of the quartz or glass cylindrical chamber. The substrates are held in the vertical position by a slice holder and are immersed in the plasma with no electrical bias applied.

The planar system consists of two flat and parallel electrodes of the same size. The substrates are placed flat on the lower electrode which is also used as a heating stage. Electrons are created in the plasma by the dissociation of atoms into ions. Since they have a greater mobility than the positive ions, they move from the plasma onto the electrode surfaces, thus giving them a negative charge with respect to the plasma. This results in an electric field across the plasma sheath, between the plasma and the electrodes. This field then causes the ions at the edge of the plasma sheath to be accelerated toward the electrodes.

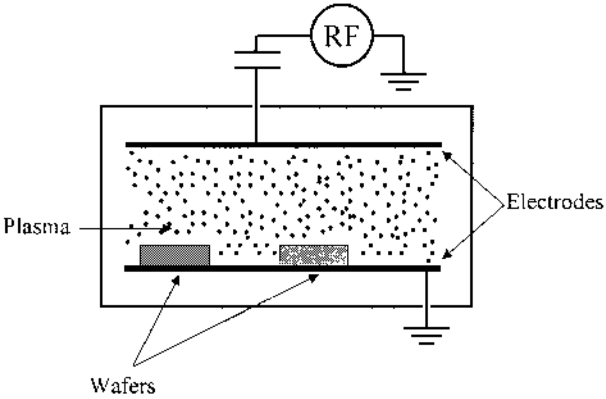


Fig. 16.26. A typical planar reactor, consisting of two flat and parallel electrodes of the same size. The substrates are placed flat on the lower electrode which is also used as a heating stage. The electric field which appears between the plasma sheath and the electrodes causes the ions at the edge of the plasma sheath to be accelerated toward the electrodes, thus impinging on the wafers.

Because of the geometry of the planar reactor, the ions are accelerated perpendicularly to the electrodes except near the outer radius where deviations can lead to corresponding distortions in the etch profile. This perpendicular impingement of energetic ions makes the anisotropic etching possible.

Fig. 16.27 illustrates an anisotropic plasma etch. This process can also lead to the formation of a passivating film on the vertical sidewalls. For example, when a fluoride compound is used in the etching chemistry, e.g. carbon tetrafluoride or CF_4, a fluorocarbon film deposits on all surfaces. Since the ions mostly follow a vertical path, there is little ion bombardment on the sidewalls and the fluorocarbon film can accumulate there, as the etching proceeds. The nature of such a film depends on the plasma conditions.

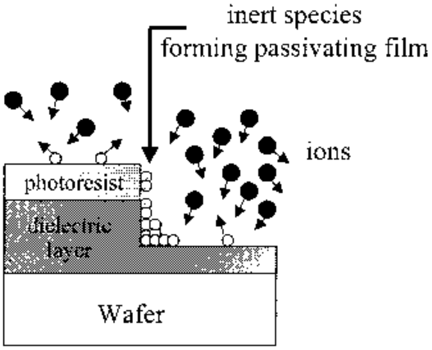


Fig. 16.27. Schematic diagram of an anisotropic plasma etch, showing the formation of a passivating film on the sidewall from the products of the chemical reactions which occur during the etch.

Table 16.3 lists the advantages and disadvantages that are commonly associated with in plasma etch using either type of reactors. It is important to be aware that the plasma etch rate for a process is measured for a given set of process conditions, from which the duration needed to etch a layer with a particular thickness can subsequently be determined.

	Advantages	Disadvantages
Barrel reactor	- good to remove resist	- poor uniformity due to gas flow RF fields - etch rates tend to increase from the center of wafer to the edges
Planar reactor	- good uniformity - suitable for selective etching through masking patterns	- ionic bombardment can damage wafer surface layers - can lead to significant undercutting

Table 16.3. Advantages and disadvantages in plasma etch using barrel or planar reactors.

Plasma etching provides a poor reproducibility because subtle changes in the etch process or in the film properties can result in poor uniformity or rough surfaces. Fig. 16.28 is an example of a planar reactor plasma etch where the surface of the sample has been damaged due to an increase in the energy of the bombarding ions.

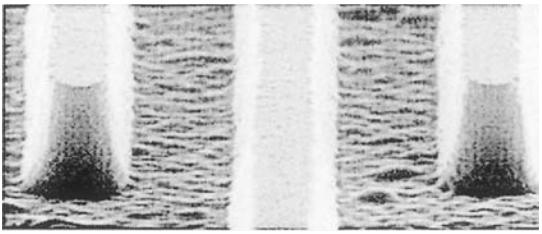


Fig. 16.28. Scanning electron microscope image of a damaged semiconductor surface obtained when excessive ion bombardment is used. The damage can be seen from the roughness of the surface.

16.4.3. Reactive ion etching

Reactive ion etching (RIE) is very similar to plasma etching and a clear distinction is difficult to make. In the RIE process, the emphasis is primarily put on the directionality of the etch. The etch parameters such as the pressure and the configuration of the etching equipment are modified to ensure a directional ion bombardment. A typical RIE system is illustrated in Fig. 16.29.

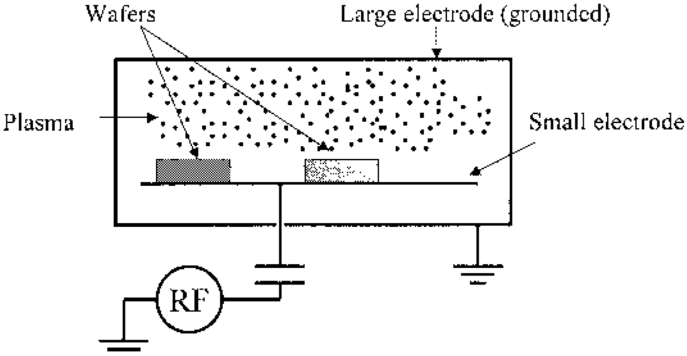


Fig. 16.29. A typical RIE etching system which is similar to plasma etching. It includes a chamber under vacuum, the wafers are placed on a small electrode and the inside walls of the chamber constitute the ground electrode.

By contrast to plasma etching, RIE systems operate at much lower pressures: 0.01 to 0.1 Torr. The advantages of RIE are its highly anisotropicity and directionality, but its disadvantages are that the stage needs to be cooled in order to resist the temperature rise.

16.4.4. Sputter etching

Sputter etching is a purely mechanical process in which energetic ions from the plasma of inert gases, such as argon, strike the wafers and physically blast atoms away from the surface. The sputtering technique and the reactive ion etching are both carried out with the wafer placed directly on an electrically powered electrode in contact with the plasma, as shown in Fig. 16.30. In this configuration, the ions impinge on the sample at near normal incidence. In sputter etching, the chamber is maintained at a low vacuum (10^{-2} Torr) and, as it contains the plasma discharge, it is exposed to ultraviolet radiation, x-rays, and electrons as part of the plasma environment.

The advantages of sputter etching include its high anisotropy, whereas one of its disadvantages is its poor chemical selectivity: all materials are nearly etched at the same rate in this process.

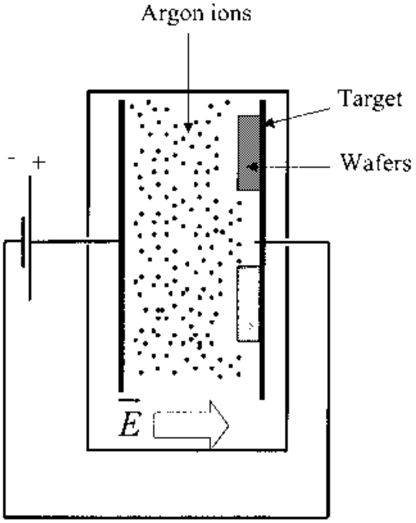


Fig. 16.30. Schematic diagram of a sputter etching system. Energetic ions from the plasma of inert gases are accelerated between two electrodes in a chamber under vacuum and strike the wafers to be etched which are placed directly on one of the electrodes in contact with the plasma.

16.4.5. Ion milling

Ion milling is also a purely mechanical process in which the incoming ions are energetic enough to sputter material from the surface of the wafer. Inert gases are also generally used here as well. One key parameter in this process is the sputtering yield which is defined as the number of sputtered atoms per incident ion. This quantity depends on the material being etched, the nature of the impinging ions, their energy, angle of incidence, and the composition of the background atmosphere

In ion milling, the positive ions are generated in a confined plasma discharge and accelerated in the form of a beam towards the sample to be etched, as shown in Fig. 16.31. Inert gases, such as argon, are usually used in these systems because they exhibit a higher sputtering yield than other atoms and also because they do not participate in chemical reactions. A neutralizer, usually a hot filament, emits a flux of electrons to cancel the positively charge of the ions while keeping their kinetic energy for etching. The sample can thus be kept neutral, so that no deleterious charging effects occur. The sample resides in a moderate vacuum (10^{-4}~10^{-6} Torr) environment and can be attached to a cooled substrate holder to maintain its surface at low temperatures during etching. The sample can also be positioned at any angle with respect to the incoming ion beam.

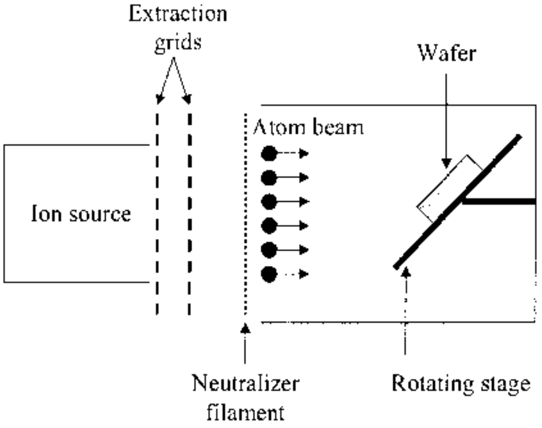


Fig. 16.31. Schematic diagram of an ion-beam milling system. The entire system is under moderate vacuum. The positive ions of inert gas are generated in a confined plasma discharge, away from the wafer, and accelerated in the form of a beam towards the sample to be etched. A neutralizer, usually a hot filament, emits a flux of electrons to cancel the positively charge of the ions. The wafer can be attached to a cooled substrate holder and be positioned at any angle with respect to the incoming ion beam.

Although there are similarities between ion milling and sputter etching, the ion milling process offers more flexibility, can be carried out at a lower temperature, in a less harsh etch environment, and results in a reduced redeposition of contaminants.

The main advantage of ion milling over other methods is the absence of undercutting in this process. However, this technique suffers from a number of disadvantages: it is a slow process, it generates a good amount of heat which makes the subsequent removal of the resist difficult, the sputtered material can redeposit anywhere on the wafer surface, scattering effects make the etching of vertical edges much faster, and trenching effects can occur as a result of the sample tilting.

16.5. Metallization

16.5.1. Metal interconnections

In addition to lithography and etching, a third important step in the fabrication process of a semiconductor device is the deposition of metals in certain areas of the wafer, through a process called metallization. This is done in order to allow the surface wiring of individual semiconductor layers and metal interconnection between contacts in a microcircuit as shown in Fig. 16.32.

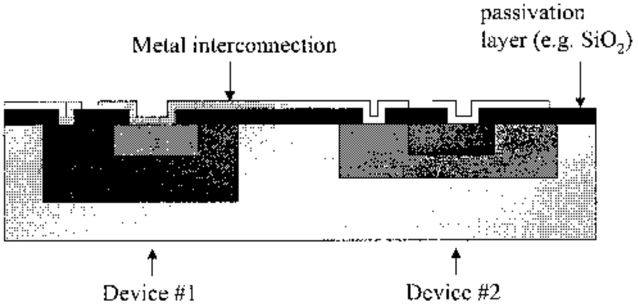


Fig. 16.32. Single layer metal interconnection between two devices. A passivation layer, such as silicon dioxide, is generally first deposited and patterned to isolate areas of the devices which must be electrically connected. Then a metal layer is deposited and patterned to connect electrically one part of a device to another part of a second device.

The metal interconnection is first deposited in the form of thin films of various materials on the surface of the semiconductor wafer. The thicknesses of these films are typically on the order of 1500 to 15000 Å. There exist several deposition techniques which will be discussed below. However, in order to be useful, the metal must not cover the entire wafer uniformly, but only certain areas. Lithography is then used to define the areas where the metal will remain. The direct patterning and the lift-off techniques are equally used in this process. A few examples of wet chemical etching solutions for various metals were given in Table 16.2. In order to isolate metal interconnects from one another, to prevent current leakage and short circuits, a protective or passivation layer of dielectric material (e.g.

silicon dioxide) is often used, as shown in Fig. 16.32. In most cases, following the formation of the metal interconnects, a heat treatment step called alloying is performed typically between 200 and 400 °C in order to ensure good mechanical and electrical contact between the metal and the semiconductor wafer surface.

The metal materials used must ideally satisfy a number of properties, such as: have a good electrical current-carrying capability, a good adhesion to the top surface of the wafer, a good electrical contact with the wafer material, be easy to pattern (etch or lift-off), be of high purity, be corrosion resistive, and have long-term stability.

Most of these characteristics are met by either gold or aluminum. In microelectronic circuit technology, aluminum is the most commonly used metal interconnect because it adheres well to both silicon and silicon dioxide, although it is less conductive than copper or gold. Aluminum also has a good current-carrying capability and is easy to pattern with conventional lithography process.

However, one metal material is often not sufficient to satisfy all the properties mentioned previously. This is why most circuit designs require the use of multilayer metal films, such as platinum and gold or a combination of titanium, platinum and gold. A multilayer metal stack makes it possible to avoid the use of gold as a direct electrical contact with silicon because gold adheres badly to the semiconductor surface and is at the origin of significant current leakage which can impair the device performance.

16.5.2. Vacuum evaporation

The deposition of metal thin films on a semiconductor wafer is commonly accomplished through vacuum deposition. Fig. 16.33(a) shows a typical vacuum deposition system. It consists of a vacuum chamber, maintained at a reduced pressure by a pumping system. The shape of the chamber is generally a bell jar that is made of quartz or stainless steel, inside which many components are located, including: the metal sources, a wafer holder, a shutter, a thickness rate monitor, heaters, and an ion gauge to monitor the chamber pressure.

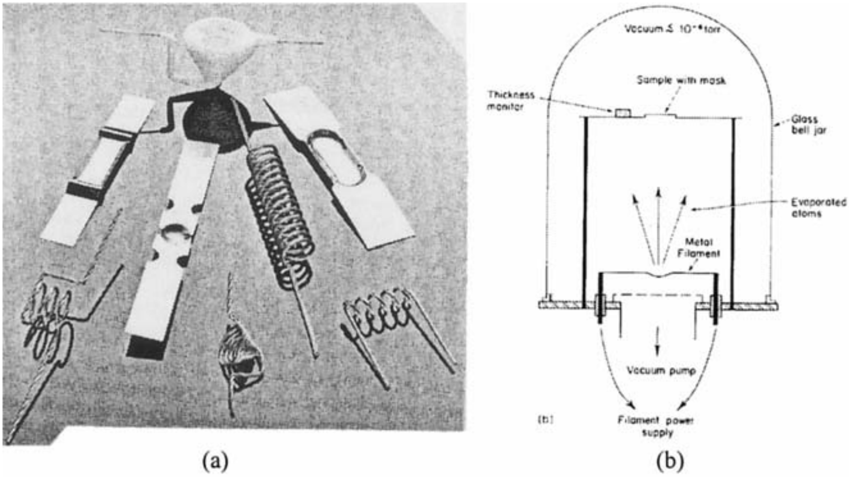


Fig. 16.33. (a) Examples of evaporation sources which can be used in vacuum evaporation systems; (b) cross-section of a typical vacuum evaporation system which includes a glass bell jar under vacuum, a sample holder, a metal filament (evaporation source), and a thickness monitor. [An Introduction to Semiconductor Microtechnology, Morgan, D.V. and Board, K. Copyright 1990. ©John Wiley & Sons Limited. Reproduced with permission.]

It is important to operate at a reduced pressure for a number of reasons. First is a chemical consideration. If any air or oxygen molecule is found in the vacuum chamber during the evaporation of aluminum, the metal would readily oxidize and aluminum oxide would form in the depositing film. Reducing the pressure ensures that the concentration of residual oxygen molecules is small enough to minimize the oxidation reaction. Secondly, the coating uniformity is enhanced at a higher vacuum. Indeed, at low pressures, the mean free path of the evaporated metal atoms is increased, that is the distance traveled by the atoms before collision with another atom. When the mean free path exceeds the dimensions of the chamber, this ensures that the metal atoms will strike the wafers before hitting another atom which would have caused non-uniform depositions.

Two ranges of vacuum conditions are typically used for vacuum evaporation: the low and medium vacuum range (10^5~10^{-1} Pa) and the high vacuum range (10^{-1}~10^{-4} Pa). For the low-medium vacuum range, a mechanical roughing pump is sufficient. To attain the high vacuum range, a roughing pump is first used to evacuate the chamber to the medium vacuum range, then a high capacity pump takes over. The most common high capacity vacuum pumps include diffusion pumps, cryo pumps, and turbomolecular pumps.

The physical laws governing the evaporation of metal are those of the kinetic theory of gases in which each particle, e.g. metal atom or gas molecule, is modeled as moving freely in space with a momentum and energy, which is subject to instantaneous collision events with other particles, the probability for a collision to occur is proportional to the interval of time since the previous collision, and the particles reach thermal equilibrium only through such collisions. In this model, the mean free path of a particle is given by:

Eq. (16.4) $$\lambda = \frac{k_b T}{\sqrt{2}\pi P d^2}$$

where k_b is the Boltzmann constant, T the absolute temperature, d is the diameter of the particle, and P is its partial pressure in the chamber which is related to the concentration of the particles n in the chamber through the ideal gas law:

Eq. (16.5) $$n = \frac{PV}{k_b T}$$

where V is the volume of the chamber. A high quality film can only be obtained with a clean environment, e.g. a clean chamber, pure source material and clean wafer surface.

There are three different types of evaporation techniques, depending on the method used to physically evaporate the metal from its solid state: filament evaporation, electron-beam evaporation, and flash hot plate.

The filament evaporation is the simplest of these methods. The metal can be in the form of a wire wrapped around a coiled tungsten that can sustain high temperatures and current as shown Fig. 16.33(b). The metal can also be stored in tungsten boats if large quantities of material are required. An electrical current is passed through the tungsten boat, thus heating and melting the metal into a liquid which can then evaporate into the chamber at low pressure. Filament evaporation is not very controllable due to temperature variations along the filament. Another drawback when using filaments is that the source material can easily be contaminated and the contaminants can subsequently be evaporated onto the wafers. Moreover, mixtures of metal alloys containing for example titanium, platinum, nickel, and gold are difficult to achieve using the filament evaporation method because each metal has a different evaporation rate at a given temperature.

To avoid such problems, the electron-beam evaporation technique was developed. Fig. 16.34 is an illustration of the principle of an electron-beam source, which consists of a copper holder or crucible with a center cavity which contains the metal material. A beam of electrons is generated and bent by a magnet flux so that it strikes the center of the charge cavity as shown in Fig. 16.34. In addition, the solid metal within the crucible is heated to its melting point such that it presents a smooth and uniform surface where the electron beam hits, thus ensuring that the deposition on the wafer is uniform. The crucible is cooled with water to maintain the edges of the metal in a solid state. Electron-beam evaporation is relatively controlled for a variety of metals such as aluminum and gold. This method had its own limitations: it can only evaporate one alloy at a time. But, over the years electron-beam systems have incorporated multiple guns so that each material will have its own electron beam.

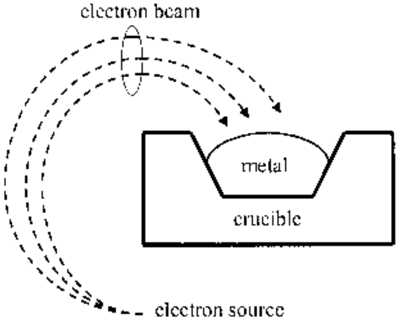


Fig. 16.34. Schematic diagram of an electron-beam evaporation source in which electrons are generated by a high temperature filament and accelerated into an electron beam. A magnetic field bends the electron beam so that it hits a metal charge in a water-cooled crucible. The impact melts the metal and allows it to evaporate in the vacuum of the deposition chamber.

The *flash hot plate* method uses a fine wire as the source material. This fine wire which contains an alloy material is fed automatically onto a hot plate surface. Upon contact the tip of the wire melts and the material "flashes" into a vapor and coats the wafers in the chamber. Since all of the elements are flashed simultaneously, the composition of the metal film deposited on the wafer is close to the alloy composition of the wire.

16.5.3. Sputtering deposition

Sputtering deposition is also called physical vapor deposition and is a physical process. A typical sputtering deposition system is shown in Fig. 16.35. It contains a slab or target of the desired metal which is electrically grounded and serves as the cathode. Under vacuum conditions, argon gas is introduced into the chamber and is ionized into a positively charged ion. These are accelerated toward the cathode target. By impacting the target, enough metal atoms are dispersed such that they deposit onto the wafer surface. The main feature of the sputtering method is that the target material is deposited on the wafer without chemical or compositional change.

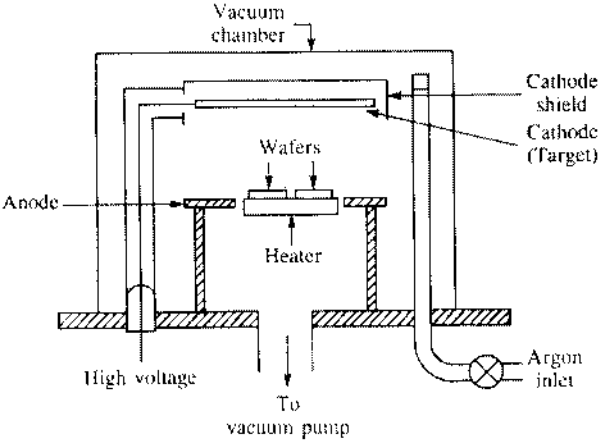


Fig. 16.35. Cross-section schematic diagram of a typical sputtering system, which is enclosed in a vacuum chamber and includes the wafers which are placed on a heater, and a set of electrodes, one of which is made from the target material to be sputtered. Argon gas is supplied and ionized so that ions can impact on the target to release atoms of the material to be deposited. [JAEGER, RICHARD C., INTRODUCTION TO MICROELECTRONICS FABRICATION: VOLUME 5 OF MODULAR SERIES ON SOLID STATE DEVICES, 2nd Edition, © 2002. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

Sputtering has several advantages over other traditional evaporation techniques. For example, the composition of the deposited film is precisely determined by that of the target material, step coverage is improved, and sputtered films have a higher adhesion.

As with the evaporation technique, a high quality film can only be obtained with a clean environment, e.g. clean chamber, pure source material and clean wafer surface.

16.6. Packaging of devices

The final step in the fabrication of a semiconductor device consists of separating the individual components on a same wafer and packaging them.

16.6.1. Dicing

At the industrial scale, mass produced wafers contain a large number of equivalent integrated circuits which need to be separated from one another.

Each resulting circuit is called a die chip. This can be accomplished for example by using a diamond saw as shown in Fig. 16.36 (a). As the demand for accuracy becomes important and tolerances get tighter, other forms of separation have been developed including for example a laser beam as shown in Fig. 16.36(c).

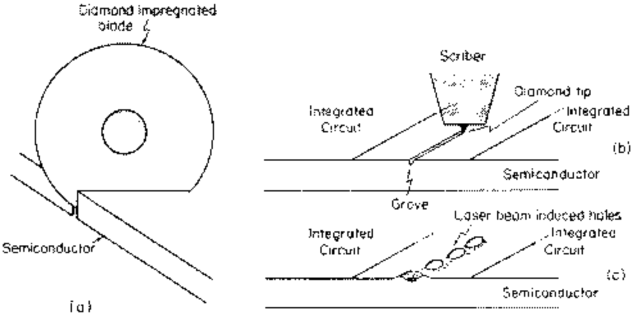
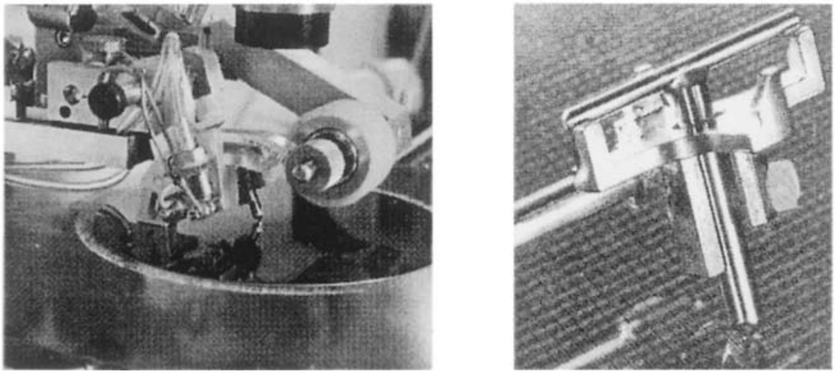


Fig. 16.36. Illustration of the various methods that can be used to separate individual devices from a semiconductor wafer by using: (a) a diamond saw, (b) a scriber, or (c) a laser beam. [An Introduction to Semiconductor Microtechnology, Morgan, D.V. and Board, K. Copyright 1990. © John Wiley & Sons Limited. Reproduced with permission.]

Fig. 16.37(a) is a photograph of a modern commercial scribing tool, while Fig. 16.37(b) is a close up look at the scriber tip.



(a) (b)

Fig. 16.37. (a) Photograph of a modern commercial scribing tool showing the scriber tip, with its positioning wheels, and a camera. (b) Close up photograph of the scriber tip. [Reprinted with permission from Loomis Industries, Inc.]

Fig. 16.38 is a photograph illustrating a wafer after scribing on which one can see delimitated chip-scale die components. Following the dicing of the wafer, each individual die chip is sorted and inspected under a microscope before wire bonding and packaging.

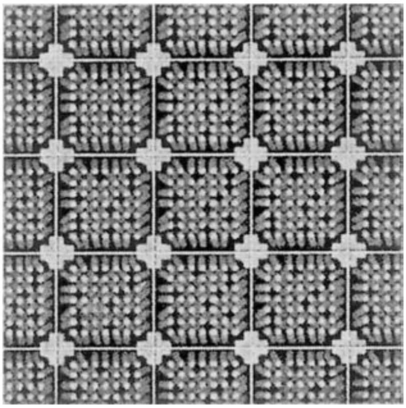


Fig. 16.38. Photograph of chip-scale die components delimitated on a wafer after scribing. [Reprinted with permission from Kulicke & Soffa Industries.]

16.6.2. Wire bonding

It is necessary to link the metal interconnects which have microscopic sizes to a macroscopic electrical connector. The method used is called wire bonding. The wire used consists of gold or aluminum with a diameter of about 10 to 50 micrometers. Gold wire is generally used in industry as it welds readily to both aluminum and gold contact pads by heat and pressure. This process is known also as thermocompression bonding.

A fine wire of gold is fed through a resistance-heated tungsten carbide capillary tube as shown in Fig. 16.39(a). Applying an electric spark melts the exposed end of the wire which is brought down with pressure upon the area of the metal contact where it is then welded. Under manual or automated control, the capillary is moved to another contact pad where the second bond will be made. The capillary is then raised and the wire is broken near the edge of the bond by an electric spark which forms a ball.

A variation of this technique is the pulse-heated thermocompression bonding method, as shown in Fig. 16.39(b), in which the tungsten carbide bonding tool is heated by a pulse of current rather than an electric resistance.

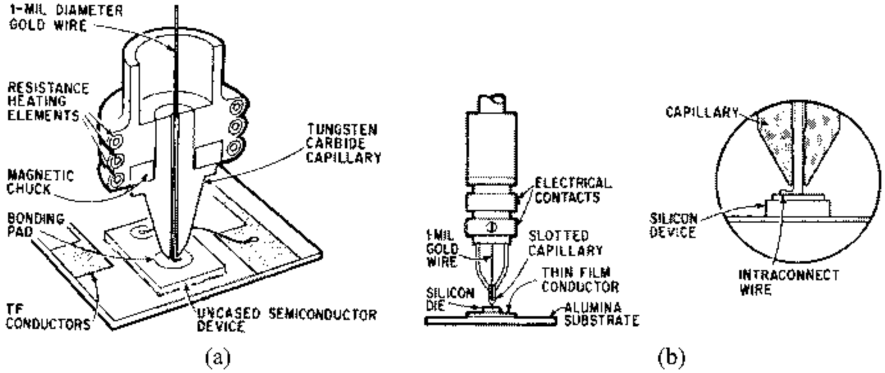


Fig. 16.39. Schematic diagrams of (a) a resistance-heated thermocompression wire bonder, and (b) a pulse-heated thermocompression wire bonding tool. Applying an electric spark melts the exposed end of the wire which is brought down with pressure upon the area of the metal contact where it is then welded. Under manual or automated control, the capillary is moved to another contact pad where the second bond will be made. [Fogiel, M., Microelectronics-Principle, Design Techniques, and Fabrication Processes. Copyright © 1968 by Research & Education, Inc. Reprinted by permission of Research and Education Association, New York.]

The sequence of steps during these thermocompression processes is schematically illustrated in Fig. 16.40. An example of wire-bonded die viewed at high magnification under electron microscopy is shown in Fig. 16.41.

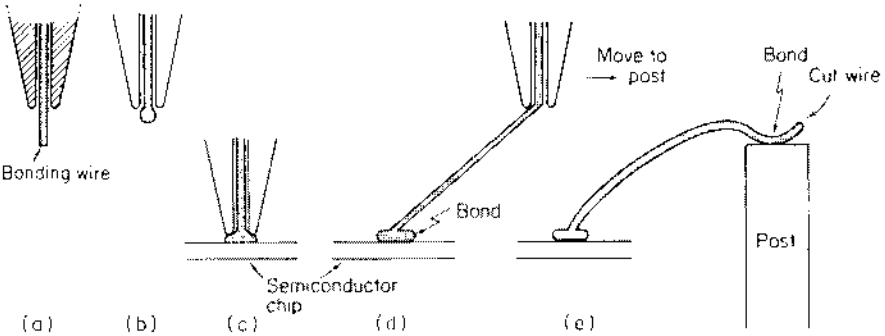


Fig. 16.40. Schematic diagrams showing the sequence in the thermocompression wire bonding process. [An Introduction to Semiconductor Microtechnology, Morgan, D.V. and Board, K. Copyright 1990. © John Wiley & Sons Limited. Reproduced with permission.]

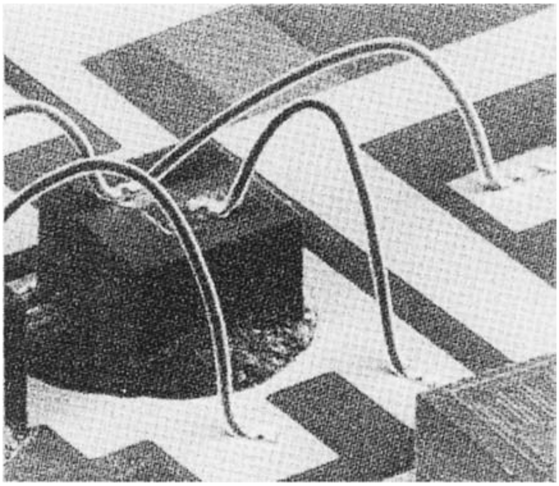


Fig. 16.41. An example of wire-bonded die viewed under electron microscopy. [Reprinted with permission from Kulicke & Soffa Industries.]

While heating of the wire works quite well for gold, when aluminum wire is used the high temperatures that are required cause oxidation which makes it difficult to form a good ball at the end of the wire. Thus an alternative process is needed, it is known as ultrasonic bonding. In this technique the bond is formed by pressure and mechanical vibration. In this case as the wire leaves the spool, the tip is pushed against the surface and the vibration removes any existing oxide and allows the metal to deform and flow under pressure at room temperature to create a strong bond. The result is a good bond with little to now oxide formation.

16.6.3. Packaging

Once the die chip is fully wire bonded, it is ready to be encapsulated in a package. Integrated circuit devices can be mounted in a wide variety of packages which have a specialized shape and nature. In this sub-section, we will briefly review three examples of packages shown in Fig. 16.42.

Fig. 16.42(a) shows a round TO-style package which is commonly used for low power transistors. The package utilizes a pie shape header where the silicon chip or die is mounted to the center of the gold plated header. Wires are connected from the die pad to the Kovar lead posts that protrude through the header. A glass-to-metal cap is sealed over the die chip to protect the device.

Another form of packages is the dual line package (DIP) as illustrated in Fig. 16.42(b). The dual line package is considered the least expensive package and the most popular one in industry. The design of the DIP package is such that it eliminates the waste of volume of the TO can, and brings the die closer to the metal leads. One advantage of the dual line package over TO packages is the amount of leads that can pass through the walls. Typically, dual line packages contain four to eighty leads. Leads projecting from the walls of the package rather than at the base can bring out more leads from a package of a given size, and still maintain reasonable space between the leads.

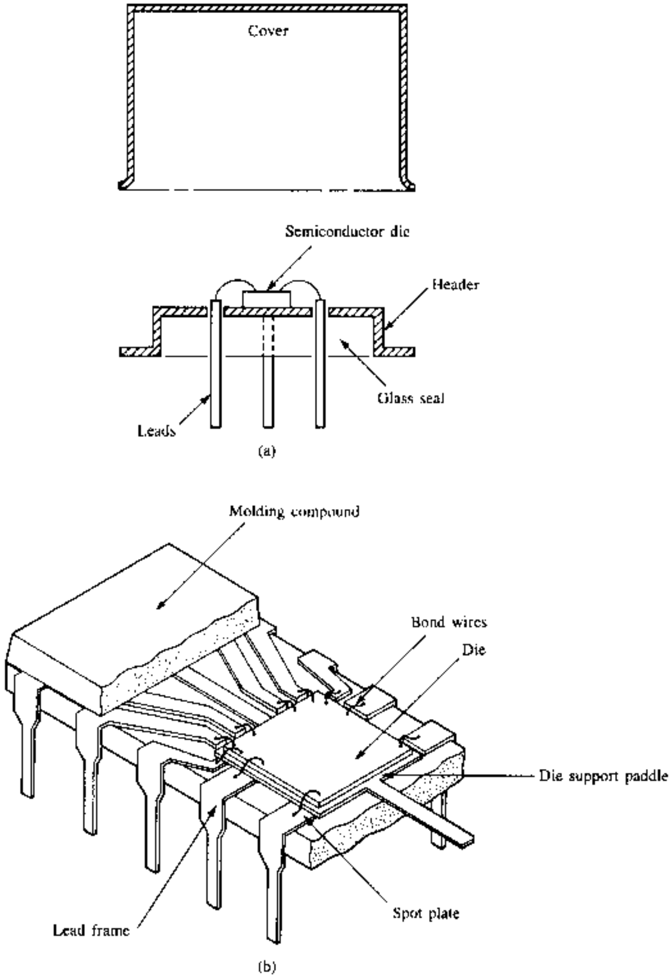


Fig. 16.42. Schematic diagram of (a) a TO-style package and (b) a dual line package. [Reprinted with permission from 19th IEEE International Reliability Physics Symposium Proceedings. Howell, J.R., "Reliability study of plastic encapsulated copper lead frame epoxy die attach packaging system," pp. 104-110. © 1981 IEEE.]

Both "TO"-style packages and dual-in-line packages are packages designed for surface mounting; that is they are both designed to be mounted into prepatterned holes on printed circuit boards (PCBs). While these types of mounts are used for making most systems, they require PCBs to be made before any testing can be performed; this is quite expensive. A method that was developed to permit processing of batch fabricated controls electronics with the desired circuit is known as flip-chip bonding. In this method the control electronics and active system are fabricated separately then sandwiched together to form both the package and interconnection. A schematic diagram of a hybridized focal plane array is shown in Fig. 16.43.

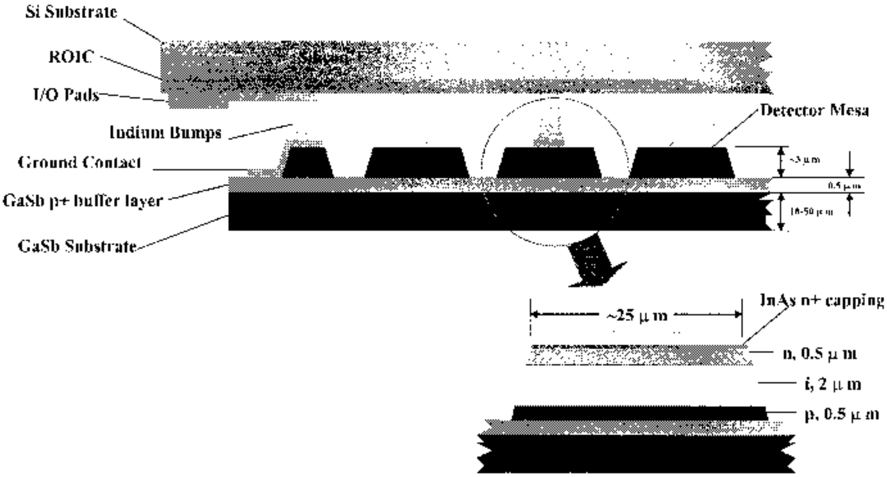


Fig. 16.43. In this system the Si-based control electronics and the p-i-n photodetectors are formed separately. They both have indium solder ball formed on their contacts. They are then aligned, the temperature of the device is then heated, to allow the solder to reflow and create both the electrical contact and the die connection to occur simultaneously.

16.7. Summary

In this Chapter, we have reviewed the important steps involved in the fabrication of a semiconductor device. We described the photolithography and the electron-beam lithography processes. We showed the difference between positive and negative resists, and between the direct patterning and lift-off techniques. We have discussed the various etching process which are commonly used, including wet chemical etching, plasma etching, reactive ion etching, sputter etching and ion milling. We described the metallization process, including the deposition of metal thin films and the formation of metal interconnections. Finally, we presented in broad lines the packaging of semiconductor devices, which involves the dicing of the wafer into chip dies, their wire bonding and packaging into standard packages.

References

Campbell, S.A., *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, New York, 1996.

Fogiel, M., *Microelectronics-Principle, Design Techniques, and Fabrication Processes*, Research and Education Association, New York, 1968.

Howell, J.R., "Reliability study of plastic encapsulated copper lead frame epoxy die attach packaging system," *19th IEEE International Reliability Physics Symposium Proceedings*, pp. 104-110, 1981.

Jaeger, R.C., *Introduction to Microelectronic Fabrication*, 2nd Edition, Prentice-Hall, Upper Saddle River, NJ, 2002.

Morgan, D.V. and Board, K., *An Introduction to Semiconductor Microtechnology*, John Wiley & Sons, Chichester, UK, 1990.

Piner, R.D., Zhu, J., Xu, F., Hong, S., and Mirkin, C.A., "Dip-pen nanolithography," *Science* 283, pp. 661-663, 1999.

Wu, W., Cui, B., Sun, X.Y., Zhang, W., Zhuang, L., Kong, L., and Chou, S.Y., "Large area high density quantized magnetic disks fabricated using nanoimprint lithography," *Journal of Vacuum Science and Technology* 16, pp. 3825-3829, 1998.

Further reading

Choudhury, P.R., *Handbook of microlithography, micromachining & microfabrication*, vol. 1: Microlithography, SPIE Optical Engineering Press, Bellingham, WA, 1997.

Castaño, J.L., Piqueras, J., Gomez, L.J., and Montojo, M.T., "Chemical cleaning of GaSb (1,0,0) surfaces," *Journal of the Electrochemical Society* 136, pp.1480-1484, 1989.

D'Agostino, R., Cramarossa, F., Fracassi, F., Desimoni, E., Sabbatini, L., Zambonin, P.G., and Caporiccio, G., "Polymer film formation in C2F6-H2 discharges," *Thin Solid Films* 143, pp. 163-175, 1986.

Diaz, J.E., *Fabrication of High Power Aluminum-Free 0.8 μm to 1.0 μm InGaP/InGaAsP/GaAs Lasers for Optical Pumping*, PhD dissertation, Northwestern University, 1997.

Elliott, D.J., *Microlithography: Process Technology for IC Fabrication*, McGraw-Hill, New York, 1986.

Elliott, D.J., *Integrated Circuit Fabrication Technology*, 2nd edition, McGraw-Hill, New York, 1989.

Ghandi, S., *VLSI Fabrication Principles*, John Wiley & Sons, New York, 1983.

Hatzakis, M., "High sensitivity resist system for lift-off metallization," U.S. Patent No. 4024293, 1975.

Levinson, H.J., *Principles of Lithography*, SPIE Optical Engineering Press, Bellingham, WA., 2001.

Mackie, S. and Beaumont, S.P., "High sensitivity positive electron resist," *Solid State Technology* 28, pp. 117-122, 1985.

Madou, M.J., *Fundamental of Microfabrication*, CRC Press, Boca Raton, FL, 1997.

Moreau, W. and Ting, C.H., "High sensitivity positive electron resist," US Patent 3934057, 1976.

Plummer, J.D., Deal, M., and Griffin, P.B., *Silicon VLSI Technology: Fundamentals, Practice and Modeling*, Prentice-Hall, Upper Saddle River, NJ, 2000.

Propst, E.K., Vogt, K.W., and Kohl, P.A., "Photoelectrochemical Etching of GaSb," *Journal of the Electrochemical Society* 140, pp. 3631-3635, 1993.

Sheats, J.R. and Smith, B.W., *Microlithography: Science and Technology*, Marcel Decker, New York, 1998.

Williams, R.E., *Gallium Arsenide Processing Technique*, Artech House, Dedham, MA, 1984.

Wong, A.K., *Resolution Enhancement Techniques in Optical Lithography*, SPIE Optical Engineering Press, Bellingham, WA, 2001.

Problems

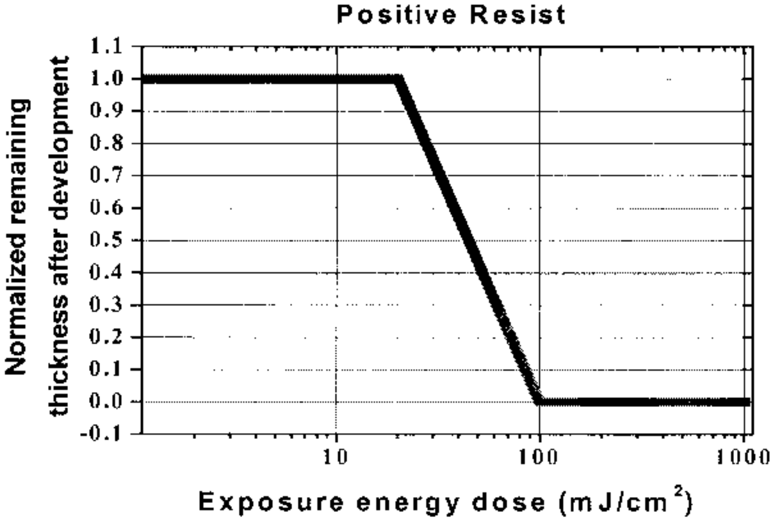
1. Draw a layout, in top view and perspective, where the mask would be opaque to realize an "+" shaped metal line using a positive resist and the lift-off technique.

2. Do the same as Problem 1, but using a negative resist.

3. Design a photolithography and metallization sequence of steps to obtain a Au square and a Ti circle on the surface of a semiconductor wafer. Draw the shape of the mask to be used. At each appropriate step, indicate whether you use the direct patterning or the lift-off technique, you use positive or negative resist.

4. Indicate the advantages and disadvantages of using gold or aluminum for wire bonding. Which metal has a higher electrical conductivity?

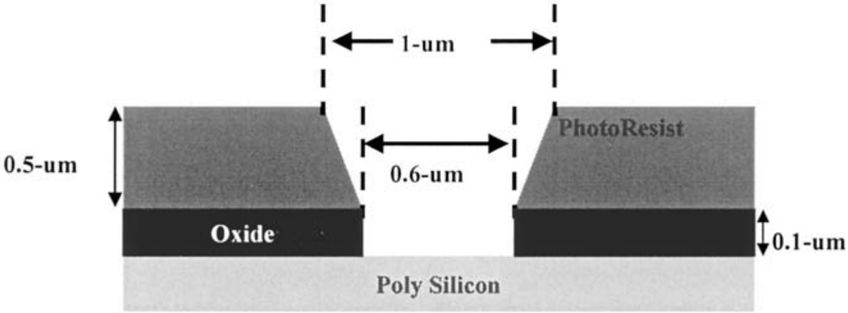
5. For a particular positive resist, the normalized remaining thickness after development versus photoexposure energy density is plotted below. Calculate the contrast (γ) of this resist (γ=1/log(E_T/E_I)).



6. Draw a layout, in side view of a set of laser bars (i.e. a comb structure) if you were using a isotropic etch and if you were using an anisotropic etchant. Which etchant gives a higher aspect ratio?

7. Discuss the major advantages and disadvantages of electron beam lithography.

8. Poly-Si of the following structure is to be etched using a *completely anisotropic* dry-etch process, to remove poly-Si at a rate of 0.1 μm /min. However, this etch process has poor selectivities: selectivity to SiO_2 is 5, selectivity to photoresist is 2.
   (a) Sketch the cross-section after 5 minutes of etching.
   (b) Calculate the angle of the SiO2 sidewalls after 5 minutes of etching.



9. In this Chapter we have discussed several different forms of lithography (i.e. photo, electron beam, x-ray, high energy ion beam). Why do companies like Intel keep developing new lithographic techniques?

10. What is the wavelength regime of each of the lithographic techniques listed in Problem 9? (red-visible light, green-visible light, blue light, UV (E=5eV), e-beam (acceleration voltage of 10 keV), and x-ray).

11. Compare wet and dry etching in terms of its directionality, selectivity, cleanliness, feature size, and controllability.

17. Transistors

17.1. Introduction

Modern semiconductor electronics was revolutionized by the invention of the bipolar transistor at the Bell Telephone Laboratories in 1948. The impact of transistors can be best understood when one realizes that, without them, there would have been no progress in such diverse areas of everyday life as computers, television, telecommunications, the Internet, air travel, space exploration, as well as the tools necessary to study and understand the biological process.

Transistors can be classified in two major categories: they can be either bipolar transistors or field effect transistors. Each is fundamentally different from the other in its operation mechanisms. A bipolar transistor operates through the injection and collection of *minority* carriers utilizing p-n junctions. By contrast, a field effect transistor is a *majority* carrier device and is thus a unipolar device.

In this Chapter, we will first review the general motivation and principles for electrical amplification and switching. We will then describe qualitative and quantitatively the direct-current (DC) operation mechanisms of bipolar junction transistors (BJT), their modes of operation, second order effects, as well as practical applications of BJTs in amplifier configurations. A variation of the bipolar transistor, utilizing a heterojunction, will subsequently be discussed. Next, the DC operating principles of field effect transistors (FET) will be presented along with corresponding second order behaviors of FET-based devices. Once again, practical circuit configurations and applications will be introduced. Finally, application specific transistors will be presented, including single electron as well as high electron mobility transistors.

17.2. Overview of amplification and switching

Transistors are capable of serving as switches and amplifiers, depending upon their configuration. The term "transistor" comes from "transfer resistor" and alludes to a transistor's behavior as a resistor that amplifies signals as they are transferred from the input to the output terminal of the device.

Before trying to understand the amplification and switching mechanisms in a transistor, it is important to comprehend the idea of operating current and voltage of a given device. Let us consider the simple electrical circuit in Fig. 17.1(a), which includes a voltage source (e.g. a battery) V_0, a resistance R and the electrical device under consideration.

The current-voltage characteristic of the device under consideration is shown as the solid line in Fig. 17.1(b), which is the illustration of the mathematical function:

Eq. (17.1) $I_T=f(V_T)$

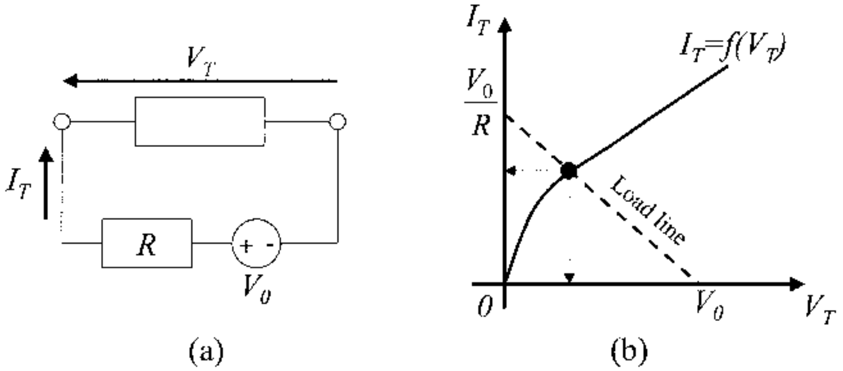


Fig. 17.1. (a) Example of electrical circuit. (b) Illustration of the current-voltage characteristic of the device (solid line) and the load line (dashed line).

In addition, the voltage-loop equation around the circuit shown in Fig. 17.1(a) yields:

Eq. (17.2) $V_0 - RI_T = V_T$

This equation is shown as the dashed line in Fig. 17.1(b). The steady state current I_T and voltage V_T are determined by solving the system formed by these two equations. The solution can be easily visualized graphically and corresponds to the intersection point of the two curves in Fig. 17.1(b).

Let us now consider an electrical device with three terminals or electrical connections, as shown in Fig. 17.2(a). Let us further assume that the current I_T through two of the terminals can be controlled by changes in the current $I_{control}$ or the voltage $V_{control}$ applied at the third terminal as illustrated in Fig. 17.2(b) by the collection of current-voltage characteristic. Eq. (17.2) is still valid and is still represented by the dashed line along which the steady state values of the current I_T and voltage V_T are located. This line is called the circuit load line.

As we can see graphically in Fig. 17.2(b), the significant changes in the current I_T (e.g. ~50 mA) can be achieved by only small changes in the control current $I_{control}$ (e.g. ~0.3 mA). This feature is called amplification, through which a small signal variation, such as that of $I_{control}$, can be amplified into a large signal change such as that of I_T.

Another important feature of this type of electrical circuit is the possibility to turn *on* and *off* the device through changes in $I_{control}$. This is achieved by switching the current I_T between the two extremes on the load lines, from I_T=0 to I_T=V_0/R. This feature is called switching.

A transistor is an example of a three terminal device that exhibits amplification and switching capabilities. These are the basis of all electronic device functions, which makes transistors the basic elements in modern electronics.

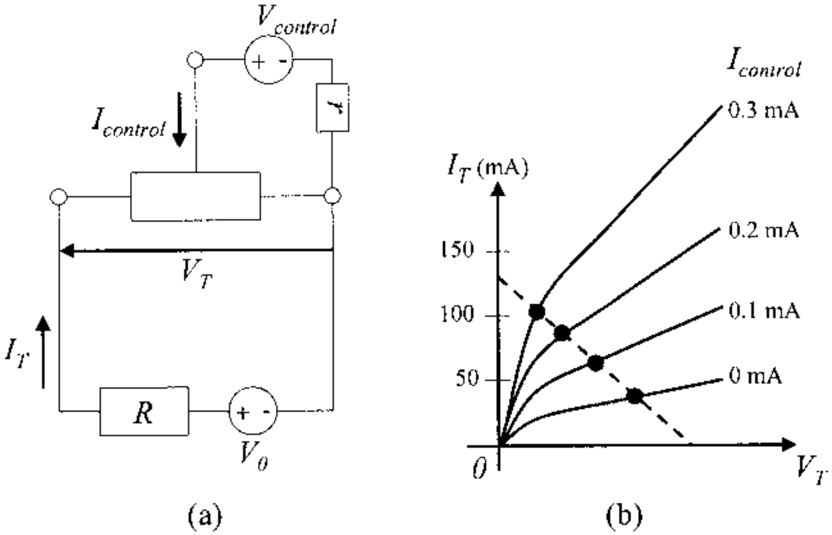


Fig. 17.2. (a) Example of electrical circuit utilizing a three terminal device. (b) Illustration of the load line (dashed line) and the current-voltage characteristic of the device (solid lines) as a function of the control current. The intersection of the solid and dashed lines gives the steady state values of current and voltage across the device.

17.3. Bipolar junction transistors

Bipolar junction transistors consist of two back-to-back p-n junctions which share a common terminal. Such a transistor can be a p-n-p or an n-p-n transistor. In this section, we shall use the p-n-p configuration for most illustrations and analysis. The main advantage of the p-n-p for discussing transistor action is that hole flow and current are in the same direction. This makes the various mechanisms of charge transport somewhat easier to visualize in a preliminary explanation. Once these basic ideas are established for the p-n-p device, it is simple to relate them to the more widely used n-p-n transistors. The corresponding current and bias polarities need only to be switched for the n-p-n case. The common schematic diagrams for n-p-n and p-n-p bipolar junction transistors are shown in Fig. 17.3. The direction of the arrow on the emitter leg indicates the direction of current flow, during the forward active mode of operation, and the p-n transition; thus, it can be used to easily identify the transistor type in circuit diagrams. We will start by discussing the BJT from a qualitative viewpoint and gain physical understanding on how the device operates.

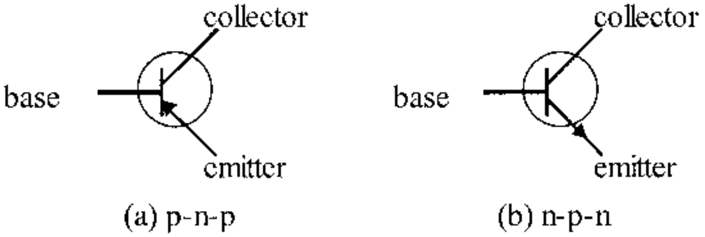


Fig. 17.3. Schematic diagrams for (a) p-n-p and (b) n-p-n bipolar junction transistors.

