

Albrecht Winnacker

The Physics Behind Semiconductor Technology



Springer

The Physics Behind Semiconductor Technology

Albrecht Winnacker

The Physics Behind Semiconductor Technology

 Springer

Albrecht Winnacker
Centre for Advanced Materials
Heidelberg University
Heidelberg, Germany

Institute Materials for Electronics
and Energy Technology
University of Erlangen-Nuernberg
Erlangen, Germany

ISBN 978-3-031-10313-1 ISBN 978-3-031-10314-8 (eBook)
<https://doi.org/10.1007/978-3-031-10314-8>

© Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Preface

This book goes back to a two-semester lecture given at the University of Heidelberg for master's and advanced bachelor's students of physics, and it summarizes many aspects of a lecture given to students at the Engineering faculty of the University of Erlangen specializing in the fields of electrical engineering and electronic materials. This dual origin of the book reflects the fact that on the one hand semiconductor technology is certainly an outstanding application of solid state physics, and on the other hand it represents an important sector of materials science and engineering, dealing with processes and devices based on semiconductor materials. So, having a foothold in one of the basic fields of physics, and another one in the engineering fields of materials science and electronics, it operates at the interface between pure and applied science. The distinction made between these two manifestations of science, sometimes resulting in controversial assessments of their respective merits, often overlooks the close interrelation between the two. Basic facts and discoveries of pure science regularly have resulted in technical achievements of high relevance. On the other hand, new technical tools or experimental observations have provided the key to fundamental scientific insights, and technical challenges have initiated basic understanding. So the distinction between pure and applied science is not so much a matter of the objects and topics under consideration, but, as it seems, one of the meanings of "understanding". Pure science in general and pure physics in particular aim at realizing the complex interplay between cause and effect, while application and engineering rather aim at ensuring "function", taking care of reliability and yield of technical processes.

This textbook somehow has in mind to bridge that gap. Whenever dealing with basic technological processes and facts like the doping of semiconductors, lattice defects, or the field effect transistor, it includes considerations and statements on the underlying basic physical principles in the form of "physics comments". When dealing with fundamental physics like dispersion relations, carrier statistics, or the two-dimensional electron gas, it immediately includes their most practical consequences like the indirect or direct nature of semiconductors, the question of high-temperature electronics, or the definition of the unit Ohm.

In this way, it may help engineering students to get an insight into the basic, sometimes profound, character of the underlying physics, and to help the student of physics to get an impression of the sophistication of engineering and fabrication with their requirements on perfection and reliability.

The book is divided into two parts, the first one dealing essentially with materials, their structure, and their function. The second part deals with “devices”, making use of the structure and function of the materials and components.

So, in summary, this is not a book on the latest achievements of electronics (in spite of often referring to them); it is a book on “understanding”, in the sense given above, in the sense of physics, and in the sense of engineering. In this way, it may promote mutual respect between the two disciplines.

I am deeply grateful to my wife Eva for her continuous support and patience.

Heidelberg, Erlangen, Germany
September 2022

Albrecht Winnacker

Contents

Part I Semiconductor Materials: Structure, Processes, Fabrication

1	Crystal Structure and Energy Bands	3
1.1	The Lattice Structure	3
1.2	Energy Bands	6
1.3	Metals and Insulators	8
1.4	Electrons and Holes	9
1.5	The Origin of Energy Bands according to the Schroedinger Equation	10
1.6	Band Structure	12
1.7	Direct and Indirect Semiconductors	17
1.8	Basic Properties of a Semiconductor Material to Be Taken from the Band Structure	19
1.9	Questions and Answers	23
2	Transport of Charge Carriers	25
2.1	Ohm's Law	25
2.2	Mobility and Hall Effect	30
2.3	Factors Determining the Mobility of Charge Carriers	33
2.4	Saturation Velocity	35
2.5	Mobility and Frequency Limit of Devices	35
	References	37
3	Donors and Acceptors	39
3.1	Creation of Free Electrons and Holes by Thermal Excitation	39
3.2	Doping of Si	41
3.3	The Hydrogen Model of Donors and Acceptors	43

3.4	Shallow Donors and Acceptors in Compound Semiconductors	46
3.5	Isoelectronic Impurities	47
3.6	Questions and Answers	49
4	Carrier Statistics	51
4.1	Density of States	51
4.2	Free Carrier Concentration in Thermal Equilibrium	53
4.3	The Law of Mass Action	57
4.4	The Intrinsic Case	57
4.5	Semi-insulating Semiconductors	61
4.6	The Temperature Dependence of Carrier Concentration. Why Si-Electronics Does not Work at Elevated Temperatures	64
4.7	Questions and Answers	67
5	Fabrication of Electronic Silicon	69
5.1	Silicon Valley, Si-Technology, and Si-Age	69
5.2	The Siemens Process	70
5.3	Crystal Growth of Silicon	73
5.4	The Float Zone Process as a Purification Method	75
5.5	Wafering	75
5.6	Electronic Silicon, a Material of Extreme Purity and Crystalline Perfection	77
5.7	Questions and Answers	79
6	Lattice Defects	81
6.1	Intrinsic and Extrinsic Defects	81
6.2	Point Defects	82
6.3	Line Defects (Dislocations)	85
6.4	The Role of Dislocations in Electronic Materials	86
6.5	Two-Dimensional Defects	88
6.6	Three-Dimensional Defects	89
6.7	Questions and Answers	94
	References	94
7	Compound Semiconductors	95
7.1	Introduction, Some History	95
7.2	The Potential of III-V-Compounds	99
7.3	Growth of GaAs and InP by the LEC and VGF Process	102
7.4	“New” Compound Semiconductors: SiC, AlN, and GaN	104
7.5	Epitaxy	108
7.6	Questions and Answers	112
	References	113

8	Amorphous Semiconductors	115
8.1	Short-Range Order and Band Structure	115
8.2	On the Distinction Between Direct and Indirect Semiconductors in Amorphous Semiconductors	119
8.3	Variable Range Hopping Conduction	121
8.4	The Problem of Doping in a-Si	122
8.5	Applications of a-Si:H	124
8.6	Questions and Answers	125
	References	126
 Part II Devices		
9	The pn-Junction	129
9.1	pn-Junction Without External Voltage	129
9.2	pn-Junction with External Voltage	134
9.3	Breakthrough Voltage	139
9.4	Thickness of Depletion Layers	140
9.5	Questions and Answers	140
10	Solar Cells	143
10.1	The Sun, a Gigantic Energy Source	143
10.2	The Solar Cell, a pn-Junction Device	144
10.3	Operating Principle of the Solar Cell in Detail	148
10.4	The Optimum Bandgap for a Solar Cell Material	151
10.5	Thin Film Solar Cells	153
10.6	Cheaper Silicon for Photovoltaics	154
10.7	Questions and Answers	157
	References	158
11	Light Emitting Diodes (LEDs)	159
11.1	The Light Emitting Process in Semiconductors	159
11.2	The Linewidth of LEDs	161
11.3	Radiative and Nonradiative Processes	163
11.4	External and Internal Quantum Efficiency	165
11.5	LEDs for All Colors	166
11.6	The Story of the Blue LED	168
11.7	White LEDs	170
11.8	UV LEDs	172
11.9	Brightness of LEDs	173
11.10	Questions and Answers	177
	References	178

12	Semiconductor Lasers	179
12.1	The Laser Principle	179
12.2	Light Amplification and Lasing	185
12.3	Double Heterojunction Laser	189
12.4	Lasers for Optical Data Storage and Optical Data Transmission	192
12.5	Laser Materials for Optical Data Transmission, an Example of Bandgap Engineering	194
12.6	Questions and Answers	197
	References	198
13	Transistors	199
13.1	The Bipolar Transistor	200
13.2	Field Effect Transistor	204
13.3	A MOSFET Problem Related to Miniaturization of Modern Electronics: High k-dielectrics	211
13.4	The Two-Dimensional Electron Gas, Realized in a MOSFET	214
13.5	Questions and Answers	219
	Reference	220
14	Integrated Circuits	221
14.1	Moore's Law	221
14.2	Photolithography: Components and Steps of Fabrication	223
14.3	Exposure in the Photolithography	227
14.4	Some Important Constituents of Integrated Circuits	228
14.5	Light Sources for Photolithography	230
14.6	Resolution Enhancement Techniques	234
14.7	Some Concluding Remarks on the Present Situation of Chip Technology	235
14.8	Questions and Answers	236
	References	237
15	Organic Electronics	239
15.1	Organic Materials for Electronics, a Totally Different Approach	239
15.2	Conjugated Double Bonds	240
15.3	Band Structure and Conductivity in Polymers	241
15.4	Polymer Conductivity by "Doping"	243
15.5	Conductivity of Polymers by Charge Carrier Injection	246
15.6	Charge Carrier Transport	247
15.7	Organic Light-Emitting Diodes (OLEDs)	248
15.8	Organic Solar Cells	255
15.9	Questions and Answers	257
	References	259

16	Towards Molecular Electronics	261
16.1	From “Top-Down” to “Bottom-Up”	261
16.2	Carbon Allotropes, Fullerenes	262
16.3	Carbon Nanotubes	263
16.4	Graphene	267
16.5	Single Molecule Interconnects and Anchor Groups	271
16.6	Molecular Switches	272
16.7	C-Based Electronics, Complementary to Si-Technology	273
16.8	Questions and Answers	273
	References	274
	List of Physics Comments	277
	Name Index	279
	Subject Index	281

About the Author



Albrecht Winnacker was born in Frankfurt/Main, Germany. He studied Physics in Freiburg, Göttingen, Paris, and Heidelberg. After finishing his PhD work at the University of Heidelberg under the supervision of Prof. Gisbert zu Putlitz, he joined the University of Berkeley as a postdoctoral fellow. In 1972, he returned to Heidelberg, where he finished his “Habilitation” in 1976 and was appointed Professor of Experimental Physics in 1980.

From 1984 to 1985, he worked as “Visiting Scientist” at the IBM Research Almaden Laboratories; in 1986, he joined the Corporate Research Laboratories of the Siemens Corporation in Erlangen/Germany, where he was in charge of the division on compound semiconductors and phosphors. At the end of 1991, he returned to academia by taking over the Chair of Electronic Materials at the University of Erlangen. His research work focused on the development of SiC and AlN as semiconductor materials in particular, and compound semiconductors in general. He was cofounder of the start-up company SiCrystal, which is now a leading company on the world market for the production of SiC wafers.

After his retirement from Erlangen University, he was appointed Senior Professor and founding director of the Centre for Advanced Materials CAM at the University of Heidelberg, an innovative center dedicated to materials of organic electronics. He is still serving this institute as an Honorary Director.

Albrecht Winnacker published more than 350 papers in refereed journals and published a book on “The physics of lasers and masers”. He is an Honorary Professor of Physics at the University of Heidelberg, a member of the “German Academy of Engineering Sciences acatech”, and a member of the Heidelberg Academy of Sciences.

Part I
**Semiconductor Materials: Structure,
Processes, Fabrication**

Chapter 1

Crystal Structure and Energy Bands



Abstract Crystalline solids can be described in terms of a lattice consisting of a large number of identical unit cells, characterized by the lattice points and the lattice vectors. The group of atoms assigned to each lattice point forms the “basis”. The energy levels of the electrons are grouped in “energy bands”, each of those bands consisting of a large number of discrete energy levels given by the number of unit cells in the crystal. Basic properties of solids can be concluded from this elementary picture, like the existence of two types of charge carriers, i.e. negatively charged electrons and positively charged “holes”; the difference between metals, semiconductors, and insulators; the existence of a “valence band” and a “conduction band”; and others. For a more detailed discussion of the electronic properties of solids, the Schroedinger equation has to be considered. The characteristic feature of the Schroedinger equation in a crystalline solid is the periodicity of the potential. This results in the grouping of the energy levels of the electrons in energy bands, as well as the relation between the energy and momentum of these electrons. The relation between energy and momentum, the “dispersion relation”, defines the “band structure” from which important properties like bandgap, effective mass, and direct or indirect nature of the semiconductor can be concluded.

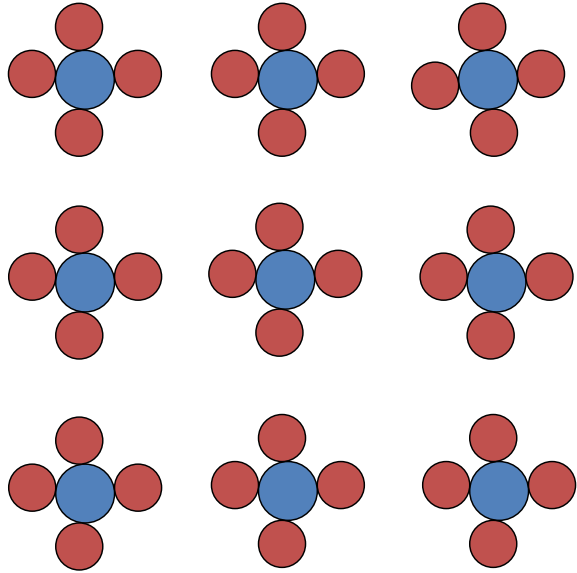
Many properties of semiconductors which we are dealing with in this book are just due to the basic fact that semiconductors are solid state systems, usually in crystalline, occasionally in the amorphous state. Some of the basic facts about crystals will be discussed in this chapter.

1.1 The Lattice Structure

The crystal lattice is defined by a periodic pattern of lattice points, the “point lattice”. This three-dimensional periodic pattern of lattice points can be described by

$$\mathbf{R}_{nlm} = n\mathbf{a} + l\mathbf{b} + m\mathbf{c} \quad (n, l, m \text{ integers})$$

Fig. 1.1 Point lattice and basis. The blue atoms mark the lattice points, and the basis is formed by the group of one blue atom and four surrounding red atoms



where $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are the so-called lattice vectors.

Each lattice point is associated with an atom or a group of atoms, called the “basis”.

In this sense, we can state that the crystal is described by a point lattice and a basis:

$$\text{Crystal} = \text{point lattice} + \text{basis}$$

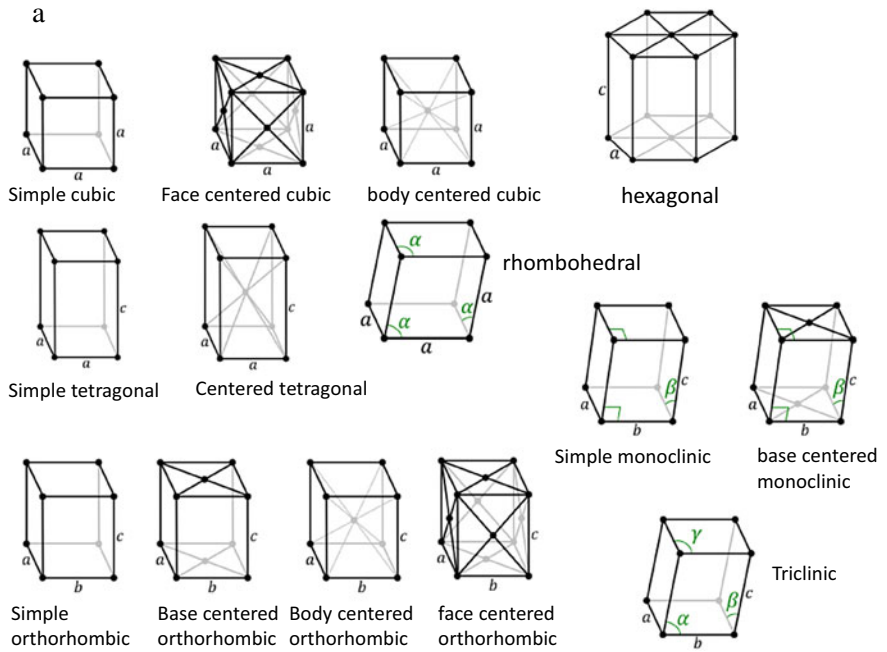
In the crystal shown very schematically in Fig. 1.1, the basis would consist of the atom drawn in blue surrounded by the four atoms drawn in red. The lattice vectors $\mathbf{a}$ and $\mathbf{b}$ of this two-dimensional lattice are also included.

The crystals found in nature can be classified into seven “crystal systems” according to their “point symmetry elements”, which are essentially crystal rotation axes and mirror planes. Just to give an example, the hexagonal lattice obviously has a six-fold rotation axis; after each rotation step of 60° , the lattice looks identical, and it is “transformed into itself”. Mirror planes are easily identified from the figures as well.¹

By subdividing the crystal systems into groups of the same point symmetry elements but different unit cells (again we have to refer to the general textbooks), one obtains the 14 Bravais lattices of Fig. 1.2a.

In this way, e.g. one can distinguish between the three cubic lattices “simple cubic”, “face-centered cubic”, and “body-centered cubic” as seen in the figure.

¹ For more details, we refer to the general textbooks on solid state physics, e.g. Siegfried Hunklinger, Festkörperphysik, 5th edition 2018, De Gruyter; Berlin, Boston, Siegfried Hunklinger, Christian Enss, Solid State Physics (De Gruyter Textbook) 2022.



b

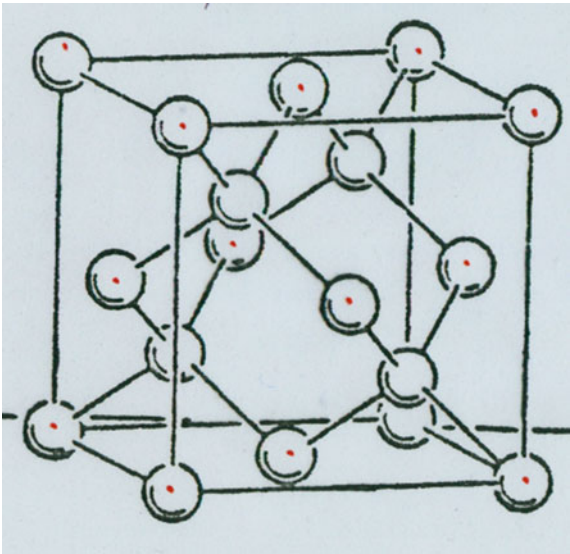


Fig. 1.2 **a** The seven “crystal systems”: Triclinic, monoclinic, orthorhombic, tetragonal, cubic, trigonal, and hexagonal, and the 14 Bravais lattices. **b** Unit cell of Si

One may ask why the large number of different crystals found in nature are catalogued according to symmetry properties. Why not according to hardness, color, size of the unit cells, etc.? The answer is that many interesting physical properties of solids are directly a consequence of their symmetry, like piezoelectricity, pyroelectricity, birefringence, capability of frequency doubling, etc.

Example Figure 1.2b shows, as an example, the unit cell of Silicon.

The lattice is face-centered cubic, as the figure shows. In each corner and in each center of a face is an atom (marked red), and the basis consists of two atoms, one in the corner or in the face center, and the other one shifted by a quarter of a diagonal of the cube. As a result, one obtains the famous Silicon- or Diamond structure: Each atom is surrounded by four atoms in the corners of a tetrahedron.

1.2 Energy Bands

The crucial point in understanding the energy level scheme of electrons in a crystal concerns the role of periodicity.

The basic unit of a crystal is the “unit cell”. One or more atoms can be assigned to that unit cell such that the total crystal is an ensemble of N unit cells. The energy level scheme corresponding to just one unit cell is shown schematically in Fig. 1.3a.

Let us now (Fig. 1.3b) add a second unit cell. What will the energy level scheme of these two unit cells look like? It will certainly look different, because the electrons of unit cell 1 will “feel” the electric fields of the electrons of unit cell 2. Perturbation

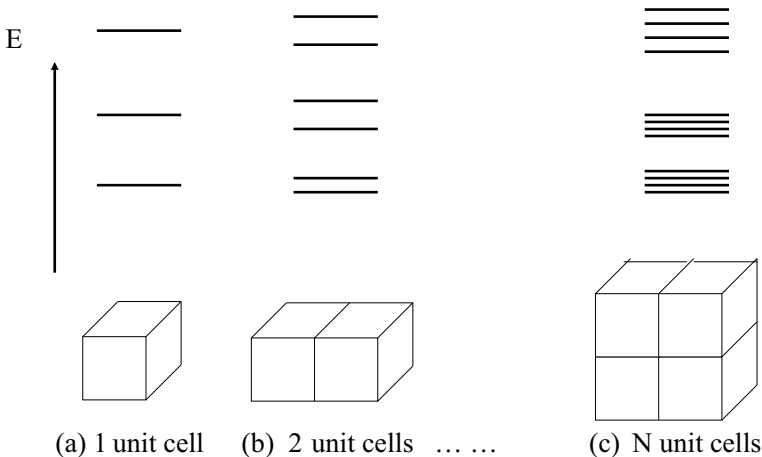


Fig. 1.3 **a** One unit cell with corresponding energy levels. **b** Two unit cells with corresponding energy levels. **c** N unit cells with corresponding energy levels

theory tells that the levels related to unit cell 1 will be split due to the presence of unit cell 2, as seen in Fig. 1.3b.

If, now, we consider the level scheme of N unit cells, each level will be split into N sublevels as shown in Fig. 1.3c.

Physics Comment: The Splitting of Energy Levels in a Periodic Lattice

A handwaving argument may illustrate the fact that each energy level of a single unit cell will be split into two close-lying levels by the presence of a second cell (Fig. 1.3b) and N close-lying levels by N neighboring cells. When a second unit cell is added to the first one, it is clear that the total number of energy levels of that system now is doubled, because the interaction between the cells certainly does not create nor annihilate levels. If the interaction between the two unit cells is only weak (a realistic approximation contained in our argument), levels are only weakly shifted by that interaction, which means that the energy level scheme is essentially doubled into pairs by this interaction, as in Fig. 1.3b. The same argument applies to the case of N unit cells.

From this very schematic consideration, it is seen how the energy bands of a crystal develop out of the level scheme of a unit cell forming the basic building block of the solid. No matter how qualitative in character this consideration is, a number of important conclusions can be drawn from it as follows:

1. Each energy band consists of N sublevels if N is the number of unit cells in the crystal.
2. According to Pauli's principle, each level can only be occupied by two electrons (spin up and spin down).
3. This amounts to stating that there is something like a "full" band. One energy band can only host $2N$ electrons.
4. A full band does not contribute to the current flowing through the crystal. This is due to the following reason: For any electronic state in a band whose electron moves through the crystal with velocity v , there must exist a state with the same energy with velocity $-v$ for symmetry reasons. So the contributions of these two states to a current compensate each other and add up to zero

This latter fact (point 4) evidently facilitates enormously the discussion of electronic conductivity in solids. Let us consider the case that all the electrons of a solid have been accommodated on the available electronic levels, starting from the lowest level and ending at a level in some higher lying band. In a discussion of the conductive behavior of the material, we then can forget about all the full bands and only have to consider the highest, only partially filled band. For this reason, the lowest not completely filled band is called the "conduction band" CB, and the highest completely filled band is called the "valence band" VB. (At that point, we forget about thermal excitation and consider the situation at $T = 0$.)

1.3 Metals and Insulators

When all the electrons have been accommodated in the available electronic levels, one can distinguish between the two situations shown in Fig. 1.4.

- When all the electrons are accommodated, a band is partially filled (left side of Fig. 1.4), i.e. the valence band is full and the conduction band partially empty.
- One band just is completely filled, all the higher bands are empty (right side of Fig. 1.4), i.e. the valence band is full and the conduction band empty.

The first case (statement a) is characteristic of a metal. It means that there is conductivity even at $T = 0$, because there is a partially filled band.

The second case (statement b) is characteristic of semiconductors and insulators. It means that the conductivity is zero at $T = 0$, because all bands are either full or empty.

In the case of a semiconductor (right side in Fig. 1.4), an important change in the situation will occur at $T > 0$ due to thermal excitation; see Fig. 1.5.

The excitation process leaves holes in the VB and produces occupied states in the CB. In this way, an electronic conductivity shows up because now the conduction band is no longer empty, and, very important for the further discussion, the valence

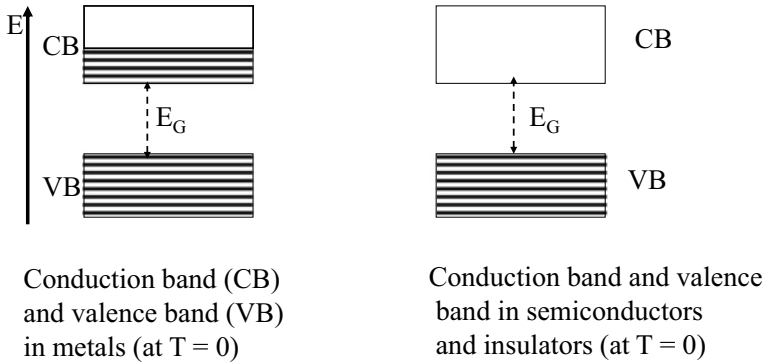


Fig. 1.4 Occupancy of conduction and valence band in different materials. The black lines in each band mark schematically the energy levels in the bands, CB: Conduction band; VB: Valence band. Left side: Metal. Right side: Insulator or semiconductor

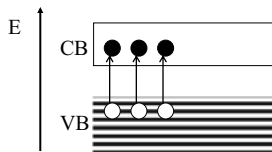


Fig. 1.5 Thermal excitation of charge carriers. The arrows symbolize the excitation process, leaving holes (empty circles) in the VB and electrons (full circles) in the CB

band is no longer full! So there are two contributions to the current: One from the CB which is no longer empty, and another one from the VB, which is no longer full. This latter contribution of the valence band has an important consequence on the electronic properties of semiconductors, as discussed in the next section.

1.4 Electrons and Holes

Of great importance for the understanding of the electronic properties of solids is the following fact: A missing electron in an otherwise fully occupied band contributes to the current in the same way as an electron with equal speed but opposite charge, i.e. a missing electron acts like a positively charged particle.

This fact will be formally proven in the following:

As stated above (conclusion 4), a full band does not contribute to the current. Because the contribution of a single electron (called electron number n) to a current is proportional to its drift velocity (the constant of proportionality is called K here), we have

$$i_n = KeV_n \quad (1.1a)$$

where V_n is the drift velocity of electron n . The current of the total band is

$$I_{\text{total}} = K \sum_n eV_n = 0$$

where the sum extends over all electrons of the full band.

Now we can write the sum over all electrons of the band as the sum over all electrons except electron number m plus the contribution of electron no. m . That means

$$I_{\text{total}} = K \sum_{\text{all } n \text{ except } m} KeV_n + eV_m = 0$$

or

$$K \sum_{\text{all } n \text{ except } m} eV_n = -KeV_m \quad (1.1b)$$

This is just what we wanted to prove: The current of a band with one missing electron (electron m) equals the current that the missing electron would provide, but with the opposite sign (compare 1.1a and 1.1b).

So one can conclude that the current in a semiconductor can have two contributions, one by negatively charged carriers (i.e. the electrons in the CB) and one by positively charged carriers (i.e. the missing electrons in the VB). Because they correspond to “missing electrons”, the positive carriers are usually called “holes”.

This fact is of paramount importance for the electronic properties of semiconductors. Materials with predominantly positive carriers (“holes”) are called p-type;

otherwise, when electrons in the CB dominate the electronic behavior, they are called n-type. Many properties of electronic devices are based on the existence of the two types of carriers, positive ones (holes) and negative ones (electrons).

For example,

the pn-rectifying diode,
 the pnp-and npn- transistor (bipolar transistor),
 the solar cell,
 the photodetector,
 the light-emitting diode (LED),
 and others.

1.5 The Origin of Energy Bands according to the Schroedinger Equation

The correct treatment of the system “electrons in a solid” requires, of course, the consideration of Schroedinger’s equation

$$-(\hbar^2/2m)\Delta\psi + V(\mathbf{r})\psi(\mathbf{r}) = E\psi(\mathbf{r}) \quad (1.2a)$$

Characteristic for a (crystalline) solid is the periodicity of the potential:

$$V(\mathbf{r}) = V(\mathbf{r} + \mathbf{R}_{nlm}) \quad (1.2b)$$

where $\mathbf{R}_{nlm} = n\mathbf{a} + l\mathbf{b} + m\mathbf{c}$ is any lattice vector.

Figure 1.6 illustrates this fact for the two-dimensional case.

The point given by vector $\mathbf{r}$ and the point given by $\mathbf{r}$ plus, e.g. the linear combination of lattice vectors $2\mathbf{a} + \mathbf{b}$ have the same potential, because they have an identical environment due to the periodicity of the lattice.

It was shown by Felix Bloch, one of the pioneers of quantum mechanics, that the solution of the Schroedinger Equation in a periodic potential (1.1a, 1.1b) can be written in the form

$$\psi(\mathbf{r}) = u_k(\mathbf{r}) e^{i\vec{k}\vec{r}} \quad (1.3a)$$

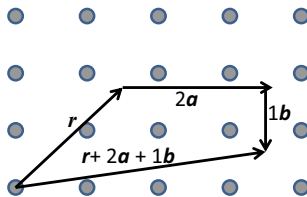


Fig. 1.6 Illustration of periodicity of potential. The full points outlined in blue are the lattice points, and $\mathbf{a}$ and $\mathbf{b}$ are the lattice vectors. $\mathbf{r}$ and $\mathbf{r} + 2\mathbf{a} + \mathbf{b}$ mark two points with identical potential

where $u_k(\mathbf{r})$ has the same periodicity as the lattice, i.e.

$$u_k(\mathbf{r}) = u_k(\mathbf{r} + \mathbf{R}_{nlm}) \quad (1.3b)$$

Figure 1.7 shows that this is a very plausible result, indeed, by presenting separately the two components forming the Bloch function.

Equation (1.3a) describes a plain wave $e^{i\mathbf{k}\mathbf{r}}$ with wavevector $\mathbf{k}$ (Fig. 1.7a) whose amplitude $u_k(\mathbf{r})$ (Fig. 1.7b) is modulated with the same periodicity as the lattice, as expressed in (1.2b). This latter fact, the amplitude having the same periodicity as the lattice, is a very intuitive one: The amplitude $u_k(\mathbf{r})$ must have the same value at equivalent points of the lattice. This results in the complete wavefunction Fig. 1.7c.

We now insert (1.3a) into the Schroedinger equation (1.2a). It is obvious that this transforms the Schroedinger equation from being a differential equation for $\psi(\mathbf{r})$ into a differential equation for $u_k(\mathbf{r})$:

$$-(\hbar^2/2m)\Delta(u_k(\mathbf{r})e^{i\mathbf{k}\mathbf{r}}) + V(\mathbf{r})u_k(\mathbf{r})e^{i\mathbf{k}\mathbf{r}} = Eu_k(\mathbf{r})e^{i\mathbf{k}\mathbf{r}} \quad (1.4a)$$

This differential equation for $u_k(\mathbf{r})$ will contain the wavevector $\mathbf{k}$ as a parameter. We see immediately for instance that $\Delta u_k(\mathbf{r}) e^{i\mathbf{k}\mathbf{r}}$ will result in a term $k^2 u_k(\mathbf{r}) e^{i\mathbf{k}\mathbf{r}}$ (among others). As a consequence the Eigenvalues E_n of the differential equation (1.4a) and consequently of the Schroedinger equation (1.2a) will depend on $\mathbf{k}$, that is they should be written as

$$E_n = E_n(\mathbf{k}) \quad (1.4b)$$

In other words, to each wavevector $\mathbf{k}$ (or momentum $\hbar\mathbf{k}$) belongs a series of Eigenvalues $E_n(\mathbf{k})$. Figure 1.8 shows how this results in energy bands.

In Fig. 1.8, use has been made of two facts concluded from Bloch's theory:

Fact 1. The functions $E_n(\mathbf{k})$ are symmetric to $\mathbf{k} = 0$, i.e. $E_n(\mathbf{k}) = E_n(-\mathbf{k})$.

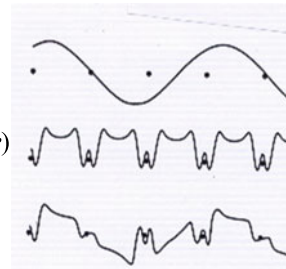
Fact 2. They are periodic as a function of $\mathbf{k}$. This means, e.g.: when we consider the function $E_n(\mathbf{k})$ in the x-direction and $\mathbf{a}$ is the lattice vector in this direction, the period is $2\pi/a$, as plotted in Fig. 1.8. Because of this periodicity, the function $E(\mathbf{k})$ has only to be shown in one periodicity region, which is called the “1st Brillouin zone”. This is indicated in the figure.

Fig. 1.7 Characteristic features of Bloch functions.
a Plain wave component of Bloch function. **b** Amplitude factor. **c** Total Bloch function

(a) plain wave $\text{Re}(e^{i\mathbf{k}\mathbf{r}})$

(b) amplitude factor $\text{Re}u_k(\mathbf{r})$

(c) Total wavefunction:
 $\psi(\mathbf{r}) = \text{Re}(u_k(\mathbf{r})e^{i\mathbf{k}\mathbf{r}})$



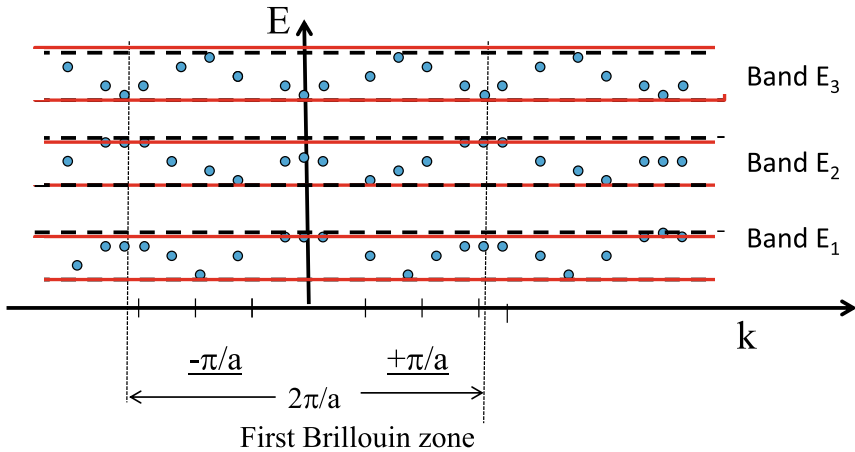


Fig. 1.8 Energy bands in solids: the Eigenvalues $E_n(\mathbf{k})$ are symmetric with regard to $\mathbf{k} = 0$ and periodic

The first statement can be understood easily: An electron with wavevector $+\mathbf{k}$ has a momentum $\hbar\mathbf{k}$, and an electron with wavevector $-\mathbf{k}$ has a momentum $-\hbar\mathbf{k}$, which means the electrons move with equal velocity in opposite directions. For symmetry reasons then they must have the same energy, because the two $\pm$ directions are, of course, equivalent.

The second fact is more sophisticated and will be discussed in the next “Physics comment”.

1.6 Band Structure

The total set of functions $E_n(\mathbf{k})$ is called the band structure of the corresponding semiconductor. Remember that $\hbar\mathbf{k}$ is the momentum $\mathbf{p}$ of the electron. So the band structure gives the relation between Energy E and momentum $\mathbf{p}$ of the electrons in the different bands. The relation between energy and momentum is generally called the “dispersion relation” in physics. So the band structure represents the dispersion relations of the different energy bands in the semiconductor material.

Physics Comment: Dispersion Relations of Phonons and Electrons in Comparison

“Dispersion relation” is a general notion for the relation $\omega(\mathbf{k})$ between frequency ω and wavenumber $\mathbf{k}$ of a wave. When—according to the rules of quantum mechanics—we describe a particle with Energy E and momentum $\mathbf{p}$ in the wave picture, the frequency is given by $\omega = E/\hbar$ and the momentum

Fig. 1.9 **a** The monoatomic chain. **b** Dispersion relation in one dimension. **c** Example of two different oscillations of the monoatomic chain with identical amplitude at the atomic positions

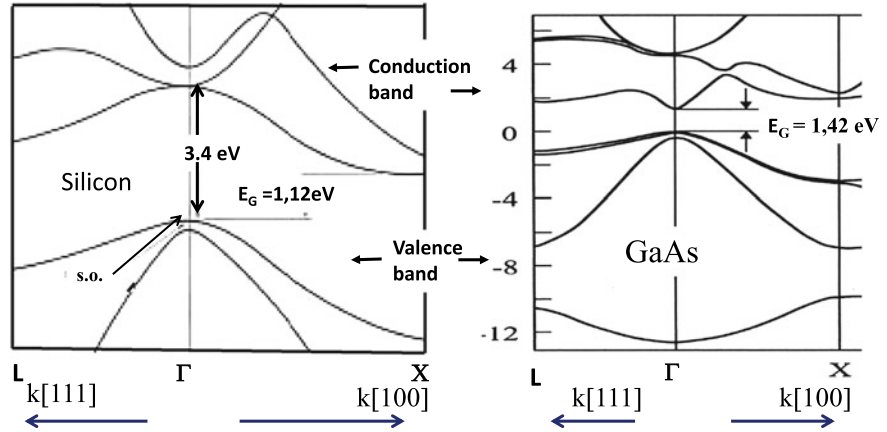
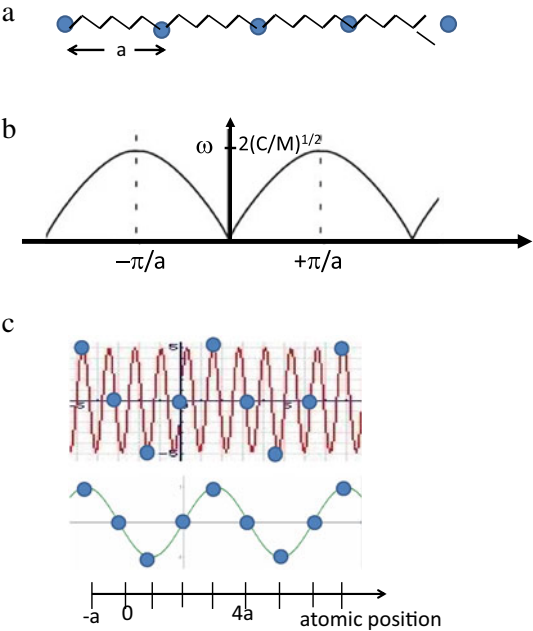


Fig. 1.10 Band structures of Si and GaAs

by $p = \hbar k$, so in this case, $E(p)$ or $E(k)$ is also called “dispersion relation”, as stated above.

In this chapter, the periodicity of the dispersion relation of the electrons plays an important role in the context of Bloch waves (Fig. 1.8 and corresponding

text). “Periodicity” of dispersion relations is a general phenomenon in solid state physics as a consequence of the periodicity of the lattice, and it applies, e.g. to phonons (lattice vibrations) and conduction electrons. In this “physics comment”, we shall discuss the fact, that in spite of this “superficial” general character, the reason for this periodicity is different for lattice vibrations and conduction electrons.

To illustrate this, we remind at the case of lattice vibrations and—for simplicity—consider the case of a monoatomic chain (Fig. 1.9a).