17.3.1. Principles of operation for bipolar junction transistors

We will now qualitatively describe how the current control can be achieved using a BJT. Let us consider the p-n-p transistor shown in Fig. 17.4 with the left p-n junction under forward bias and the right one under reverse bias, otherwise known as the forward active mode of operation.

According to the analysis of carrier transport in a p-n junction done in Chapter 9 (section 9.3.1), the left p-n junction is biased such that holes are injected from its heavily doped *p*-type region into its *n*-type region (current i_E) where they become minority carriers and will diffuse to reach the other side of the *n*-type region. The left *p*-type region electrode is therefore called the emitter. The width, W, of the base region is typically thinner than the minority carrier diffusion length, to maximize the number of minority carriers that diffuse across it and to minimize base recombination.

As the n-p junction on the right is under reverse bias, the electrical current flowing through it (I_C) is mostly determined by the drift current across the depletion region and its magnitude is determined by the concentration of minority carriers present at the boundaries, $x=W$ and $x=x_c$, of this depletion region.

Therefore, the holes which were injected by the p-n junction (left) and which succeeded to reach the edge of the base depletion, $x=W$, region for the n-p junction (right) via diffusion will determine the magnitude of the electrical current through the n-p junction. These holes are, in a way, collected by the right *p*-type region electrode, which is therefore called the collector. The I_C current corresponds to the saturation current given in Eq. (9.53) and is independent of the applied reverse bias voltage, neglecting any leakage.

The exact amount of holes that diffuse through the *n*-type region is affected by a few parameters which are intrinsic to the *n*-type region (e.g. diffusion lengths), as well as by other parameters which are extrinsic to it such as the current flowing through the base (I_B) which acts as the control current of Fig. 17.2. The mechanisms of this current control will now be illustrated in further details in the case of amplification in a bipolar junction transistor.

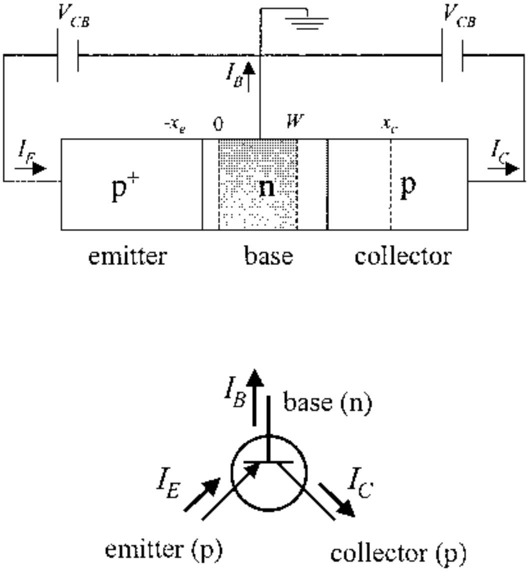


Fig. 17.4. A p-n-p bipolar junction transistor with its emitter-base junction under forward bias and the base-collector under reverse bias. The conventions for the signs of the electrical currents I_E, I_B and I_C are shown: the electrical current will be positive if it actually flows in the direction of the arrow.

17.3.2. Amplification process using BJTs

In order to understand how the amplification process is carried out, we will first develop a qualitative picture of the current mechanisms in a BJT transistor. Then, we will take a more analytical approach to expressing the current mechanisms, as well as the amplification and transport factors relevant to BJT devices.

Let us begin by considering a p-n-p transistor with a heavily doped p$^+$ region for the emitter. The *p*-type emitter is taken to be highly doped in order to be a more efficient hole emitter. A schematic diagram of such a structure is shown in Fig. 17.5.

For the holes that are injected from the emitter electrode into the base, a portion will undergo recombination with the electrons present in the base region (1 in Fig. 17.5). The probability of recombination is proportional to the density of electrons available and the density of holes that are injected.

When no external electrons are injected into the base region, the recombining holes will lead to the apparition of fixed positive ions. The density of electrons will then decrease, thus causing less and less electron-hole recombination, and the fixed positive ions will build up. The resulting electric field will reduce the injection of the holes into the base region, and hence, the portion of the holes injected from the emitter and reaching the collector will decrease as well.

However, when electrons are injected into the base region through the base current I_B (4 in Fig. 17.5), they will recombine with the holes and reduce the build up of positive charges. The base barrier will therefore decrease and a larger amount of holes will reach the collector (2 in Fig. 17.5).

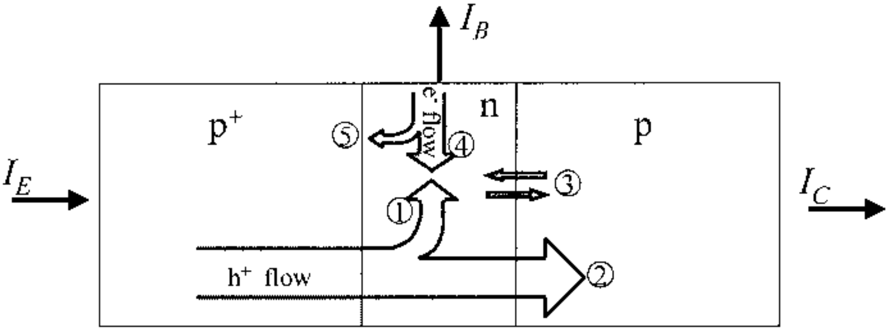


Fig. 17.5. Schematic diagram showing the flows of holes and electrons within a p^+-n-p bipolar junction transistor. The conventions for the signs of the electrical currents i_E, i_B and i_C are shown: the electrical current will be positive if it actually flows in the direction of the arrow.

This phenomenon provides us with a method to control the current flow from the emitter to the collector via the amount of electrons injected into the base. Since only a small portion of holes will be recombined with the injected electrons, we can use a small injection current (I_B) to control a much bigger current (I_C). And the current gain can be very high if the recombination rate is low, which can be done by engineering the base region adequately.

17.3.3. Electrical charge distribution and transport in BJTs

We will now try to quantitatively examine the operation of bipolar junction transistors. We will consider a p-n-p transistor shown Fig. 17.6. The case of forward active operation will be considered. The objective will be to determine the minority carrier distributions and the terminal currents. In order to simplify the calculations, a few assumptions are made:

(1) Holes diffuse from the emitter to the collector and drift is negligible in the base region.

(2) The emitter current is contributed entirely by holes, i.e. the emitter injection efficiency γ is 1 and I_{En}=0.
(3) The collector saturation current is negligible, i.e. component 3 in Fig. 17.5.
(4) The active part of the base and the two junctions are of uniform cross-sectional area A and the current flow in the base is essentially one-dimensional from the emitter to the collector.
(5) All currents and voltages are considered at steady state.

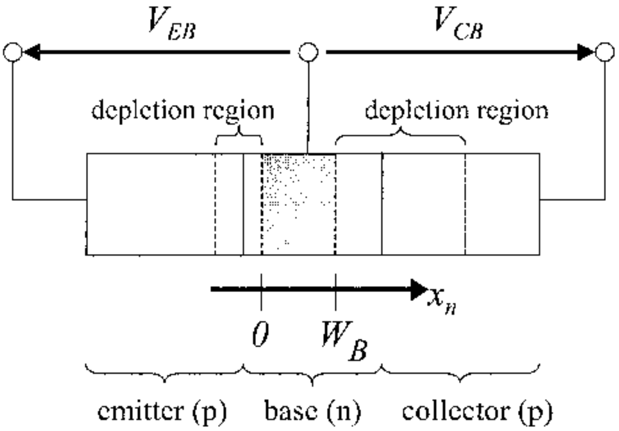


Fig. 17.6. Schematic diagram of a p-n-p bipolar junction transistor showing the voltages and convention for the position variable x_n.

The excess hole concentration on the collector side of the base Δp_C and that on the emitter side of the base Δp_E are given by Eq. (9.42):

Eq. (17.3)
$$\begin{cases} \Delta p_E = p_n\left(e^{\frac{qV_{EB}}{k_bT}} - 1\right) \\ \Delta p_C = p_n\left(e^{\frac{qV_{CB}}{k_bT}} - 1\right) \end{cases}$$

where p_n is the equilibrium hole concentration in the n-type base region. If the emitter junction is strongly forward biased (V_{EB}>>k_bT/q) and the collector junction is strongly reverse biased (V_{CB}<<0), these expressions can be simplified and become:

Eq. (17.4)
$$\begin{cases} \Delta p_E \approx p_n e^{\frac{qV_{EB}}{k_bT}} \\ \Delta p_C \approx -p_n \end{cases}$$

The diffusion equation is given by:

Eq. (17.5) $$\frac{d^2\delta p(x_n)}{dx_n^2}=\frac{\delta p(x_n)}{L_p^2}$$

where $\delta p(x_n)$ is the concentration of excess holes at x_n, and L_p is the hole diffusion length in the n-type base region. The general solution of this equation is:

Eq. (17.6) $$\delta p(x_n)=C_1e^{\frac{x_n}{L_p}}+C_2e^{-\frac{x_n}{L_p}}$$

where C_1 and C_2 are integration constants. Expressing the boundary conditions, we get:

Eq. (17.7) $$\begin{cases}\delta p(x_n=0)=C_1+C_2=\Delta p_E\\[2ex]\delta p(x_n=W_b)=C_1e^{\frac{W_b}{L_p}}+C_2e^{-\frac{W_b}{L_p}}=\Delta p_C\end{cases}$$

where W_b is the width of the base region. Solving for C_1 and C_2 we get:

Eq. (17.8) $$\begin{cases}C_1=\dfrac{\Delta p_C-\Delta p_E e^{-\frac{W_b}{L_p}}}{e^{\frac{W_b}{L_p}}-e^{-\frac{W_b}{L_p}}}\\[4ex]C_2=\dfrac{\Delta p_E e^{\frac{W_b}{L_p}}-\Delta p_C}{e^{\frac{W_b}{L_p}}-e^{-\frac{W_b}{L_p}}}\end{cases}$$

If we assume that the collector junction is strongly reverse biased and the equilibrium hole concentration p_n is negligible compared with the injected concentration Δp_E, the concentration of excess holes at x_n within the base region becomes:

Eq. (17.9) $$\delta p(x_n) = \Delta p_E \frac{e^{\frac{W_b}{L_p}} e^{-\frac{x_n}{L_p}} - e^{-\frac{W_b}{L_p}} e^{\frac{x_n}{L_p}}}{e^{\frac{W_b}{L_p}} - e^{-\frac{W_b}{L_p}}} \text{ (for } \Delta p_C \approx 0)$$

Having solved for the excess hole distribution in the base region, we can now evaluate the emitter and collector currents from the gradient of the hole concentration at each depletion region edge:

Eq. (17.10) $$I_p(x_n) = -qAD_p \frac{d\delta p(x_n)}{dx_n}$$

This expression evaluated at x_n=0 gives the hole component of the emitter current (i.e. I_{Ep}), and evaluated at x_n=W_b gives the collector current (i.e. I_C):

Eq. (17.11) $$\begin{cases} I_{Ep} = I_p(x_n = 0) = qA\dfrac{D_p}{L_p}(C_2 - C_1) \\ \qquad = qA\dfrac{D_p}{L_p}\left(\Delta p_E \operatorname{ctnh}\dfrac{W_b}{L_p} - \Delta p_C \operatorname{csch}\dfrac{W_b}{L_p}\right) \\ I_C = I_p(x_n = W_b) = qA\dfrac{D_p}{L_p}\left(C_2 e^{-\frac{W_b}{L_p}} - C_1 e^{\frac{W_b}{L_p}}\right) \\ \qquad = qA\dfrac{D_p}{L_p}\left(\Delta p_E \operatorname{csch}\dfrac{W_b}{L_p} - \Delta p_C \operatorname{ctnh}\dfrac{W_b}{L_p}\right) \end{cases}$$

Then I_B is obtained by current summation:

Eq. (17.12) $$I_B = I_E - I_C = qA\frac{D_p}{L_p}\left[(\Delta p_E + \Delta p_C)\tanh\frac{W_b}{2L_p}\right]$$

17.3.4. Current gain

We can now define a few parameters which characterize the amplification mechanism. For simplicity, we will neglect the saturation current at the collector (3 in Fig. 17.5) and recombination in the depletion regions.

We first start by expressing the collector current I_C as a function of the emitter current I_E. The total emitter current I_E has two separate components: a net hole and a net electron diffusion currents (5 in Fig. 17.5), or i_{Ep} and i_{En} respectively:

Eq. (17.13) $$I_E = I_{Ep} + I_{En}$$

An emitter injection efficiency γ can thus be defined as:

Eq. (17.14) $$\gamma = \frac{I_{Ep}}{I_{Ep} + I_{En}}$$

The emitter injection efficiency can be considered as the portion of total emitter current that is due solely to minority carriers being injected into the base (see Fig. 17.7). γ can be closer to unity when the p-type emitter region is highly doped (p^+). It can be shown that the emitter injection efficiency of a p-n-p transistor can also be written in terms of the emitter (L_p^n, p_p) and base material properties (L_n^p, n_n):

Eq. (17.15) $$\gamma = \left[1 + \frac{L_p^n n_n \mu_n^p}{L_n^p p_p \mu_p^n} \tanh \frac{W_b}{L_p^n}\right]^{-1} \approx \left[1 + \frac{W_b n_n \mu_n^p}{L_n^p p_p \mu_p^n}\right]^{-1}$$

In this equation we use superscripts to indicate which side of the emitter-base junction is referred to.

The ratio of the collector hole current I_{Cp} to the hole component of the emitter current I_{Ep}, called the base transport factor, is denoted α_T and is given by:

Eq. (17.16) $$\alpha_T = \frac{I_{Cp}}{I_{Ep}} = \frac{\operatorname{csch} \frac{W_b}{L_p}}{\operatorname{ctnh} \frac{W_b}{L_p}} = \operatorname{sech} \frac{W_b}{L_p}$$

This factor reflects the amount of recombination occurring in the base region. Finally, the ratio of the collector current to the total emitter current is called the current transfer ratio and is denoted α_0:

Eq. (17.17) $$\alpha_0 = \alpha_T \gamma = \left[\cosh \frac{W_b}{L_p^n} + \frac{L_p^n n_n \mu_n^p}{L_n^p p_p \mu_p^n} \sinh \frac{W_b}{L_p^n} \right]^{-1}$$

or expressed otherwise as:

Eq. (17.18) $$\frac{I_C}{I_E} = \frac{\alpha_T I_{Ep}}{I_{En} + I_{Ep}} = \alpha_T \gamma \equiv \alpha_0$$

An efficient transistor is one such that $\alpha_T \approx 1$ and $\gamma \approx 1$, and therefore the current transfer ratio, α_0 is close to unity too.

Now, let us consider the relationship between the collector current I_C and the base current I_B. By taking into account all the currents going into and out of the base region, we can express the base current as:

Eq. (17.19) $$I_B = I_{En} + I_{Ep} - I_C = I_{En} + (1 - \alpha_T) I_{Ep}$$

Using the above relation and the fact that $I_C = \alpha_T I_{Ep}$, we successively get:

Eq. (17.20) $$\begin{aligned} \frac{I_C}{I_B} &= \frac{\alpha_T I_{Ep}}{I_{En} + (1 - \alpha_T) I_{Ep}} \\ &= \frac{\alpha_T [I_{Ep} / (I_{En} + I_{Ep})]}{1 - \alpha_T [I_{Ep} / (I_{En} + I_{Ep})]} \\ &= \frac{\alpha_T \gamma}{1 - \alpha_T \gamma} = \frac{\alpha_0}{1 - \alpha_0} \\ &= \beta \end{aligned}$$

where β is called the base-to-collector current amplification factor. This factor is also commonly seen as h_{FE} in datasheets and the literature. For an efficient transistor, as α_0 is close to unity, the factor β can be large. This means that the collector current is large compared to the base current.

We mentioned earlier that the amplification can be large if the base region is engineered correctly. This can be illustrated by expressing the amplification factor β in terms of two characteristic times: τ_p and τ_t, which are the average hole lifetime in the base and the average transit time of holes from emitter to collector, respectively. To do so, we will also assume a unity emitter injection efficiency (γ=1) and a negligible collector saturation current.

Under these conditions, the average hole recombination lifetime τ_p is also the average time that an electron injected from the base contact spends within the base region. Furthermore, the average time that a hole stays within the base region, W_b, is the transit time τ_t given by:

$$\tau_t = \frac{W_b^{\,2}}{2D_h}$$

This transit time can be made much shorter in comparison to the recombination lifetime τ_p by reducing the dimension of the base region, the origin of the Early effect, discussed later. This means in particular that an injected electron can "outlive" an injected hole in the base region. Thus, in order to ensure the overall charge neutrality of the base region, more holes need to be injected from the emitter into the base. In other words, for each injected electron, there will be τ_p/τ_t holes which can traverse the base region before recombination occurs. This in particular means that:

Eq. (17.21) $$\frac{\tau_p}{\tau_t} \approx \frac{I_{Ep}}{I_B}$$

We can therefore qualitatively understand how the amplification process takes place. Since I_{Ep}-I_B=I_C when γ=1, we get:

Eq. (17.22) $$I_B(\frac{\tau_p}{\tau_t} - 1) = I_C$$

Using Eq. (17.20), we get:

Eq. (17.23) $$\frac{I_C}{I_B} = \frac{\tau_p}{\tau_t} - 1 = \beta$$

Since usually $\frac{\tau_p}{\tau_t}$ is generally large, we get:

Eq. (17.24) $\beta = \frac{I_C}{I_B} \approx \frac{\tau_P}{\tau_t}$

Therefore, as the base current I_B can be controlled independently as shown in Fig. 17.2 and is mainly determined by the external circuit parameters, the collector current I_C will be the base current I_B multiplied by the current amplification factor β, which represents current gain.

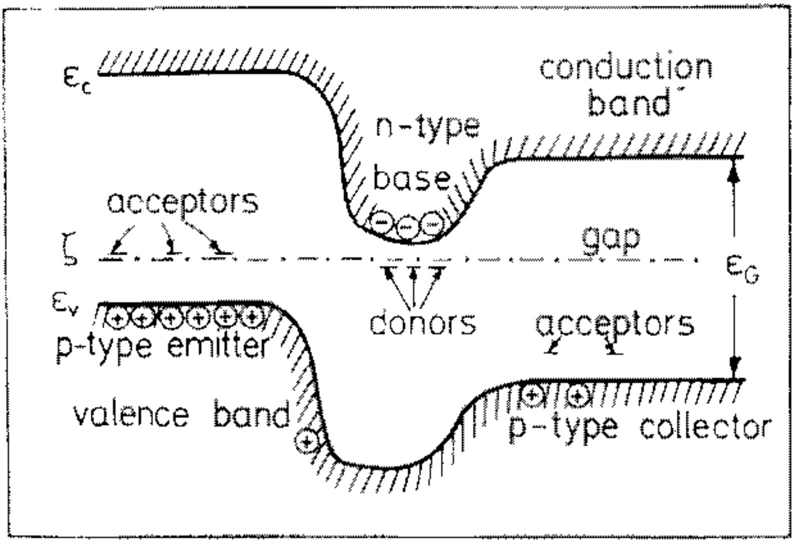


Fig. 17.7. Energy band edges in a p-n-p type transistor at thermal equilibrium. [Semiconductor Physics: An Introduction, 1997, p. 144, Seeger, K., Fig. 5.12. © Springer-Verlag Berlin Heidelberg 1973, 1982, 1985, 1989, 1991 and 1997. With kind permission of Springer Science and Business Media.]

17.3.5. Typical BJT configurations

Four possible modes of operation exist for BJT biasing. The forward active mode is the most commonly used operational mode when using a BJT for amplification purposes (see Fig. 17.7 for the equilibrium state). The three remaining modes are saturation, cutoff, and reverse active modes. The junction biasing is configured as shown in Table 17.1.

Mode of operation	Base-emitter bias	Base-collector bias
Active	Forward	Reverse
Cutoff	Reverse	Reverse
Saturation	Forward	Forward
Reverse Active	Reverse	Forward

Table 17.1. The four modes of operation for a BJT and the corresponding junction biasing for each mode.

The active, cutoff, and saturation modes will be explained in the following sections but for now we will shortly discuss the reverse active mode. The reverse active mode is analogous to forward active in terms the equations applicable to it, except the emitter is replaced with the collector and vice-versa. This mode is not often used due to its poor efficiency arising from the doping configuration and the corresponding depletion widths dimensions.

Fig. 17.8 shows the three common BJT amplifier configurations known as common base, common emitter, and common collector. These configurations can be easily identified by determining which terminal is connected to ground or circuit "common."

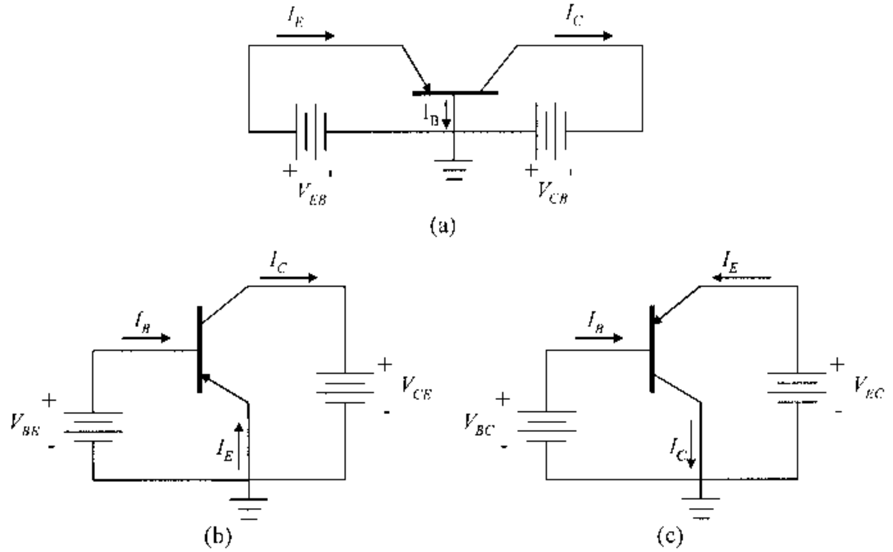


Fig. 17.8. Typical BJT amplifier configurations: (a) common base, (b) common emitter, and (c) common collector

Consider the case of a common base (CB) configuration. In forward active mode the emitter-base junction is forward biased and the collector-base junction is reverse bias. A typical family of curves for such a configuration is shown in Fig. 17.9.

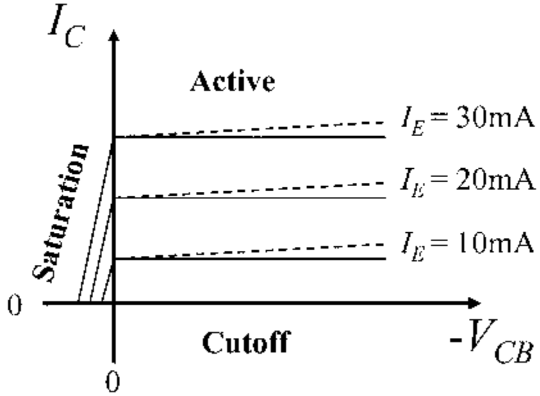


Fig. 17.9. A family of curves demonstrating the dependence of collector current, I_C, on the collector-base voltage, V_{CB}. Curves are shown for a range of emitter currents, I_E. The modes of operation are labeled and the dashed lines indicate the consequence of the Early effect.

In Fig. 17.9 the Early effect is shown with dashed lines. This effect originates from the increasing base-collector depletion width with increasing reverse bias of V_{CB}. This reduces the width of the base region resulting in an increased charge gradient across the base as well as decrease in the recombination probability in the base. The result of the former is increased injection of minority carriers from the emitter and the latter enhances the base transport factor, α_T. If one extrapolates in the positive V_{CB} direction the dashed output curves of Fig. 17.9 they will converge at the Early voltage, V_A. If the base width is much larger than the depletion region extending into the base the Early voltage can be expressed as:

Eq. (17.25) $$V_A \cong \frac{qN_b W_b^2}{\varepsilon_s}$$

where N_b is the base doping.

In the active region of operation the collector-base junction is reverse biased and the emitter-base junction is forward biased. If an emitter current, I_E is allowed to flow then $\sim\alpha I_E$ flows from the collector. In this region the Early effect should be considered but its overall effect is not very significant. If I_E goes to 0 then the collector current is equal to the reverse bias saturation current, I_{CO}, a few nano- to micro-amperes depending upon the material comprising the BJT.

The saturation region refers to the case where both junctions are forward biased. The forward bias behavior of the collector junction dictates the strong dependence of I_C on small changes in V_{CB}.

In the cutoff region of operation both junctions are reverse biased and the collector current is negligible.

The common emitter (CE) configuration is encountered much more often in electronic circuits than the common base configuration and will be considered next. One advantage of the CE configuration, compared to the CB, is that the input current I_B can be much smaller than the output current I_C by nearly β. In the active region it can be shown that:

Eq. (17.26) $I_C = \beta I_B$

when the collector leakage current is taken to be negligible and the Early effect is not considered. The active region is the normal region of operation for a CE amplifier. The Early effect has much more dramatic consequences in the case of the CE configuration. A very small change of a fraction of one percent in α due to a decrease in W_b can increase β by tens of percents or more. Proof of this concept is left for the exercises.

The cutoff region is defined as the case when the collector current is equal to the saturation current and the reverse-biased emitter junction's current is zero. If V_{CE} drops below V_{BE} then the collector junction is forward biased and the device is considered to be saturated.

17.3.6. Deviations from the ideal BJT case

As with just about any electrical device, the practical behavior of BJTs deviates, to some respect, from the ideal models presented in the previous sub-section. In this sub-section we will discuss some divergences of BJT behavior from the ideal case. We have already discussed the Early effect so that will not be addressed here but some explanation of base spreading, current crowding, depletion region recombination, and breakdown mechanisms will be offered.

In a p-n diode, the occurrence of high injection conditions increases the carrier density of injected carriers to levels similar to that of the collector doping concentration. When a BJT is placed into high injection conditions, by forward biasing the emitter-base junction a reduction in the current gain is realized according to the following relationship with V_{BE}:

$$\beta = \beta_0 \exp\left(-\frac{V_{BE}}{2V_t}\right)$$

Punchthrough is the case when the low-doped base width is reduced to zero and a short circuit between the p-type (or n-type) collector and emitter is created upon sufficient V_{CB}. This condition can also arise from very narrow base widths. The result of punchthrough is a large current through the current and emitter.

Breakdown in BJT devices usually originates from avalanche multiplication. Collector doping levels are typically not high enough to cause direct tunneling, or Zener breakdown. The avalanche process in BJTs is nearly identical to that in a p-n junction. In the CE configuration avalanche breakdown is caused by impact ionization. The ionized carriers appear as an increase in the base current which causes even more collector current to flow in a positive feedback fashion. One can show that the breakdown voltage in the CE configuration is:

$$V_{B,CE} = V_{B,CB}(1-\alpha_0)^{1/n}$$

where n is a constant and $V_{B,CB}$ is the common-base breakdown voltage. The common-base breakdown voltage is commonly determined by the open-emitter reverse breakdown voltage of the base-collector p-n junction. This breakdown voltage is similar to that for a typical p-n junction and, again, is generally due to avalanche breakdown.

17.4. Heterojunction bipolar transistors

In a homojunction BJT, the emitter injection efficiency is limited by the fact that carriers can flow from the base into the emitter region, over the emitter junction barrier, which is reduced by the forward bias. It is necessary to use lightly doped base and heavily doped p$^+$ emitter for the optimum injection of holes. But this will result in higher base resistance. Degenerate doping can lead to a slight decrease of E_g in the emitter, which will decrease the emitter injection efficiency. For high frequency applications, a heavily doped base and a lightly doped emitter are desirable. There are better ways to accomplish the design instead of doping only, i.e. to use heterojunctions instead of homojunctions. We then talk about a heterojunction bipolar transistor or HBT.

For example, if we use a wider bandgap material for the emitter than the base, then it is possible that for an n-p-n transistor, the barrier for electron injection is smaller than the hole barrier. Since carrier injection rate varies exponentially with the barrier height, even a small difference in these two barriers can make a very large difference in the transport of electrons and holes across the emitter junction. Neglecting differences in carrier mobilities and other effects, we can approximate the dependence of carrier injection across the emitter as:

$$\text{Eq. (17.27)} \quad \frac{I_n}{I_p} \propto \frac{N_D^E}{N_A^B} e^{\frac{\Delta E_g}{k_b T}}$$

A relatively small value of ΔE_g will have significant effects to the current ratio. This allows us to choose the doping terms for lower base resistance and emitter junction capacitance. In particular, we can choose a heavily doped base to reduce the base resistance and a lightly doped emitter to reduce junction capacitance. However there will usually be spike and notch at heterojunction interface. This can be eliminated by graded interface.

In the following sub-sections, we will describe two of the most widely used heterojunction bipolar transistors: AlGaAs/GaAs and GaInP/GaAs HBTs.

17.4.1. AlGaAs/GaAs HBT

Thanks to the excellent lattice-match between $Al_xGa_{l-x}As$ and GaAs over the entire compositional range, the A1GaAs/GaAs system has been the most widely used system for heterojunction bipolar transistors. Fig. 17.10 shows the cross-section structure of a typical AlGaAs/GaAs HBT.

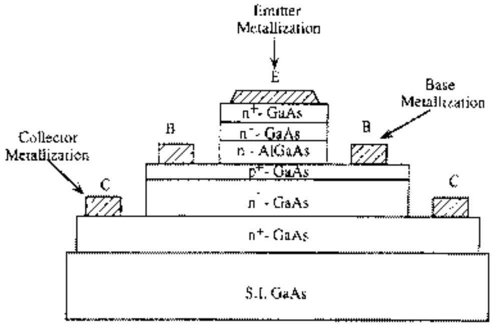


Fig. 17.10. Cross-section structure of an AlGaAs/GaAs heterojunction bipolar transistor. [Copyright © 1995 From The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

In order to avoid DX (unknown defect) center problems, the Al mole fraction x in $Al_xGa_{1-x}As$ is usually kept around 0.25 which results in a conduction-band discontinuity of 0.2 eV and a valence-band discontinuity of 0.1 eV. Due to the large conduction-band discontinuity, the emitter-base junction is usually computationally graded.

Device isolation is performed through deep ion implantation to make the layers outside the device semi-insulating or by using a mesa structure. By means of composition-selective etches, vias are etched to the base and collector layers to make the corresponding contacts.

In order to obtain good device performance, the contact resistance of the ohmic contacts should be minimized. One of the ways that has been used to reduce the emitter contact resistance is to use lattice-mismatched InGaAs cap layers grown on the emitter. Due to the small metal-semiconductor barrier height good ohmic contacts can be achieved. Commonly used contacts are AuGe/Ni and Ge/Au/Cr. Minimization of the parasitic resistance is also very important in obtaining a good device performance. The base-emitter separation should be a few tenths of a micron. This can be accomplished by using self- aligned techniques ([Nagata *et al.* 1987] [Hayama *et al.* 1987] [Chang *et al.* 1987]). By using a shallow proton implant into the collector region under the base contacts, extrinsic collector doping and therefore the base-collector capacitance can be reduced [Ginoudi *et al.* 1992].

It is important that the base-emitter p-n junction coincides with the heterojunction between AlGaAs and GaAs. Therefore good doping profile and material composition control is required in the growth of the epilayers. Most of the AlGaAs/GaAs HBT research has been done on MBE-grown devices. Since the early 1980s the performance of MOCVD-grown AlGaAs/GaAs HBTs has increased significantly. A f_{max} of 94 GHz and an f_t of 45 GHz have been obtained by Enquist and Hutchby [1989] using a self-aligned structure. One of the difficulties in HBT fabrication is the diffusion of impurities from the heavily doped GaAs base into the AlGaAs emitter at high temperatures during or subsequent to growth. This causes the p-n junction to move into the AlGaAs layer and the current gain of the device is reduced due to the reduction in the barrier to hole injection. This problem can be avoided by introducing a thin undated GaAs spacer layer between the base and emitter or by reduction of the growth temperature before the AlGaAs layer is grown. Common p-type dopants in MOCVD are magnesium, zinc and carbon. Mg doping shows abnormal memory effects, which requires growth interruptions in order to obtain an abrupt doping profile ([Kuech *et al.* 1988] [Landgren *et al.* 1988]). Zn has a large diffusion coefficient and carbon doping needs a low growth temperature both of which are incompatible with the growth of high-quality AlGaAs which requires high temperatures. However, very high base doping levels are possible with carbon doping due to the very low diffusion coefficient of carbon [Ashizawa *et al.* 1991]. Using carbon doping in the base (p=4×10^{19} cm^{-3}), Twynam *et al.* [1991] have reported MOCVD-grown AlGaAs/GaAs microwave HBTs with an f_t of 42 GHz, f_{max} of 117 GHz and a current gain of 50.

17.4.2. GaInP/GaAs HBT

The $Ga_{0.51}In_{0.49}P$/GaAs system has some major advantages over AlGaAs/GaAs. For n-p-n heterojunction bipolar transistors, the GaInP/GaAs system has an additional advantage when compared with the widely used AlGaAs/GaAs structure. The valence-band discontinuity in the $Ga_{0.51}In_{0.49}P$/GaAs system is about 0.28 eV and conduction-band discontinuity is 0.2 eV [Biswas *et al.* 1990]. A large valence-band discontinuity is an exciting property for n-p-n HBTs. In the AlGaAs/GaAs system, the same amount of valence-band discontinuity requires that the Al mole fraction be about 0.6, in which case there would be a very large conduction-band spike at the emitter-base junction together with an indirect-gap emitter, neither being acceptable. In the AlGaAs/GaAs system, about 60 per cent of the energy gap difference occurs in the conduction band and the emitter-base junction of the device is usually graded to eliminate the conduction-band spike which decreases the emitter injection efficiency and increases the emitter switch-on voltage. However, theoretical investigations [Das and Lundstrom 1988] have shown that grading of the emitter-base junction increases the recombination in the emitter-base junction and therefore the current gain may not be increased considerably by junction grading. Because of the relatively small conduction-band discontinuity and large valence-band discontinuity of $Ga_{0.51}In_{0.49}P$/GaAs, it can be estimated that the current gain of n-p-n HBTs based on this material system will be significantly higher than that of AlGaAs/GaAs HBTs. Modry and Kroemer [1985] have reported a GaInP/GaAs HBT grown by MBE. The current gain was low at small current densities suggesting a high recombination rate at the emitter-base junction due to a large number of defects at the heterojunction interface. A maximum current gain of 30 was obtained at 3000 A.cm^{-2}. Later, MOCVD and chemical beam epitaxy grown GaInP/GaAs HBTs with better performances were reported ([Kobayashi *et al.* 1989] [Razeghi *et al.* 1990] [Alexandre *et al.* 1990] [Bachem *et al.* 1992]). In addition, Razeghi *et al.* [1990] reported a current gain of 400 for a low-pressure MOCVD-grown GaInP/GaAs HBT. Three different HBT structures were grown: (i) conventional, (ii) double heterojunction, (iii) pseudo-graded base. The details of collector, base and emitter thicknesses, carrier concentrations and a typical x-ray diffraction pattern for the structures around the (400) reflection peak are shown in Fig. 17.11.

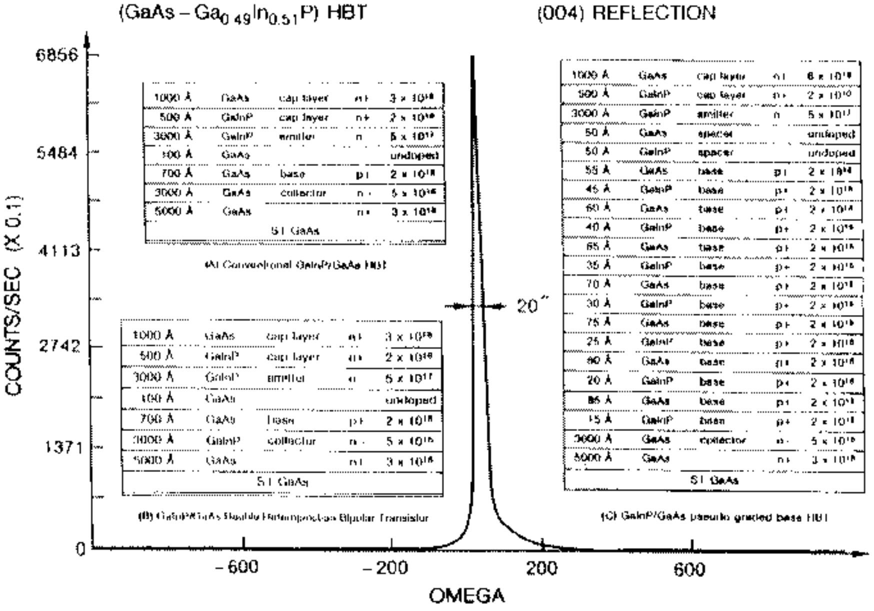


Fig. 17.11. X-ray diffraction spectrum of a GaInP/GaAs heterojunction bipolar transistor. (A) conventional HBT, (B) double heterojunction HBT, and (C) pseudo-graded base HBT, with the layer thicknesses and carrier concentrations of the collector, base and emitter. [Copyright © 1995 From The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

The diffraction peak is very intense and has a full width at half maximum of 20 seconds, demonstrating that GaInP is perfectly lattice-matched to GaAs and the pseudo-graded base has excellent crystallographic properties, which is necessary to allow optimal transport properties of the injected minority carriers. Fig. 17.12 shows the doping profile demonstrating an abrupt and perfectly controlled transition from emitter to base and from base to the collector.

The device structure was a conventional mesa type. NH_4:H_2O_2:H_2O (10:4:500) and HCl:H_3PO_4 (1:1) were used to etch GaAs and GaInP respectively. The emitter and collector contacts were defined by depositing and annealing Ge/AuNi/Au. The base contact was defined by the deposition and annealing of Zn/Au. Fig. 17.13 shows the emitter-grounded current-voltage characteristic of the device which exhibited a current gain of 400 at 20 mA.

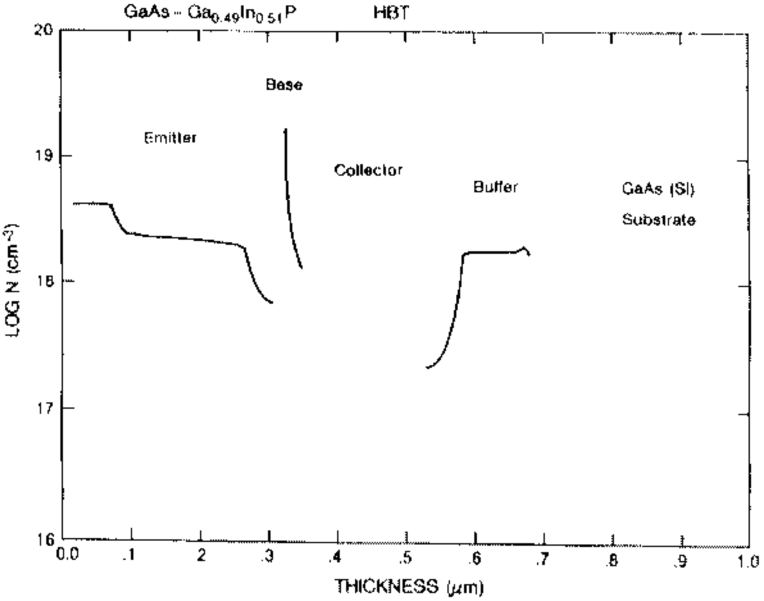


Fig. 17.12. Example of doping profile of a GaInP/GaAs heterojunction bipolar transistor. [Copyright © 1995 From The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

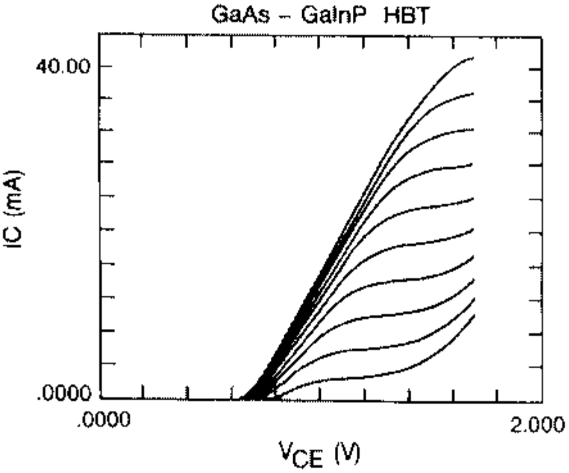


Fig. 17.13. Emitter-grounded current-voltage characteristic of the conventional GaInP/GaAs heterojunction bipolar transistor shown in (A) of Fig. 17.11. [Copyright © 1995 From The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

17.5. Field effect transistors

A field effect transistor or FET is a three terminal device in which the current flow between two terminals can be controlled by the third terminal. However, unlike a bipolar junction transistor, the control is through the voltage, not current, of the third terminal. There are several types of FETs depending on the junction of the controlling terminal or "gate". The first type is a junction FET or JFET, where the gate junction is a simple p-n junction. If this junction is replaced with a metal-semiconductor Schottky contact, the device is called a metal-semiconductor FET or MESFET. Also, if an insulator is placed between the metal and the semiconductor, the device is a metal-insulator-semiconductor FET or MISFET. Oxides are the common insulator, and the devices based on oxide insulators are metal-oxide-semiconductor FET or MOSFET.

17.5.1. JFETs

The operation of FET is based on the change of the thickness of a conducting layer or channel, and hence the current flow through it. Fig. 17.14(a) shows the schematic diagram of a JFET.

The device is made from an *n*-type channel sandwiched between two *p*-type "gate" layers. The two ends of the channel are attached to metal contacts and are named drain and source. There are two depleted layers that are naturally formed between the *n*-type channel and the *p*-type gates. Under a zero bias, the thicknesses of the depleted layers are constant. However, if current *I* passes through the channel, the resistance of the channel results in a voltage gradient across it (Fig. 17.14(b)). This means that the voltage between the gate and the channel is higher at the drain compared to the source, and hence the thickness of the depleted layer is higher at the drain accordingly. The thickness of the depleted layer increases for higher gate biases, and thus higher channel currents, and at some point the depleted layers from both sides of the channel reach together. This situation is called pinch-off (Fig. 17.14(c)) and it prevents a further increase of the channel current even if a higher voltage is applied between the drain and the source.

After the initial pinch-off and upon further gate bias, the pinch-off point near the drain moves towards the source. The *n*-type channel from the source to the pinch-off point dominates the resistance of the electron flow through the *n*-type channel until the electrons are quickly swept across the highly resistive depletion region by the large electric field.

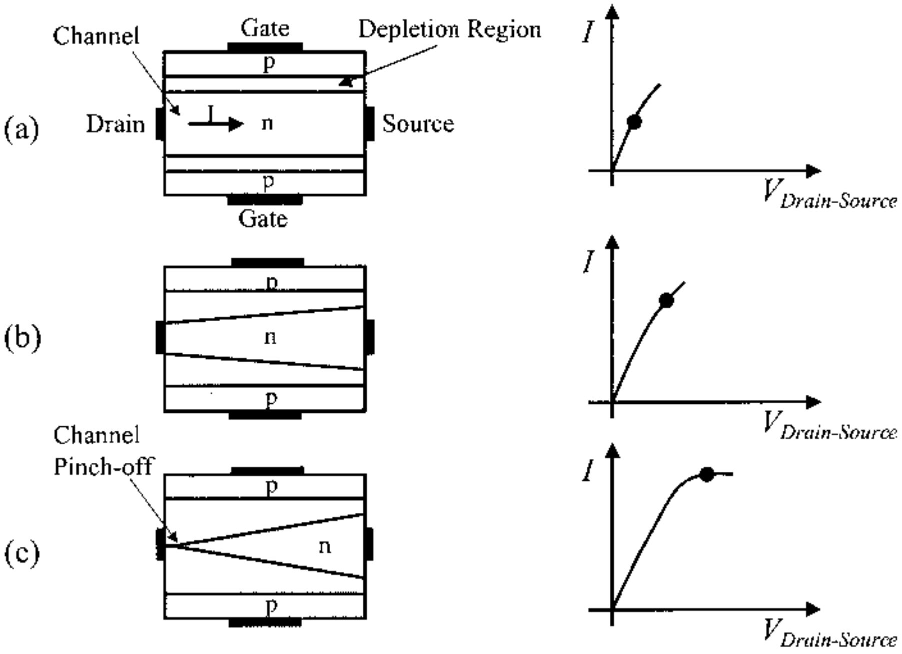


Fig. 17.14. Schematic diagrams of a junction field effect transistor under different operation conditions. (a) With a low bias between the drain and the source, the thicknesses of the depleted layers are nearly constant. (b) When a larger bias is applied and a higher current results across the channel, the depleted layers get thicker near the drain. (c) At some point, the depleted layers from both sides of the channel reach, resulting in the pinch-off condition when no further increase in the channel current is possible even if a higher bias is applied.

17.5.2. JFET gate control

A negative bias of the gate can simply increase the thickness of the depleted layer, and change the effective thickness of the channel. This means that the conductance of the channel can be reduced with a negative bias on the gate. More importantly, the pinch-off effect happens at a lower drain-source current. Fig. 17.15 shows the current-voltage relationship of the drain-source with different gate voltages.

Note that if the drain-source voltage is higher than the pinch-off condition, the drain-source current only depends on the gate voltage. Therefore, the device behaves as a current source that is controlled by the gate voltage. Such a characteristic is very useful in the design of AC amplifiers.

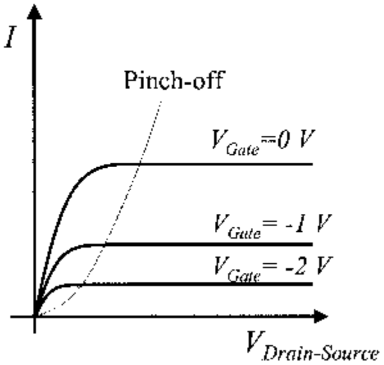


Fig. 17.15. Current-voltage relationship between the drain and the source of a field effect transistor as a function of gate voltage. The dotted curve shows the characteristic points where the pinch-off occurs.

The pinch-off voltage of the device can be calculated using our knowledge about the gate-channel p-n junctions. Assuming that the gates are heavily doped (p^+) and the built-in voltage of the junction V_0, is negligible compared to the gate-drain voltage V_{GD}, the depletion layer thickness is:

Eq. (17.28) $$W = \left(\frac{2\varepsilon(-V_{GD})}{qN_D} \right)^{1/2}$$

where ε is the permittivity of the semiconductor and N_D is the donor concentration in the channel. Now at the gate-drain voltage that pinch-off happens or $-V_p$ the depletion thickness is equal to the channel width a, we have:

Eq. (17.29) $$a = \left(\frac{2\varepsilon(V_p)}{qN_D} \right)^{1/2} \rightarrow V_p = \frac{qa^2 N_D}{2\varepsilon}$$

17.5.3. JFET current-voltage characteristics

Now we are in a position to calculate the current-voltage characteristic of a JFET. Fig. 17.16 shows a simplified diagram of the device.

Considering the symmetry of the device, the half of the channel width is called a, and the depletion layer from one side is $W(x)$ where the origin of the coordinate x is placed at the drain. The conducting part of the channel is $a\text{-}W(x)=h(x)$. Now if the width of the channel is Z, the total area of the channel at position x is:

Eq. (17.30) $A = 2h(x)Z$

and assuming that the resistivity of the channel is ρ, the resistance of the channel over a differential thickness *dx* is:

Eq. (17.31) $R(x) = \rho\frac{dx}{A} = \rho\frac{dx}{2h(x)Z}$

and the drain-source current *I*, is simply the voltage drop dV_x over the differential distance *dx* divided by the resistance *R(x)*:

Eq. (17.32) $I_{DS} = \frac{-dV_x}{R(x)} = -\frac{2Zh(x)dV_x}{\rho dx}$

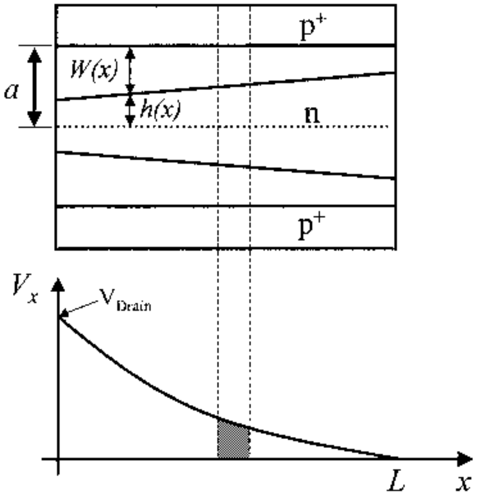


Fig. 17.16. Schematic diagram of a junction field effect transistor, showing the depletion layer width W(x) as a function of the distance from the drain x.

Now *h(x)* can be replaced with:

Eq. (17.33)

$$h(x) = a - W(x) = a - \left(\frac{2\varepsilon(V_x - V_G)}{qN_D}\right) = a\left[1 - \left(\frac{V_x - V_G}{V_p}\right)^{1/2}\right]$$

Inserting Eq. (17.33) into Eq. (17.32), we have:

Eq. (17.34) $$I_{DS}= -\frac{2ZdV_x}{\rho dx} a\left[1-\left(\frac{V_x - V_G}{V_p}\right)^{1/2}\right]$$

and now we can separate I and dV_x as:

Eq. (17.35) $$I_{DS}dx = -\frac{2Z}{\rho} a\left[1-\left(\frac{V_x - V_G}{V_p}\right)^{1/2}\right]dV_x$$

Integrating from both sides yield:

Eq. (17.36) $$I_{DS} = G_0V_p\left[\frac{V_D}{V_p} + \frac{2}{3}\left(-\frac{V_G}{V_p}\right)^{3/2} - \frac{2}{3}\left(\frac{V_D - V_G}{V_p}\right)^{3/2}\right]$$

17.5.4. MOSFETs

We will now take a look at the metal oxide semiconductor (MOS)-based FETs. The operation of MOSFETs is based on the effect of an electric field penetrating into the conductive channel between two highly-doped contact regions for the source and drain. The basic structure of an n-type MOSFET is shown in Fig. 17.17 along with a typical circuit schematic representation of a MOSFET.

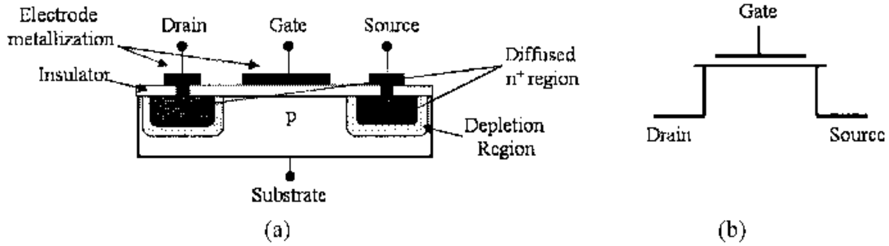


Fig. 17.17. (a) A schematic depiction of an NMOS FET. (b) The corresponding typical circuit schematic symbol.

The highly-doped n$^+$ regions are typically diffused into the p-type substrate and act as contact regions for the source and drain metal electrodes. The gate is electrically insulated from the substrate by an insulator, an oxide in the case MOSFETs.

By applying a positive bias to the gate electrode an electric field will extend into the substrate and deplete holes from the region directly under the gate electrode. As the gate bias is increased past some threshold value, V_{th}, an n-channel inversion layer forms under the gate and a conductive n-type current path between the n$^+$ source and drain is created.

Based on the biasing configuration there are two basic modes of operation for MOSFET devices; linear and saturation. The situation explained in the previous paragraph describes the linear mode of operation where an increase of the drain-source voltage, V_{DS}, will result in a linear increase in the drain current, I_D, depending upon the resistance of the channel. Similar to the case of the JFET, if V_{DS} is increased further the channel begins to pinch-off and the drain current saturates. This mode of operation is called the saturation region. These two modes of operation are depicted in Fig. 17.18(a) and (b).

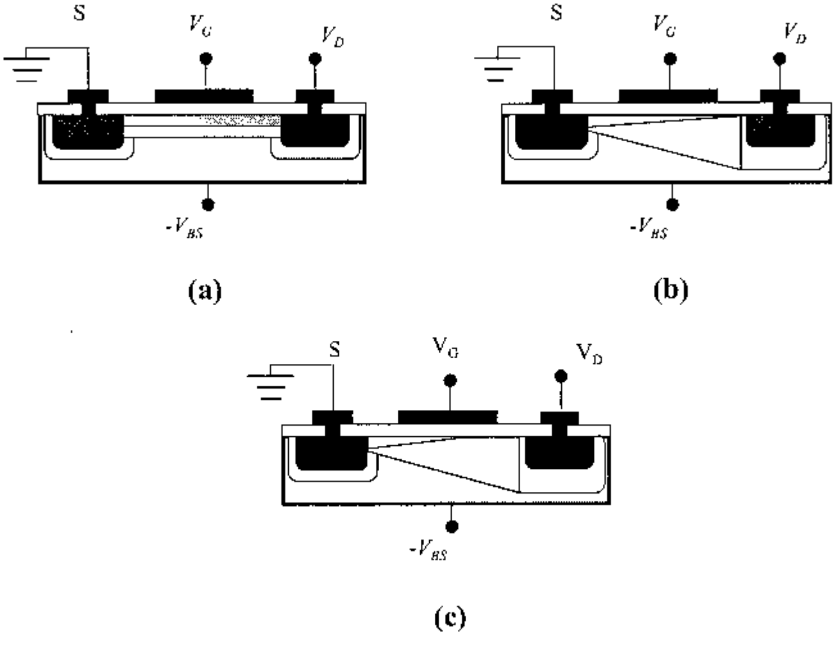


Fig. 17.18. (a) In the linear mode of operation of a MOSFET, where $V_{DS}<V_{GS}-V_{th}$, the drain current increases linearly with the drain voltage, V_D. (b) In the saturation mode of operation, where $V_{DS}>V_{GS}-V_{th}$, pinch-off occurs and the drain current saturates. (c) Upon further increase of V_{DS} the channel length is shortened and drain current increases even above that for the saturation case .

The relationship between drain current and V_{DS} for both modes of operation can be shown to be:

Eq. (17.37) $$I_D = \mu C_{ox}\frac{W}{L}\left[(V_{GS}-V_{th})V_{DS}-\frac{V_{DS}^2}{2}\right]$$

when $V_{DS}<(V_{GS}-V_{th})$ and the MOSFET is operating in the linear region and:

Eq. (17.38) $$I_D = \mu C_{ox}\frac{W}{L}\frac{(V_{GS}-V_{th})^2}{2}$$

when $V_{DS}>(V_{GS}-V_{th})$ and the MOSFET is in its saturation mode of operation, μ is the channel mobility, C_{ox} is the gate oxide capacitance, W is the channel width (into the page), L is the channel length, V_{GS} is the gate-source voltage, and V_{th} is the threshold voltage (~0.5-1 V for silicon-based devices).

17.5.5. Deviations from the ideal MOSFET case

Similar to BJT devices, the simplified MOSFET principles of operation and corresponding relationships do not fully explain the practical behavior of actual MOSFET devices. In this sub-section we will discuss velocity saturation, channel-length modulation, and insulator breakdown.

In an effort to keep pace with Moore's law, MOSFET transistor sizes have continuously followed a trend of decreasing minimum feature size. This scaling reduces parameters such as oxide thickness, gate length, transit time, current, power consumption, voltage, etc. Such scaling has strong benefits in terms of economics and performance but can cause a variety of short channel effects that complicate transistor and highly-integrated circuit design.

One of the most significant effects encountered with decreasing channel lengths is velocity saturation. At low electric fields, a linear relationship between carrier velocity and electric field strength is observed, even for short channels. But at higher field strengths, the carrier velocities begin to saturate and are on the order of the thermal velocity. The velocity of carriers under a high electric field saturates due to increased optical phonon emission. An approximation of this effect is given by the following analytical expression:

Eq. (17.39) $$v_d = \frac{\mu_n E}{1 + E / E_c}$$

where v_d is the carrier velocity, μ_n is the low-field mobility, E is the electric field, and E_c is the critical field value.

As discussed in the previous sub-section, there is a drain-source voltage, V_p, above which the conduction channel begins to pinch off. Increasing V_{DS} to a value greater than V_p, moves the pinch-off point towards the source, as shown in Fig. 17.18(c). This effectively reduces the channel length and increases the drain current, I_D. This effect can be mitigated by increasing the substrate carrier concentration. If V_{DS} is increased even further, punch-through can occur and the drain and source are effectively short circuited, similarly to the punch-through case of bipolar transistors.

An additional limitation on applied terminal voltages pertains to the gate voltage. Excessive gate voltages can cause the gate dielectric to catastrophically breakdown. This voltage is dependent upon the dielectric type and thickness, but is typically around 25~50V.

17.6. Application specific transistors

A brief summary of a few application specific transistor types will now be given. We will discuss single electron transistors, power transistors, and high electron mobility transistors.

Single electron transistors (SET) are metal-insulator-metal (MIM)-based devices that operate on the concept of electron tunneling. By placing two such MIM junctions in series with a gate capacitor connected to a third electrode between the two MIM junctions the SET device structure is realized. By increasing the voltage on the gate capacitor, electrons tunnel more quickly and the current through the device is increased. This makes an SET similar to a MOSFET but on a much smaller scale. If the gate capacitor is made even smaller, fewer electrons are involved in the tunneling current and quantization effects become more prominent.

Power transistors exist for both bipolar and MOSFET transistor types. Bipolar transistors are more traditionally used due to their robust ability to withstand high currents. This is generally due to the larger active area of these devices which allow for low current densities but high, up to 1000 A, currents. Silicon BJTs are most commonly encountered due to their relatively low cost of manufacturing but silicon carbide (SiC) is fundamentally capable of higher breakdown voltages and better thermal conductivity than silicon, but at a higher cost. Other bipolar power transistors include darlington transistors, thyristors, insulated gate bipolar transistors, and triacs.

High electron mobility transistors (HEMTs) are used in very low noise amplifiers and very high frequency applications such as microwave radio frequency, space communications, radio-based telescopes, and digital broadcasting systems. They are constructed of III-V compound semiconductor materials such as GaAs/AlGaAs, for example. In HEMTs, the heterojunction formed creates a two-dimensional electron gas (2DEG) which confines electrons to a very thin, high-mobility conduction layer resulting in a very low channel resistivity. By applying a gate voltage the conductivity of this channel is changed, effecting the current flow through the device, giving it its transistor-like behavior.

17.7. Summary

In this Chapter, we have described the general principles for electrical amplification and switching. We then modeled the amplification mechanisms, the charge distribution and transport in bipolar junction transistors. The advantages of heterojunction bipolar transistors have been discussed and illustrated with transistors based on AlGaAs/GaAs or GaInP/GaAs. Finally, the principles and electrical properties of field effect transistors were presented.

References

Alexandre, F., Benchimol, J.L., Dangla, J., Dubon-Chevallier, C., and Amarger, V., "Heavily doped based GaInP/GaAs heterojunction bipolar-transistor grown by chemical beam epitaxy," *Electronics Letters.* 26, pp. 1753-1755, 1990.

Ashizawa, Y., Noda, T., Morizuka, K., Asaka, M., and Obara, M., "LPMOCVD growth of C-doped GaAs-layers and AlGaAs/GaAs heterojunction bipolar-transistors," *Journal of Crystal Growth* 107, pp. 903-908, 1991.

Bachem, K.H., Lauterbach, Th., Maier, M., Pletschen, W., and Winkler, K., "MOVPE growth, technology and characterization of Ga0.5In0.5P/GaAs heterojunction bipolar-transistors," *Gallium Arsenide and Related Compounds* (*Institute of Physics Conference Series* 120), ed. G.B. Stringfellow, Institute of Physics, Bristol, pp. 293-298, 1992.

Biswas, D., Debhar, N., Bhattacharya, P., Razeghi, M., Defour, M., and Omnes, F., "Conduction-band and valence-band offsets in GaAs/Ga0.51In0.49P single quantum-wells grown by metalorganic chemical vapor-deposition," *Applied Physics Letters* 56, pp. 833-835, 1990.

Chang, M.F., Asbeck, P.M., Wang, K.C., Sullivan, C.J., Sheng, N.H., Higgins J.A., and Miller, D.L., "AlGaAs/GaAs heterojunction bipolar-transistors fabricated using a self-aligned dual lift-off process," *IEEE Electron Device Letters* 8, pp. 303-305, 1987.

Das, A. and Lundstrom, M.S., "Numerical study of emitter-base junction design for AlGaAs GaAs heterojunction bipolar-transistors," *IEEE Transactions on Electron Devices* 35, pp. 863-870, 1988.

Enquist, P.M. and Hutchby, J.A., "High-frequency performance of MOVPE npn AlGaAs/GaAs heterojunction bipolar-transistors," *Electronics Letters* 25, pp. 1124-1125, 1989.

Ginoudi, A., Paloura, E.C., Kostandinidis, G., Kiriakidis, G., Maurel, Ph., Garcia, J.C., and Christou, A., "Low-temperature DC characteristics of S-doped and Si-doped Ga0.51In0.49P/GaAs high electron-mobility transistors grown by metalorganic molecular-beam epitaxy," *Applied Physics Letters* 60, pp. 3162-3164, 1992.

Hayama, N., Okamoto, A., Madihian, M., and Honjo, K., "Submicrometer fully self-aligned AlGaAs/GaAs heterojunction bipolar-transistor," *IEEE Electron Devices Letters* 8, pp. 246-248, 1987.

Kobayashi, T., Taira, K., Nakamura, F., and Kawai, H., "Band lineup for a GaInP/GaAs heterojunction measured by high-gain npn heterojunction bipolar-transistor grown by metalorganic chemical vapor-deposition," *Journal of Applied Physics* 65, pp. 4898-4902, 1989.

Kuech, T.F., Wang, P.J., Tischler, M.A., Potenski, R., Scilla, G.J., and Cardone, F., "The control and modeling of doping profiles and transients in MOVPE growth," *Journal of Crystal Growth* 93, pp. 624-630, 1988.

Landgren, G., Rask, M., Anderson, S.G., and Lundberg, A., "Abrupt Mg doping profiles in GaAs grown by metalorganic vapor-phase epitaxy," *Journal of Crystal Growth* 93, pp. 646-649, 1988.

Mondry, M.J. and Kroemer, H., "Heterojunction bipolar-transistor using a (Ga,In)P emitter on a GaAs base, grown by molecular-beam epitaxy," *IEEE Electron Device Letters* 6, pp. 175-177, 1985.

Nagata, K., Nakajima, O., Nittono, T., Ito, H., and Ishibashi, T., "Self-aligned AlGaAs/GaAs HBT with InGaAs emitter cap layer," *Electronics Letters* 23, pp. 64-65, 1987.

Razeghi, M., Omnes, F., Defour, M., Maurel, Ph., Hu, J., Wolk, E., and Pavlidis, D., "High-performance GaAs GaInP heterostructure bipolar-transistors grown by low-pressure metalorganic-chemical vapor-deposition," *Semiconductor Science and Technology* 5, pp. 278-280, 1990.

Razeghi, M., *The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications*, Institute of Physics, Bristol, UK, 1995.

Seeger, K., *Semiconductor Physics: An Introduction*, Springer-Verlag, New York, 1997.

Twynam, J.K., Sato, H., and Kinosada, T., "High-performance carbon-doped base GaAs AlGaAs heterojunction bipolar-transistor grown by MOCVD," *Electronics Letters* 27, pp. 141-142, 1991.

Problems

1. Explain why BJTs are considered minority carrier devices and FETs are majority carrier devices?

2. Why is an n-p-n BJT used for high speed applications rather than a p-n-p BJT?

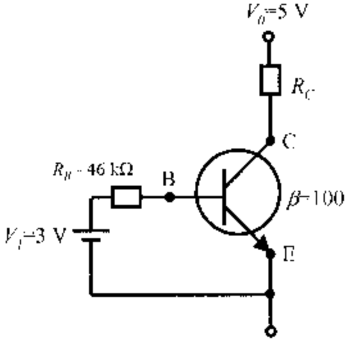
3. What is the typical difference in doping between the emitter and collector in a BJT and why?

4. What is the origin of the Early voltage in a BJT?

5. Thoroughly explain why a BJT performs so poorly in reverse active mode.

6. Summarize the four different modes of operation for a BJT.

7. Consider the electrical circuit shown below, which is designed to deliver a constant current through the collector. Estimate the range of values for the resistance R_C in order to keep this constant current source working properly. Assume that the collector to emitter voltage should be greater than 2 V and that V_{BE}=0.7 V.



8. Taking into account carrier recombination in the depletion region the current transfer ratio can be expanded to:

$$\alpha_0 = \alpha_T \gamma \delta$$

where δ is the depletion region recombination factor.

(a) Calculate the collector current, base current, DC current gain, and the current transfer ration of a BJT with an I_E of 2.5 mA and the following performance parameters:

α_T=0.998 (base transport factor)

γ=0.999 (emitter efficiency)

δ=0.997 (depletion recombination factor)

(b) Due to the Early effect the value of α_0 in part a is reduced by 0.004. By what percent does this affect the DC current gain of the BJT?

9. Consider a symmetrical p^+-n-p^- Si bipolar junction transistor with the following properties: $A=10^{-4}$ cm^2, W_b=0.8 μm, $N_A=3\times10^{17}$ cm^{-3}, and $N_D=3\times10^{15}$ cm^{-3}.

The characteristics of the emitter material are:

τ_n=0.15 μs

μ_p=300 cm^2/Vs

μ_n=800 cm^2/Vs.

The characteristics of the base material are:

τ_p=8 μs

μ_n=1500 cm^2/Vs

μ_p=500 cm^2/Vs

Calculate the saturation current on the collector side.

10. Calculate the pinch-off voltage for a silicon nMOSFET with a channel half width of 1.5 μm, and a donor concentration of 2.0×10^{15} cm^{-3}, ε_s=12.

11. Calculate I_D for an enhancement-mode nMOSFET with a length of L=1.3 μm, width of W=15 μm, an oxide thickness of t_{ox}=25 nm, and a threshold voltage V_T=0.75 V. The drain-source voltage V_{DS}=5 V and the gate voltage V_{GS}=3.5 V. Assume zero substrate bias and a mobility of 350 cm^2/Vs.

18. Semiconductor Lasers

18.1. Introduction

The word "laser" is an acronym for "light amplification by stimulated emission of radiation". The principles of lasers were understood at the end of 1950's [Schawlow *et al.* 1958]. The first working laser was built by Maiman in 1960, and used a ruby crystal optically pumped by a flash lamp. The Nobel Prize for fundamental work in the field of quantum electronics, which has led to the construction of oscillators and amplifiers based on the laser principle was awarded in 1964 to N.G. Basov, A.M. Prokhorov, and C.H. Townes.

The use of carrier injection across a p-n junction for stimulated emission from semiconductors was suggested as early as 1961. The stimulation emission itself was observed in a GaAs p-n junction one year later [e.g. Holonyak *et al.* 1962], and then the first semiconductor lasers were fabricated. The light emitted from a laser can be a continuous beam of low or medium power, or it can consist of short bursts of intense light delivering millions of watts.

This Chapter will first review the fundamental mechanisms of a laser. It will then describe the first laser, which was realized using a ruby crystal and subsequently focus on more sophisticated semiconductor lasers.

18.2. Types of lasers

Over the past forty years, scientists have investigated and developed many types of lasers. These lasers fall into several broad categories, as categorized in Fig. 18.1.

Solid Laser
Ruby
Nd:YAG
Ti:Sapphire

Gas Laser
HeNe
Argon-Ion
CO_2

Liquid (Dye) Laser
Polyphenyl
Rhodamine
Oxazine

Semiconductor Laser
III-V
II-VI
IV-VI

Fig. 18.1. Different types of lasers with several examples for each type.

Solid lasers are typically crystals that are doped with specific impurities, which introduce energy levels in the band structure of the crystal. These energy levels determine the energy (or wavelength) of the light emitted by the laser. Solid lasers need to be optically pumped in order to emit light, i.e. the energy needed to make the laser emit light is provided by illuminating the crystal with intense light. This is done typically with an incoherent white flash lamp, though many commercial lasers are now incorporating semiconductor lasers tuned to the optimum absorption frequency for higher efficiency. The power conversion efficiency of a laser is the ratio of the output power it emits to the total power used. Typical power conversion efficiency for a solid laser ranges from 0.1~5 %.

Gas lasers are very similar to solid lasers. However, instead of impurities, the energy of the light emitted depends on the gas mixture used. The gas mixture is excited by an electrical discharge. Due to a low absorption efficiency, high voltage discharge (typically 2~4 kV) is used to transfer energy to the gas mixture. The mixture normally consists of an inert gas that absorbs the discharge energy and transfers it to an active gas atom, whose allowed energy levels (as discussed in Chapter 2) determine the emission energy. Due to a large variety of gas mixtures, the power conversion efficiency of these lasers ranges from 0.01~15 %.

Liquid lasers are typically based on organic dyes dissolved in a solvent. These liquids exhibit groups of many closely spaced energy levels, which can provide a significant amount of emission wavelength tuning (around 90 nm in the visible range). Dye lasers are optically pumped, either with a flash lamp or another laser. The power conversion efficiency of this type of laser can be as high as 20 %.

The last type of laser relevant to this discussion is the semiconductor lasers. The band nature of semiconductor energy levels has already been explained in Chapter 4. When electrons in a direct bandgap material relax from the conduction band to the valence band, a photon i.e. light can be emitted. The wavelength of the emitted light can be changed widely through the use of various semiconductor materials with different bandgaps. Semiconductor lasers can be based on III-V, II-VI, and IV-VI compound semiconductors. This Chapter will deal primarily with III-V semiconductor lasers.

Although semiconductor lasers can emit light when optically pumped, they provide a significant advantage of the other types of lasers in that they can also be pumped electrically. Further, unlike gas lasers, the voltage requirements are minimal (1~2 V), which enables semiconductor lasers to be operated with the aid of only a simple battery. Semiconductor lasers are very efficient (up to 80 % power efficiency) and come in very small packages, similar to electrical transistors.

18.3. General laser theory

All the lasers discussed above have several common characteristics. Originally, in order to qualify as a laser, stimulated emission had to be demonstrated. The convention is now that a laser must demonstrate both stimulated emission and positive optical feedback. These concepts will be addressed in this section.

18.3.1. Stimulated emission

Most materials exhibit some kind of optical absorption. Absorption is the process by which incident light is converted to electrical potential energy. For example, an electron can be excited from the valence band to the conduction band by absorbing the energy of a photon. Logically, most materials also exhibit some form of light emission. For example, in an ordinary light source such as a light bulb, the emission of radiation occurs through a process called spontaneous emission. In this process, an electron which had been excited into a higher energy state for "some time" falls down to a lower state by emitting a photon. This "time" is called the radiative recombination lifetime.

However, in the presence of photons, an excited electron can be forced or stimulated to fall down to a lower state much faster than in a spontaneous event. The stimulus is provided by a photon with the proper wavelength. This process is called stimulated emission and produces an additional photon of exactly the same direction of propagation, frequency, and polarization (direction of the electric field in a wave) as the stimulating photon.

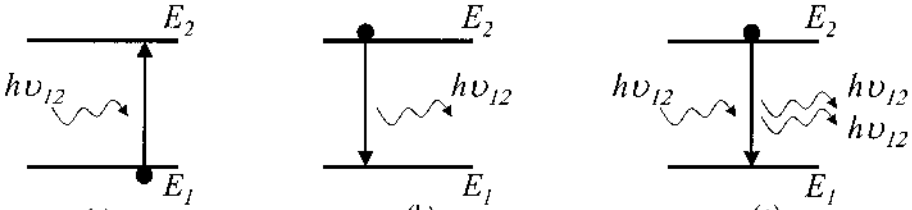


Fig. 18.2. Interaction of photons and electrons in a two energy level system: (a) optical absorption, an electron is excited from the lower level into the upper level by absorbing the energy of a photon of the correct energy; (b) spontaneous emission, the relaxation of the excited electron from the upper level to the lower level by release of energy in the form of a photon; (c) stimulated emission, the relaxation process is triggered by an incident photon and results in an additional photon with the same energy as the incident photon.

Spontaneous and stimulated processes are illustrated in Fig. 18.2. An electron in state E_2 drops spontaneously to E_1 emitting a photon with energy $h\upsilon_{12}=E_2-E_1$. Assuming that the system is immersed in an intense field of photons, each having energy $h\upsilon_{12}=E_2-E_1$, the electron is induced to transit from E_2 to E_1, contributing a photon whose wave is *in phase* with the radiation field. If this process continues and other electrons are stimulated to emit photons in the same fashion, a large radiation field due to stimulated emission can build up. This radiation is monochromatic since each photon has an energy $h\upsilon_{12}=E_2-E_1$ and is coherent because all the photons are released in the same phase.

Let us consider the general conditions necessary for stimulated emission to occur. We assume the populations of energy levels E_1 and E_2 ($E_1<E_2$) to be n_1 and n_2, respectively. At thermal equilibrium the relative population will be:

Eq. (18.1) $$\frac{n_2}{n_1}=e^{-(E_2-E_1)/k_bT}=e^{-h\nu_{12}/k_bT}$$

At equilibrium most electrons are in the lower energy level, i.e. $n_2<<n_1$. If the atoms exist in a radiation field of photons with energy $h\nu_{12}$ such that the energy density of the radiation field is $\rho(\upsilon_{12})$, i.e. the photon density of states, then stimulated emission can occur along with absorption and spontaneous emission. The rate of stimulated emission is proportional to the number of electrons in the upper level n_2 and to the energy density of the radiation field $\rho(\upsilon_{12})$. It can then be written as $B_{21}n_2\rho(\upsilon_{12})$ where B_{21} is the proportionality coefficient.

The rate at which the electrons make upward transitions from E_1 to E_2 (photon absorption) should also be proportional to $\rho(\upsilon_{12})$ and to the electron population in E_1. This rate is given by $B_{12}n_1\rho(\upsilon_{12})$, where B_{12} is a proportionality factor for upward transitions.

Finally, the rate of spontaneous emission is proportional only to the population of the upper level, $A_{21}n_2$.

The coefficients B_{21}, B_{12}, and A_{21} are called the Einstein coefficients. The ratio of the stimulated to spontaneous emission rates is:

Eq. (18.2) $$\frac{\text{Stimulated emission rate}}{\text{Spontaneous emission rate}}=\frac{B_{21}n_2\rho(\upsilon_{12})}{A_{21}n_2}=\frac{B_{21}}{A_{21}}\rho(\upsilon_{12})$$

which is generally very small, so the contribution of stimulated emission is negligible. From Eq. (18.2) it follows that in order to enhance the stimulated emission over spontaneous emission one has to provide large photon field energy density $\rho(\upsilon_{12})$. In a laser, this is encouraged by providing a resonant optical cavity in which the photon density can build up to a large value through multiple internal reflections.

Under thermal equilibrium the total rates of upward and downward transitions are equal:

Eq. (18.3) $$B_{12}n_1\rho(\upsilon_{12}) = A_{21}n_2 + B_{21}n_2\rho(\upsilon_{12})$$

Eq. (18.3) means that in equilibrium the ratio of the stimulated emission and upward transition rates is given by:

Eq. (18.4) $$\frac{\text{Stimulated emission rate}}{\text{Absorption rate}} = \frac{B_{21}}{B_{12}}\frac{n_2}{n_1} < 1$$

So, stimulated emission may dominate over absorption only when, (if $B_{12} = B_{21}$) $n_2 > n_1$, i.e. in a non-equilibrium state. This state is called population inversion. In summary, Eq. (18.2) and Eq. (18.3) indicate that stimulated emission can dominate if two requirements are met: (i) there is an optical resonant cavity to encourage the photon field to build up, (ii) there is population inversion.

The first requirement implies multiple passes of the light through the amplifying medium and the second requirement is a necessary condition for the medium to amplify the input light (Fig. 18.2(c)). Fig. 18.3 illustrates these two conditions for the generation of high intensity, coherent light.

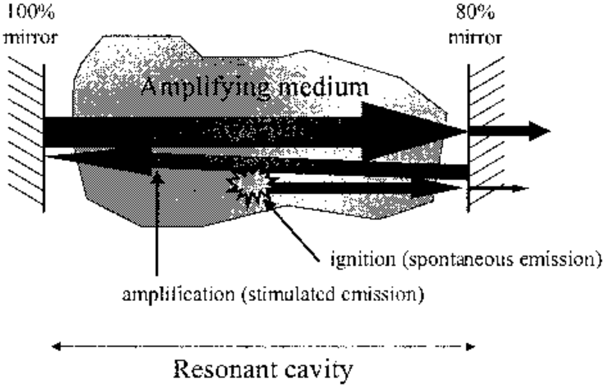

100% mirror
80% mirror
Amplifying medium
ignition (spontaneous emission)
amplification (stimulated emission)
Resonant cavity

Fig. 18.3. The elements necessary to a laser system: a resonant cavity which ensures the build up of a photon field, an amplifying medium in which the population inversion has been established, and an external pump which supplies the energy necessary to the system for amplification.

18.3.2. Resonant cavity

As shown in Fig. 18.4, light propagating inside the amplifying medium and perpendicularly to the mirrors (on-axis) with a particular frequency can be reflected back and forth within the resonant cavity in a reinforcing or coherent manner if an integral number of half-wavelengths fit between the end mirrors. Thus the length of the cavity for stimulated emission must be:

Eq. (18.5) $$L = \frac{m\lambda}{2}$$

where m is an integer. In this equation, λ is the photon wavelength within the laser material. It should be noted that this wavelength is related to the vacuum wavelength λ_0 through the relation:

Eq. (18.6) $$\lambda_0 = \lambda \bar{n}$$

where $\bar{n}$ is the refractive index of the resonant cavity material. The allowed wavelengths within the optical cavity are referred to as longitudinal optical modes.

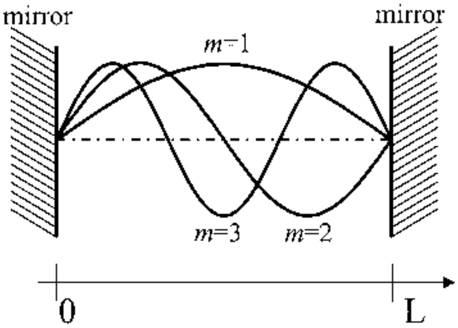


Fig. 18.4. Longitudinal optical modes in an optical resonant cavity delimited by two parallel mirrors. The length of the cavity is equal to an integral number of half-wavelengths of the light being amplified. This configuration leads to the constructive interference or reinforcing reflection of the light by the cavity, and thus amplification.

The amplifying medium is characterized by a definite wavelength region in which stimulated emission can occur. This is referred to as the material gain curve and is shown in Fig. 18.5(a). In a given resonant cavity including this amplifying medium, only the longitudinal modes will experience amplification, which leads to a characteristic laser output as shown in Fig. 18.5(c).

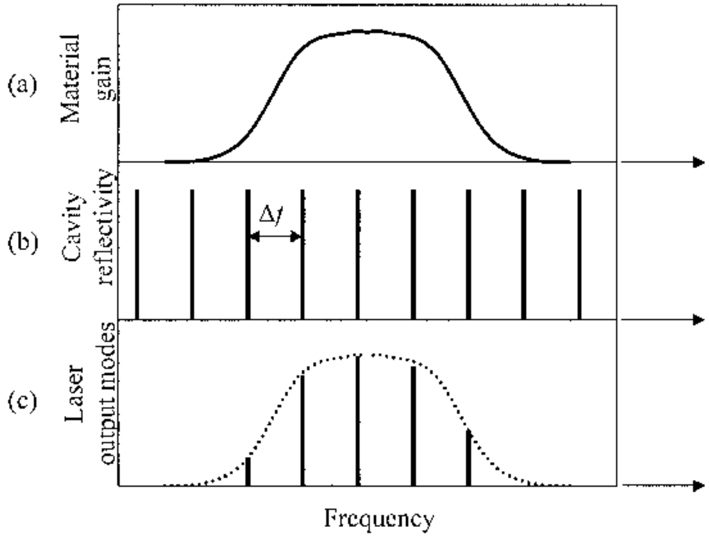


Fig. 18.5. Illustration of the effect of the cavity longitudinal modes on the laser output spectrum: (a) gain curve of the amplifying medium, (b) the wavelengths of the allowed longitudinal modes inside a given resonant cavity, and (c) the laser output modes.

Light output from a laser is also characterized by off-axis transverse modes. The transverse modes refer to the spatial intensity distribution at the exit mirrors. The origin of these modes is similar to the longitudinal modes, but is related to the spatial interference of light in the cavity. Fundamental modes are typically most intense on the axis of the cavity. Higher order modes exhibit multiple intensity peaks of increasing spatial frequency. Semiconductor lasers in particular demonstrate very specific transverse modes.

18.3.3. Waveguides

In order to achieve low power consumption and high efficiency, modern semiconductor lasers include thin layers (<1 μm) deposited by epitaxial techniques (Chapter 12) such that the electrons and holes are confined into a narrow region. In addition, when different types of materials with different refractive indices are used, these layers can also confine light, i.e. constitute a waveguide which is a medium in which a wave can propagate in one direction and is confined in others.

Light is an electromagnetic wave, composed of oscillatory electric and magnetic fields perpendicular to each other and the direction of propagation. The propagation of light in a medium can be described by Maxwell's equations which give the relation between the electric and magnetic fields in the wave. In this part, the propagation of the light inside a dielectric waveguide will be discussed.

Maxwell's equations introduced in Chapter 10 can be simplified in a dielectric since there is a negligible electrical current inside such a material. In the single frequency mode approximation and source free region, they are:

Eq. (18.7) $$\begin{cases} \vec{\nabla} \times \vec{E} = i\omega \vec{B} \\ \vec{\nabla} \times \vec{H} = -i\omega \vec{D} \\ \vec{\nabla} \cdot \vec{B} = 0 \\ \vec{\nabla} \cdot \vec{D} = 0 \end{cases}$$

where $\vec{E}$ is the electric field, $\vec{B}$ is the magnetic induction or flux density, $\vec{H}$ is the magnetic field strength, $\vec{D}$ is the electric displacement, and ω is the angular frequency of the electromagnetic wave (i.e. light). In an isotropic material, the displacement $\vec{D}$ and electric field strength $\vec{E}$ are related through the absolute permittivity ε of the material (in this Chapter the permeability is always the absolute permeability):

Eq. (18.8) $$\vec{D} = \varepsilon \vec{E}$$

Similarly, the magnetic field strength $\vec{H}$ and the magnetic flux density $\vec{B}$ are related through the permeability μ of the material:

Eq. (18.9) $$\vec{B} = \mu \vec{H}$$

Using Eq. (18.9) in the first relation in Eq. (18.7), we have:

Eq. (18.10) $$\vec{\nabla} \times \vec{E} = i\omega\left(\mu \vec{H}\right)$$

After applying $\vec{\nabla}\times$ to both sides of Eq. (18.10), we get:

Eq. (18.11) $$\vec{\nabla} \times \left(\vec{\nabla} \times \vec{E}\right) = i\omega\mu\left(\vec{\nabla} \times \vec{H}\right)$$

On the other hand, we have the mathematical relation:

Eq. (18.12) $\vec{\nabla}\times\left(\vec{\nabla}\times\vec{E}\right)=\vec{\nabla}\times\left(\vec{\nabla}\cdot\vec{E}\right)-\nabla^2\vec{E}$

where ∇^2 represents the Laplacian operator and is such that:

Eq. (18.13) $\nabla^2\vec{E}=\frac{\partial^2\vec{E}}{\partial x^2}+\frac{\partial^2\vec{E}}{\partial y^2}+\frac{\partial^2\vec{E}}{\partial z^2}$

In addition, using Eq. (18.8) in the last relation in Eq. (18.7), we get:

Eq. (18.14) $\vec{\nabla}\cdot\vec{D}=\vec{\nabla}\cdot\left(\varepsilon\vec{E}\right)=\varepsilon\vec{\nabla}\cdot\vec{E}=0$

which means, because the permittivity ε cannot be zero, that:

Eq. (18.15) $\vec{\nabla}\cdot\vec{E}=0$

Combining Eq. (18.11), Eq. (18.12) and Eq. (18.15) we find that:

Eq. (18.16) $-\nabla^2\vec{E}=i\omega\mu\left(\vec{\nabla}\times\vec{H}\right)$

Now, by inserting the second relation of Eq. (18.7) into this last equation, we get:

Eq. (18.17) $-\nabla^2\vec{E}=i\omega\mu\left(-i\omega\vec{D}\right)=\mu\omega^2\vec{D}$

Using Eq. (18.8), this becomes:

Eq. (18.18) $\nabla^2\vec{E}+\mu\varepsilon\omega^2\vec{E}=\vec{0}$

This equation is sometimes called the wave equation and governs the behavior of the electric field strength component $\vec{E}$ of an electromagnetic wave (i.e. light) in a medium. Knowing $\vec{E}$, one can use the first relation in Eq. (18.7) to calculate $\vec{B}$, the magnetic component of the propagating wave.

Solving the wave equation in any such non-isotropic structure also requires the knowledge of the boundary conditions. These are in part determined by the geometry of the waveguide. The wave equation is of the second order in $\vec{E}$ and, for a symmetric waveguide, yields even and odd parity solutions, which are called even and odd modes.

The simplest waveguide is a slab waveguide and consists of a high index core layer sandwiched by two other parallel layers, called cladding layers, with different refractive indices, as shown in Fig. 18.6. As a result, the propagation of the electromagnetic wave depends on whether the electric field is parallel to the layers or perpendicular to them. The former case is called TE polarization while the latter is called TM polarization.

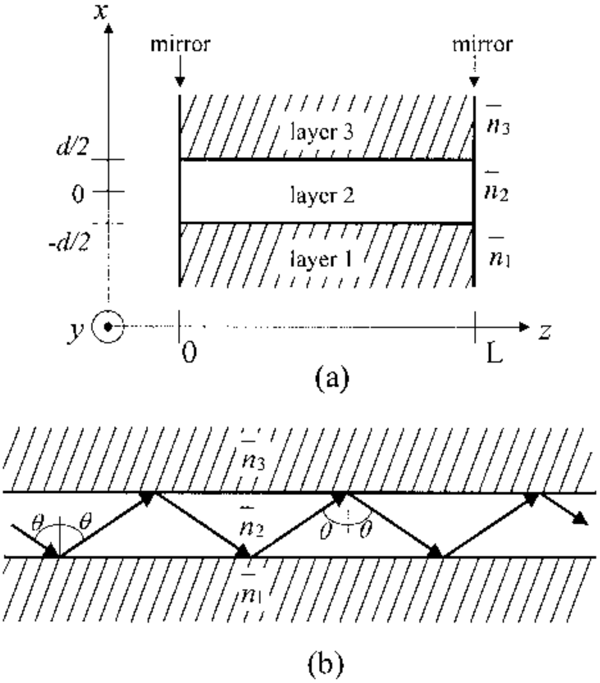


Fig. 18.6. (a) Representation of a three-layer dielectric waveguide, with three different refractive indices. (b) Ray trajectories of the guided wave, when the refractive index of the center layer is larger than those of the surrounding layers. The ray of light can experience total internal reflection at the interfaces between the dielectric materials, confining light to the core material.

To model the wave propagation in such a structure, one needs to consider two general solutions to the wave equation Eq. (18.18). The fundamental mode of most semiconductor lasers is TE. A TE mode must satisfy the wave equation and takes on the general form in the waveguide core:

Eq. (18.19) $E_y(x,z) = A\exp(ik_z z)\cos(k_x x + \phi)$

where A is a normalization constant, and:

Eq. (18.20) $k_x{}^2 + k_z{}^2 = \omega^2 \mu\varepsilon$

where k_z is the propagation wavevector in the waveguide, ε is the permittivity and μ the permeability of the waveguide. In the slab waveguide, some values of k_z cause k_x to become imaginary. This leads to a decay of the mode in the waveguide cladding, and is called the evanescent solution:

Eq. (18.21) $E_y(x,z) = \exp(ik_z z)[B\exp(-\alpha x) + C\exp(\alpha x)]$

where B and C are normalization constants, and:

Eq. (18.22) $-\alpha^2 + k_z^2 = \omega^2 \mu\varepsilon$

and:

Eq. (18.23) $\alpha = ik_x$

Let us assume that the waveguide is limited in the x-direction to within the region ($-d/2<x<d/2$), that it extends to infinity in the y-direction and that the waves propagate in the z-direction, as shown in Fig. 18.6. By solving the wave equation, one can find that the electric field in the even TE-polarization mode is along the y-direction and is given by:

Eq. (18.24)

$$E_y(x,z) = e^{ik_z z}\begin{cases} A_1 \exp\left(-\alpha_1\left(x - \frac{d}{2}\right)\right) & for\ x > \frac{d}{2} \\ A_2 \cos(k_{2x}x + \phi) & for\ -\frac{d}{2} < x < \frac{d}{2} \\ A_3 \exp\left(\alpha_3\left(x + \frac{d}{2}\right)\right) & for\ x < -\frac{d}{2} \end{cases}$$

where (α_1, k_{2x}, α_3) are the components of the wavevector in the x-direction defined by:

Eq. (18.25) $$\begin{cases} k_{2x}{}^2 + k_z^2 = \omega^2\mu_2\varepsilon_2 \\ -\alpha_1{}^2 + k_z^2 = \omega^2\mu_1\varepsilon_1 \\ -\alpha_3{}^2 + k_z^2 = \omega^2\mu_3\varepsilon_3 \end{cases}$$

and k_z is the propagation wavevector for the confined mode.

The magnetic field strength can be determined from the electric field strength using Eq. (18.10), given that $\vec{E} = E_y\vec{y}$:

Eq. (18.26) $$\vec{H} = \frac{1}{i\omega\mu_l}\left(-ik_zE_y\vec{x} + \frac{\partial E_y}{\partial x}\vec{z}\right)$$

where μ_l is the permeability of the material in layer *l*. The boundary conditions for the electric and magnetic field strengths in a waveguide rely on the continuity of their tangential components, i.e. E_y and H_z. Applying these boundary conditions at $x=\pm d/2$ yields a transcendental equation, similar to finding the bound states for an electron in a finite potential well (sub-section 3.3.3). By further assuming the permeabilities are such that $\mu_l=\mu_2=\mu_3$, which is the case for most III-V semiconductors, this transcendental equation has the form:

Eq. (18.27) $$\tan(k_{2x}d) = \frac{(\alpha_1+\alpha_3)k_{2x}}{k_{2x}{}^2 - \alpha_1\alpha_3}$$

Using Eq. (18.25), this can be solved graphically as a function of k_{2x}, as shown in Fig. 18.7.

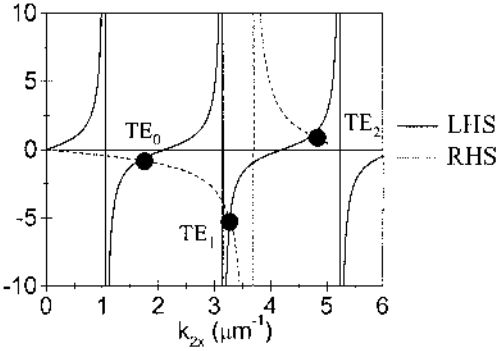


Fig. 18.7. Plot of the left-hand side (LHS) and right-hand side (RHS) of Eq. (18.27). The intersection points represent allowed optical modes referenced as TE_0 through TE_2.

The complete expression of the electric field strength can be obtained after applying the normalization condition:

Eq. (18.28) $$P = -\frac{1}{2}\mathrm{Re}\int_{-\infty}^{\infty}\left(E_y H_x^{\,*}\right)dx = 1$$

where $H_x^{\,*}$ is the complex conjugate of the quantity H_x. This gives:

Eq. (18.29)

$$E_y(x,z) = Ae^{ik_z z}\begin{cases}\cos\left(k_{2x}\frac{d}{2}+\phi\right)\exp\left(-\alpha_1\left(x-\frac{d}{2}\right)\right) & for\, x > \frac{d}{2}\\ \cos\left(k_{2x}x+\phi\right) & for\, -\frac{d}{2} < x < \frac{d}{2}\\ \cos\left(-k_{2x}\frac{d}{2}+\phi\right)\exp\left(\alpha_3\left(x+\frac{d}{2}\right)\right) & for\, x < -\frac{d}{2}\end{cases}$$

where:

Eq. (18.30) $$A = \sqrt{\frac{4\omega\mu}{k_z\left(d+\frac{1}{\alpha_1}+\frac{1}{\alpha_3}\right)}}$$

and:

Eq. (18.31) $$\phi = \cot^{-1}\left(\frac{k_{2x}}{\alpha_1}\right) - k_{2x}\frac{d}{2}$$

Each of these TE modes solutions of the above equations will be indexed with an integer p and has its own wavenumber k_z, varying from that in the cladding layer $k_z = \frac{\omega\bar{n}_1}{c}$ at low frequency to that in the core $k_z = \frac{\omega\bar{n}_2}{c}$ at high frequency.

When $\bar{n}_1 = \bar{n}_3 = \bar{n}_{clad}$ and $\bar{n}_2 = \bar{n}_{core}$, the slab waveguide is said to be symmetric. In this case, the allowed modes take on an even or odd parity and look similar to Fig. 18.8. All the TE$_p$ modes, for p>0, exhibit a cutoff frequency, ω_p, such that a wave with a lower frequency cannot propagate through the waveguide in that mode. For the symmetric waveguide, the predictable cutoff condition is:

Eq. (18.32) $$\frac{\omega}{c}\frac{d}{2}\sqrt{\bar{n}_{core}^2-\bar{n}_{clad}^2}=p\frac{\pi}{2} \qquad p=0,1,2,...$$

which indicates that the TE$_0$ mode always has a solution, while higher order modes may not.

The equations for TM modes are derived in a similar fashion to the TE. This is done by using the duality principle, which consists of replacing $\vec{E}$ and $\vec{H}$ with $\vec{H}$ and $-\vec{E}$, respectively, and swapping μ with ε in all the formulas. It should be realized that, in this case, E_x will be discontinuous due the change of permittivity at material interfaces.

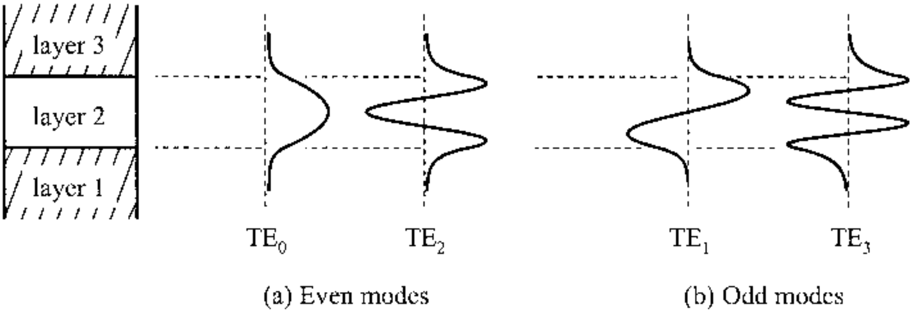


Fig. 18.8. Electric field profiles in TE modes in a symmetric waveguide: (a) even modes, (b) odd modes.

In practice, the waveguide in a semiconductor laser is formed by an active layer or active region surrounded by two cladding layers of different material. These layers typically have a smaller refractive index than the active layer, which confines the light within the active layer and guides it.

The extent of the confinement is mathematically described by the optical confinement factor, Γ, which in turn influences the threshold current and other laser characteristics. In terms of the above solution, the confinement factor can be expressed as:

Eq. (18.33) $$\Gamma = \frac{-\frac{1}{2}\mathrm{Re}\int_{-d/2}^{d/2}\left(E_y H_x^*\right)dx}{-\frac{1}{2}\mathrm{Re}\int_{-\infty}^{\infty}\left(E_y H_x^*\right)dx}$$

where E_y and H_x are the components of the electric and magnetic field strengths in the y- and x-directions, respectively. The confinement factor is always less than or equal to unity. It represents the percentage of the optical mode that is confined within the waveguide core.

A waveguide with a high confinement factor makes efficient use of emitted light and tends to have a low threshold gain/current. For a symmetric waveguide, a thick core and/or high refractive index difference between layers leads to a high confinement factor. This effect is shown graphically in Fig. 18.9.

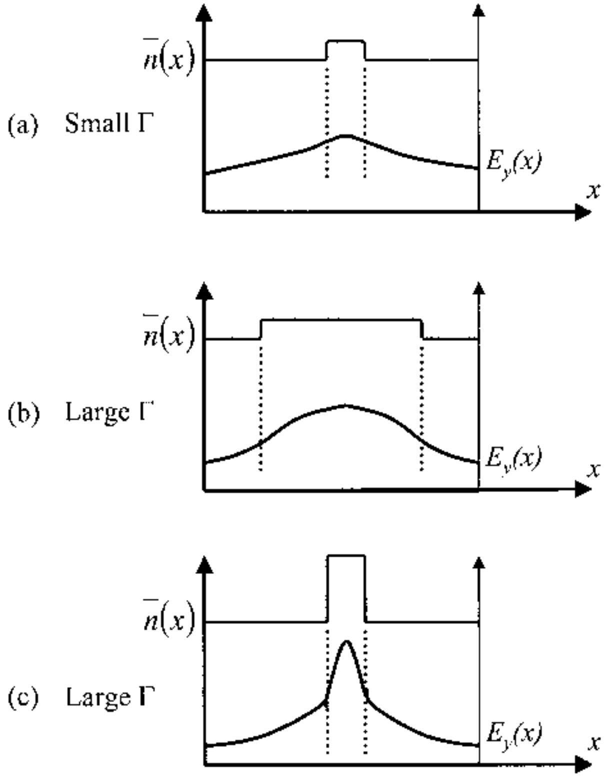


Fig. 18.9. TE$_0$ mode and confinement factor dependence on the refractive index difference $\Delta\bar{n}$ and core thickness d. (a) small d, small $\Delta\bar{n}$, small Γ, (b) large d, small $\Delta\bar{n}$, large Γ; (c) small d, large $\Delta\bar{n}$, large Γ.

18.3.4. Laser propagation and beam divergence

Regardless of the optical confinement, laser light exiting a semiconductor laser ($\overline{n}\approx3.4$) into air ($\overline{n}=1$) diverges in the x-direction due to diffraction. This happens for all lasers, but due to their small emitting aperture, the effect is much more pronounced in semiconductor lasers.

Let us consider the laser geometry shown in Fig. 18.10. At some distant observation point from the mirror (z=0), such that r>>r', the far-field approximation holds, and is given as:

Eq. (18.34) $$\left|\vec{r}-\vec{r}'\right| \approx \left|\vec{r}\right| - \vec{r}'\cdot\frac{\vec{r}}{\left|\vec{r}\right|}$$

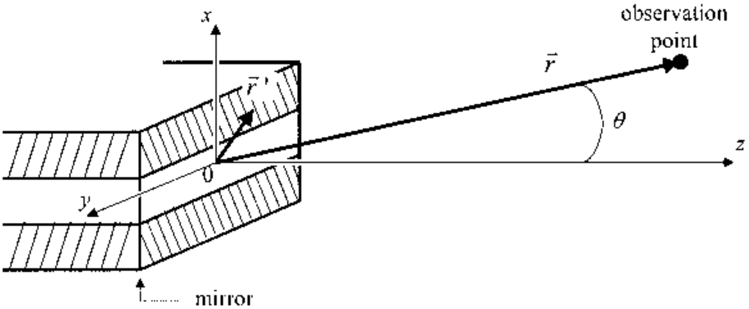


Fig. 18.10. Coordinates for calculating the far-field pattern at a distance r.

Following a derivation from Chuang [1995] based on the propagating radiation field, and the definition of the radiated power $I = \dfrac{1}{2\sqrt{\dfrac{\mu_0}{\varepsilon_0}}}\left|E(\vec{r})\right|^2$ the angular dependence of the power distribution for a TE mode is given by the following transform:

Eq. (18.35) $$I(\theta)_\perp \propto \cos^2\theta\left|\int_{-\infty}^{\infty} e^{-ikx'\sin\theta}E_y(x',z=0)dx'\right|^2$$

This intensity distribution is called the far-field pattern. The integral in Eq. (18.35) is effectively a spatial Fourier transform, as a small aperture translates to a large divergence.

In a symmetric slab waveguide characterized by a core thickness of d, a Gaussian function can be used to approximate the TE$_0$ transverse mode. This leads to an effective spot size, s_e, given by:

Eq. (18.36)
$$s_e = \begin{cases} \dfrac{d}{2}\left[1+\dfrac{2}{\alpha d}\ln\left(\sqrt{e}\cos\left(\dfrac{k_x d}{2}\right)\right)\right] & for\ \cos\left(\dfrac{k_x d}{2}\right) > \dfrac{1}{\sqrt{e}} \\ d\left[\dfrac{1}{\alpha d}\cos^{-1}\left(\dfrac{1}{\sqrt{e}}\right)\right] & for\ \cos\left(\dfrac{k_x d}{2}\right) < \dfrac{1}{\sqrt{e}} \end{cases}$$

This function is easier to transform, and leads to a simpler solution for the divergence angle, $\theta_\perp$, given by:

Eq. (18.37)
$$\theta_\perp \approx \frac{2}{ks_e} = \frac{\lambda_0}{\pi s_e}$$

In other words, the divergence angle is inversely proportional to the spot size. This is very important with regard to application. In general, the higher the divergence, the harder it is to collect all the light.

Unlike the y-direction of a slab waveguide, the transverse dimension of a real laser diode is not truly uniform. Various gain- and index-guided designs are used to control the transverse mode output in this direction as well, which makes

Eq. (18.38) $E_y(x,z) \rightarrow E_y(x,y,z)$

As a consequence, even in the transverse direction of the waveguide, there is divergence, as shown in Fig. 18.11. The resulting power intensity pattern in the two directions can be given as:

Eq. (18.39)
$$\begin{cases} I(\theta)_\perp \propto \cos^2\theta \left|\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-ikx'\sin\theta} E_y(x',y',z=0)dx'dy'\right|^2 \\ I(\theta)_{//} \propto \left|\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-iky'\sin\theta} E_y(x',y',z=0)dx'dy'\right|^2 \end{cases}$$

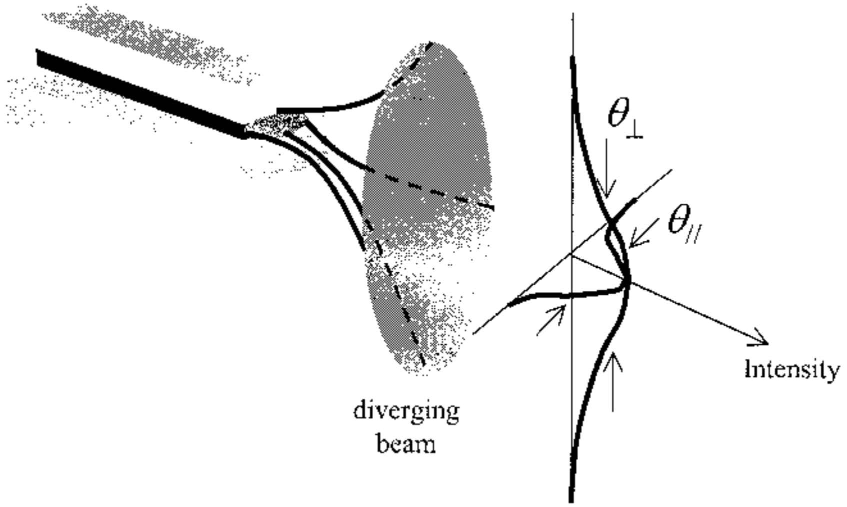


Fig. 18.11. Divergence characteristic of a non-infinite semiconductor laser waveguide.

18.3.5. Waveguide design considerations

It should be noted that higher order transverse modes will all have different divergence characteristics. When multiple modes exist, interference effects can produce non-Gaussian, poor quality, far-field patterns. This condition makes it extremely difficult to focus the laser output to a tight spot, which decreases the usable output.

The waveguide needs to be designed for a high confinement factor, low divergence, and good beam quality. Care must be taken to suppress higher order transverse modes, which appear more readily when there is a large index difference or a thick waveguide core. For the most usable output, a laser of given frequency ω, needs a waveguide such that $\omega<\omega_1$, where ω_1 is the cutoff frequency for the TE_1 mode.

For semiconductor lasers in general, even when higher modes do exist, it is the ellipticity of the semiconductor laser output, as shown in Fig. 18.11 that makes collection and focusing of light difficult. This is especially true for laser arrays in which there may be a very large aspect ratio to overcome. Cylindrical lenses are needed to first circularize the beam before the output can be utilized. A laser designed with a circular output can take advantage of simpler, and in general, high quality optics.

18.4. Ruby laser

The first working laser was built in 1960 by Maiman, using a ruby crystal as the amplifying or active medium. Ruby belongs to the family of gems consisting of sapphire or alumina ($A1_2O_3$) with various types of impurities. For example, pink ruby contains about 0.05 % Cr atoms. Similarly, $A1_2O_3$ doped with Ti, Fe, or Mn results in variously colored sapphire. Most of these materials can be grown as single crystals.

Ruby crystals are available in rods several inches long, convenient for forming an optical cavity (Fig. 18.12). The crystal is cut and polished so that the ends are flat and parallel, with the end planes perpendicular to the axis of the rod. These ends are coated with a highly reflective material, such as Al or Ag, producing a resonant cavity in which light intensity can build up through multiple reflections. One of the end mirrors is constructed to be partially transparent so that a fraction of the light will "leak out" of the resonant system. This transmitted light is the output of the laser. Of course, in designing such a laser one must choose the amount of transmission to be a small perturbation on the resonant system. The gain in photons per pass between the end plates must be larger than the transmission at the ends, as well as any other losses due to light scattering and absorption. The arrangement of parallel plates providing multiple internal reflections is similar to that used in the Fabry-Perot interferometer; thus the silvered ends of the laser cavity are often referred to as Fabry-Perot faces.