Let the atoms be coupled by a spring with a force constant C . In this case, it turns out that the dispersion relation is given by (Fig. 1.9b).

$$\omega = 2(C/M)^{1/2}|\sin(ka/2)| \quad (1.5a)$$

This dispersion has a periodicity with a period of $2\pi/a$. So the question is why do two vibrations that differ in their k -value by $2\pi/a$ have the same frequency ω ?

The answer is given in the following way:

The amplitude of a vibration with wavenumber k is given by

$$y = A \sin(kx - \omega t) \quad (1.5b)$$

Because the positions of the atoms of the monoatomic chain are given by $x = na$, the amplitude of atom n at time t is given by

$$y = A \sin(nka - \omega t) \quad (1.5c)$$

We compare this with a vibration with wavenumber $k + 2\pi/a$. In this case,

$$y = A \sin[(k + 2\pi/a)x - \omega t] \quad (1.5d)$$

Again we consider the amplitudes of the atoms at their positions na

$$y = A \sin[(k + 2\pi/a)na - \omega t] = A \sin(nka - \omega t + 2\pi n) \quad (1.5e)$$

Because

$$\sin(nka - \omega t) = \sin(nka - \omega t + 2\pi n)$$

(given the periodicity of the sinus function), we see that the vibrations (1.5b) and (1.5d) at any moment have the same amplitudes at the points $n = na$, where the atoms are situated. At the positions between the atoms, they don't have a physical meaning.

So the answer to our question “Why do two vibrations that differ in their k -value by $2\pi/a$ have the same frequency ω ”? is simply given by “Because they are the same vibration”, in the sense that in any point occupied by atoms they have the same amplitude. Figure 1.9c gives an illustration of two vibrations of different wavelengths but the same amplitudes at lattice points occupied by atoms.

(Figure 1.9c shows the atoms of a monoatomic chain, separated by a distance a . The green curve corresponds to a lattice vibration $A_1(x)$ with wavelength $\lambda_1 = 4a$, i.e. wavenumber $k_1 = 2\pi/4a$. Because of the periodicity with period $2\pi/a$, a “neighboring” oscillation $A_2(x)$ with $k_2 = k_1 + 2\pi/a = \frac{5}{4}(2\pi/a)$ should have the same amplitude at all points na where atoms are situated. The wavelength of $A_2(x)$ is $\lambda_2 = 2\pi/k_2 = 4a/5$, which is $1/5$ of λ_1 . $A_2(x)$ is plotted in red in Fig. 1.9c. Evidently, the amplitudes of $A_1(x)$ and $A_2(x)$ are identical at the positions $x = na$)

It is obvious that this argument cannot be applied to the situation of conduction electrons in a crystal because the amplitude of the wave is defined at any point in the crystal, not only at certain distinct positions. For simplicity, we consider the situation in one dimension with coordinate x and lattice constant a . The line of argument can easily then be applied to the three-dimensional situation:

We shall demonstrate the periodicity of $E(k)$ with period $2\pi/a$ (a is the lattice constant) by showing the following:

Whenever there is a Bloch function with wavenumber k and Eigenvalue $E(k)$, there will also exist a Bloch function with wavenumber $k + 2\pi/a$ having the same Eigenvalue $E(k)$. We will show this by explicitly giving this Bloch function.

The argument runs as follows:

Let

$$\psi_k(x) = u_k(x)e^{ikx} \quad (1.6)$$

be a Bloch function in the first Brillouin zone, with wavenumber k and Eigenvalue $E(k)$. Being such a Bloch function means per definition:

1. $\psi_k(x)$ is a solution of the Schroedinger equation with Eigenvalue $E(k)$.
2. $u_k(x) = u_k(x + a)$.

We will consider now another Bloch function with $k' = k + 2\pi/a$. This, by definition, would be a function of type

$$v(x)e^{i(k+2\pi/a)x} \quad (1.7)$$

where the amplitude function $v(x)$ would have the periodicity as the lattice, i.e. $v(x) = v(x + a)$. As announced as our line of proof, we will show now that such a Bloch function exists with the same Eigenvalue $E(k)$.

Indeed, the wavefunction

$$\psi_k + 2\pi/a(x) = u_k(X)e^{-i(2\pi/a)x}e^{i(k+2\pi/a)x} \quad (1.8)$$

is evidently such a function, indeed, because of the following reasons:

- (a) It is a solution of the Schrodinger equation with Eigenvalue $E(k)$, because it is the same function as $\psi_k(x)$, as is easily seen:

$$\text{Evidently, } \psi_k(x) = u_k(X)e^{-i(k+2\pi/a)x}e^{i(k+2\pi/a)x}.$$

which, by inserting (1.6), can be written as

$$\psi_k(x) = u_k(x)e^{ikx}e^{-i(k+2\pi/a)x}e^{i(k+2\pi/a)x} = u_k(x)e^{-i(2\pi/a)x}e^{i(k+2\pi/a)x}$$

the latter expression being identical to (1.8). So we have shown that the function (1.8) is identical to function (1.6), i.e. the two functions have the same Eigenvalue $E(k)$.

- (b) The amplitude factor $v(x) = u_k(x)e^{-i(2\pi/a)x}$ of the plane wave $e^{i(k+2\pi/a)x}$ evidently is periodic with periodicity a , because both factors $[u_k(x) \text{ and } e^{-i(2\pi/a)x}]$ have this property.

So (1.8) is a Bloch function with wavenumber $k + 2\pi/a$ and the same Eigenvalue $E(k)$ as $\psi_k(x)$. Q.E.D.

So the two cases, phonons in a crystal and Bloch waves in a crystal, both show periodic dispersion relations, but the cases are not identical. The argument for them has to run differently!

Figure 1.10 shows the band structures of Si and GaAs, two of the most important semiconductors.

Only the two bands which are crucial for the electronic properties are shown, the valence band and the conduction band. The figures make practical use of the facts 1 and 2 given above (in the context of Fig. 1.8) in the following way: Because the functions $E_n(k)$ are symmetric relative to $k = 0$, they are plotted only for positive k (in the 100-direction) and negative k (in the 111-direction). And because they are periodic functions of k , they are plotted only in one periodicity region. This “first” periodicity region, as already mentioned in the context of Fig. 1.8), is called the “first Brillouin zone”. So in Fig. 1.10, the first Brillouin zones are plotted in 100-direction for Si and GaAs, see right side of the plot, and in 111-direction, see left side of the plot. Γ , L, and X are established notations for certain points in k -space. Γ marks the center of the k -space ($k = 0$), X the edge of the first Brillouin zone in $\langle 100 \rangle$ -direction (at $k = 2\pi/a$ for the fcc lattice like Si and GaAs), and L the edge of the first Brillouin zone in $\langle 111 \rangle$ -direction.

The maximum of the valence band of Si is positioned at $k = 0$ (the Γ -point), and the minimum of the conduction band of Silicon is situated in $[100]$ -direction near the X-point (at $k_x = 0, 85 * 2\pi/a$), 1.12 eV above the maximum of the valence band. So the bandgap is 1.12 eV. For GaAs both, the minimum of the conduction band and the maximum of the valence band are at $k = 0$, and the bandgap is $E_G = 1.42$ eV.

1.7 Direct and Indirect Semiconductors

At this point, an important distinction should be introduced, the one between direct and indirect semiconductors. The basic idea is shown in Fig. 1.11.

When the minimum energy of the conduction band is situated right above the maximum energy of the valence band (i.e. at the same k -value), we speak of a “direct semiconductor” (see the left side of Fig. 1.11), otherwise of an indirect semiconductor. Because $\hbar k$ is the momentum of the electron, the former case implies that electrons at the lower edge of the conduction band and the upper edge of the valence band have the same momentum. As a consequence, when an electron jumps from the edge of the conduction band to a free state at the edge of the valence band, the momentum is conserved, and the process is “allowed”. The energy E_G released in this process can be used to create a photon. This is the light-emitting process in optoelectronic devices like light-emitting diodes (LEDs) and semiconductor lasers. In other words, **the optical emission by a band-band transition of an electron is “allowed” in a direct semiconductor**. The frequency ν of the emitted photon in this case is equal to

$$\nu = E_G/h \quad (1.9)$$

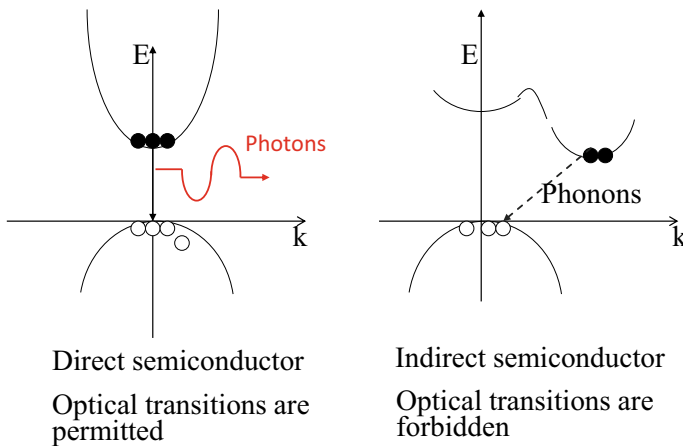


Fig. 1.11 Schematic band structure of direct (left) and indirect (right) semiconductor

(Equation 1.9 only holds approximately; for details, we refer to Sects. 11.1 and 11.2).

In an indirect semiconductor, the k -values of an electron at the lower edge of the conduction band and the one at the upper edge of the valence band are different (right side of Fig. 1.11). In this case, the band-band recombination via photon emission violates the law of momentum conservation. **The emission of a photon by a band-band transition in an indirect semiconductor is forbidden.** So indirect semiconductors cannot be used for LEDs and semiconductor lasers (A list summarizing the direct or indirect nature of important semiconductors is given in Table 1.1).

From Fig. 1.10, it is obvious that Si is an indirect semiconductor and consequently cannot be applied for LEDs and semiconductor lasers. While Si is the dominant material in most of the modern electronic applications, it is largely excluded from optoelectronics. This opens a wide area for semiconductor materials other than Si. As a matter of fact, the optoelectronic market is one of the most rapidly expanding ones in modern electronics, opening a wide door of applications for compound semiconductors!

What holds for emission processes also holds for the corresponding absorption process. So the direct excitation of an electron from the valence band to the conduction band by a photon is forbidden for indirect semiconductors and allowed for direct semiconductors. This process is the basic process in solar cells. So, for the reason given above, the absorption coefficient for a solar photon is orders of magnitude lower in indirect semiconductor materials as compared to direct semiconductors. In order to absorb the solar light quantitatively, a thickness of the solar cell of about 100 μm is required in a Si solar cell, and for GaAs, about 10 nm is sufficient. A Si solar cell is a “thick film solar cell”, and a GaAs solar cell a thin film solar cell”. This issue will be treated in Sect. 10.5.

Physics Comment: The Role of the Momentum of the Photon in the Recombination Process

The argument that in the case of an indirect semiconductor the band-band transition under photon emission is forbidden is based on the fact that the momenta of the electrons are different in the initial and final state of the process. We did,

Table 1.1 Direct and indirect semiconductors

Semiconductor	E_{Gap}/eV	Direct/indirect	Semiconductor	E_{Gap}/eV	Direct/indirect
Ge	0.664	i	3C-SiC	2.42	i
Si	1.12	i	6H-SiC	3.02	i
InP	1.34	d	4H-SiC	3.33	i
GaAs	1.42	d	GaN	3.44	d
AlAs	2.15	i	Diamond	5.50	i
GaP	2.27	i	AlN	6.28	d

however, not take into account that the photon emitted in the process also carries a momentum given by $p_{\text{photon}} = h\nu/c$. Could it be that the photon corrects the “missing momentum” and thus closes the gap between the momenta of the electron in the initial and the final state of the emission process?

Let us give a short estimate of the size of the momenta involved, e.g. for the case of Si. Looking at Fig. 1.10 (left side), we see that the minimum of the conduction band is close to the X-point at $0, 85 \cdot 2\pi/a = 5, 3/a$. The difference between the wavenumbers of the electrons in conduction and valence band hence is close to $5, 3/a$, where $a = 0.54 \cdot 10^{-9}$ is the lattice constant of Si. The corresponding difference in the momenta then is

$$\Delta p_e = \hbar \cdot \Delta k = h/2a \cong 0.6 \cdot 10^{-24} \text{ kgm/s} \quad (1.10)$$

We compare this with the momentum of the photon emitted in the process, which (with $E_G = 1.12 \text{ eV}$) is

$$p_{\text{photon}} = h\nu/c \cong E_{\text{Gap}}/c = 0.6 \cdot 10^{-27} \text{ kgm/s} \quad (1.11)$$

By comparing (1.10) and (1.11), it is obvious that the momentum of the emitted photon is a factor of 1000 smaller than the “gap” between the momenta of electrons and thus cannot come up for this difference. As a matter of fact, the numbers show that the **photon momentum can be neglected in the discussion of “allowed” or “forbidden” in connection with the electron recombination process.**

1.8 Basic Properties of a Semiconductor Material to Be Taken from the Band Structure

Three important properties can be extracted from the band structure as follows:

1. The bandgap E_G .
A look at Fig. 1.10a, b shows, e.g. that the bandgap of GaAs is 1.42 eV, And the one for Si is 1.12 eV.
2. The direct or indirect nature of the bandgap. Again Fig. 1.10 may be used as an example: GaAs is a direct semiconductor because the minimum of the conduction band and the maximum of the valence band have the same k-value ($k = 0$). Si, however, is seen to be an indirect semiconductor.
3. The so-called effective mass of the electrons and holes.

This point has not been discussed so far. It turns out that the electrons, called “free electrons” in the sense that they are delocalized and move freely in the crystal, nevertheless behave as having a mass different from the mass of the free electron. This is not surprising because at any given moment, when moving through the crystal,

they are exposed to the internal electric fields of the lattice. The value of the effective mass (called here m_{eff}) can be obtained as follows:

The definition of the (inert) mass according to Newton's law is given by

$$b = (1/m_{\text{eff}})F \quad (1.12)$$

where b is the acceleration, m_{eff} the effective mass, and F the acting force.

In the wave picture, i.e. describing the momentum via a wavevector k , we write

$$b = dv/dt = \underset{(B)}{(dv/dk)} \underset{(C)}{(dk/dt)} \quad (1.13)$$

(B) denotes the first factor on the right side of the equation, and (C) the second one.

To obtain (B), we argue the following:

v is the velocity of the moving particle. The velocity corresponds to the group velocity of the wave given by

$v = d\omega/dk$, which, because of $E = \hbar\omega$, equals.

$v = (1/\hbar)dE/dk$, and thus $dv/dk = (1/\hbar)d^2E/dk^2$, which is our result for (B).

Furthermore, according to Newton's laws, $dp/dt = F$, and because of $p = \hbar k$ in the wave picture,

$dk/dt = (1/\hbar)F$, which is our expression for (C).

By inserting our expressions for (B) and (C) into (1.13), one obtains

$$b = (1/\hbar^2)d^2E/dk^2F \quad (1.14)$$

Comparison of (1.14) and (1.12) shows

$$m_{\text{eff}} = \hbar^2/d^2E/dk^2 \quad (1.15)$$

d^2E/dk^2 is the curvature of the $E(k)$ curve. **So the effective mass of the electron is inversely proportional to the curvature of the $E(k)$ curve** and thus can be directly obtained from the band structure.

By looking at the band structures of Si and GaAs in Fig. 1.10, one sees that there are two kinds of holes at the upper end of the valence band, which differ by the curvature of their $E(k)$, and the one with slightly higher energies having the smaller curvature. So these holes are the "heavy holes", and the other ones the "light holes" of Si and GaAs (see numbers in Table 1.2). (Please note that some of the notions given in Table 1.2 like m_l , m_t and m_{cond} will only be explained in the following.)

It is worthwhile to note that the result (1.15), derived from the basic notions of wave mechanics like $p = \hbar k$, $E = \hbar\omega$, and group velocity $v = d\omega/dk$, is in perfect agreement with our intuitive understanding that an electron in the conduction behaves just like a free particle, with a mass m_{eff} different from the mass of a free electron. Within this "particle picture", we would write

Table 1.2 Effective (conductivity masses) of electrons and holes in Si and Ge and effective mass of electrons in GaAs in units of the free electron mass m_0 . Data taken, e.g. from <http://apachepersonal.miun.se/~gorthu/halvledare/Effectivemas sinsemiconductors.htm> m_{cond} has been calculated from (1.19)

	Electron	Hole
Si	$m_l = 0.89^*$ $m_t = 0.19$ $m_{\text{econd}} = 0.26$	Heavy hole $m_{\text{hh}} = 0.49$ Light hole $m_{\text{lh}} = 0.16$
GaAs	$m_e = 0.067$	Heavy hole $m_{\text{hh}} = 0.45$ Light hole $m_{\text{lh}} = 0.082$
Ge	$m_l = 1.64$ $m_t = 0.082$ $m_{\text{econd}} = 0.12$	Heavy hole $m_{\text{hh}} = 0.28$ Light hole $m_{\text{lh}} = 0.044$

* This value has been taken from HD Barber, [https://doi.org/10.1016/0038-1101\(67\)90122-0](https://doi.org/10.1016/0038-1101(67)90122-0)

$$E(k) = E_G + p^2/2m_{\text{eff}} \quad (1.16)$$

Again, by replacing p by $p = \hbar k$, we obtain

$$E(k) = E_G + p^2/2m_{\text{eff}} = E_G + (\hbar k)^2/2m_{\text{eff}} \quad (1.17)$$

which immediately leads to (1.15).

It follows from Eq. 1.15 that m_{eff} has a constant value as long as d^2E/dk^2 is constant, i.e. as long as $E(k)$ is parabolic. Figure 1.10 shows that that is the case near the edge of the energy band, where electrons and holes usually are positioned as long as the doping concentration is not extremely large. This parabolic behavior is not so much a physical as a mathematical phenomenon because the first member in the Taylor expansion around a minimum or maximum E_m (where first derivatives vanish) is parabolic:

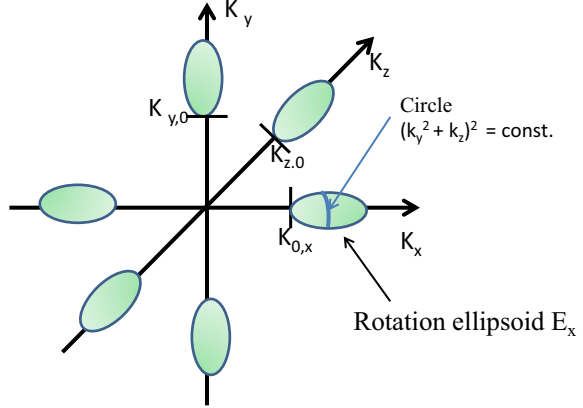
$$E(k) = E_m + a_2(k - k_m)^2 + \dots$$

Further away from the band edge, however, m_{eff} is no longer constant.

In one respect, our discussion of the effective mass has been strongly simplified. For simplicity, and to make the train of arguments more transparent, we have considered $E(k)$ to be an isotropic function, while, in reality, $E(\mathbf{k})$ is a function of the wave vector $\mathbf{k}$. There are, however, important cases where $E(\mathbf{k})$ is isotropic indeed and thus can be written as $E(k)$. Such a case is met, e.g. for the electrons in the conduction band of GaAs. Here, the minimum of $E(\mathbf{k})$ is at $\mathbf{k} = 0$ (Fig. 1.10). For this reason, and because, due to the cubic symmetry, the three axes k_x , k_y , and k_z are equivalent, $E(\mathbf{k})$ can be written as (1.17)

$$E(k) = E_G + (\hbar k)^2/2m_{\text{eff}}$$

Fig. 1.12 Constant energy ellipsoids for the six minima of the electron energy in the conduction band of Si



Hence, from Fig. 1.10, a clear and well-defined value for m_{eff} of the electrons in the conduction band can be given, which, by the way, is $m_{\text{eff}} = 0.067 m_0$ (see Table 1.2).

The situation, however, is different for Si. As Fig. 1.10 shows, there is a minimum of the conduction band for some k -value $k_{0,x}$ along the $[100]$ -axis, 1.12 eV above the conduction band, which defines the bandgap of Si (for illustration see also Fig. 1.12). We call this axis the x -axis. Because of the cubic symmetry of the lattice, there exists an analogue minimum along the y -axis, the z -axis, the $(-x)$ -axis, the $(-y)$ -axis, and the $(-z)$ -axis, altogether six minima. Let us consider the electrons in the minimum around the x -axis. Their energy as a function of $k = [k_x, k_y, k_z]$ then would be given by

$$E(k_x, k_y, k_z) = E_G + (\hbar^2/2m_l)(k_x - k_{0,x})^2 + (\hbar^2/2m_t)(k_y - k_{0,y})^2 + (\hbar^2/2m_t)(k_z - k_{0,z})^2 \quad (1.18a)$$

This expression is visualized in Fig. 1.12 in the way that the rotation-ellipsoid E_x shown there is the surface of constant energy as obtained from (1.18a).

It is plausible and has been taken into account in (1.18a) that in the expression for the dependence of the energy of the electrons around the x -minimum, the term $(k_x - k_{0,x})^2$ in (1.18a) has a different prefactor $\hbar^2/2m_l$ than the one of $(k_y - k_{0,y})^2$ and $(k_z - k_{0,z})^2$ given by $(\hbar^2/2m_t)$. We call m_l the longitudinal effective mass of the electron, and m_t the transverse mass of the electron.

Correspondingly, for the electrons in the minimum around $k_{0,y}$, one obtains

$$E(k_x, k_y, k_z) = E_G + (\hbar^2/2m_l)(k_y - k_{0,y})^2 + (\hbar^2/2m_t)(k_x - k_{0,x})^2 + (\hbar^2/2m_t)(k_z - k_{0,z})^2 \quad (1.18b)$$

and so on for the other four minima.

Table 1.3 Density of state masses in Si and Ge Data taken from <http://apacheper sonal.miun.se/~gorthu/halvle dare/Effectivemassinsemico nductors.htm> $m_{e,dos}$ has been calculated from (1.20)

	Electron	Hole
Si	$m_{e,dos} = 1.08$	$m_{h,dos} = 0.81$
GaAs	$m_{e,dos} = 0.067$	$m_{h,dos} = 0.47$
Ge	$m_{e,dos} = 0.56$	$m_{h,dos} = 0.29$

The surfaces of constant energy described by (1.18a, 1.18b ...) are the ellipsoids of revolution in Fig. 1.12.

So in Si, different from GaAs, the electrons in the conduction band are characterized by two effective masses, the longitudinal and the transvers mass. In describing the movement of the electron under the influence of an electric field, some average value of m_l and m_t has to be applied for the electron mass, the so-called “conductivity mass m_{cond} ”, which can be shown to be given by

$$m_{cond} = 3[1/m_l + 2/m_t]^{-1} \quad (1.19)$$

Let us mention for completeness that in the context of calculations where the density of states plays a role, the averaging over m_t and m_l has to be done in a different way, resulting in different values for m_{eff} , called “density of state mass” m_{edos} :

$$m_{edos} = 6^{2/3}(m_t m_l m_l)^{1/3} \quad (1.20)$$

Factor 6 refers to the fact that in Si there are six equivalent energy minima; see Fig. 1.12.

Values for “density of state masses” in Si and GaAs are given in Table 1.3.

The “density of state masses” comes into play, e.g. when calculating the “effective density of states” N_C and N_V (4.3b and 4.5b).

So different “effective masses” exist which have to be clearly distinguished from one another. In systems of lower symmetry, the “effective masses” have to be given by the “effective mass tensor m_{eij} ”, the components of which depend on the crystallographic directions involved.

1.9 Questions and Answers

Questions

- Question 1: What is the point lattice of Silicon? What is the basis?
 Question 2: What is the conduction band, and what is the valence band of a material?
 Question 3: Which property of $V(\mathbf{r})$ is characteristic of a crystalline solid?
 Question 4: Name two characteristic properties of the dispersion relation $E(\mathbf{k})$.
 Question 5: What is a direct, and what is an indirect semiconductor? Give examples!

Question 6: What is the “band structure” of a semiconductor material?

Question 7: Which basic properties of a semiconductor can be deduced from the band structure?

Answers

Answer 1: The point lattice is face-centered cubic. The basis consists of two atoms, one atom in a corner of the cube or the center of a face, plus a second atom shifted by a quarter of a space diagonal.

Answer 2: The conduction band is the lowest unoccupied or partly occupied band of a material, and the valence band is the highest fully occupied band at $T = 0$.

Answer 3: The periodicity of $V(\mathbf{r})$: $V(\mathbf{r}) = V(\mathbf{r} + \mathbf{R}_{nlm})$.

Answer 4: Symmetry with regard to $\mathbf{k} = 0$, periodicity in $\mathbf{k}$ -space.

Answer 5: In a direct semiconductor, the radiative band-band transition is allowed, and in an indirect semiconductor, it is forbidden for reasons of momentum conservation. Indirect: Si, Ge, GaP, and SiC. Direct: AlAs, GaAs, InP, and GaN.

Answer 6: The “band structure” is the whole of dispersion relations $E_n(\mathbf{k})$.

Answer 7: The bandgap E_{Gap} , the direct or indirect nature of the semiconductor, and the effective mass.

Chapter 2

Transport of Charge Carriers



Abstract Ohm's law describes the relation between voltage and current in metals and semiconductors. In this context, the notions of resistance, conductance, resistivity, conductivity, and—very important—mobility are introduced. Ohm's law is the result of a dissipative process, where the charge carriers are continuously accelerated by an electric field and slowed down by scattering processes. In the stationary state, they assume an average “drift velocity” the size of which, at a given E-field, is determined by the mobility of the carriers in the material. As a consequence of this picture, the conditions of validity of Ohm's law turn out to be that a) such dissipative processes exist, and b) the space charge remains zero. This is different from the case of space charge limited currents as they exist in, e.g. a vacuum tube or in most organic semiconductors. Mobility is an important property of semiconductors, relevant for the performance of many devices. It can be determined via the “Hall effect”. The Hall effect also reveals whether the semiconductor is n-type or p-type. Various factors determine the size of the mobility, like temperature, lattice perfection, and purity. It can be stated as a general rule that everything which perturbs the strict periodicity of the crystal lattice will reduce the mobility. One characteristic feature of devices determined by the mobility is their frequency limit. The mobility, e.g., determines the frequency limit of an important component of integrated circuits, the field effect transistor. However, when the length of the device gets comparable to or smaller than the scattering length, a completely different behavior sets in, described as “ballistic transport”.

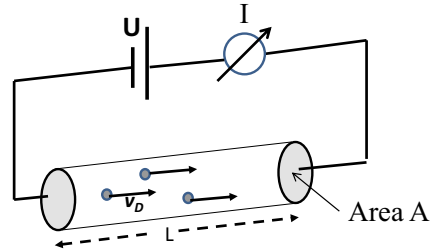
2.1 Ohm's Law

Ohm's law is general textbook knowledge. Nevertheless, it is not trivial, and it requires some analysis to define its range of validity.

Georg Simon Ohm stated this law in a treatise published in 1827. He investigated the current–voltage dependence of metallic wires of different materials and lengths and found

$$U = IR \quad (2.1)$$

Fig. 2.1 Electrons as carriers of a current in a wire



where U is the voltage and I the current. The constant of proportionality R is called the “resistance” of the wire.

Because the unit of U is Volt and of I is Ampere, it follows from (2.1) that the unit of R is Volt/Ampere which is called $1\ \Omega$ in honor of Georg Simon Ohm¹, the unit being symbolized by the Greek large letter Omega

$$1\Omega = 1\text{V/A} \quad (2.2)$$

Some comments on the modern definition of the unit Ohm will be given in Sect. 13.4.

A comparison of wires with different lengths L and cross section A (see Fig. 2.1) shows that

$$R = \rho L/A \quad (2.3)$$

That means R is a product of a “geometrical” factor L/A and a factor \ characteristic for the material, called resistivity. The inverse of ρ is called the conductivity σ of the material

$$\sigma = 1/\rho \quad (2.4)$$

By introducing the current density $i = I/A$, and remembering that the electric field equals $E = U/L$, (2.1) transforms into

$$U/L = (1/\sigma)/A)iA$$

or

$$i = \sigma E \quad (2.5)$$

(2.5) is just a different way of writing Ohm’s law (2.1) Table 2.1

¹ Georg Simon, German Physicist, born Erlangen, a small, but renowned university town in southern Germany in the year 1789 as son of a craftsman. He died in Munich 1854, Professor of Experimental Physics at University of Munich

Table. 2.1 Electrical conductivity of some materials All values at 20 °C. The values of Ge and Si refer to carrier concentrations of about 10^{14} cm^{-3} . According to 2.13 (derived below), the values roughly vary linearly with the carrier concentration

Material	Ag	Cu	Al	Fe	NiCr	Ge
$\sigma \text{ (S/m)}$	6×10^7	$5, 8 \times 10^7$	$3, 5 \times 10^7$	$1, 5 \times 10^7$	10^6	3

Material	Si	Glass	Mica	Teflon	Quartz
$\sigma \text{ (S/m)}$	1	10^{-10} - 10^{-14}	10^{-11} - 10^{-15}	$< 10^{-13}$	$1, 3 \times 10^{-18}$

Obviously, the electrical conductivity varies over nearly 26 orders of magnitude between very good and very poor conductors. No other material property shows such a wide variation!

In order to understand the physical process behind Ohm's law, we consider an ensemble of electrons moving in a wire under the influence of an electric field with an average velocity v_D ; see Fig. 2.1).

The current is defined as the charge moving through a cross section per second. From this definition, it is easily seen that in the case shown in Fig. 2.1, the current I equals $neAv_D$, where n is the concentration of charge carriers, each carrying charge e . So the current density is

$$i = nev_D \quad (2.6)$$

Now we calculate the average velocity v_D under the assumption that the flow of the electrons consists of a sequence of "stop and go". That means the electrons are accelerated by the electric field E , then stopped by some collision with an obstacle, then reaccelerated, then stopped again, etc.

Let τ_j be the time between one stopping event and the next one of a particular electron j . The acceleration in field E is given by $b = (e/m) E$, so after time τ_j the electron has achieved a velocity

$$v_j = b\tau_j = (e/m)E\tau_j \quad (2.7)$$

By averaging over all collision times (the collisions are assumed to be a statistical process), one obtains the average velocity, i.e. the drift velocity v_D :

$$v_D = (e/m)\tau E \quad (2.8)$$

where τ is the average time between two collisions of the electron, the "mean collision time". By inserting (2.8) into (2.6), it follows

$$i = (ne^2/m)\tau E \quad (2.9)$$

or, by comparison with (2.5), the conductivity σ turns out to be

$$\sigma = (ne^2/m)\tau \quad (2.10)$$

Equation (2.9) says that the current density i is proportional to E , thus it corresponds to Ohm's law in the form (2.5). So it turns out that Ohm's law can be understood on the basis of a “stop and go” mechanism, a process, where the electrons are accelerated by the electric field and then, in a stochastic process, stopped by some collision process. It turns out that the dominating collision process in metals is the one with some other free electron due to the high free electron concentration, while in semiconductors collisions with phonons are dominant at least in the room temperature region and above.

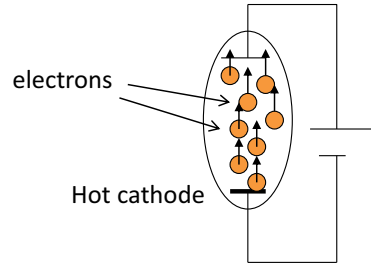
Physics Comment: Current Flow as a Dissipative Process

According to (2.8), as a consequence of the “stop and go” process, the drift velocity v_D is proportional to the field E . It is worthwhile to think for a moment about this fact. Keep in mind that the force on an electron in an electric field E is eE . So the velocity is proportional to the force on the electrons. But isn't, according to Newton, the acceleration proportional to the force? What does it mean that in this system not the acceleration, but the velocity is proportional to the force?

Let us put it that way: When a particle is accelerated, its kinetic energy increases. The fact that the particle is no longer accelerated but achieves a constant velocity v_D means that its kinetic energy no longer increases but gets constant. This evidently is the case because, via the collisions, the energy introduced by the force is lost to the environment, thus that in the steady state, the increase of kinetic energy by acceleration is just compensated by the loss of kinetic energy to the environment. The energy transferred to the environment, i.e. to the solid, ends up as thermal energy. In the case of conduction in semiconductors, this transformation into heat occurs by production of and collisions with phonons. These phonons then thermalize, and the material heats up. This is the reason for Joule's heat which develops when current flows. When energy is continuously lost to the environment ending up in a final form, in this case, thermal energy, we speak of a “dissipative process”. **In that sense, we may say that Ohm's law is the expression of a dissipative process underlying the current flow.**

Physics Comment: On the Validity of Ohm's Law

Ohm's law is so “elementary” that it is easily overlooked that it is by no means valid in all materials and systems. A familiar example where that is not the case is provided by the old vacuum tube (Fig. 2.2).

Fig. 2.2 The vacuum tube

The process of charge transport here is the following: At low voltages, the electrons emitted from the heated filament form a cloud which, due to its negative charge, repels further electrons evaporating from the filament, resulting in a small current only. When the voltage is increased, more and more electrons are pulled out of the cloud toward the anode, the cloud is thinned out, and the current increases at an accelerated rate. This results in the $I \sim U^{3/2}$ characteristic typical for the so-called “space charge limited current”, evidently differing from Ohm's law.

Thus, the second condition for the validity of Ohm's Law (apart from the existence of a dissipative process) is maintaining a space charge zero or, in other words, charge neutrality in the volume crossed by the current.

Let us note at this point that violation of space charge neutrality not only happens in the vacuum tube, it also regularly occurs in organic electronics. In a copper wire, as an example, the negative charge of the free carriers, the electrons, is at any time compensated by the positive charges of the copper ion cores, so charge neutrality is permanently conserved. However, the organic materials normally do not have a noticeable free carrier concentration, and the charge carriers are injected via the electrodes just like in the vacuum tube, so the organic materials also usually show a “space charge limited current”. For details, we refer to Sect. 15.5.

Let us now come back to the basic equations of transport underlying Ohm's law. We may write (2.8) as

$$v_D = (e/m)\tau E = \mu E \quad (2.11)$$

with

$$\mu = (e/m)\tau \quad (2.12)$$

The constant of proportionality μ between v_D and E is called the “mobility” of the charge carriers. The mobility of its charge carriers (electrons and holes) is a very important characteristic feature of a semiconductor, as many examples will teach us.

Inserting (2.12) into (2.10) gives

$$\sigma = ne\mu \quad (2.13)$$

which expresses the obvious fact that the conductivity σ depends on two quantities: On the concentration n of charge carriers and on the velocity which they assume under the action of an electric field (given by μ).

Because in semiconductors two kinds of carriers, electrons and holes, exist, the conductivity depends on the contributions of both, i.e.

$$\sigma = ne\mu_e + pe\mu_h \quad (2.14)$$

where n and p are the concentrations of electrons and holes, respectively, and μ_e and μ_h the mobilities. However, in most practical cases, one charge carrier type is dominant. In this case, we speak of an n-type or a p-type semiconductor, depending on the dominant carrier type.

2.2 Mobility and Hall Effect

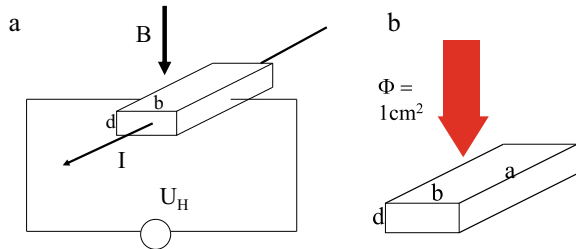
The experimental setup of the Hall effect is shown in Fig. 2.3a.

A current I flows through the sample, the drift velocity of the carriers being v_D . The magnetic field B is vertical to the current. Then, in the setup of Fig. 2.3a, the Lorentz force on the carriers is

$$K_L = qv_D B \quad (2.15)$$

if we denote the charge of the carriers by q . Under this force, the free carriers (electrons or holes) are deflected in a direction perpendicular to their velocity, i.e. to

Fig. 2.3 a: Experimental setup for observing the Hall effect. b: Illustration of the notion of “areal or surface density”



the current, and the B-field. As a consequence, an electric field builds up perpendicular to the current because the free carriers are spatially separated from the immobile ion cores which carry the opposite charge. This “Hall field” E_H produces a force on the carriers of size qE_H which is directed opposite to the Lorentz force. In the stationary situation, the two forces compensate each other, so

$$qE_H = qv_D B \quad (2.16)$$

Because $U_H = E_H b$ and $v_D = i/nq$ (2.6) $= (I/bd)/nq$, one gets for the “Hall voltage”

$$U_H = (1/nq)(IB)/d = R_H(IB)/d \quad (2.17)$$

with the “Hall constant” $R_H = 1/nq$.

A short comment may be given regarding the question, why, among the geometrical parameters of the sample, just the thickness d shows up in (2.17). When n is the number of charge carriers per volume in a sample, the product $nd = n_s$ corresponds to the areal or “surface density” of the sample (for illustration, see Fig. 2.3b).

We consider a numerical example: When the beam with crosssection $F = 1 \text{ cm}^2$ crosses the sample with, e.g. $n = 10^{22} \text{ cm}^{-3}$ atoms and $d = 0, 1 \text{ cm}$, it passes $n_s = nd$ atoms, so it sees an “areal density” of $n_s = 10^{21} \text{ cm}^{-2}$ atoms.

So (2.17) can be written as

$$U_H = R_H * IB \quad \text{with } R_H = 1/n_s q \quad (2.18)$$

This equation tells us that the Hall voltage depends on n_s , which is the areal density of carriers in the plain vertical to B . Some short consideration, which we shall leave to the reader, will show that indeed U_H can only depend on this areal density n_s , not the spatial density n , because “everything happens in the same plain”, the deflection of the electrons as well as the build-up of the Hall field.

So the Hall constant, measured in the setup of Fig. 2.3a, gives two information:

- (a) From the sign of R_H (i.e. from the sign of U_H), we obtain the sign of q . When R_H is negative, the dominant charge carriers are electrons with charge $-e$, and if it is positive, holes are dominant in the sample with charge $+e$.
- (b) From the absolute value of R_H , one obtains n , the carrier concentration. (In the case of holes, we would denote the carrier concentration rather by the letter p instead of n . So, with this notation, we obtain either n or p .)

If the carrier concentration is known, and if we measure in the same setup as shown in Fig. 2.3a the conductivity σ , then from (2.13), one gets the mobility

$$\mu_e = \sigma / ne \text{ or } \mu_h = \sigma / pe \quad (2.19)$$

for electrons or holes, respectively.

It is often stated that “the mobility is measured through the Hall effect”. This is not quite correct. To be precise, through the Hall effect one measures the carrier

Table 2.2 Carrier mobilities and bandgaps at $T = 300\text{ K}$

Material	μ_e (m^2/Vs)	μ_h (m^2/Vs)	E_{gap} (eV)
Si	0,15	0,0475	1,12
GaAs	0,85	0,04	1,42
InP	0,46	0,015	1,35
InAs	3,3	0,046	0,36
InSb	8,80	0,125	0,17

concentration, and when the carrier concentration is known, μ can be obtained from (2.19) after a measurement of σ .