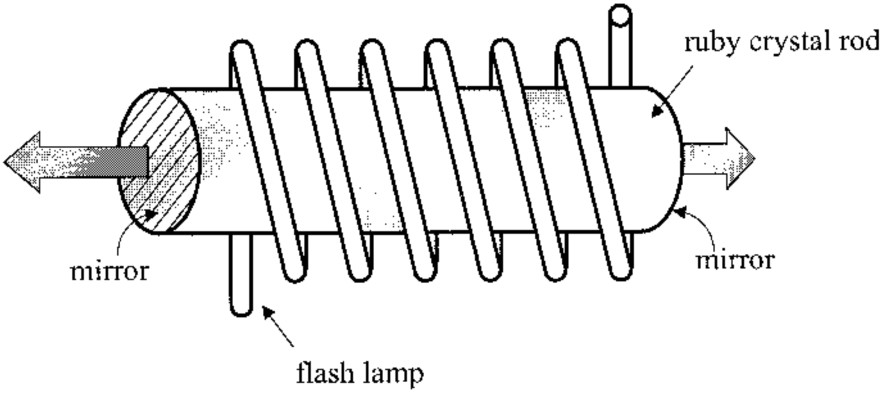


Fig. 18.12. Schematic diagram of a ruby laser. A ruby crystal rod is cut and polished so that its ends form mirrors to create the resonant cavity. A flash lamp supplies the necessary energy in the form of photons to pump the rod.

In the case of ruby, chromium (Cr) atoms in the crystal have their energy levels as shown in Fig. 18.13, where only the energy levels that are important for stimulated emission are depicted.

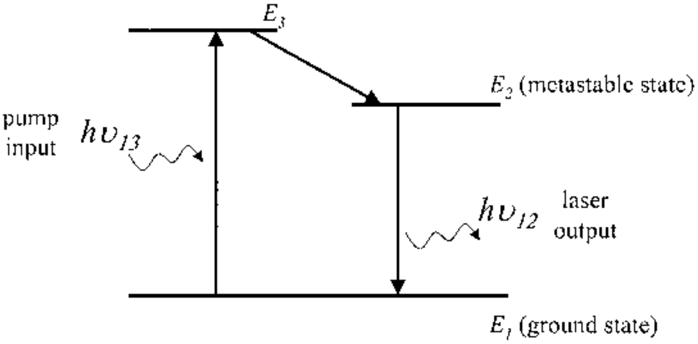


Fig. 18.13. Energy levels for chromium ions in ruby. The three level system includes a ground level at E_1, an excited level at E_3, and a metastable state at E_2 where the excited electrons relax rapidly to. The mean lifetime of the metastable state is long enough to ensure that population inversion can be achieved between the levels E_1 and E_2.

This is basically a three-level system. Absorption occurs in the green part of the spectrum, exciting electrons from the ground state E_1 to the band of levels designated E_3 in the figure. Then electrons decay rapidly to the level E_2. This transition is non-radiative. The level E_2 is very important for the stimulated emission process since electrons in this level have a mean lifetime of about 5 ns before they fall to the ground state. Because this lifetime is relatively long, E_2 is called a metastable state. If electrons are excited from E_1 to E_3 at a rate faster than the radiative rate from E_2 back to E_1, the population of the metastable state E_2 becomes larger than that of the ground state E_1 (we assume that electrons fall from E_3 to E_2 in a negligibly short time).

In the experiment done by Maiman in 1960, population inversion is obtained by optical pumping of the ruby rod with a flash lamp such as the one shown in Fig. 18.12. A common type of flash lamp is a glass tube wrapped around the ruby rod and filled with xenon gas. A capacitor can be discharged through the xenon-filled tube, creating a pulse of very intense light over a broad spectral range. If the light pulse from the flash tube is several milliseconds in duration, we might expect an output from the ruby laser over a large fraction of that time. However, the laser does not operate continuously during the light pulse but instead emits a series of very short spikes (Fig. 18.14).

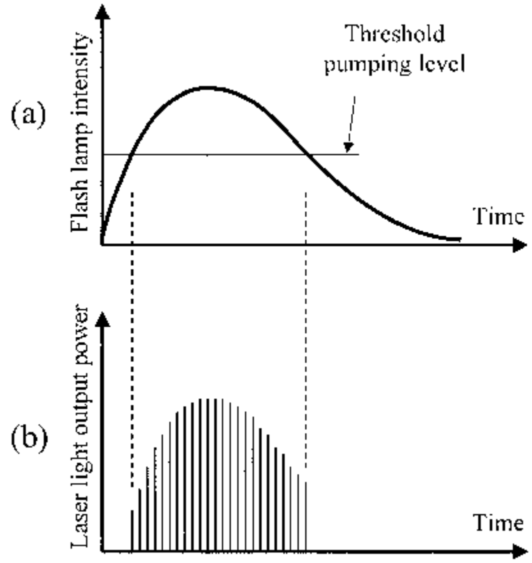


Fig. 18.14. Laser spikes in the output of a ruby laser: (a) typical variation of the intensity of the flash lamp with time: the intensity is above the threshold pumping level only during a certain period of time, not during the entire duration when the lamp is powered. (b) Laser spikes occurring while the flash intensity is above the threshold pumping level. As soon as population inversion is achieved between the levels E_1 and E_2, the laser emits a pulse of light. This process results in the series of laser intensity spikes.

When the flash lamp intensity becomes large enough to create population inversion (the threshold pumping level), stimulated emission from the metastable level to the ground level occurs, with a resulting laser emission. Once the stimulated emission begins, the metastable level is depopulated very quickly. Thus the laser output consists of an intense spike lasting from a few nanoseconds to microseconds. After the stimulated emission spike, population inversion builds up again and a second spike results. This process continues as long as the flash lamp intensity is above the threshold pumping level.

In this situation, one can easily understand that the metastable level never receives a highly inverted population of electrons. Whenever the population of E_2 reaches the minimum required for stimulated emission, these electrons are depleted quickly in one of the laser emission spikes.

To prevent this, we must somehow keep the coherent photon field in the ruby rod from building up (and thus prevent stimulated emission) until after a larger population inversion is obtained. This can be accomplished if we temporarily interrupt the resonant character of the optical cavity.

This process is called Q-switching, where Q is the quality factor of the resonant structure. A straightforward method for doing this is illustrated in Fig. 18.15. The front face of the ruby rod is silvered to be partially reflecting, but the back face is left un-silvered. The back reflector of the optical cavity is provided by an external mirror, which can be rotated at high speeds. When the mirror plane is aligned exactly perpendicular to the laser axis, a resonant structure exists; but as the mirror rotates away from this position, there is no buildup of photons through multiple reflections, and no laser action can occur. Thus during a flash from the xenon lamp, a very large inverted population builds up while the mirror rotates off-axis. When the mirror finally returns to the position at which light reflects back into the rod, stimulated emission can occur, and the large population of the metastable level is given up in one intense laser pulse. This structure is called a giant pulse laser or a Q-switched laser. By saving the electron population for a single pulse, a large amount of energy can be given up in a very short time. For example, if the total energy in the pulse is 1 Joule and the pulse width is 100 ns (10^{-7} s), the peak pulse power is 10^7 $J.s^{-1}$=10 MW.

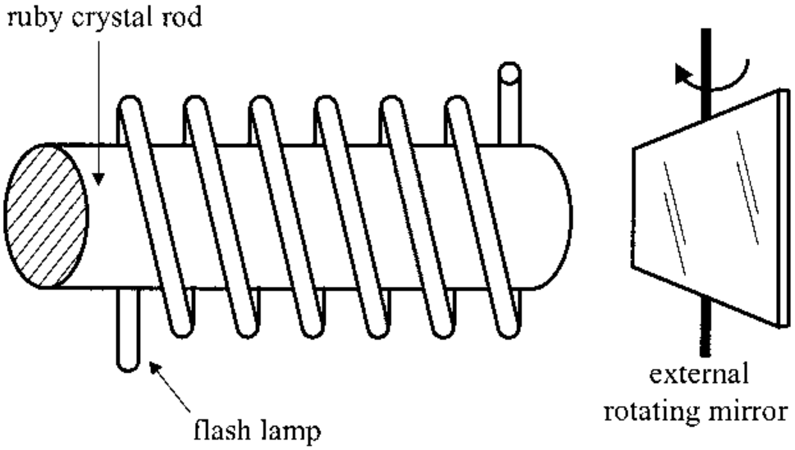


Fig. 18.15. Schematic diagram of a Q-switched ruby laser in which one face of the resonant cavity is an external rotating mirror. The purpose of the rotating mirror is to prevent stimulated emission by interrupting the resonant nature of the laser cavity, thus preventing the photon field from building up. This allows a larger population inversion to be achieved, and consequently, a higher intensity laser light emission.

18.5. Semiconductor lasers

In a semiconductor laser, population inversion mechanism is realized through a very unique method: by injecting electrical current directly into a p-n junction. This method of achieving population inversion is very efficient when compared to the process in ruby lasers or gas lasers. The semiconductor laser itself is also very compact (a typical size of the active laser part is only 100 μm x 1000 μm x 100 μm = one part in a hundred thousand cubic centimeters!).

Moreover, semiconductor lasers can be easily integrated with other types of semiconductor devices such as transistors or even large-scale integrated circuits, and the laser output can be easily modulated by controlling the junction current. It is no surprise that semiconductor lasers are now widely used for high speed optical processing and optical communication.

Another great advantage for these lasers is an inherent optical cavity. Most popular semiconductors (III-V, II-VI) have natural cleavage planes, which are the crystallographic planes along which the atomic bonds are weakest and therefore most easily broken. For zinc-blende crystals, cleavage parallel to (110) and $(1\overline{1}0)$ planes can produce atomically flat mirrors for use in a Fabry-Perot optical cavity. The reflectivity of the mirrors is limited by the refractive index of the semiconductor, and is given by:

$$\text{Eq. (18.40)} \quad R = \frac{\left(\overline{n}-1\right)^2}{\left(\overline{n}+1\right)^2}$$

where $\overline{n}$ is the refractive index of the semiconductor. A typical semiconductor has a refractive index of 3.4 in the mid-infrared region, which gives a natural mirror reflectivity of 29.8 %.

18.5.1. Population inversion

If a p-n junction is formed between degenerate materials, the bands under forward bias appear as shown in Fig. 18.16. If the injected current is large enough, electrons and holes are injected into and travel across the transition region in considerable concentrations. A large concentration of electrons is then present in the conduction band, while a large concentration of holes is present in the valence band, which satisfies the condition for population inversion.

Unlike the case of the three-level system discussed earlier (the ruby laser is essentially a three-level system), the condition for population inversion in semiconductors is more complicated. Both electrons and holes experience strong intraband scattering. The rapid intraband scattering and separate injection of electrons and holes allows for thermal equilibrium in each band separately. This situation is often called quasi-equilibrium and the particle distributions are described by Fermi distribution functions using quasi-Fermi energies for electrons and holes E_{F_n} and E_{F_p} (Chapter 8).

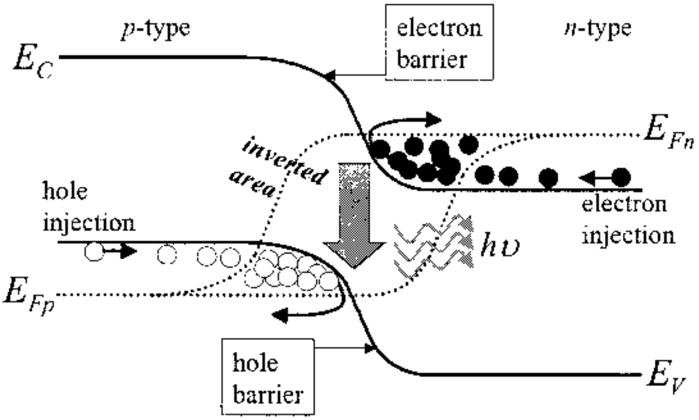


Fig. 18.16. Band diagram of a p-n junction laser under forward bias. The electrons and holes injected into the space charge region recombine radiatively to emit photons with an energy close to the bandgap of the semiconductor in the inversion region. When enough electrons and holes are injected, population inversion can be achieved, making laser emission possible.

Let us consider the population inversion in a semiconductor in more detail. We saw in Chapter 9 that, when an external bias was applied, minority carriers are injected on either side of the p-n junction and the quasi-Fermi levels go deep into each band apart from their equilibrium position somewhere in the bandgap. The absorption coefficient depends on the quasi-Fermi energies E_{F_n} and E_{F_p}. It changes sign when $E_{F_n} - E_{F_p} = h\upsilon$, for a given photon energy $h\upsilon$.

As the quasi-Fermi levels move apart from each other, this leads to a negative absorption coefficient which means the medium amplifies the light of frequency υ. The condition $E_{F_n} - E_{F_p} = h\upsilon$ is known as the Bernard-Durafforg condition, or transparency point, because at this point the absorption coefficient is zero. Reaching the transparency point, or population inversion, is a necessary condition for lasing. When $E_{F_n} - E_{F_p} > h\upsilon$, light of frequency υ is subject to amplification and is characterized by a gain, which is the opposite of the absorption coefficient.

As a result of this condition, the frequency of the light emitted by semiconductor lasers is larger than $\frac{E_g}{h}$. This in turn means that for lasing to occur in semiconductor lasers, it is necessary to apply a voltage $V = \frac{E_{F_n} - E_{F_p}}{q}$ at least higher than $\frac{E_g}{q}$.

18.5.2. Threshold condition and output power

The photon wavelengths which participate in stimulated emission are determined by the length of the resonant cavity as described in Eq. (18.5). Fig. 18.17 illustrates a typical plot of the light emission intensity versus photon energy for a semiconductor laser.

At low current levels, a spontaneous emission spectrum is observed, as shown in Fig. 18.17(a). As the current is increased to the threshold value, stimulated emission occurs at light frequencies corresponding to the cavity modes as shown in Fig. 18.17(b). Finally, at a higher current level, a most preferred mode or set of modes will dominate the spectral output, as shown in Fig. 18.17(c).

This very intense emission represents the main laser output of the device, where the output light will be composed of almost monochromatic radiation superimposed on a relatively weak radiation background, due primarily to spontaneous emission.

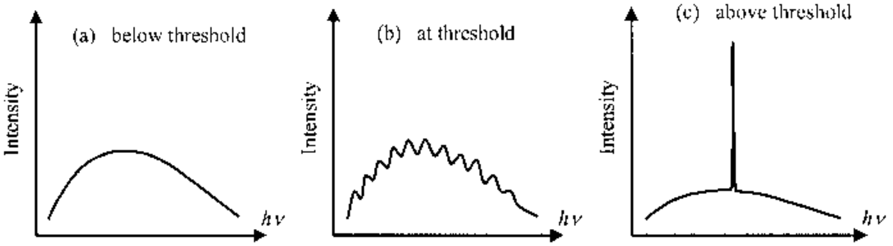


Fig. 18.17. Emission spectrum plotted as light intensity versus the energy of photons for a semiconductor laser: (a) below the threshold, an incoherent emission occurs with many photons emitted at several values of energy; (b) at threshold, laser modes appear which are determined by the dimensions of the resonant cavity; (c) above threshold, one dominant laser mode remains.

When the transparency condition is satisfied, the region becomes active, which means it can amplify light. The peak value g of the gain as function of frequency plays a major role in laser action. Typically, the peak gain is a linear function of carrier density:

Eq. (18.41) $g(n) = a(n - n_0)$

where n_0 is the transparency density and a is called the differential gain.

Although gain leads to light amplification, the optical losses, such as absorption outside the active region, α_i, and mirror loss, α_m, prevent the domination of the stimulated emission. So, in real lasers, the threshold current I_{th} has to provide not only the transparency condition but also has to compensate the optical losses in the laser cavity. This means that the real threshold current is larger than that which simply maintains the population inversion. In other words, the threshold gain g_{th} must compensate losses:

Eq. (18.42) $$\Gamma g_{th} = \alpha_i + \alpha_m$$

where Γ is the confinement factor which is the fraction of stimulated output mode power guided by the active region.

For a current density *J*, the carrier density rate equation is:

Eq. (18.43) $$\frac{\partial n}{\partial t} = D(\nabla^2 n) + \frac{J}{qd} - \frac{n}{\tau}$$

where *n* denotes the carrier density, *d* is the thickness of the active region, and τ is the lifetime of the non-equilibrium carriers. The first term of the above equation accounts for carrier diffusion with a diffusion coefficient *D*. The second term governs the rate at which the carriers are injected into the active layer. Since the active region dimensions are usually much smaller than the diffusion length, we assume the carrier density does not vary significantly over the active region so that the diffusion term can be neglected. Therefore, from $\frac{\partial n}{\partial t} = 0$ at steady-state, we get:

Eq. (18.44) $$J = \frac{qnd}{\tau}$$

When the threshold condition is reached, the carrier density is pinned at threshold value n_{th}, and the threshold current density can be expressed as:

Eq. (18.45) $$J_{th} = \frac{qn_{th}d}{\tau}$$

An important factor which determines laser output power is the internal quantum efficiency η_i which is the percentage of the injected carriers that contribute to radiative transitions. So, in the cavity, the photon density can be written as:

Eq. (18.46) $$N_{pn} = \eta_i \cdot \frac{J - J_{th}}{qd} \cdot \tau_p$$

where τ_p is the photon lifetime which is defined by $\tau_p^{-1} = v_g(\alpha_m + \alpha_i)$. $v_g = c/\overline{n}$ is the group velocity of the light. Since photons escape out of the cavity at a rate of $v_g\alpha_m$, the output power is related to the photon density by the relation:

Eq. (18.47) $$P = \hbar\omega \cdot v_g\alpha_m \cdot VN_{ph} = \hbar\omega \cdot v_g\alpha_m \cdot V \cdot \eta_i \frac{J - J_{th}}{qd} \cdot \frac{1}{v_g(\alpha_m + \alpha_i)}$$

where V is the volume of the active region. If we neglect current leakage, i.e. we assume that all the injected current passes through the active region, then the current can be written as $I = JS$, where S is the area of the active region. Considering $V = d \times S$, the output power depending on driven current $I > I_{th}$ is rewritten as:

Eq. (18.48) $$P = \frac{\hbar\omega}{q} \frac{\alpha_m}{\alpha_m + \alpha_i} \eta_i (I - I_{th})$$

Typical electrical and laser output power characteristics are shown in Fig. 18.18, in which V_t is the turn-on voltage for the diode, R_s is the series resistance above diode turn-on, and η_s is the slope efficiency.

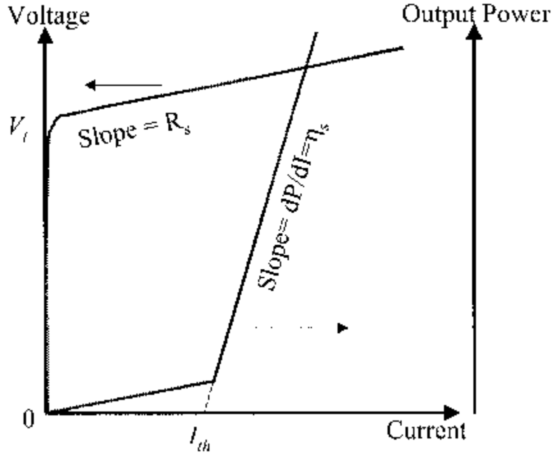


Fig. 18.18. Typical electrical and laser output power characteristic of a semiconductor laser, visualized as the current-voltage and output power-current characteristic. The linear part of the current-voltage curve gives the diode series resistance (R_s), while the linear part of the output power-current curve yields the slope efficiency (η_s).

In real lasers, the linear current dependence in Eq. (18.48) saturates due to such factors as leakage of carriers from the active region, heating, and current induced increases in the internal loss α_i.

The external differential quantum efficiency is defined as:

Eq. (18.49) $$\eta_d = \frac{dP/dI}{\hbar\omega/q} = \eta_i \frac{\alpha_m}{\alpha_i + \alpha_m}$$

Since the optical field should reproduce itself after each round trip under steady-state, the mirror loss α_m can be determined by:

Eq. (18.50) $$R_1 R_2 \cdot \exp(-\alpha_m \cdot 2L) = 1$$

where R_1, R_2 are the facet reflectivities at the two ends and L is the cavity length. From the above equation we obtain:

Eq. (18.51) $$\alpha_m = \frac{1}{2L} \ln(\frac{1}{R_1 R_2})$$

When $L \to 0$, we have $\alpha_m \to \infty$ and $\eta_d \to \eta_i$. Therefore, by plotting $1/\eta_d$ versus L and extrapolating to L=0, the internal quantum efficiency can be determined. Further the slope of the curve is proportional to the internal loss, α_i. In III-V double-heterostructure lasers η_i is close to unity and α_i ranges from 1~100 cm^{-1}.

18.5.3. Linewidth of semiconductor laser diodes

The linewidth in laser diodes depends on the instantaneous changes of phase and intensity in the lasing field [Henry 1982]. These instantaneous changes of phase and intensity have two components: (1) The components directly related to the spontaneous emission and (2) The components directly related to the coupled relationship between the phase and intensity of the field. These changes in field induce a perturbation in the carrier distribution (affecting the real and imaginary parts of the refractive index) which is incorporated into equation by a linewidth enhancement factor denoted as α. The mathematical derivation is rather lengthy and the reader is referred to the book by Chuang [1995]. The linewidth Δf of a typical laser in terms of frequency is then given as:

Eq. (18.52) $$\Delta f = \frac{v_g^2 h\upsilon g n_{sp} \alpha_m (1+\alpha^2)}{8\pi P}$$

where n_{sp} is the spontaneous emission factor which is related to the spontaneous emission rate R_{sp} via $R_{sp} = v_g g n_{sp}$.

18.5.4. Homojunction lasers

The first semiconductor laser was realized using a simple p-n junction as shown in Fig. 18.16. This is referred to as a homojunction, i.e. using the same semiconductor material for the active and surrounding layers. In this case, the difference in refractive index between the active layer and the adjacent layers is only 0.1~1 %. The result of this is that a lot of the emitted light escapes without undergoing feedback and amplification.

The primary benefits of the homojunction laser rely on simplicity of design and compactness compared to gas and traditional solid state lasers. The low confinement factor and high absorption loss lead to a very high threshold current density at room temperature (>100 kA.cm^{-2}) and low power conversion efficiency. These problems are solved in part by making use of a more elegant design, the heterojunction laser.

18.5.5. Heterojunction lasers

To obtain more efficient lasers, it is necessary to use multiple layers with different optical properties in the laser structure. When dissimilar materials are combined, a heterojunction laser can be formed.

An example of a single heterojunction laser is shown in Fig. 18.19. Carrier confinement is obtained in this single-heterojunction laser by using an AlGaAs layer grown epitaxially on GaAs. In this structure the injected carriers are confined to a narrow region so that population inversion can be built up at lower current levels. Further, because there is a noticeable refractive index change at the GaAs/AlGaAs interface, some waveguiding inside the epilayer and substrate is possible. These two effects help to reduce the average threshold current density at room temperature to ~10 kA.cm^{-2}.

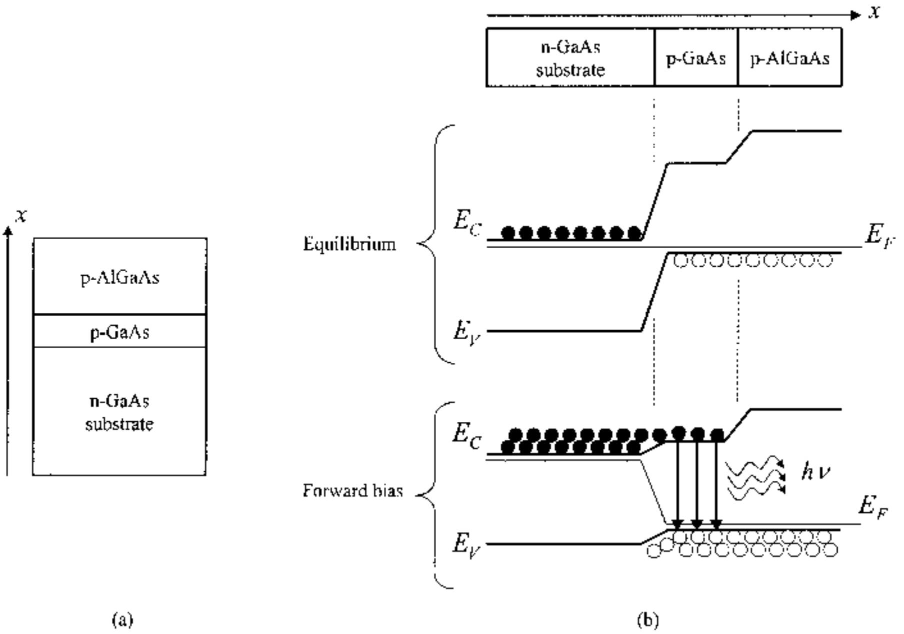


Fig. 18.19. Illustration of the use of a single heterojunction for carrier confinement in laser diodes: (a) cross-section schematic of an AlGaAs heterojunction grown on a thin p-type GaAs layer and on an n-type GaAs substrate; (b) energy band diagrams for this structure at equilibrium and under high forward bias, showing the confinement of electrons into the thin p-type region under forward bias.

A further improvement can be obtained by sandwiching the active GaAs layer between two AlGaAs layers. This double-heterojunction (DH) structure further confines injected carriers to the active region, and refractive index steps at the GaAs-AlGaAs boundaries form the waveguide that confines the generated light waves. A double-heterojunction laser, also called double-heterostructure laser, is shown in Fig. 18.20.

To date, the most extensively used heterostructure lasers are in the GaAs-AlGaAs and GaAs-InGaAsP systems. The ternary alloy $Al_xGa_{1-x}As$ has a direct bandgap for x up to $x\approx 0.45$, then becomes an indirect bandgap semiconductor. For heterostructure lasers, the composition region $0<x<0.35$ is of most interest and the direct energy gap of the ternary compound can be expressed as:

$$E_g = 1.424 + 1.247x \text{ (eV)}$$

The compositional dependence of the refractive index can be represented by:

$$\bar{n}(x) = 3.59 - 0.71x + 0.091x^2$$

For example, for x=0.3 the bandgap of $Al_{0.3}Ga_{0.7}As$ is 1.798 eV which is 0.374 eV larger than GaAs; its refractive index 3.385 is about 6 % smaller than the GaAs.

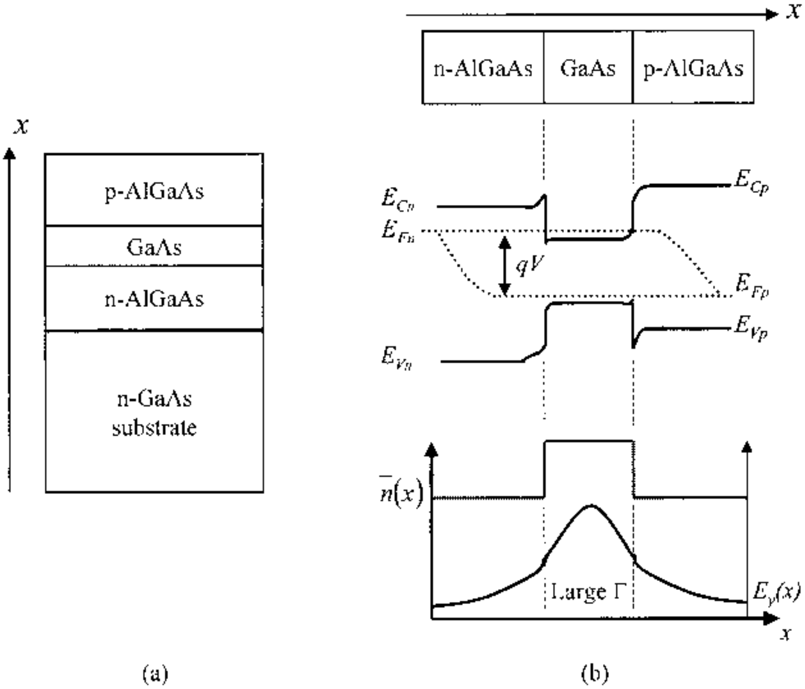


Fig. 18.20. Illustration of a double-heterojunction laser structure used to confine injected carriers and provide waveguiding for the light: (a) an n-type doped AlGaAs layer has been added between the n-type GaAs substrate and the GaAs active layer in the structure of Fig. 18.19 to form the double-heterostructure; (b) band structure and optical waveguide properties of the resulting laser structure at forward bias (V). The Solid lines in the band structure represent band edges, while the dashed lines represent quasi-Fermi levels.

Because a DH laser requires several different materials and a controlled doping profile, more sophisticated epitaxial techniques had to be developed, as described in see Chapter 12. The deposition can be based on liquid, vapor, or atomic beam processes. All processes allow some degree of multilayer growth with accurate n- and p-type doping done in-situ. Lastly, in order to avoid threading dislocations, the various crystal layers should be lattice-matched to the substrate.

The optical and electrical confinement of the double heterostructure give it a significant advantage over the homojunction and single-heterojuction laser. Indeed, the threshold current density is reduced by an order of magnitude for most structures. However, in order to maintain a high confinement factor and minimize loss in the cladding regions, the active region thickness must be quite large (0.1~0.5 μm). The thick layer, combined with a large density of states in the active region, requires a large number of carriers to maintain a sufficient population inversion. Modern DH lasers have thresholds current densities ~1 kA.cm^{-2}.

18.5.6. Device fabrication

After the laser structure is designed and epitaxially grown, a laser device must be fabricated using the photolithographic and metallization processes discussed in Chapter 16. Metal contacts are used to inject electrons and holes into the active region, therefore allowing the creation of a population inversion. In order to realize a low contact resistance between the metal and the semiconductor film, most laser structures employ a highly doped cap layer directly on top of the waveguide cladding layers. This will tend to form an ohmic rather than Schottky contact when metal is evaporated onto the surface.

The other feature a laser must exhibit is optical feedback. After the contact region is formed, the laser must be diced into separate cavities for testing. As discussed earlier, this is often accomplished by taking advantage of the natural cleavage planes in the crystal material. Cleaving is most easily achieved when the substrate is thinned down.

Subsequently, for testing and excess heat removal, the laser chip must be mounted on some type of heat sink. The heat sink also supplies mechanical stability, while allowing electrical connection to an outside circuit.

Example 1. Broad area laser fabrication

The simplest semiconductor laser to fabricate is a broad area laser. The typical fabrication steps for such a laser are shown in Fig. 18.21. No lithography is needed and the fabrication is just enough to satisfy the above requirements. A thin layer of metal is evaporated or electrochemically deposited on the top surface. The type of metal used depends on the material and doping type of the semiconductor, and is chosen to yield a low resistance, ohmic metal-semiconductor contact. The metal is typically annealed to both increase the metal adhesion and establish the ohmic contact.

In order to decrease electrical and thermal resistance, as well as make the wafer more conducive to cleaving of Fabry-Perot cavities, the back side of the wafer is thinned. Lapping and polishing brings the wafer thickness down to ~100 μm. The polished surface is then cleaned and the bottom metal contact is deposited and annealed, as shown in Fig. 18.21(c) and (d).

After both contacts are formed, the wafer is cleaved into individual laser cavities with a well-defined length and width, as shown in Fig. 18.21(e).

After cavity formation, the laser is then die bonded to a submount, or directly onto a heat sink (Fig. 18.21(f)). Besides removing waste heat, the heat sink provides mechanical stability and allows macroscopic external electrical contact.

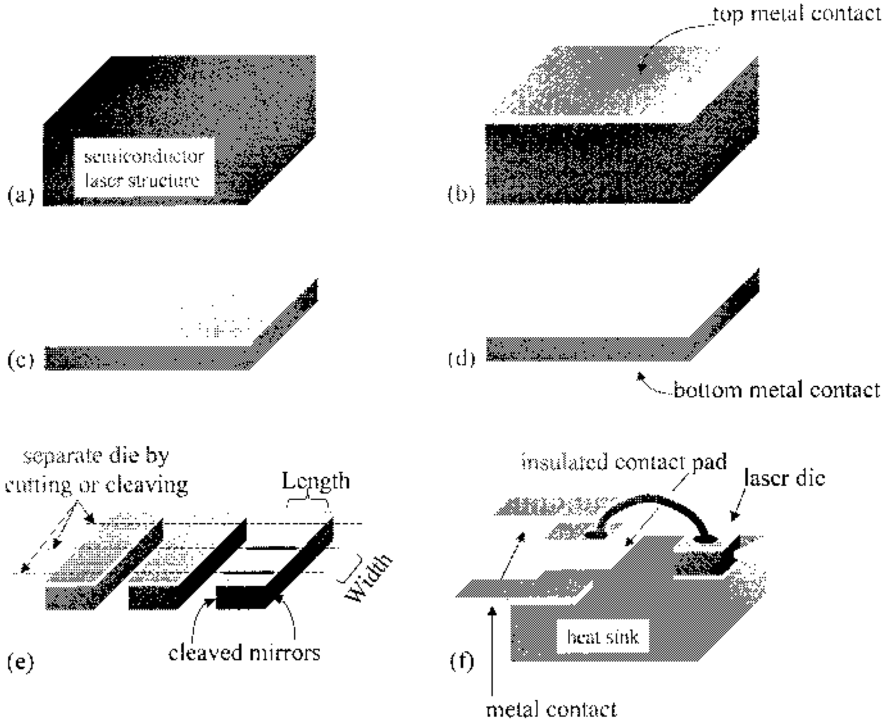


Fig. 18.21. Broad area laser fabrication steps: (a) bare semiconductor laser structure, (b) top contact metallization, (c) substrate lapping and polishing to thin down the substrate, (d) bottom contact metallization, (e) cleaving of laser cavities, (f) bonding of laser die to heat sink and external contact formation.

Broad area lasers use one of the fastest fabrication methods. The device features confinement of light and carriers within the large slab formed by the cleaving process. Unfortunately, mechanical cleaving can only reliably produce lengths and widths >200 μm. Further, the cleaving produces minor damage where the surface was scribed along the lateral edges. The net effect is a large current requirement and an increased chance of failure.

Example 2. Stripe-geometry laser fabrication

Stripe-geometry lasers are lasers in which the current is restricted along the junction plane. In this technique, metal contact stripes are defined which are typically 5~200 μm wide. This technique, while more difficult to fabricate, does allow for a lower operating current and reduced failure rate by keeping the injected area small and far from the (lateral) edges of the chip.

Stripe-geometry lasers can be fabricated in a variety of ways. One of the simplest techniques is shown in Fig. 18.22. The cavity width is defined by a patterned insulator, such as SiO_2. The insulator is typically deposited using chemical vapor deposition, rf sputtering, or e-beam evaporation. Patterning is done using standard optical photolithography and etching, as described in Chapter 16.

The etching of the insulator can typically be done with chemicals or plasma. For SiO_2, a buffered hydrofluoric acid (HF) solution is generally a good selective chemical etchant. The plasma process varies, but typically uses CF_4 or a similar fluorocarbon species.

After the stripe definition, the process followed is similar to the broad area laser. The top contact metallization is followed by substrate thinning and bottom contact metallization. As shown in Fig. 18.22(d), the current injection can be well confined to the area under the insulator opening. Laser die bonding is straightforward, and device operation is relatively insensitive to the lateral edges of the chip.

One of the disadvantages of the stripe-geometry laser is lateral current leakage. Even though the width of the cavity is defined at the surface, the injected current spreads out as it travels toward the substrate. The relative amount of spreading at a given current depends on the stripe width as well as the lateral conductivity and carrier diffusion length of the layers.

Because the gain of the laser varies along the junction plane due to this current spreading, the effective complex refractive index is also non-uniform. The complex refractive index is highest at the center of the stripe and decreases in a quadratic manner with distance, until it reaches its equilibrium value. A weak waveguiding effect is anticipated, as derived in Casey and Panish [1978]. This is referred to as a gain-guided laser.

While helpful in reducing the threshold current requirements, gain-guiding typically shows multiple transverse mode output in the junction plane.

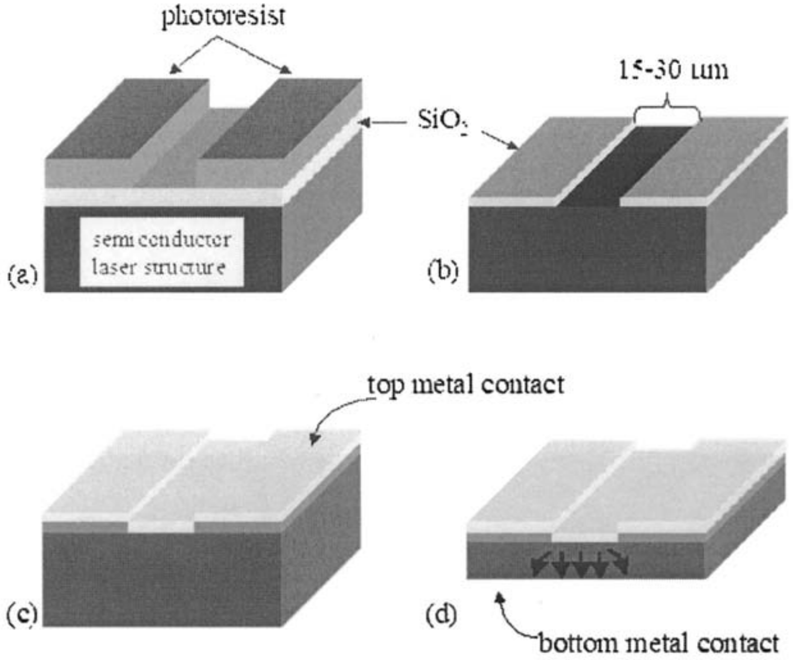


Fig. 18.22. Stripe-geometry laser fabrication steps and schematic: (a) photolithography on SiO_2 insulation, (b) patterned SiO_2, (c) top contact metallization, (d) schematic of final device showing localized current injection paths.

Example 3. Buried-heterostructure laser

The last structure that will be discussed is referred to as a buried-heterostructure laser. This device goes one step further in complexity in order to completely confines both current and the optical mode around a small emitting core.

The fabrication of the buried-heterostructure lasers, as shown in Fig. 18.23, starts with the growth of the *n*-type waveguide cladding and active layer(s). Using photolithography, the core is patterned with photolithography and etching into very narrow stripes (Fig. 18.23(b)). The width is small in order to confine only a single transverse optical mode, and depends on the laser emission wavelength as well as the index difference between the core and cladding regions.

After the core is defined, epitaxial regrowth is used to cover the core with a low index, high-bandgap, *p*-type cladding (Fig. 18.23(c)). Following this step, standard thinning, metallization, and die bonding is performed to complete the device.

While epitaxial regrowth is technologically very challenging, this fabrication procedure has the potential of producing the most efficient lasers with a single transverse mode. The current is confined thanks to the cladding-core material band offset, and the light is confined thanks to the refractive index change. The output beam quality is generally very high, which makes these lasers attractive for fiber coupling and telecommunications. The only drawback is a low output power, which scales as the core volume.

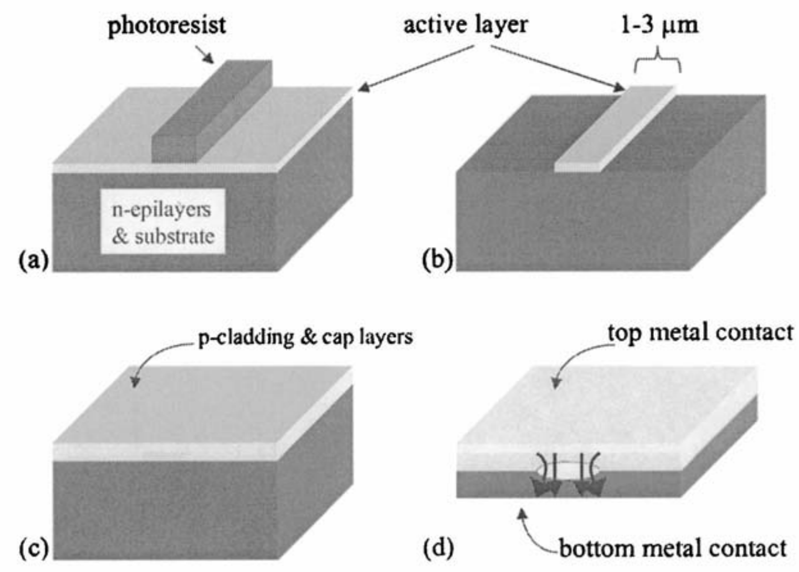


Fig. 18.23. Fabrication steps and schematic of buried-heterostructure laser: (a) photolithography to define waveguide core, (b) patterning of core, (c) semiconductor regrowth of waveguide cladding, (d) schematic of final device showing the confined current and optical confinement to achieve single transverse mode output.

18.5.7. Separate confinement and quantum well lasers

A further advancement on the double heterostructure laser uses several different kinds of materials to separate the optical and electrical confinement into separate regions. The separate confinement heterostructure (SCH) typically uses a thin or quantum well-based active region surrounded by an intermediate waveguide layer, all of which are embedded in the standard high-bandgap cladding region. A schematic of the heterostructure and the optical waveguide properties under forward bias are shown in Fig. 18.24.

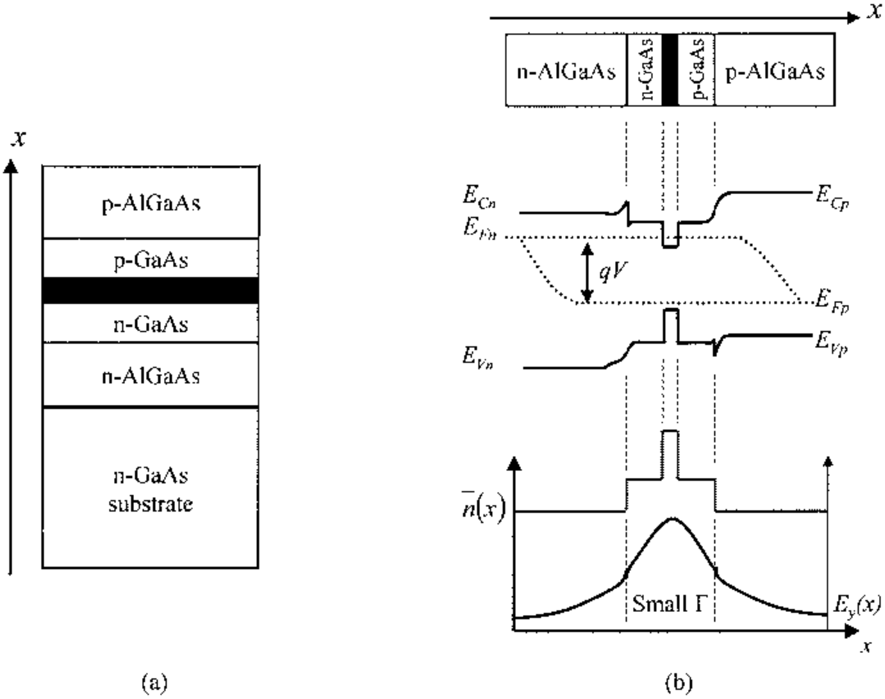


Fig. 18.24. Illustration of a separate confinement heterostructure laser: (a) cross-section of the device structure; (b) band structure, index and optical mode profiles of the resulting laser structure at forward bias (V). The Solid lines in the band structure represent band edges, while the dashed lines represent quasi-Fermi levels.

For quantum well (QW) or multi-quantum well (MQW) active regions, the density of states is reduced. This fact, combined with a narrow width for the well (<30 nm) leads to a population inversion at very low current densities. The gain also takes on another general form as a function of carrier density and is given by:

Eq. (18.53) $$g_w = g_0\left[\ln\left(\frac{J_w}{J_0}\right)+1\right]$$

where g_0 and J_0 are the transparency gain and current density respectively. Furthermore, because the carrier injection is more uniform, the internal quantum efficiency can be higher. The inserted waveguide layer helps distribute the optical mode in order to reduce the divergence of the laser beam. In addition, as the waveguide layer is nominally undoped, free-carrier absorption is reduced.

Unfortunately, the confinement factor is rather small for the active region (1~5 %). Despite this apparent disadvantage, for a low modal threshold gain, defined as:

Eq. (18.54) $G_{th}=\Gamma g_{th}$,

the SCH can still be designed to reach threshold significantly earlier than a double heterostructure, as shown schematically in Fig. 18.25.

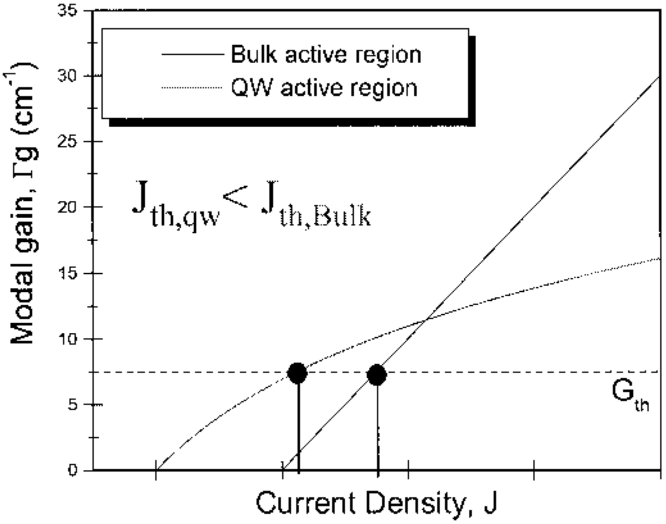


Fig. 18.25. Schematic comparison of QW and DH gain behavior as a function of current density. For a low threshold modal gain, the SCH can have a significantly lower threshold current density, thanks partially to the reduced density of states in the active region.

Another benefit of using thin and/or QW active regions is the possibility to realize strained layers. Indeed, unlike the DH laser where all materials needed to be lattice-matched, the SCH can incorporate strained materials (such as $In_xGa_{1-x}As$ on GaAs) as long as the layer is thinner than the critical thickness above which threading dislocations start to form. Strained layers can have two important benefits. The first improvement has to do with the band structure of a strained semiconductor. For compressively strained material, the heavy-hole mass becomes lighter than bulk, which leads to a reduced hole density of states, and further reduction of the transparency current density.

Finally, the accessible wavelength range for a given substrate is increased. For example, a DH laser based on GaAs has a maximum emission wavelength of 870 nm. Using strained quantum wells, a SCH laser can extend the emission wavelength well past 1 μm.

18.5.8. Laser packaging

Regardless of the fabrication procedure, the packaging of the laser diode depends on the final application.

The simplest package consists of an open heat sink, similar to the type used in the broad area laser fabrication (Fig. 18.21(f)). These come in many different designs, depending on the size restrictions in the final product. One example of an open package is shown in Fig. 18.26(a). Another popular product, which also protects the laser, is a transistor-type can with an output window, as shown in Fig. 18.26(b). These typically are sold in 5 and 9 mm diameter sizes as well as a larger TO-3 package. For high power lasers or laser bar packaging, heat management is very important. In this case, a larger mass submodule is used, which frequently incorporates a thermoelectric cooler (TEC) and controller to keep the temperature stable. A representative high heat load package, is shown in Fig. 18.26(c). Other applications benefit from having the laser coupled directly to a fiber optic cable. In this case, the fiber is carefully aligned and welded into place inside a hermetically sealed package, as shown in Fig. 18.26(d).

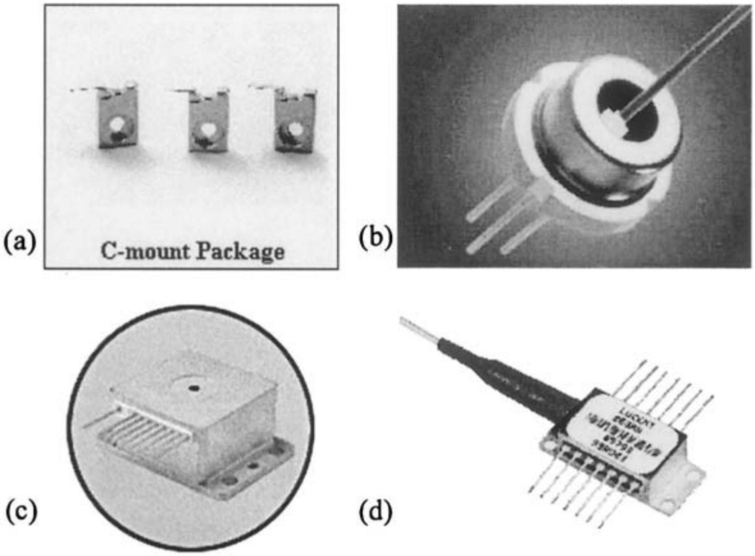


Fig. 18.26. Examples of commercial laser package designs. (a) open "c-mount" type heat sinks; (b) 9 mm transistor can package; (c) high heat load package incorporating laser and cooler; (d) fiber-coupled "butterfly" package.

18.5.9. Distributed feedback lasers

Semiconductor lasers typically exhibit multiple wavelength emission at high current because of the presence of a Fabry-Perot cavity, as illustrated in Fig. 18.27(b). This occurs when multiple longitudinal modes reach the laser threshold gain. When only one wavelength is desired, a common technique is to realize a secondary feedback within the cavity. A corrugated grating positioned inside the waveguide is one means to this end, as shown in Fig. 18.27(a). This type of laser is referred to as a distributed feedback (DFB) laser, because the feedback effect is distributed over most, if not all, of the laser cavity.

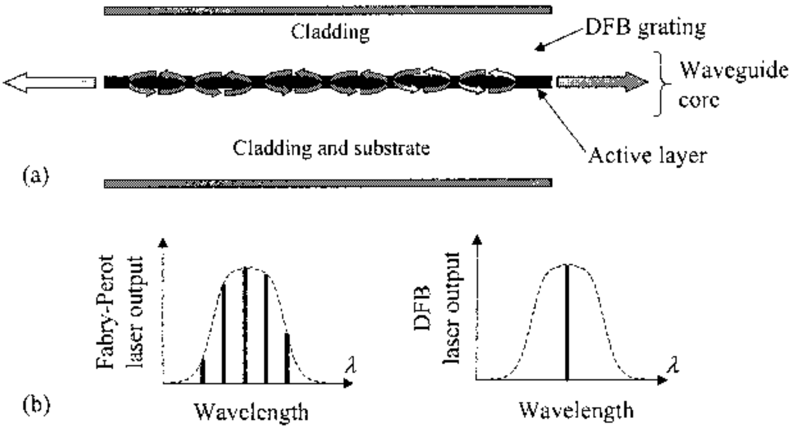


Fig. 18.27. (a) Schematic of the cross-section of a DFB laser. Periodic variations in the effective refractive index leads to distributed optical feedback; (b) proper grating design can change the output spectrum of a laser structure from multi- to single-wavelength.

The grating geometrical parameters are chosen to satisfy Bragg's law of diffraction:

$$\text{Eq. (18.55)} \quad m\frac{\lambda_0}{\overline{n}_{eff}} = 2\Lambda\sin(\theta)$$

where m is the grating order, λ_0 is the free space wavelength, $\overline{n}_{eff}$ is the effective refractive index in the waveguide, Λ is the grating period, and θ is the diffraction angle. The minimum requirement for optical feedback is diffraction at 180° relative to the propagation of the waveguide mode, i.e. the diffraction angle is θ=90°. For a first order grating (m=1), Bragg's law is fulfilled only when:

Eq. (18.56) $\Lambda = \dfrac{\lambda_0}{2\bar{n}_{eff}}$

The diffraction of light by the grating serves to enhance the cavity reflectivity at the designed wavelength. The threshold gain is then reduced for this wavelength, allowing favored laser oscillation of a specific longitudinal mode, in some cases (Fig. 18.27(b)).

Unfortunately, for many near-infrared lasers, the period of a first order grating is in the 100~300 nm range, a resolution which requires a high-end lithography setup. To make fabrication easier, second, third, and fourth order gratings have also been explored for this application, though their efficiency is somewhat lower due to lower order diffraction loss.

Despite this technological difficulty, DFB lasers are frequently incorporated into buried-heterostructure designs in order to demonstrate true single mode (longitudinal and transverse) behavior. This is especially useful in telecommunications and chemical spectroscopy, where a stable, monochromatic laser (single longitudinal mode wavelength) is preferred.

18.5.10. Material choices for common interband lasers

Table 18.1 summarizes some of the most common III-V material systems used in various parts of a laser structure in order to achieve a specific range of laser emission. The parts of the laser considered include the substrate, cladding and active region materials.

Substrate	Cladding material	Active Material	Wavelength Range
GaN(*)	$Al_xGa_{1-x}N$	strained $In_xGa_{1-x}N$ QWs	400~600 nm
GaAs	$Al_xGa_yIn_{1-x-y}P$	$Ga_{0.51}In_{0.49}P$	660 nm
GaAs	$Al_xGa_{1-x}As$ or $Ga_{0.51}In_{0.49}P$	$Ga_xIn_{1-x}As_yP_{1-y}$	660~870 nm
GaAs	$Al_xGa_{1-x}As$ or $Ga_{0.51}In_{0.49}P$	strained $In_xGa_{1-x}As$ QWs	0.87~1.1 μm
InP	$Al_{0.48}In_{0.52}As$	InP	920 nm
InP	InP	$Ga_xIn_{1-x}As_yP_{1-y}$	1~1.7 μm
InP	InP	$Ga_{0.47}In_{0.53}As$ QWs	1.3~1.7 μm
InP	InP	strained $In_xGa_{1-x}As$ QWs	1.5~2 μm
InAs	$Al_xGa_yIn_{1-x-y}As_zSb_{1-z}$ or $InAs_xSb_yP_{1-x-y}$	InAs QW	2~3 μm
InAs	$Al_xGa_yIn_{1-x-y}As_zSb_{1-z}$ or $InAs_xSb_yP_{1-x-y}$	strained $InAs_ySb_{1-y}$ QWs	3~5 μm

Table 18.1. Materials for substrates, cladding and active regions in semiconductor lasers for various emission wavelength ranges. (*) Not a mature substrate technology. Lasers are often grown on sapphire or SiC substrates.

18.5.11. Interband lasers

GaAs-based lasers in general have found many applications from simple laser pointers to CD/DVD players. In addition, significant development has also been invested in achieving very high power efficiency and output powers from these devices for applications such as welding/cutting, frequency doubling, and solid-state laser pumping.

The last application especially has drawn a lot of interest, as diode-pumped laser systems show significantly higher efficiency and lifetime compared to conventional flashlamp-pumped systems. The primary reason for this is that the diode laser emission wavelength can be closely matched to the narrow absorption peak of the solid-state laser medium. This allows most of the emitted radiation from the laser diode to be directly absorbed. Flashlamps, on the other hand, are broadband sources, and a large fraction of the light goes directly to waste heat. Further, flashlamps have a limited lifetime, on the order of five hundred to several thousand hours. On the other hand, the durability of aluminum-free diode lasers was challenged by operating them under continuous wave at 1 W, 60 °C for an extended period of time. They exhibited no degradation over 30,000 hours [Diaz *et al.* 1996] [Diaz *et al.* 1997]. Under normal operating conditions (20 °C) the projected lifetimes are on the order of several million hours.

The aluminum-free GaInAsP technology was used to achieve high performance semiconductor lasers emitting at 980 nm [Mobarhan *et al.* 1992] through the use of strained InGaAs quantum wells inside a separate confinement heterostructure (SCH). These lasers exhibited a low J_{th}~70 A/cm^2, high differential efficiency ~1.0 W/A, and low internal loss of 1.5 cm^{-1}. They also yielded high output power and a very high characteristic temperature T_0, over 350 K in the range of 20~40 °C. These 980 nm lasers could operate under continuous wave at high power (1.4 W) at 100 °C.

Optimized 808 nm laser diodes based on GaAs/GaInAsP with uncoated facets emitted high output powers of 10 W and 7 W in pulse and continuous wave operation, respectively. Laser bars yielded output powers of 70 W in quasi-continuous wave operation [Razeghi 1994] [Yi *et al.* 1995]. A properly designed GaInAsP laser structure provides a narrow transverse beam with a divergence of only 26°, which is convenient for efficient laser light coupling into the optical fiber or the pumped Nd:YAG crystal. For comparison, 32-48° are typical values of beam divergence for commercial AlGaAs lasers.

InP based laser have also received a lot of attention thanks to the accessible wavelength range of lattice-matched quaternaries (0.92~1.65 μm). This range makes the system ideal for fiber optic communications, which relies on lasers designed for low-loss and low-dispersion fiber propagation.

Before these lasers could be considered for use in any application, however, several demonstrations had to be made. These initial demonstrations include:

- Double heterostructure λ=1.3 μm and 1.55 μm lasers with threshold current densities of 430 and 500 A/cm^2 respectively [Razeghi *et al.* 1983a] [Razeghi *et al.* 1983b].
- Buried ridge 1.3 and 1.55 μm lasers with threshold currents as low as 6 mA [Razeghi 1985a].
- Buried ridge lasers with distributed feedback for single wavelength emission (Fig. 18.28) [Razeghi *et al.* 1985b] [Razeghi *et al.* 1985c].
- High power phase-locked arrays of 1.3 μm lasers with 600 mW output [Razeghi 1987].

- Separate confinement heterostructure GaInAsP/GaInAs/InP waveguide for improved performance [Razeghi *et al.* 1985c].

Many other groups took advantage of this technology to produce advanced lasers for telecommunications. [Kuznetsov *et al.* 1989] With the groundwork laid by these achievements, GaInAsP/InP lasers are now in mass production for the telecommunication industry. There is still active research to extend the wavelength to 2 μm and beyond using strained layer epitaxy.

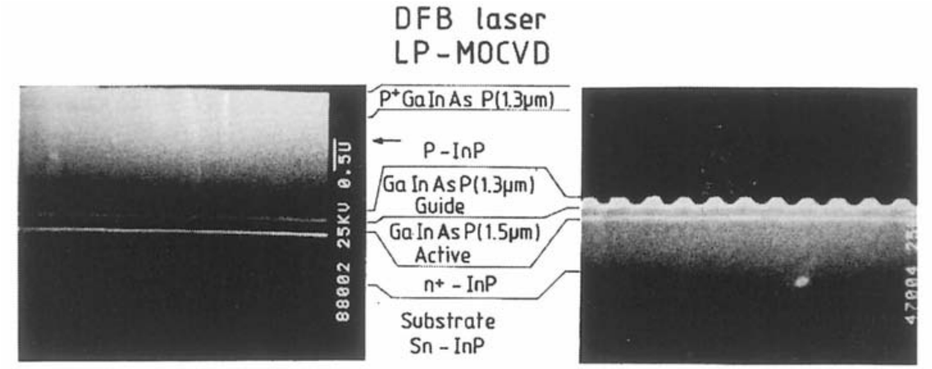


Fig. 18.28. Scanning electron microscope cross-sections of the corrugated structures after and before LP-MOCVD regrowth..

In general, the quest for longer wavelength (λ>2 μm) diode lasers has been fought with difficulty. There are many applications for longer wavelength lasers, including infrared spectroscopy, infrared countermeasures, free-space communication, and low-loss fiber communication. The only commercially available semiconductor alternative, up until a few years ago, were actually based on IV-VI materials, commonly referred to as lead-salt lasers. This technology has several liabilities, not the least of which are low operating temperature (80 K) and low output power (~1 mW).

Clearly, for most applications it is advantageous to overcome one or both of these liabilities. Unfortunately, increased losses and reduced radiative efficiency are a fact of life for low bandgap zinc-blende semiconductors as well. Despite this challenge, extensive work has been done on InAs-based laser diodes at cryogenic temperatures. This technology has the potential for significantly higher output power powers and is based on an intrinsically more robust material system.

Double heterostructure (DH) lasers based on the InAs/InAsSb/InAsSbP material system were grown by low-pressure Metal-Organic Chemical Vapor Deposition (LP-MOCVD) for high power lasers emitting with

3.1 ≤λ≤3.4 μm. A novel, asymmetric, heterostructure was employed, as shown in Fig. 18.29, which allowed for reduced electron leakage and improved performance. Laser bars consisting of four 100 μm stripes exhibited a maximum peak output power of 6.7 W [Wu *et al.* 1999]. This registers the highest output power in this wavelength range. Furthermore, optimizing the efficiency of these lasers allowed over 450 mW to be obtained under continuous operation [Razeghi 1998] [Razeghi *et al.* 1999].

The first mid-infrared superlattice injection lasers have been designed with InAs/InAsSb, InAsP/InAsSb, and InAsSb/InAsSb superlattices. Although complex, SLS lasers benefit from better optical confinement and more emission wavelength flexibility than a DH laser and a larger gain region than a MQW laser. Another advantage that these superlattice lasers have is the reduction of non-radiative recombination (Auger) mechanisms. The first lasers based on the superlattice active region were designed and fabricated for emission from 4~4.8 μm [Lane *et al.* 1999] [Lane *et al.* 2000]. This is the longest reported emission from electrically injected lasers based on InAsSb interband transitions.

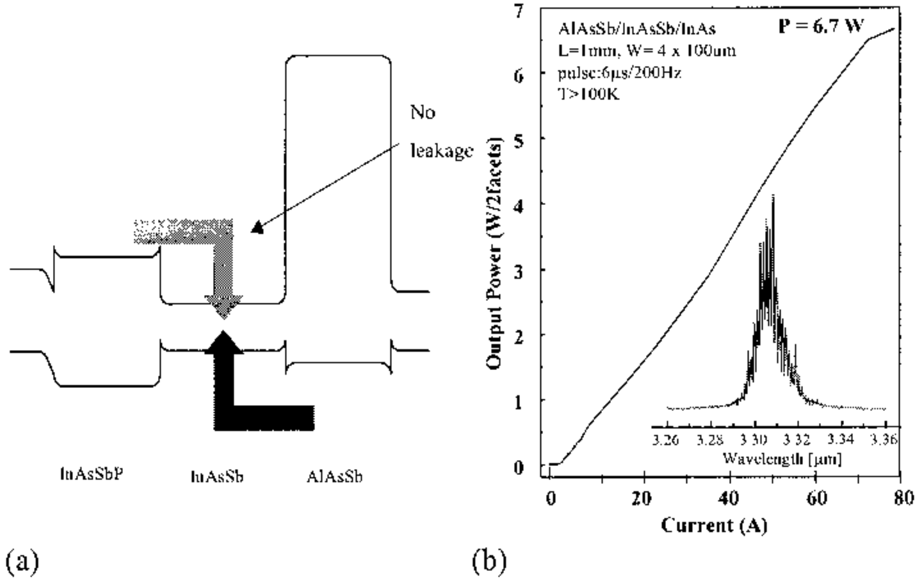


Fig. 18.29. (a) InAsSbP/InAsSb/AlAsSb double heterostructure. (b) P-I curve of InAsSbP/InAsSb/AlAsSb laser bar at T> 80 K.

18.5.12. Quantum cascade lasers

In conventional (interband) semiconductor lasers, light is generated from the recombination of electrons and holes separated by the energy gap. The energy of the photons, or the wavelength of laser light is therefore essentially determined by the energy gap of the semiconductors used. This means, however, that only a certain range of wavelengths can be obtained, which correspond to the energy gap of semiconductors that exist. It turns out that it is quite difficult to have low energy gap semiconductors (E_g<0.3 eV) with high enough quality suitable for semiconductor lasers, and thus it is difficult to make lasers with low photon energy which corresponds to infrared wavelengths (λ>3 μm). Because of this limitation, totally different types of semiconductor lasers have been recently fabricated. The first one is the quantum cascade laser, and the other is the type II superlattice laser which will be discussed in the next sub-section. Here, we will briefly describe the operating principle and underlying basic physics of these lasers.

Unlike the interband lasers which have bipolar device characteristics, relying on electron-hole recombination to generate light, the quantum cascade laser (QCL) is a unipolar device which uses only conduction-band electrons. This means that the electron is the dominant carrier and that emitted light is solely due to electron transitions from upper to lower subbands in the conduction band (semiconductor quantum wells have the property of splitting a bulk band structure into a series of subbands). The benefit of this design is that, using the same material, the emission energy can be varied over a wide range (limited by the conduction band offset between the barrier and well) simply by varying the quantum well widths.

The idea of generating light from intersubband transitions within semiconductor quantum wells was first proposed in 1971 by Kazarinov and Suris [1971]. Since then, many groups have tried to produce devices based on similar models. The first electrically pumped intersubband laser at λ =4.2 μm was demonstrated in 1994 by Faist *et al.* [1994].

In order to have stimulated emission, there has to be a population inversion. This means that there must be a steady injection of high-energy electrons and steady collection of low energy electrons from the same quantum well region. The principle of operation is based on a multi-quantum well structure in an electric field and is illustrated in Fig. 18.30.

Electrons come to the active region from the injector (left) and go out to a subsequent region (right) which acts as the injector of another active region. All these regions are made from multi-quantum wells. The combination of active+injector regions forms a period of the structure which repeats many times (typically 25~30) forming a cascade structure. This allows a single injected electron to emit multiple photons, which leads to differential efficiencies much greater than unity.

The active region of a typical QCL is a triple AlInAs/GaInAs quantum well structure that supports three subbands as illustrated above. The injector region is made up of the same material and has many wells depending on the emission wavelength. The QCL emission wavelength is controlled solely by the layer thicknesses in the active region, and thus has the potential for a large variety of emission wavelengths in a given material system.

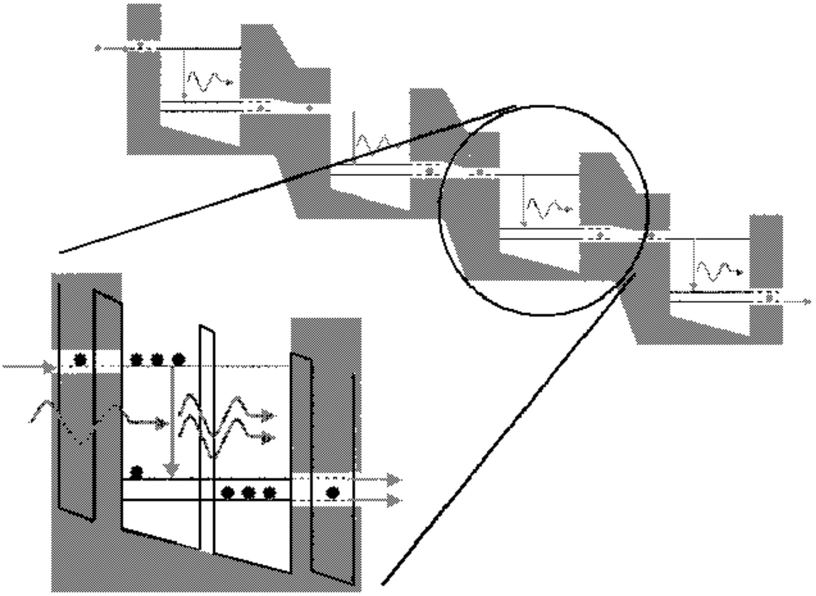


Fig. 18.30. Example of quantum well-based active region for an electrically injected quantum cascade laser. Electrons are injected into the higher energy level in the quantum well, from where they can be stimulated to relax down to the lower energy level and emit a photon. The relaxed electrons can then be transmitted to the next quantum well on the right. This process can be continued by serially stacking active regions for higher power output.

A real example of QCL designed for λ=8.5 μm is shown in Fig. 18.31. Due to quantum mechanical selection rules, the light emitted from a QCL is of TM polarization, i.e. the electric field strength in the wave is perpendicular to quantum well plane.

The most important advantage of the quantum cascade laser is its insensitivity to temperature changes. Laser operation is not limited by Auger recombination due to the unipolar nature of the device, which gives a much higher theoretical operating temperature. This high temperature operation will potentially reduce package size and cost due to the absence of elaborate cooling systems. Another significant advantage of QCL is that, as the operation is not directly related to the bandgap of the constituent materials, relatively mature InP or GaAs technology can be used. This ensures that the growth and the physical characteristics of the materials are relatively well understood for QCLs. It also allows for excellent uniformity across 2" wafers, which has already been demonstrated in laboratory situations.

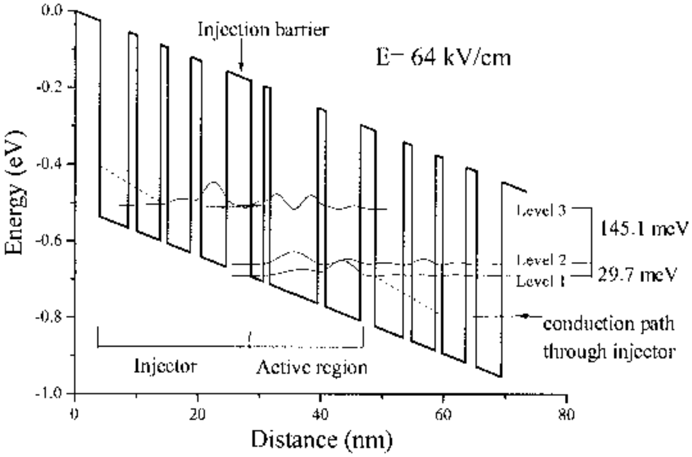


Fig. 18.31. Conduction band profiles for quantum cascade active region surrounded by two injectors. Wavy lines are electron wavefunctions corresponding to levels 1, 2, 3.

Many important milestones have been made in quantum cascade laser technology. The total demonstrated wavelength range of operation spans from 3.4-160 μm at cryogenic temperatures. At room temperature, this range is still impressive at 3.6-16 μm. No interband laser technology has come close to this range at room temperature. High power, thanks to the cascade nature of the device, is also a hallmark of this technology. At λ=9 μm, over 3.5 W of power per facet has been demonstrated at room temperature in pulsed mode [Slivken *et al.* 2002].

Still, due to relative inefficiency of the intersubband process compared to near-infrared interband lasers, it has been very difficult to achieve continuous operation at room temperature. With a room temperature threshold current density of >1 kA/cm^2 and an operating voltage on the order of 10 V, it is difficult to efficiently remove all the heat from the waveguide core.

Many design and laser geometry improvements were necessary, but quantum cascade lasers have been demonstrated with >100 mW of continuous power at room temperature at many wavelengths. [Evans *et al.* 2004] [Razeghi *et al.* 2005] Even single-mode, DFB quantum cascade lasers have been demonstrated with similar power output and 30 dB side mode suppression [Yu *et al.* 2005].

Lastly, another nice property of intersubband lasers is the ability to emit multiple wavelengths from the same laser core. Unlike interband lasers, there is no intrinsic absorption around the laser emission energy. Besides developing high power, room temperature laser sources, future development of this technology is likely to include demonstration of broadband, tunable laser sources, which can eventually replace thermal sources for infrared spectroscopy.

18.5.13. Type II lasers

Type II band alignment and some of its interesting physical behavior were originally suggested by Sai-Halasz *et al.* [1977]. Soon after that they reported the optical absorption of type II superlattices [Sai-Halasz *et al.* 1978a] and later the semimetal behavior of the superlattice [Sai-Halasz *et al.* 1978b]. The applications of such a superlattice was proposed only after several years [Smith *et al.* 1987]. The flexibility of the material used to cover a huge infrared range (2 to >50μm) and the reduced Auger recombination rate [Youngdale *et al.* 1994] caught the attention of many research groups. Type II heterojunctions have found many applications in electronic devices such as resonant tunneling diodes and hot electron transistors. However, perhaps the most important of these applications has been in the optoelectronics and recently many significant results have been achieved in type II modulator [Xie *et al.* 1994], detectors [Johnson *et al.* 1996] [Fuchs *et al.* 1997], and laser diodes [Yang *et al.* 1998] [Felix *et al.* 1997].

Type II lasers are based on active layers which exhibit a type II band alignment. In general, the band alignment of any semiconductor heterojunction can be categorized as type I, type II staggered or type II misaligned, as illustrated in Fig. 18.32. The main difference between type I and type II staggered band alignments resides in the fact that, in the former case, one (same) side of the heterojunction presents a lower energy for both electrons and holes. The electrons and the holes will thus be preferentially found on the same side of the junction. If the bottom of the conduction band of one material is located at a lower energy than the top of the valence band of the other one, we obtain a type II misaligned heterojunction, as shown in Fig. 18.32.

The type of heterojunction depends on the nature of the semiconductors brought in contact. Most semiconductor junctions are of the type I. The special band alignment of the type II heterojunctions provides three important physical phenomena or features which are illustrated in Fig. 18.33: (a)a lower effective bandgap, (b) the separation of electrons and holes and (c) tunneling. These properties are used in many devices to improve their overall performance.

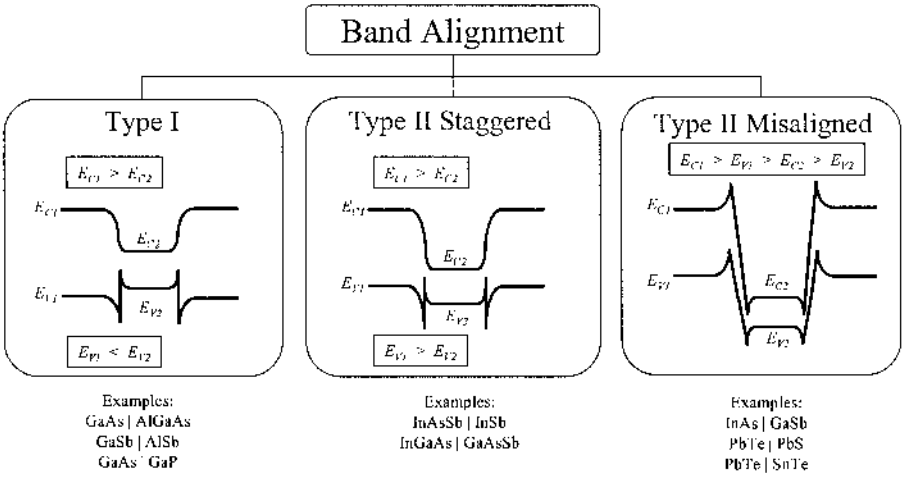


Fig. 18.32. Possible band alignment configurations of a semiconductor heterojunction. The most common band alignment is the type I alignment shown on the left in which one same side of a heterojunction presents a lower energy for both electrons and holes. The type II band alignment is shown in the middle and the right, in which the electrons and holes have a lower energy on different sides of a heterojunction.

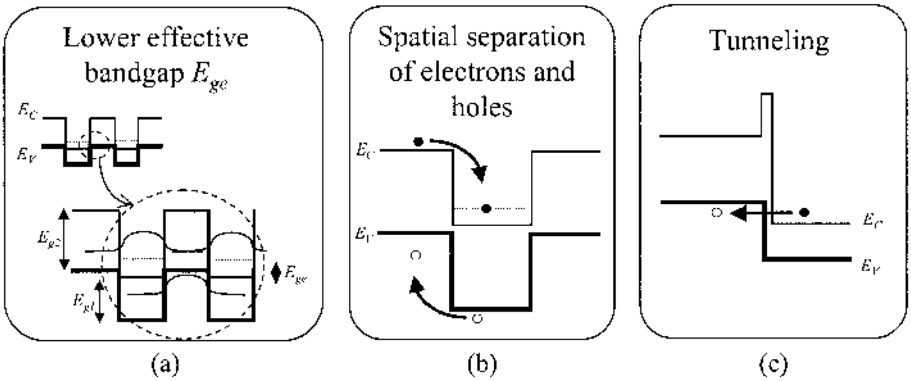

Lower effective bandgap E_{ge}
E_C
E_V
E_{g2}
E_{g1}
E_{ge}
Spatial separation of electrons and holes
E_C
E_V
Tunneling
E_C
E_V
(a)
(b)
(c)

Fig. 18.33. Unique features of type II heterojunctions and superlattices: (a) they have an effective bandgap energy which is lower than those of the constituting semiconductors, (b) electrons and holes are spatially separated, (c) an electron can tunnel from the conduction band on one side of the heterojunction into the valence band on the other side of the heterojunction.

The first feature involves a superlattice with type II band structure. A superlattice is a structure consisting of closely spaced quantum wells, such that the localized discrete energy levels in the quantum wells become delocalized minibands across the entire structure, in both the conduction and the valence bands. These minibands thus form an effective bandgap E_{ge} for the superlattice considered as a whole. Because of the type II band alignment, these minibands can exhibit a lower effective bandgap than the bandgap of either constituting layer, as shown in Fig. 18.33(a), which is a most interesting property. In addition, this effective bandgap is tunable to some extent by changing the thickness of the layers. This is very important for the applications in the mid- and long infrared wavelength range since one can generate an artificial material (the superlattice) with a fixed lattice parameter but with a different bandgap. For example, recently, very successful detectors [Mohseni *et al.* 1997] and lasers [Felix *et al.* 1997] have been implemented in the 2 to 15 μm wavelength range with InAs/InGaSb superlattices.

The second feature is the spatial separation of the electrons and holes in a type II heterojunction, as shown in Fig. 18.33(b). This phenomenon is a unique feature of this type of band alignment and is due to the separation of the electron and hole potential wells. As a result, a huge internal electrical field exists in the junction without any doping or hydrostatic pressure. High performance optical modulators have been implemented based on this feature [Johnson *et al.* 1996].

The third feature is the Zener type tunneling of a type II misaligned heterojunction, as shown in Fig. 18.33(c). Electrons can easily tunnel from the conduction band of one layer to the valence band of the other layer, since the energy of the conduction band of the former layer is less than the energy of the valence band of the later layer. However, unlike Zener tunneling, no doping is necessary for such a junction. Therefore, even a semi-metal layer can be implemented with very high electron and hole mobility since the impurity and ion scattering are very low. This feature of type II heterojunctions has been successfully used for resonant tunneling diodes and, recently, for type II unipolar lasers [Lin *et al.* 1997].

Type II lasers can generally be categorized as bipolar and unipolar lasers. In the bipolar type II lasers, the structure of a conventional III-V heterostructure laser diode is used except that the active layer is a type II superlattice, as shown in Fig. 18.34. The electron-hole recombination occurs between the first conduction miniband and the first heavy-hole miniband. Since the active layer is a type II superlattice, the energy difference between the minibands can be adjusted by changing the thickness of the layers. This is an important advantage since the laser can potentially cover a wide range of wavelengths (from ~2 μm to beyond 50 μm) without changing the chemical composition of the materials in different layers.

In a unipolar type II laser, electrons are injected into the active layer through an injection layer, similar to a quantum cascade laser. They then radiatively recombine with a hole, unlike a quantum cascade laser. The electrons can also tunnel through a type II tunneling junction, and go to the next injection layer and repeat a similar transition in a cascade fashion. The interband transition in these lasers leads to an important advantage compared to the quantum cascade lasers. This type of transition is much more immune to the phonon scattering than the intersubband transition, and hence the efficiency of unipolar type II lasers is much higher than that of quantum cascade lasers.

It should be noted that type II lasers are interband lasers, meaning that it involves the radiative recombination of an electron from the conduction band with a hole in the valence band. In such types of lasers, Auger recombination (sub-section 8.6.4) can generally play an important role and limit the high temperature operation of such infrared lasers. Fortunately in the case of a bipolar type II laser, the spatial separation of electrons and holes and the band structure of the superlattices lead to a much lower Auger recombination rate than in conventional semiconductor infrared lasers. Recently, bipolar type II lasers yielded the highest operating temperature for light emission at 3.2 μm.

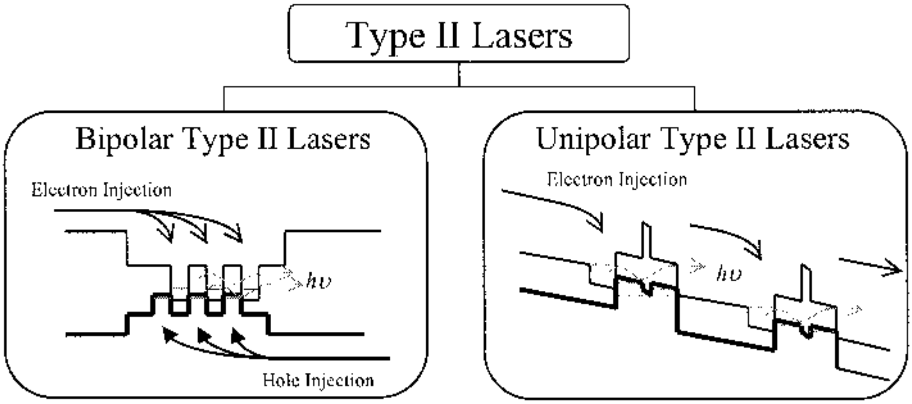


Fig. 18.34. Bipolar and unipolar type II lasers. In a bipolar laser, the electrons and holes are injected and recombine radiatively in the quantum well or superlattice type II structures. In the unipolar type II laser, electrons are injected and emit photons by relaxing into a lower energy state in the heterostructure.

18.5.14. Vertical cavity surface emitting lasers

In all the lasers discussed so far the light is emitted from the edges of active region. There has always been a desire for surface-emitting lasers which would allow free-space communication from chip to chip between specific locations and 2D arrays of high-power light sources. Earlier attempts have introduced mirrors or gratings to turn the edge-emitted light by 90°, but recently major advances have been made by using GaAs-AlGaAs or GaAs-AlAs multilayers (grown at the same time as the laser structure) as integral mirrors.

The multilayer mirrors, known as Bragg reflectors, reflect light at certain wavelengths, in particular that of the laser. The GaAs and AlGaAs (or AlAs) layers forming the mirrors are each typically a quarter-wavelength thick (Fig. 18.35).

The structure is shown in Fig. 18.35, together with a photograph of an array of such devices. It is straightforward to place ohmic contacts on the top and the bottom, and there is no need for cleaving and polishing end facets. The three main features that determine their operation are (i) the high reflectivity of the mirrors, (ii) the means of lateral confinement of the active region, and (iii) the limitations imposed by laser heating from carrier injection through the resistive mirrors.

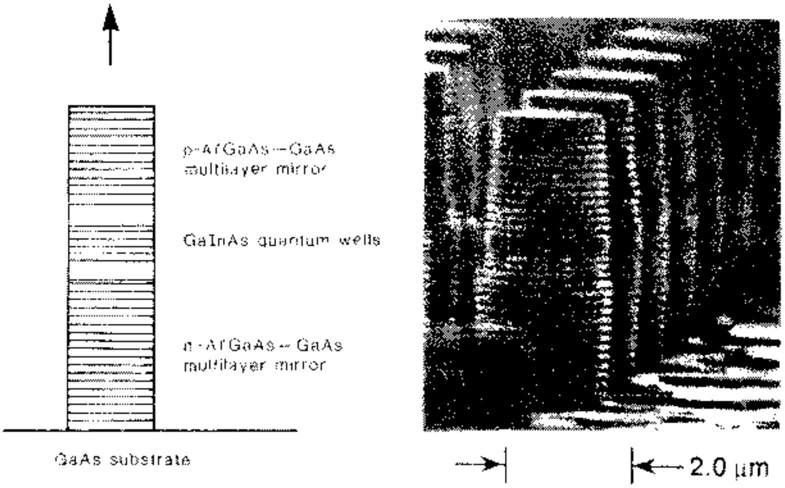


Fig. 18.35. Example of a multilayer structure for a surface-emitting laser: GaInAs quantum wells are sandwiched between two GaAs-AlAs multilayer mirrors, one n-type doped and the other p-type doped. ["Figure 18.10", from LOW-DIMENSIONAL SEMICONDUCTORS: MATERIALS, PHYSICS, TECHNOLOGY, DEVICES by M.J. Kelly; taken after Applied Physics Letters Vol. 55, Scherer, A., Jewell, J.L., Lee, Y.H., Harbison, J.P., and Florez, L.T., "Fabrication of microlasers and microresonator optical switches," p. 2724-2726. Copyright 1989, American Institute of Physics. Reprinted with permission of Oxford University Press, Inc. and American Institute of Physics.]

Surface-emitting lasers have comparable losses and gains to edge-emitting lasers but are much more compact. The need to confine the current to achieve a narrow beam presents problems too. Lateral p-n junctions may be used to confine the current. Moreover, the current flowing to the cavity can heat up the mirrors through which it passes, which can lead to severe heating during continuous-wave operation. At present the devices are about 10 % efficient in electrical to optical power conversion.

The surface-emitting geometry has many attractions for future systems, including a smaller, circular beam divergence making it easier to couple with optical fibers and integrate with other optoelectronic components.

18.5.15. Low-dimensional lasers

Further improvement in the performance of semiconductor lasers is expected by using higher degrees of quantum confinement, such as the quantum wires and quantum dots discussed in Chapter 11. One of the driving forces toward such structures is to achieve a "threshold-less" semiconductor laser, i.e. which can reach lasing threshold with minimal electrical current.

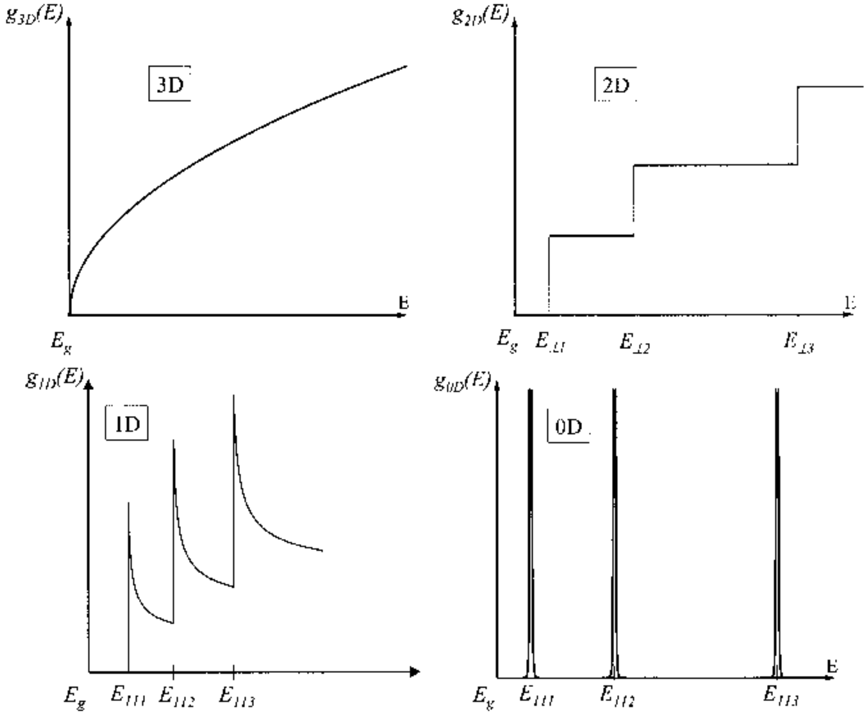


Fig. 18.36. Density of states versus dimensionality: in a bulk (3D) semiconductor crystal, a quantum well (2D), a quantum wire (1D) and a quantum dot (0D).

Fig. 18.36 compares the density of states in bulk crystals (3D), quantum wells (2D), quantum wires (1D), and quantum dots (0D). The very narrow density of states distribution in lower dimensional structures achieves a narrower energy distribution for carriers than in bulk crystals, which results in narrower luminescence spectra, higher differential gain, lower threshold current density and wider modulation bandwidth in lasers using such structures.

Another important feature in quantum dots is the dispersion-less behavior of its electronic states. In other words, the allowed energy levels are independent of the momentum and have a constant energy, which makes it possible to avoid Auger recombination, as long as the positions of the energy levels are positioned adequately. This is illustrated in Fig. 18.37 which compares the Auger recombination process in a bulk semiconductor with that in a quantum dot. If we pretend that the discrete energy levels of the quantum dot lie on a pseudo momentum curve as shown (of course in reality there is no momentum in a quantum dot), we see that, in a quantum dot, the energy and "momentum" conservation cannot be achieved simultaneously during an Auger process. In other words it is difficult for the band to band transition to exactly match an intersubband transition. This ensures that Auger processes in a quantum dots are less likely, which is an important property because a high Auger recombination rate is the major obstacle for the high operating temperature of infrared interband lasers.

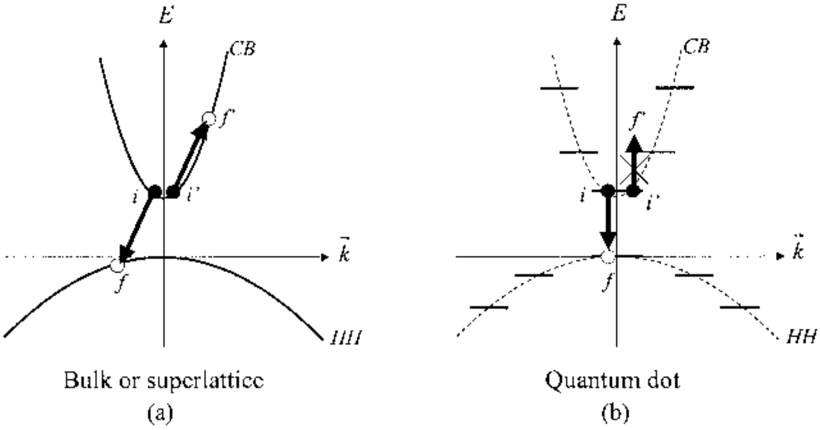


Fig. 18.37. Illustration of the Auger process in a (a) bulk semiconductor or (b) superlattice in comparison with a quantum dot. In the former case, there is a continuum of available states where an Auger electron can be excited into. This makes the conservation of the total "momentum" and energy possible to be achieved in an Auger process. In a quantum dot, there are only discrete energy levels allowed, as a result of the shape of the density of states. Many transitions are therefore forbidden since the" momentum" and energy conservation laws can be satisfied simultaneously only for a few states.

The use of lower dimensional structures has advantages for QCLs. Indeed, in quantum wires and dots, the LO phonon scattering rates can be considerably lower than in quantum wells. The main reason behind this property is the fact that the scattering rate between two energy bands is proportional to the overlap of the density of states of these bands. Fig. 18.38

shows that such overlap is smaller in a quantum wire and can even be non-existent in a quantum dot. A quantum dot based QCL can therefore have an excellent efficiency through the reduction of phonon scattering rates.

Fabrication of lower dimensional structures is in general extremely difficult. Currently, one of the main targets of atomic engineering research is indeed the fabrication of quantum structures and surfaces with quantum features. The average quantum dot is only 10~50 nm in diameter. Sophisticated growth and/or fabrication techniques are required to produce uniform features in this size regime.

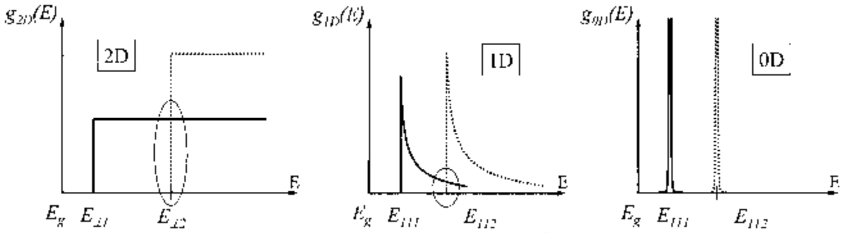


Fig. 18.38. Illustration of the overlap in the density of states for a quantum well (2D), quantum wire (1D), and quantum dot (0D).

18.5.16. Raman lasers

The Raman effect occurs when incident monochromatic light hits the material and a photon is generated with an energy that is different from the incident photon by the energy of a phonon. Specifically, the interaction of an incident photon with the material leads to inelastic scattering where the photon-phonon interactions cause the generation of a photon, which has an energy that is exactly one phonon energy higher or lower than the incident photon. In fact, once there is significant net Raman gain, or photon generation, achieved in a material it is possible to observe lasing action from that material. This effect can be exploited to make lasers from indirect bandgap materials such as silicon, where lasing can be achieved by optically pumping a silicon waveguide. This kind of Raman lasing was first demonstrated in silicon using pulsed lasers in 2004 [Boyraz *et al.* 2004] and with continuous-wave lasers in 2005 [Rong *et al.* 2005]. The development of Raman lasing in silicon is considered to be an important milestone in the development of silicon based optoelectronic devices, since silicon is already the most widely used semiconductor in electronic circuits.

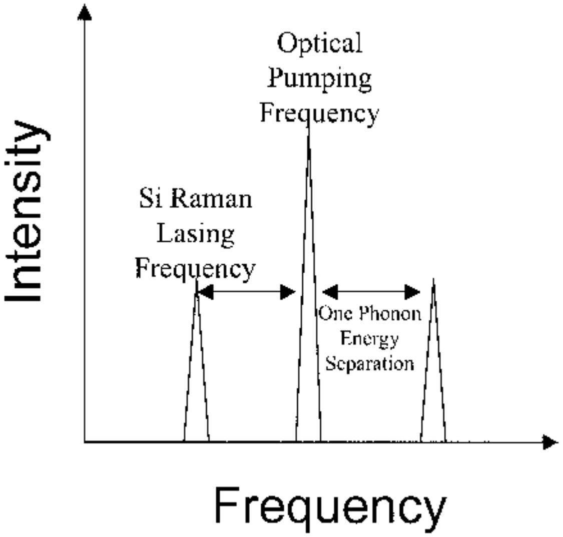

Optical Pumping Frequency
Si Raman Lasing Frequency
One Phonon Energy Separation
Intensity
Frequency

Fig. 18.39. The Raman effect and observed lasing in an optically-pumped silicon waveguide. The separation between the central peak and Raman scattered peaks is the frequency of optical phonons in silicon.

18.6. Summary

In this Chapter, we reviewed the fundamental physical concepts relevant to lasers, including stimulated emission, resonant cavity, waveguide, propagation of an electromagnetic wave in a waveguide and the laser beam divergence, and waveguide design. We introduced the notion of absorption, spontaneous and stimulated emission, the Einstein coefficients, resonant cavity, population inversion and threshold. The example of the ruby laser was used to illustrate these concepts.

The discussion was then focused on semiconductor lasers, which are becoming dominant for numerous modern applications. The concepts of gain, threshold current density, transparency current density, linewidth, external differential quantum efficiency, mirror loss and internal loss were introduced. The different types of semiconductor lasers were then described, including homojunction, single and double heterojunction, and separate confinement and quantum well lasers. The fabrication and packaging technology of semiconductor lasers were then briefly described. Finally, a few specific examples of lasers were presented, including quantum cascade, type II, vertical cavity surface emitting lasers, low-dimensional lasers, and Raman lasers.

References

Casey Jr., H.C., and Panish M.B., *Heterostructure Lasers, parts A & B*, Academic Press, New York, 1978.

Chuang, S.L., *Physics of Optoelectronic Devices*, John Wiley & Sons, New York, p. 497, 1995.

Diaz, J., Yi, H.J., Kim, S., Wang, L.J., and Razeghi, M., "High Temperature Reliability of Aluminum-free 980 nm and 808 nm Laser Diodes," *Compound Semiconductors 1995* (*Institute of Physics Conference Series* 145), eds. J.C. Woo and Y.S. Park, Institute of Physics Publishing, Bristol, UK, pp. 1041-1046, 1996.

Diaz, J., Yi, H.J., and Razeghi, M., "Long-term reliability of Al-free InGaAsP/GaAs (λ=808nm) lasers at high-power high-temperature operation," *Applied Physics Letters* 71, pp. 3042-3044, 1997.

Evans, A., Yu, J.S., David, J., Doris, L., Mi, K., Slivken, S., and Razeghi, M., "High-temperature high-power continuous-wave operation of buried heterostructure quantum-cascade lasers," *Applied Physics Letters* 84, pp. 314-316, 2004.

Faist, J., Capasso, F., Sivco, D.L., Hutchinson, A.L., and Cho, A.Y., "Quantum cascade laser," *Science* 264, pp. 553-556, 1994.

Felix, C.L., Meyer, J.R., Vurgaftman, I., Lin, C.H., Murry, S.J., Zhang, D., and Pei, S.S., "High-temperature 4.5 μm type II quantum-well laser with Auger suppression," *IEEE Photonics Technology Letters* 9, pp. 734-736, 1997.

Fuchs, F., Weimer, U., Pletschen, W., Schmitz, J., Ahlswede, E., Walther, M., Wagner, J., and Koidl, P., "High performance InAs/$Ga_{1-x}In_xSb$ superlattice infrared photodiodes," *Applied Physics Letters* 71, pp. 3251-3253, 1997.

Henry, C.H., "Theory of the Linewidth of Semiconductor Lasers", *IEEE Journal of Quantum Electronics* 18, pp. 259-264, 1982.

Holonyak Jr., N. and Bevacqua, S.F., "Coherent (visible) light emission from $Ga(As_{1-x}P_x)$ junctions)," *Applied Physics Letters* 1, pp. 82-83, 1962.

Johnson, J.L., Samoska, L.A., Gossard, A.C., Merz, J., Jack, M.D., Chapman, G. R., Baumgratz, B.A., Kosai, K., and Johnson, S.M., "Electrical and optical properties of infrared photodiodes using the InAs/$Ga_{1-x}In_xSb$ superlattice in heterojunctions with GaSb," *Journal of Applied Physics* 80, pp. 1116-1127, 1996.

Kazarinov, R.F. and Suris, R.A., "Possibility of the amplification of electromagnetic waves in a semiconductor with a superlattice," *Soviet Physics Semiconductors* 5, pp. 707-709, 1971.

Kelly, M.J., *Low-Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Oxford University Press, New York, 1995.

Kuznetsov, M., Willner, A.E., Okaminow, I.P., "Frequency-modulation response of tunable 2-segment distributed feedback lasers," *Applied Physics Letters* 55, pp. 1826-1828, 1989.

Lane, B., Wu, A., Stein, A, Diaz, J., and Razeghi, M., "InAsSb InAsP strained-layer superlattice injection lasers operating at 4.0 μm grown by metal-organic chemical vapor deposition," *Applied Physics Letters* 74, pp. 3438-3440, 1999.

Lane, B., Tong, S., Diaz, J., Wu, Z., and Razeghi, M., "High power InAsSb/InAsSbP electrical injection laser diodes emitting between 3 and 5 μm," *Material Science and Engineering B* 74, pp. 52-55, 2000.

Lin, C.H., Yang, R.Q., Zhang, D., Murry, S.J., Pei, S.S., Allerman, A.A. and Kurtz, S.R., "Type II interband quantum cascade laser at 3.8 μm," *Electronics Letters* 33, pp. 598-599, 1997.

Maiman, T.H., "Stimulated Optical Radiation in Ruby," *Nature* 187, pp. 493-494, 1960.

Mobarhan, K., Razeghi, M., Marquebielle, G., Vassilaki, E., "High-Power 0.98 μm, $Ga_{0.8}In_{0.2}As/GaAs/Ga_{0.51}In_{0.49}P$ Multiple Quantum-Well Laser," *Journal of Applied. Physics* 72, pp. 4447-4448, 1992.

Mohseni, H., Michel, E., Sandven, J., Razeghi, M., Mitchel, W., and Brown, G., "Growth and characterization of InAs/GaSb photoconductors for long wavelength infrared range," *Applied Physics Letters* 71, pp. 1403-1405, 1997.

Boyraz, O. and Jalali, B., "Demonstration of a silicon Raman laser," *Optics Express* 12, pp. 5269-5273, 2004.

Razeghi, M., Hirtz, P., Blondeau, R., and Duchemin, J.P., "Aging Test of MOCVD Shallow Proton Stripe GaInAsP-InP, DH Laser Diode Emitting at 1.5 μm," *Electronics Letters* 19, p. 481, 1983a.

Razeghi, M., Hersee, S., Blondeau, R., Hirtz, P., and Duchemin, J.P., "Very Low Threshold GaInAsP/InP DH Lasers Grown by MOCVD," *Electronics Letters* 19, p. 336, 1983b.

Razeghi, M., in *Lightwave Technology for Communication*, ed. W.T. Tsang, Academic Press, New York, 1985a.

Razeghi, M., Blondeau, R., Boulay, J.C., de Cremoux, B., and Duchemin, J.P., "LP-MOCVD growth and cw operation of high quality SLM and DFB semiconductor $Ga_xIn_{1-x}As_yP_{1-y}$–InP lasers," in *GaAs and Related Compounds 1984* (*Institute of Physics Conference Series* 74), UK Adam Hilger, Bristol, UK, p. 451, 1985b.

Razeghi, M., Blondeau, R., Krakowski, M., Bouley, J.C., Papuchon, M., de Cremoux, B., and Duchemin, J.P. "Low-Threshold Distributed Feedback Lasers Fabricated on Material Grown Completely by LP-MOCVD," *IEEE Journal of Quantum Electronics* QE-21, pp. 507-511, 1985c.

Razeghi, M., "CW Phase-Locked Array GaInAsP-InP High Power Semiconductor Laser Grown by Low- Pressure Metalorganic Chemical Vapor Deposition," *Applied Physics Letters* 50, p. 230, 1987.

Razeghi, M., "High-power laser diodes based on InGaAsP alloys," *Nature* 369, pp. 631-633, 1994.

Razeghi, M., "High Power InAsSb/ InAsSbP Laser Diodes Emitting in the 3-5 μm Range," in *1998 Army Research Office Highlights*, Physical Sciences Directorate, 1998.

Razeghi, M., Wu, D., Lane, B., Rybaltowski, A., Stein, A., Diaz, J., and Yi, H., "Recent achievement in MIR high power injection laser diodes (λ=3 to 5 μm)," *LEOS Newsletter* 13, pp. 7-10, 1999.

Razeghi, M., Evans, A., Slivken, S., Yu, J.S., Zheng, J.G., Dravid, V.P., "High Power continuous-wave mid-infrared quantum cascade lasers based on strain-balanced heterostructures," in *Photonic Materials, Devices and Applications*, eds. G. Badanes, D. Abbott, and A. Serpengüzel, *SPIE Proceedings Series* 5840, SPIE-The International Society for Optical Engineering, Bellingham, WA, pp. 54-63, 2005.

Rong, H., Jones, R., Liu, A., Cohen, O., Hak, D., Fang, A., and Paniccia, M., "A continuous-wave silicon Raman laser," *Nature* 433, pp. 725-728, 2005.

Sai-Halasz, G.A., Tsu, R., and Esaki, L., "A new semiconductor superlattice," *Applied Physics Letters* 30, pp. 651-653, 1977.

Sai-Halasz, G.A., Chang, L.L., Welter, J.M., Chang, C.A., and Esaki, L., "Optical absorption of $In_{1-x}Ga_xAs$-$GaSb_{1-y}As_y$ superlattices," *Solid State Communications* 27, pp. 935-937, 1978a.

Sai-Halasz, G.A., Esaki, L., and Harrison, W.A., "InAs-GaSb superlattice energy structure and its semiconductor-semimetal transition," *Physical Review B* 18, pp. 2812-2818, 1978b.

Schawlow, A.L. and Townes, C.H., "Infrared and optical masers," *The Physical Review* 112, pp. 1940-1949, 1958.

Slivken, S., Huang, Z., Evans, A., Razeghi, M., "High-power (~9 μm) quantum cascade lasers," *Applied Physics Letters* 80, pp. 4091-4093, 2002.

Smith, D.L., and Mailhiot, C., "Proposal for strained type II superlattice infrared detectors," *Journal of Applied Physics* 62, pp. 2545-2548, 1987.

Wu, D., Lane, B., Mosheni, H., Diaz, J., and Razeghi, M., "High power asymmetrical InAsSb/ InAsSbP/ AlAsSb double heterostructure lasers emitting at 3.4 μm," *Applied Physics Letters* 74, pp. 1194-1196, 1999.

Xie, H., Wang, W.I., and Meyer, J.R., "Infrared electroabsorption modulation at normal incidence in asymmetrically stepped AlSb/InAs/GaSb/AlSb quantum wells," *Journal of Applied Physics* 76, pp. 92-96, 1994.

Yang, B.H., Zhang, D., Yang, R.Q., Lin, C.H., Murry, S.J., and Pei, S.S., "Mid-infrared interband cascade lasers with quantum efficiencies > 200%," *Applied Physics Letters* 72, pp. 2220-2222, 1998.

Yi, H., Diaz, J., Wang, L.J., Kim, S., Williams, R., Erdtmann, M., He, X., and Razeghi, M., "Optimized structure for InGaAsP/GaAs 808 nm high power lasers," *Applied Physics Letters* 66, pp. 3251-3253, 1995.

Youngdale, E.R., Meyer, J.R., Hoffman, C.A., Bartoli, F.J., Grein, C.H., Young, P.M., Ehrenreich, H., Miles, R.H., and Chow, D.H., "Auger lifetime enhancement in InAs-$Ga_{1-x}In_xSb$ superlattices," *Applied Physics Letters* 64, pp. 3160-3162, 1994.

Yu, J.S., Slivken, S., Darvish, S.R., Evans, A., Gokden, B., and Razeghi, M., "High-power, room-temperature and continuous-wave operation of distributed-feedback quantum-cascade lasers at λ~4.8 μm," *Applied Physics Letters* 87, p. 041104, 2005.

Further reading

Agrawal, G. and Dutta, N., *Semiconductor Lasers*, Van Nostrand Reinhold, New York, 1993.

Felix, C.L., Meyer, J.R., Vurgaftman I., Lin, C.H., Murry, S.J., Zhang, D., and Pei, S.S., "High-temperature 4.5 μm Type-II quantum-well laser with Auger suppression," *IEEE Photonics Technology Letters* 9, pp. 734-736, 1997.

Iga, K., *Fundamentals of Laser Optics*, Plenum Press, New York, 1994.

Johnson, J.L., Samoska, L.A., Gossard, A.C., Merz, J., Jack, M.D., Chapman, G. R., Baumgratz, B.A., Kosai, K., and Johnson, S.M., "Electrical and optical properties of infrared photodiodes using the InAs/$Ga_{1-x}In_xSb$ superlattice in heterojunctions with GaSb," *Journal of Applied Physics* 80, pp. 1116-1127, 1996.

Kim, S. and Razeghi, M., "Recent advances in quantum dot optoelectronic devices and future trends," in *Handbook of Advanced Electronic and Photonic Materials and Devices*, ed. H.S. Nalwa, Academic Press, London, pp. 133-154, 2001.

Lin, C.H., Yang, R.Q., Zhang, D., Murry, S.J., Pei, S.S., Allerman, A.A. and Kurtz, S.R., "Type-II interband quantum cascade laser at 3.8 μm," *Electronics Letters* 33, pp. 598-599, 1997.

Mohseni, H., Michel, E., Sandven, J., Razeghi, M., Mitchel, W., and Brown, G., "Growth and characterization of InAs/GaSb photoconductors for long wavelength infrared range," *Applied Physics Letters* 71, pp. 1403-1405, 1997.

O'shea, D., *Introduction to Lasers and Their Applications*, Addison-Wesley, Reading, MA, 1978.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, 1989.

Razeghi, M., *The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for Photonic and Electronic Device Applications*, Institute of Physics, Bristol, UK, pp. 21-29, 1995.

Razeghi, M., "Optoelectronic Devices Based on III-V Compound Semiconductors Which Have Made a Major Scientific and Technological Impact in the Past 20 Years," *IEEE Journal of Selected Topics in Quantum Electronics*, 2000.

Razeghi, M., Wu, D., Lane, B., Rybaltowski, A., Stein, A., Diaz, J., and Yi, H., "Recent achievements in MIR high power injection laser diodes (λ = 3 to 5 μm)," *LEOS Newsletter* 13, pp. 7-10, 1999.

Razeghi, M., "Kinetics of Quantum States in Quantum Cascade Lasers: Device Design Principles and Fabrication," *Microelectronics Journal* 30, pp. 1019-1029, 1999.