The crucial step in understanding the Hall effect is included in the statement that in the setup of Fig. 2.3a, the Lorentz force is compensated by the force due to the Hall field; see (2.16).

Physics Comment: The Independence of the Conductivity of the Magnetic Field

This compensation of the two forces explains an interesting and, in a way, amazing fact concerning currents in solids: In the normal case, **the resistance of a metal or semiconductor is independent of the magnetic field**. At first sight, one might think that the resistance gets larger in a magnetic field, because as a consequence of the Lorentz force, the charge carriers have deviated from their straight path, and, due to a longer path, the resistance increases. This, however, is not true! Because the Lorentz force is just counteracted by the “Hall force” (2.16), the path of the carriers is unchanged in a magnetic field, and consequently, the resistivity remains the same.

There are various special situations that result in deviations from this general rule. In this case, the material is said to show a “magnetoresistance”. One, for instance, is that more than one type of carrier with different mobilities contribute to the current. Then (2.16) cannot be fulfilled for both carriers simultaneously.

Table 2.2 gives values for the mobilities in some semiconductors, together with the bandgaps.

At this point, only the following comments should be given with regard to the data in Table 2.2:

InAs and InSb are not used for microelectronics because of their small bandgaps.

Among the “useful” semiconductors GaAs has by far the largest mobility of carriers. $\mu_e(\text{GaAs})$ is 5–6 times as large as $\mu_e(\text{Si})$.

Generally, electron mobilities are much larger than hole mobilities.

2.3 Factors Determining the Mobility of Charge Carriers

It must be pointed out that the mobilities given in Table 2.2 refer to specific conditions. As given in the table caption, they refer to room temperature, but there are factors other than temperature determining mobilities.

A general statement should precede the more detailed discussion: **Essentially everything which perturbs the periodicity of the crystal will also reduce the mobility of the charge carriers.** A few examples of practical importance shall illustrate this as follows:

- (a) **Doping dependence of mobility:** As will be discussed in detail in chapter 3, semiconductors are usually “doped” by certain impurity atoms in order to control the charge carrier concentration. Evidently, any impurity atom (e.g. a phosphor atom on a Si-site in Si) disturbs the periodicity of the lattice and thus, according to our general statement above, will reduce the mobility. This is shown for n-doped GaAs in Fig. 2.4

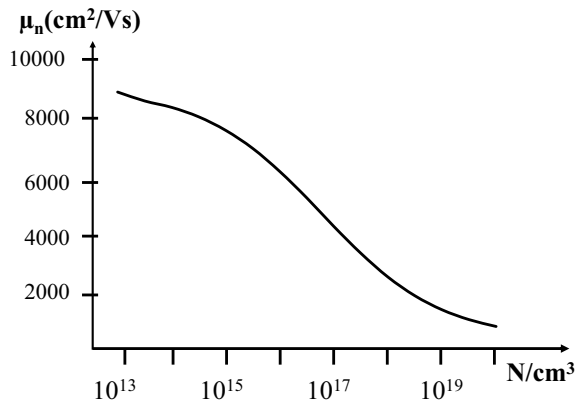
Essentially the same curve shape is seen in p-type GaAs. At dopant concentrations above roughly 10^{17}cm^{-3} , the mobility is significantly reduced. The underlying physical process is the scattering of the free carriers at the ion cores of the dopants.

As well as dopants any other kind of impurity or lattice defect (like dislocations) perturbing the periodic structure of the lattice may reduce the mobility.

- (b) **Temperature dependence of mobility:** As discussed above, μ is determined by scattering processes characterized by a scattering time τ (see 2.12). In the context of the temperature dependence, two contributions have to be taken into account: Scattering by lattice defects and scattering by phonons.

It is plausible that scattering processes by lattice defects are temperature dependent, given the fact that the velocity of the charge carriers is larger at higher temperatures (we are not talking here about the drift velocity v_D , but about the total velocity

Fig. 2.4 Mobility of electrons in n-doped GaAs versus dopant concentration



including the thermal contribution), so their trajectory is less affected in the scattering process, and the mobility will increase with temperature. As a matter of fact, the scattering process will result in a temperature dependence of μ given by $\mu \sim T^{+3/2}$.

But the temperature will also in another way affect the mobility.

Because of the thermal vibrations of the atoms, the atoms of the lattice at any given moment are no longer occupying the exact lattice points of the crystal, but are vibrating around those “ideal” mean positions. In this way, the strong periodicity of the lattice is perturbed, and the mobility will be reduced due to the heat movement of the atoms. In the phonon picture, the underlying physics process in this case is scattering by the lattice phonons (the quantized lattice vibrations). The theoretical treatment reveals that for this process μ is proportional to $T^{-3/2}$.

The results on the T-dependence of the mobility are summarized in Fig. 2.5 in a double logarithmic scale.

At low temperatures, impurity and lattice-defect scattering dominate (i.e. $\mu \sim T^{+3/2}$), and at higher temperatures, the phonon scattering ($\mu \sim T^{-3/2}$) is dominant. The difference between the two curves (low defect concentration as compared to high defect concentration) is explained as follows: The effect of impurities and lattice defects on the mobility is an “extrinsic” property of the crystal: So the reduction of the mobility is smaller for a crystal of high perfection and purity and larger for a crystal of low perfection and purity as shown in Fig. 2.5. The phonon contribution, however, is an intrinsic property characteristic for the crystal “as such” as shown in the figure by the fact that at high temperatures the curves for crystals with different lattice perfection join in one line.

Fig. 2.5 Mobility as a function of temperature

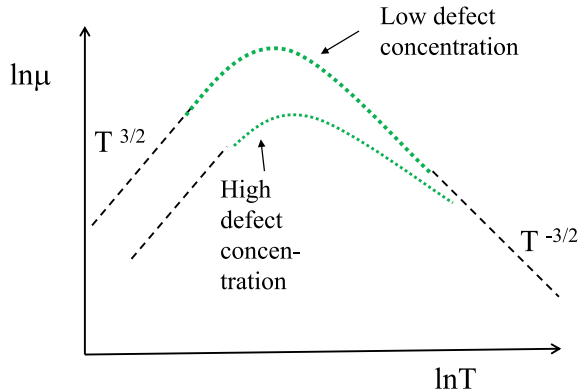
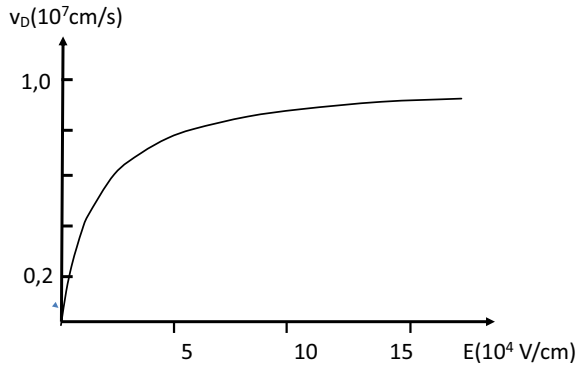


Fig. 2.6 Electron velocity in Si as a function of electric field strength E



2.4 Saturation Velocity

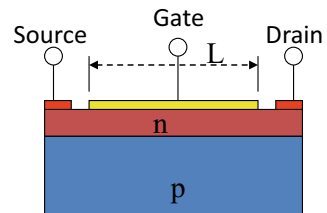
The notion of mobility goes back to the fact that, as expressed in (2.11), the drift velocity v_D is proportional to the electric field E , the constant of proportionality being called “mobility”. At a sufficiently high electric field, however, a wider spectrum of higher and higher frequencies (i.e. of higher and higher energies) of phonons can be excited by scattering processes of the charge carriers, and the slope of the $\mu(E)$ curve gets smaller until finally the drift velocity assumes a constant value (Fig. 2.6), the so-called saturation velocity.

As the figure shows the transition region toward the saturation region is roughly between 1–5 MV/m, and the saturation value of the drift velocity is around 10^5 m/s. This value for the saturation velocity, by the way, is similar for other important semiconductors like GaAs or SiC.

2.5 Mobility and Frequency Limit of Devices

The mobility is a very important property of a semiconductor. It determines the speed (or frequency limit) of a device. This will be explained for the important case of a field effect transistor (Fig. 2.7). (The following discussion is anticipating the detailed presentation of the field effect transistor (FET) in Sect. 13.4.)

Fig. 2.7 Field effect transistor



In the field effect transistor, electrons are injected from the source electrode and run through the n-type “channel” to the drain electrode. The charge carrier density in the channel is controlled by the “gate” voltage V_g , so the current from source to drain I_{sd} at a given source-drain voltage V_{sd} can be controlled by the gate voltage V_g . It follows from this working scheme that, in order to change the current I_{sd} , the carrier density in the channel has to be changed. It is obvious that the time to change this carrier concentration is essentially given by the time τ_{cross} required to cross the channel from source to drain because the channel has to be cleaned from the former carrier concentration in order to achieve the new carrier concentration. So the switching time T_s is roughly given by

$$T_s = \tau_{\text{cross}} \quad (2.20)$$

resulting in a frequency limit of $f_{\text{limit}} = 1/\tau_{\text{cross}}$. Because $\tau_{\text{cross}} = L/v_D$, this would result in $f_{\text{limit}} = v_D/L$.

A precise technical definition of the notion of “frequency limit” (i.e. the frequency for which the amplification of the transistor equals 1) results in the modification

$$f_{\text{limit}} = v_D/2\pi L \quad (2.21)$$

or, because of $v_D = \mu E$ (see 2.11) and $E = U_{sd}/L$, one finally obtains

$$f_{\text{limit}} = \mu U_{sd}/(2\pi L^2) \quad (2.22)$$

Numbers

With $L = 1 \mu\text{m}$ (a typical structure size) and $\mu = 0,1 \text{ m}^2/\text{Vs}$, and $U_{sd} = 1 \text{ V}$ (2.22) results in $f_{\text{limit}} = 16 \text{ GHz}$. $\mu = 0,1 \text{ m}^2/\text{Vs}$ would be a realistic value for doped Si.

With a value of $\mu = 0,6 \text{ m}^2/\text{Vs}$ (realistic for doped GaAs), 100GHz could be achieved.

So high mobility results in a high-frequency limit of the device. Because of its higher mobility, this gives GaAs an advantage over Si for high-frequency applications, e.g. in modern mobile phones.

Ballistic transport

The drift velocity as given by (2.11) is the result of a “stop and go” process, i.e. of the interplay between the acceleration of the electron by the electric field and the scattering processes in the material which slow down the electron. The mean spatial distance between two such scattering processes is called the “scattering length”. Let us now again consider the field effect transistor in Fig. 2.7. An interesting new situation arises when in the course of the miniaturization of modern devices the separation L between source and drain gets smaller than the scattering length. In this case, the carriers just “fall freely” without any scattering event from one electrode to the other, and the device enters the regime of “ballistic transport”. In this regime, velocities much larger than the ones resulting from $v_D = \mu E$ can be achieved which,

as explained above, favors shorter switching times and a higher limiting frequency. In the case of GaAs, ballistic transport sets in at about $L = 0,1 \mu\text{m}$, in Si at roughly $0,01 \mu\text{m}$. So GaAs in present technology reaches already the limit of ballistic transport, not to the same extent, however, at present Silicon. In summary, we can state that at larger dimensions, the high mobility of carriers favors GaAs for high-frequency devices (see Table 2.2), for very small dimensions, the advantage may get even larger because GaAs may enter the regime of ballistic transport, while Si does not yet.)

Questions

- Question 1: How is the “mobility” of charge carriers defined?
- Question 2: How are mobilities measured?
- Question 3: Give an example of a semiconductor with high mobility?
- Question 4: Which factors determine the mobility?
- Question 5: For which kind of devices is a high mobility particularly important?
- Question 6: Why does Ohm’s law usually not hold in organic semiconductors?
- Question 7: What is “ballistic transport”?

Answers

- Answer 1: The mobility μ is the proportionality factor between drift velocity and electric field.
- Answer 2: Via the Hall effect.
- Answer 3: GaAs.
- Answer 4: Doping concentration, lattice defects, and temperature.
- Answer 5: For high-frequency devices.
- Answer 6: Because in organic semiconductors space charges are built up by the injected charge carriers. So currents are “space charge limited”.
- Answer 7: Ballistic transport sets in when the scattering length of the charge carrier gets comparable or larger as compared to the channel length of the FET, i.e. the gate length.

References

1. Georg Simon, German Physicist, born Erlangen, a small, but renowned university town in southern Germany in the year 1789 as son of a craftsman. He died in Munich 1854, Professor of Experimental Physics at University of Munich.
2. Data in fig 2.4 have been calculated from eq. 4 in M. Sotoodeh, A.H. Khalid and A.A. Razazadeh: Empirical low-field mobility model for III-V-compounds applicable in device simulation codes, J. Appl. Phys. **87** (6) 2890. These data coincide well with the experimental data quoted therein.

Chapter 3

Donors and Acceptors



Abstract Free electrons and holes in a semiconductor can be created by thermal excitation. In this case, the concentrations of electrons and holes are the same. However, a material can be made “n-type” or “p-type” by “doping”.

Si can be made n-type by doping with an element of the Vth group of the periodic table (exception N!), i.e. with P, As, and Sb. It can be made p-type by doping with an element of the IIIrd group (B, Al, and Ga). These elements give rise to energy levels in the bandgap so close to the conduction band and valence band, respectively, that thermal excitation of carriers to the bands is possible at room temperature.

The fact that the donor or acceptor levels are so close to the corresponding band can be explained within the frame of the Hydrogen model: The ionization energy of a donor or acceptor, as compared to the one of free Hydrogen, is reduced by the shielding of electric fields described by the permittivity ϵ_r and by the fact that the free electron mass has to be replaced by their effective mass for the electrons in the conduction band as well as for the holes in the valence band. Donors and acceptors which can be described by the “Hydrogen model” are called “shallow donors” or “shallow acceptors”, respectively.

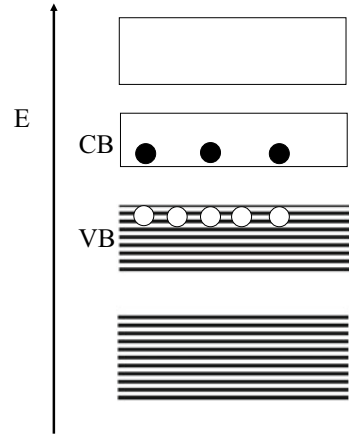
The analogous doping “mechanism” works in compound semiconductors. An interesting case is Si in GaAs. Because its electronegativity is between the one of Ga and As it can go on the Ga- as well as the As-site, it is “amphoteric”.

Not all impurities in a semiconductor are “electrically active”. When an impurity replaces substitutionally an atom of the host material, both belonging to the same group of the periodic table, no free electrons or holes are produced. Interesting cases of “isoelectronic impurities” are B in GaAs or Ge in Si.

3.1 Creation of Free Electrons and Holes by Thermal Excitation

In chapter 1 it was discussed that two kinds of free carriers exist in a semiconductor: Electrons in the conduction band and “holes” in the valence band (see Fig. 1.5). Free

Fig. 3.1 Conduction band CB and valence band VB in a semiconductor



carriers provide the conductivity to semiconductor materials, which then, according to (2.14), is given by

$$\sigma = ne\mu_n + pe\mu_p$$

It was also shown in chapter 1 that full energy bands do not contribute to the current flow in the material. So the relevant energy band providing the conductivity of the material are the conduction band (i.e. the lowest not empty band) and the valence band (i.e. highest not fully occupied band). Figure 3.1. shows this schematically.

The crucial point in making a material a useful semiconductor is that the concentration of electrons and holes can be adjusted by proper preparation of the material, thus adjusting its conductivity in general and the type of conductivity, n-type or p-type, in particular. How can that be done? This is the topic of this chapter.

One way to generate free charge carriers has been shown already. By **thermal excitation** of electrons from the valence band (Fig. 1.5), which in a semiconductor or insulator is full at $T = 0$, to the conduction band which is empty at $T = 0$. Thermal excitation sets in for $T > 0$. It is obvious that conductivity results from this process because now the valence band is no longer full and the conduction band no longer empty. To put it differently, the conductivity of a semiconductor will increase with temperature due to the (exponential) increase of the free carrier concentration with temperature. For a quantitative discussion of this effect, see Sect. 4.4. Nevertheless, the crucial point about thermal excitation from band to band is that in this way the number of electrons produced in the conduction band necessarily is equal to the number of holes in the valence band because every electron excited into the conduction band leaves a hole in the valence band. In order to produce a well-defined type of conductivity, “n-type” by an excess of electrons in the conduction band, or “p-type” by an excess of holes in the valence band, other measures have to be taken. This can be done by “doping” the material with certain impurity atoms.

3.2 Doping of Si

The method is explained in the most important case of Silicon. Si is a member of the fourth group of the periodic table (Table 3.1), so it has four valence electrons.

In the Si lattice, four orbitals are shared with the four neighbors positioned in the corners of the surrounding tetrahedron (Fig. 1.2b). Now let one Si atom be substituted by a Phosphorus atom (Fig. 3.2a).

Phosphor is a member of the fifth group, so it has five valence electrons. One of those is not required to form the orbitals to the four Si neighbors, so it is plausible that this electron is only loosely bound and can be easily removed away from the phosphorus atom as presented schematically in Fig. 3.2a.

In Fig. 3.2b, this situation is presented within the energy band model. **The fifth (“redundant”) electron gives rise to an additional level in the (otherwise forbidden) energy gap, called the “P^{0/+} level”** for reasons explained below. The fact that it can easily be removed from the Phosphorus atom is expressed in the figure by positioning this level close to the conduction band, so only a small amount of

Table 3.1 Excerpt of the periodic table

III	IV	V
B	C	N
Al	Si	P
Ga	Ge	As
In	Sn	Sb

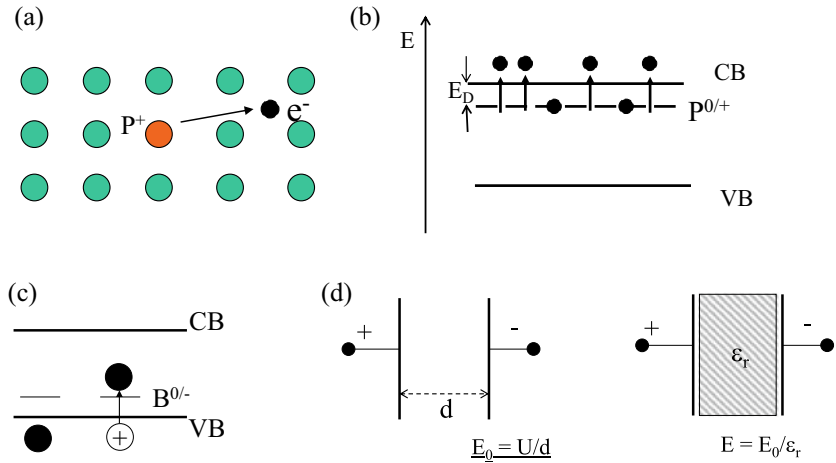


Fig. 3.2 (a): Phosphorus atom substituting a Silicon atom. (b): Energy level in the gap provided by substituting a group-V-atom (like P) for a Si atom. (c): Energy level in the gap provided by substituting a group-III-atom (like B) for a Si atom. (d): Electric field within a medium with permittivity ϵ_r

energy is required to free the redundant electron from the Phosphorus atom. In order to provide a free electron to the conduction band at room temperature, it is obvious that the separation of the $P^{0/+}$ level from the conduction band must be of the order of kT . From $k = 1,38 \times 10^{-23} \text{ JK}^{-1}$, it follows that $kT = 26 \text{ meV}$ at $T = 300 \text{ K}$. This number, by the way, a physicist should know by heart, will often save him from tedious little calculations.

The question of the position of donor and acceptor levels in the bandgap will be considered in Sect. 3.3 on the “Hydrogen model”.

The notation $P^{0/+}$ describes the two possible charge states of the P atom in Si, neutral or positive. The atom as such is neutral, as expressed by the upper index 0 in $P^{0/+}$. When it delivers an electron to the conduction band, it gets positively charged with a charge $+e$, indicated by the upper index $+$ in $P^{0/+}$. Let us make perfectly clear at this point what the energy level marked by $P^{0/+}$ precisely means. The energy difference between this level and the lower edge of the conduction band is the energy needed to change the charge state of P from 0 to $+e$, i.e. to release an electron of the P atom into the conduction band.

The same argument that makes P a donor in Si holds for other members of the Vth group of the periodic table (see Table 3.1). Attention should be given, however, to the fact that N is an exception. As and Sb, however, will do as donors.

Physics comment: Why Nitrogen is not a donor in Si?

The reason for Nitrogen not acting as a donor in Silicon is that, due to the strong binding force between Si and N, a Si–N complex forms such that the N no longer occupies a substitutional site in the Si lattice. Occupation of a substitutional site, however, is a prerequisite for the argument given above for group V atoms acting as donors in Si.

It should be remarked, as a marginal note, that the situation is different in the semiconductor SiC. Here, N sitting substitutionally on the Carbon site acts as a donor.

Let us now consider the question of how to obtain a free hole in the valence band. If an atom of a group III element of the periodic table (we take B as an example) substitutes a Si atom, one electron is missing to form the four orbitals extending to the four Si neighbors. Only a small amount of energy is required for the B atom to take up an electron from the VB, leaving a hole in the VB. The situation is shown in Fig. 3.2c in the frame of the energy band model:

The Boron atom “gives rise to a level in the otherwise forbidden energy gap situated closely to the valence band”. Again the energy separation between the valence band and the “ $B^{0/-}$ -level” should be of the order kT to let this process occur with some probability at temperature T .

3.3 The Hydrogen Model of Donors and Acceptors

As mentioned above, the energy difference between the donor or acceptor levels and the corresponding bands should be of the order kT . Is that really the case?

Let us consider the case of P in Si (see Figs. 3.2a,b). In order to separate an electron from the P atom, leaving behind the ionic core P^+ , a charge $-e$ (of the electron) has to be separated from a charge $+e$ (of the core). This is exactly the problem to be dealt with when the ionization energy E_I of the Hydrogen atom has to be calculated. According to Bohr's model, E_I is given by

$$E_I = (m_0 e^4 / 8h^2 \varepsilon_0^2) = 13,6 \text{ eV} \quad (3.1)$$

where ε_0 is the vacuum permittivity and m_0 the electron rest mass. This is evidently much higher than the value of $kT_{300K} = 0,026 \text{ eV}$ so no separation of the electron from the P-atom would be energetically possible. However, two corrections will significantly (and decisively) lower this value.

- (a) Electrons in a semiconductor may behave like free electrons, but with an “effective mass m_{eff} ” different from the mass of the electron in the vacuum. m_{eff} is determined by the band structure (Chapter 1, 1.15). For Si, a value of $m_{\text{eff}} = 0,26m_0$ is taken (see Table 1.2).
- (b) Fig. 3.2d reminds us of a basic fact of electrostatics:

Let the electric field in an empty capacitor be $E_0 = U/d$. If a material fills the capacitor, the electric field between the capacitor plates is reduced because of the polarization of the medium. The “reduction factor” is the “permittivity ε_r ” of the medium. This is generally valid. Electrical fields within a medium are reduced by this factor. As a consequence, the factor ε_0^2 in the denominator of (3.1) has to be replaced by $(\varepsilon_0 \varepsilon_r)^2$. By combining statements a) and b), one obtains the ionization energy of two charges $+e$ and e

$$\begin{aligned} E_I &= (m_{\text{eff}} e^4 / 8h^2 \varepsilon_r^2 \varepsilon_0^2) = (m_0 e^4 / 8h^2 \varepsilon_0^2) (m_{\text{eff}}/m_0) / \varepsilon_r^2 \\ &= 13,6 \text{ eV} \times (m_{\text{eff}}/m_0) / \varepsilon_r^2 \end{aligned} \quad (3.2a)$$

So the ionization energy of the hydrogen atom is reduced by a “correction factor”.

$$K = (m_{\text{eff}}/m_0) / \varepsilon_r^2 \quad (3.2b)$$

Taking literature values for Si: $m_{\text{eff}}/m_0 = 0,26$ (see Table 1.2) and $\varepsilon_r = 11,9$, we obtain $K = 0,00184$ and thus an ionization energy of the phosphor donor in Si of $0,00184 \times 13,6 \text{ eV} = 0,025 \text{ eV}$. This is, indeed, comparable to thermal energies $kT_{300} = 0,026 \text{ eV}$ which means that the phosphor donor in Si will release its electron into the conduction band with high probability.

As far as the holes are concerned, we have to take into account the existence of two kinds of holes, light holes and heavy holes, the effective masses of which are $0,16m_0$

and $0,49m_0$, respectively. Because m_{eff} enters into (3.2a) and (3.2b) as first-order factor ($\sim m_{\text{eff}}$), it has to be replaced by the geometrical mean $(m_{\text{lh}}m_{\text{hh}})^{1/2}$, which, with the values of Table 1.2, results in $m_{\text{heff}} = 0,28m_0$ and thus in a correction factor $K = 0,0020$, which gives an ionization energy of $0,0020 \times 13,6 \text{ eV} = 27 \text{ meV}$, very well in the range of thermal energies at $T = 300 \text{ K}$.

In summary, the main reason for the ionization energies of the typical donors and acceptors being close to thermal energies at room temperature is the reduction of electric fields in solids due to their polarizability described by ϵ_r . A smaller additional reduction results from the effective mass.

Physics Comment: Spatial Extension of a Donor in Si

The description of the polarization effect in the matter by a permissivity ϵ_r is the result of an averaging process. The electric field around an individual atom, of course, depends strongly on the position relative to that atom. Only as a result of an averaging process over a volume containing many atoms, say a few hundred, we can speak of a mean electric field E_0/ϵ_r . Is this “continuum approximation” justified when we discuss the ionization energy in a solid like Si?

We check this by considering, in the frame of the Hydrogen model, the Bohr radius of the donor. The first Bohr radius of the H-atom is given by

$$r_1 = h^2 \epsilon_0 / \pi m_0 e^2 = 0,0529 \text{ nm} \quad (3.3)$$

Again, by replacing m_0 by m_{eff} and ϵ_0 by $\epsilon_r \epsilon_0$, we obtain

$$r_1(\text{Si}) = 0,0529 \text{ nm} \times \epsilon_r \times m_0 / m_{\text{eff}} = 0,0529 \text{ nm} \times 11,9 / 0,26 = 2,42 \text{ nm}$$

How many Si atoms are wrapped up in a sphere of radius $r_1(\text{Si}) = 2,42 \text{ nm}$?

Such a sphere has a volume of $(4/3)\pi r_1(\text{Si})^3 = 0,6 \times 10^{-25} \text{ m}^3$. Because the density of atoms in Si is $N = 5 \times 10^{28} \text{ m}^{-3}$, a volume of $0,6 \times 10^{-25} \text{ m}^3$ contains

$$Z = N \times 0,6 \times 10^{-25} = 5 \times 10^{28} \times 0,6 \times 10^{-25} = 3000 \text{ Atoms.}$$

In other words, the first Bohr radius of the Phosphor donor encloses about 3000 atoms, and the continuum approximation obviously can be applied.

This treatment of donors and acceptors within the hydrogen model can be considered as a nice example of what is considered a “self-consistent model” in physics. An approximation is made and, in this case, the continuum approximation implied in the use of a permissivity ϵ_r , and on this basis, a result is obtained consistent with this approximation, i.e. a Bohr radius enclosing a large number of atoms.

Tables 3.2 and 3.3 give the ionization energies of donors and acceptors in Si and Ge.

The following observations can be stated from Tables 3.2 and 3.3:

- The values are all roughly of the order of thermal energies at room temperature ($kT_{300K} = 0,026$ eV) as predicted by the hydrogen model.
- The values for P, As, and Sb are pretty close to each other, as well as the ones of B, Al, and Ga. Indeed, the argument provided by the hydrogen model does not differentiate between different donors or acceptors of the same group, but just discusses the ionization energy of a hydrogen-like center in a solid.
- The Ge values are about a factor of four smaller than the ones for Si. This can also be explained in terms of the hydrogen model. Let us consider the correction factor $K = (m_{\text{eff}}/m_0)/\epsilon_r^2$ for Ge. With $m_{\text{eff}} = 0,12m_0$ (Table 1.2) for electrons in Ge and $\epsilon_r = 16,0$, we obtain $K = 4,7 \times 10^{-4}$ to be compared with $K = 18,4 \times 10^{-4}$ for Si (see text below 3.2b), which corresponds to a ratio close to the factor four, almost too good a verification of the hydrogen model!

A donor or acceptor which can be described by the hydrogen model is called a “shallow” donor or acceptor. The notion “shallow” originally refers to the fact that the energy levels are so close to the bands that a thermal excitation at room temperature is possible, i.e. the separation of the bands is of the order kT_{300K} as is the case, e.g. for Si.

It should be noted at this point, however, that a donor or acceptor called “shallow” according to this definition does not necessarily give rise to a level that is close to the conduction or valence band in semiconductors other than Si, Ge, or GaAs, due to a smaller permittivity. Boron, for instance, can be considered as the only controllable dopant in diamond, with an activation energy of 0,37 eV (J.W. Glesener, Appl. Phys. Lett. 64 (1994). 217.). This is considerably larger than the value, e.g. of B in Si (0,045 eV), which is largely due to the lower permittivity of diamond ($\epsilon_r = 5,6$) as compared to $\epsilon_r = 11,9$ in Si.

Table 3.2 Donor levels of group V elements in Si (in eV) at room temperature

	P	As	Sb
Si	0,045	0,049	0,039
Ge	0,012	0,013	0,0096

Table 3.3 Acceptor levels of group III elements in Si (in eV) at room temperature

	B	Al	Ga
Si	0,045	0,057	0,065
Ge	0,0104	0,0102	0,0108

3.4 Shallow Donors and Acceptors in Compound Semiconductors

Considerations analogous to the ones for Si apply to other semiconductor materials. In order to consider the situation in III-V-semiconductors, let us discuss the case of the widely used III-V-material GaAs.

The relevant part of the periodic table in this context is shown in Table 3.4.

A group IIb element like Zn substituting a Ga atom would act as an acceptor, and a group VIb element substituting As would act as a donor. Similar to the case of N in Si, Oxygen provides an exception. It forms a Ga-O complex sitting off-site, so the hydrogen model is not applicable.

An interesting case arises when a group IV element is doped into GaAs. Evidently, when it replaces Ga it will act as a donor, and when it substitutes As, it will act as an acceptor. Si can do both. At high temperatures near the melting point of Si, it occupies a Ga-site, thus acting as a donor, and at lower temperatures it will replace As thus acting as an acceptor. **Such a dopant that may act as a donor as well as an acceptor is called amphoteric.** Si in GaAs is amphoteric. Carbon, however, is not amphoteric. Due to its large electronegativity, it always occupies an As-site thus producing an acceptor. (The notion of electronegativity will be discussed in the next section.)

The amphoteric nature of Si in GaAs plays an important role in GaAs-technology. For instance, a pn-junction in a GaAs film can easily be produced by growing a film at a high temperature in the presence of SiH_4 , resulting in n-material (Si on the Ga-site), and then just reducing the growth temperature such that the Si occupies the As-site resulting in a p-layer.

In order to be considered a useful semiconductor, it should be possible to prepare it by doping in both modifications, n-type and p-type. This, however, is by no means trivial and, according to present knowledge, turns out to be impossible in some cases. As already mentioned above, diamond can only be doped in a controllable way to be p-type by Boron. ZnSe was for a long time considered as a promising semiconductor but it turned out to be difficult to produce it p-type. For decades no way was found to make GaN p-type. This story will be told in more detail in the context of the blue light-emitting diode. The solution to this problem constituted a breakthrough in optoelectronics (see Sect. 11.6).

Table 3.4 Excerpt of the periodic table around the elements of the IIIrd and Vth group

IIb	IIIb	IVb	Vb	VIb
	B	C	N	O
	Al	Si	P	S
Zn	Ga	Ge	As	Se
Cd	In	Sn	Sb	Te

3.5 Isoelectronic Impurities

A donor provides an additional electron to the semiconductor which is thermally released to the conduction band. An acceptor provides a “missing electron” which then is taken out of the valence band, leaving a hole behind. It follows from this understanding that an impurity that has a valence electron configuration identical to the one of the atoms which it substitutes will not be electrically active. An example would be Boron in GaAs replacing a Ga atom because both are third group elements of the periodic table. Another example would be P on an As-site in GaAs, because both, P and As, belong to the fifth group. So in these cases, B and P would be “isoelectronic” impurities which do not affect the carrier concentration, and they are “electrically inactive”.

In the growth of bulk semiconductor crystals from a melt, a general problem is the choice of the crucible material. The crucible tends to be a source of impurities entering the melt and the growing crystal. The solution in the case of, e.g. GaAs is provided by using Boron nitride crucibles (Sect. 7.3). In spite of the relatively high concentration of Boron penetrating the growing crystal (10^{-4} atomic %) out of the crucible, the electronic properties will be very little affected, because the B, substituting a Ga atom, is an isoelectronic impurity, thus overcoming an old problem in GaAs-crystal growth, the choice of a suitable crucible material. In the “old days” when BN-crucibles were not available, GaAs was grown out of quartz crucibles resulting always in n-type material because Si entered the crystal and (at the high growth temperature) occupied a Ga-site.

Isoelectronic impurities, without strongly affecting the electronic properties of a semiconductor, might nevertheless have a strong effect on optical and optoelectronic properties. So GaP:N was used as material for a green-yellow light-emitting diode. Here, the N acts as a formation center of a bound exciton (see Sect. 10.3 for “excitons”).

Physics comment: Electronegativity determines the site of the impurity

In the preceding chapter, it was mentioned that Si in GaAs is amphoteric. It can either go on a Ga-site acting as a donor, or on an As-site acting as an acceptor. Carbon, however, will always occupy an As-site, thus always acting as an acceptor. What makes the difference?

A qualitative way to discuss this is the following:

Consider the case of C in GaAs. Ga is a metal, As a nonmetal. C also is a nonmetal, so it will occupy the As-site and act as an acceptor.

Si, however, is rather something between a metal and a nonmetal. So it can go either on the Ga- or on the As-site.

To express this in a more quantitative way, use is made of the concept of “electronegativity”. **Electronegativity is a measure of the tendency of an element to attract an electron.** The element with the highest electronegativity

is fluorine, the strong tendency of which to form F^- is well known. The element with the smallest electronegativity is, as one would assume, Cesium which rather has a strong tendency to lose an electron and form Cs^+ .

The concept of electronegativity has been given a quantitative turn by Linus Pauling¹ by defining the difference in electronegativity χ of two atoms A and B by

$$|\chi_A - \chi_B| = [E_d(AB) - (E_d(AA) + E_d(BB))/2]^{1/2} (eV)^{-1/2} \quad (3.4)$$

where the dissociation energies E_d of compounds AA, AB, and BB are given in eV and the factor $(eV)^{-1/2}$ is introduced to make χ dimensionless.

Because (3.4) defines only differences, one absolute value can be fixed, which is done by setting χ of hydrogen to be 2.2.

For illustration, a numerical example is given for a simple system.

We consider for illustration of the concept the system H, Br.

From literature, we obtain the dissociation energy of H_2 as $E_d(HH) = 4,53$ eV, $E_d(BrBr) = 2,01$ eV, and $E_d(HBr) = 3,81$ eV. Inserting these numbers into (3.4) results in

$$|\chi_{Br} - \chi_H| = 0,7$$

With χ_H defined to be 2.2, this gives $\chi_{Br} = 2.92$ in good agreement with the literature value of 2.96. Along these lines, a list of the electronegativity of all elements has been worked out.

We now apply the concept of electronegativity to the important case of Si in GaAs.

From the literature values² $\chi_{Ga} = 1.81$, $\chi_{As} = 2.18$, and $\chi_{Si} = 1.90$, we see the electronegativity of Si is fairly close to both, the one of Ga and the one of As, and lies between the two. So Si can occupy the Ga- as well as the As-site, acting as donor or acceptor, respectively. The situation is completely different from the one of C in GaAs. $\chi_C = 2.55$, which is much further away from χ_{Ga} than from χ_{As} . So C will certainly occupy the As-site and act as an acceptor.

¹ Chemist, born 1901 in Portland, Oregon, and died 1994 in Big Sur, California. Nobel chemistry prize 1954; Nobel peace prize 1963.

² See, e.g. wikipedia, "electronegativity" of the elements.

3.6 Questions and Answers

Questions

- Question 1: What is the definition of a donor or an acceptor in terms of the energy band model?
- Question 2: Give examples of shallow donors and acceptors in Si.
- Question 3: Why is the ionization energy of P in Si so much smaller than the ionization energy of the hydrogen atom (which is 13.6 eV)?
- Question 4: What is a shallow donor (or a shallow acceptor)?
- Question 5: What is an amphoteric impurity? Give an example!
- Question 6: Carbon in GaAs, is it a donor or acceptor?

Answers

- Answer 1: A donor is an impurity that gives rise to an energy level in the bandgap which tends to supply an electron to the conduction band. An acceptor impurity is one that gives rise to a level in the bandgap which tends to take an electron out of the valence band, leaving a "hole" there.
- Answer 2: Donors: P, As, and Sb. Acceptors: B, Al, and Ga.
- Answer 3: Because in a solid state system a factor $1/\epsilon_r^2$ has to be included into the Rydberg constant due to the polarizability of the solid, and because the free electron mass has to be replaced by the effective mass.
- Answer 4: One which can be described by the hydrogen model.
- Answer 5: One which can act as a donor as well as an acceptor. Example: Si in GaAs. On the Ga-site, it acts as a donor, and on the As-site as an acceptor.
- Answer 6: Carbon in GaAs goes on the As-site and thus is an acceptor.

Chapter 4

Carrier Statistics



Abstract The total number of free electrons and holes in a semiconductor as well as their distribution over the energy levels depend on the doping level, the density of states in the conduction and the valence band, and the temperature, which enters via the distribution functions of statistical physics (Fermi–Dirac or Boltzmann, depending on carrier concentration). Interesting insights are provided by the “law of mass action” which tells that the product of electron and hole concentrations in thermal equilibrium is independent of the doping level. It follows that the lowest possible total carrier concentration at a given temperature is realized in the intrinsic case and that a “highest possible resistivity” of a semiconductor material can be specified, strongly depending on the bandgap. This highest possible resistivity can be realized for certain applications in so-called “semi-insulating” materials. Carrier statistics explain why Si cannot be used for high-temperature electronics. For this application, high bandgap materials like SiC are required.

4.1 Density of States

In the previous Chap. 3 “Donors and Acceptors”, we learned how the carrier concentration in semiconductors can be controlled by doping. In this chapter, we will treat this issue more quantitatively in the sense that the carrier concentration established in thermal equilibrium in a doped as well as in an undoped semiconductor will be calculated and discussed. A notion to be applied in this context is the one of “density of states”.

To fix ideas, let us consider the electrons in a solid. In principle, the eigenstates of the electrons can be calculated via the Schrodinger equation. Now imagine that all the possible energies of the electrons in the system are represented on a one-dimensional “energy ray” (Fig. 4.1a).