Scherer, A., Jewell, J., Lee, Y.H., Harbison, J., and Florez, L.T., "Fabrication of microlasers and microresonator optical switches," *Applied Physics Letters* 55, pp. 2724-2726, 1989.

Siegman, A.E., *Lasers*, University Science Book, Mill Valley, Calif., 1986.

Silfvast, W.T., *Laser Fundamentals*, Cambridge University Press, New York, 1996.

Streetman, B.G., *Solid States Electronic Devices,* Prentice-Hall, Englewood Cliffs, NJ, 1990.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

Problems

1. In a space mission to Mars, it is needed to equip the research robot with a single mode laser to make sample absorbance/transmittance analysis to identify the unknown gases. Assume that all types of lasers (e.g. gas, solid state, liquid, semiconductor lasers) specified in the Chapter are offering wavelength ranges and power levels adequate for this application. Which one would you choose for this application? Specify which properties of this type of laser make it more suitable?

2. Using the definition of the coefficients B_{12} and B_{21} given in sub-section 18.3.1, show that B_{12}=B_{21}.Take the photon field $\rho(E_{21})$ as $N(E_{21}).n_{ph}$ where $N(E_{21})$ is the density of energy states for the photon field and n_{ph} is the average number of photons defined as $n_{ph} = \dfrac{1}{e^{E_{21}/k_bT} - 1}$.

3. A laser diode has a gain peak between 4.45-4.65 μm, and we want to have only a single mode at 4.5 μm as output of this laser diode. What should be the length of the laser so that only a single mode is allowed within the laser? Take the refractive index within the active region as 3.2. (This type of laser is called short cavity laser).

4. Find the spacing (free spectral width) of wavelengths between adjacent modes in a Fabry Perot cavity. Compare this with the mode spacing in terms of frequency. Which one would you prefer to use if you are comparing free spectral widths of two cavities designed for different wavelength regimes?

5. Consider Eq. (18.18) and Eq. (3.36). It is interesting to note that confining light in a cavity via the use of total internal reflection is analogous to confining electrons in a finite quantum well. Observe that Eq. (18.18) and Eq. (3.36) have similar structures and identify the variables in Eq. (18.18) which are analogous to potential, mass and wave function in Eq. (3.36).

6. *Duality principle.*
   Once the solutions for E field are known, solutions for the H field can be found easily by using duality principle. Change the source free Maxwell's equations (Eq. (18.17)) by using the following substitutions: $E \to H$, $H \to -E$, $\varepsilon \to \mu$, $\mu \to \varepsilon$. Does anything change?

7. Consider an edge-emitting laser with identical facets. Find the mirror loss and threshold gain if $\overline{n}_{active}$=3.2, α_i=5 cm^{-1} and L=1 mm. Now consider that one of the facets is coated with a dielectric/metal layer such that the reflection coefficient is R=0.98, find threshold gain and mirror loss for this case and compare with the previous result. Which one do you think will have a smaller threshold current and higher power when the output from the facet with the same R is measured? Why?

8. Plot the external quantum efficiency η_e^{-1} vs. L when L is varying from 0 to 3 mm, given the values α_i=5 cm^{-1} and $\overline{n}_{active}$=3.2 and $\eta_e^{-1}(L)$=2.0327, 2.8154 and 3.5981 for L=1, 2 and 3 mm respectively. Find the internal quantum efficiency η_i^{-1}.

9. For a semiconductor laser diode, it is possible to make an edge-emitting laser be surface emitting at the same time by using second order corrugations on the waveguide and laterally inducing current so the surface of the ridge is not coated with metal. Consider the far field characteristics of the edge emission given the aperture sizes around 2 μm by 40 μm for the edge. Compare to a surface emission aperture with the same width, but a length of 35 μm. Which one do you expect to have less divergence and what is the beam shape in each case?

10. Quantum cascade lasers do not suffer Auger processes due to the fact that intersubband interactions involve electrons only, and in the cladding layers the majority carriers are electrons. What is the dominant loss mechanism in this type of laser?

19. Photodetectors - General Concepts

19.1. Introduction

A detector can be defined as a device that converts one type of signal into another as illustrated in Fig. 19.1. Various forms of input signal can be entered into the detector, which then generates the measurable output signal, such as an electrical current or voltage. There exist many different types of detectors depending on the objects or physical properties that they sense. The input signal can be mechanical vibrations, electromagnetic radiation, small particles, and other physical phenomena. Smoke detectors can sense the soot particulates caused by fire and seismometers sense the mechanical vibrations caused by the earth. The human body has various types of detectors: the eyes can sense electromagnetic radiation in the visible range, the ears detect sound from pressure variations through a medium such as atmospheric air or water, the tongue senses various types of chemicals, and the skin can detect temperature and pressure. Our natural sensory skills have been augmented through the development of advanced instruments such as the microscope and the thermometer, that were made possible thanks to the development of technology. Furthermore technology has made it possible for humans to detect things that could not be naturally sensed by the human body. For example, we can observe the infrared (IR) light emitted from warm objects and the ultraviolet (UV) light from hot objects with the help of photodetectors which will be the focus of the this and the next Chapters.

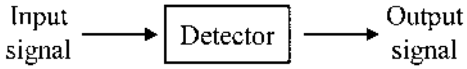


Fig. 19.1. Concept of a detector. The input signal usually has the form of electromagnetic radiation and the output signal is often an electrical signal. The detector is a device which converts one type of signal into another which can then be processed.

Human eyes respond only to visible light, from violet to red. However, the light spectrum is much broader and includes radiation beyond violet (e.g. ultraviolet, gamma rays) and red (e.g. infrared, microwaves). If the temperature of the object is larger than 6,000 K, it will emit predominantly in the ultraviolet. However, colder objects (<2,000 K) emit predominantly in the infrared. Most of the objects on earth emit IR light, and by choosing the correct materials and growth and fabrication techniques, photodetectors can be designed to sense light in this wavelength range. Using infrared photodetectors, we can obtain information on the objects emitting this radiation to determine their geometry, temperature, surface quality, and chemical content. We can also get information on the atmosphere through which the IR light is propagated.

Due to the fact that some wavelengths of infrared light are transmitted with little loss within the Earth's atmosphere, the IR spectrum offers some attractive advantages for photodetection purposes. Because of this and other advantages, IR photodetectors have been in active development over the last several decades and found numerous applications, such as night vision, missile guidance, and range finders. As the cost of these IR photodetectors has decreased, they have become more available for civilian and industrial applications where they are used in hazardous gas sensing, security systems, thermal imaging for medical purposes, hot spot monitoring and optical communications. Specialized infrared imagers have recently been used to detect malignant cancers and have acted as collision and ranging sensors in automobiles. Due to their prominence in commercial and military applications, in this and the next Chapters we will focus on photodetectors designed for the infrared regime.

Regardless of sensing wavelength, photodetectors are usually integrated into a system that generates a signal which can be easily recognized and interpreted by humans. A few elements of such a system are shown in a block diagram form in Fig. 19.2. The system may be designed to detect the target, to track it as it moves, or to measure its temperature. If the radiation from the target passes through any portion of the earth's atmosphere, it will be *attenuated* because the atmosphere is not perfectly transparent. The *optical receiver* collects some of the radiation from the target and delivers it to a detector which converts it into an electrical signal. Before reaching the *detector*, the radiation may pass through an *optical modulator* where it is coded with information concerning the direction to the target or information destined to assist in the discrimination of the target from unwanted details in the background. Since some detectors must be cooled, one of the system elements may be a *cooler*. The electrical signal from the detector then passes through a *processor* where it is amplified and the coded target information is extracted. The final step is the use of this information to automatically control some process or to display the information for interpretation by a human observer.

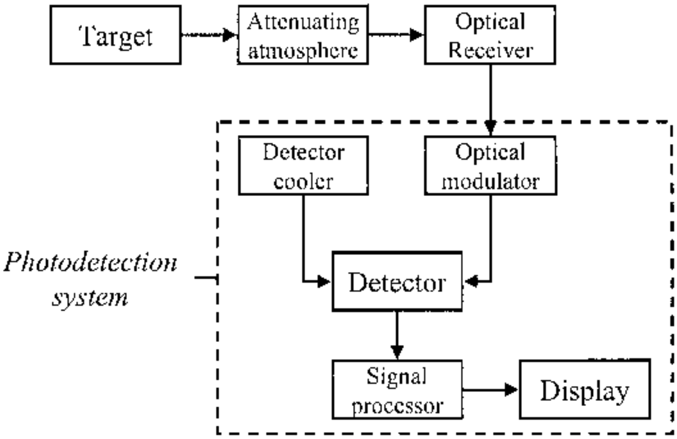


Fig. 19.2. The major elements constituting a photodetection system.

In this Chapter, we will first review the fundamental concepts of electromagnetic radiation emitted by a body, which will allow us to better understand the principles of photodetection. Next, we will describe the theory of operation of photodetectors and introduce the important parameters that characterize and compare their performance. Most photodetectors can be divided in two types: thermal detectors and photon detectors, and the difference lies in the detection mechanism. We will explore thermal detectors here and discuss the operation and some specific examples of photon detectors in the following Chapter.

19.2. Electromagnetic radiation

The schematic in Fig. 19.3 shows various types of electromagnetic radiations along with their associated wavelength and frequency ranges. The borders between visible, infrared, far-infrared, and millimeter waves are not absolute, and have been introduced primarily for convenience. For example, visible light is that portion of the spectrum to which the human eye is sensitive, and a statement such as "infrared extends from 0.7 μm to 1000 μm" is only a convention. Typically, IR radiation does not penetrate metals unless these are very thin, but passes through many crystalline, plastic, and gaseous materials-including the earth's atmosphere. There does not exist a detector that can detect all types of radiation with the same sensitivity. Thus, a photodetector has to be designed to operate within a specific spectral bandwidth.

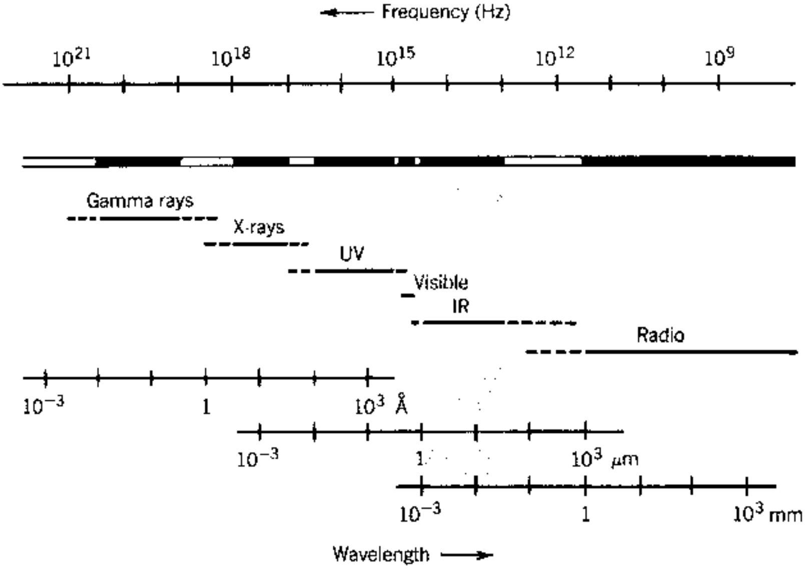


Fig. 19.3. The electromagnetic spectrum. The major spectral regions of interest here are shown with their limits in terms of frequency and wavelength, including the ultraviolet (UV), visible and infrared (IR).

A source of electromagnetic radiation generally emits over a broad range of wavelengths, and some wavelengths are emitted with more power than others. For example, different bodies emit radiation differently depending on their surface properties, due to varying emissivities. To have an absolute scale for proper comparison, we typically use a blackbody which is a perfect absorber of all radiant energy and which is also a perfect emitter of electromagnetic radiation. The intensity of the blackbody radiation depends on the wavelength of emitted light. This dependence is called the spectral distribution of intensity and is a function of the blackbody's absolute temperature.

An analytical expression of this spectral intensity distribution is shown in Fig. 2.3 and was determined by Max Planck in 1901. His theory assumed that the energy carried by an electromagnetic radiation is composed of discrete or quantized energy packages proportional to the frequency considered. The idea answered so many unsolved physics problems that Planck's hypothesis quickly became the basis for modern quantum theory. Planck's law, for which he received the Nobel Prize, quantifies the spectral radiation M from a blackbody as

$$\text{Eq. (19.1)} \quad M(\lambda) = \frac{2\pi hc^2}{\lambda^5}\frac{1}{\exp\left(\frac{ch}{k_b T\lambda}\right) - 1} \quad (\mathrm{Wm^{-2}\mu m^{-1}})$$

where λ is the wavelength in microns, c is the velocity of light in vacuum in m.s^{-1}, T is the absolute temperature in K, h is Planck's constant and k_b is Boltzmann's constant. This relation expresses thermal radiation as a function of wavelength and temperature for all wavelengths. The peak wavelength of the ideal blackbody emission is described by Wien's law:

$$\text{Eq. (19.2)} \quad \lambda_{Peak}(BB_T) = \frac{hc}{4.965k_b T} = \frac{2.898\times 10^{-3}\,Km}{T}$$

When considering the entire spectral power R of an ideal blackbody (integrating Planck's Law over all wavelengths), the result is the Stephan-Boltzmann law for total blackbody emittance:

$$\text{Eq. (19.3)} \quad \int_0^\infty M(\lambda,T)d\lambda = R(\lambda,T) = \sigma T^4$$

where σ is the Stephan-Boltzmann constant 5.67 × 10^{-8} Wm^{-2}K^{-4}. Further details can be found in sub-section 3.1.1.

19.3. Photodetector parameters

As mentioned earlier, photodetectors are devices which sense light (as an input signal) and generate a measurable output signal in the form of an electrical current or voltage. The performance of these photodetectors can be quantified and compared using several parameters, which will be discussed in this section.

19.3.1. Responsivity

The responsivity of a photodetector is the ratio of its output electrical signal, either a current I_{out} or a voltage V_{out}, to the input optical signal expressed in terms of the incident optical power P_{in}. One can define a current responsivity and a voltage responsivity using respectively:

Eq. (19.4) $$R_i = \frac{I_{out}}{P_{in}} \text{ and } R_v = \frac{V_{out}}{P_{in}}$$

The current responsivity R_i is expressed in terms of A/W, while the voltage responsivity R_v is expressed in units of V/W. The output signals I_{out} and V_{out} in Eq. (19.4) can be expressed in detail for the case of specific light detection mechanisms of the detector in question. The output current and voltage are often called photocurrent (I_{ph}) and photovoltage (V_{ph}) as they arise in the presence of light. The incident input power P_{in} on the photodetector can be expressed as:

Eq. (19.5) $$P_{in} = A\Phi_{ph}\frac{hc}{\lambda}$$

where A is the area of the detector, Φ_{ph} is the incident photon flux density expressed in units of photons.m^{-2}.s^{-1}, h is Planck's constant, c is the velocity of light in vacuum, and λ is the wavelength of the incident light.

19.3.2. Noise in photodetectors

As the output signal of a photodetector is an electrical signal, it is not a strictly stable quantity over time, but fluctuates due to electrical noise. Electrical noise is a random variable that results from stochastic or random processes associated with various particles, i.e. discrete, nature of electrons, phonons, and photons and their interactions. For example, the instantaneous voltage signal *V(t)* is shown schematically in Fig. 19.4. A quantitative measure of the noise in the voltage signal is then its root-mean-square and is given by:

Eq. (19.6) $$I_{noise} = \sqrt{\frac{1}{T}\int_0^T \left(I(t) - \langle I \rangle\right)^2 dt}$$

This relation expresses the average value of the square of the fluctuation of the variable *I(t)* around its average <*I*> over a long period of time *T*.

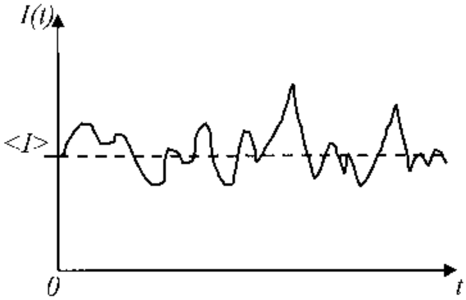


Fig. 19.4. Illustration of the instantaneous current signal I(t) exhibiting electrical noise about its average value <I>.

It is also useful to define a power signal-to-noise ratio as:

Eq. (19.7) $$\frac{S}{N}=\frac{I_{ph}^{2}}{\left\langle I_{noise}^{2}\right\rangle}$$

where $\langle ... \rangle$ denotes the statistical average of a random variable, I_{ph} is the instantaneous photocurrent and I_{noise} is the instantaneous noise current. We talk about a power ratio because the expression in Eq. (19.7) involves the squares of the electrical current. The effect of the signal-to-noise ratio is illustrated in Fig. 19.5.

The noise current I_{noise} is usually a function of the frequency considered, and the noise power I_{noise}^{2} depends on the frequency bandwidth Δf over which the statistical average is measured. For example, if the output signal is centered at a frequency of 200 Hz, eliminating all electrical signal outputs with significant components above 250 Hz and below 150 Hz would decrease the noise power in the output because the remaining 100 Hz electrical bandwidth contains less noise than the entire frequency band in the case of white noise. This is why, in order to compare different noise mechanisms in different photodetectors, with different frequency bandwidths, it is convenient to use the concept of noise spectral density, expressed in terms of A.Hz$^{-½}$, which consists of normalizing the noise power to the frequency bandwidth considered:

Eq. (19.8) $$\text{noise spectral density}=\frac{\sqrt{\left\langle I_{noise}^{2}\right\rangle}}{\sqrt{\Delta f}}$$

A similar expression can be obtained using the voltage noise V_{noise}.

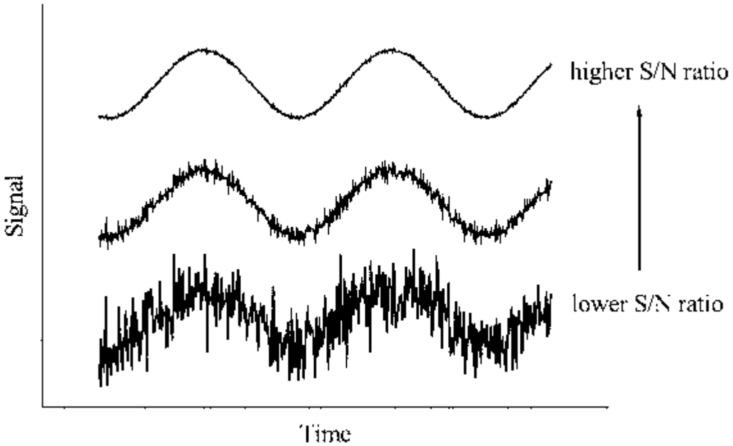


Fig. 19.5. Detector output with varying signal-to-noise ratios. When the S/N ratio is high, the signal is clear but when the S/N ratio is low, the random modulation signal (noise) is superimposed which reduces the accuracy of the actual signal.

A useful concept in the evaluation of the electrical noise is to consider a noise equivalent circuit that consists of a small signal electrical circuit model of the device including the noise source. This equivalent circuit makes use of the differential resistance R_{diff} of the device when a bias voltage $V=V_b$ is applied:

Eq. (19.9) $$R_{diff} = \left(\frac{\partial V}{\partial I} \right)_{|V=V_b}$$

where V and I are the voltage and current of the device, respectively. The noise equivalent circuits are shown in Fig. 19.6.

These noise equivalent circuits can only be used when the noise in the devices is a white noise, i.e. does not depend on the frequency. These noise equivalent circuits can be used like any small signal circuits when several contributions to the total noise are considered, with the exception that the noise powers need to be added instead of summing or subtracting noise voltages and noise currents, e.g.:

Eq. (19.10) $$\left\langle I_{noise}^2 \right\rangle_{total} = \left\langle I_{noise}^2 \right\rangle_1 + \left\langle I_{noise}^2 \right\rangle_2$$

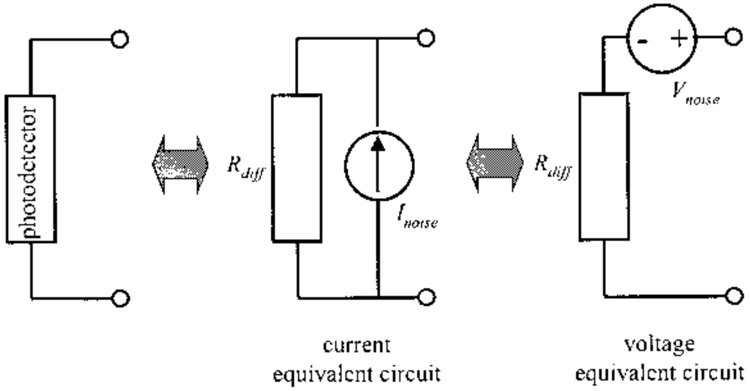


Fig. 19.6. Current and voltage equivalent circuits of a photodetector, including the sources of noise I_{noise} and V_{noise}.

19.3.3. Noise mechanisms

There exist several contributions to the noise in photodetectors that will be briefly described in this sub-section.

Johnson noise.

Thermal or Johnson noise occurs in all electrically conductive materials regardless of the conduction mechanism and results from the random motion of thermally-activated charge carriers through the conductor. The total noise current is proportional to the sum of the carrier movement occurring within a short time frame. Johnson noise is named after the scientist who experimentally investigated it [Johnson 1928]. The mean-square voltage of such noise can be calculated from:

Eq. (19.11) $$\left\langle V_{noise}^2 \right\rangle_{Johnson} = 4k_b TR\Delta f$$

where k_b is the Boltzmann constant, T is the temperature of the conductor, R is its electrical resistance, and Δf is the frequency bandwidth of the noise which is the frequency range in which the noise exists or is considered. This equation shows that the value of the noise in a given bandwidth is independent of the value of frequency, therefore the Johnson noise is a white noise. As a result, there is a constant maximum noise power P_{noise} from any resistor R at a given temperature T:

Eq. (19.12) $$\left(P_{noise}\right)_{Johnson\max} = \frac{1}{2}\frac{\left\langle V_{noise}^2 \right\rangle}{2R} = k_b T\Delta f$$

and the value of this power at room temperature is about 4×10^{-21} W.

Shot noise.

In 1918, Schottky showed that the random arrival of electrons on the collecting electrode of a vacuum tube was responsible for a so-called shot noise. At the origin of shot noise is the process of charge carriers being thermally or optically excited over a junction barrier. The current shot noise in a simple temperature limited vacuum diode is given by the following expression:

Eq. (19.13) $\left\langle I_{noise}^{2}\right\rangle_{shot}=2qI_{DC}\Delta f$

where q is the elementary charge and I_{DC} is the DC current bias of the vacuum diode. This equation shows that shot noise is also independent of the frequency and is a white noise. However, this is only valid if the inverse of the frequency of operation, $1/f$, is much larger than the traveling time of the electron in the device.

The shot noise in a p-n junction diode can be estimated by the following equation for the low frequency region:

Eq. (19.14) $\left\langle I_{noise}^{2}\right\rangle_{shot}=2q\left(I_{D}+I_{0}\right)\Delta f$

where $I_{D}=I_{0}\left(e^{\frac{qV_b}{k_BT}}-1\right)$ and I_0 is the saturation current of the diode. This relation is equivalent to Eq. (19.11) when no bias is applied (V_b=0), and to Eq. (19.13) for a high enough current bias.

1/f noise.

There also exists an important type of noise which has a power spectrum inversely proportional to the frequency of operation f. This noise mechanism often dominates other mechanisms at low frequencies. This is a process dependent noise in the sense that it can be affected by the contact type and preparation, as well as the surface preparation and passivation. For example, carrier trapping and re-emission to and from defects at surfaces and contacts may contribute to this $1/f$ noise. It should be noted that a variety of names have been used for this type of noise in the literature such as excess noise, modulation noise, contact noise, and flicker noise. The power spectrum of this noise is given by:

Eq. (19.15) $P_{noise}(f) = \gamma \frac{I_{DC}^{\alpha}}{f^{\beta}}$

where γ is a constant, I_{DC} is the DC current bias of the device, f is the frequency of operation, α and β are exponents characteristic of the particular device considered. The value of α is usually near 2, while the value of β ranges from 0.8 to 1.5.

Generation-recombination noise.

The generation and recombination of charge carriers in semiconductors are random processes as they are associated with the creation and annihilation of electron-hole pairs in the material. The number of free carriers during a given period of time is therefore not constant but fluctuates in a random manner too. This results in a random change of the voltage of the device which is called the generation-recombination or G-R noise. For a near intrinsic semiconductor with a moderate bias, this noise can be expressed as [Long 1967]:

Eq. (19.16) $(V_{noise})_{G-R} = \frac{2V_b}{(lwt)^{1/2}} \left(\frac{1+b}{bn+p} \right) \left[\frac{np}{n+p} \frac{\tau \Delta f}{1+\omega^2 \tau^2} \right]^{1/2}$

where V_b is the voltage bias, l is the length, w is the width, and t is the thickness of the device, b is the ratio of mobilities μ_e / μ_h, n and p are the electron and hole concentrations respectively, τ is the carrier lifetime in the semiconductor, ω is the angular frequency of operation, and Δf is the frequency bandwidth within which the noise is measured.

Temperature noise.

If the conductivity of the device strongly depends on the temperature, any random temperature fluctuation would result in a so called temperature noise in the device. This is an important source of noise for all infrared thermal detectors as well as low noise preamplifiers. This is the reason why even uncooled infrared thermal detectors, i.e. which can operate without cooling, usually have a thermoelectric cooler to stabilize their temperature.

Photon noise.

This noise mechanism arises from the random arrival of background photons onto the surface of the photodetector and is included in the incident photon flux. Unlike the previously described noise mechanisms, photon noise is a source of noise which is extrinsic to the device. The equivalent power of this noise or noise-equivalent-power for a given detector with an area A and a quantum efficiency η is:

$$\text{Eq. (19.17)}\quad \left(P_{NEP}\right)_{photon} = h\nu\left[\frac{2A\Phi_{bkg}\Delta f}{\eta}\right]^{1/2}$$

where Φ_{bkg} is the total background photon flux density reaching the detector, and υ is the frequency of the photon.

19.3.4. Detectivity

Although the responsivity of a photodetector gives a measure of the output signal of the detector for a given optical input signal, it does not give any information about the sensitivity of the device. The sensitivity of the detector can be defined as the minimum detectable optical input power that can be sensed with a signal-to-noise ratio of unity. This power is called the noise-equivalent-power (P_{NEP}) of the detector and is given by:

$$\text{Eq. (19.18)}\quad P_{NEP} = \frac{I_{noise}}{R_i} \text{ or } P_{NEP} = \frac{V_{noise}}{R_v}$$

Jones suggested to define the detectivity of a detector as the inverse of this noise-equivalent-power [Jones 1953]:

$$\text{Eq. (19.19)}\quad D = \frac{1}{P_{NEP}}$$

which is expressed in units of W^{-1}. This quantity is very useful when measuring the sensitivity of photodetectors. However, it is not a fair means of comparing the overall performance of different detectors because it neglects the effects of the detector area and frequency bandwidth. For example, photodetectors with different sizes and thus detection areas will have different noise-equivalent-powers. In addition, a detector with low electrical bandwidth can have higher detectivity than an otherwise identical detector with wider electrical bandwidth. This is despite the fact that higher bandwidth is desired for faster devices. To address these issues, Jones introduced the concept of specific detectivity, which is denoted D^* and is the detectivity of a photodetector with an area of 1 cm^2 and an electrical bandwidth of 1 Hz:

Eq. (19.20) $D^* = D\sqrt{A\Delta f} = \dfrac{\sqrt{A\Delta f}}{P_{NEP}}$

where A is the area of the detector in cm^2. D^* is expressed in units of cm.Hz^{½}.W^{-1}. Since it is independent of the device dimensions and the electrical configuration used for the measurement, D^* is widely used to compare photodetectors with very different physical and operational characteristics and is often simply called detectivity. One can easily express D^* in terms of the detector responsivity:

Eq. (19.21) $D^* = \dfrac{R_i}{I_{noise}}\sqrt{\dfrac{A}{\Delta f}} = \dfrac{R_v}{V_{noise}}\sqrt{\dfrac{A}{\Delta f}}$

where Δf is the frequency bandwidth of the measurement setup, I_{noise} and V_{noise} are the total root-mean-square current and voltage noises of the detector in the given frequency bandwidth of Δf.

19.3.5. Detectivity limits and BLIP

In order to ascertain the maximum performance of a photodetector, it is important to understand the role of the background noise current, I_d. This current can arise from blackbody radiation absorbed by the detector from the environment with temperature T_B. Additionally background noise can be produced by the detector device itself in the form of dark currents, that is, noise contributions from conduction or leakage currents under no illumination. These two components define the upper detectivity limit photodetectors as shown in Fig. 19.7.

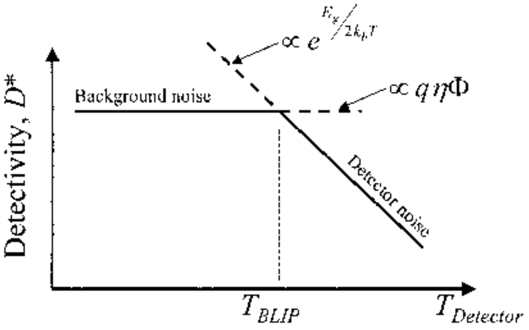


Fig. 19.7. Detectivity limit with respect to temperature for a photodetector.

As the detector temperature increases, the dark current increases exponentially as $e^{-E_g/2k_bT}$. Conversely, it is logical that the background noise stays constant with temperature and is proportional to the quantum efficiency of the device and the photon flux incident on the detector. The temperature at which the background noise begins to limit the device performance is known as the background limited infrared performance (BLIP) temperature. The BLIP threshold can be found from solving an equality between the constant background blackbody noise current and the device dark current:

Eq. (19.22) $$I_{sat}(T_{BLIP}) = I_B = q\int_{\lambda_1}^{\lambda_2}\eta(\lambda)\frac{d\Phi_B}{d\lambda}d\lambda$$

where η is the detector quantum efficiency (essentially the ratio between the absorbed and total incident photons) with respect to wavelength and $\frac{d\Phi_B}{d\lambda}$ is the photon flux over the wavelength range from λ_1 to λ_2. Using this formalism, one can also calculate the absolute detectivity limit for ideal photodiodes, for example:

Eq. (19.23) $$D^*_{BLIP,\max}(\lambda_0, T_B) = \frac{1}{h\nu}\frac{1}{2^{1/2}[\Phi_B(\lambda_0)]^{1/2}}$$

It should be noted that this limit should be modified by $1/\sin(\varphi)$ when considering a real detection system with cryostat, and an acceptance angle of φ. A complete explanation of detectivity limits and BLIP can be found in [Rosencher and Vinter 2002].

19.3.6. Frequency response

When the optical input signal is periodic with a fixed amplitude and a frequency f, the amplitude of the detector electrical output signal is not necessarily a constant but may vary with the frequency as shown in Fig. 19.8. This phenomenon is usually due to the electrical resistive-capacitance or RC delay of the device in the case of the photon detectors, and the thermal RC delay of the thermal detectors. The frequency dependent responsivity can be approximated by:

Eq. (19.24) $$R_{i,v}(f) = R_{i0,v0}\frac{1}{\sqrt{1+\left(\frac{f}{f_c}\right)^2}}$$

where $R_{i0,v0}$ is the responsivity of the photodetector at very low frequencies, and:

Eq. (19.25) $$f_c = \frac{1}{2\pi RC}$$

where R is the electrical (or thermal) resistance and C is the electrical (or thermal) capacitance of the photon (or thermal) detector.

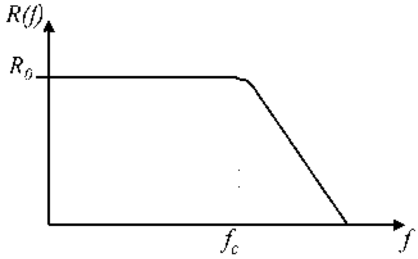


Fig. 19.8. Illustration of the frequency dependence of detector responsivity.

19.4. Thermal detectors

In thermal detectors, the absorption of IR light leads to a change in the temperature of the detector, thus resulting in a change in resistance (bolometer) or electrical polarization (pyroelectric detectors). This change is then recorded by an electrical circuit. Because they can operate at room temperature, thermal detectors are mostly used whenever cooling systems are not possible and to minimize cost. It is however often necessary to use thermoelectric coolers in order to stabilize the temperature of the detectors, in an effort to minimize the Johnson noise. In general, thermal detectors have a large spectral bandwidth compared to their photon detector counterparts. This is because thermal-type devices typically absorb photons like a blackbody, or with a wide, flat response function with respect to wavelength. There exist several types of thermal detectors that will be briefly discussed here: bolometers, thermopiles, thermocouples, and pyroelectric detectors.

A bolometer is a thermal detector whose resistance depends on its temperature. Since the resistance of most semiconductors is a strong function of temperature, the resistance of a semiconductor chip can tell us how much radiant energy is falling on it. The Golay cell uses the expansion of a gas when heated to sense radiated power: the gas is contained in a chamber that is closed with a reflecting membrane. When the gas is warmed, the membrane distorts and deflects a beam of light that has been focused on it.

A thermopile typically consists of several thermocouple stacks in series. The principle of a thermocouple resides in the thermoelectric effect which yields an voltage proportional to the temperature difference between two dissimilar metal junctions. Therefore, an increase in the incident infrared radiation power causes an increase in the thermopile voltage.

When ferroelectric polar crystals are exposed to a change in temperature, their internal electric polarization changes, electrical charges accumulate and can be measured on opposite sides of the crystal. The capacitance of the detector material also changes and can be electrically measured. Pyroelectric detectors use this material property to detect IR radiation.

Among all these thermal detectors, bolometers are the most widely used because of their advantages such as easy fabrication, stability, light-weight, ruggedness, and an easy array capability. To illustrate some principles of thermal detectors, consider the carbon bolometer shown in Fig. 19.9. Such devices are very sensitive and detect radiation over a very wide spectral range.

The resistance of an ordinary carbon resistor is a strong function of temperature, which makes a carbon resistor an inexpensive temperature sensor. To make a bolometer, we would mount the resistor in such a way that it is cooled to a low temperature but also isolated. As radiation strikes it, it warms up and the resistance decreases. An external electrical circuit detects the resistance change. To make the bolometer more sensitive, we would want to make its heat capacity smaller so that a small amount of energy could heat it faster.

When a modulated optical signal with power amplitude P_{in} and angular frequency ω hits a pixel with a heat capacity C, a temperature change ΔT can be recorded. The heat dissipation inside a solid is characterized by its thermal conductance K. If we define the quantum efficiency η as the fraction of the incident optical power that is absorbed by the solid, the temperature change is given by:

Eq. (19.26) $$\Delta T = \frac{\eta P_{in}}{K\left(1+\omega^2\tau_{th}^2\right)^{1/2}}$$

where τ_{th} is the thermal response time defined as the heat capacity of the material divided by its thermal conductance:

$$\text{Eq. (19.27)} \quad \tau_{th} = \frac{C}{K}$$

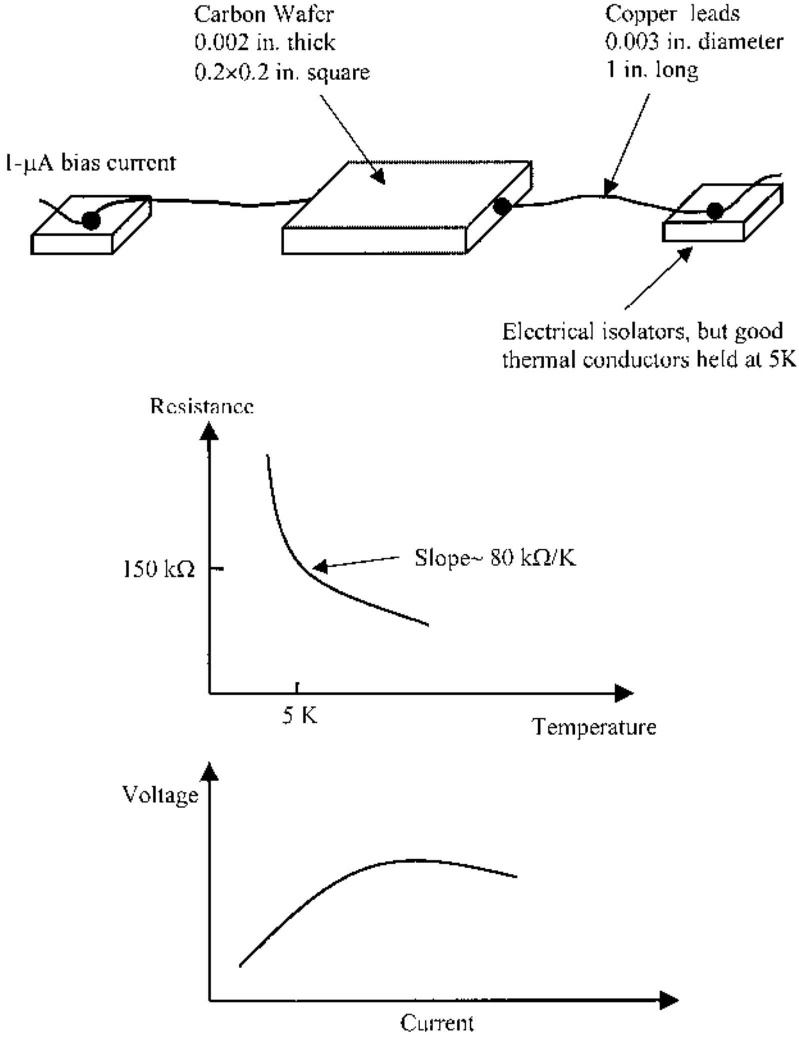


Fig. 19.9. Bolometer and its operating principle. A bolometer is a thermal detector whose resistance depends on its temperature. By precisely measuring the change in the resistance, one can determine how much radiant energy has reached the device.

A few key points can be understood from Eq. (19.26) which can optimize the thermal detector performance. First, it is very important to minimize the thermal conductance K, by maximizing thermal isolation, in order to sense low power infrared radiation. Improved temperature isolation is often accomplished by lengthening or thinning the support legs in bolometer bridge structures. However, minimizing K also leads to a low frequency response, which is proportional to the heat dissipation rate. Secondly, the quality of the detector material is important in maximizing η, which is the ratio of incident to absorbed radiation. And lastly, the heat capacity C, of the element must be low enough in order to meet the response time requirement.

Recent research on bolometric arrays has demonstrated good room temperature performance. Wang *et al.* [2005] have deposited vanadium oxide on silicon substrates using ion beam sputtering. After micromachining, a bolometer array of 128 elements showed a detectivity of 2×10^8 cm.Hz^{½}.W^{-1} with a responsivity of 5 kV/W in the 8~12 μm regime. An SEM image of an array pixel is shown in Fig. 19.10.

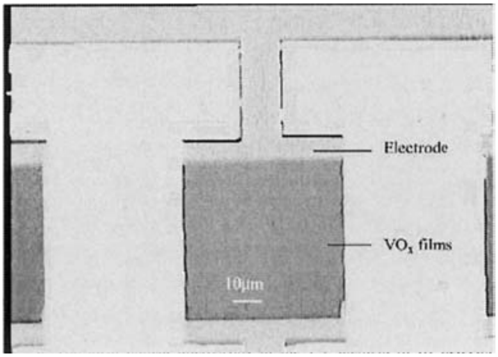


Fig. 19.10. Example of a bolometer array pixel. The active material is vanadium oxide. [Reprinted from Sensors and Actuators A Vol. 117, Wang, S.B., Xiong, B.F., Huang, G., Chen, S.H., and Yi, X.J., "Preparation of 128 element of IR detector array based on vanadium oxide thin films obtained by ion beam sputtering," p. 113, Copyright 2005, with permission from Elsevier.]

19.5. Summary

In this Chapter, the photodetector was established as a device that converts photon energy into electrical energy. The electromagnetic spectrum and blackbody concepts were initially revisited to provide a good background for the reader. The parameters used to describe the performance and limits of photodetectors were outlined, including responsivity, noise and detectivity. Finally, one specific type of photodetector, the thermal detector, and a carbon bolometer example were presented.

References

Johnson, J.B., "Thermal agitation of electricity in conductors," *Physical Review* 32, pp. 97-109,1928.

Jones, R.C., *Advances in Electronics*, Vol. V, Academic Press, New York, 1953.

Long, D., "On generation-recombination noise in infrared detector materials," *Infrared Physics* 7, pp. 169-170, 1967.

Rosencher, E. and Vinter B., *Optoelectronics*, Cambridge University Press, Cambridge, 2002.

Wang, S.B., Xiong, B.F., Zhou, S.B., Huang, G., Chen, S.H., Yi, X.J., "Preparation of 128 element of IR detector array based on vanadium oxide thin films obtained by ion beam sputtering," *Sensors and Actuators A* 117, pp. 110-114, 2005.

Further reading

Boyd, R.W., *Radiometry and the Detection of Optical Radiation*, John Wiley & Sons, New York, 1983.

Hudson Jr., R.D., *Infrared System Engineering*, John Wiley & Sons, New York, 1969.

Kingston, R.H., *Detection of Optical and Infrared Radiation*, Springer-Verlag, Berlin, 1978.

Problems

1. Sort the following types of electromagnetic radiation by order of increasing wavelength: radio, infrared, near-infrared, red, gamma rays, blue, ultraviolet, x-rays.

2. What is the total radiative emission of a surface (like a classroom wall) with area of 10 m^2, emissivity of 0.5 and temperature of 300 K? The emissivity is simply a non-ideality factor for an object that does not emit and absorb perfectly. If this emission power seems high to you, remember that this surface is usually in thermal equilibrium and is also absorbing approximately the same amount of power from its surroundings.

3. Using Wien's law, calculate the peak wavelength of emission from these sources: The Sun (6000 K), tungsten filament (3000 K), red hot source (1000 K), 300 K ambient.

4. Name the following detector parameters:
(a) Electrical output for a given light input.
(b) The signal-to-noise ratio that would result if the performance of your detector were scaled to a detector of a standard size, under standardized test conditions.
(c) The "clutter" or unwanted electrical variation that tends to hide the true signal.
(d) The minimum infrared power that a detector can accurately "see".
(e) A measure of the "cleanliness" of a signal pattern.
(f) The condition when a detector's performance is not limited by intrinsic device noise, but rather the incident photon noise.

5. Assume you have a photodetector with a current responsivity of 5 A/W. A continuous wave HeNe laser with wavelength of 632.8 nm and spot size of 0.5 mm^2 is incident on the active area of the detector. Assume that 6.37×10^{22} photons per m^2 arrive on the detector each second. Calculate the expected photocurrent under these conditions.

6. A coaxial cable is used to transmit data with a bandwidth of 100 MHz. Calculate the maximum peak Johnson noise power for 300 K and 80 K.

7. A certain HgCdTe detector has a specific detectivity of 1×10^{11} cm.Hz^{½}.W^{-1} and is designed to operate at 10 μm. It is

incorporated into a focal plane array that needs to operate with a bandwidth of 1 kHz. This particular detector has a pixel size of 30 μm×30 μm. Calculate the minimum power that can be sensed by this detector pixel.

8. You are tasked with designing a night vision infrared imaging system that should be able to detect a incident signal power level of 1×10^{-7} W.cm^{-2}. Assume that you need to use a focal plane array with pixels of 30μm×30μm operating at 60 Hz. What minimum specific detectivity does your system need to have?

9. Describe briefly how thermal detectors work.

10. What specific property changes with temperature in the following thermal detectors?
    (a) mercury-in-glass thermometer
    (b) carbon bolometer
    (c) thermopile
    (d) Golay cell

20. Photon Detectors

20.1. Introduction

In the previous Chapter, the basic concepts of photodetectors were outlined. Furthermore, thermal detectors and the bolometer, specifically, were described in detail. In photon detectors, incident photons interact with the electrons in the material and change the electronic charge distribution. This perturbation of the charge distribution generates a current or a voltage that can be measured by an electrical circuit. Because the photon-electron interaction is "instantaneous", the response speed of photon detectors is much higher than that of thermal detectors. Indeed, by contrast to thermal detectors, quantum or photon detectors respond to incident radiation through the excitation of electrons into a non-equilibrium state.The mechanisms of electron excitation are shown in Fig. 20.1.

Semiconductor photon detectors may rely on interband electron excitation (Fig. 20.1(a)) (intrinsic detectors), on impurity-band transition (Fig. 20.1(b)) (extrinsic detectors) or intersubband transitions in a quantum well (Fig. 20.1(c)) (quantum well intersubband photodetectors). In intrinsic semiconductors, an electron in the valence band absorbs the energy of an incident photon and is excited into the conduction band. Extrinsic semiconductor detection occurs when an electron from a trap level inside the bandgap absorbs the energy of an incident photon and is excited into the conduction band. Intersubband detectors feature an electron in a quantum-confined state that can be excited into a higher energy state or continuum level. In the following sub-sections, the two main types of photodetectors, photoconductive and photovoltaic detectors, will be discussed in more detail.

The output signal of a photon detector due to incident light is called the photoresponse. It is strongly dependent on the frequency or wavelength of the incident light. When the wavelength of the incident light becomes longer than a critical wavelength, the photoresponse decreases abruptly. This particular wavelength generally corresponds to the bandgap of the semiconductor material in intrinsic detectors or to the activation energy of defect states in extrinsic detectors. For those wavelengths longer than this critical point, the energy of the incident photons is no longer sufficient to excite an electron-hole pair across the bandgap or to overcome the activation energy. The wavelength at which this abrupt decrease in responsivity occurs is called the cut-off wavelength.

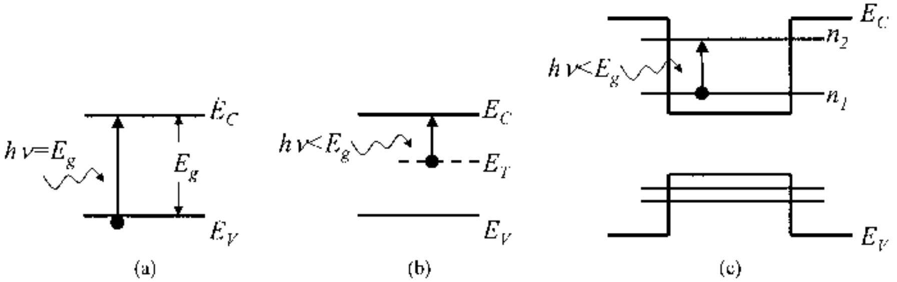


Fig. 20.1. Different mechanisms of excitation of an electron: (a) intrinsic semiconductor; (b) extrinsic semiconductor; and (c) intersubband transitions in a quantum well.

For example, InSb has a bandgap of 0.22 eV, which corresponds to a wavelength of 5.6 μm. Photons with longer wavelengths will pass through InSb undetected, i.e. InSb is transparent in spectral region beyond 5.6 μm, while photons of wavelengths shorter than 5.6 μm are effectively absorbed by InSb and contribute to the responsivity. Thus we expect to see a cut-off wavelength of 5.6 μm. This property is true for both photoconductors and photovoltaic detectors.

Examples of photoconductive detectors include doped germanium (Ge:X) and silicon (Si:X), and lead salts (PbS, PbSe). The ternary compounds HgCdTe and PbSnTe can be used as photoconductors and also as photovoltaic detectors.

20.2. Types of photon detectors

20.2.1. Photoconductive detectors

A photoconductive detector (also called a photoconductor) is essentially a radiation-sensitive resistor. The operation of a photoconductor is shown in Fig. 20.2. A photon of energy $h\upsilon$ greater than the bandgap energy E_g is absorbed to produce an electron-hole pair, thereby changing the electrical conductivity of the semiconductor. In almost all cases, electrodes attached to the sample measure the change in conductivity. Photoconductors are usually biased using a battery and a load resistor. An increase in the detector conductance both increases the current and decreases the voltage across the detector. For low resistance material, the photoconductor is usually operated in a constant current circuit as shown in Fig. 20.2. The series load resistance is large compared to the sample resistance, and the signal is detected as a change in the voltage developed across the sample. For high resistance photoconductors, a constant voltage circuit is preferred, and the signal is detected as a change in current.

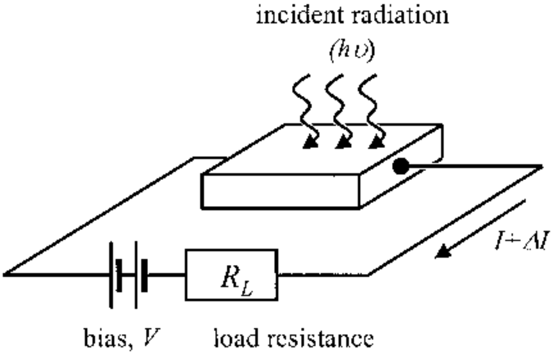


Fig. 20.2. Photoconductor and its biasing circuit.

The photoconductivity $\Delta\sigma$ is the difference between the electrical conductivity when a photoconductive material is illuminated and non-illuminated:

Eq. (20.1) $\Delta\sigma = q(\mu_e \Delta n + \mu_h \Delta p)$

where Δn and Δp are the excess electron and hole concentrations resulting from the incident light, μ_n and μ_p are the electron and hole mobility, respectively. In this case, the photoresponse takes the form of a change in the electrical current flowing through the sample, as a result of the change in electrical conductivity. The additional current component is called photocurrent. In a sample with contacts area A and length l, the photocurrent under the bias V is given by:

Eq. (20.2) $\Delta I = q(\mu_e \Delta n + \mu_h \Delta p)\frac{A}{l}V$

The excess carrier concentrations generated under a steady-state illumination are:

Eq. (20.3) $\Delta n = \Delta p = G\tau_n$

where τ_n is the recombination lifetime of the excess carriers, as defined in section 8.6, and G is the excess carrier generation rate. This quantity is further related to the incident optical power P_{in} through:

Eq. (20.4) $G = \eta \frac{P_{in}}{h\upsilon}\frac{1}{lA}$

where the quantum efficiency η represents the fraction of the incident optical power that contributes to electron-hole pair generation, and υ is the frequency of the incident light. Eq. (20.2), Eq. (20.3) and Eq. (20.4) can be combined and give:

Eq. (20.5) $\Delta I = q\left(\frac{\eta P_{in}}{h\upsilon}\right)\tau_n(\mu_e + \mu_h)\frac{V}{l^2}$

For $\mu_e >> \mu_h$, this expression becomes:

Eq. (20.6) $\Delta I = q\left(\frac{\eta P_{in}}{h\upsilon}\right)\frac{\tau_n}{\tau_t}$

where we have defined the quantity:

Eq. (20.7) $$\tau_t = \frac{l^2}{\mu_e V}$$

which physically represents the electron transit time across the electrodes, i.e. the time taken by an electron to travel or transit from one end of the semiconductor to the other, separated by a distance l, when a bias V is applied. The ratio $\frac{\tau_n}{\tau_t}$ in Eq. (20.6) characterizes how fast the electrons can transit from one electrode to another electrode and contribute to the photocurrent before recombination occurs. It is called the photoconductive gain.

The current responsivity R_i of a photoconductor is defined as the ratio of the output signal, i.e. the photocurrent ΔI, to the input signal, i.e. the incident optical power P_{in}:

Eq. (20.8) $$R_i = \frac{\Delta I}{P_{in}} = q\left(\frac{\eta}{h\upsilon}\right)\frac{\tau_n}{\tau_t}$$

The units of the current responsivity are A.W^{-1}. It is also common to express the responsivity as a function of the wavelength λ of the incident light:

Eq. (20.9) $$R_i = \frac{\Delta I}{P_{in}} = q\left(\frac{\eta\lambda}{hc}\right)\frac{\tau_n}{\tau_t}$$

When the incident optical power is modulated by a sinusoidal signal of frequency ω, the photoresponse can be considered in terms of the root-mean-square (rms) of the photocurrent, i.e.:

Eq. (20.10) $$I_{rms} = q\left(\frac{\eta P_{rms}}{h\upsilon}\right)\frac{\tau_n}{\tau_t}\frac{1}{\sqrt{1+(\omega\tau_n)^2}}$$

where $P_{rms} = \frac{P_{in}}{\sqrt{2}}$ is the rms of the incident optical power.

Depending on the electrical circuit considered, the photoresponse can sometimes be expressed as the ratio of the output voltage to incident optical power. The choice of the output signals for a photoconductor, either current or voltage, generally depends on the application in which the photodetector will be used. As shown in Eq. (20.9), the responsivity is a linear function of the wavelength of the incident light, up until the cut-off wavelength is reached. Beyond that wavelength, the responsivity abruptly decreases as discussed previously.

20.2.2. Photovoltaic detectors

Photoconductors are passive devices that need an external electrical bias in order to operate and do not generate a voltage by themselves. They can consist of a simple block of semiconductor material. By contrast, a photovoltaic detector needs a more complex structure that uses a p-n junction. Such a detector is also called a photodiode. This allows the detector to exhibit a voltage when photons are absorbed as we will now briefly discuss. As a result, photovoltaic detectors are usually operated with a low or zero external bias.

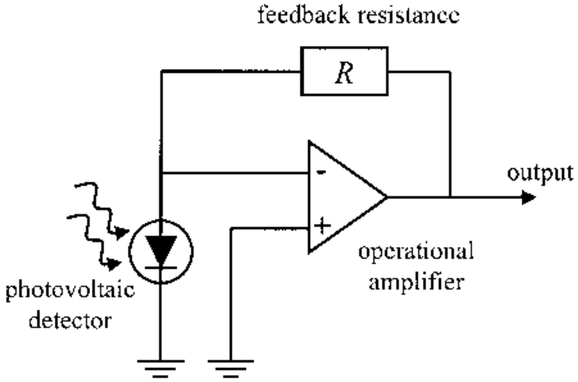


Fig. 20.3. Photovoltaic detector circuit using an operational amplifier in the feedback mode. When incident radiation is absorbed by the photovoltaic detector, a voltage is generated which is then collected through the shown circuit.

In such a detector, as a result of the built-in electric field in the depletion region, the photogenerated carriers drift to opposite sides of the depletion region: holes toward the *p*-type side and electrons toward the *n*-type side. There, they increase the majority carrier densities on both sides of a junction. An open-circuit voltage generated by this build-up of charge can then be measured. Fig. 20.3 depicts an example of the electrical circuit commonly used with photovoltaic detectors. No specific bias circuit is necessarily used in the photovoltaic detector operation.

The simple calculation conducted in the case of a photoconductor can be easily applied in the case of a photovoltaic detector to obtain an expression for the current responsivity:

Eq. (20.11) $R_i = \frac{q\lambda\eta}{hc}$

where the parameters have the same meaning as defined earlier.

A more precise calculation can be conducted as follows. Since we are considering a p-n junction, we can make use of the diode equation, which relates the current to the applied external voltage as obtained in Eq. (9.52) in Chapter 9. The current-voltage characteristic of a photovoltaic detector is illustrated in Fig. 20.4. The current given by this equation is termed the dark current, i.e. the current which would be flowing through the device without illumination. To obtain the total current across the detector, we must add to the dark current the photocurrent which is the component directly resulting from the photogenerated electron-hole pairs. The total current is then given by:

Eq. (20.12) $I = I_0\left(e^{\frac{qV}{k_bT}} - 1\right) - I_{ph}$

where V is applied external voltage, I_0 is the saturation current, and I_{ph} is the photocurrent, which is also called the short-circuit current and has the following expression:

Eq. (20.13) $I_{ph} = qAG(L_n + L_p)$

where A is the cross-sectional area of the p-n junction diode. G is the excess carrier generation rate. The gain of the device is usually considered to be unity for photodiodes, L_n and L_p are the diffusion lengths of electrons and holes, respectively. The current-voltage characteristic of a photodiode under varying degrees of illumination is depicted in Fig. 20.4.

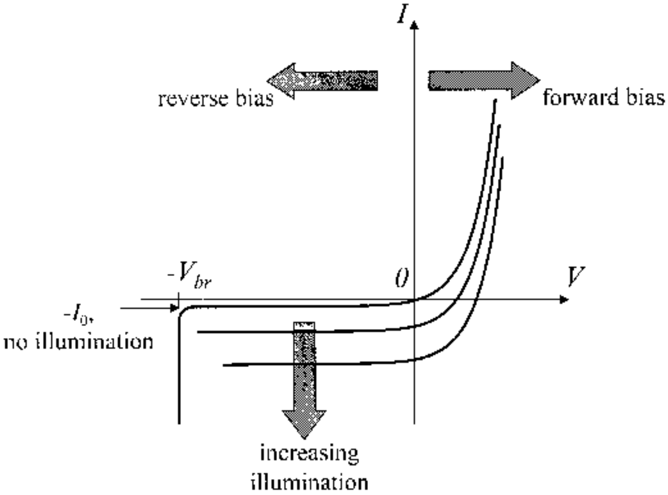


Fig. 20.4. Current-voltage characteristic of a photovoltaic detector with and without illumination.

The voltage that could be measured across the photodetector in an open-circuit situation can be found from Eq. (20.12) by setting I=0:

Eq. (20.14) $$V_{oc} = \frac{k_b T}{q} \ln\left(1 + \frac{I_{ph}}{I_0}\right)$$

Since a photovoltaic detector can operate without external voltage, one important characteristic of the detector is its differential resistance R_0 at zero bias, defined by:

Eq. (20.15) $$\frac{1}{R_0} = \left(\frac{dI}{dV}\right)_{V=0}$$

By differentiating Eq. (20.12) with respect to the voltage V, and calculating it at V=0, we can express the saturation current as a function of the differential resistance R_0:

Eq. (20.16) $$\frac{1}{R_0} = \frac{qI_0}{k_b T}$$

Using Eq. (20.13), we can express the ratio of the photocurrent to the saturation current which appears in Eq. (20.14):

Eq. (20.17) $$\frac{I_{ph}}{I_0} = \frac{qAG(L_n + L_p)}{k_bT \big/ qR_0} = \frac{q^2G(L_n + L_p)}{k_bT} R_0A$$

We therefore observe that this ratio is proportional to the product R_0A, which is a useful figure of merit for photovoltaic detectors.

A more simple analytical expression for the detectivity than the one given in Eq. (19.21) can be obtained in the case of a p-n junction photovoltaic detector when the thermal noise is dominant over all other noise sources:

Eq. (20.18) $$D^* = \frac{q\eta}{h\upsilon}\sqrt{\frac{R_0A}{4k_bT}}$$

in which all the terms have been defined previously. This equation directly relates the R_0A product to the thermally-limited detectivity.

Si, InSb, and HgCdTe are examples of materials commonly used for photovoltaic infrared photodetectors. Some of the advantages of photovoltaic detectors over photoconductive ones include a better theoretical signal-to-noise ratio, simpler biasing, and a more predictable responsivity. However, photovoltaic detectors are generally more fragile than photoconductors. Indeed, they are susceptible to electrostatic discharge and to physical damage during handling. In addition, because they are thin (10 μm for a backside illuminated InSb p-n junction), the insulating layers are susceptible to electrical breakdown. Surface effects may also lead to leakage between the *p*-type and the *n*-type regions, which can then degrade detector performance.

20.3. Examples of photon detectors

In addition to the simple photoconductive and p-n junction photovoltaic detectors discussed previously, there are other types of photon detectors which will be briefly described in this section, including the p-i-n photodiode, avalanche photodiode (APD), Schottky photodiode, photoelectromagnetic (PEM) detectors, and quantum well and dot detectors.

20.3.1. P-i-n photodiodes

A p-i-n photodiode consists of a p-n junction diode within which an undoped intrinsic or i-region is inserted between the doped regions. Because of the very low density of free carriers in the i-region and its high resistivity, any applied bias is dropped almost entirely across this i-layer, which is then fully depleted at low reverse bias. The p-i-n diode has a "controlled" depletion layer width, which can be tailored to meet the requirements of photoresponse and frequency bandwidth. The absorption and carrier generation processes in a p-i-n photodiode are shown in Fig. 20.5.

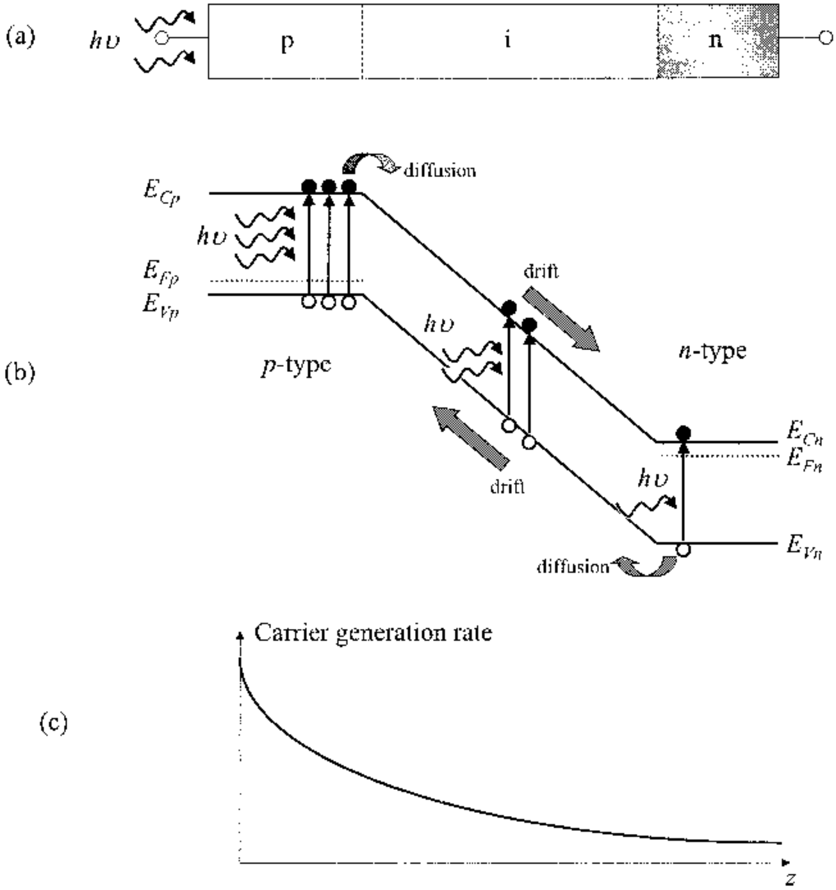


Fig. 20.5. (a) Schematic structure of a reverse-biased p-i-n diode, with the incident light arriving on the p-type side; (b) the absorption of photons creates electron-hole pairs in the p-type, n-type and the i-region where they then become spatially separated through the electric field across the space charge region; and (c) graph of the carrier generation.

For practical applications, photoexcitation is provided either through an etched opening in the top contact, or an etched hole in the substrate, as schematically shown in Fig. 20.6. The latter reduces the active area of the diode to the size of the incident light beam.

By careful choice of the material parameters and device design, large bandwidths can be attained using p-i-n photodetectors. The response speed and bandwidth are ultimately limited either by the transit time or by circuit parameters. The transit time of carriers across the depletion or i-layer depends on its width and the carrier velocity and can be reduced by making the i-layer more thin, at the possible expense of reducing the overall photosignal.

The key elements in achieving high-performance with a p-i-n diode (high quantum efficiency and large bandwidth) is to illuminate the diode through the substrate, also called back-side illumination, ensure the total depletion of the i-layer, and use the device at a low reverse bias. The latter is important for digital operation and for low-noise performance.

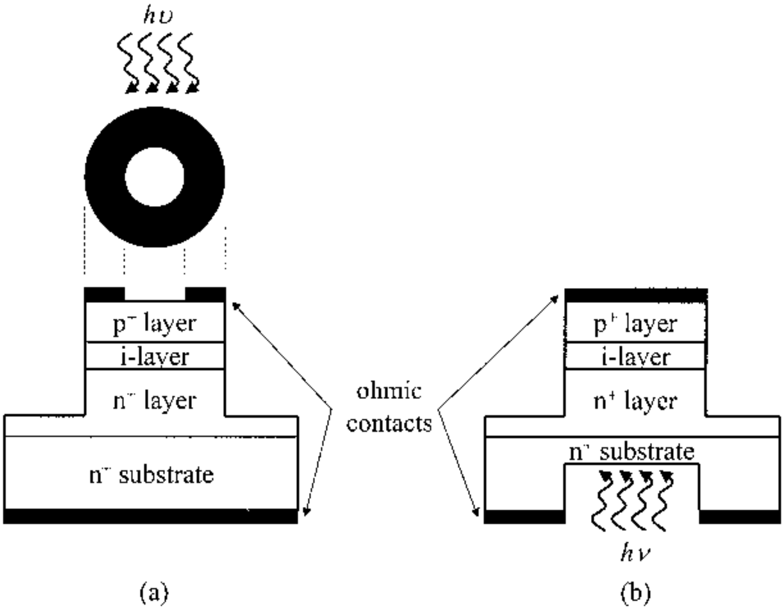


Fig. 20.6. Examples of mesa-etched p-i-n photodiodes for (a) top illumination and (b) back illumination. Top and back illuminations refer to the direction of the incident radiation to be detected with respect to the substrate on which the photodiode is fabricated.

20.3.2. Avalanche photodiodes

An avalanche photodiode (APD) operates by converting each absorbed photon into a cascade of electron-hole pairs. The device is a strongly reverse-biased photodiode in which the junction electric field is large. The charge carriers therefore accelerate in the space charge region, acquiring enough energy to generate additional electron-hole pair through impact ionization. This phenomenon has been somewhat discussed in the context of avalanche breakdown in a p-n junction in sub-section 9.4.4.

The avalanche multiplication process is illustrated in Fig. 20.7. A photon absorbed at point 1 creates one electron-hole pair. The electron accelerates in the strong electric field. The acceleration process is constantly interrupted by random collisions with the lattice in which the electron loses some of its acquired energy. These competing processes cause the electron to reach an average saturation velocity. When the electron can gain enough kinetic energy to ionize an atom, it creates a second electron-hole pair. This is called impact ionization (at point 2). The newly created electron and hole both acquire kinetic energy from the electric field and create additional electron-hole pairs (e.g. at point 3). These in turn fuel the process, creating other electron-hole pairs. This process is therefore called avalanche multiplication.

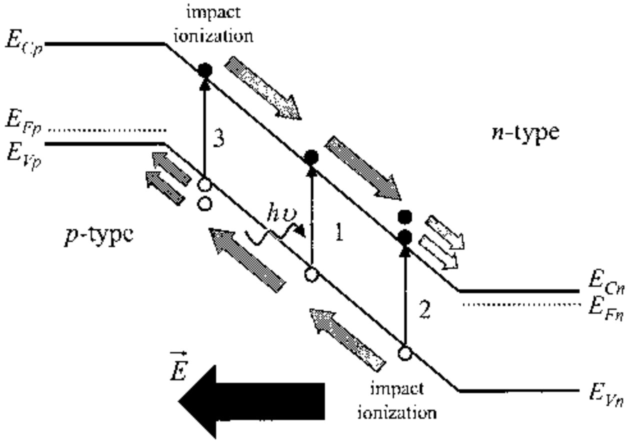


Fig. 20.7. Schematic representation of the multiplication process in an avalanche photodiode.

The abilities of electrons and holes to ionize atoms are characterized by their ionization coefficients α_e and α_h. These represent the ionization probabilities for electrons and holes per unit length. The ionization coefficients increase with the electric field in the depletion layer and decrease as the device temperature is raised. An important parameter which characterizes the performance of an APD is the ionization ratio $k=\alpha_h/\alpha_e$. If holes do not ionize effectively (k<<1), most of the ionization events are caused by electrons. The avalanching process then proceeds principally from left to right, i.e. from the p-type side to n-type side, in Fig. 20.7. If electrons and holes both ionize appreciably ($k\approx1$), the gain of the device, i.e. the total charge generated in the circuit per photogenerated carrier pair, increases. However, this situation is undesirable for several reasons: it is time consuming and therefore reduces the device bandwidth, it is a random process and therefore increases the device noise, and it may cause avalanche breakdown. It is therefore generally desirable to fabricate APDs from materials that permit only one type of carrier (either electrons or holes) to ionize effectively. The ideal case of single-carrier multiplication is achieved when k is 0 or ∞.

In an optimally designed photodiode, the geometry of the APD should maximize photon absorption and the multiplication region should be thin to minimize the possibility of localized uncontrolled avalanches.

Due to their speed and extreme sensitivity (they can easily count individual photons), avalanche photodiodes have been used extensively for commercial and military applications. Recently, APD devices have been used heavily in laser range finding or Light Detection And Ranging (LIDAR), which is a technique that allows measurements of distance, speed, rotation and even chemical composition and concentration. An aerial image from the LIDAR intensity data is shown in Fig. 20.8.

Fig. 20.8. LIDAR light intensity data. [From http://www.sbgmaps.com/lidar.htm. Reprinted with permission from Spencer B. Gross, Inc.]

20.3.3. Schottky barrier photodiodes

Schottky barrier photodiodes have been studied extensively and have found various applications. These devices have some advantages over p-n junction photodiodes such as their simplicity in fabrication, absence of high-temperature diffusion processes, and their high speed of response. The rectifying property of the metal-semiconductor junction, which is called a Schottky contact, has been reviewed in detail in sub-section 9.5.2. Briefly, the rectification arises from the presence of an electrostatic barrier between the metal and the semiconductor, which is due to the difference in work functions Φ_m and Φ_s of the metal and semiconductor, respectively, as shown in Fig. 20.9 for an *n*-type and a *p*-type semiconductor.

As also discussed in Chapter 9, the current transport across metal-semiconductor junctions is mainly due to the majority carriers, in contrast to p-n junctions where current transport is mainly due to minority carriers. It is now widely accepted that thermionic emission is the main process of carrier transport across Schottky barriers, and the current density is given by Eq. (A.36) in Appendix A.10. Knowing the current-voltage relationship, it is possible to calculate the R_0A product, as it was done in sub-section 20.2.2:

Eq. (20.19) $$\frac{1}{R_0A}=\frac{1}{A}\left(\frac{dI}{dV}\right)_{V=0}=\left(\frac{dJ}{dV}\right)_{V=0}$$

where A is the area of the Schottky junction. Using Eq. (A.36), we get:

Eq. (20.20) $$R_0A=\frac{k_bT}{qJ_{ST}}=\frac{k_b}{qA^*T}e^{\frac{q\Phi_B}{k_bT}}$$

where J_{ST} is the thermionic emission saturation current, k_b is the Boltzmann constant, A^* is the effective Richardson constant, and Φ_B is the Schottky potential barrier height for electron injection and is defined as:

Eq. (20.21) $$\Phi_B=\frac{\Phi_m-\chi}{q}$$

This quantity is illustrated in Fig. 20.9(a).

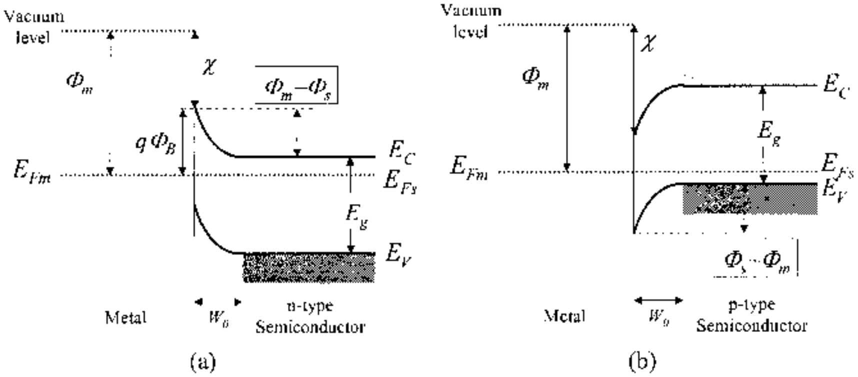


Fig. 20.9. Equilibrium energy band diagram of Schottky contacts: (a) metal-(n-type) semiconductor ($\Phi_m > \Phi_s$); (b) metal-(p-type) semiconductor ($\Phi_m < \Phi_s$). A Schottky contact is obtained in each case because the majority carriers in the semiconductor experience a potential barrier which prevents their free movement across the metal-semiconductor junction.

20.3.4. Metal-semiconductor-metal photodiodes

A metal-semiconductor-metal (MSM) photodiode consists of two Schottky contacts on an undoped semiconductor layer. Unlike a p-n junction diode, it uses a planar structure. It can be designed such that the region between the metal fingers is almost completely depleted. When the semiconductor absorbs an incident photon, an electron-hole pair is created. The electron and the hole are spatially separated and accelerated under the influence of the applied electric field until they reach the metal contacts where they enter the biasing electrical circuit. When a hole exits the semiconductor on one side of the device, another hole is injected from the opposite contact in order to maintain the overall electrical charge neutrality in the semiconductor. An illustration of the energy band diagram for a MSM photodiode with an applied bias V is shown in Fig. 20.10.

The frequency response and bandwidth of a MSM photodiode are determined primarily by the transit time of the photogenerated carriers and the charge-up time of the diode. The MSM photodiode exhibits gain, has a low dark current, a large bandwidth, and is amenable to simple and planar integration schemes.

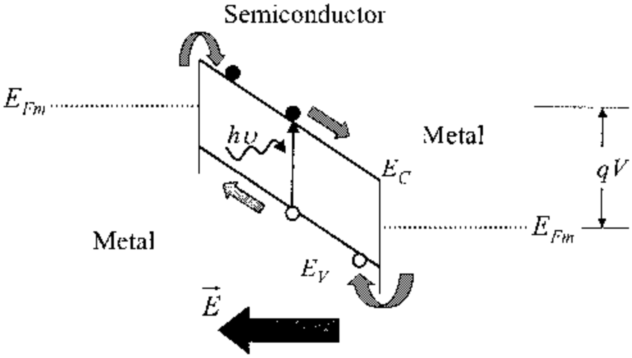


Fig. 20.10. Energy band diagram of an MSM photodiode under bias.

20.3.5. Type II superlattice photodetectors

The active layers of type II superlattice photodetectors are based on the type II band alignment in semiconductor heterojunctions as discussed in sub-section 11.2.2. The basic physical properties behind such heterojunctions have already been discussed in sub-section 18.5.13 and have been illustrated in the case of semiconductor lasers.

Fig. 20.11 shows a photoconductive InAs/GaSb type II superlattice detector and schematic diagram of the detection mechanisms inside the active region. Although the electrons and the holes are mostly confined in different layers, their wavefunctions can extend into the thin superlattice barriers. As a result, the overlap of the electron and hole wavefunctions is not strictly zero. The optical matrix element will have a high enough value to yield a considerable optical absorption only in the regions near the layer interfaces. Although this means a lower absorption than in type I quantum wells, the spatial separation of electrons and holes is advantageous in reducing the Auger recombination rate, which results in a longer lifetime of the photogenerated electron-hole pairs. This longer carrier lifetime, especially near room temperature, is the main advantage of this type of detector and has been experimentally demonstrated by Youngdale *et al.* [1994].

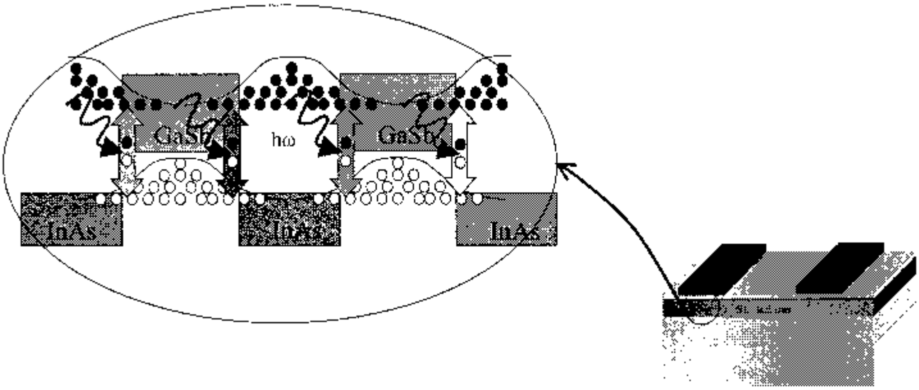


Fig. 20.11. Example of an InAs/GaSb type II photoconductive detector and the schematic diagram of the optical generation of electrons and holes in the active layer of the device.

Using type II superlattices as the narrow bandgap, active layer of a p-i-n photodiode, a detectivity of better than 10^{13} cm.IIz$^{1/2}$.W^{-1} has been recently achieved at 77 K with material designed for 5.4 μm cutoff wavelength [Walther *et al.* 2005]. Fig. 20.12 shows part of a focal plane array consisting of these devices. Other researchers at the Center for Quantum Devices have demonstrated good control of the cutoff wavelength of type II materials out to 32 μm [Wei *et al.* 2002]. At a wavelength of 8 μm, although the detectivity of HgCdTe is more than an order of magnitude higher than the type II photodetector, the latter still benefits from an easier growth process and a higher uniformity than HgCdTe, which can lead to less expensive focal plane arrays with comparable noise equivalent temperature difference (NEΔT) performance.

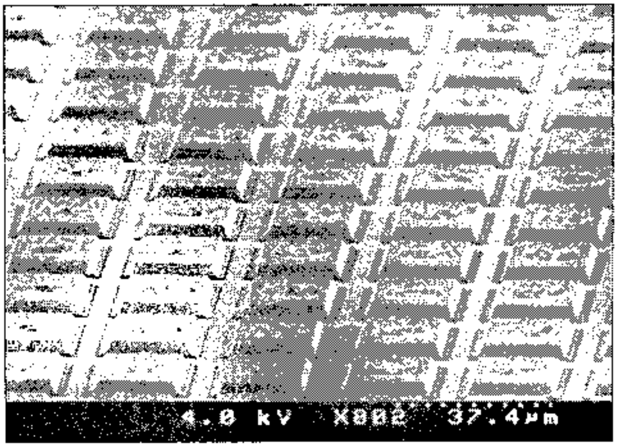


Fig. 20.12. Scanning electron microscope image of type II focal plane array pixels with indium bumps deposited on each pixel.

Currently, the only commercially available and fast room temperature photodetectors operating in the long wavelength infrared spectral region are based on HgCdTe and HgCdZnTe. In spite of their high detectivity, microbolometers have a time response which is at least three orders of magnitude slower than that of intrinsic detectors based on type II or HgCdTe. The new generation of HgCdZnTe detectors are now available with a special design to suppress the Auger recombination. Nevertheless, it has been shown that a type II superlattice with only 50 periods can have a higher detectivity at room temperature thanks to an lower Auger recombination rate (see Table 20.1 [Mohseni *et al.* 1998]).

Material	Type of detector	Wavelength (μm)	Detectivity (cm.Hz^{½}.W^{-1})
HgCdZnTe	Photovoltaic	10.6	1×10^7
HgCdTe	Photoelectromagnetic	10.6	5×10^6
HgCdZnTe	Photoconductive	10.6	6×10^6
Microbolometer	Thermal	8~14	5×10^8
Type II superlattice	Photoconductive	11	1×10^8

Table 20.1. Values of detectivity from typical infrared photon detectors at selected wavelengths at room temperature.

20.3.6. Photoelectromagnetic detectors

The origin of the photoelectromagnetic (PEM) effect is the diffusion of photogenerated carriers resulting from the carrier concentration gradient and their deflection in opposite directions due to a magnetic field, as shown in Fig. 20.13. If the sample ends are open circuited in the x-direction, a space charge builds up, giving rise to an electric field directed along the x-axis (open-circuit voltage). If the sample ends are short circuited in the x-direction, current flows through the circuit (short-circuit current).

The measured voltage or current can be directly related to the incident radiation which generated the carriers.

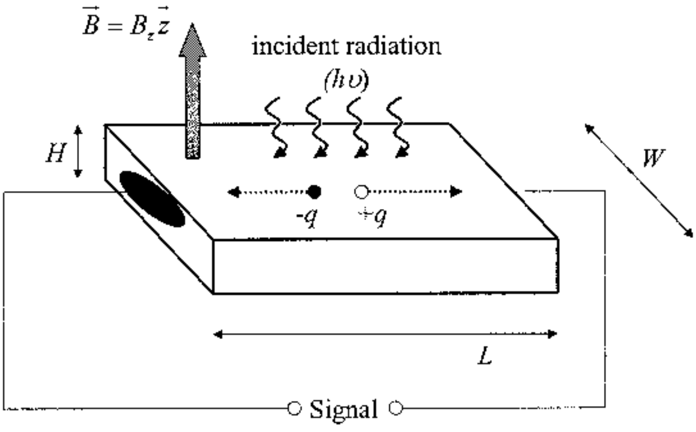


Fig. 20.13. Schematic of photoelectromagnetic effect.

20.3.7. Quantum well intersubband photodetectors

A quantum well intersubband photodetector or (QWIP) is a device whose operation is based on the absorption of photons through the intersubband transition of carriers which are confined in multiple quantum wells. The process is the reverse of that described in a quantum cascade laser in sub-section 18.5.11.

QWIPs have a narrow absorption spectrum that can be tailored to match optical transitions in the 3~20 μm wavelength range by adjusting the quantum well width and barrier height or barrier layer composition. More importantly, it can be made using mature III-V semiconductors based on gallium arsenide (GaAs) or indium phosphide (InP) substrates.

The study of intersubband optical transitions in doped multiple quantum wells was motivated by the possibility of realizing high-speed infrared photodetectors. Both conduction band (*n*-type) and valence band (*p*-type) based quantum wells have been studied, although the larger hole effective mass results in poorer responsivity for the *p*-type devices.

The schematic for designing QWIPs is shown in Fig. 20.14(a). The absorption of photons having energy equal to the intersubband separation leads to transition of carriers between these subbands. For example, for III-V quantum wells of width L=100 Å, the intersubband energy separation is in the range of 100~200 meV. The quantum mechanics selection rules allow absorption of electromagnetic radiation when the incident light polarization is parallel to the growth direction, i.e. TM polarization. This causes difficulties in detecting a two-dimensional image, since there is no effective absorption of light directed perpendicular to quantum well plane. However, illumination geometries using 45° facets as well as the use of surface gratings can circumvent the problem. For photoconductive devices utilizing intersubband absorption, the photogenerated carriers travel out of the quantum wells and contribute to the photocurrent.

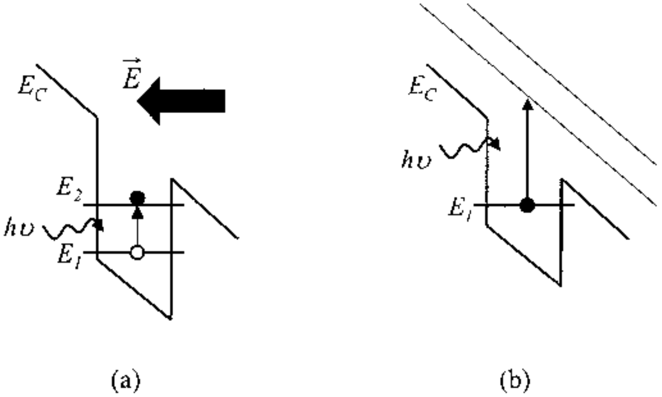


Fig. 20.14. Absorption of long-wavelength light in a quantum well due to (a) intersubband transitions in a wide well and (b) transition from a quasi-bound state to the continuum in a narrow well.

20.3.8. Quantum dot infrared photodetectors

Although quantum well intersubband photodetectors have been applied extensively in commercial and military applications, they suffer from low operating temperatures and non-normal incidence light absorption. Researchers have been developing quantum-dot infrared photodetectors (QDIP) in order to overcome these disadvantages. In practice, several layers of III-V quantum dots are grown on lattice-mismatched matrix spacers and intersubband photon absorption takes place within the dots themselves. Due to the three-dimensional confinement of carriers, QDIP devices offer several potential advantages including the absorption of normally incident light. Furthermore, the quantum dots greatly reduce the rate of electron-phonon relaxation transitions, which leads to a longer average lifetime of photoexcited carriers and better overall device responsivity. Finally, lower dark currents may eventually allow high temperature operation.

One of the challenges of the QDIP architecture is in controlling the quantum dot size and shape. Most conventional fabrication techniques involve Stranski-Krastanow random growth in molecular beam epitaxy (MBE) or metalorganic chemical vapor deposition (MOCVD) and the dot geometry depends strongly on such parameters as growth temperature, V-III ratio, growth rate, etc. To add another set of variables, researchers can adjust the doping of quantum dot and/or surrounding epitaxial layers in order to control the carrier population of the device, for example.

Recent efforts in this detector architecture have resulted in room temperature operation with photoresponse out to 17 μm with a specific detectivity of 1.5×10^7 cm.Hz^{½}.W^{-1} and 1 V bias voltage at room temperature [Bhattacharya *et al* 2005]. Future work in this field will probably involve improving the geometrical control of the nanometer-scale quantum dots as well as the periodicity and density of the dots.

20.4. Focal Plane Arrays

Solid state imaging systems consist of several different elements including the detector, optics, interconnects, and readout and clock circuitry. One type of detector architecture is the focal plane array (FPA). An FPA is an array of photodetectors placed at the focal plane of the imaging system's optics, which enables it to capture a two dimensional image. There are many application areas in which focal plane arrays are useful. These imagers find many uses in broadcast television, commercial photo and video devices (camcorders, for instance), machine vision, military and scientific applications. Arrays used for astronomy can boast pixel (individual image elements) numbers as large as a gigapixel, or one billion individual sensing elements!

Today, FPAs are available in monochromatic or multicolor systems, depending on the material type and wavelength range of interest. The most common types of imaging read-out architectures (essentially the manner in which the photosignal is handled within the device) include charge-coupled device (CCD) and complementary-metal-oxide-semiconductor (CMOS) arrays. One potential advantage to the CMOS design is the possibility of "per-pixel" signal processing, amplification and image correction.

Although focal plane array imagers are very common in our lives with products such as digital still and video cameras, they are quite complex to fabricate. Depending on the array architecture, the process can include over 150 individual fabrication steps. Contemporary visible imagers consist of silicon photodiodes integrated to an on-chip read-out integrated circuit (ROIC). When a detector substrate material different from the read-out circuit's silicon substrate is needed for different spectral regimes, the active sensing devices are often "hybridized" to the silicon-based read-out circuit. This hybridization process involves flip-chip indium bonding between the "top" surfaces of the ROIC and detector array. The indium bond must be uniform between each sensing pixel and its corresponding read-out element in order to insure high-quality imaging. After hybridization, a backside thinning process is usually performed to reduce the amount of substrate absorption. Some advanced FPA fabrication processes involve complete removal of the substrate material.

Fig. 20.15 shows two images taken using two different infrared FPAs. The left image was taken using a 256×256 hybridized infrared focal plane array with GaInAs/InP QWIP pixels operating at 8μm [Jiang *et al* 2003]. The right image was taken using a 256×256 hybridized FPA with InAs/GaSb type II superlattice pixels operating with a cutoff wavelength of 8μm [Razeghi *et al* 2005]. Both arrays were operated at liquid nitrogen temperature.

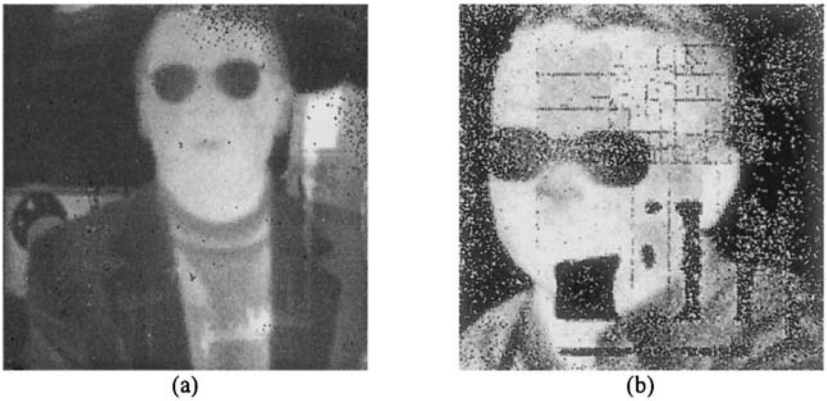


Fig. 20.15. Images from a: (a) GaInAs/InP QWIP and (b) InAs/GaSb type II superlattice (right) focal plane array camera operating in the 8-12 μm wavelength regime.

20.5. Summary

In this Chapter, we have explored the topic of photon detectors. In particular, photoconductive and photovoltaic detector types were discussed in detail. Specific photodetector examples were also described, including p-i-n, avalanche, Schottky barrier, metal-semiconductor-metal, type II superlattice, and photoelectromagnetic detectors, as well as quantum well and quantum dot photodetectors. Finally, the topic of focal plane arrays was investigated, including a brief overview of the complex fabrication process.

References

Bhattacharya, P., Su, X.H., Chakrabarti, S., Ariyawansa, G., Perera, A.G.U., "Characteristics of a tunneling quantum-dot infrared photodetector operating at room temperature," *Applied Physics Letters* 86, 191106, 2005.

Jiang, J., Mi, K., McClintock, R., Razeghi, M., Brown, G.J., and Jelen, C., "Demonstration of 256 × 256 focal plane array based on Al-free GaInAs-InP QWIP," *IEEE Photonics Technology Letters* 15, pp. 1273-1275, 2003.

Mohseni, H., Michel, E.J., Razeghi, M., Mitchel, W.C., and Brown, G.J., "Growth and characterization of InAs/GaSb type II superlattices for long wavelength infrared detectors," in *Photodetectors: Materials and Devices III*, ed. G.J. Brown, *SPIE Proceedings Series* 3287, SPIE-The International Society for Optical Engineering, Bellingham, WA, pp. 30-37, 1998.

Razeghi, M., Wei, Y., Gin, A., Hood, A., Yazdanpanah, V., Tidrow, M.Z., and Nathan, V., "High performance Type II InAs/GaSb superlattices for mid, long, and very long wavelength infrared focal plane arrays," in *Infrared Technology and Applications XXXI*, eds. B.F. Andresen and G.F. Fulop, *SPIE Proceedings Series* 5783, SPIE-The International Society for Optical Engineering, Bellingham, WA, pp. 86-97, 2005.

Wei, Y., Gin, A., Razeghi, M., Brown, G.J., "Advanced InAs/GaSb superlattice photovoltaic detectors for very long wavelength infrared applications," *Applied Physics Letters* 80, pp. 3262-3264, 2002.

Youngdale, E.R., Meyer, J.R., Hoffman, C.A., Bartoli, F.J., Grein, C.H., Young, P.M., Ehrenreich, H., Miles, R.H., and Chow, D.H., "Auger lifetime enhancement in InAs-$Ga_{1-x}In_xSb$ superlattices," *Applied Physics Letters* 64, pp. 3160-3162, 1994.

Further reading

Dereniak, E.L. and Crowe, D.G., *Optical Radiation Detectors*, John Wiley & Sons, New York, 1984.

Dereniak, E.L., and Boreman, G.D., *Infrared Detectors and Systems*, John Wiley & Sons, New York, 1996.

Holst, G.C., *CCD Arrays, Cameras and Displays, Second Edition*, SPIE Optical Engineering Press, Bellingham, WA, 1998.

Leigh, W.B., *Devices for Optoelectronics*, Marcel Dekker, New York, 1996.

Rogalski, A., *Infrared Photon Detectors*, SPIE Optical Engineering Press Bellingham, Washington, 1995.

Problems

1. What arc thc fundamcntal diffcrcnccs bctwccn a thermal and a photon (quantum) detector?

2. Assuming an interband absorption mechanism, what is thc cutoff wavelength for a semiconductor with a bandgap of 0.2 eV? What is it if thc bandgap is 0.02cV?

3. Calculate the electron transit time and device gain for a photoconductor undcr thesc conditions: distance between contacts=50 μm, applied voltage=10 V, minority carrier lifetime=2×10^{-7} s, $\mu_n = 10000cm^2/Vs$, and $\mu_p = 1000cm^2/Vs$.

4. Describe briefly the spectral characteristics of the following detectors and cxplain thc rcasons for thcir particular spcctral rcsponsc shapc.
   (a) InSb photovoltaic detector
   (b) Bolometer
   (c) Quantum well intersubband photodetector

5. Consider a hypothetical IIgCdTe photodiode with an active area of 500 μm×500 μm, R_0=500 MΩ, quantum efficiency=0.6 operating at 80 K. Calculatc thc dcvicc spccific dctcctivity at 8 μm wavclcngth.

6. List a few advantages and disadvantages of photovoltaic and photoconductivc dctcctors.

7. Explain thc possiblc advantagcs of a p-i-n photodiode over an abrupt pn design.

8. Describe the quantum selection rule as it pertains to n-type QWIP dcviccs. How do conventional QWIP dcviccs ovcrcomc this limitation? Can you list some more novel solutions to avoid this specific selection rulc?

9. List the advantages and drawbacks of quantum well and quantum dot intcrsubband photodctcctors.

10. Using knowledge gained from Chapter 16, describe the various processing steps that must be added to the conventional detector fabrication steps in order to generate a focal plane array. Assume pitch of 28 μm, InSb FPA with 640×320 pixels and Si ROIC. What are some of the complications added in the FPA hybridization process?

Appendix

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}
Caloric	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	$4\pi10^{-9} = 1.25633\times10^{-6}$ H.m^{-1}
Permittivity in vacuum	ε_0	8.85418×10^{-12} F.m^{-1} (=$1/\mu_0c^2$)
Plank's constant	h	6.62617×10^{-34} J.s
Reduced Plank'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 300K	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

A.2. International System of 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

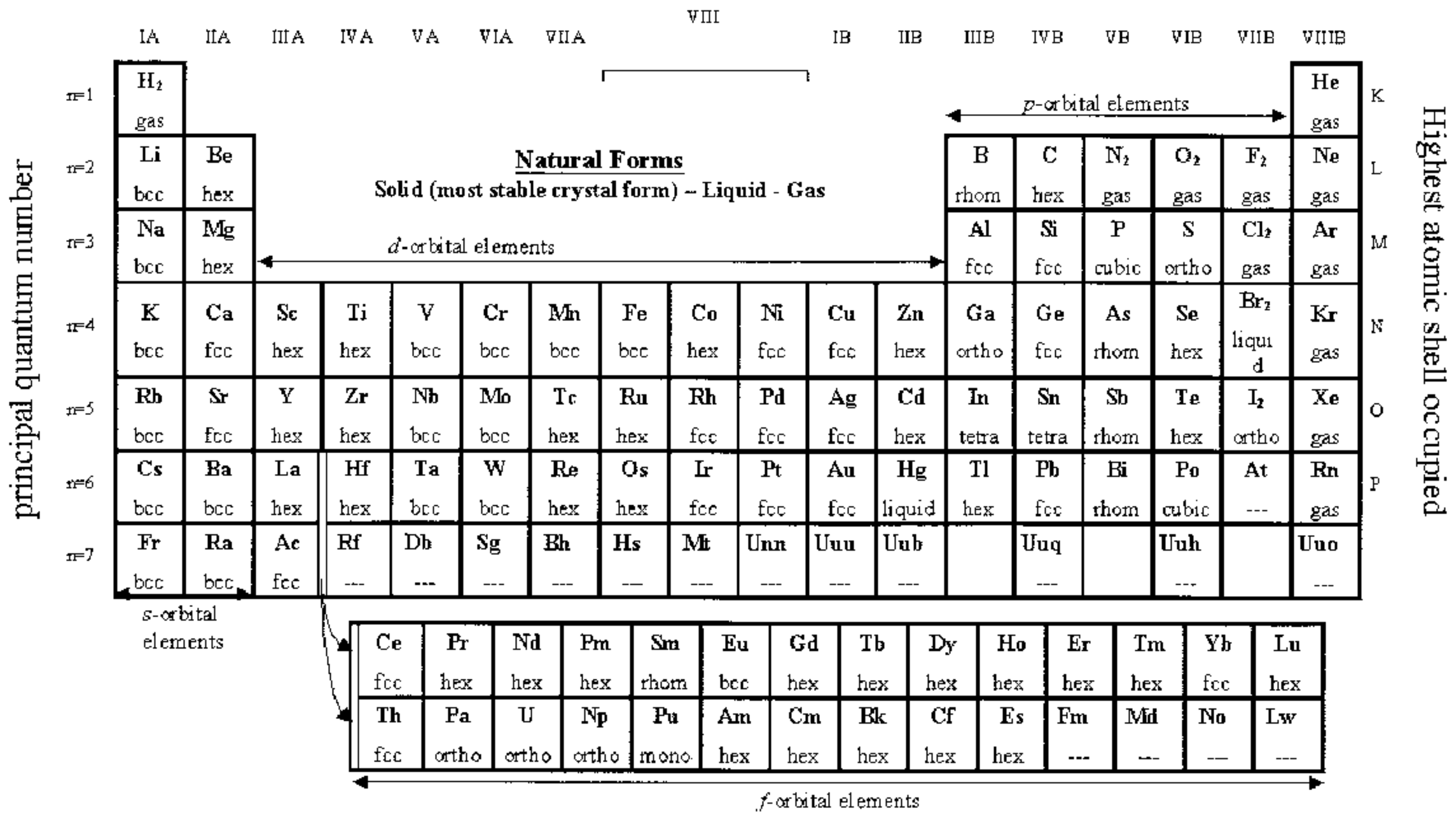
Factor	*Prefix*	*Symbol*	*Factor*	*Prefix*	*Symbol*
10^{24}	yotta	Y	10^{-1}	deci	d
10^{21}	zetta	Z	10^{-2}	centi	c
10^{18}	exa	E	10^{-3}	milli	m
10^{15}	peta	P	10^{-6}	micro	µ
10^{12}	tera	T	10^{-9}	nano	n
10^{9}	giga	G	10^{-12}	pico	p
10^{6}	mega	M	10^{-15}	femto	f
10^{3}	kilo	k	10^{-18}	atto	a
10^{2}	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.s^{-1}
Acceleration	-	-	m.s^{-2}
Angular velocity, frequency		-	rad.s^{-1}
Angular acceleration	-	-	rad.s^{-2}
Frequency	hertz	Hz	s^{-1}
Force	newton	N	kg.m.s^{-2}
Pressure, stress	pascal	Pa	N.m^{-2}
Work, energy, heat	joule	J	N.m, kg.m^{2}.s^{-2}
Power	watt	W	J.s
Electric charge	coulomb	C	A.s
Electric potential	volt	V	J.C^{-1}, W.A^{-1}
Resistance	ohm	Ω	V.A^{-1}
Conductance	siemens	S	A.V^{-1}, Ω^{-1}
Magnetic flux	weber	Wb	V.s
Inductance	henry	H	Wb.A^{-1}
Capacitance	farad	F	C.V^{-1}
Electric field strength	-	-	V.m^{-1}, N.C^{-1}
Magnetic induction	tesla	T	Wb.m^{-2}, N.A^{-1}.m^{-1}
Electric displacement	-	-	C.m^{-2}
Magnetic field strength	-	-	A.m^{-1}
Celsius temperature	degrees Celsius	°C	K
Luminous flux	lumen	lm	cd.sr
Illuminance	lux	lx	lm.m^{-2}
Radioactivity	becquerel	Bq	s^{-1}
Catalytic activity	katal	kat	mol.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).



Highest atomic shell occupied

principal quantum number

Natural Forms

Solid (most stable crystal form) -- Liquid - Gas

p-orbital elements
d-orbital elements
s-orbital elements
f-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	H_2 gas																	He gas	K
n=2	Li bcc	Be hex											B rhom	C hex	N_2 gas	O_2 gas	F_2 gas	Ne gas	L
n=3	Na bcc	Mg hex											Al fcc	Si fcc	P cubic	S ortho	Cl_2 gas	Ar gas	M
n=4	K bcc	Ca fcc	Sc hex	Ti hex	V bcc	Cr bcc	Mn bcc	Fe bcc	Co hex	Ni fcc	Cu fcc	Zn hex	Ga ortho	Ge fcc	As rhom	Se hex	Br_2 liquid	Kr gas	N
n=5	Rb bcc	Sr fcc	Y hex	Zr hex	Nb bcc	Mo bcc	Tc hex	Ru hex	Rh fcc	Pd fcc	Ag fcc	Cd hex	In tetra	Sn tetra	Sb rhom	Te hex	I_2 ortho	Xe gas	O
n=6	Cs bcc	Ba bcc	La hex	Hf hex	Ta bcc	W bcc	Re hex	Os hex	Ir fcc	Pt fcc	Au fcc	Hg liquid	Tl hex	Pb fcc	Bi rhom	Po cubic	At ---	Rn gas	P
n=7	Fr bcc	Ra bcc	Ac fcc	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

Ce fcc	Pr hex	Nd hex	Pm hex	Sm rhom	Eu bcc	Gd hex	Tb hex	Dy hex	Ho hex	Er hex	Tm hex	Yb fcc	Lu hex
Th fcc	Pa ortho	U ortho	Np ortho	Pu mono	Am hex	Cm hex	Bk hex	Cf hex	Es hex	Fm ---	Md ---	No ---	Lw ---

Fig. A.1. Natural forms of elements in the periodic table.

principal quantum number

Highest atomic shell occupied

Density: solid, liquid in g.cm^{-3} (20°C, 1 atm)
gas in g.l^{-1} (0°C, 1 atm)

s-orbital elements

d-orbital elements

p-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	H_2 0.089																	He 0.179	K
n=2	Li 0.53	Be 1.85											B 2.34	C 2.62	N_2 1.25	O_2 1.43	F_2 1.70	Ne 0.90	L
n=3	Na 0.97	Mg 1.74											Al 2.70	Si 2.33	P 1.82	S 2.07	Cl_2 3.21	Ar 1.78	M
n=4	K 0.86	Ca 1.55	Sc 3.0	Ti 4.54	V 6.11	Cr 7.19	Mn 7.44	Fe 7.87	Co 8.90	Ni 8.90	Cu 8.96	Zn 7.13	Ga 5.90	Ge 5.32	As 5.72	Se 4.80	Br_2 3.12	Kr 3.73	N
n=5	Rb 1.63	Sr 2.54	Y 4.50	Zr 6.50	Nb 8.57	Mo 10.2	Tc 11.5	Ru 12.4	Rh 12.4	Pd 12.0	Ag 10.5	Cd 8.65	In 7.31	Sn 7.3	Sb 6.69	Te 6.24	I_2 4.92	Xe 5.89	O
n=6	Cs 1.87	Ba 3.5	La 6.15	Hf 13.31	Ta 16.6	W 19.3	Re 21.0	Os 22.6	Ir 22.4	Pt 21.5	Au 19.3	Hg 13.55	Tl 11.85	Pb 11.4	Bi 9.8	Po 9.3	At ---	Rn 9.73	P
n=7	Fr ---	Ra 5.0	Ac 10.07	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

Ce 6.77	Pr 6.77	Nd 7.01	Pm 7.30	Sm 7.52	Eu 5.24	Gd 7.90	Tb 8.23	Dy 8.55	Ho 8.80	Er 9.07	Tm 9.32	Yb 6.90	Lu 9.84
Th 11.7	Pa 15.4	U 18.95	Np 20.3	Pu 19.8	Am 13.6	Cm 13.5	Bk ---	Cf ---	Es ---	Fm ---	Md ---	No ---	Lw ---

f-orbital elements

Fig. A.2. Density of mass of elements in the periodic table.

Boiling point (°C)

Liquid <-> Gas, 0°C, 1 atm; (sp= sublimation point)

Highest atomic shell occupied

principal quantum number

p-orbital elements

d-orbital elements

s-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	H_2 -253																	He -269	K
n=2	Li 1342	Be 2472											B 4002	C 3367 sp	N_2 -196	O_2 -183	F_2 -188	Ne -246	L
n=3	Na 883	Mg 1090											Al 2520	Si 3267	P 280	S 445	Cl_2 -34	Ar -186	M
n=4	K 759	Ca 1494	Sc 2836	Ti 3289	V 3409	Cr 2672	Mn 2062	Fe 2862	Co 2928	Ni 2914	Cu 2563	Zn 907	Ga 2205	Ge 2834	As 615	Se 685	Br_2 59	Kr -153	N
n=5	Rb 688	Sr 1382	Y 3338	Zr 4409	Nb 4744	Mo 4639	Tc 4265	Ru 4150	Rh 3697	Pd 2964	Ag 2163	Cd 767	In 2073	Sn 2603	Sb 1587	Te 988	I_2 184	Xe -108	O
n=6	Cs 671	Ba 1805	La 3464	Hf 4603	Ta 5458	W 5555	Re 5596	Os 5012	Ir 4428	Pt 3827	Au 2857	Hg 357	Tl 1473	Pb 1750	Bi 1564	Po 962	At 337	Rn -62	P
n=7	Fr 677	Ra ---	Ac 3200	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

Ce 3443	Pr 3520	Nd 3074	Pm 3000	Sm 1794	Eu 1527	Gd 3273	Tb 3230	Dy 2567	Ho 2700	Er 2868	Tm 1950	Yb 1196	Lu 3402
Th 4788	Pa 4134	U 3902	Np 3230	Pu ---	Am ---	Cm ---	Bk ---	Cf ---	Es ---	Fm ---	Md ---	No ---	Lw ---

f-orbital elements

Fig. A.3. Boiling point of elements in the periodic table.