If the number of states dZ between the energies E and $E + dE$ is written as

$$dZ = D(E)dE$$

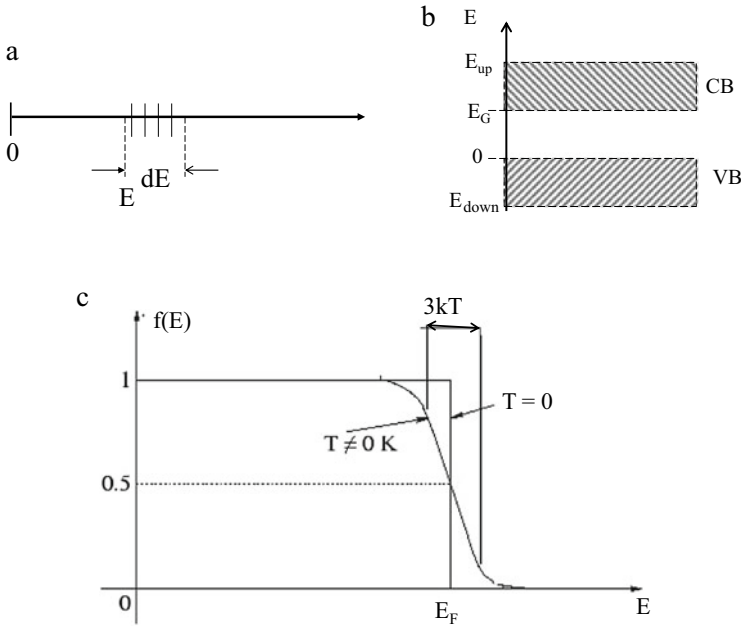


Fig. 4.1 **a** Definition of “density of states”. **b** Scale of energies, the definition of $E = 0$. **c** The Fermi–Dirac distribution function given by eq. 4.2b

then the function $D(E)$ is called the “density of states” of the system. Evidently, $D(E)$ characterizes the distribution of the allowed electronic states on the energy axis.

The electrons in the conduction band of a solid may be considered as free electrons in the potential well formed by the solid. Therefore, use is made of the fact known from solid state theory that the density of states of free electrons in a box is given by

$$D(E) = 8\pi\sqrt{2}(m_e/h^2)^{3/2}\sqrt{E} \quad (4.1)$$

where E —in the case of free electrons—is the kinetic energy of the electron, and m_e its effective mass. Equation (4.1) takes into account the spin degeneracy, which means $D(E)dE$ is the number of electrons that can be accommodated between E and $E + dE$, two on each level. It refers to the unit volume of the sample under consideration. Here and in the following, the scale of the energies is the one shown in Fig. 4.1b.

Because kinetic energy zero corresponds to the edge of the conduction band, E in (4.1) has to be replaced by $E - E_G$, which means $D(E) = 8\pi\sqrt{2}(m_e/h^2)^{3/2}(E - E_G)^{1/2}$.

4.2 Free Carrier Concentration in Thermal Equilibrium

With this notion of “density of states”, we can write for the number of electrons in the conduction band:

$$n = \int_{E_G}^{E_{up}} D(E)f(E)dE \quad (4.2a)$$

with

$$f(E) = 1/\left[\exp(E - E_F)/kT + 1\right] \quad (4.2b)$$

$f(E)$ is the probability that an electronic state with energy E is occupied, see Fig. 4.1c, according to the Fermi–Dirac statistics.

Equation (4.2a) expresses the fact that the total number of electrons in the conduction band equals the number of states between E and $E + dE$, given by $D(E)dE$, multiplied by the probability $f(E)$ that the states are occupied, and summed over all states between the lower and upper edge of the conduction band (see Fig. 4.1b).

By introducing $D(E)$ from (4.1) and $f(E)$ from (4.2b) into (4.2a), we obtain for n

$$n = 2(2\pi m_e kT/h^2)^{3/2} \exp[-(E_G - E_F)/kT] = N_C \exp[-(E_G - E_F)/kT] \quad (4.3a)$$

with

$$N_C = 2(2\pi m_e kT/h^2)^{3/2} \quad (4.3b)$$

For m_e , the “density of states mass” has to be taken (Table 1.3).

Because $D(E)$ in (4.2a) marks the density of states per unit volume, n in (4.3a) is the electron density, i.e. the number of electrons per unit volume.

N_C is called the “**effective density of states**” of the conduction band. (4.3a) can be easily obtained from (4.2a) by looking up the corresponding integrals. It is important to note, however, that two approximations have been made in obtaining (4.3a) as follows:

- (a) In the expression for the Fermi–Dirac distribution $f(E)$ (see 4.2b), we assumed $\exp[(E - E_F)/kT] \gg 1$ for the energies in the conduction band, i.e. for all energies between the lower and upper edge of the integral. So the 1 in the denominator can be neglected. This amounts to replacing the Fermi–Dirac statistics by the Boltzmann Statistics. In the next physics comment “Boltzmann or Fermi–Dirac Statistics in a semiconductor”, we shall come back to that important point.
- (b) When $\exp[(E - E_F)/kT] \gg 1$ for all energies of the conduction band, including the ones close to the lower edge of the conduction band, we may replace the upper edge E_{up} of the integral by ∞ . This is so because $\exp[(E - E_F)/kT]$ will

assume very large values at the upper edge of the conductivity band and beyond, given the fact that the width of the band is typically a few eV, while kT is only some 10 s of meVs at decent temperatures, such that $f(E) = 1/\exp[(E-E_F)/kT] \rightarrow 0$ in this energy region. So the integral (4.2a) almost vanishes at the upper end of the integral and beyond, and the upper edge of the integral can be put to ∞ without significant change of the integral.

The same kind of calculation can be done to calculate the total number of holes in the valence band in thermal equilibrium. In the case of the holes, we have to start out with the equation

$$p = \int_{E_{\text{low}}}^0 D(E)[1 - f(E)]dE \quad (4.4a)$$

Equation (4.4a) differs from (4.2a) basically by the fact that $f(E)$ is replaced by $1-f(E)$, because, evidently, the number of holes between E and $E + dE$ equals the number of levels $D(E)dE$, multiplied by the probability $1-f(E)$ that the level is not occupied. The total number of holes in the valence band is then obtained by summing up all contributions from E_{low} up to $E = 0$. With the analogous approximations as above (we take E_{low} as $-\infty$ in (4.4a) and $\exp(E_F - E) \gg 1$ in (4.2b)), one obtains the following:

$$p = 2(2\pi m_h kT/h^2)^{3/2} \exp(-E_F/kT) = N_V \exp(-E_F/kT) \quad (4.5a)$$

where m_h is the effective mass of the holes and

$$N_V = 2(2\pi m_h kT/h^2)^{3/2} \quad (4.5b)$$

N_V is called the “Effective density of states” of the valence band. Again, p is the number of holes per unit volume because $D(E)$ in (4.4a) stands for the density of states per unit volume. For m_h , the “density of states mass” has to be taken in this context (Table 1.3).

Physics Comment: Boltzmann or Fermi–Dirac Distribution in a Semiconductor?

In textbooks on semiconductor physics, sometimes Boltzmann statistics and sometimes the Fermi-Statistics are used to describe the distribution of charge carriers in the bands. It is quite instructive to consider in a more detailed and quantitative way this issue.

In the following, we focus our discussion on the electrons in the conduction band. The one for the holes in the valence band would run correspondingly.

In deriving (4.3a), the approximation was made

$$\exp(E - E_F)/kT \gg 1 \quad (4.6a)$$

Only under this condition, (4.3a) holds.

When this approximation (i.e. 4.6a) is made, in the Fermi–Dirac distribution (which has to be applied in a strict manner because electrons are fermions)

$$f(E) = 1/[\exp(E - E_F)/kT + 1] \quad (4.6b)$$

the 1 in the denominator can be omitted and thus $f(E)$ is written as

$$f(E) \cong \exp[-(E - E_F)/kT] \quad (4.6c)$$

which amounts to replacing the Fermi–Dirac distribution by the Boltzmann distribution.

Because $E > E_G$ for all electrons in the conduction band, approximation (4.6a) is fulfilled when

$$\exp(E_G - E_F)/kT \gg 1 \text{ or } \exp[-(E_G - E_F)/kT] \ll 1 \quad (4.6d)$$

so we might as well say that in case that approximation (4.6d) is valuable the 1 in the denominator of the Fermi–Dirac distribution (4.6b) can be omitted and thus the Fermi–Dirac distribution can be replaced by the Boltzmann distribution in a good approximation.

It was stated above that (4.3a) only holds when approximation (4.6a) and correspondingly (4.6d) hold. By stating this, we put a limit on the range of n values for which this equation holds. Because of (4.6d) this range is given by $n \ll N_C$. So the question arises: **Why does n have to be much smaller than N_C such that the Fermi–Dirac distribution can be replaced by the Boltzmann distribution? In which way is N_C a border stone separating the Fermi–Dirac statistics from the Boltzmann statistics?** The answer to this question leads us directly to the basic ideas of statistical mechanics, i.e. to the fact that Fermi–Dirac statistics apply to nondistinguishable particles, while Boltzmann statistics apply to distinguishable particles. **It turns out, indeed, that the electrons in the conduction band can be considered distinguishable when $n \ll N_C$, and thus objects to Boltzmann statistics, and indistinguishable when $n \geq N_C$.**

To explain this, we have to refer to the theory of the free electron gas, as the electrons in the conduction band are forming one. In principle, even identical particles like electrons can be distinguished via their positions in space and their momenta. However, because of Heisenberg's uncertainty relation, this is no longer the case if their density in phase space is comparable or larger than 1

per phase space cell $(\hbar)^3$ because in this case their orbits cannot be described separately by identifying their space and momentum coordinates. If, however, their density is much smaller than 1 per phase space volume $(\hbar)^3$, they can be distinguished. Now the theory of the free electron gas tells us that the density of the electrons in phase space is just 1 per cell $(\hbar)^3$ when its spatial density at temperature T is given by the “density of the degenerate electron gas” n_{deg} :

$$n_{\text{deg}} = 2(2\pi m_e kT)^{3/2} \quad (4.7)$$

Comparison of (4.7) with (4.3b) for N_C shows that N_C is just the density of the degenerate electron gas for the electrons in the conduction band. So it is clear that as long as $n \ll N_C$ the conduction band can be considered as distinguishable and the Boltzmann statistics be applied, otherwise not.

Some numbers

N_C in Si at room temperature is $2,8 \cdot 10^{19} \text{ cm}^{-3}$; see Table 4.1. So as long as the density of electrons in the conduction band (i.e. the density of donors) is much smaller than that, the Boltzmann statistics can be applied at room temperature. For the vast majority of electronic devices, this is the case (an exception, to give an example, being the conditions in semiconductor lasers; see the end of Sect. 12.1).

Another way to see that $(E_G - E_F)/kT \gg 1$ in practical cases is to calculate E_F explicitly for such an example:

Let us consider a doping concentration of (order of magnitude) $n = 10^{15} \text{ cm}^{-3}$. Because this is much smaller than $N_C = 2,8 \cdot 10^{19} \text{ cm}^{-3}$, (4.3a) can be applied. From (4.3a), we obtain

$$E_G - E_F = kT(\ln N_C - \ln n) = kT \ln(N_C/n)$$

Because $kT = 0,26 \text{ eV}$ at room temperature

$$E_G - E_F = 0,026 \text{ eV} \times 10,24 = 0,27 \text{ eV}$$

Table 4.1 gives numbers for some important materials (data taken from eesemi.com/sigeGaas.htm)

	Ge	Si	GaAs
E_G/eV	0,66	1,12	1,42
$N_C/10^{19} \text{ cm}^{-3}$	1,04	2,8	0,047
$N_V/10^{19} \text{ cm}^{-3}$	0,6	1,04	0,7
n_i/cm^{-3}	$2,4 \cdot 10^{13}$	$1,45 \cdot 10^{10}$	$1,79 \cdot 10^6$

at room temperature. The Fermi energy E_F is 0,27 eV below the band edge of the conduction band (with a bandgap of Si of 1,12 eV) and $\exp(E_G - E_F)/kT = e^{10.4} = 3,2 * 10^4 \gg 1$.

For a ten times higher doping level ($n = 10^{16} \text{ cm}^{-3}$), we get $E_G - E_F = 0,026 \text{ eV} \times 7,94 = 0.21 \text{ eV}$ and $\exp(E_G - E_F)/kT = e^{8.08} = 3,2 * 10^3$.

4.3 The Law of Mass Action

At a first glimpse, (4.3a) and (4.5a) might be considered as equations to calculate n and p at a given temperature. This, however, is an illusion because first E_F would have to be known, which in turn depends on the doping. So the two equations rather can be considered as relations between the carrier concentrations and E_F .

A very useful expression, however, is obtained by multiplying n and p , which results in

$$np = N_C N_V \exp(-E_G/kT) \quad (4.8)$$

So the product of n and p no longer contains E_F . Because E_F is determined by the doping level of the material, we conclude the following:

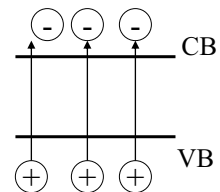
The product of the carrier concentrations n and p in a semiconductor material (at thermal equilibrium) is independent of the doping! (4.8) in connection with this statement is called the “law of mass action”.

A couple of interesting conclusions can be drawn from the law of mass action, which will be given in terms of four statements in the following section

4.4 The Intrinsic Case

Because (4.8) applies to any level of doping, it applies in particular to the case of completely undoped samples, i.e. to the case of “intrinsic material”. If no dopants are incorporated, free charge carriers are produced by thermal excitation of electrons from the valence band to the conduction band, a process resulting in electron–hole pairs (Fig. 4.2).

Fig. 4.2 Thermal excitation from VB to CB



So the intrinsic case is characterized by $n = p$. For $n = p$, (4.8) transforms into $n^2 = p^2 = N_C N_V \exp(-E_G/kT)$, or

$$n_i = p_i = (N_C N_V)^{1/2} \exp(-E_G/2kT) \quad (4.9)$$

The suffix i stands for intrinsic.

Mind the factor $1/2$ in the exponential, different from what we are “accustomed” to from the Boltzmann factor.

So let us keep in mind the following:

Statement 1: The intrinsic carrier concentration increases exponentially with temperature, and the exponent is given by $-E_G/2kT$, not by the exponent of the “usual” Boltzmann factor $\exp(-E_G/kT)$.

Because of the expression (4.9) for n_i , the law of mass action (4.8) is often in the textbooks written in the form

$$np = n_i^2 \quad (4.10)$$

This is certainly correct but can give rise to some misunderstanding. It somehow suggests that the intrinsic carrier concentration has a certain special significance such that it fixes the product of np for any doping situation. The fact, however, is that the law of mass action (4.8) applies equally to any doping, including the intrinsic case $n = n_i$. This is the way one should read (4.10).

It is noteworthy how strongly the intrinsic carrier concentration depends on the bandgap due to the exponential dependence in (4.9).

The intrinsic carrier concentration is by definition the one encountered when the semiconductor is not doped. For almost all semiconductor materials used in practice, however, this case does not occur in reality, due to the fact that they just cannot be produced so pure that the carrier concentration approaches the intrinsic value. Let us consider the case of the most widely used semiconductor Si. The purest material available has an impurity concentration of about $5 \cdot 10^{11} \text{ cm}^{-3}$, some of those impurities acting as donors or acceptors. So the intrinsic value $n_i = 1,4 \cdot 10^{10} \text{ cm}^{-3}$ (at room temperature) cannot be achieved by aiming for the highest possible purity. This latter statement is even more true for a material like GaAs with its intrinsic carrier concentration of some 10^6 cm^{-3} . An exception is Ge, where the intrinsic carrier concentration of $\approx 10^{13} \text{ cm}^{-3}$ can be approached by producing a material of maximum purity, given the circumstance that here the intrinsic carrier concentration is particularly high due to the relatively small bandgap of 0,66 eV.

Physics Comment: A Remark Concerning the Definition of a Semiconductor

Often a semiconductor is defined as a material whose bandgap is so small that a useful number of charge carriers is produced by thermal excitation from the

valence band to the conduction band. As our previous discussion shows, this definition describes a situation that in reality essentially never occurs (at least at room temperature). The charge carriers originate from impurity atoms, shallow dopants, not from the intrinsic thermal excitation process. So a better definition of a semiconductor seems to be as follows: A semiconductor material is a solid which can be doped with shallow donors or acceptors.

So we have to choose between these two seemingly totally different definitions:

- (a) A semiconductor is a solid with a bandgap so small that a carrier concentration providing a useful conductivity can be achieved by thermal excitation from the valence band to the conduction band, or
- (b) a semiconductor is a solid which can be doped with shallow donors or acceptors.

The choice, however, is not that fundamental as it might seem at first sight. We are faced with an interesting “reconciliation” between these two points of view: Only materials with fairly small bandgaps can be doped with shallow donors and acceptors! So there is no such thing as a shallow donor or acceptor in, e.g. KCl or Al_2O_3 !

We shall not enter a discussion on this interesting point here, related to the chemical bonding situation.

The Fermi Energy in the intrinsic case

As mentioned above, the intrinsic case is characterized by $n = p$. By putting

$n = p$, taken from (4.3a) and (4.5a), we obtain for E_F :

$$E_F = (E_G/2) + (3/4)kT \ln(m_h/m_e) \quad (4.11)$$

Let us insert typical numbers from Table 1.3, case of Si:

With $m_{e,\text{dos}} = 1.08m_0$, $m_{h,\text{dos}} = 0.81m_0$ (m_0 = free electron mass), and for room temperature ($kT = 0.026$ eV), we obtain

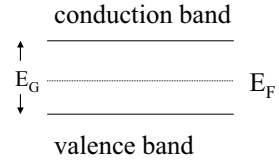
$$E_F = E_G/2 - 0,006 \text{ eV}$$

Because $E_G = 1,12$ eV for Si, this means that the Fermi energy for intrinsic Si is very close to the middle of the bandgap. Actually, this is generally true for semiconductors because the ratio of the two effective masses is roughly somewhere between 0,1 and 1, and because of the smallness of kT even at relatively high temperatures. In the special case that $m_e = m_h$, E_F would be just at $E_G/2$ (Fig. 4.3).

So let us summarize the following:

Statement 2: In an intrinsic semiconductor, the Fermi energy is close to the middle of the bandgap. It is exactly in the middle if $m_h = m_e$ ($\ln(m_h/m_e) = 0$).

Fig. 4.3 Fermi energy E_F in the intrinsic semiconductor



The smallest total carrier concentration in a semiconductor at a given temperature:

Let $z = n + p$ be the total carrier concentration. It is large at high doping and small at low doping levels. From the law of mass action, it follows immediately:

Statement 3: The intrinsic carrier concentration is the smallest one that a semiconductor can have at a given temperature.

It is known, and most easily shown, that the sum of two quantities a and b for a fixed size of the product ab is the smallest for $a = b$. (The best-known application of this fact is that the square has the smallest circumference of all rectangles of a given area). Applied to our problem, this means that $z = n + p$ is the smallest for $n = p$, because np is fixed by the law of mass action. $n = p$, however, characterizes the intrinsic case.

The lowest possible conductivity of a semiconductor at a given temperature:

An important fact for semiconductor technology, closely related to the previous statement 3, is the following:

Statement 4: At a given temperature, a semiconductor material has a minimum value of conductivity and (correspondingly) a maximum value of its resistivity at carrier concentrations close to the intrinsic ones.

The conductivity of a semiconductor with a free electron concentration n and a free hole concentration p is given by (see 2.14)

$$\sigma = ne\mu_e + pe\mu_h$$

Because of the law of mass action in the form (4.10), we can write

$$\sigma = ne\mu_e + pe\mu_h = ne\mu_e + (n_i^2/n)e\mu_h$$

From $d\sigma/dn = 0$, one gets that the minimal conductivity $\sigma_{\min}$ is achieved for an electron concentration

$$n_{\min} = n_i(\mu_h/\mu_e)^{1/2} \quad (4.12a)$$

and that the minimum of σ is

$$\sigma_{\min} = 2n_i e (\mu_h \mu_e)^{1/2} \quad (4.12b)$$

Equation (4.12a) teaches that for the case that $\mu_e = \mu_h$, the minimum $\sigma_{\min}$ of the conductivity is achieved for $n = n_i$, otherwise, for $\mu_e \neq \mu_h$, it is close to n_i , because μ_e and μ_h generally do not differ by more than a factor of about 10.

Because of $\rho = 1/\sigma$, we obtain from (4.12b) the maximum possible resistivity at a given temperature

$$\rho_{\max} = 1/[2n_i e (\mu_h \mu_e)^{1/2}] \quad (4.12c)$$

With the numbers of the mobilities from Table 2.2 and for n_i from Table 4.1, one gets for Si ($\mu_e = 0.15 \text{ m}^2/\text{Vs}$, $\mu_h = 0.0475 \text{ m}^2/\text{Vs}$, $n_i = 1.45 \cdot 10^{16} \text{ m}^{-3}$)

$$\sigma_{\min} = 0.38 \cdot 10^{-3} \text{ S/m}, \quad 1/\sigma_{\min} = \rho_{\max} = 2.6 \cdot 10^3 \Omega\text{m at room temperature}$$

and for GaAs ($\mu_e = 0.85 \text{ m}^2/\text{Vs}$, $\mu_h = 0.04 \text{ m}^2/\text{Vs}$, $n_i = 1.79 \cdot 10^{12} \text{ m}^{-3}$)

$$\sigma_{\min} = 1.95 \cdot 10^{-7} \text{ S/m}, \quad 1/\sigma_{\min} = \rho_{\max} = 0.96 \cdot 10^7 \Omega\text{m}$$

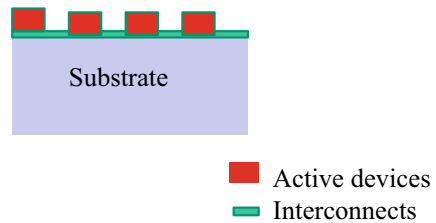
It is obvious that the upper limit of the resistivity of a semiconductor is strongly determined by the bandgap because n_i depends so strongly on E_G (Table 4.1).

4.5 Semi-insulating Semiconductors

The fact that the resistivity of a semiconductor material has an upper limit specific for that semiconductor given by (4.12c) is relevant for electronics insofar as there are important applications where a high resistivity is desirable. This is related to the fact that electronic devices and integrated circuits (“chips”) in many cases consist of active layers which are grown on a substrate, as plotted schematically in Fig. 4.4.

In microwave devices in the Gigahertz range, for example, the resistivity of the substrate should be large in order to avoid losses from heating by Eddy currents. Furthermore, in integrated circuits, the individual components should be electrically isolated from each other, which can only be achieved by sufficiently high resistivity

Fig. 4.4 Active electronic components fabricated on a “substrate”



of the substrate. Experience shows that for this purpose a resistivity of $10^5 - 10^6 \Omega\text{m}$ is required. As the numbers given above (in the context of (4.12c) and Table 4.1) show, this cannot be achieved at room temperature for Si, but can well be achieved for GaAs. Therefore, in Si integrated circuits, always a blocking pn-junction (Chap. 9) between active layers and substrate is required to stop currents from flowing into the substrate, which gives rise to additional complications in fabrication and circuit design.

When, as stated above (statement 4 in the previous Sect. 4.4), the highest values of resistivity are achieved in the semiconductor with carrier concentrations close to the intrinsic ones, does that mean that they cannot be achieved in reality, because, as stated in the preceding section, the intrinsic case cannot be realized for most semiconductors because the necessary purity cannot be achieved?

The answer is no. Very low concentrations of carriers can, as a matter of fact, be achieved not only by getting rid of shallow donors and acceptors, but by a process called “compensation” of donors and acceptors. The idea is the following: If a semiconductor contains N_D shallow donors and N_A shallow acceptors (with, e.g. $N_D > N_A$), the electrons of the donor fall onto the acceptor levels such that the resulting “effective” donor concentration is close to $N_D - N_A$ (Fig. 4.5a).

Because the concentration of electrically active impurities cannot be pushed under something like 10^{18} m^{-3} , two concentrations of order 10^{18} m^{-3} would have to be equilibrated to about $n_i = 2 \cdot 10^6 \text{ cm}^{-3} = 2 \cdot 10^{12} \text{ m}^{-3}$ in the case of GaAs (see Table 4.1) in order to achieve approximately the intrinsic carrier concentration. This is

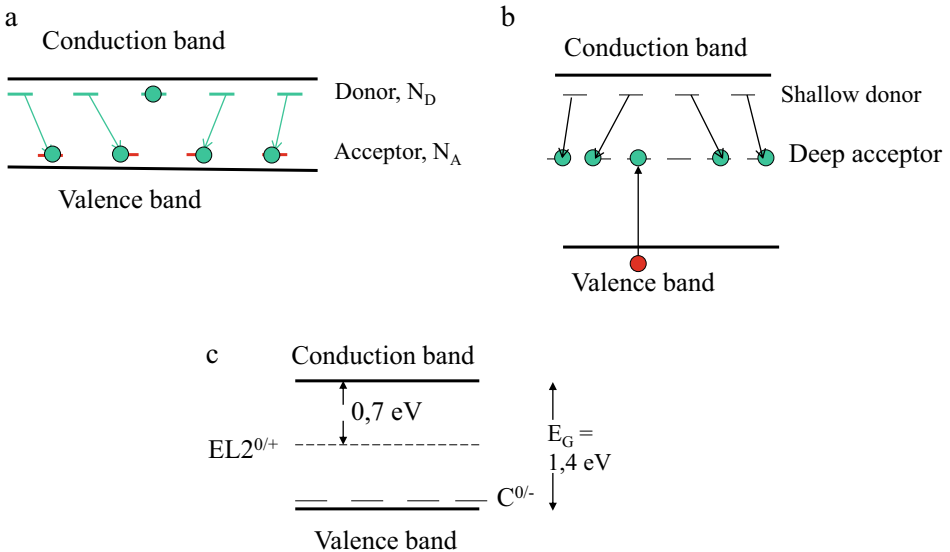


Fig. 4.5 **a** Mutual compensation of shallow donors and acceptors. **b** Compensation of a shallow donor by a deep acceptor, resulting in semi-insulating material. **c** Semi-insulating GaAs. The residual shallow acceptor carbon is compensated by the deep donor EL2

impossible, of course. **Instead, however, carrier concentrations near the intrinsic ones can be achieved by compensation of a residual shallow donor by a deep acceptor in excess concentration with its energy level close to the middle of the bandgap (Fig. 4.5b), while a residual shallow acceptor can be compensated by a deep donor.**

In the first case, for instance, the electrons would fall on the deep acceptor instead of filling the conduction band, so few electrons would arrive in the conduction band. The excess acceptor level would only scarcely be occupied by electrons from the valence band, leaving holes behind, because the level is so deep, so only a few holes would occur in the valence band. So the overall carrier concentration would be very low.

For clarification, the method is explained in more detail in Fig. 4.5c for the case of GaAs, which is of high practical and commercial importance: Semi-insulating GaAs is strongly used as substrate material for high-frequency devices.

After optimum purification of GaAs, the remaining electrically active impurity is carbon (Fig. 4.5c). Carbon in GaAs occupies an As-site, so it acts as a shallow acceptor and thus would have to be compensated by a deep donor with a level near the middle of the bandgap. Such a deep donor is the so-called EL2-defect. The EL2-defect is a famous example of an intrinsic defect center: It is a so-called “antisite”, where an As-atom is sitting on a Ga-site. Thus, an element of the fifth group is sitting on the site of an atom of the third group, providing two excess electrons. The level $EL2^{0/+}$ (one electron getting lifted to the conduction band) at 0.7 eV (Fig. 4.5c) acts as the deep donor compensating for the shallow level $C^{0/-}$.

The “antisite” (As on a Ga-site), by the way, in practice is produced by adding a small excess of As to the melt in the growth process (for growth of GaAs, see Sect. 7.3).

Physics Comment: Semi-insulating Semiconductors, an “Imitation” of the Intrinsic Case

Why can a carrier concentration near the intrinsic concentration be achieved following the recipe given above, i.e. by compensating a residual shallow donor by an excess deep acceptor, or a residual shallow acceptor by an excess deep donor?

We discuss this issue for the first case (shallow donor, deep acceptor). The argument for the second case runs correspondingly.

Because an excess of the deep acceptor is introduced, the deep acceptor is only partially occupied by the electrons released from the shallow donor. Due to the narrow transition region of the Fermi–Dirac function (only a few kT , see Fig. 4.1c), this means that the Fermi energy E_F is close to the level of the deep acceptor. (By the way, this is a fairly general fact: The Fermi energy E_F is usually close to the level of the dominating defect center). Because the deep level is chosen such that its energy level is close to the middle of the bandgap, this means that the Fermi level is close to where it is in the intrinsic

case (see 4.11 and corresponding statement). Because E_F defines the carrier concentration (see 4.3 and 4.5), this means that the carrier concentration is close to the intrinsic one.

So the recipe given for “semi-insulating” semiconductors simulates the intrinsic case in the sense that it fixes the Fermi level close to the energy where it is in the intrinsic case.

4.6 The Temperature Dependence of Carrier Concentration. Why Si-Electronics Does not Work at Elevated Temperatures

Let us first consider the intrinsic case, given by (4.9). By plotting $(\ln n_i)$ versus $1/T$, we get an “Arrhenius plot”, i.e. a straight line with slope $-E_G/2 T$, as shown in Fig. 4.6 for three different semiconductors (Si, $E_G = 1,12$ eV; GaAs, $E_G = 1,42$ eV; and 6H-SiC, $E_G = 3,28$ eV).¹ The figure clearly shows that the larger the bandgap, the steeper the slope, and the smaller the intrinsic carrier concentration at a given temperature.

An interesting conclusion can be taken from Fig. 4.6. In spite of the terrific achievements of Si-electronics, a drawback is that it does not work at temperatures above 200–250 °C. This is a considerable drawback in view of the fact that there are many applications in technology where electronics at elevated temperature are required, like in chemical engineering, in turbines, and even in motor vehicles, where sensor electronics close to the cylinder of the engine would be useful. What is the problem with Si-electronics at high temperatures?

The properties and performance of essentially all electronic devices are determined by the doping of its constituents, as, e.g. in a pn-junction acting as a rectifier (Sect. 9.2), a bipolar transistor of npn-pnp-structure (Sect. 13.1), a field effect transistor (Sect. 13.2), and many others. Typical doping concentrations would be, roughly speaking, 10^{14} cm⁻³ up to 10^{18} cm⁻³. As Fig. 4.6 shows, the intrinsic carrier concentration reaches about 10^{14} cm⁻³ at 200 °C and quickly approaches much higher values. The consequence is that the dopant concentration, n- or p-type, which has been introduced by doping in order to warrant the proper performance of the device is “flooded” by the intrinsic carrier concentration which, making things even worth, kills the n- or p-type character because it tends toward the equal population of electrons and holes (Sect. 3.1 and Fig. 1.5).

For operation at higher temperatures evidently, an electronic material of higher bandgap is required, as, e.g. SiC or GaN. As Fig. 4.6 shows, the “critical” intrinsic

¹ Data are taken from J.C. Zelter, Solid-State Electronics 42 (1998) 2153, quoted in Willander, M., Friesel, M., Wahab, Qu. et al. Silicon carbide and diamond for high temperature device applications. *J Mater Sci: Mater Electron* 17, 1 (2006). <https://doi.org/10.1007/s10854-005-5137-4>.

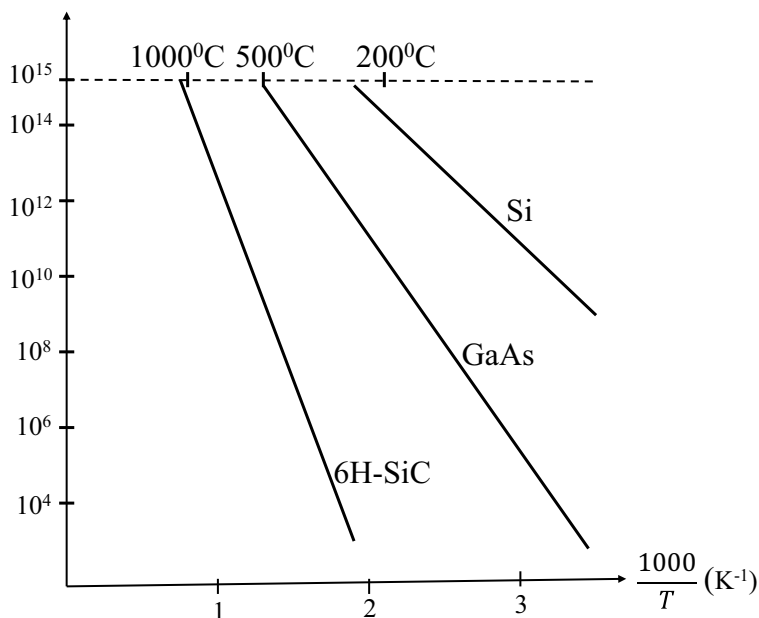


Fig. 4.6 Intrinsic carrier concentration versus reciprocal temperature for Si, GaAs, 6H-SiC (Data are taken from Willander, M., Friesel, M., Wahab, Qu. et al. Silicon carbide and diamond for high temperature device applications. J Mater Sci: Mater Electron 17, 1 (2006). <https://doi.org/10.1007/s10854-005-5137-4>)

concentration of about $10^{14} - 10^{15} \text{ cm}^{-3}$ is only reached at about 1000°C for 6H-SiC. So high-temperature electronics is a domain of “large bandgap materials” like SiC and GaN. The notion “large bandgap materials”, by the way, often refers to materials with bandgaps larger than 3 eV, like SiC and GaN. Transistors working at red heat (ca. 700°C) have been realized in SiC.

Physics Comment: General Features of the Law of Mass Action

In the following, some comments shall be given on the physics behind the law of mass action. The term “law of mass action” is familiar from chemistry, a simple case being the one of H^+ and $(\text{OH})^-$ ions in water. For aqueous solutions, it says

$$[\text{H}^+][\text{OH}^-] = A(T)$$

which means that the product $[H^+][OH^-]$ is just a function of temperature. For a strong acid, $[H^+]$ is large and $[OH^-]$ is small; for a strong base, it is the other way round, but the product at a given temperature is always the same.

The basics of the law of mass action, be it the one for carriers in semiconductors or the one valid in chemistry, can be understood as follows (we discuss the semiconductor case exemplarily):

Let there be a temperature-dependent mechanism of creating electron and hole pairs in the semiconductor. So

$$(dn/dt)_{\text{generation}} = (dp/dt)_{\text{generation}} = g(T) \quad (4.13)$$

At the same time, we have a recombination process where electrons in the conduction band recombine with holes in the valence band under the emission of a photon or of phonons. The rate of such a recombination process, of course, is proportional to the concentration of electrons n and the concentration of free levels in the valence band, i.e. proportional to the concentrations of holes:

$$(dn/dt)_{\text{rec}} = (dp/dt)_{\text{rec}} = -r(T)np \quad (4.14)$$

where $r(T)$ is some function of T . In the presence of both, the generation of electron–hole pairs and electron–hole recombination, we get

$$dn/dt = dp/dt = g(T) - r(T)np$$

In equilibrium (stationary state) $dn/dt = dp/dt = 0$, so

$$np = g(T)/r(T) \quad (4.15)$$

So the product of n and p just depends on T , which is the characteristic feature of the law of mass action.

The crucial point in the above simple derivation is that the generation rate (4.13) does neither depend on n nor p , which leads directly to (4.15). This is so because the electrons for the creation of electron–hole pairs are taken from the essentially unlimited reservoir of electrons from the valence band and thus have nothing to do with the concentration of the already excited electrons. Because this generation process is driven by phonons or by photons (black body radiation!), it is just T -dependent.

Basically, the same consideration holds for the example of aqueous solutions in chemistry. The point here is that the H^+ ions are produced by ionization of the essentially unlimited reservoir of H_2O molecules in the water, so their concentration has nothing to do with the concentration of already existing $[H^+]$ or $[OH^-]$ ions.

So the law of mass action holds if two components of a system are annihilated by mutual recombination and are generated from some essentially unlimited pool, meaning that this pool is not noticeably affected by the generation process.

4.7 Questions and Answers

- Question 1: What is the density of states in a semiconductor?
- Question 2: What does the law of mass action say?
- Question 3: Where is the Fermi energy situated in an intrinsic semiconductor?
- Question 4: What is the smallest possible carrier concentration in a semiconductor at a given temperature?
- Question 5: Is a concentration of residual shallow donors and acceptors smaller than the intrinsic carriers concentration required to realize the intrinsic carrier concentration?
- Question 6: What is the compensating deep donor in commercial semi-insulating GaAs?
- Question 7: Why do Si-devices not work at elevated temperatures (above -300°C)?
- Question 8: Which kind of semiconductor materials are required for high-temperature electronics?
- Answer 1: When the number dZ of electronic states with energies between E and $E + dE$ is written as $dZ = D(E)dE$, then $D(E)$ is called the “density of states”.
- Answer 2: The product of electron and hole concentration is independent of the Fermi level, i.e. of the doping.
- Answer 3: Near the middle of the bandgap. And exactly in the middle of the bandgap if the effective masses of electrons and holes are equal.
- Answer 4: The intrinsic one.
- Answer 5: No. The intrinsic carrier concentration can be achieved by compensation in “semi-insulating” semiconductors. In those, residual shallow donors are compensated by a deep acceptor or residual shallow acceptors by a deep donor.
- Answer 6: The EL2-defect, i.e. the As-antisite.
- Answer 7: Because the intrinsic carriers “flood” the material beyond the p- and n-doping level generally required for the performance of the device.
- Answer 8: Large bandgap materials with $E_G \geq 3\text{eV}$. For example, 6H-SiC ($E_G = 3.0\text{ eV}$ and 4H-SiC ($E_G = 3.23\text{ eV}$) or GaN ($E_G = 3.4\text{ eV}$).

Chapter 5

Fabrication of Electronic Silicon



Abstract Electronics provide extreme demands on the purity and crystalline perfection of materials. Such a “state of the art” has to an extreme degree been achieved in the case of Silicon, the most widely used semiconductor material. Electronic Si is produced via the Siemens process. In this process, first “metallurgical Si” is converted into SiHCl_3 (trichlorosilane), the SiHCl_3 then is purified to the necessary degree of about 10^{-10} by distillation, and then it is reconverted into Si. Subsequently out of this electronic grade polycrystalline Si single crystals of high perfection have to be produced. This is done either by growth from the melt, the “Czochralski process”, or by the float zone process, the former one being most widely used for industrial applications. The special feature of the float zone process of Si is that it works crucible free, so essentially no impurities are incorporated via the growth process. Extreme demands are set for the “wafering” process. Wafers of extreme geometrical perfection (the wafer shape being described by “bow” and “warp”) and almost perfect flatness are obtained after production steps including sawing, lapping, and polishing. “Wafering” is the most important step in the total value chain of electronic grade Si.