Melting point (°C)

Solid <-> Liquid, 0°C, 1 atm;   (tp= triple point)

principal quantum number	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII	VIII	VIII	IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	Highest atomic shell occupied
n=1	H_2 -259tp																	He -272	K
n=2	Li 181	Be 1289											B 2092	C 3550	N_2 -210tp	O_2 -219tp	F_2 -220tp	Ne -249	L
n=3	Na 98	Mg 650											Al 661	Si 1414	P 44	S 115	Cl_2 -101tp	Ar -189tp	M
n=4	K 64	Ca 842	Sc 1541	Ti 1670	V 1910	Cr 1863	Mn 1246	Fe 1538	Co 1495	Ni 1445	Cu 1085	Zn 420	Ga 30tp	Ge 939	As 615sp	Se 221	Br_2 -7tp	Kr -157	N
n=5	Rb 40	Sr 769	Y 1522	Zr 1855	Nb 2469	Mo 2623	Tc 2204	Ru 2334	Rh 1963	Pd 1555	Ag 962	Cd 321	In 157	Sn 232	Sb 631	Te 450	I_2 114tp	Xe -112tp	O
n=6	Cs 28	Ba 729	La 918	Hf 2231	Ta 3020	W 3422	Re 3186	Os 3033	Ir 2447	Pt 1769	Au 1064	Hg -39	Tl 304	Pb 328	Bi 271	Po 254	At 302	Rn -71	P
n=7	Fr 27	Ra 700	Ac 1051	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

s-orbital elements

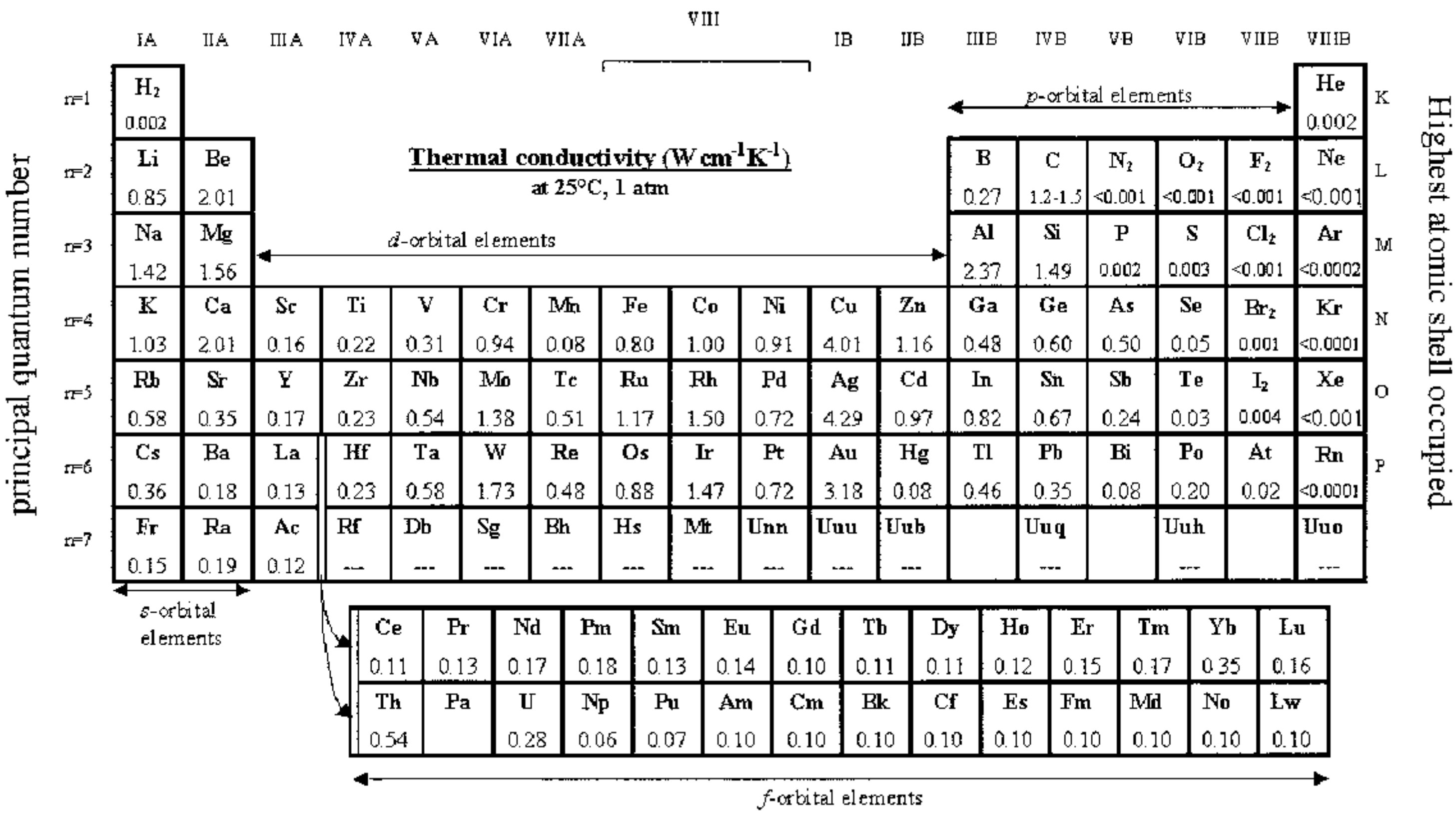
d-orbital elements

p-orbital elements

Ce 798	Pr 931	Nd 1021	Pm 1042	Sm 1074	Eu 822	Gd 1313	Tb 1356	Dy 1412	Ho 1474	Er 1529	Tm 1545	Yb 819	Lu 1663
Th 1755	Pa 1572	U 1135	Np 639	Pu 640	Am 1176	Cm 1345	Bk 1050	Cf 900	Es ---	Fm ---	Md ---	No ---	Lw ---

f-orbital elements

Fig. A.4. Melting point of elements in the periodic table.

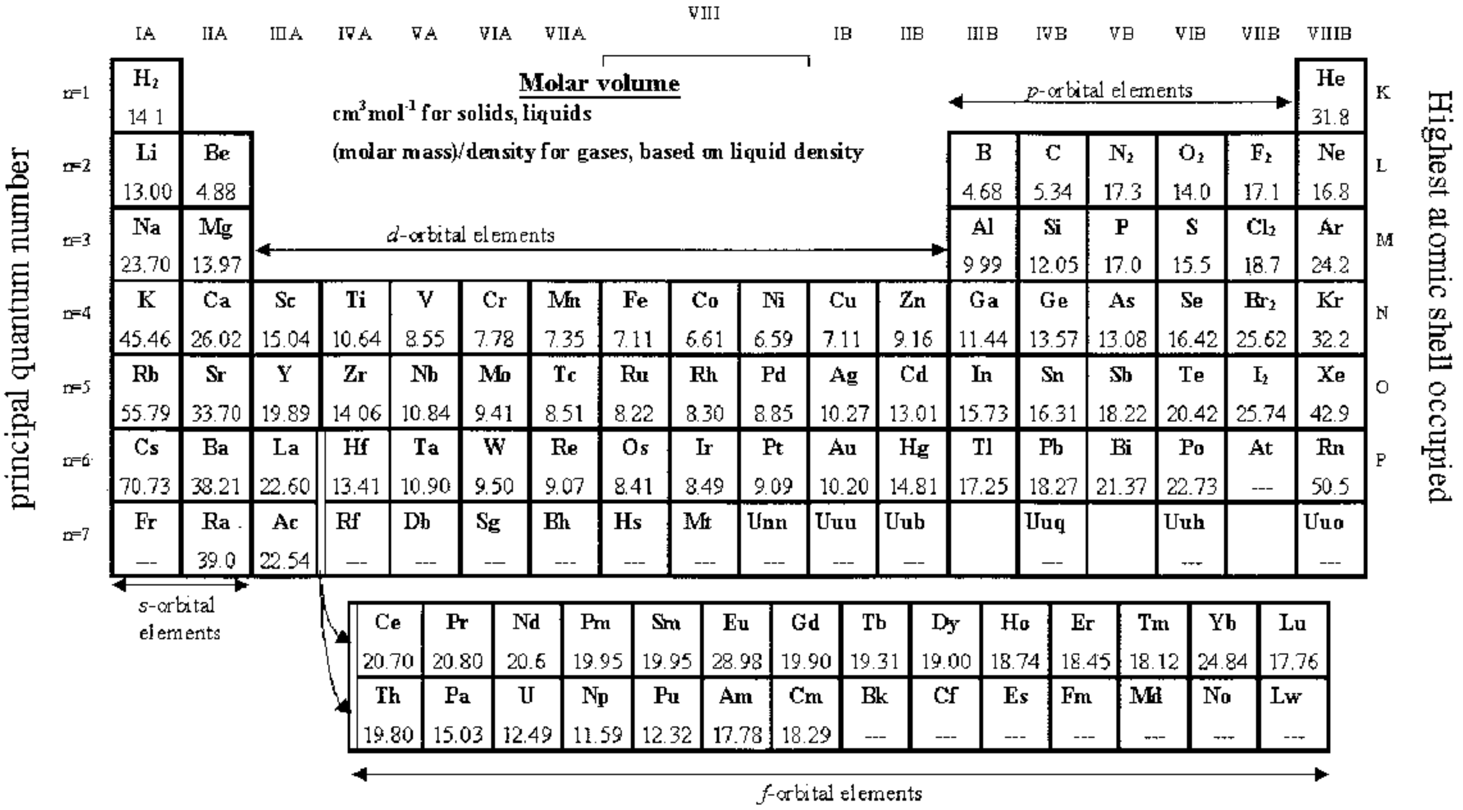

principal quantum number
Highest atomic shell occupied
p-orbital elements
d-orbital elements
s-orbital elements
f-orbital elements

Thermal conductivity (W cm^{-1}K^{-1})
at 25°C, 1 atm

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	H_2 0.002																	He 0.002	K
n=2	Li 0.85	Be 2.01											B 0.27	C 1.2-1.5	N_2 <0.001	O_2 <0.001	F_2 <0.001	Ne <0.001	L
n=3	Na 1.42	Mg 1.56											Al 2.37	Si 1.49	P 0.002	S 0.003	Cl_2 <0.001	Ar <0.0002	M
n=4	K 1.03	Ca 2.01	Sc 0.16	Ti 0.22	V 0.31	Cr 0.94	Mn 0.08	Fe 0.80	Co 1.00	Ni 0.91	Cu 4.01	Zn 1.16	Ga 0.48	Ge 0.60	As 0.50	Se 0.05	Br_2 0.001	Kr <0.0001	N
n=5	Rb 0.58	Sr 0.35	Y 0.17	Zr 0.23	Nb 0.54	Mo 1.38	Tc 0.51	Ru 1.17	Rh 1.50	Pd 0.72	Ag 4.29	Cd 0.97	In 0.82	Sn 0.67	Sb 0.24	Te 0.03	I_2 0.004	Xe <0.001	O
n=6	Cs 0.36	Ba 0.18	La 0.13	Hf 0.23	Ta 0.58	W 1.73	Re 0.48	Os 0.88	Ir 1.47	Pt 0.72	Au 3.18	Hg 0.08	Tl 0.46	Pb 0.35	Bi 0.08	Po 0.20	At 0.02	Rn <0.0001	P
n=7	Fr 0.15	Ra 0.19	Ac 0.12	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

Ce 0.11	Pr 0.13	Nd 0.17	Pm 0.18	Sm 0.13	Eu 0.14	Gd 0.10	Tb 0.11	Dy 0.11	Ho 0.12	Er 0.15	Tm 0.17	Yb 0.35	Lu 0.16
Th 0.54	Pa	U 0.28	Np 0.06	Pu 0.07	Am 0.10	Cm 0.10	Bk 0.10	Cf 0.10	Es 0.10	Fm 0.10	Md 0.10	No 0.10	Lw 0.10

Fig. A.5. Thermal conductivity of elements in the periodic table.


principal quantum number
Highest atomic shell occupied
p-orbital elements
d-orbital elements
s-orbital elements
f-orbital elements

Molar volume

cm^{3}mol^{-1} for solids, liquids

(molar mass)/density for gases, based on liquid density

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	H_2 14.1																	He 31.8	K
n=2	Li 13.00	Be 4.88											B 4.68	C 5.34	N_2 17.3	O_2 14.0	F_2 17.1	Ne 16.8	L
n=3	Na 23.70	Mg 13.97											Al 9.99	Si 12.05	P 17.0	S 15.5	Cl_2 18.7	Ar 24.2	M
n=4	K 45.46	Ca 26.02	Sc 15.04	Ti 10.64	V 8.55	Cr 7.78	Mn 7.35	Fe 7.11	Co 6.61	Ni 6.59	Cu 7.11	Zn 9.16	Ga 11.44	Ge 13.57	As 13.08	Se 16.42	Br_2 25.62	Kr 32.2	N
n=5	Rb 55.79	Sr 33.70	Y 19.89	Zr 14.06	Nb 10.84	Mo 9.41	Tc 8.51	Ru 8.22	Rh 8.30	Pd 8.85	Ag 10.27	Cd 13.01	In 15.73	Sn 16.31	Sb 18.22	Te 20.42	I_2 25.74	Xe 42.9	O
n=6	Cs 70.73	Ba 38.21	La 22.60	Hf 13.41	Ta 10.90	W 9.50	Re 9.07	Os 8.41	Ir 8.49	Pt 9.09	Au 10.20	Hg 14.81	Tl 17.25	Pb 18.27	Bi 21.37	Po 22.73	At ---	Rn 50.5	P
n=7	Fr ---	Ra 39.0	Ac 22.54	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

Ce 20.70	Pr 20.80	Nd 20.6	Pm 19.95	Sm 19.95	Eu 28.98	Gd 19.90	Tb 19.31	Dy 19.00	Ho 18.74	Er 18.45	Tm 18.12	Yb 24.84	Lu 17.76
Th 19.80	Pa 15.03	U 12.49	Np 11.59	Pu 12.32	Am 17.78	Cm 18.29	Bk ---	Cf ---	Es ---	Fm ---	Md ---	No ---	Lw ---

Fig. A.6. Molar volume of elements in the periodic table.

Specific heat ($Jg^{-1}K^{-1}$)
at 25°C, 1 atm

principal quantum number

Highest atomic shell occupied

s-orbital elements

d-orbital elements

p-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	**H_2** 14.30																	**He** 5.19	K
n=2	**Li** 3.58	**Be** 1.83											**B** 1.03	**C** 0.71	**N_2** 1.04	**O_2** 0.918	**F_2** 0.82	**Ne** 1.03	L
n=3	**Na** 1.23	**Mg** 1.02											**Al** 0.90	**Si** 0.71	**P** 0.77	**S** 0.71	**Cl_2** 0.48	**Ar** 0.52	M
n=4	**K** 0.76	**Ca** 0.65	**Sc** 0.568	**Ti** 0.52	**V** 0.49	**Cr** 0.45	**Mn** 0.48	**Fe** 0.45	**Co** 0.42	**Ni** 0.44	**Cu** 0.39	**Zn** 0.39	**Ga** 0.37	**Ge** 0.32	**As** 0.33	**Se** 0.32	**Br_2** 0.27	**Kr** 0.25	N
n=5	**Rb** 0.36	**Sr** 0.30	**Y** 0.30	**Zr** 0.28	**Nb** 0.27	**Mo** 0.25	**Tc** 0.21	**Ru** 0.24	**Rh** 0.24	**Pd** 0.24	**Ag** 0.24	**Cd** 0.23	**In** 0.23	**Sn** 0.23	**Sb** 0.21	**Te** 0.20	**I_2** 0.15	**Xe** 0.16	O
n=6	**Cs** 0.24	**Ba** 0.20	**La** 0.20	**Hf** 0.14	**Ta** 0.14	**W** 0.13	**Re** 0.14	**Os** 0.130	**Ir** 0.13	**Pt** 0.13	**Au** 0.13	**Hg** 0.14	**Tl** 0.13	**Pb** 0.13	**Bi** 0.12	**Po** ---	**At** ---	**Rn** 0.09	P
n=7	**Fr** ---	**Ra** ---	**Ac** 0.12	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---	

Ce 0.192	**Pr** 0.193	**Nd** 0.190	**Pm** ---	**Sm** 0.20	**Eu** 0.18	**Gd** 0.24	**Tb** 0.18	**Dy** 0.17	**Ho** 0.17	**Er** 0.17	**Tm** 0.16	**Yb** 0.16	**Lu** 0.15
Th 0.11	**Pa** ---	**U** 0.116	**Np** 0.12	**Pu** 0.13	**Am** 0.11	**Cm** ---	**Bk** ---	**Cf** ---	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

f-orbital elements

Fig. A.7. Specific heat of elements in the periodic table.

Atomic radius (0.01 Å)

principal quantum number	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	Highest atomic shell occupied
n=1	H 37																	He —	K
n=2	Li 152	Be 111											B 80	C 77	N 74	O 74	F 71	Ne ---	L
n=3	Na 186	Mg 160											Al 143	Si 118	P 110	S 103	Cl 99	Ar ---	M
n=4	K 227	Ca 197	Sc 161	Ti 145	V 131	Cr 125	Mn 137	Fe 124	Co 125	Ni 125	Cu 128	Zn 133	Ga 122	Ge 123	As 125	Se 116	Br 114	Kr ---	N
n=5	Rb 248	Sr 215	Y 178	Zr 159	Nb 143	Mo 136	Tc 135	Ru 133	Rh 135	Pd 138	Ag 145	Cd 149	In 163	Sn 141	Sb 145	Te 143	I 133	Xe ---	O
n=6	Cs 265	Ba 217	La 187	Hf 156	Ta 143	W 137	Re 137	Os 134	Ir 136	Pt 139	Au 144	Hg 150	Tl 170	Pb 175	Bi 155	Po 118	At —	Rn —	P
n=7	Fr ---	Ra ---	Ac 188	Rf ---	Db —	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq —		Uuh ---		Uuo ---	

s-orbital elements (IA–IIA)

d-orbital elements (IIIA–IIB)

p-orbital elements (IIIB–VIIIB)

Ce 183	Pr 182	Nd 181	Pm 181	Sm 180	Eu 199	Gd 179	Tb 176	Dy 175	Ho 174	Er 173	Tm 173	Yb 194	Lu 172
Th 180	Pa 161	U 139	Np ---	Pu 151	Am 131	Cm ---	Bk ---	Cf ---	Es —	Fm —	Md ---	No ---	Lw ---

f-orbital elements

Fig. A.8. Atomic radius of elements in the periodic table.

Highest atomic shell occupied

Oxidation states

principal quantum number

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	H 1																	He ---	K
n=2	Li 1	Be 2											B 3	C 4,2	N ±3,5,4,2	O -2,-1	F -1	Ne ---	L
n=3	Na 1	Mg 2											Al 3	Si 4	P ±3,5,4	S ±2,4,6	Cl ±1,3,5,7	Ar ---	M
n=4	K 1	Ca 2	Sc 3	Ti 4	V 5,3	Cr 6,3,2	Mn 7,6,4,2,3	Fe 2,3	Co 2,3	Ni 2,3	Cu 2,1	Zn 2	Ga 3	Ge 4	As ±3,5	Se ±2,4,6	Br ±1,5	Kr —	N
n=5	Rb 1	Sr 2	Y 3	Zr 4	Nb 5,3	Mo 6,5,4,3,2	Tc 7	Ru 2,3,4,6,8	Rh 2,3,4	Pd 2,4	Ag 1	Cd 2	In 3	Sn 4,2	Sb ±3,5	Te ±2,4,6	I ±1,5,7	Xe —	O
n=6	Cs 1	Ba 2	La 3	Hf 4	Ta 5	W 6,5,4,3,2	Re 7,6,4,2,-1	Os 2,3,4,6,8	Ir 2,3,4,6	Pt 2,4	Au 3,1	Hg 2,1	Tl 3,1	Pb 4,2	Bi 3,5	Po 4,2	At ±1,3,5,7	Rn ---	P
n=7	Fr 1	Ra 2	Ac 3	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

s-orbital elements

d-orbital elements

p-orbital elements

Ce 3,4	Pr 3,4	Nd 3	Pm 3	Sm 3,2	Eu 3,2	Gd 3	Tb 3,4	Dy 3	Ho 3	Er 3	Tm 3,2	Yb 3,2	Lu 3
Th 4	Pa 5,4	U 6,5,4,3	Np 6,5,4,3	Pu 6,5,4,3	Am 6,5,4,3	Cm 3	Bk 4,3	Cf —	Es ---	Fm —	Md ---	No ---	Lw ---

f-orbital elements

Fig. A.9. Oxidation states of elements in the periodic table.

Highest atomic shell occupied

principal quantum number

Ionic radius (0.01 Å) (for atoms with multiple oxidation states, the oxidation state considered for the ionic radius is given in parenthesis)

s-orbital elements

d-orbital elements

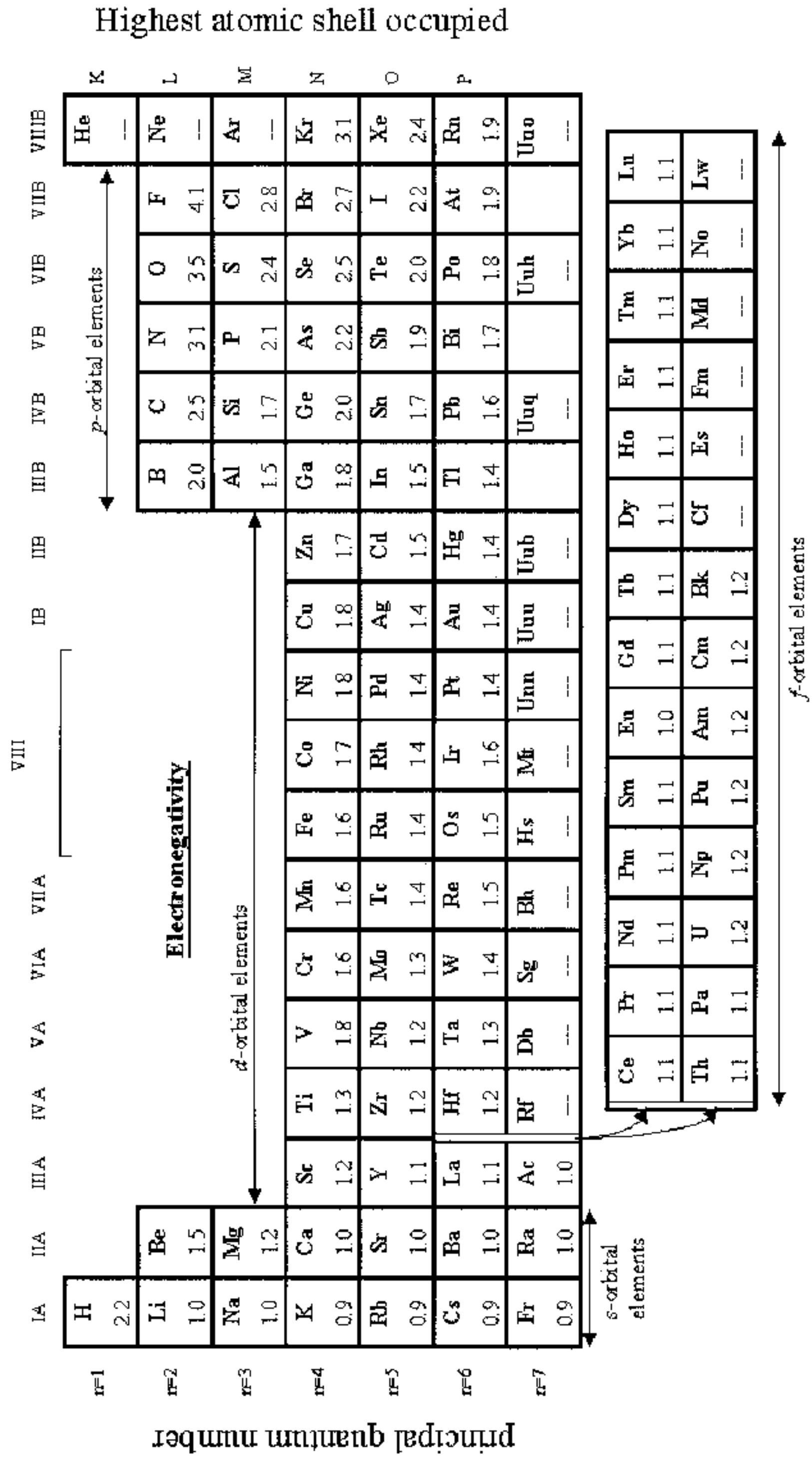
p-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	**H** ---																	**He** ---	K
n=2	**Li** 76	**Be** 45											**B** ---	**C** ---	**N** ---	**O** (-2) 140	**F** 133	**Ne** ---	L
n=3	**Na** 102	**Mg** 72											**Al** 54	**Si** 26	**P** (+5) 17	**S** (-2) 184	**Cl** (-1) 181	**Ar** ---	M
n=4	**K** 151	**Ca** 100	**Sc** 75	**Ti** 61	**V** (+5) 54	**Cr** (+3) 62	**Mn** (+2) 67	**Fe** (+3) 55	**Co** (+2) 65	**Ni** (+2) 69	**Cu** (+2) 73	**Zn** 74	**Ga** 62	**Ge** 53	**As** (+3) 58	**Se** (-2) 198	**Br** (-1) 196	**Kr** ---	N
n=5	**Rb** 161	**Sr** 126	**Y** 102	**Zr** 84	**Nb** (+5) 64	**Mo** (+6) 59	**Tc** ---	**Ru** (+3) 68	**Rh** (+3) 67	**Pd** (+2) 64	**Ag** 115	**Cd** 95	**In** 80	**Sn** (+4) 45	**Sb** (+3) 76	**Te** (-2) 107	**I** (-1) 220	**Xe** ---	O
n=6	**Cs** 174	**Ba** 142	**La** 116	**Hf** 83	**Ta** 64	**W** (+6) 60	**Re** (+7) 53	**Os** (+4) 63	**Ir** (+4) 63	**Pt** (+4) 63	**Au** (+3) 85	**Hg** (+2) 102	**Tl** (+1) 159	**Pb** (+2) 119	**Bi** (+3) 103	**Po** ---	**At** ---	**Rn** ---	P
n=7	**Fr** ---	**Ra** 162	**Ac** ---	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---	

Ce (+3) 114	**Pr** (+3) 113	**Nd** 114	**Pm** 109	**Sm** (+3) 108	**Eu** (+3) 107	**Gd** 105	**Tb** (+3) 118	**Dy** 103	**Ho** ---	**Er** 100	**Tm** (+3) 109	**Yb** (+3) 99	**Lu** 98
Th 105	**Pa** ---	**U** (+6) 81	**Np** ---	**Pu** ---	**Am** ---	**Cm** ---	**Bk** ---	**Cf** ---	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

f-orbital elements

Fig. A.10. Ionic radius of elements in the periodic table.


Highest atomic shell occupied
principal quantum number
Electronegativity
p-orbital elements
d-orbital elements
s-orbital elements
f-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	H 2.2																	He —	K
n=2	Li 1.0	Be 1.5											B 2.0	C 2.5	N 3.1	O 3.5	F 4.1	Ne —	L
n=3	Na 1.0	Mg 1.2											Al 1.5	Si 1.7	P 2.1	S 2.4	Cl 2.8	Ar —	M
n=4	K 0.9	Ca 1.0	Sc 1.2	Ti 1.3	V 1.8	Cr 1.6	Mn 1.6	Fe 1.6	Co 1.7	Ni 1.8	Cu 1.8	Zn 1.7	Ga 1.8	Ge 2.0	As 2.2	Se 2.5	Br 2.7	Kr 3.1	N
n=5	Rb 0.9	Sr 1.0	Y 1.1	Zr 1.2	Nb 1.2	Mo 1.3	Tc 1.4	Ru 1.4	Rh 1.4	Pd 1.4	Ag 1.4	Cd 1.5	In 1.5	Sn 1.7	Sb 1.9	Te 2.0	I 2.2	Xe 2.4	O
n=6	Cs 0.9	Ba 1.0	La 1.1	Hf 1.2	Ta 1.3	W 1.4	Re 1.5	Os 1.5	Ir 1.6	Pt 1.4	Au 1.4	Hg 1.4	Tl 1.4	Pb 1.6	Bi 1.7	Po 1.8	At 1.9	Rn 1.9	P
n=7	Fr 0.9	Ra 1.0	Ac 1.0	Rf —	Db —	Sg —	Bh —	Hs —	Mt —	Unn —	Uuu —	Uub —		Uuq —		Uuh —		Uuo —	

Ce 1.1	Pr 1.1	Nd 1.1	Pm 1.1	Sm 1.1	Eu 1.0	Gd 1.1	Tb 1.1	Dy 1.1	Ho 1.1	Er 1.1	Tm 1.1	Yb 1.1	Lu 1.1
Th 1.1	Pa 1.1	U 1.2	Np 1.2	Pu 1.2	Am 1.2	Cm 1.2	Bk 1.2	Cf —	Es —	Fm —	Md —	No —	Lw —

Fig. A.11. Electronegativity of elements in the periodic table.

Electron affinity (eV)
(N/S = not stable)

s-orbital elements

d-orbital elements

p-orbital elements

principal quantum number	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB
n=1	**H** 0.75																	**He** N/S
n=2	**Li** 0.62	**Be** N/S											**B** 0.28	**C** 1.26	**N** N/S	**O** 1.46	**F** 3.40	**Ne** N/S
n=3	**Na** 0.55	**Mg** N/S											**Al** 0.44	**Si** 1.39	**P** 0.75	**S** 2.08	**Cl** 3.61	**Ar** N/S
n=4	**K** 0.50	**Ca** 0.04	**Sc** 0.19	**Ti** 0.08	**V** 0.53	**Cr** 0.67	**Mn** N/S	**Fe** 0.151	**Co** 0.66	**Ni** 1.16	**Cu** 1.24	**Zn** N/S	**Ga** 0.3	**Ge** 1.23	**As** 0.81	**Se** 2.02	**Br** 3.36	**Kr** N/S
n=5	**Rb** 0.49	**Sr** 0.11	**Y** 0.31	**Zr** 0.43	**Nb** 0.90	**Mo** 0.75	**Tc** 0.55	**Ru** 1.05	**Rh** 1.14	**Pd** 0.56	**Ag** 1.30	**Cd** N/S	**In** 0.30	**Sn** 1.11	**Sb** 1.07	**Te** 1.97	**I** 3.06	**Xe** N/S
n=6	**Cs** 0.47	**Ba** 0.15	**La** 0.5	**Hf** ~0	**Ta** 0.32	**W** 0.86	**Re** 0.15	**Os** 1.10	**Ir** 1.57	**Pt** 2.13	**Au** 2.31	**Hg** N/S	**Tl** 0.2	**Pb** 0.36	**Bi** 0.95	**Po** 1.9	**At** 2.8	**Rn** N/S
n=7	**Fr** 0.46	**Ra** ---	**Ac** ---	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---

Ce ---	**Pr** ---	**Nd** ---	**Pm** ---	**Sm** ---	**Eu** ---	**Gd** ---	**Tb** ---	**Dy** ---	**Ho** ---	**Er** ---	**Tm** ---	**Yb** ---	**Lu** ---
Th ---	**Pa** ---	**U** ---	**Np** ---	**Pu** ---	**Am** ---	**Cm** ---	**Bk** ---	**Cf** ---	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

Fig. A.12. Electron affinity of elements in the periodic table.

A.4. Physical properties of important semiconductors

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

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	holes
				m_e	m_h
Element	C	1800	1200	0.2	0.25
	Si	1500	450	0.98[a]	0.16[c]
				0.19[b]	0.49[d]
	Ge	3900	1900	1.64[a]	0.04[c]
				0.082[b]	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
	GaSb	5000	850	0.42	0.04[c]
					0.4[d]
	InAs	33000	460	0.023	0.40
	InSb	80000	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:

Eq. (A.1)
$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots$$
$$\dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + R_n$$

where R_n is called the remainder and is equal to:

Eq. (A.2)
$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}$$

for an appropriate value ξ such that $|\xi - x_0| \le |x - x_0|$.

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.1). 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:

Eq. (A.3)
$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Eq. (A.4) $$\sin x \quad = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + - \cdots \qquad = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$$

Eq. (A.5) $$\cos x \quad = 1 - \frac{x^2}{2!} + \frac{x^3}{3!} - \frac{x^5}{5!} + - \cdots \qquad = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}$$

Eq. (A.6) $$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - + \cdots \qquad = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^n}{n}$$

There exist convergence ranges in evaluating the infinite sums in Eq. (A.3) to Eq. (A.6). 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:

Eq. (A.7) $$f(t)=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left(a_n\cos(nwt)+b_n\sin(nwt)\right)$$

where we have denoted $w=\frac{2\pi}{T}$, and:

Eq. (A.8) $$a_0=\frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}f(t)dt$$

Eq. (A.9) $$a_n=\frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}f(t)\cos(nwt)dt$$

Eq. (A.10) $$b_n=\frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}f(t)\sin(nwt)dt$$

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.13, is:

Eq. (A.11)

$$f(t)=\frac{\tau}{T}+\sum_{n=1}^{\infty}\frac{1}{n\pi}\left[\sin nw\tau\cos nwt+(1-\cos nw\tau)\sin nwt\right]$$

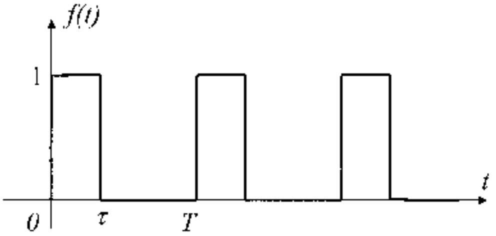


Fig. A.13. Example of periodic function, used to illustrate the concept of Fourier series.

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:

Eq. (A.12) $$F(w)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}f(t)e^{-iwt}dt$$

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.7), 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:

Eq. (A.13) $$f(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}F(w)e^{iwt}dw$$

For example, the Fourier transform of the function shown in Fig. A.14 is:

Eq. (A.14) $$F(w)=\frac{1-e^{-iwT}}{iw\sqrt{2\pi}}$$

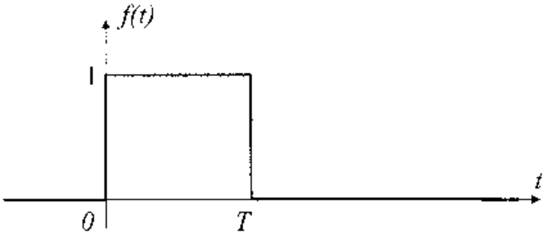


Fig. A.14. Example of an arbitrarily chosen function used to illustrate the concept of Fourier transform.

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 Chapter 3 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 eigenfunctions solution of the Schrödinger equation.

This procedure then makes the envelope wavefunction of the Bloch wave $u_{\vec{k}}$ in:

Eq. (A.15) $$\Psi_{\vec{k}}(\vec{r}) = u_{\vec{k}}(\vec{r})e^{i\vec{k}.\vec{r}}$$

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 Chapter 4. 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 electrons 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 $b_{n\dot{k}} = \sum_{\dot{s}} e^{i\bar{k}.s} b_n(\vec{r}-\vec{s})$ 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

Eq. (A.16) $$\Psi_{\vec{k}} = \sum_{\vec{g}} \alpha_{\vec{k}-\vec{g}} \chi_{\vec{k}-\vec{g}}$$

Eq. (A.17) $$\chi_{\vec{k}} = e^{i\vec{k}.\vec{r}} - \sum_j a_j b_{j\vec{k}}$$

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, give 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 be 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., *Energy Band Theory*, Academic Press, New York, 1964.

Chuang, S.L., *Physics of Optoelectronic Devices*, John Wiley & Sons, New York, 1995.

Harrison, W.A., *Pseudopotentials in the Theory of Metals*, W.A. Benjamin, New York, 1966.

Ziman, J.M., *Principles of the Theory of Solids*, Cambridge University Press, Cambridge, 1998.

A.8. The Kane effective mass method

In the Chapter 3 on energy band structures, we made the observation (and indeed used this later also throughout the book) that the band dispersion in most semiconductors near $\vec{k} = \vec{0}$ could be approximated as a parabola in $\vec{k}$ but with an effective mass which is determined by rigorous band structure calculation. One finds in practice that the scheme works very well, and that true effective masses can be very different from the free electron masses. From the "exact results" shown in Chapter 4, 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 bandgap for example. Also it would be nice to have a scheme, which could predict the effective mass, which was versatile, and which could be applied to confined, and multilayer structures as well. Some years ago 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 of the effective mass near the $\vec{k} = \dot{0}$ points in semiconductors, and a correlation between the effective mass and the bandgap.

Kane's method is a brilliant example on how one "piece of information", normally obtained and obtainable by experiment, can be used to derive another piece of information using the logical structure of the theory. The Kane argument goes as follows.

Consider the full Hamiltonian and Schrödinger equation of the electron in the periodic potential $V(\vec{r})$ of the lattice. Now assume that the wavefunction is a Bloch wave:

Eq. (A.18) $\psi_{n\vec{k}}(\vec{r}) = u_{n\vec{k}}(\vec{r})e^{i\vec{k}.\vec{r}}$

with energy $E_n(\vec{k})$ .We know that this must be true, so we substitute it in the Schrödinger equation:

Eq. (A.19) $[\frac{p^2}{2m_0} + V(\vec{r})]\Psi_{n\vec{k}}(\vec{r}) = E_{n\vec{k}}\Psi_{n\vec{k}}(\vec{r})$

Then, we differentiate, collect the terms and find:

Eq. (A.20) $$[\frac{p^2}{2m_0}+\frac{\hbar}{m_0}\vec{k}.\vec{p}+V(\vec{r})]u_{n\vec{k}}(\vec{r})=[E_{n\vec{k}}-\frac{\hbar^2}{2m_0}k^2]u_{n\vec{k}}(\vec{r})$$

This is now an equation for the unknown modulating part of the wavefunction $u_{n\vec{k}}(\vec{r})$. The known part has been incorporated and has given an energy shift (second term in the right-hand side of Eq. (A.20)) and a new term in the Hamiltonian (second term in the left-hand side of equation Eq. (A.20)). We can rewrite Eq. (A.20) as:

Eq. (A.21) $$[H_0+\frac{\hbar}{m_0}\vec{k}.\vec{p}]u_{n\vec{k}}(\vec{r})=[E_{n\vec{k}}-\frac{\hbar^2}{2m_0}k^2]u_{n\vec{k}}(\vec{r})$$

and taking the limit $\vec{k}=\vec{0}$ we have the eigenvalue equation:

Eq. (A.22) $$H_0u_{n0}(\vec{r})=[E_n(0)]u_{n0}(\vec{r})$$

for the $\vec{k}=\vec{0}$ envelope function. 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*.
- The number of basis eigenstates is equal to the number of energy bands in the semiconductors. In particular there is a valence band function and a conduction band function. The energy difference between each band is finite. We could use these functions, even though we do not know them, as basis functions and expand the $\vec{k}$ dependent term of the Hamiltonian (Eq. (A.20)) as a perturbation near $\vec{k}=\vec{0}$. In this way we derive the additional $\vec{k}$-dependence of the energy and the $\vec{k}$-dependence of the core wavefunction $u_{n\vec{k}}(\vec{r})$. 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 (see Chapter 11) to the $\vec{k}$-dependent term in Eq. (A.20), we have for the energy and wavefunction:

Eq. (A.23)

$$E_n(\vec{k}) = E_n(0) + \frac{\hbar^2 k^2}{2m_0} + \frac{\hbar}{m_0}\vec{k}.\vec{p}_{nn} + \frac{\hbar^2}{m_0^2}\sum_{n'\neq n}\frac{\left|\vec{k}.\vec{p}_{nn'}\right|^2}{E_n(0)-E_{n'}(0)}$$

Eq. (A.24) $$u_{n\vec{k}}(\vec{r}) = u_{n0}(\vec{r}) + \sum_{n'\neq n}[\frac{\hbar}{m_0}\frac{\vec{k}.\vec{p}_{n'n}}{E_n(0)-E_{n'}(0)}]u_{n'0}(\vec{r})$$

where $\vec{p}_{nn}' = \int u_{n0}(\vec{r})\vec{p}u_{n'0}(\vec{r})d^3\vec{r}$ and $u_{n0}(\vec{r})$ are the eigenfunctions at $\vec{k} = \vec{0}$.

The second term in the right-hand side of Eq. (A.24) comes from the energy shift term in Eq. (A.23), and the two others are deduced from the perturbation theory.

Remember we need to multiply this function $u_{n\vec{k}}(\vec{r})$ by $e^{i\vec{k}.\vec{r}}$ to get the real wavefunction. Now we see that progress has been made. We notice that $\vec{p}_{nn}\alpha\int u_{n0}\vec{\nabla}u_{n0}\alpha\left[u_{n0}^2\right]_{-\infty}^{\infty} = 0$ by symmetry. We deduce from Eq. (A.23) that the effective mass near $\vec{k} = \vec{0}$ is given by (see section 4.2.6):

Eq. (A.25) $$(\frac{1}{m*})_{ij} = \frac{1}{m_0}\delta_{ij} + \sum_{n'\neq n}\frac{p^i_{nn'}p^j_{n'n} + p^j_{nn'}p^i_{n'n}}{E_n(0)-E_{n'}(0)}$$

The inverse of the effective mass is a sum of the free electron mass and a term, which depends on the momentum matrix elements of the $\vec{k} = \vec{0}$ envelope. But it also depends on the energy difference between the bands. To simplify the problem, we now consider just two bands: the conduction and valence bands. 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 bandgaps 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 bandgap and effective mass. 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 bandgap 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 interplay between theory and experiment gives us useful and simple empirical rules and numbers for the above matrix element in Eq. (A.25). For example one finds that the Kane parameter $E_P = \frac{2m_0}{\hbar^2}P^2$ where $P = \frac{\hbar}{m_0}p_{cv}^z$, is roughly 20~25 eV for most semiconductors of interest. Note that we have used this parameter to calculate the optical absorption in Chapter 10. The Kane method of expanding around the $\vec{k} = \vec{0}$ envelope states can be extended to treat also the spin-orbit interaction. The spin orbit coupling is of the form:

Eq. (A.26) $$V_{so} = \frac{\hbar}{4m_o c^2}\vec{\sigma}.(\vec{\nabla}V(\vec{r})\times\vec{p})$$

where $\vec{\sigma}$ is the electron spin operator, and $V(\vec{r})$ 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, involve 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 is known and the second of which 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. The bandgap which also enters the formula however, changes quite a lot. If therefore for some reason, such as strain or confinement, the bandgap 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 bandgap 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, strain can vary locally and give rise to local effective mass. The reader is referred to the book by Chuang [1995] for a detailed treatment of the Kane model and its applications.

Further reading

Chuang, S.L., *Physics of Optoelectronic Devices*, John Wiley & Sons, New York, 1995.

A.9. 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 sub-section 4.2.6, the motion of the electron is described by:

Eq. (A.27) $$\hbar\frac{d\vec{k}}{dt}=q(\vec{E}+\vec{v}\times\vec{B})$$

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:

Eq. (A.28) $$\frac{\partial f}{\partial t}+\vec{v}\cdot\vec{\nabla}_{\vec{r}}f+\frac{d\vec{k}}{dt}\cdot\vec{\nabla}_{\vec{k}}f=\left(\frac{\partial f}{\partial t}\right)_{coll}$$

where, together with Eq. (A.27), 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:

Eq. (A.29) $$f(\vec{r}+d\vec{r},\vec{k}+d\vec{k},t+dt)=f(\vec{r},\vec{k},t)+\left(\frac{\partial f}{\partial t}\right)_{coll} dt$$

to first order:

Eq. (A.30)
$$f(\vec{r},\vec{k},t)+\frac{\partial f}{\partial t}dt+\vec{\nabla}_{\vec{r}}f\cdot d\vec{r}+\vec{\nabla}_{\vec{k}}f\cdot d\vec{k}=f(\vec{r},\vec{k},t)+\left(\frac{\partial f}{\partial t}\right)_{coll} dt$$

Eq. (A.31) $$\frac{\partial f}{\partial t}+\vec{\nabla}_{\vec{r}}f\cdot\frac{d\vec{r}}{dt}+\vec{\nabla}_{\vec{k}}f\cdot\frac{d\vec{k}}{dt}=\left(\frac{\partial f}{\partial t}\right)_{coll}$$

Using $\vec{F}=\hbar\frac{d\vec{k}}{dt}=q(\vec{E}+\vec{v}\times\vec{B})$, we get:

Eq. (A.32) $$\frac{\partial f}{\partial t}+\vec{\nabla}_{\vec{r}}f\cdot\vec{v}+\vec{\nabla}_{\vec{k}}f\cdot\frac{\vec{F}}{\hbar}=\left(\frac{\partial f}{\partial t}\right)_{coll}$$

This equation states that the changes of the distribution function with time (represented by the first term on the LHS of this equation) is determined by the flow of electrons in real space (the second term in the LHS of the equation), by the flow of electrons in $\vec{k}$ -space (the last term in the LHS of the equation) and the collisions (RHS of the equation). The RHS 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 $f(\vec{k},\vec{r},t)$ 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 1960's (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, we consider that the motion of the electron is well described by Eq. (A.27) 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.15). 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 $f(\vec{k},\vec{r},t)$.

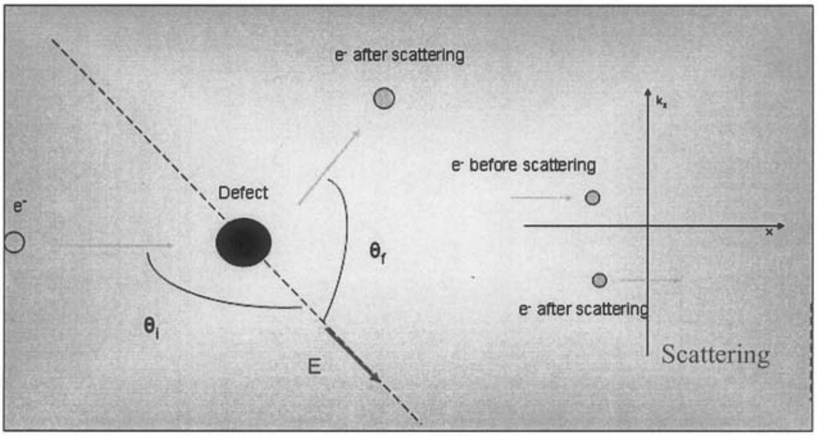


Fig. A.15. On the left, sketch of the scattering of an electron by an impurity. On the right, illustration of the disappearance and the appearance of the electron in the phase space.

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.27). Finally, we generate randomly the new direction of the wavevector.

- For the purpose of this simulation, we introduce a scattering rate $\Gamma = \sum_{i=1}^{n} \lambda_i(\vec{k}) + \lambda_0$ where we have introduced an artificial scattering mechanism, with a rate $\lambda_{0,}$ 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 $W_0(\vec{k},\vec{k}') = \lambda_\circ(\vec{k})\delta(\vec{k}-\vec{k}')$. 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 ($t_s = -\frac{1}{\Gamma}\ln(1-r)$, 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 equation Eq. (A.27) 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 $\vec{r}(t) - \vec{r}_0 = \int_0^t \vec{v}_g(t')dt' = \int_{t_0}^{t} \frac{1}{\hbar}\nabla_{\vec{k}} E dt'$, 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.

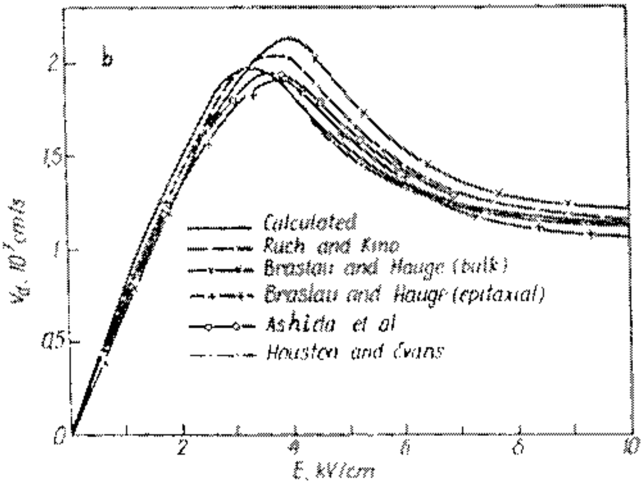


Fig. A.16. Measured and calculated drift velocity. [Reprinted from Solid State Electronics Vol. 23, Pozhela, J. and Reklaitis, A., "Electron transport properties in GaAs at high electric fields," p. 931, Copyright 1980, with permission from Elsevier.]

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.16. 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 one 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.17.

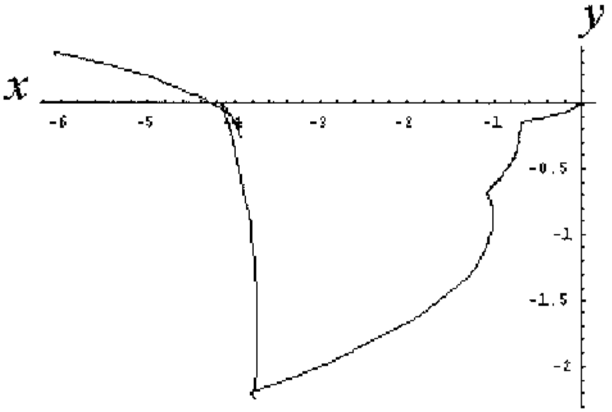


Fig. A.17. Simulation of the motion of an electron under an electric field E in the x-direction (10 collisions are simulated). The motion of the electron starts at the origin and evolves randomly. This figure represents the trajectory of the electron in real space.

References

Pozhela, J. and Reklaitis, A., "Electron transport properties in GaAs at high electric fields," *Solid States Electronics* 23, pp.927-933, 1980.

Further reading

Shur, M., *Physics of Semiconductor Devices*, Prentice-Hall, Englewoods Cliff, NJ, 1990.

A.10. The thermionic emission

The thermionic emission theory is a semi-classical 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.18.

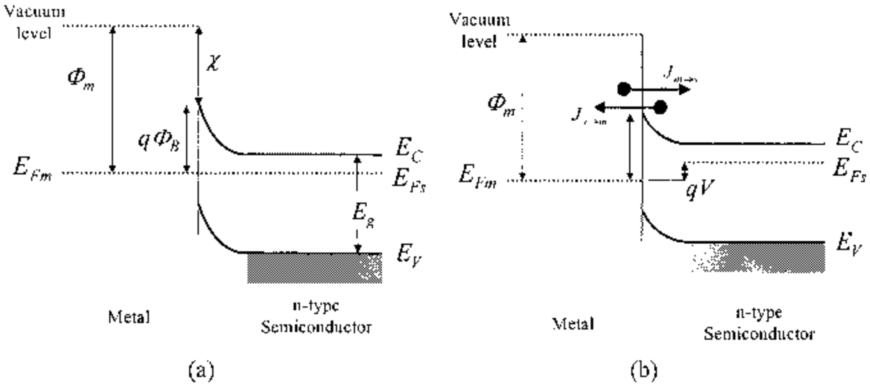


Fig. A.18. Energy band diagram of a Schottky metal-(n-type) semiconductor junction: (a) at equilibrium and (b) under forward bias (V>0), showing the transport of electrons over the potential barrier as the main transport process under forward bias.

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, (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:

Eq. (A.33) $$J_{s\to m} = \left(\frac{4\pi q m^* k_b^2}{h^3} \right) T^2 e^{-\frac{q\Phi_B}{k_b T}} e^{\frac{qV}{k_b T}}$$

or:

Eq. (A.34) $$J_{s\to m} = A^* T^2 e^{-\frac{q\Phi_B}{k_b T}} e^{\frac{qV}{k_b T}}$$

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 Plank's constant and m^* is the electron effective mass in the direction perpendicular to the junction plane, and $A^* = \frac{4\pi q m^* k_b^2}{h^3}$ is called the effective Richardson constant for thermionic emission. This quantity can be related to the Richardson constant for free electrons, A=120 A.cm^{-2}.K^{-2}, as discussed below.

For n-type semiconductors with an isotropic electron effective mass m^* in the minimum of the conduction band, we have $\frac{A^*}{A} = \frac{m^*}{m_0}$, 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 $\frac{A^*}{A} = \frac{\left(l_x^2 m_y^* m_z^* + l_y^2 m_z^* m_x^* + l_z^2 m_x^* m_y^* \right)^{\frac{1}{2}}}{m_0}$, 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 $\dfrac{A^*}{A} = \dfrac{\left(m_{lh}^* + m_{hh}^*\right)}{m_0}$, where m_{hh}^* and m_{lh}^* are the heavy-hole and light-hole effective masses, respectively. A few examples of values for $\dfrac{A^*}{A}$ are given in Table A.1.

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

Table A.1. Examples of values for $\dfrac{A^*}{A}$ *in a few semiconductors. [Sze 1981]*

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.18(b)), $J_{m\to s}$ 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.34), we obtain:

Eq. (A.35) $$J_{m\to s} = -A^* T^2 e^{-\frac{q\Phi_B}{k_b T}}$$

The total current density is therefore:

Eq. (A.36) $$\begin{aligned} J &= J_{s\to m} + J_{m\to s} = A^* T^2 e^{-\frac{q\Phi_B}{k_b T}} \left[e^{\frac{qV}{k_b T}} - 1 \right] \\ &= J_{ST} \left[e^{\frac{qV}{k_b T}} - 1 \right] \end{aligned}$$

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:

Eq. (A.37) $$J_{ST} = A^{*}T^{2}\, e^{-\frac{q\Phi_B}{k_b T}}$$

References

Bethe, H.A., "Theory of the boundary layer of crystal rectifiers," *MIT Radiation Laboratory Report* 43-12, 1942.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

A.11. 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 non-essential 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: 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.

Human health

Metalorganics cause severe burns. Do not get in eyes, on 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.

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 stomach. Do not give anything by 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 minutes. 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 minutes. 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 minutes 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 one 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, 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

Compound	Formula	Abbreviation	Melting point (°C)	Boiling point (°C)	Log₁₀ *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	-82.3	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.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)	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	$(C_2H_6)_4Ge$	TESn	-112	181		
Group V sources						
Diethylarsine Hydride	$(C_2H_6)_2AsH$	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	$(CH_3)_3P$	TMP	-85	37.8	7.7329 – 1512/T	
Triethylphosphorus	$(C_2H_6)_3P$	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	$(C_2H_6)_3Sb$	TESb	-98	116	17	75
Group VI sources						
Diethylselenide	$(C_2H_6)_2Se$	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.3. Physical properties of some organometallics used in MOCVD. [Ludowise 1985, and http://electronicmaterials.rohmhaas.com]

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 towards 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.4. Chemical properties of diethylzinc. [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 mmIIg
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.

Table A.5. Chemical properties of trimethylindium. [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 towards 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.6. Chemical properties of triethylindium. [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 towards 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.

Table A.7. Chemical properties of trimethylgallium. [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 towards 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.

Table A.8. Chemical properties of triethylgallium. [Razeghi 1989]

References

Ludowise, M., "Metalorganic chemical vapor deposition of III-V semiconductors," *Journal of Applied Physics* 58, R31-R55, 1985.

Mudry, W.L., Burleson, D.C., Malpass, D.B., and Watson, S.C., *Journal of Fire Flammability* 6, p. 478, 1975.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, 1989.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

Index