5.1 Silicon Valley, Si-Technology, and Si-Age

Silicon is by far the dominant electronic material, as commonly used notions like “Si-technology” as a pseudonym for modern electronic technology in general, or “Silicon valley”, “Silicon age” demonstrate. This does not mean that it is necessarily the best and most suitable material for all applications. On the contrary, depending on the application, often a different semiconductor material might be preferable. However, as far as materials are concerned, the development of electronics proceeds essentially following the rule: Whatever can be done with Si will be done with Si. This is because the material can be supplied at a reasonable price in a quality fulfilling the extreme standards of modern electronics and, equally important as a key factor, because for most requirements the suitable technology is at hand.

It has not been that way throughout history. The first steps toward modern semiconductor technology were done with Germanium. Due essentially to some disadvantages resulting mostly from its smaller bandgap (see, e.g. Sect. 9.2), Si took over once it was available in sufficient quality.

The key issue regarding quality was purity. “Purity” in the context of semiconductor technology has a meaning not to be met in other branches of materials science and technology, a point a semiconductor technologist has to keep in mind when she communicates with chemists, metallurgists, etc. In the field of semiconductor technology, we speak of impurity levels far below parts per million, rather than parts per billion or below.

Another factor, apart from purity, in favor of Si is its availability. Behind oxygen, Si is the most common element in the earth’s crust (20.4 atom percent), and absolutely no poisonous properties are known.

5.2 The Siemens Process

Some history: It is not commonly known that the early history of semiconductor technology is connected with Erlangen, a University town near Nürnberg, and the site of a large part of the corporate research of the Siemens company, not to forget the little village of Pretzfeld nearby. The laboratory in Pretzfeld (Fig. 5.1a) gave birth to the Siemens process, still the standard process for the purification of Si, and was home of important contributions to semiconductor technology.

The Siemens process is connected with the names of Walter Schottky (Fig. 5.1b) and Eberhard Spenke (Fig. 5.1c) (Courtesy Siemens Historical Institute)

When the second world war approached its end, these two, together with many other Siemens scientists, left the permanently bombed city of Berlin and finally arrived in the small village of Pretzfeld near Erlangen, where they started a small laboratory in the old castle of the place to continue their pioneering work on semiconductors.

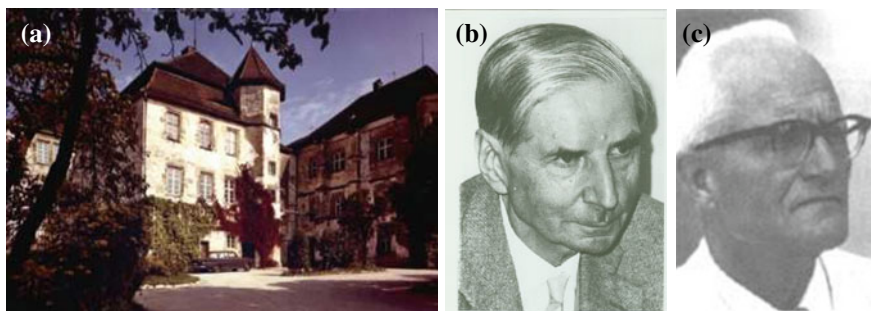


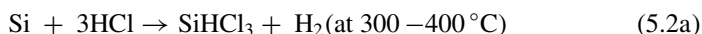
Fig. 5.1 a The castle in the little village of Pretzfeld. b Walter Schottky. c Eberhard Spenke

Purifying Si Via the Siemens Process

The most common source of Si is quartz sand, SiO_2 , the white stuff covering so many beautiful beaches of the oceans. From that sand, or from quarry material, Si is produced via reduction with coke in the electric arc furnace by the reaction



This Si, called “metallurgical grade Si” (purity 98–99%), is widely used in technology (construction work, ceramics, cement, and fillers). For electronic purposes, it has to be purified. This is done in the so-called Siemens Process, which consists essentially of the following procedure: First Si is converted into the compound trichlorosilane (SiHCl_3) by the following reaction:



SiHCl_3 is a liquid with a boiling temperature slightly above room temperature (at 31.8°C), which can then be purified with moderate energy consumption by distillation. Subsequently, the purified SiHCl_3 has to be reconverted into Si.

This is done by letting reaction (5.2a) run in the reverse direction. Reaction (5.2a) is highly exothermic, so—according to Le Chatelier’s principle—it can be reversed by going to a high temperature. So, at $1000\text{--}1100^\circ\text{C}$, reaction (5.2a) runs reversely:



H_2 is added to the liquid SiHCl_3 in a bubbler. The mixture of gas and droplets then enters the “glass dome”, a huge glass dome vessel of 2–3 m in dimensions (Fig. 5.2a).

Here, it meets the rods of polycrystalline Si shown in green in the figure. The rods are heated to $1000\text{--}1100^\circ\text{C}$ electrically (see “power supply” in Fig. 5.2a). The reverse reaction (Eq. 5.2b) only happens at high temperatures, so Si is deposited only on those rods, not anywhere else in the glass dome. The question arises as to the chemical nature of the rods, they should not introduce new impurities into the material. The point is that the rods are made of polycrystalline silicon which has been highly purified by a float zone process (see Sect. 5.4). In this way, an incorporation of impurities is avoided which would result from the usage of a different material for the rods at this high temperature. Remember that high temperature is the enemy of purity!

The originally slim rods of about 1 cm diameter grow to a thickness of 15 to 18 cm before they are taken out of the dome. They have a purity of up to $1:10^{10}$ and form the basic material for the subsequent crystal growth process or the float zone process. The crystal growth process provides the single crystals required for technology.

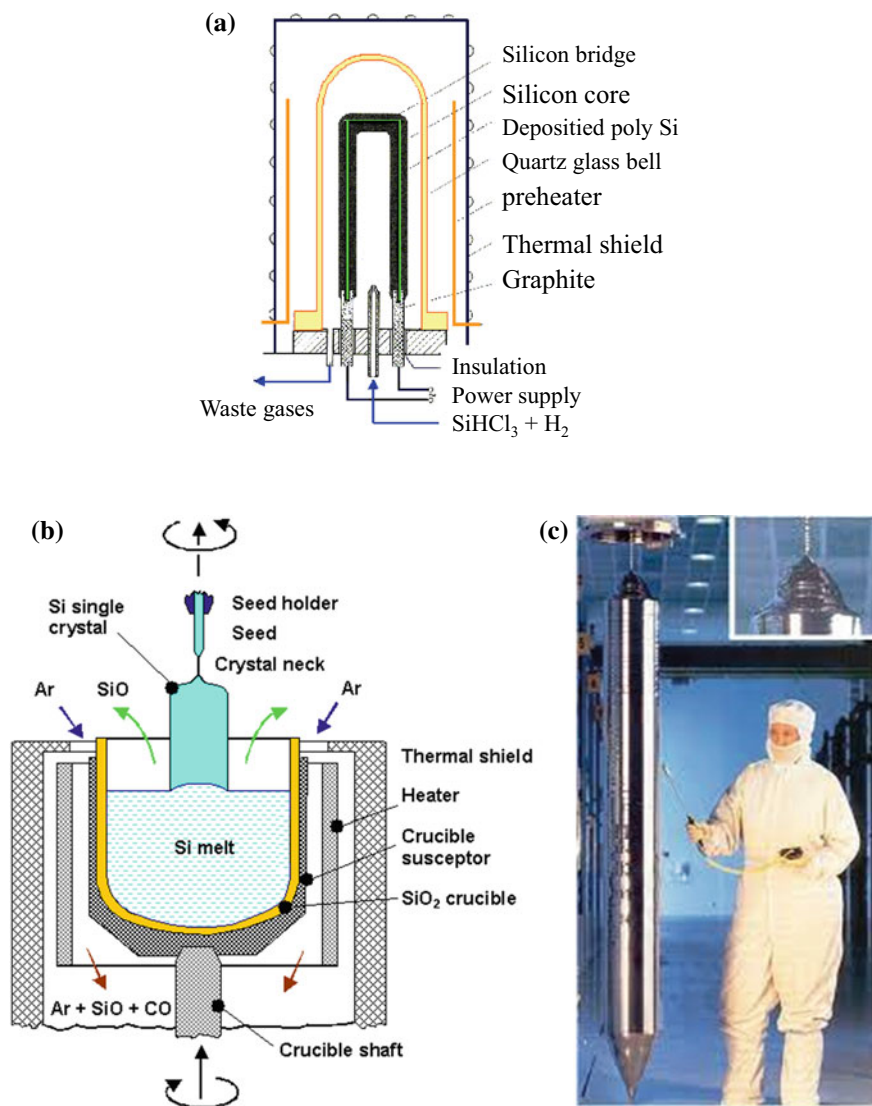


Fig. 5.2 **a** Siemens process reactor (from: <https://www.tf.uni-kiel.de> Hyperscripts AMAT Electronic Materials, courtesy by Helmut Foell). **b** Czochralski process for the growth of Si. **c** Si state-of-the-art single crystal (**b** and **c** from (from <https://www.tf.uni-kiel.de> Hyperscripts AMAT Electronic Materials, courtesy by Helmut Foell))

5.3 Crystal Growth of Silicon

Two growth processes are in use: Czochralski growth and float zone growth.

Czochralski Growth

The principle setup is shown in Fig. 5.2b.

In the Czochralski process, the crystal is grown from the melt. Si melts at 1414 °C. The melt is contained in a quartz crucible. Because quartz starts to soften at the Si melting temperature, the crucible is supported by a graphite susceptor. To start the process, a Si seed crystal is immersed into the melt and slowly pulled out. By the speed of the pulling process, the diameter of the crystal is controlled. The present state of the art is a diameter of 40 cm, and crystals longer than 2 m can be grown (Fig. 5.2c). In fabrication, a diameter of about 30 cm is achieved, and in the laboratory about 40 cm.¹

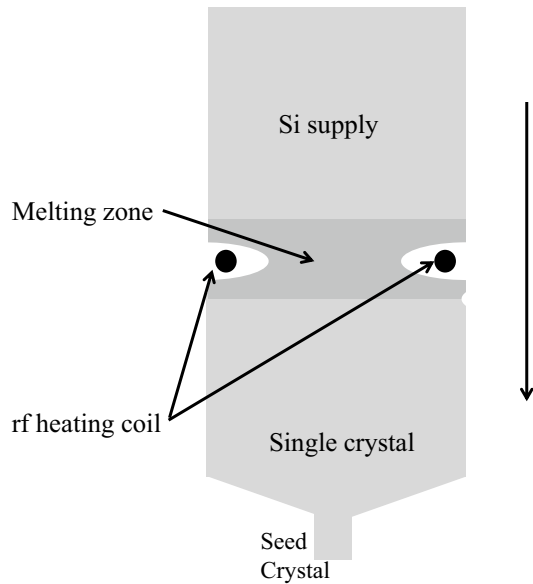
In crystal growth, generally the choice of the crucible material provides a major challenge. It must withstand the melt temperature, of course, and from the crucible material impurities will be transported into the growing crystal. In the case of Si quartz is chosen as crucible material. It can be provided at high purity, so the only impurity heavily incorporated into the growing Si-crystal is oxygen. The oxygen is dispersed and may act as a donor, making the material n-type, and act as recombination center for charge carriers. The situation can be improved by thermal annealing steps, resulting in precipitation of the oxygen, but nevertheless, certain devices don't tolerate oxygen. For those float zone, Si is used (later in this section). A few remarks concerning O in Si are given in Sect. 5.6.

A source of inhomogeneity of the growing crystal is convection in the melt. The hottest region in the melt is in the lower part of the crucible, in particular near the bottom, close to the heating system (see Fig. 5.2b). The coldest part is at the top, where the melt crystallizes and the crystal is formed. Consequently, the melt has the lowest specific weight near the bottom and the lower part of the side walls, and the highest specific weight at the top. So a convection sets in providing unstable conditions including temperature variations at the growth front of the crystal. These instabilities result, e.g. in variations of the doping concentration.

A way to suppress this convection is by putting the crystal growth setup into a magnetic field. The Si melt is metallic, so the convection will be slowed down according to the principle of the Eddy current brake. At the present state of Si growth (crystal diameter 30–40 cm, crystal length close to 2 m), this would require a gigantic magnet and is not economic on a production scale. Interesting experiments are being performed at present on a laboratory scale by using the magnetic field provided by the heating current system of the crystal growth apparatus.² Whether this will be

¹ 40 cm diameter has been under development for many years now without entering fabrication. It seems that the geometry (bow, warp) at this large dimensions is at least part of the problem.

² See, e.g. the use of heater-magnet module for Czochralski growth of PV silicon crystals with quadratic cross section, P. Rudolph, M. Czupalla, B. Lux, F. Kirscht, C. Frank-Rotsch, W. Müller, M. Albrecht, J. Crystal Growth 318(1) 249–254.

Fig. 5.3 Float zone process

used on a larger scale is an open question, but an elegant idea it is anyway. It will be shortly mentioned in the following Sect. 5.3 that suppression of convection is also an argument for doing Czochralski growth experiments under microgravity.

Float Zone Process

This setup is shown in Fig. 5.3.

The crystal is heated in a narrow zone via a rf coil and then pulled slowly through the coil opening. It is remarkable that the diameter of the coil can be smaller than the one of the crystal due to the concave shape of the liquid's surface. Another special feature of the float zone process of Si is the fact that no crucible is required. Due to the high surface tension of Si, the liquid Si stays in the melt zone without flowing downwards. This is of course a high advantage in terms of material purity, and it is a fairly unique feature of Si.

Crystal Growth in Microgravity

As a motivation for running a space station circulating our planet, we often hear about crystal growth experiments in "microgravity". One of the reasons for such experiments is that a crucible-free float zone process can be applied to many other materials under these conditions. Let us remark at this point that microgravity conditions can also be useful for the Czochralski process because the convection due to temperature fluctuations is suppressed. It should be noted expressively that even in the complete absence of gravity, convection would not be completely suppressed due to another phenomenon, the "Marangoni convection" which occurs as a result of surface tension. It will not be discussed here.

5.4 The Float Zone Process as a Purification Method

As mentioned above, purity is promoted, as compared to Czochralski growth, in the float zone process for Si by the fact that Si can be grown crucible free. Moreover, it has to be pointed out that the float zone process is by itself a purification process.

In a solidification process, when a material makes a transition from the liquid to the crystalline state, usually impurities accumulate in the liquid phase, with an often much lower concentration in the solid. The effect is characterized by the “segregation coefficient”

$$K = c_{\text{solid}}/c_{\text{liquid}} \quad (5.3)$$

The segregation coefficient is for many impurity atoms much smaller than one (it is often in the region of 10^{-2} to 10^{-6}). When the float zone moves along the axis of the setup (Fig. 5.3), the material recrystallizes at the back front, the impurity content in the recrystallized material being strongly reduced. On the contrary, the impurities will accumulate at the front edge, where the material is melting. So in the end, one will obtain a rod of material with most of the impurities concentrated at the upper end, which then can be cut off. The process can be repeated a few times in order to further improve the purity.

Having just purification in mind as the objective of the process, one generally speaks of the method of “zone refining” instead of “float zone”. In this case, of course, no seed crystal is required at the lower end of the rod.

Some history: Actually the zone refining process was originally invented as a purification method, only later use as a method of crystal growth was commonly installed. The float zone process was invented by William Gardner Pfann at the Bell Laboratories in 1950/51. It constituted a major progress in semiconductor technology at the time because semiconductor materials like Si and others subsequently could be produced at sufficient purity for the first time.

5.5 Wafering

After crystal growth, “wafers” have to be fabricated out of the crystal. Actually, this process of “wafering” is at least as important a step in the value chain as the crystal growth itself. It consists of several steps as follows:

1. Cylindrical Grinding: The Crystal is Ground to a Precisely Equal Diameter Everywhere.
2. A little mark is set to the outer diameter of the crystal to mark the crystallographic orientation. When Si-crystal diameters were smaller than 200 mm this was a “flat”, since then it is a little “notch” (Fig. 5.4a).
3. Sawing: This Was Traditionally Done with an Annular Saw (Fig. 5.4b).

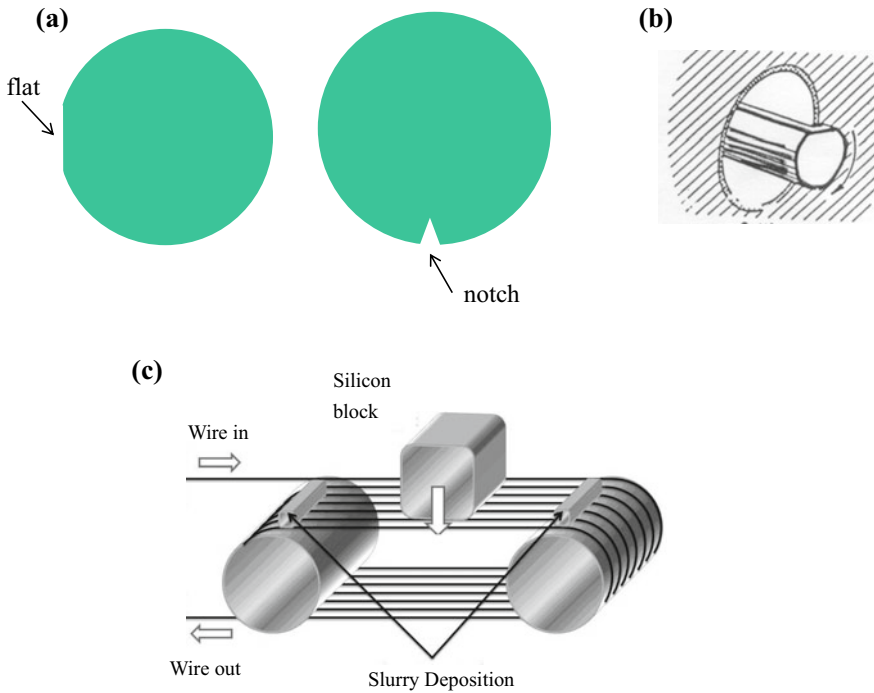


Fig. 5.4 **a** A “flat” or a “notch” marking the crystallographic orientation. **b** Annular saw. **c** Multi-wire saw (After: Slurry reprocessing options: W. Scott Radeker, Scott W. Cunningham, TMS Conference papers, Febr. 14-18 2010, Seattle). The wires are covered with diamond nanoparticles to provide sufficient hardness

The advantage of an annular saw as compared to the traditional circular saw is that the flutter of the saw blade is much quieter at the inner circumference than at the outer circumference. Nowadays, for the present large diameters, multiwire saws are used in fabrication lines (Fig. 5.4c). In the figure, not a Si-crystal is sawed, but a block, as, e.g. used for solar cell production (Sect. 10.6).

With those multiwire saws up to 100 wafers can be sawed at a time. A major problem in the development of the wire saw process was that, once a wire broke, the whole crystal had to be thrown away. This problem is overcome now.

After sawing, about 50 μm of material is lapped away in order to remove the sawing damage from the surface. The thickness of wafers then was typically 725 μm for 200 mm wafers and is typically 775 μm now for 300 mm wafers.

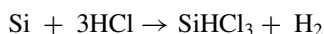
4. A final and very decisive step in the process is polishing. Nowadays a flatness of about 3 nm or even less is achieved in 300 mm wafers. An incredible achievement required in actual semiconductor technology! Let's just scale this up by a factor of 300. Then a disk of 10 m in diameter would have a flatness of about 1 μm ! The main parameters describing the geometrical perfection are “bow” and “warp” for the overall shape of the wafer.

5.6 Electronic Silicon, a Material of Extreme Purity and Crystalline Perfection

At the end of this chapter, we want to point out a few of the specifics of the fabrication process which make Si the purest and most perfect material in modern technology.

Impurities in Si can be harmful in different ways. They may act as shallow donors or acceptors and in this way give rise to an uncontrolled conductivity. Or they may give rise to energy levels deep in the bandgap and thus affect the recombination time of carriers, and they act as “lifetime killers”. And, last but not least, they might form precipitates and, in this way, affect crystalline quality. In the course of the various technology steps discussed in this chapter, impurities are eliminated to a degree often less than 0,1 ppb, resulting in a material of absolutely unique purity and quality. Let us shortly summarize the major purification steps.

The biggest contribution arises from the transformation of metallurgical Si into trichlorosilane SiHCl_3 via the reaction (Eq. 5.2a)



which subsequently is purified by (multiple) distillations.

Without going into details, let us remark that the biggest merit of that process consisted in the elimination of the “Boron problem”. In the processes used before, Boron had always remained as a detrimental impurity (B acts as a shallow acceptor), while in the “Siemens process”, B is carried away with the gas stream. Many other impurities are carried away by the gas stream as volatile chlorides.

When semiconductor crystals are pulled from the melt, a critical source of impurities is always the crucible. This is particularly true for Si which as a melt reacts with every conceivable crucible material. SiO_2 (silica) turns out to be the most suitable material. For the reason of purity, synthetic silica is used for that purpose. The main impurity then introduced into the Si melt is Oxygen in form of SiO . It is of great help in that context that most of that SiO , more than 99%, evaporates at the free surface of the melt.

One widely used way of dealing with the remaining oxygen consists in the creation of a “denuded zone” by out-diffusion of the oxygen out of the surface-near region of the wafer. Because the active region of a device or an integrated circuit is located very close to the surface, the rest of the wafer acts essentially as a substrate, so the influence of oxygen is eliminated.

Another interesting aspect is that the presence of oxygen might be turned into an advantage. Oxygen bubbles might get impurities and thus lower the effective impurity content of the material.

It was pointed out in Sect. 5.4 that solidification is in most cases accompanied by a purification process because of “segregation”. Most impurities accumulate in the melt, resulting in a purification of the crystallized material described by the “segregation coefficient”. The purification effect in zone refining is based on this fact. It is noteworthy that this segregation effect also plays a role in the Czochralski

growth. The impurity concentration in the melt is much higher than it is in the growing crystal, so also the Czochralski growth, like the float zone crystal growth process, is accompanied by purification of the material.

As a striking example, we consider Fe as an impurity. Fe acts as a “lifetime killer”. The segregation coefficient of Fe in Si is (see Eq. 5.3) $\kappa = c_{\text{solid}}/c_{\text{liquid}} \approx 8 \times 10^{-6}$. This means that only 1 out of about 125,000 Fe atoms in the melt will end up in the growing crystal.

It was mentioned in Sect. 5.4 that κ is generally very small, in most cases between 10^{-2} and 10^{-6} , important exceptions being B ($\kappa = 0,8$) and oxygen ($\kappa = 1,2$). These cases have been discussed above.

Another way of saying that the segregation coefficient of most elements is very small is to state that their solubility in crystalline Si is small. This is mainly a matter of the atomic radius, the atomic radius of the atoms under consideration being much larger than the one of Si. So those atoms don’t “fit” into the Si lattice.

The last point to be discussed when talking about the extraordinary quality of Si single crystals is the fact **that Si can be grown dislocation free**. This is a remarkable achievement indeed. To give a comparison to other crystals grown from the melt, single crystals of GaAs or InP have typically a few thousand dislocations per cm^2 . Remember that dislocations are line defects (Sect. 6.3), so a few thousands of dislocation lines cross an area of 1cm^2 in these crystals. In single crystals of electronic Si, this number is zero. This achievement is of high technological significance. As discussed in Sect. 6.4, dislocations give rise to local variations of conductivity and carrier lifetime. Integrated circuits, however, require a large electrical homogeneity of electronic properties all over the wafer, to make sure that devices have the same properties everywhere on the chip. **Only because of this lack of dislocations, Si wafers are suitable for highly integrated circuits in microelectronics.**

The growth of dislocation-free Si-crystals by the Czochralski process is achieved by the proper “seeding technique” developed by W.C. Dash³ in 1958. After immersing the seed crystal of about 6–8 mm diameter into the melt, he first increased the pulling speed such that the diameter of the crystal gets as low as about 3 mm and kept this small diameter for a few centimeter lengths of the crystal. Under these conditions, the growth rate is so fast that dislocations cannot follow the propagating growth interface. They are diverted to the lateral crystal surface. After a few centimeters, the crystal is dislocation free and the crystal diameter can continuously be increased by reducing the pulling speed of the seed, up to the desired value. The zone of the smallest diameter, a few cm long, called the “seed neck”, is seen in Fig. 5.2c. It is very amazing, indeed, that this thin seed neck can carry the grown crystal with its 150 kg or more, demonstrating the extraordinary mechanical properties characteristic of Si, apart from its outstanding electronic properties which make it the dominant electronic material.

³ W.C. Dash: J. Appl. Phys. 29 (1958) 736.

5.7 Questions and Answers

- Question 1 : What is the basic procedure in the purification of metallurgical Silicon?
- Question 2 : Which are the two main crystal growth processes for Si?
- Question 3 : What is the main impurity in the crystals grown by the Czochralski process?
- Question 4 : Why do Si-crystals from the float zone process have a higher purity as compared to the Czochralski crystals?
- Question 5 : Which kind of saw is used in wafer production?
- Question 6 : Which method provided the breakthrough in the purification of semiconductor materials?
- Question 7 : What is the “segregation coefficient”?
- Question 8 : Which relevant impurities in Si have a segregation coefficient so close to 1 such that zone refining is not applicable for those?
- Question 9 : Why is B a relevant impurity in Si, having provided a major problem in the development of Si-technology?

Answers

- Answer 1 : Metallurgical Si is converted into trichlorosilane (SiHCl_3). Subsequently, trichlorosilane is purified by distillation. Finally, the purified trichlorosilane is reconverted into Si.
- Answer 2 : Czochralski growth and float zone process.
- Answer 3 : Oxygen, from the SiO_2 crucible.
- Answer 4 : Because they are grown without crucible, and because segregation of impurities.
- Answer 5 : A multiwire saw in state-of-the-art fabrication.
- Answer 6 : The zone refining process, invented by William Gardner Pfann.
- Answer 7 : The ratio of the concentration of an impurity in the solid to concentration in the melt (Eq. 5.3) in thermal equilibrium.
- Answer 8 : Oxygen and B.
- Answer 9 : Because it is “electrically active”. B acts as a shallow acceptor (see Chap. 3).

Chapter 6

Lattice Defects



Abstract Lattice defects are deviations from the strictly periodic arrangement of atoms in a solid. Those could be related to impurities. In this case, we speak of extrinsic defects. Or they may just represent a disarrangement of the regular atoms of the solid, those are “intrinsic defects”. Defects often determine the essential electronic properties of a semiconductor. This chapter is dealing with intrinsic defects. They are often categorized according to their dimensionality. In that way, one distinguishes between point defects like vacancies and interstitials, line defects like edge dislocations and screw dislocations, two-dimensional defects like stacking faults, and three-dimensional defects like precipitates and microtubes. Point defects play an important role in diffusion processes. Dislocations may give rise to an inhomogeneous distribution of donors or acceptors, and they may act as recombination centers. Screw dislocations play a role in crystal growth, resulting in growth spirals. Stacking faults can give rise to the growth of twin crystals. Famous examples of three-dimensional defects are the microtubes in SiC or the As precipitates in semiinsulating GaAs. An important fact to realize is that there is no such thing like a perfect crystal. Due to the entropy term in the thermodynamic potentials, intrinsic lattice defects are inevitable, they must always be present in thermodynamic equilibrium.

6.1 Intrinsic and Extrinsic Defects

Lattice defects, in a way, are the essence of semiconductor material. In a positive sense, they can be used to control a wide spectrum of electronic functions, like conductivity, p- or n-type character, mobility, and charge carrier recombination. On the other hand, as undesired perturbations of the ideal lattice, they may impair desired and useful functions. In many cases, already very small deviations from the ideal crystal may significantly perturb the electronic properties of a semiconductor material. As a matter of fact, essentially up to the 50ties of the last century, it was impossible to fabricate useful semiconductor materials with controlled properties due to undesired and uncontrolled impurities and imperfections. In those days, Wolfgang Pauli, one of the pioneers of quantum mechanics, had good reasons for saying: “Man

Cardona(1)

sollte sich mit Halbleitern nicht beschaeftigen, das sind Dreckeffekte”. In a way, Pauli was correct by declaring that most of the effects we see are due to defects (if not “dirt effects”), but he was certainly wrong by claiming we should not bother! Only after defects could be either eliminated or controlled by improvement of crystal quality and purity, a semiconductor technology could develop.

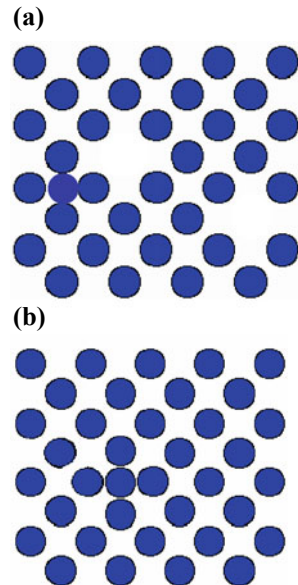
It was discussed in Chap. 1 that a crystal is defined by a periodic scheme of points, i.e. the “lattice points”, and a group of atoms associated with each of these points, called the “basis”. Crystals, however, never are perfect, they contain “lattice defects”. Basically, any perturbation of the strict periodicity of the lattice may be called a lattice defect. Such a lattice defect may e.g. be an impurity atom, in this case, we speak of an “extrinsic” defect. It may also be an “intrinsic” defect, i.e. just a deviation from the perfect arrangement of the regular atoms. Lattice defects affect in many ways the electronic properties of crystals. Under this aspect, they will be dealt with in this chapter.

Defects are often categorized according to their “dimensionality”, so there are pointlike defects (often just called point defects), one-dimensional, two-dimensional, and three-dimensional defects. In these terms, they will be discussed in the following.

6.2 Point Defects

Figure 6.1a, b show the two types of point defects: Vacancies (Fig. 6.1a) and interstitials (Fig. 6.1b)

Fig. 6.1 **a** Vacancies: Frenkel defect and Schottky defect. **b** Interstitial



A vacancy just results from a missing atom on a regular lattice site, an interstitial is an additional atom between regular lattice sites. There are two species of vacancies: The Frenkel defect, which is a complex of two defects: The vacancy and the nearby atom whose removal has created the vacancy. The other one is the plain vacancy, called “Schottky defect”, left over when the “missing” atom has disappeared by diffusion to the surface.

Figure 6.1b shows an “interstitial”. An interstitial is an atom not sitting at a regular lattice site, but somewhere in between.

Vacancy diffusion and interstitial diffusion

Vacancies and interstitials play an important role in diffusion. Figure 6.2a shows the case where an impurity jumps from a substitutional site to the next one “assisted” by the presence of a neighboring vacancy. Figure 6.2b shows the “kick-out process” where diffusion occurs via repeated exchange of substitutional and interstitial sites of impurity- and host-lattice atom.

Diffusion is an important process in semiconductor technology, either as a useful and intended one when it is used to introduce dopants into the semiconductor (see

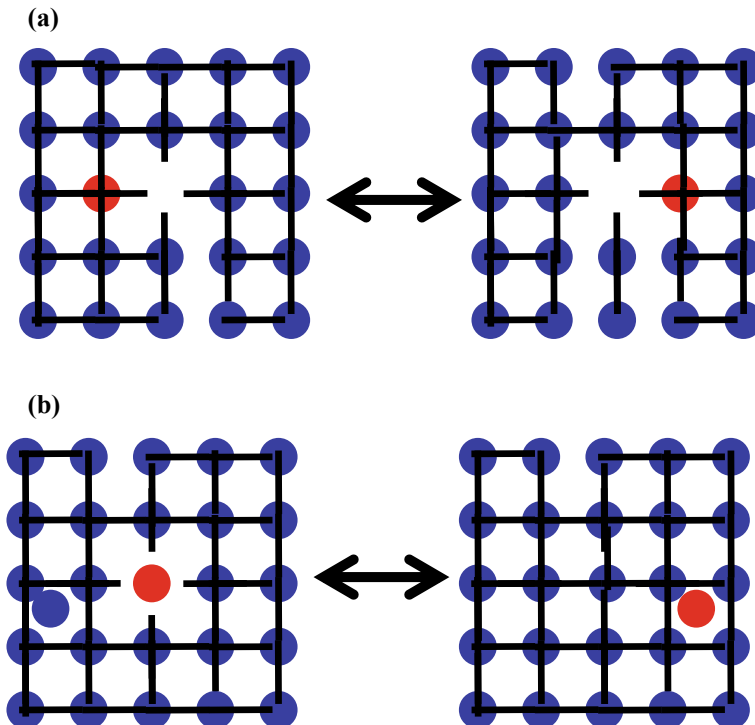


Fig. 6.2 a Vacancy assisted diffusion. b Kick-out mechanism

Chap. 3), or as a detrimental process, which incorporates undesired impurities into the material with undesired consequences.

The speed of diffusion, according to Fick's first law of diffusion, is given by

$$\vec{J}(\vec{r}) = -D(T) \text{grad}n(\vec{r}) \quad (6.1)$$

$\vec{J}$ is the particle current in unit “number of particles per sec and area”, and n is the concentration of the diffusing particles. So the diffusion current always follows the gradient of the concentration, going from regions of higher to regions of lower concentration, with the diffusion constant $D(T)$ being the factor of proportionality between $\vec{J}$ and the gradient. By looking at the processes shown in Fig. 6.2a, b (vacancy diffusion and interstitial diffusion), it is plausible that the atom, by hopping from one site to the next, has to overcome an energy barrier, because, in a way, it has to “squeeze” through the arrangement of the neighboring atoms.

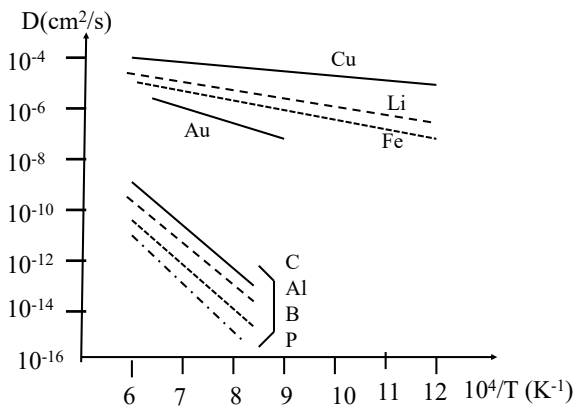
If the height of this energy barrier is E_a (E_a stands for “activation energy” of the diffusion process), the temperature dependence of $D(T)$, according to Arrhenius' law, is given by

$$D(T) = D_0 \exp(-E_a/kT) \quad (6.2)$$

where D_0 is a constant of proportionality depending on material and doping species. Because of the exponential factor, the speed of diffusion increases strongly with temperature. In case of an undesired diffusion of harmful impurities, this gives rise to the saying: “Temperature is the enemy of semiconductor technology.” High-temperature process steps tend to be harmful to purity!

Figure 6.3 shows a plot of $\log D(T)$ versus $1/T$, a so-called Arrhenius plot, for a number of impurities in Silicon. If the Arrhenius law is applied, the plot should show a linear dependence of $\log D(T)$ with a slope of $(1/\ln 10)E_a/k^1$.

Fig. 6.3 Diffusion constants as a function of reciprocal temperature for different atoms in Si



¹ Goesele(2)

Figure 6.3 clearly shows that the diffusion constants of the different atoms all show an Arrhenius behavior, but there are two clearly distinct cases, the case of “slow diffusors” in the lower part of the figure, and “fast diffusors” in the upper part of the figure. The fast diffusors such as Cu, Li, and Fe diffuse by jumping from interstitial to interstitial, a process requiring no breaking of bonds and thus only have low activation energies around 1 eV. The slow diffusors such as Al, B, P (and other members of the IIIrd or Vth group widely used as dopants in Si) as well as C show vacancy diffusion (Fig. 6.2a) with high activation energies of roughly 3–4 eV. The fast diffusor such as Au is a special case involving the kick-out mechanism shown in Fig. 6.2b. Because of their extremely high diffusion coefficients, Cu and Au (and also Ag) easily enter Si under fabrication process conditions. Special precautions have to be taken to keep them out because they strongly act as electron–hole recombination centers affecting the carrier lifetimes and thus the performance of electronic devices. Nevertheless, Cu is nowadays used as interconnect in integrated Si circuits by making use of a diffusion barrier as described in Sect. 14.2.

6.3 Line Defects (Dislocations)

Figure 6.4a shows a so-called edge dislocation.

In an edge dislocation a lattice plane ends, so evidently, a whole line of lattice points (symbolized by L in the figure) is perturbed, this corresponds to the notation as a “line defect”. Figure 6.4b shows the second species of dislocations, the “screw dislocation”.

Dislocations are the result of a “slipping process”, developing from internal stress. Figure 6.4a illustrates this for the case of the edge dislocation, where the dislocation is formed as a result of a “lattice mismatch”. The figure shows a semiconductor material with different lattice constants in the upper and the lower part. This may, e.g., be a consequence of a temperature gradient in a material. Due to thermal expansion, the lattice constants will be different in areas of different temperatures. So in the transition region between the two regions of different temperatures, the lattice constants have

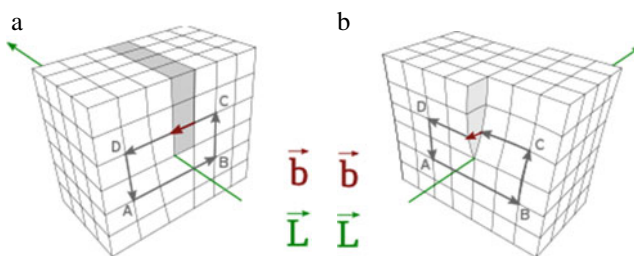


Fig. 6.4 a Edge dislocation. b Screw dislocation (both from Wikipedia)

to adapt to each other, some lattice plains of the area with lower lattice constant (the cooler one) have to end there.

A contact between regions of different lattice constants arises in semiconductor technology when layers of different materials with slightly different lattice constants are grown one on top of the other in the so-called epitaxial processes (see Sect. 7.5 Epitaxy). In the interface region of the two materials, stress develops because the lattice constants tend to adapt to each other. The lattice will partly give in to that stress, so, as shown in the figure, some of the lattice plains of the material with a smaller lattice constant will have to end in the transition region, and an edge dislocation forms. Here the “slipping process” evidently occurs in the plain between larger and smaller lattice constants, its size and direction being described by the “Burger’s vector” $\mathbf{b}$.

As seen in Fig. 6.4b, screw dislocations are also the result of a slipping process. The “Burger’s vector” $\mathbf{b}$ gives the size and the direction of that slipping process. Figures 6.4a, b show that for the edge dislocation, the burger’s vector is vertical to the dislocation line, whereas for the screw dislocation, it is parallel to the dislocation line.

6.4 The Role of Dislocations in Electronic Materials

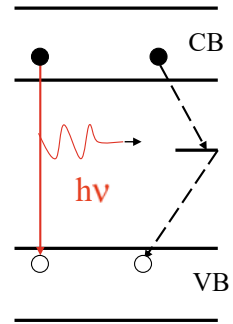
Dislocations play a very important role in the mechanical properties of materials like elasticity and strength. This is not the issue in our context. However, they are also of great relevance for electronic properties for the following reasons:

- (a) They attract impurities, in a way they act as a “vacuum cleaner” in the crystal, providing an energetically favorable site for impurity atoms. If those impurities are electrically active and act as donors or acceptors, there shall be a change in electrical conductivity around the dislocation, so an electrical inhomogeneity of the crystal shall result. Such an inhomogeneity, by the way, can arise just from the stress field around the dislocation, because electrical parameters of the material are also stress dependent.

This inhomogeneity could adversely affect, e.g. the performance of integrated circuits. Integrated electronic circuits and their components, intended to perform identically everywhere on a wafer, will show different electronic properties, depending on the location on the wafer where they are fabricated.

- (b) Being a structural defect in a crystal dislocations perturb the translational symmetry of the crystal. As discussed in Chap. 2, “everything which perturbs the periodicity (i.e. the “translational symmetry”) of the crystal will also reduce the mobility of the charge carriers. This may often be an undesired effect in the semiconductor material, given the importance of mobility in, e.g. high-frequency devices.
- (c) Dislocations may act as recombination centers for electrons and holes, as shown in Fig. 6.5.

Fig. 6.5 Carrier recombination via a deep state in the bandgap



The dislocation may give rise to an energy level deep in the bandgap. This level will open up a second recombination path, in addition to the direct recombination shown in the left part of the figure. In other words: It shortens the carrier's lifetime. Charge carrier lifetime is an important parameter in various electronic devices, so their electronic characteristics may be affected. Dislocations may be particularly harmful in optoelectronic devices. The light-producing process in light-emitting diodes and semiconductor lasers is shown on the left side of the figure: The energy released by recombination is transformed into a photon (Sect. 11.1). So a photon is emitted as a result of a “radiative” recombination process. The dislocation, however, will open a competing recombination path, as shown in the figure. In this process, usually, the energy is released via phonons (resulting in a “nonradiative process”, i.e. in local heat production), so this recombination process is lost for the production of light.

Screw dislocations and crystal growth

Screw dislocations may play an important role in crystal growth. If a crystal with an atomically smooth surface is expected to grow out of a liquid or gaseous phase, the atoms have to stick on a plain surface. In many cases, the probability for this to happen is small because the binding force of an atom to such a plain is much weaker than the one in the bulk of the crystal: An atom in the interior of a solid experiences a binding force in all three spatial directions. On the surface, however, the binding force acts only in one direction, i.e. toward the surface. The situation is different, however, at a screw dislocation as seen in Fig. 6.6a.

The screw dislocation forms a step such that an atom is bound by forces in two directions resulting in a higher sticking probability as compared to an atom on the plain surface. As a result of that growth process along the steps of a screw dislocation, the growth spirals form as seen in Fig. 6.6b.

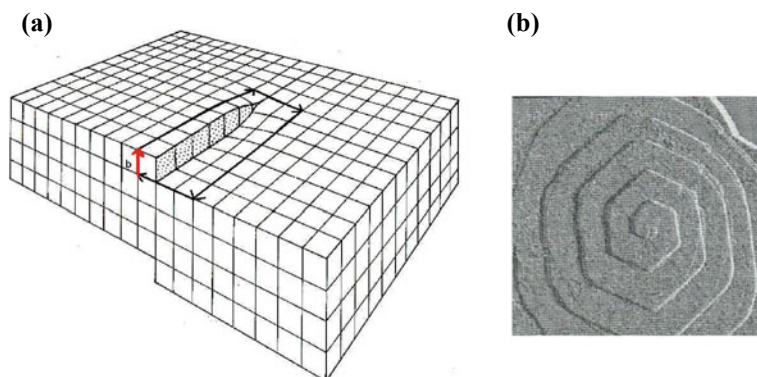
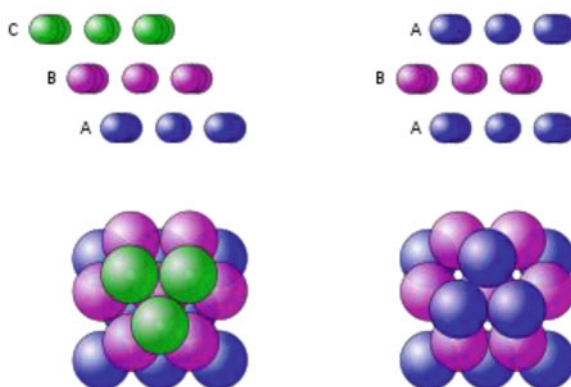


Fig. 6.6 (a) Atoms sticking on a plain surface (right) and at a screw dislocation. (b) Growth spirals. The growth direction is vertical on the plain seen in the figure. From: B. E. Epelbaum, D. Hofmann, M. Müller, A. Winnacker: Hollow defect elimination during solution growth of SiC <https://doi.org/10.1557/PROC-640-H2.2>

6.5 Two-Dimensional Defects

A commonly appearing two-dimensional defect in crystals is the stacking fault. A stacking fault arises when a whole plain is incorrectly incorporated within a crystal. A well-known example is seen in Fig. 6.7, which shows the two possible arrangements of plains in a closely packed crystal.

If the third layer is positioned right above the first one (i.e. sequence ABABAB), a hexagonal crystal is formed (see right side of the figure). If, however, only the fourth layer is positioned above the first one (ABCABCA...), a face-centered cubic lattice is formed (left side of the figure). Because the two lattice types only differ in the



Left: Cubic densely sphere packing
Sequence ABCABCA.....

Right: Hexagonal densely sphere packing
Sequence ABABAB.....

Fig. 6.7 Sequence of plains in closely packed crystals. From: (https://de.wikipedia.org/wiki/Dichteste_Kugelpackung)

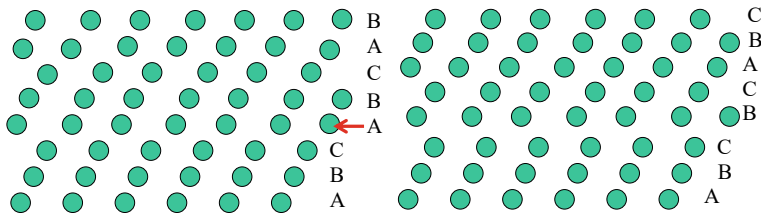


Fig. 6.8 Formation of a stacking fault

second neighboring layer, there is only a small difference in binding energy between the two lattices, and a so-called stacking fault can easily develop by a switch in the lattice sequence as shown in Fig. 6.8.

Another case of misplaced planes naturally occurs at grain boundaries in polycrystalline samples when grains of different orientations touch each other. So grain boundaries are a very common example of two-dimensional defects.

Stacking faults or grain boundaries can affect electronic properties in the same way as dislocations. They can, e.g. reduce the mobility, act as recombination centers or attract impurities. Donors or acceptors can accumulate at grain boundaries and stacking faults.

Stacking faults can also adversely affect crystal growth. At a stacking fault, twin growth can develop.

6.6 Three-Dimensional Defects

Examples of three-dimensional defects in crystals are inclusions, precipitates, voids, and microtubes.

Figure 6.9a shows a C-inclusion in 4H-SiC.

The development of such a carbon inclusion arises as a result of “carbonization” during the high-temperature growth process of SiC (Sect. 7.4, context of Fig. 7.7a). Some SiC is decomposed into liquid Si and solid C.

Another well-known three-dimensional defect in the context of semiconductor growth is As precipitates in GaAs. As discussed in Chap. 4.5, in order to grow semiinsulating GaAs, a small excess of As is added to the melt during crystal growth, a fraction of which ends up as As precipitate (the other fraction of the excess As, by the way, will form the EL2 Antisite defect). The art of growing semiinsulating GaAs consists of partly controlling the distribution of the excess As on these two “deposits”, precipitates and antisites.

As an example of gas-filled voids, let us mention the oxygen bubbles in Si due to the oxygen expelled from the SiO₂ crucible during Czochralski growth (see Chap. 5.6).

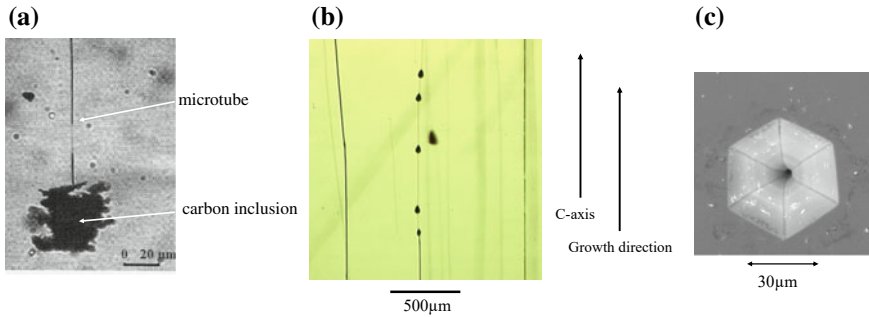


Fig. 6.9 (a) Carbon inclusion in 4H-SiC from D. Hofmann, E. Schmitt, M. Bickermann, M. Kölbl, P.J. Wellmann, A. Winnacker: *Materials Science and Engineering B* 61–62 (1999) 48–53. Additionally, a “micropipe” (see below) is seen. The stress field around the inclusion has given rise to a micropipe. (b) Microtube in 4H-SiC. Their formation might be related to the presence of the Si droplet seen also in the figure. (c) Scanning electron microscope picture of a microtube in 4H-SiC after KOH etching (b and c by kind permission of SiCrystal AG)

A kind of three-dimensional defects which played an important role in the development of SiC as a semiconductor, so-called microtubes, are shown in Fig. 6.9b, c.

These hollow tubes formed during the crystal growth of SiC in the PVT process. It was found that these microtubes arise from screw dislocations. When the Burger’s vector is sufficiently large, the stress around the dislocation line gets so strong that the dislocation line opens into a microtube.

The diameter of those holes typically is a few μm. We shall come back to SiC growth in Sect. 7.4.

Physics Comment: Lattice Defects Are Always Present in Thermal Equilibrium

At first sight, one might think that the ideal, perfect crystal represents the equilibrium state. After all, the reason to form a strictly crystalline structure is energy: The internal energy is lowest if each atom is positioned in its proper site in the perfectly periodic lattice. The point, however, is, that in thermodynamics the equilibrium state is not characterized by a minimum of internal energy. Nature, generally speaking, favors a certain degree of disorder, expressed by the entropy of the system. For instance, under conditions of a given temperature and volume, the thermal equilibrium is characterized by a minimum of the function

$$H = U - TS \quad (6.4)$$

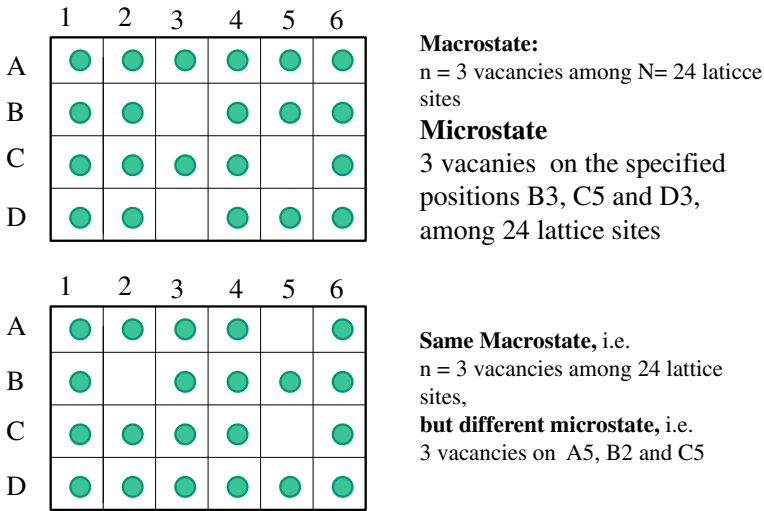


Fig. 6.10 On the definition of “macrostate” and “microstate”

where U is the internal energy, T is the absolute temperature, and S is the entropy of the system. The “state quantity” H is called the “Helmholtz free energy”. To illustrate the consequence of that law for the occurrence of defects in crystals, we shall discuss a simple, but representative example by asking:

Let us consider a crystal with N regular lattice sites and composed of identical atoms, each one associated with one lattice site, i.e. one atom per unit cell. How many vacancies would form in such a crystal at temperature T in thermal equilibrium?

The question is not only of interest in the context of defect center physics in the solid state, but also can give a most instructive insight into the principles of statistical mechanics as developed by Ludwig Boltzmann².

Let ϵ be the formation energy of a vacancy in the said crystal, i.e. the energy to remove an atom from its regular lattice site to the surface and away into free space. We are free to define the zero of U , so we ascribe it to the perfect crystal with zero vacancies. So, in the presence of n vacancies

$$U = n\epsilon \quad (6.5)$$

and thus

$$H(n) = \epsilon n - TS(n) \quad (6.6)$$

To find its minimum, i.e. the number $n_{\min}$ of vacancies in thermal equilibrium, the derivative of $H(n)$ has to be set to zero. To carry this through, we

have to find out: What is $S(n)$, the entropy of the crystal in the presence of n vacancies on N lattice sites? This brings us to some basic concepts of statistical mechanics.

According to Boltzmann, the entropy of a state is given by the equation $S = k \ln W$, where k is the constant nowadays named after Boltzmann, and W is the thermodynamic probability of that state. This “thermodynamic probability” is something very different from the notion of probability in day-to-day life, as seen already from the fact, that its maximum value is not one. The general definition of W as introduced by Boltzmann is:

The probability W of a state, called macrostate in the following, is the number of microstates resulting in that macrostate.

This sounds fairly mysterious and has to be clarified for our example. The “macrostate” in our example is defined by the description “ n vacancies among N lattice sites”. The microstate is defined by stating not only that we have n vacancies on N lattice sites, but also by specifying on which lattice sites the vacancies are situated. These definitions can best be clarified by a simple example as given in Fig. 6.10.

The upper and the lower graph show the same macrostate of a lattice consisting of 24 lattice sites, i.e. 3 vacancies on 24 lattice sites. However, the microstates are different. In the lower graph, the 3 vacancies are on sites different from the ones in the upper graph.

So, in order to find the probability of a macrostate with n vacancies on N lattice sites one has to find the number $W(n)$ of possibilities to distribute n vacancies on N lattice sites. Each of those represents a microstate. According to elementary combinatorics, this number is

$$W(n) = N! / (N - n)! n! \quad (6.7)$$

In order to calculate the first derivative of $W(n)$, we first transform (6.7) by making use of Stirling’s formula which perfectly applies to the huge numbers n and N

$$\ln(n!) = n \ln(n) \quad (6.8)$$

hence

$$\begin{aligned} \ln W(n) &= N \ln N - (N - n) \ln(N - n) - n \ln(n), \text{ so} \\ S(n) &= k[N \ln N - (N - n) \ln(N - n) - n \ln(n)] \end{aligned} \quad (6.9)$$

So we obtain the Helmholtz free energy (see 6.6)

$$H(n) = \varepsilon n - kT[N \ln N - (N - n) \ln(N - n) - n \ln(n)] \quad (6.10)$$

To find the minimum of $H(n)$ as a function of n , we put $dH(n)/dn$ equal to zero, which, by elementary calculation, results in

$$dH/dn = 0 = \varepsilon - kT[\ln(N - n) - \ln n] = \varepsilon - kT \ln\left(\frac{N}{n} - 1\right) \quad (6.11)$$

Hence, $kT \ln\left(\frac{N}{n} - 1\right) = \varepsilon$ and thus

$$n/N = 1/(1 + e^{\varepsilon/kT}) \quad (6.12)$$

A numerical example

In Germanium the formation energy of a vacancy is $\varepsilon = 2.1$ eV, its melting point is 1211 K. Remembering, as every physicist should, that kT is 0.026 eV at $T = 300$ K, we obtain for a temperature slightly below the melting point, say at 1200 K

$$\varepsilon/kT = (2.1/0.026)(300/1200) = 20.2, \text{ and thus (from eq. (6.12))}$$

$$n/N \approx 10^{-9}$$

As seen from this typical example, usually $e^{\varepsilon/kT} \gg 1$, so we might as well replace (6.12) by

$$n/N = \exp(-\varepsilon/kT) \quad (6.13)$$

Because at “realistic” temperatures $\exp(-\varepsilon/kT) \ll 1$, this means that the number of vacancies is many order of magnitude smaller than the number of lattice sites.

We now consider the situation at a lower temperature, e.g. $T = 300$ K, again taking Ge as an example.

$\varepsilon/kT = 2.1/0.026 = 81$, so $n/N = 10^{-35}$. This is a totally negligible number, given the fact that 1 Mol = 72.6 g of Ge contains $6 \cdot 10^{23}$ Ge atoms which means that even in a gigantic piece of Ge, there would be no vacancy at all in thermal equilibrium at room temperature.

However, this consideration is not realistic. The point is that in most cases, at temperatures considerably below the melting point, the probability of formation of a vacancy and the speed of diffusion are so strongly reduced that the vacancy concentration and distribution existing at temperatures near the melting point are “frozen in”.

This is an example of a situation encountered often in physics: The thermal equilibrium of a system and the question of whether it can realistically be achieved are two different matters.

² Ludwig Boltzmann (1844 – 1906), theoretician, founding father of statistical mechanics.

6.7 Questions and Answers

Question 1: Give examples of

- (a) Point defects
- (b) One-dimensional defects
- (c) Two-dimensional defects
- (d) Three-dimensional defects?

Question 2: In which way do dislocations affect electronic properties?

Question 3: What is the difference between extrinsic and intrinsic defects?

Question 4: Give the expression for the concentration of vacancies in a crystal in thermal equilibrium!

Question 5: Why is the concentration of lattice defects at room temperature usually orders of magnitude higher than the thermal equilibrium value?

Answer 1:

- (a) Vacancies and interstitials
- (b) Edge dislocations and screw dislocations
- (c) Stacking faults, grain boundaries
- (d) Inclusions, voids, microtubes.

Answer 2: They affect the homogeneity of the distribution of donors and acceptors. They affect the mobility and recombination time of carriers.

Answer 3: Extrinsic defects are related to impurities. Intrinsic defects are the consequence of a wrong arrangement of the atoms relative to the ideal crystal.

Answer 4: $n/N \approx e^{-\varepsilon/kT}$, where ε is the formation energy of the defect.

Answer 5: Since the formation of the defect and the diffusion of atoms slow down considerably, already slightly below the melting point, thus the concentration and distribution stay close to the one encountered at thermal equilibrium slightly below the melting point.

References

1. One shouldn't bother with semiconductors, those are dirt effects." Wolfgang Pauli 1931, quoted, e.g. as entrance motto in: Manuel Cardona, Theory of defects in semiconductors. Springer Book series Topics of Applied Physics Vol. **104** (2006).
2. The data are taken from U. Goesele: Annual Review of Material Science **18** 257–282 (1988).
3. Ludwig Boltzmann (1844–1906), theoretician, founding father of statistical mechanics.

Chapter 7

Compound Semiconductors



Abstract III-V-compound semiconductors are chemical compounds of two elements, one of the 3rd group and one of the 5th group of the periodic table. Most III-V-compounds have the same crystallographic as well as electronic structure as Silicon, with each atom sharing all together eight valence electrons with its four neighbors. The combination of III and V elements results in a large number of potential semiconductor materials, some of which have achieved a high industrial significance, among them GaAs, InP, and GaN. SiC as a IV-IV-compound may be added to this list of important compound semiconductors. Bulk crystals can be grown from the melt, like GaAs and InP, or in a gas-phase process like SiC, GaN, and AlN. The potential of III-V-semiconductors is significantly increased by the almost unlimited miscibility of the materials, resulting in ternary and quaternary compounds with properties “in between” the ones of the constituents. This opens a wide “playing ground” for materials engineering for a wide range of applications, e.g. in optoelectronics (GaAs, InP, and GaN), in high frequency (GaAs and GaN), and in high-power and high-temperature electronics (SiC and GaN). Mixed crystals are grown as epitaxial layers on an appropriate “lattice-matched” substrate, preferentially by Metalorganic Vapor Phase Epitaxy (MOVPE) or Molecular Beam Epitaxy (MBE).

7.1 Introduction, Some History

As already stated in the introduction to Chap. 5, Silicon is by far the most important semiconductor material. This is not just due to the fact that for most conceivable applications Si is the most appropriate material, but also due to availability, costs, and most advanced technology. For quite a number of applications, other materials would be more suitable judged on physical properties alone. There are even applications where Si cannot be used at all, as already mentioned in Sect. 1.7: Because Si is an indirect semiconductor, it is not suitable for light-emitting devices like light-emitting diodes and semiconductor lasers, nor can it be used for high-temperature electronics (Sect. 4.6).

This is where the III-V-compound materials come into play. Those are compounds consisting of an element of the IIIrd and the Vth group. The idea of compound semiconductors goes back to Heinrich Welker.

Heinrich Welker, see Fig. 7.1a, certainly is one of the great pioneers of semiconductor physics and technology.

He was born in 1912, studied physics, and was the scientific assistant of Arnold Sommerfeld from 1935 to 1939. Later, he worked at the Institute of Chemical Physics of the University of Munich, where between 1942 and 1945 he worked on the Germanium rectifying diode. In 1945, he developed the theoretical concept of the field effect transistor. During his time as chief of Westinghouse Laboratory in Paris, he developed the first transistor (a field effect transistor) together with Herbert F. Matare. It was essentially a matter of company publishing policy that J. Bardeen, W. B. Shockley, and W. Brattain preceded them with their paper on the first bipolar transistor. In 1951, he entered the Siemens research laboratory, where he invented the compound semiconductors. In 1961, he became chief of Siemens Corporate Research. He died in Erlangen in 1981.

Figure 7.1b shows the patent on the new class of semiconductors from the year 1951.

The basic idea is shown in Fig. 7.2a, b.

Figure 7.2a shows the crystal structure of Si as compared to the one of GaAs. They are both of fcc (Zinc blende) type. The difference between the two is that in GaAs the “basis” (see Sect. 1.1) is formed by two different atoms, Ga and As. This corresponds to the bonding situation shown schematically in Fig. 7.2b.

In Si as well as in GaAs, each atom shares altogether eight valence electrons with its four neighbors, two with each one, forming the bonding orbital.

Figure 7.3a gives an excerpt of the periodic table around the IIIrd and Vth groups.

Figure 7.3b shows the possible III-V-compounds that can be obtained by the combination of the corresponding elements. The ones marked in bold letters have the fcc structure like Si or GaAs, the other ones the Wurtzite structure

Actually, by analyzing the situation in more detail, based on Linus Pauling’s theory of chemical bonding, Heinrich Welker concluded from properties like ionic radius, ionicity, atomic separation, and others that the compounds marked in bold letters in Fig. 7.3b would show the fcc (Zinc blende) structure. Welker called them “copies” (Nachbildungen) of Si and estimated from the parameters just mentioned their bandgaps, mobilities, and other properties. It is obvious that in this way, by combining the different elements of the IIIrd and the Vth group, a great number of new semiconductor materials could be postulated and subsequently be confirmed as such. Table 7.1 summarizes bandgap and mobility data as well as the direct or indirect nature of those compounds. For electronic applications, the (stable) wurtzite modification of GaN is used, not the also existing less stable cubic modification.

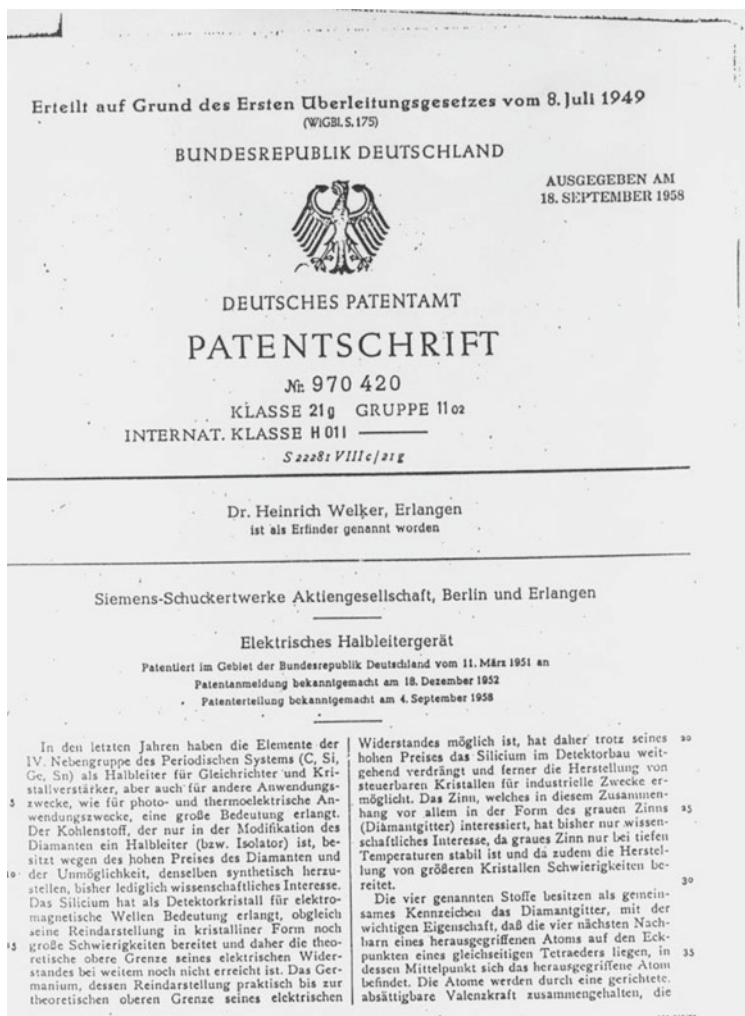


Fig. 7.1 a Heinrich Welker, pioneer of semiconductor technology (Courtesy „Siemens Historical Institute“). b Patent of Heinrich Welker from the year 1951

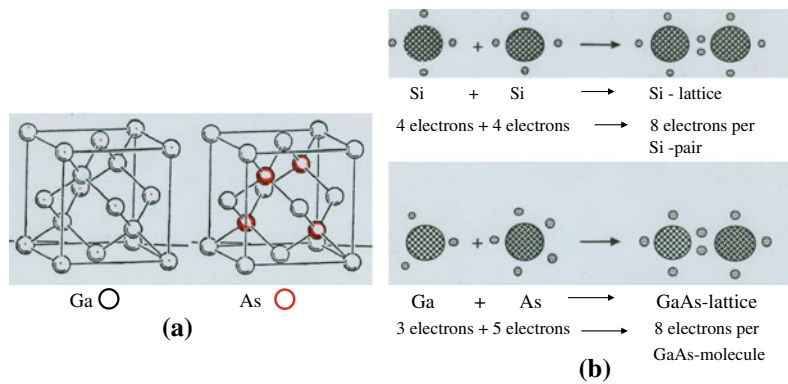


Fig. 7.2 **a** Si and GaAs crystal lattice. **b** Electronic structure of Si and GaAs

Group III	Group IV	Group V
B	C	N
Al	Si	P
Ga	Ge	As
In	Sn	Sb

a

b

BN, BP, AlN, GaN, InN, AlP, GaP, InP, AlAs, GaAs, InAs, AlSb, GaSb, InSb, BSb, AIP, GaP, InP

Fig. 7.3 **a** Excerpt of the periodic table around IIIrd, IVth, and Vth group. **b** List of III-V-compounds

Table 7.1 Data on the most important III-V-compounds.¹ For comparison, the data on 4H-SiC as well as Ge and Si are also given. Some of the data are already presented in Table 2.2 and are repeated here for comparison

Material	E _G /eV	μ _n /cm ² /Vs	μ _p /cm ² /Vs	Direct/Indirect
GaAs	1.42	8500	400	d
GaP	2.27	300	100	i
InP	1.34	5000	150	d
GaN (wurtzite)	3.44	1000	200	d
4H-SiC	3.26	900	120	i
Ge	0.66	3800	1200	i
Si	1.12	1500	400	i

¹ See, e.g. [1].

Physics Comment: Similarity of the Band Structure of Direct and Indirect Semiconductors

It might at first sight seem surprising that materials which are so similar in terms of their chemical bonding situation, and which correctly have been declared by Welker to be “copies” (Nachbildungen) of each other, differ in a property, which looks so fundamental, at least in terms of possible applications, like “direct” or “indirect” semiconductor. Isn’t this property (direct or indirect) a consequence of the band structure, and shouldn’t the band structures be very similar to each other?

The answer is as follows: The band structures of the cubic III-V-semiconductors are very similar to each other, indeed, nevertheless they may easily differ in a property like “direct” or “indirect”. To see this, let us have a look at Fig. 1.10 showing the band structures of Si and GaAs in comparison, the former being an indirect, and the latter being a direct semiconductor. The comparison shows that the band structures are qualitatively very similar, indeed (existence of light and heavy holes, and many other similarities), the main difference being the relative height of the various conduction band minima. In GaAs, the minimum at the Γ -point is the absolute lowest point of the conduction band. Because the maximum of the valence band is at the Γ -point, too, this makes GaAs a direct semiconductor. In Silicon, the Γ -point minimum is above the minimum near the X-point, the latter being the absolute minimum of the conduction band. This makes Si an indirect semiconductor. That makes all the difference! The band structures might be quite similar, and they are, **but the different positions of the absolute minima of the conduction band make the difference between direct and indirect.**

7.2 The Potential of III-V-Compounds

The number of new semiconductor materials, introduced by Welker’s invention, is even more increased by the fact that mixed crystals can be formed due to the complete miscibility of most of them. The situation is schematically shown in Fig. 7.4 for a mixture of 60–40 of GaAs and AlAs.

Layers of such mixed crystals can be grown one on top of the others to form electronic devices.

The point is that the practically relevant properties of those mixed crystals, like bandgaps or mobilities, usually lay between the “corner stones”, so, e.g. the bandgap of $\text{Ga}_{1-x}\text{Al}_x\text{As}$ is between the one of GaAs and AlAs (depending on x , of course). As a consequence, a large number of materials with a large range of “properties made to order” is made available. This fascinating aspect seems at first sight limited by the fact that those mixed crystals cannot be grown as bulk material, but only in

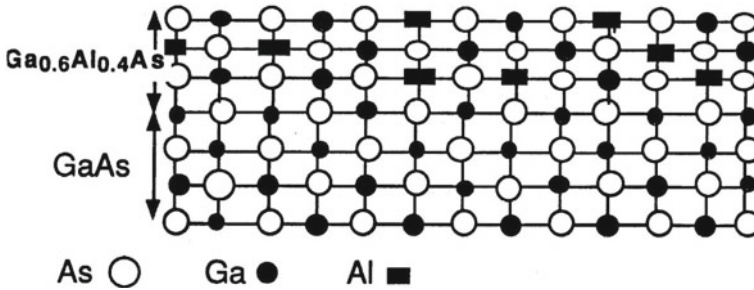


Fig. 7.4 Layers of different III-V-semiconductors

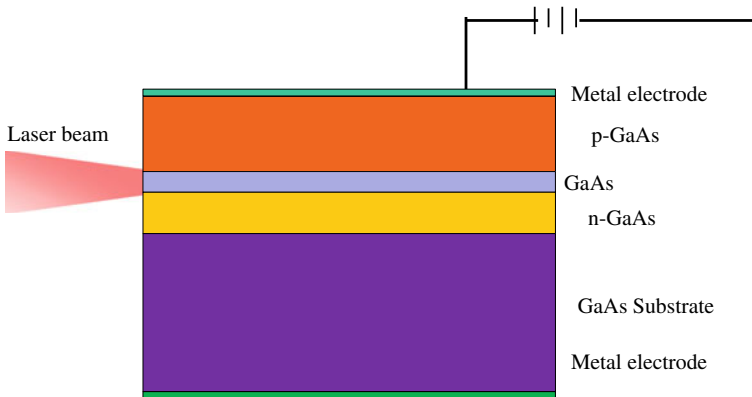


Fig. 7.5 Modern semiconductor laser as an example of a multi-layer-structure

thin layers by methods called “epitaxy” (see Sect. 7.5). This limitation, however, is not as serious as it may sound because most electronic devices are constructed as a system of thin layers that “carry” the electronic functions. For illustration, Fig. 7.5 just anticipates a typical semiconductor laser (discussed later in Chap. 12). It consists of multiple “epitaxial” layers which provide the electronic functions, grown on an appropriate “substrate”, here GaAs.

The requirement for perfect crystal growth of such “epilayers” is identical lattice structure and sufficiently close lattice constant of neighboring layers (or layer and “substrate”). Deviations in the lattice constants between the two layers (or the layer and its substrate) result in crystal defects, mainly dislocations, as already explained in Sect. 6.3. These defects can strongly affect the electronic functions and have to be avoided by a good “lattice match”. **So a modern electronic device, based on compound semiconductors, as a rule requires not just one material, but a whole “system” of materials, including the substrate and various layers of mixed**

crystals, all being identical in their crystal structure and close to each other in their lattice constants.

An example of a good lattice match is provided by the system GaAs, AlAs, and their mixtures:

Their lattice constants a are very close to each other:

$a(\text{GaAs}) = 0.56533 \text{ nm}$ and $a(\text{AlAs}) = 0.56605 \text{ nm}$, while the bandgaps are considerably different: $E_G(\text{GaAs}) = 1.42 \text{ eV}$ and $E_G(\text{AlAs}) = 2.16 \text{ eV}$. So any bandgap can be tailored between these two values by a proper mix $\text{Ga}_{1-x}\text{Al}_x\text{As}$ of the two materials. This mixed crystal can be grown on a GaAs substrate, because the lattice constants of the two materials GaAs and AlAs differ only by $(\Delta a/a) = 1.3 \times 10^{-3}$.

For $x = 0.2$, the difference between the layer and the substrate would be only approximately a fifth of that, i.e. $\text{ca. } 2.6 \times 10^{-4}$. Such a mismatch would be perfectly alright for a film of that composition grown on a GaAs substrate. Because of the good lattice match between GaAs and AlAs this system has been used for quite a number of pioneering applications of epilayers like semiconductor lasers. The semiconductor laser shown schematically in Fig. 7.5 is an example.

It should be mentioned at this point, however, that by certain modifications of epitaxial methods nowadays systems with large mismatch are being in use and of great technical significance. This will be discussed in Sect. 11.6 for the case of the blue light-emitting diode, which can be fabricated from GaN on substrates like SiC (mismatch $\text{ca. } 3.5\%$) or even sapphire (mismatch 20% !).

In the patent of Welker (Fig. 7.1b), the III-V-compounds are largely considered as an alternative to Ge and Si. At that time, only Ge as a semiconductor material was ready for use on an industrial scale, Si being under development. So a demand was seen for alternative materials. For that reason, the patent met very high attention in the community, as—by the way—illustrated by the fact that 145 objections were raised against it. It turned out, however, that the III-V-semiconductors did not act as replacements for Si or Ge, but opened the way to new applications, due to their specific physical and electronic properties.²

The “classical”, widely used compound semiconductors are **GaAs and InP**.

GaAs, as mentioned earlier, is a direct semiconductor with a bandgap of 1.42 eV at room temperature. A major application is the manufacturing of microwave devices. The high electron mobility as well as the large saturation velocity (see Chap. 2) make transistors possible up to the frequency region of 250 GHz (!). So GaAs is more and more used in mobile phones, satellite communication, and high-frequency radar.

Due to its direct bandgap, it can be applied in all kinds of optoelectronic devices like near-infrared light-emitting diodes and lasers (Chaps. 11 and 12). In these applications, usually epilayers of mixed crystalline material $\text{Ga}_{1-x}\text{Al}_x\text{As}$ are used with GaAs as a substrate.

InP has developed into a material for a large variety of electronic applications in optical communication. It provides the possibility to integrate light generation,

² Bergner [2].

light detection, light modulation, etc. on one chip. It is also widely used for high-speed switches and amplifiers. The most important applications are substrates for semiconductor lasers in the wavelength region of optical data transmission around $1.54\ \mu\text{m}$ (Sect. 12.4).

7.3 Growth of GaAs and InP by the LEC and VGF Process

The classical way to grow single crystals of GaAs or InP for electronic applications is the Liquid Encapsulated Czochralski (LEC) process. The LEC process is essentially identical to the Czochralski growth process of Si (Sect. 5.3), i.e. the crystals are pulled from the melt. The special feature of GaAs and InP is that at the melting temperature ($1240\ ^\circ\text{C}$ for GaAs and $1065\ ^\circ\text{C}$ for InP), these materials start to disintegrate already, which means As evaporates from the GaAs melt and P from the InP melt. This has to be suppressed by putting a “lid” on the melt in form of a layer of liquid Boron Oxide (B_2O_3), which in this way “encapsulates” the melt (Fig. 7.6a).

Boron Oxide is a solid which melts at $450\ ^\circ\text{C}$ and thus covers the GaAs or InP melt as a transparent liquid. “Transparent” is an important advantage because it permits the crystal grower to observe the formation of the crystal at the interface between melt and crystal. Actually, what the experienced grower does is observing the lamella developing at this interface which permits conclusions about the realization of the proper growing conditions. At present, GaAs wafers of 8 in. diameter are on the market.

The crystals are usually grown in pBN crucibles (so-called pyrolytic BN). BN can be produced at high purity, so the main impurity incorporated from the crucible into the crystal is Boron. Boron, however, does essentially not affect the electronic

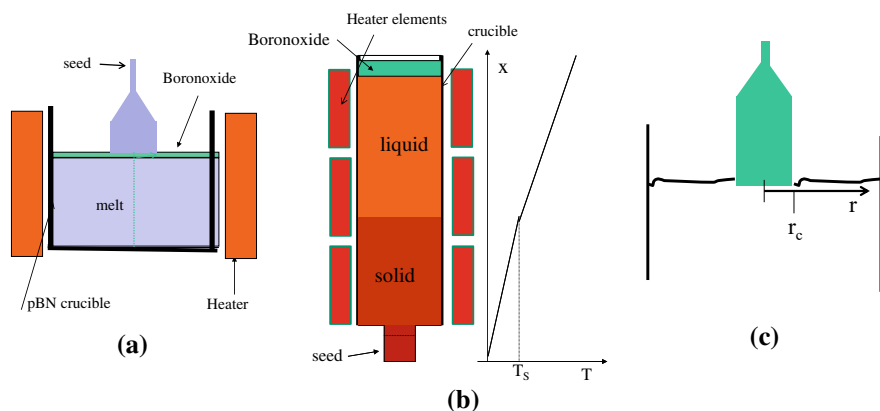


Fig. 7.6 **a** The Liquid Encapsulated Czochralski (LEC) process. **b** The Vertical Gradient Freeze (VGF) process. **c** Fixation of the radius of the growing crystal in the LEC growth

properties. This point has already been discussed in Sect. 3.5. It is an “isoelectronic impurity” which does not give rise to a donor or acceptor.

The other important (and more recent) method for the growth of GaAs and InP is the Vertical Gradient Freeze (VGF) process (Fig. 7.6b).

In a multizone oven, the temperature profile (right side of Fig. 7.6b) is slowly shifted, such that the melting temperature and—correspondingly—the interface between melt and solid move upwards, leaving behind a single crystal if the process has been started with the interface within the seed crystal at the bottom. The advantage of the process, as compared to the LEC method, is twofold:

1. Crystals with much lower dislocation density can be grown due to the fact that thermal gradients in the radial direction can be kept much smaller than in the LEC process. Thus, thermal tensions, which give rise to dislocations, can be kept much smaller than in the LEC process. This is a consequence of a basic difference between the two processes: In the LEC process, a certain radial temperature gradient is unavoidable because the spatial temperature field at any moment defines the diameter of the crystal. In Fig. 7.6c, the radial temperature gradient is such that for radii $r < r_c$ the temperature is below the melting temperature of the material, so in this region, it is a solid, and for $r > r_c$, the temperature is above the melting temperature, so there we have a liquid. The point $r = r_c$, where the temperature is precisely equal to the melting temperature, defines the momentary radius of the growing crystal. So, by principle, a temperature gradient in the radial direction is required in the LEC process in order to define the crystal radius. If the temperature gradient is too small, the radius is not well defined; it fluctuates during the growth of the crystal, and the crystal in the end looks like a “Christmas tree”, as crystal growers say.

The situation is different, however, in the case of the VGF process. Here, the diameter of the growing crystal is defined by the diameter of the crucible. So in principle, a crystal can be grown in a zero radial gradient temperature field.

This results in a higher crystalline quality.

2. A second advantage of the VGF process as compared to the LEC process is that there is no temperature inversion: The low temperature is at the bottom, the lower temperature upwards (see temperature profile Fig. 7.6b right). So no convection driven by the vertical T-gradient sets in, different from the situation in the Czochralski process.

Specific problems of the VGF process, however, result from the contact of the growing crystal with the crucible wall. One problem is purity: Impurities can be incorporated from the crucible. This problem, however, is not very severe nowadays because of the high purity of the BN-crucible material. More severe, however, is the interaction of the growing crystal with the crucible wall, because irregularities in the crucible surface can result in perturbations of the crystal growth, mainly in the formation of stacking faults (see Sect. 6.5). Stacking faults, in turn, may lead to the formation of crystal twins. This problem is particularly relevant for InP because of the low stacking fault formation energy in InP. Altogether the VGF process now

dominates the fabrication of bulk GaAs and InP because of the higher crystalline quality and purity.

A high section of the GaAs market is covered by so-called “semi-insulating” substrates. These are very high resistivity wafers required mainly for integrated circuits. The high resistivity of the substrate guarantees that the current flow goes through the electrically active epitaxial layers, keeping the substrate essentially free from currents. The principle of semi-insulating materials, an important issue of semiconductor materials technology, has been treated in Sect. 4.5.

7.4 “New” Compound Semiconductors: SiC, AlN, and GaN

Silicon Carbide SiC

So far the crystal growth of traditional III-V-compounds like GaAs and InP has been discussed. In the following, more recent developments will be treated concerning relatively new compound semiconductors like SiC, GaN, and AlN.

SiC is evidently not a III-V-, but a IV-IV-compound. It is a semiconductor material with properties making it most suitable for power electronics. Table 7.2 summarizes some relevant properties of SiC and GaN, as compared to Si.

SiC exists in various crystal modifications (“polytypes”). They differ in the way the Si- and C-layers are stacked. The one most suitable for power electronics is the hexagonal 4H-polytype. Another polytype, 6H-SiC with a bandgap of 2.86 eV, achieved a major importance as a substrate for GaN LEDs (see Sect. 11.6). It should be noted that more recently also the cubic polytype ($E_G = 2.2$ eV) with interesting properties is getting available.

The high breakthrough voltage and the high saturation velocity (as compared to Si) make SiC an ideal material for high-power electronics at fast switching rates. A huge market is opened by hybrid and electrical cars, a development only in its starting phase, but—obviously—with huge potential, resulting in enormous growth rates of SiC electronics. Rectifying diodes and MOSFETs are the most important devices.

As Table 7.2 shows GaN shares the advantages of 4H-SiC in terms of breakdown field strength and saturation velocity. However, its thermal conductivity is

Table 7.2 Summary of properties of Si, 4H-SiC, and GaN

Properties	Silicon	4H-SiC	GaN
Crystal structure	fcc (diamond structure)	Hexagonal	Hexagonal
Energy gap (E_G /eV)	1.12	3.26	3.5
Breakdown field strength (V/cm)	0.3×10^6	3×10^6	3×10^6
Saturation drift velocity (cm/s)	1×10^7	2.7×10^7	2.7×10^7
Thermal conductivity W/cmK	1.5	4.9	1.3
Direct (d) or indirect (i)	i	i	d

only in the range of the one of Si, making SiC the leading material in power electronics. The high thermal conductivity helps to dissipate the heat while, as explained already in Sect. 4.6, the wide bandgap permits the operation of electronic devices at temperatures of a few hundred centigrade.

The growth of SiC has been a challenge in crystal growth technology for decades. Different from “conventional” materials like Si, GaAs, or InP, bulk crystals cannot be pulled from the melt. SiC does not melt at all, but—as the phase diagram of the Si–C-system shows—at a temperature of about 2900 °C disintegrates into liquid Si and solid carbon. Thus, the method of choice is the one of Physical Vapor Transport (VPT) shown in Fig. 7.7a. The process is also called the “Lely process” named after Jan Anthony Lely of Philips Electronics. The patent was filed in 1955 and granted in 1958.

The supply, high purity SiC powder, is heated to a temperature of about 2000 °C in a graphite crucible (graphite is one of the few materials withstanding this temperature, and it has the advantage that it does not introduce a new element into the system, just carbon). At this temperature, the SiC supply starts to evaporate. Because the seed crystal, mounted at the lid of the crucible, is kept at a slightly lower temperature (about 20°–40° lower), the constituents of the vapor are deposited on its surface, forming the growing crystal. The name Physical Vapor Transport originates from the idea that it is SiC as a molecule which is transported from the supply to the growing crystal, with no chemistry happening in the vapor phase. This does, however, not correspond to reality. The vapor is rather composed of Si, Si₂C, SiC₂, and other constituents, reacting with each other and with the graphite wall. The complexity of this chemistry makes the modeling of the process so much of a challenge.

The growth rate is considerably slower than in Czochralski growth, basically due to the smaller density of a vapor as compared to a melt. While the growth rate is of the order of 0.1–1 mm/min in Czochralski- or LEC-growth, it is about two orders of magnitude smaller than PVT growth. So a SiC crystal of a few centimeters in length will grow within about three days.

It might be noteworthy that the first application of SiC, leading to a systematic study of the growth of 6H-SiC, was the blue LED (see Sect. 11.6). The first blue LED actually consisted of a pn-junction in 6H-SiC, with very poor intensity, though,

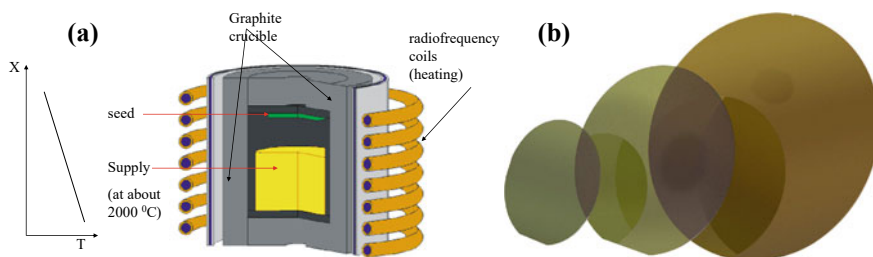


Fig. 7.7 **a** Physical Vapor Transport (VPT) growth setup of SiC. **b** 4H-SiC wafers: 100, 150, and 200 mm (state of the art). Both a) and b) by courtesy SiCrystal AG

because SiC is an indirect semiconductor. Later 6H-SiC became important as a substrate for the GaN blue light-emitting diode. These optoelectronic developments of 6H-SiC preceded the electronic applications of 4H-SiC mentioned above.

The great challenge in providing bulk crystals of SiC (either 6H or 4H) was the occurrence of micropipes (see Sect. 6.6 and Fig. 6.9b,c). Because these defects turned out to originate from screw dislocations with large Burger's vector, which in turn result from large thermal stress during growth, the solution could be found by reducing thermal gradients as much as possible during the growth process. In this context, modeling of the thermal field played a crucial role. It helped to improve the setup by the introduction of thermal shields and other geometrical measures.

The state of the art in SiC growth is most impressive, and the development is pushed forward by the enormous demand for SiC devices for high-power electronics. 150 mm wafers are dominating the market now, and 200 mm is in a test stage in the leading companies. Figure 7.7b shows such a state-of-the-art wafer of SiCrystal AG, together with the previous generations of 100 and 150 mm. One could say that SiC growth is making good the separation in diameter to Si at high speed.

At present, rapid progress in SiC technology is mainly driven by the requirements of E-mobility.

AlN

Bulk crystals of AlN can be grown by Physical Vapor Transport (PVT) as well (Fig. 7.8a)."

Temperatures are slightly higher than in SiC growth. A tungsten crucible is used instead of graphite to avoid harmful C-incorporation into the growing crystal. ΔT in the figure refers to the temperature difference between the AlN supply and seed crystal.

A severe problem in AlN growth is the unavailability of seed crystals. Large-diameter AlN wafers are not available yet. Unfortunately, the growth kinetics of the different facets of AlN are such that enlargement of crystal diameter during growth

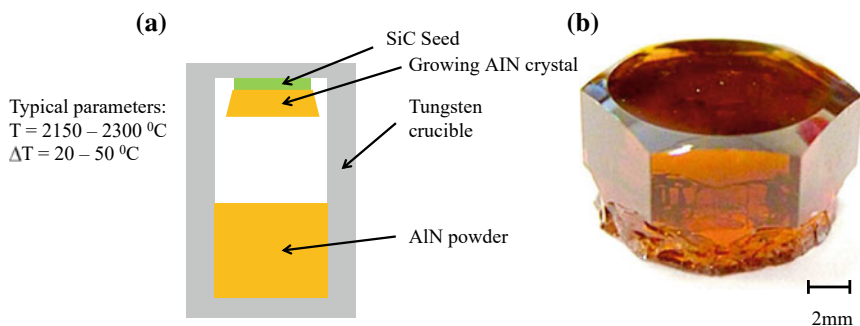


Fig. 7.8 **a** PVT growth of AlN. **b** State-of-the-art bulk AlN crystal (Courtesy Leibniz Institut für Kristallzuechtung (IKZ) Berlin, Germany)

is very hard to achieve.³ So the present state of the art is a wafer diameter of about 16–25 mm. The growth rate is in the order of a few 100 μm per hour.

Figure 7.8a refers particularly to the experiments where it was tried to solve the problem of diameter enlargement by growing AlN on a SiC seed crystal. Here, the diameter of the SiC seed crystal would define one of the growing AlN crystals. In spite of a fairly good lattice match (for a proper choice of crystal orientations), the results unfortunately were not satisfactory in terms of crystal quality.

In spite of these difficulties, efforts to grow larger and better AlN bulk crystals are continuing. Figure 7.8b shows the state of the art, as far as this can be defined in view of the rapid progress.⁴

AlN wafers might be useful, e.g. in the context of light-emitting diodes (LEDs) in the spectral region of very short-wavelength UV (UV-C, below 280 nm) to act as substrates for $\text{Al}_{1-x}\text{Ga}_x\text{N}$ LEDs (in this short-wavelength region, the Al-content x must be >0.3). This topic will be treated in more detail in the chapter on light-emitting diodes (Sect. 11.8). Apart from PVT, developments are under way to grow AlN wafers by HVPE (Hydride vapor phase epitaxy), i.e. via reactions analogous to the ones given in the following section for GaN.

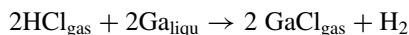
Bulk crystals of GaN

It was mentioned in Sect. 7.3 that in the growth of GaAs and InP, the disintegration of the materials at growth temperatures plays a hindering role. In order to suppress the evaporation loss of As in GaAs growth and of P in InP growth, the melt has to be covered by a layer of Boronoxid, acting as a “lid” on top of the melt. In InP, the Phosphorous pressure at growth conditions amounts already to more than 30 bar, demanding an appropriate pressure stability of the growth chamber. In GaN, the corresponding N pressure would be orders of magnitude higher, making a growth process from the melt absolutely impractical. The most promising way to fabricate bulk GaN crystals and wafers is the one of **Hydride Vapor Phase Epitaxy (HVPE)**

The process of HVPE is, as the name says, an epitaxial process. In the first step of this process, a gaseous HCl flow is sent over the surface of liquid Ga. At a temperature of 600–800 °C, gaseous GaCl is formed which subsequently (i.e. “downstreams”) reacts with NH_3 to GaN.

The underlying reactions are the following:

First step:



Second step:



³ Filip et al. [3].

⁴ Hartmann et al. [4].

In this epitaxial process, enormous growth rates of the order of $100\text{ }\mu\text{m}$ per hour can be achieved such that within a few hours wafers of bulk GaN of a few $100\text{ }\mu\text{m}$ thickness can be obtained. As a template (substrate), a thin layer (thickness some nm) of GaN grown by MOVPE (see next Sect. 7.5) on, e.g. sapphire is being used resulting in wafers of reasonable quality (some 10^6 dislocations per cm^2) suitable as substrates for high-quality optoelectronic and high-power devices. HVPE seems to be the method of choice for the fabrication of bulk GaN wafers.

7.5 Epitaxy

As discussed above in this chapter (Sect. 7.2), modern semiconductor devices usually consist of a system of thin layers (see, e.g. Fig. 7.5) providing the electronic function, grown on a substrate). The thickness of these layers is a few nanometers, as compared to a few hundred μm for the substrate. They are grown by “epitaxy”. The two most widely used methods, molecular beam epitaxy (MBE) and metalorganic vapor phase epitaxy (MOVPE), will be discussed in the following.

Physics Comment: On the Notion of “Epitaxy”

An important point to be stressed (already mentioned in Sect. 7.2) concerns the notion of “epitaxy”. An “epitaxial layer” is not just a thin layer of material on a substrate. A thin layer of a few nanometer thickness on a glass substrate would not be an epitaxial layer. The essential point in the definition of epitaxy is that the substrate acts as a template of the thin layer in the sense that the former (i.e. the substrate) determines the crystal structure of the latter. The layer “fits” on the substrate which is usually assured by the two materials having the same crystal structure with almost the same lattice constant. Other combinations of crystal structure, however, may also provide such a fit by growing appropriate crystallographic orientations one on top of the other.

Molecular beam epitaxy MBE

The method of MBE is very schematically shown in Fig. 7.9a for the example of $\text{Ga}_{1-x}\text{Al}_x\text{As}$.

The GaAs substrate is mounted on a temperature-controlled support. The materials are provided by evaporator sources, so-called effusion cells, which, due to an appropriate system of shields, produce a directional beam at sufficiently low pressure in the chamber (otherwise, the beam would be widened by scattering by the residual gas). By controlling the temperature of these evaporation sources, the Ga-Al-fraction can be controlled, as well as the concentrations of the Be acceptor and the Sn donor. By the shutters, the beams can be interrupted if desired.

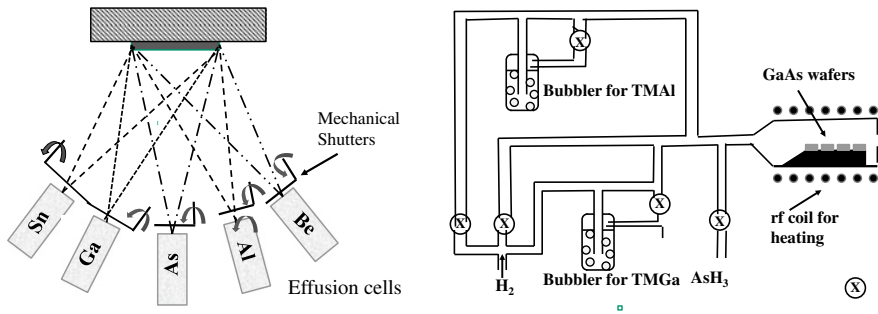


Fig. 7.9 **a** Molecular beam epitaxy MBE of $\text{Ga}_{1-x}\text{Al}_x\text{As}$. **b** Metal organic vapor phase epitaxy MOVPE

A crucial requirement for MBE is ultra high vacuum (UHV), i.e. a vacuum better than 10^{-10} mbar. This requirement makes MBE a high-cost process not well suited for industrial fabrication.

By MBE multilayer systems can be grown with 1 atomic layer precision (for clarity, e.g. a system of a few layers of GaAs alternating with a few layers of AlAs).

Physics Comment: Why Is Ultra High Vacuum (UHV) Required for MBE?

As an answer to this question, we sometimes encounter the statement that the problem is scattering of the molecular beam by the residual gas atoms, perturbing or destroying the directionality of the beam. This is, however, not the case. At an air pressure of 10^{-6} mbar, the mean free path of an atom is about 100 m, much larger than the dimensions of the MBE apparatus. The true reason is given by the purity requirement of the growing semiconductor film.

In a vessel filled with a gas, the walls, or any surface, are permanently exposed to a shower of impinging atoms or molecules of that gas. In this way, the growing semiconductor film is exposed to impinging oxygen or nitrogen atoms of the residual air in the MBE system, a fraction of which will stick to the film and thus will result in an impurity content of the growing semiconductor layer. If the residual gas pressure is too high, this will result in an intolerable impurity content of the film.

How many air molecules (oxygen or nitrogen) hit the wall of a vessel at a given pressure and temperature? This can be easily answered in the following way:

The molecules hitting the wall of the vessel give rise to what we call the “pressure” of the gas. Each atom hitting the wall transfers some of its momentum to the wall. So the shower of impinging molecules transfers a certain total amount of momentum per time to the wall. Because “force”

is transferred momentum per time, this results in a force on the wall, the “pressure”, if we consider the force per area.

The unit of pressure is $1 \text{ Pa} = 1 \text{ N/m}^2$. Let the number of molecules hitting 1 m^2 of the wall per second be called Z . Let the average momentum transmitted to the wall when hit by a molecule be called Δp . Then the total momentum transmitted to the wall per m^2 and second is $Z\Delta p$, which corresponds to the pressure P of the gas:

$$P = Z\Delta p \quad (7.1a)$$

If the pressure P is 1 Pa , then

$$Z(1 \text{ Pa}) = 1 \text{ Pa} / \Delta p = (1 \text{ N/m}^2) / \Delta p \quad (7.1b)$$

is the number of molecules hitting the wall per sec and m^2 . So we have to calculate the average momentum transferred to the wall by a molecule hitting the wall. We consider the case of room temperature $T = 293 \text{ }^\circ\text{C}$.

Because of $E = p^2/2m$, we get $p = (2mE)^{1/2}$ for the average momentum of a gas molecule. Because the average energy of translational motion of the molecule is $(3/2)kT$, we obtain

$$p_{av} = (3mkT)^{1/2} \quad (7.1c)$$

for the average momentum of the gas molecules. If the molecule hits the wall at an angle of 90° and is reflected back, the transmitted momentum is $2p_{av}$, and if it hits under a very small angle (almost zero angle of incidence), the transferred momentum is 0. Because we are only interested in the order of magnitude, we just take something between 0 and $2\Delta p_{av}$ for Δp_{av} , i.e. Δp_{av} itself, for the value of the momentum transferred to the wall by the molecules. So (7.1b)

$$Z(1 \text{ Pascal}) = 1(\text{N/m}^2) / (3mkT)^{1/2} \quad (7.2)$$

gives the number of molecules hitting 1 m^2 of the wall per second, m being the mass of the molecule.

For kT , we take the value at room temperature ($20 \text{ }^\circ\text{C}$): $kT = 0.026 \text{ eV}$.

The molecular weight of O_2 is 32, and the one of N_2 is 28. Because air consists of 80% nitrogen and 20% oxygen, the average molecular weight of air is 29.

So we take for the mass of molecules hitting the wall

$$m = 29 \times 1.68 \times 10^{-27} \text{ kg} = 4.87 \times 10^{-26} \text{ kg}$$

(1.68×10^{-27} kg is the mass of a nucleon)

Each of these molecules transfers in the average a momentum of (see 7.1c):

$$\begin{aligned} p_{av} &= (3 \times 4.87 \times 10^{-26} \text{ kg} \times 0.026 \times 1.6 \times 10^{-19})^{1/2} \\ &= 2.5 \times 10^{-23} \text{ kg m/s} \end{aligned}$$

Inserting this into (7.1c) gives

$$Z(1 \text{ Pascal}) = 0.4 \times 10^{23}$$

So 0.5×10^{23} molecules hit the wall at a pressure of 1 Pa per second and per square meter.

Or finally, because $1 \text{ Pa } 10^{-5} \text{ bar} = 10^{-2} \text{ mbar}$ and referring to 1 cm^2 :

$$Z(10^{-2} \text{ mbar}) = 0.4 \times 10^{19} \text{ Hits/cm}^2\text{s} \quad (7.3)$$

Let us consider a “high vacuum” of 10^{-7} mbar, then we have 0.5×10^{14} Hits per sec and cm^2 . What does that mean in practical terms?

A typical growth rate of MBE would be 1 atomic layer per sec (corresponding to about $1 \mu\text{m}$ per hour). In 1 atomic layer, one has about 10^{14} – 10^{15} atoms/ cm^2 . So even if only a small fraction of those 0.5×10^{14} atoms/ cm^2s hitting the surface of the growing layer would stick to it, a heavy pollution would result. However, at something like 10^{-11} to 10^{-12} mbar, i.e. in UHV, a good purity could be achieved.

So the result of this instructive little calculation is UHV is required in MBE because of purity requirements.

Metal organic vapor phase epitaxy MOVPE

In MOVPE, all components are provided by a gas flow. The method is shown in Fig. 7.9b for GaAlAs as an example. The GaAs wafers, acting as substrates, are fixed on the wafer mount.

Ga and Al are provided by a flow of Trimethylgallium $\text{Ga}(\text{CH}_3)_3$ (TMGa) and Trimethylaluminum $\text{Al}(\text{CH}_3)_3$ (TMAI), respectively. Hydrogen acts as a carrier gas. TMAI and TMG are liquids at room temperature. The gas flow is produced by “bubbling” the Hydrogen through the liquid in the “bubbler”. Arsenic is provided by a stream of AsH_3 (arsine). The compounds decompose at the hot surface of the growing AlGaAs layer. In the same way, the dopants are provided by a gas flow, e.g. the acceptor Boron as Diborane B_2H_6 .

In the early history of compound semiconductor technology, other methods than MBE and MOVPE dominated the scene. The so-called liquid phase epitaxy played a decisive role in the historical development. But this is history. Nowadays MOVPE is

the method of choice. The fact that all components are delivered in form of a gas flow gives a superb control of the fabrication processes, permitting almost a control on a one-atomic-layer scale at reasonable costs. This technology forms the basis of the tremendous development of optoelectronics in our time, permitting, for example, the high volume fabrication of huge numbers of extremely bright light-emitting diodes and of semiconductor lasers, built as multilayer devices (see Sect. 12.3).

It should be mentioned at this point already that the requirement of lattice match underlined above has been softened by modern developments, such that, e.g. nowadays GaN light-emitting diodes can be grown on sapphire substrates with a mismatch of 20%. This will be treated in more detail in Sect. 11.6.

7.6 Questions and Answers

- Question 1: Who invented the III-V-semiconductors, and where?
 Question 2: Name four of the most important compound semiconductors.
 Question 3: Give the names of the two growth processes of bulk GaAs and InP.
 Question 4: What is epitaxy?
 Question 5: For which system of III-V-compounds is the condition of lattice match very well fulfilled?
 Question 6: Which are the two most important epitaxy methods for the III-V-compounds?
 Question 7: Why does MBE require Ultra high Vacuum?
 Question 8: What are the most important applications of SiC, and which polytypes are used?
 Question 9: How are SiC bulk crystals grown?
 Question 10: What is the most promising method to produce wafers of GaN?
- Answer 1: Heinrich Welker in the Siemens Research Laboratories in Erlangen in the 1950s.
 Answer 2: GaAs, InP, GaN, and SiC.
 Answer 3: Liquid encapsulated Czochralski (LEC) process and Vertical Gradient Freeze (VGF) process.
 Answer 4: Epitaxy is the name of a growth process of thin crystalline layers on a substrate, where the substrate acts as a template for the growing layer.
 Answer 5: $\text{Ga}_{1-x}\text{Al}_x\text{As}$.
 Answer 6: Molecular Beam Epitaxy (MBE) and Metal Organic Vapor Phase Epitaxy (MOVPE).
 Answer 7: For achieving epitaxial layers of sufficiently high purity.
 Answer 8: Optoelectronics (6H-SiC substrate for the blue GaInN LED) and high-power electronics (4H-SiC).
 Answer 9: By the Physical Vapor Transport (PVT) Method, also called “Lely Process”.
 Answer 10: Hydride vapor phase epitaxy (HVPE).

References

1. S. Adachi, *Handbook on Physical Properties of Semiconductors (Vol. 2), III-V-Compounds* (Springer Science and Business Media, 2004)
2. J. Bergner, *Vom Wernerwerk zum Campeon, Sternstunden der Siemens Halbleiter und Infineon*, p. 101
3. O. Filip, B.M. Epelbaum, M. Bickermann, P. Heimann, A. Winnacker, Effects of growth direction and polarity on bulk aluminum nitride properties. *J. Cryst. Growth* **318**, 427–431 (2011)
4. C. Hartmann, L. Matiwe, J. Wollweber, I. Gamov, K. Irmischer, M. Bickermann, T. Straubinger, Favorable growth conditions for the preparation of bulk AlN single crystals by PVT. *CrystEngComm* **22**, 1762–1768 (2020)

Chapter 8

Amorphous Semiconductors



Abstract “Amorphous” materials are not completely disordered, like a liquid, but show a near range order. This results in certain modifications of the density of states like the existence of “tail states”, accompanied by local variations of the bandgap. In spite of the local variations of the bandgap, there exists a sharply defined “mobility edge” of the conduction band and valence band. It separates localized from delocalized states. Due to the existence of these localized “tail states” the dominant conduction mechanism is not via “band transport” of free charge carriers, but by “variable range hopping”. Another consequence is the elimination of the strict distinction between direct and indirect semiconductors. Another general feature of amorphous solids in general and a-Si, in particular, is the high density of deep-level states, states with energy levels deep in the bandgap, resulting mainly from dangling bonds. The existence of these deep states results in a doping problem for a-Si. This can be solved by “passivating” these states by hydrogen atoms, which, in turn, can be incorporated in high concentrations by using Silane SiH_4 as a-Si source in the plasma-discharge fabrication process. Because this process allows for the fabrication of large area films, typical applications of a-Si:H are large area devices like solar cells and the X-ray flat panel detector.

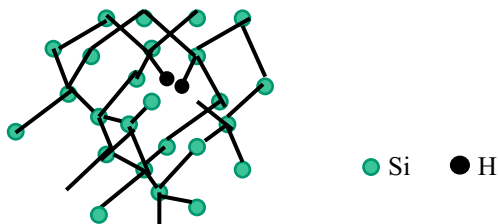
In discussing semiconductor materials, we so far have focused on crystalline solids. Nevertheless, amorphous materials play a considerable role in electronic devices, above all amorphous Si.

8.1 Short-Range Order and Band Structure

Figure 8.1 shows schematically the structure of amorphous Si, being described as “short-range order”.

While in crystalline Si (c-Si), all Si atoms are surrounded by 4 Si neighbors in a well-defined distance of 2,35 Å in a tetrahedral arrangement (Fig. 1.2b), all distances and angles being equal, in amorphous Si (a-Si), distances and angles are only approximately equal and vary slightly from bond to bond. This situation is

Fig. 8.1 Short-range order in amorphous Silicon



generally described as “short-range order”. The next neighbors are still situated in positions close to the ones they would have in the crystalline state. In that sense, the position of the nearest neighbors of a-Si atom are still strongly correlated with the one of that atom. The more one moves away from a given atom, the weaker that correlation becomes. One can roughly say that the position of the fourth or fifth nearest neighbor of an atom can hardly be predicted, the correlation is lost, in that sense, one can no longer speak of a “long-range order”, as in crystalline material, but only of a “near-range order”.

Sometimes a neighbor is completely missing resulting in a vacancy and a “dangling bond”. As explained later in this chapter, in practical applications, these dangling bonds are often saturated by H atom, as shown in Fig. 8.1.

It may look surprising at first sight that in such an amorphous system, electrical conductivity should exist. Conductivity in crystals is a consequence of the periodicity of the lattice giving rise to delocalized electrons described by “Bloch functions”. (see Sect. 1.5), i.e. free electron wave functions extending over the full volume of the crystal. But in disordered solids, delocalized states can occur as well, resulting in electrical conductivity. In disordered systems, “delocalization” can happen as a result of an overlap of wavefunctions. Such a case is shown in Fig. 8.2, where the conductivity as a function of temperature is plotted for different concentrations of the dopant.

At sufficiently low concentrations ($\leq 10^{23} \text{ m}^{-3}$), the typical behavior of a semiconductor is seen, i.e. an exponential increase of conductivity with temperature up to the “exhaustion region” where all dopants have released their electron. The higher the concentration, the higher the temperature of exhaustion. The curve for a concentration as high as 10^{24} m^{-3} , however, looks completely different which can be ascribed to the overlap of the wavefunctions of the excess electrons of the dopant. A so-called “perturbed state band” is formed.

A similar situation, by the way, is encountered in a liquid metal like mercury.

As a consequence of the “near-range order”, the energy band structure of the amorphous semiconductor shows quite a similarity with the one of the crystalline material, but also some characteristic differences.

In Fig. 8.3a, the local variations of the band edge E_G due to local structural modifications are schematically shown.

As a consequence of the local variations of the band gap, shallow electron and hole traps form, giving rise to electronic states generally called “tail states”.

Fig. 8.2 Conductivity in $(\Omega\text{m})^{-1}$ of n-doped Ge for different doping levels

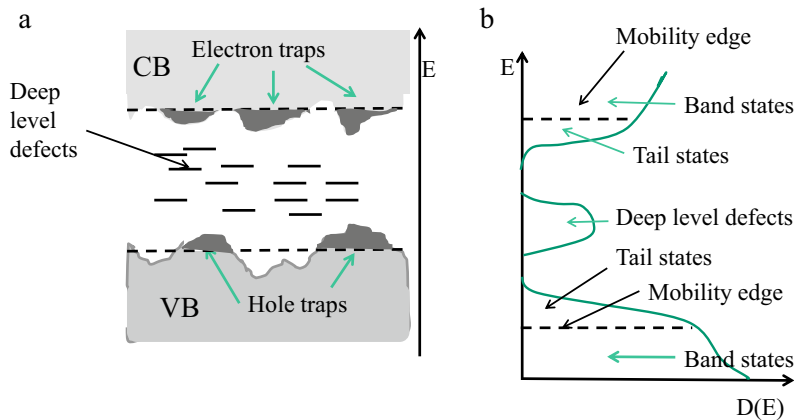
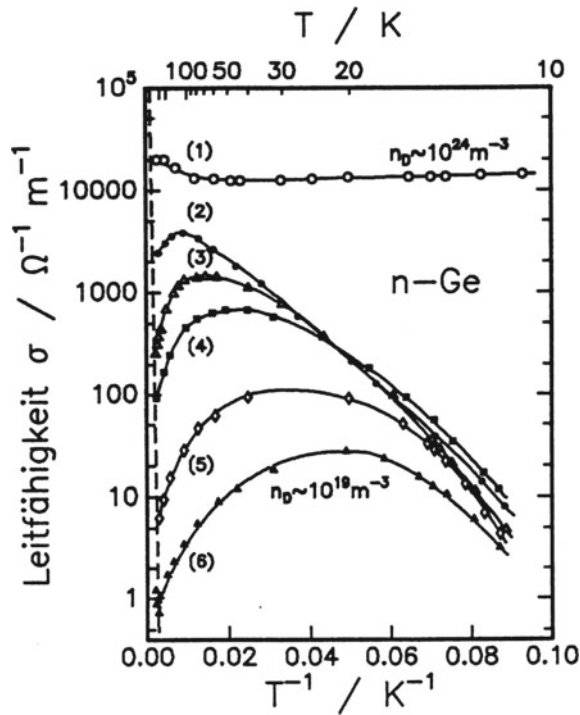


Fig. 8.3 **a** Local variations of the band edge **b** Density of states in an amorphous semiconductor.

b Density of states in an amorphous semiconductor

The “deep levels” in the bandgap shown in Fig. 8.3a mostly are caused by dangling bonds (Fig. 8.1).

Figure 8.3b shows the density of states corresponding to the situation of Fig. 8.3a.

The electrons in traps (called the “tail states” in Fig. 8.3b) result from the “softening” of the band edge of the conduction band in the amorphous state, as explained above. The electrons at sufficiently high energies (above the energy called “mobility edge” in the figure) can be considered as “delocalized band states”, like the electrons in the conduction band of a crystalline solid.

Crystal defects like dangling bonds give rise to completely new levels deep in the bandgap, the “deep defect levels”.

An analogous situation exists for the holes: Holes exist as tail states (holes in traps resulting from the local variations of the band edge of the valence band) and as delocalized band states. A very remarkable feature of the amorphous state is the “mobility edges” which separate the delocalized states from the tail states. This issue is discussed in the following physics comment

Physics comment: The mobility edge

At a superficial sight, one might think that—given the local variations of the bandgap—the notion of “bandgap” is no longer meaningful. Surprisingly, a precise quantitative meaning can be given to the notion, even in the amorphous state, based on a fact first shown by the British Physicist Sir Nevill Francis Mott¹. (The mobility edges have been already included in Fig. 8.3b). According to Mott, there is always a sharply defined energy above which the electrons in an amorphous solid are delocalized and below which they are localized. The corresponding statement holds for the holes. These specific energies are called “mobility edges”. It is a meaningful definition to call the separation of these mobility edges “energy gap” of that amorphous material.

The existence of a mobility edge can be visualized by a nice, intuitive picture. If we visualize the “energy landscape” of the electrons by a landscape of hills (the height of a point representing the energy), filled up to a certain level by water (the water representing the electrons, its surface the electrons at the Fermi energy E_F), there is always a water level below which the landscape can be considered as one of the hills with lakes in-between, and above which it has to be considered as a lake with islands. In the latter case, each point can be reached by boat without carrying it over land, any swimming object can be considered as “delocalized”. In the former case, the boat (or any swimming object) is “localized” on one or the other of the lakes, it is “localized” in its lake and cannot reach a spot on another lake without being carried over a hill.

¹ Nevill Francis Mott (1905–1996). British Physicist, Cavendish Professor of Physics at Cambridge University, Nobel Prize 1977

An interesting consequence of the existence of “electron traps” resulting from the local variation of the band edge is that the distinction between direct and indirect semiconductors is softened for amorphous semiconductors, another one is the existence of a different conduction mechanism, the so-called “hopping conduction”. Both issues are discussed in the following.

8.2 On the Distinction Between Direct and Indirect Semiconductors in Amorphous Semiconductors

The distinction between direct and indirect semiconductors is described in Sect. 1.7. In indirect semiconductors, the minimum of the conduction band and the maximum of the valence band are positioned at different k -values (i.e. different momenta of the electrons, s. Fig. 1.11). So radiative recombination of electrons and holes at the band edges would be a violation of momentum conservation.

Physics comment: Optical absorption and radiative recombination in an amorphous semiconductor

As shown in Fig. 8.3a, a situation different from the one in crystalline material exists in amorphous semiconductors. Due to the softening of the band edge, the electrons near the edge of the conduction band are trapped in local traps, as well as the holes in the valence band. These are the so-called tail states. So these are in a state of limited spatial extension Δx , which, according to Heisenberg’s uncertainty relation, means that the momentum is undefined to an extent of about $\Delta p \approx \hbar/\Delta x$. Due to this uncertainty of p , a contradiction to momentum conservation cannot be stated as long as this uncertainty is comparable to or larger than the difference between the momenta of the electrons and the holes involved in the recombination process. Whether this really happens is a question of the size Δx of the trap, the distribution of which, in turn, is a question of the degree of disorder in the amorphous material. To illustrate this idea quantitatively, we shall estimate in the following how small the spatial extension of the trap should be in order to produce a sufficiently large uncertainty of the momentum p . For that purpose, we consider the case of Si.

The lattice constant of Si is $a = 0.357$ nm. The next nearest neighbor distance is a quarter of the space diagonal of the unit cell (Fig. 1.2b): $(3)^{1/2} * 0.25 a = 0.43a$.

As mentioned in the discussion of Fig. 1.10, the X-point in Si corresponds to $k = \pi/a$. According to Fig. 1.10, the minimum of the conduction band is at $k \approx 0.85 * \pi/a$ and the maximum of the valence band is at $k = 0$, so the separation in k -space is

$$\Delta k = 0.85 * \pi/a \quad (8.1)$$

So, because of $(\hbar k) = p$, one obtains from the uncertainty relation for the uncertainty of the momentum

$$\Delta p = (\hbar)\Delta k = (\hbar)0.85 * \pi/a = 0.425 h/a \quad (8.2a)$$

Because, due to the uncertainty relation, Δp equals $h/\Delta x$ one obtains from (8.2a)

$$\Delta x = (h/0.425h) * a = 2, 4a \quad (8.2b)$$

Because the lattice constant a of Si is 0.543 nm and the nearest neighbor distance is 0.235 nm (see the context of Fig. 8.1), this is about 5.5 times the next nearest neighbor distance in Si.

This means: If the trap width Δx (i.e. the uncertainty of the position of the electron) is about this size, the uncertainty of the momentum is such that a violation of momentum conservation in radiative recombination can no longer be stated. It can be shown that this situation is achieved in moderately disordered amorphous material. If the trap is larger, we have some intermediate situation between “direct” and “indirect”.

The reflection given here is a typical application of the basic ideas of quantum mechanics. If the uncertainty of p is Δp , any statement which implies a higher accuracy of p than Δp is meaningless. So, if the trap width of the tail states is comparable or smaller than Δx given by (8.2b), a contradiction to momentum conservation in the recombination process could no longer be proclaimed. (Keep in mind: A small trap width implies large variations of the band edge on short distances, i.e. high disorder).

For the reason given above, the absorption coefficient of amorphous Si as a function of wavelength shows a behavior somehow between those of direct and indirect semiconductors as seen in Fig. 8.4.

Because in the direct semiconductors like GaAs, InP, CdTe or CdS, the optical transition from the valence band to the conduction band is allowed, the absorption coefficient steeply increases when the photon energy $h\nu$ surpasses the bandgap energy E_G . In the indirect semiconductors like Si and Ge, the slope of the absorption spectrum is comparatively shallow, as the figure shows. For amorphous Si (a-Si), however, the slope lies in between, depending on the degree of disorder. Furthermore, it is seen that, because of the more direct nature of a-Si, in the visible region of the spectrum, the absorption coefficient is much higher than that of crystalline Si (mind the logarithmic scale!). This has an important consequence for one of the major applications of a-Si: While a crystalline Si solar cell has to be fairly thick (about 100–200 μm) in order to quantitatively absorb the solar light, an a-Si solar cell can be much thinner.

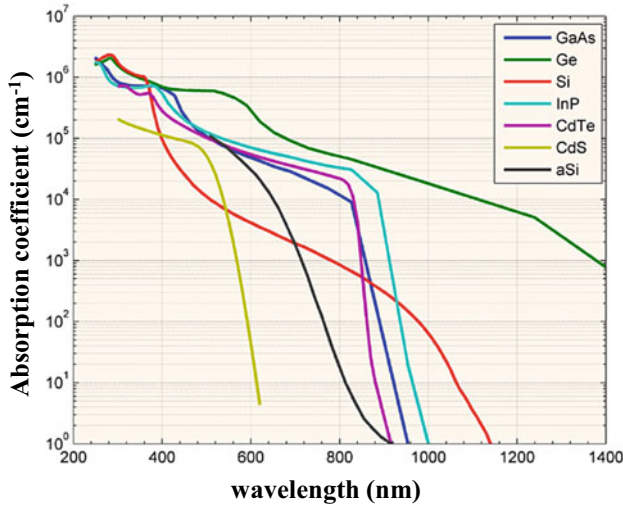


Fig. 8.4 Absorption coefficient as a function of wavelength for various semiconductors in comparison to the one of a-Si (Courtesy C.B. Honsberg and S.G. Bowden, “Absorption coefficient” Section 3.2 on www.pveducation.org2019)

Crystalline solar cells are “thick film solar cells”, a-Si solar cells are “thin film solar cells”, with film thicknesses of less than 10 nm; see also Sect. 10.5.

8.3 Variable Range Hopping Conduction

The main contribution to the conductivity of a-Si results from the electrons or holes in the band tails, i.e. the relevant electrons or holes are sitting in electron traps or hole traps, as shown in Fig. 8.3a. These traps, as explained above, result from the local variations of bonding lengths and bonding angles and thus are situated in a “four-dimensional energy landscape”, formed by the three space coordinates and the electron or hole binding energy associated with the corresponding spot.

To simplify the discussion, we refer to the following to charge transport via electrons, the one for holes would be treated correspondingly. If a voltage is applied, the electrons move across the material, and a current is produced. This movement results from hopping processes from one trapping site to the next. The electron jumps from one minimum in the “energy landscape” to a neighboring one via a tunneling process. The quantum mechanical tunneling probability depends (exponentially) on the separation of the two trapping sites and the height of the energy “hill” in between. So the probability of such a hopping process could be large even if the target site is far away, and if the hill in between is sufficiently flat. And it could be small even if the two trapping sites are close to each other, if the hill in between is sufficiently high. That is the reason, why we speak of “variable range hopping”. In analyzing this

kind of tunneling transport in a four-dimensional energy landscape, Mott arrived at a temperature dependence of conductivity in amorphous materials given by

$$\sigma = \sigma_0 \exp[(T/T_0)^{1/4}] \quad (8.2)$$

This characteristic exponential dependence on (T/T_0) is indeed found in many amorphous materials, in particular, a-Si. The parameter T_0 depends on the type and degree of disorder. So the conductivity of an amorphous semiconductor increases with T , as it does for a crystalline one, but—of course—due to a different mechanism.

8.4 The Problem of Doping in a-Si

The use of a semiconductor for electronic applications requires the availability of n- and p-doped material. However, for a long time, it was considered impossible to dope a-Si. As discussed in Chap. 3, n- or p-doping of a semiconductor is performed by the incorporation of impurities that give rise to shallow donors or acceptors, for e.g. P or B in Si. The problem in the case of amorphous Si arises from the high concentration of deep level defects (mainly dangling bonds). Imagine that a P-atom has been incorporated substitutionally on a-Si site. Then, instead of being released thermally to the conduction band, the additional electron would fall down to one of the many deep acceptors and thus this P-atom would go lost as a dopant. The corresponding process would hold for shallow acceptors like B. The “missing orbital” on the B atom would be filled by an electron from a deep level donor, not—as it should—from the valence band.

The solution was found by “passivating” the dangling bonds with Hydrogen (indicated in Fig. 8.1). By forming a complex with H atoms, the dangling bonds are no longer electrically active, i.e. no longer act as deep donors or acceptors². The straightforward way to incorporate the H into the growing a-Si film is by producing the film in a SiH_4 -discharge (Fig. 8.5).

In the gas discharge, the SiH_4 is decomposed and a film of amorphous Si with a high (10–20%) H-content is formed on the substrate. Whether this film is amorphous or polycrystalline is essentially a question of the substrate temperature. At sufficiently low temperatures (ca. $< 300^\circ\text{C}$), the film will be amorphous, at higher temperatures, it will be polycrystalline. This is, by the way, a general rule: The formation of a crystal requires a minimum surface mobility of the deposited atoms in order to “find” their proper position in the crystal lattice. If the surface mobility is too small, an amorphous film is formed because the atoms essentially stick at the site where they are deposited by chance on the growing film.

The growing a-Si:H film can be doped p-type or n-type by either injecting B_2H_6 - (diborane) or PH_3 -gas (phosphine) into the plasma. The B- or P-atoms will then be incorporated substitutionally on a-Si site and thus form a shallow acceptor or donor,

² Spear and Le Comber [1]

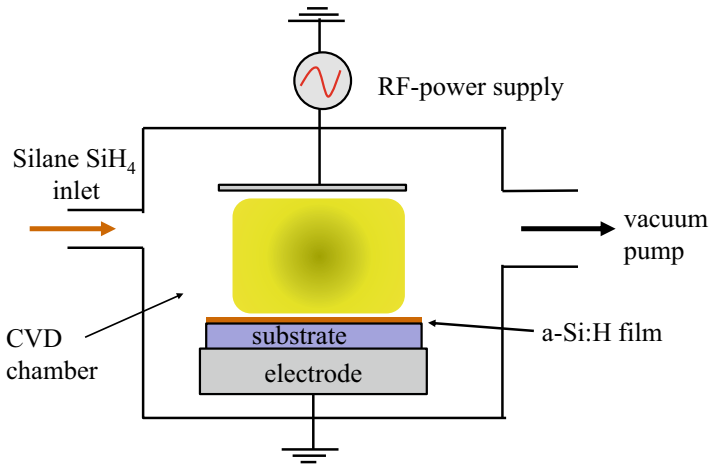


Fig. 8.5 Preparation of a-Si:H by decomposition of Silane in a plasma discharge

respectively. On this basis, electronic devices like photodiodes, solar cells, and field effect transistors can be produced.

Physics comment: Passivation of donors and acceptors by hydrogen

The passivation of a donor or acceptor by hydrogen is a fairly general phenomenon in semiconductor physics. If the donor or acceptor reacts with the H atom, the resulting H-donor- (or H-acceptor-) complex will no longer be electrically active.

This passivation effect might be desired as in the case of amorphous Si, it may be undesired and unintentional if H is present as an undesired or unnoticed impurity. For example, it is well known that the doping of GaAs with Si (Chap. 3.4) is hindered by the presence of H^3 . A famous (or rather infamous) example is the p-doping of GaN with Mg in the context of developing the blue light emitting diode, which for a long time was considered impossible due to an unnoticed H-contamination (Sect. 11.6).

³ Stavola [2]

8.5 Applications of a-Si:H

Crystalline Si is a superb semiconductor in terms of electronic properties as well as availability. Why then bother with a-Si:H?

a-Si:H can easily be deposited in large areas (Fig. 8.5) in a plasma discharge. So large area devices are the dominant applications. The first outstanding one was the X-ray flat panel detector (Fig. 8.6).

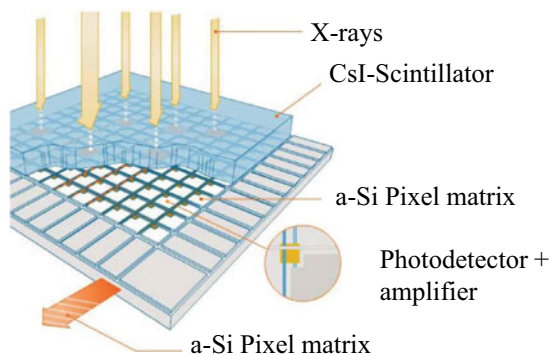
The flat panel X-ray detector replaces the traditional X-ray film. In the scintillator, the X-ray photons are transformed into photons. In the subsequent array of a-Si:H photodiodes, the light produces photocurrents. The photocurrent in each pixel, being essentially proportional to the received X-ray dose, is electronically processed in the digital board array, and the signals are given to the computer which constructs the image. In this way, digitized X-ray images are produced ready for modern ways of image processing and pattern recognition.

Typical size of the panel detector would be $40 \times 40 \text{ cm}^2$, with a pixel size of the order of $100 \text{ }\mu\text{m}$.

This application of a-Si might only be of passing interest. In view of the superior performance of electronics based on crystalline Si in terms of sensitivity and noise, developments are underway for flat panel devices based on crystalline Si electronics. Because monocrystalline Si layers of that size are not available, such a detector would consist of a system of Si-tiles. The black lines resulting from the gaps between the tiles would be eliminated by image processing.

The other large area device to be mentioned, of course, is the solar cell. Solar cells of amorphous Si (Sect. 10.5) have efficiencies of typically 8%, being much lower than the ones of crystalline solar cells. But “efficiency is not everything” in photovoltaics. Cost is also an important issue! Because the high costs of crystal growth do not enter, and because a-Si:H solar cells are thin film solar cells (Sect. 8.2), the material costs are considerably smaller. So for low cost applications like pocket calculators, cameras, etc. amorphous solar cells are common commercial devices based on a-Si. Additional interesting applications could arise from the fact that thin films a-Si:H

Fig. 8.6 Flat panel X-ray detector (Courtesy Siemens Healthineers AG)



solar cells can be produced on a flexible substrate like tissues and polymer foils, with no limits to imagination!

8.6 Questions and Answers

- Question 1: What is the difference between an amorphous material and a liquid?
Question 2: What are tail states?
Question 3: What is the dominant conduction mechanism in amorphous materials like a-Si?
Question 4: Why is the distinction between direct and indirect semiconductors less pronounced for amorphous materials?
Question 5: Given the “softening” of the band edge, is it still meaningful to speak of a bandgap?
Question 6: What is the problem with doping of an amorphous semiconductor?
Question 7: How is this problem of doping overcome in a-Si?
Question 8: How is the H incorporated in a-Si:H?
Question 9: What are typical applications of a-Si:H?

Answers

- Answer 1: In an amorphous material there is still a short-range order.
Answer 2: Tail states are the localized states originating from the “softening” of the Band edge. This softening in turn is a consequence of the local variations of the structure. These localized states generally form shallow traps.
Answer 3: Variable range hopping from the trap.
Answer 4: Because of the localization of the electrons and holes near the band edge, their momentum is no longer sharply defined due to the uncertainty relation. Hence, the contradiction to momentum conservation in the recombination or the excitation process is no longer valid.
Answer 5: Yes. According to Mott, there is always a sharply defined energy separating the localized states from the delocalized band states, called the mobility edge. The separation of the mobility edges of electrons and holes can reasonably be defined as the bandgap of the amorphous material.
Answer 6: The large concentration of deep defect levels
Answer 7: By introducing a large concentration of H. The H atoms “passivate” the deep defect centers produced by the dangling bonds. The resulting a-Si:H can be doped like crystalline Si, e.g. B (acceptors) or P (donors).
Answer 8: By depositing the a-Si film out of a plasma gas discharge of SiH₄
Answer 9: Large area devices like the flat panel X-ray detector or low-cost solar cells

References

1. W.E. Spear, P.G. Le Comber, Solid State Comm. **17**(9), 1193–1196
2. M. Stavola: Hydrogen Passivation in semiconductors, ACTA PHYSICA POLONICA A, Vol. 82 (1992) No 4 and references therein

Part II

Devices

Chapter 9

The pn-Junction



Abstract The pn-junction consists of the interface region between a p-type and an n-type semiconductor. It is the basic element of many electronic devices. The behavior is determined by two types of currents: The “diffusion current”, following the concentration gradient, of electrons and holes, respectively, and the field current, driven by the electric field originating from the potential difference that develops at the interface between p- and n-type region. If an external voltage is applied to the pn-junction, it adds to this potential difference, lowering or increasing the “potential barrier” between p- and n-region. As a result, a rectifying effect occurs, providing the most elementary and historically first important application of the pn-junction, the pn-rectifying diode. In the “forward direction” the current increases exponentially with the voltage, in the “reverse direction” the current approaches a limit, which is identical to the field current. This reverse current strongly depends on the bandgap, so it is orders of magnitudes lower in Silicon as (for example) in Germanium. If the reverse voltage exceeds a critical value, a “breakthrough” occurs, either the “avalanche breakthrough” or the “Zener breakthrough”, which would destroy the diode. This effect makes many electronic devices so sensitive to voltage spikes, sparks, etc.

The chapter “pn-junction” can be considered an introduction to the field of electronic devices. Many important devices like the rectifier diode, the solar cell, the light emitting diode, the bipolar transistor, and others are based on the functions realized in a pn-junction.

9.1 pn-Junction Without External Voltage

Figure 9.1 shows the simple scheme of a pn-junction: A semiconductor with an interface between an n-type and a p-type region.

We shall first consider the situation where no external voltage is applied. In this case, a voltage jump ΔV develops at the interface in thermal equilibrium. Qualitatively this can be understood as follows (Fig. 9.2a,b).

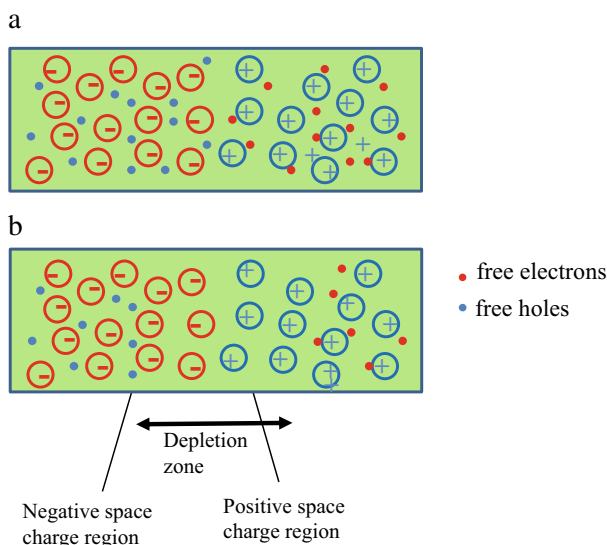
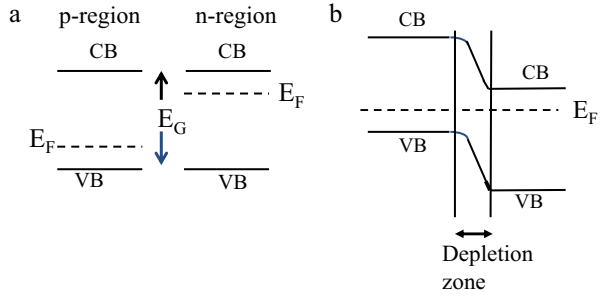
Fig. 9.1 The pn-junction**Fig. 9.2** **a** Original charge distribution before diffusion, left is the p-type region, right the n-type region. **b** Charge distribution after diffusion

Figure 9.2a shows the original situation with the negatively charged free electrons (red dots) in the n-type section (the positive ion cores are shown as blue circles) and the positive free holes (blue dots) in the p-type section (the negative ion cores are shown as red circles). Because the two regions are in contact, the free electrons will diffuse from right to left, given the fact that the concentration of free electrons is, by definition, higher in the n-type region, so the free electrons follow their concentration gradient. For the analogue reason, the free holes will diffuse from left to right. Near the interface, the diffusing electrons and holes will recombine. So the positive ion cores at the right side of the interface will form a positively charged space charge region, and the negative ion cores at the left side will form a negatively charged space charge region (Fig. 9.2b).

As a consequence, after a stationary state has been achieved, the negative space charge on the left side will repel the free electrons from the right side and thus stop the diffusion of the free electrons, and the positive space charge will stop the diffusion of the holes. So the diffusion process comes to an end. To put it more formally: A potential barrier ΔV develops at the interface between p- and n-region which inhibits further diffusion. This diffusion barrier ΔV is shown in Fig. 9.3, which presents the situation in terms of the energy band diagram.

Figure 9.3a shows the situation when the p- and n-type sections are not connected.

Fig. 9.3 **a** E-bands and Fermi-Energies without contact. **b** n- and p-region in contact



In the n-type region, the Fermi level E_F is close to the conduction band, in the p-type region, it is close to the valence band. Figure 9.3b presents the situation when the two sections are in contact with thermal equilibrium.

“Thermal equilibrium” in a system of fermions by definition means that the Fermi level is the same everywhere in the sample. From the figure, it is immediately obvious that as a consequence of this requirement, a “band bending” must occur resulting in a jump ΔV of the potential around the interface. The fact that the Fermi level continuously approaches the valence band going from right to left in the interface regions reflects the fact that we have a continuous reduction of the electron concentration from the originally high value in the n-type region (at the very far right) to the smaller value in the p-type region. A smaller value of the electron concentration means higher separation of E_F from the conduction band and—correspondingly—smaller separation from the valence band, as shown in the figure. In the following, we shall calculate the potential jump ΔV in the case of thermal equilibrium (with no external voltage applied). For this purpose, we first consider the electron currents across the pn-interface.

According to Fick’s first diffusion law, a particle current $D_n(dn/dx)$ (number of electrons per cm^2 and sec) of electrons flows from right to left (following the gradient of particle concentration), corresponding to an electrical current i_n^{Diff}

$$eD_n(dn/dx) = i_n^{\text{Diff}} \quad (9.1)$$

where D_n is the diffusion constant of free electrons in the respective semiconductor material. i_n^{Diff} is called the density of the “diffusion current”. As explained above, a potential jump ΔV builds up in the interface region pulling electrons in the opposite direction, i.e. from left to right. This is the “field current” i_n^{Field} . Because of $i = \sigma E$ (Eq. 2.5) and $\sigma = ne\mu$ (Eq. 2.13), the field current amounts to

$$i_n^{\text{Field}} = en\mu_n E$$

or, because of $E = -dV/dx$

$$i_n^{\text{Field}} = -en\mu_n(dV/dx) \quad (9.2)$$

In the stationary states, the two currents must compensate each other, that is

$$en\mu_n(dV/dx) = eD_n(dn/dx) \quad (9.3)$$

At this point, the following fact should be stated. Because in the interface region between p- and n- section, the charge carriers recombine, forming the depletion zone, new charge carriers have permanently to be created in order to maintain thermal equilibrium. The process providing these charge carriers is thermal excitation from the valence band to the conduction band, either in the region of the depletion zone, or in the neutral region. When the depletion zone is sufficiently narrow, this happens mainly in the neutral regions.

We solve the differential Eq. (9.3) by “separation of variables”, i.e. by putting the expressions containing V on one side and the expressions containing n on the other side

$$eD_n \, dn/n = e\mu_n dV \quad (9.4)$$

By integrating both sides from the “very left side” (where n equals n_p , i.e. the electron concentration in the p-type material) to the “very right side” (where n equals n_n , i.e. the electron concentration in the n-type region), we obtain

$$eD_n \ln(n_n/n_p) = e\mu_n \Delta V$$

and thus for the potential jump ΔV at the pn-interface

$$\Delta V = (D_n/\mu_n) \ln(n_n/n_p) \quad (9.5)$$

Let us assume that the donor concentration in the n-type material is given as N_D , and the acceptor concentration in the p-type part of the pn-junction as N_A . We then replace n_n and n_p in this expression with these known quantities in the following way: According to the law of mass action in semiconductors (Sect. 4.4, Eq. 4.10)

$$n_p p_p = n_i^2$$

Here n_p is the electron concentration in the p-type material, p_p the hole concentration in the p-type material of the pn-junction and n_i the intrinsic carrier concentration at the temperature under consideration. So $n_p = n_i^2/p_p$ and consequently

$$\Delta V = (D_n/\mu_n) \ln(n_n/n_p) = (D_n/\mu_n) \ln(n_n p_p / n_i^2)$$

Furthermore, we make use of the fact that usually at room temperature (and above) shallow donors and acceptors typically are fully ionized (for the notion of shallow donors and acceptors; see Sects. 3.3 and 3.4), thus the concentration of electrons n_n in the n-region equals N_D , the donor concentration, and the concentration of the holes in the p-region p_p equals N_A , the acceptor concentration. Thus, we finally obtain

$$\Delta V = (D_n / \mu_n) \ln(N_D N_A / n_i^2) \quad (9.6a)$$

Exactly the corresponding consideration can now be performed with the free holes: A diffusion current is going from left to right, a field current from right to left, and in a stationary situation both have to compensate each other, resulting in

$$\Delta V = (D_h / \mu_h) \ln(N_D N_A / n_i^2) \quad (9.6b)$$

Now it is obvious that the two potential barriers ΔV in Eqs. (9.6a) and (9.6b) have to be equal, because they stand for the same potential barrier, so we can conclude that D_n / μ_n must be equal to D_h / μ_h

$$D_n / \mu_n = D_h / \mu_h$$

The above derivation can be considered as a derivation of this equality, but in the following, we shall show that not only this equality holds, but moreover, the ratio of D/μ is equal to kT/e , where T is the temperature. This gives the final result

$$\Delta V = (kT/e) \ln(N_D N_A / n_i^2) \quad (9.7)$$

The equation $D/\mu = kT/e$ is called the Einstein relation. It first appeared in a paper by Albert Einstein on the Brownian motion, and it will be derived for our case of freely moving electrons or holes in a solid in the following.

Physics comment: The Einstein relation.

We go back to Eq. 9.3

$$en\mu_n(dV/dx) = eD_n(dn/dx) \quad (9.3)$$

Given the fact that the electrostatic energy U of a charged particle in an electric field is given by $U = -eV(x)$, and the basic law of statistical mechanics that in thermal equilibrium $n(x)$ is given by the Boltzmann distribution, we write

$$n(x) = n_0 e^{-U/kT} = n_0 e^{eV(x)/kT} \quad (9.8)$$

(in the Physics comment “Boltzmann- or Fermi–Dirac-distribution” in Sect. 4.1 it has been discussed, that the Boltzmann distribution can be applied as long as $n \ll N_C$, the “density of states of the conduction band”, the latter being $2.8 \cdot 10^{19} \text{ cm}^{-3}$ at room temperature (see Table 4.1), which is an exceptionally large carrier density).

From Eq. (9.8), we calculate

$$dn/dx = n_0 e^{eV/kT} * (e/kT) dV/dx = n(x)(e/kT) dV/dx \quad (9.9)$$

and insert this on the right side of (9.3)

$$en\mu_n(dV/dx) = eD_n n(x)(e/kT) dV/dx$$

hence

$$D_n/\mu_n = kT/e \quad (9.10)$$

The same derivation, of course, could be given for the holes, resulting in $D_h/\mu_h = kT/e$.

Looking back at this derivation, one can state that the Einstein relation follows from the idea that the two currents (diffusion and field current) have to result in a thermal equilibrium situation described by the Boltzmann distribution for both types of charge carriers.

The Einstein relation connects two properties of the free charge carriers in a semiconductor, the diffusion constant and the mobility. A closer look at the basic mechanism of diffusion (with diffusion constant D) and carrier transport (with mobility μ) immediately tells that the two phenomena are directly related to the scattering process of the carriers (for mobility see Sect. 2.1), resulting in a close relation between both, as expressed by the Einstein relation.

Numerical example

In Table 4.1, the intrinsic carrier concentration of Si at room temperature is given as ca. $n_i = 1.45 \times 10^{10} \text{ cm}^{-3}$. Remembering that kT/e is 0,026V at room temperature, and taking $N_D = 10^{18} \text{ cm}^{-3}$ and $N_A = 10^{15} \text{ cm}^{-3}$ for our numerical example, we obtain from Eq. (9.7) for a Si pn-junction under these conditions $\Delta V = 0,76 \text{ eV}$.

On the polarity of the pn-junction without external voltage

As explained at the beginning of this section, in the pn-junction without external voltage, the (negative) electrons will leave the n-type region by diffusion (“diffusion current”), thus **the n-type side of the pn-junction will develop a positive potential, and it will be the anode**. The (positive) holes will leave the p-side region by diffusion and thus leave behind a negative potential, thus **the p-type side will form the cathode** (see Fig. 9.1).

9.2 pn-Junction with External Voltage

When an external voltage U is applied to the pn-junction (Fig. 9.4a).

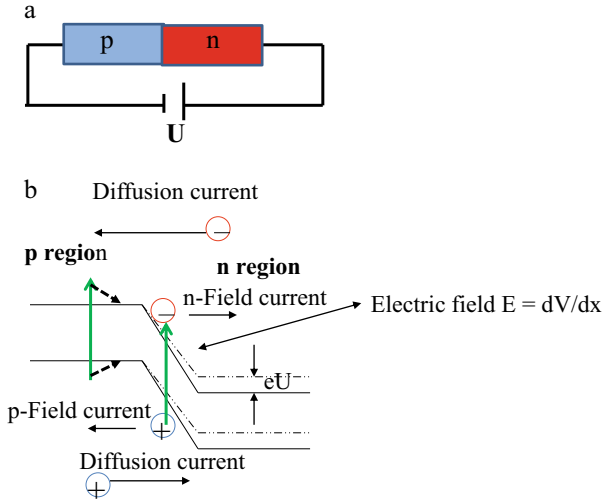


Fig. 9.4 **a** pn-junction with external voltage U. **b** Change of the barrier between n- and p-region by an external voltage U

the energy barrier between the p- and n-section is changed by eU (Fig. 9.4b). How does this affect the two currents, the diffusion current and the field current?

Diffusion current: Let us again consider the electrons first. As compared to the case of $U = 0$, the diffusion current now has to surpass an energy barrier changed by eU . According to Boltzmann's statistics, this changes the number of electrons which can overcome the barrier by $e^{eU/kT}$ (Boltzmann factor. Once again we make use of the fact that Boltzmann statistics can be used up to very large carrier densities, Sect. 4.1). So the diffusion current now is

$$i_n(U) = i_n^{\text{Diff}} \exp(eU/kT).$$

where i_n^{Diff} is the diffusion current without external voltage ($U = 0$).

Field current: A basic point to understand is that the field current is not affected by a change in the barrier height. This can be easily concluded from Fig. 9.4b. The electrons providing the field current created in the depletion zone or arriving there by diffusion from the neutral regions just fall “downhills” toward the n-region, no matter how steep the slope and how large the “falling height”, i.e. the energy barrier, is. This means that the electronic contribution to the field current is independent of the barrier and keeps its value i_n^{field} for any value of U. So the total current provided by the electrons is

$$i_n(U) = i_n^{\text{Diff}} \exp(eU/kT) - i_n^{\text{field}}$$

Because of $i_n^{\text{Diff}}_0 = i_n^{\text{field}}_0 = i_{n0}$ (at $U = 0$, the two currents compensate each other, as stated already in the context of Eq. 9.3), one gets

$$i_n(U) = i_{n0}[\exp(eU/kT) - 1] \quad (9.11a)$$

By the same arguments, we obtain for the total hole current

$$i_p(U) = i_{p0}[\exp(eU/kT) - 1] \quad (9.11b)$$

and thus for the total current $i_p(U) + i_n(U) = i(U)$ through the pn-junction

$$i(U) = i_0[\exp(eU/kT) - 1] \quad (9.12)$$

with $i_0 = i_{n0} + i_{p0}$. This is the current versus voltage characteristic of the pn-junction as shown graphically in Fig. 9.5a

For positive U , the current essentially increases exponentially. Let us compare the two contributions $i_0 \exp(eU/kT)$ and $-i_0$ in Eq. (9.12). Remembering that kT at room temperature is 0,026 eV, we obtain for a voltage as low as + 0,5 V a value for $\exp(eU/kT) \cong e^{19} \cong 10^8$, so the current–voltage characteristics for $U > 0$ (which is called the “forward direction”) is completely dominated by the exponential $\exp(eU/kT)$. This means that the currents in forward direction ($U > 0$) can be many orders of magnitudes larger than the ones for $U < 0$ (which is called the backward- or reverse direction). The latter approaches the value $-i_0$ for large negative values of U . For this reason, the pn-diode can be considered as a rectifying device, which is an important application.

Figure 9.5b shows the polarity of the diode in forward and reverse directions.

In the forward direction, the plus-electrode is connected to the p-doped side, in the reverse direction it is the other way round. This is obvious because, in the forward

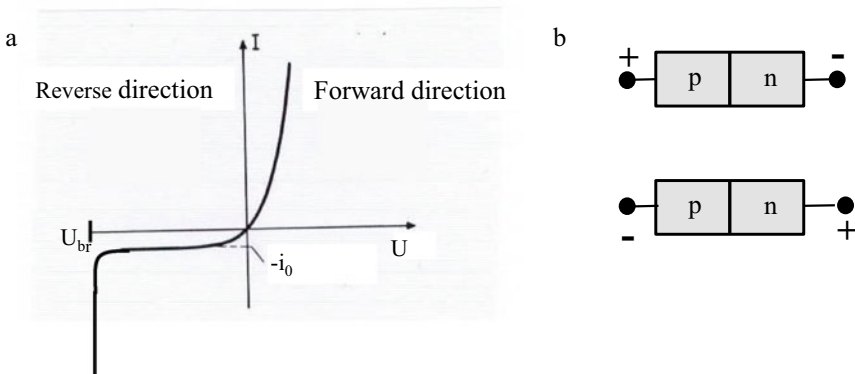


Fig. 9.5 **a** $i(U)$ -characteristic of a pn-junction. **b** Polarity of the rectifier diode in forward and in backward direction

direction, the holes and the electrons have to be pushed into the depletion layer, which in this way is narrowed, in the reverse direction, they are pulled out of the depletion layer, which then is being widened.

Historically, the pn-junction rectifier replaced the diode vacuum tube and other types of rectifiers as soon as the proper semiconductor materials were available, which in the early time of semiconductor technology was the Germanium, then followed by Si. A major factor why Ge readily was replaced by Si, indeed, was due to the fact that Si provided a much better rectifying diode as compared to Ge. Let us at this point compare the $i(U)$ characteristics of different materials.

For that purpose, we have to discuss the factor i_0 in Eq. 9.12. i_0 is identical to the field current, so we have to consider the materials parameter which determines the field current. The field current, as discussed in the context of Eq. 9.12, is independent of the external voltage U , so we might as well consider this issue for the case $U = 0$ treated in Sect. 9.1.

As explained there, the diffusion current transports electrons from the n-side to the p-side and holes from the p-side to the n-side. The space charges built up in that way (see Fig. 9.2b) produce a “band bending”, i.e. a potential jump which drives the electrons back from the p-side to the n-side, producing a current i_n^{field} , which corresponds to the contribution of the electrons to the field current, and the holes from the n-side to the p-side, producing a current i_p^{field} , which corresponds to the contribution of the holes to the field current. The total field current i_0 then is the sum of the two, i.e.

$$i_0 = i_n^{\text{field}} + i_p^{\text{field}} \quad (9.13a)$$

What is the reservoir which supplies the carriers of that field current?

As seen in Fig. 9.4b, the source of the electrons’ contribution to the field current are the electrons in the p-section of the diode, i.e. the so-called “minority carriers” in the p-section with concentration n_{p0} (the index 0 refers to the thermal equilibrium situation). (It is assumed at this point that the depletion layer is so thin that the thermal excitation of carriers within the depletion zone does not significantly contribute to the field current). So the contribution of the electrons to the field current is proportional to the minority carrier concentration in the p-doped section.

$$i_n^{\text{field}} \sim n_{p0} \quad (9.14a)$$

The source of the holes’ contribution to the field current is the holes in the n-section of the diode, i.e. the “minority carriers” in the n-section with concentration p_{n0} . So

$$i_p^{\text{field}} \sim p_{n0} \quad (9.14b)$$

Now use is made of the law of mass action, applied to the p-section of the diode (Eq. 4.10, with n_i being given by Eq. 4.9)

$$n_{p0}p_{p0} = n_i^2 \quad (9.15a)$$

As discussed in the context of Eq. 9.6a, p_{p0} can well be approached by N_A (N_A being the acceptor concentration on the p-section), we obtain

$$n_{p0} = n_i^2/N_A \quad (9.15b)$$

In the same way, by applying the law of mass action to the p-side of the diode, one gets

$$p_{n0} = n_i^2/N_D \quad (9.15c)$$

N_D being the donor concentration on the n-side of the diode.

By comparing Eqs. 9.15b, c with Eqs. 9.14a, b, one sees that the factor i_0 (which equals the total field current i^{field}) is given by

$$i_0 = i_n^{\text{field}} + i_p^{\text{field}} \sim n_i^2 = K n_i^2 \quad (9.16a)$$

where we have taken all contributing factors like N_A , N_D , diffusion lengths, etc. into the proportionality constant K .

So, other things being approximately equal, the field current i^{field} depends essentially on the intrinsic carrier concentration n_i , which according to Eq. 4.9 is given by

$$n_i^2 = N_C N_V \exp(-E_G/kT) \quad (9.16b)$$

Because n_i^2 is about 6 orders of magnitude larger for Ge than for Si at room temperature (Table 4.1), the reverse current of a Ge rectifying diode is many orders of magnitude larger in Ge than in Si, which was correctly claimed as the major reason for replacing Ge by Si as soon as the latter was available at sufficient quality.

We now switch to a comparison of pn-diodes of different materials in the forward direction.

By inserting n_i^2 from Eq. 9.16b into Eq. 9.16a, one gets

$$i_0 = K n_i^2 = K N_C N_V \exp(-E_G/kT) \quad (9.16c)$$

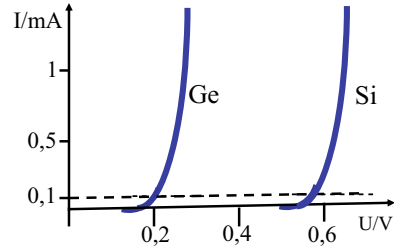
and finally, by putting this into Eq. 9.12, we get for the current–voltage characteristic

$$i(U) = K N_C N_V \exp(-E_G/kT) [\exp(eU/kT) - 1]$$

It was already pointed out in the context of Eq. 9.12 and Fig. 9.5a that, because $kT/e = 0,026$ eV at room temperature, in forward direction, for e.g. $U > 0,5$ V; this expression is completely dominated by the exponential $\exp(eU/kT)$ and thus can be written as.

$$i(U) \cong K N_C N_V \exp(-E_G/kT) \exp(eU/kT), \text{ or}$$

Fig. 9.6 $I(U)$ characteristics of Ge and Si in comparison



$$i(U) = K N_C N_V [\exp(U - E_G)/kT] \quad (9.17)$$

The result Eq. 9.17 tells us that in the forward direction, $i(U)$ looks essentially identical for different materials if U is counted from E_G on. For illustration, the $I(U)$ characteristics are shown for Ge and Si in comparison in Fig. 9.6.

Because the product $N_C N_V$ is similar in Ge and Si (see Table 4.1), so, other things being equal, the $i(U)$ plot for Si is just shifted by about 0.4 eV, relative to the one of Ge, corresponding approximately to the difference between the two bandgaps ($E_G(\text{Si}) - E_G(\text{Ge}) = 1.12 \text{ eV} - 0.66 \text{ eV} = 0.46 \text{ eV}$).

9.3 Breakthrough Voltage

If the voltage in the reverse direction is made too large (see Fig. 9.5), a breakthrough of the current occurs, resulting typically in the destruction of the diode, if the current is not limited by external resistance. There are two possible mechanisms of such a breakthrough: Avalanche breakthrough (Fig. 9.7a) and Zener- or tunneling-breakthrough (Fig. 9.7b).

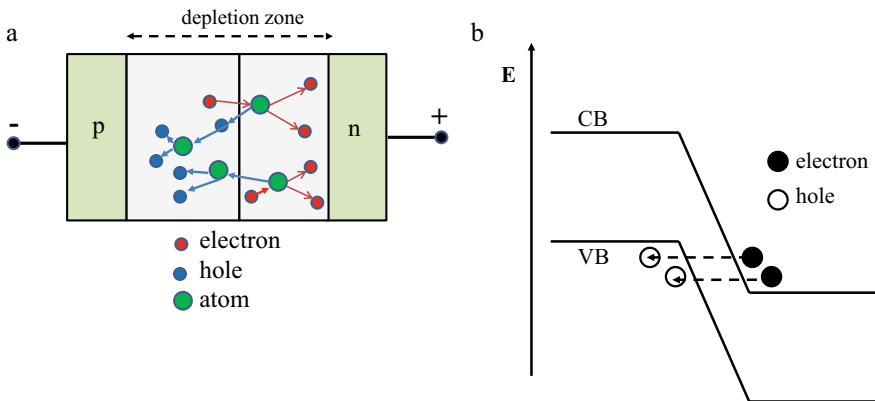
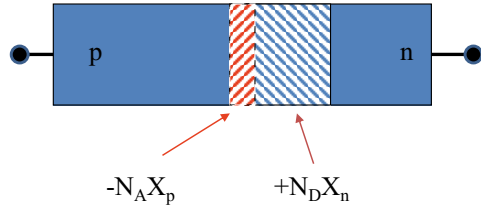


Fig. 9.7 a) Avalanche breakthrough b) Zener- or Tunneling-breakthrough

Fig. 9.8 Extensions X_n and X_p of the depletion layers on the n- and p-side



In the case of the avalanche breakthrough, the electrons are accelerated by the applied voltage such that they can ionize the atoms and thus produce further electrons, which, in turn, ionize further atoms, etc. In the case of the Zener breakthrough, the Fermi levels of n- and p-region are shifted to an extent that the occupied states of the n-region coincide in energy with empty states of the p-region, so electrons will jump from n- to p- region via a quantum mechanical tunneling process.

It can be mentioned at this point that these kinds of catastrophic events like avalanche- and Zener-process make semiconductor devices rather sensitive toward destruction by sparks, voltage spikes, etc.

9.4 Thickness of Depletion Layers

Figure 9.8 shows a pn-junction together with the depletion layer of thickness X_n on the n- and thickness X_p on the p-side.

It is immediately evident that for the reason of charge conservation (the total charge remains zero under the formation of the depletion layers), the equation.

$N_d X_d + N_a X_a = 0$ holds (for notations see Fig. 9.8), or

$$-X_d/X_a = N_a/N_d \quad (9.13)$$

So the thickness of the depletion layers on both sides of the pn-interface is inversely proportional to the respective dopant concentrations. This point is essential in many cases of device set-ups; see, e.g. in the set-up of the Si solar cell or the bipolar transistor.

9.5 Questions and Answers

- Question 1: Which are the two currents compensating each other in a pn-junction without external voltage?
- Question 2: What is the origin of the charge carriers producing the field current?
- Question 3: Which one of the two currents is changed when an external voltage is applied?

- Question 4: Give the diode characteristic $i(U)$!
- Question 5: What is the physical significance of i_0 ?
- Question 6: Why is the reverse current in Germanium larger than the one in Silicon?
- Question 7: What are the possible reasons for the breakthrough of the diode in the reverse direction?
- Question 8: What is the Einstein relation?

Answers

- Answer 1: The diffusion current, driven by the concentration gradient, and the field current driven by the electric field built up by the diffusion current.
- Answer 2: The field current is produced by the minority carriers in the n- and p-type regions.
- Answer 3: The diffusion current, because an additional potential barrier is created which has to be overcome, changing the diffusion current by the Boltzmann factor $\exp(eU/kT)$, and leaving the field current unchanged.
- Answer 4: $i(U) = i_0[\exp(eU/kT) - 1]$.
- Answer 5: It's the field current.
- Answer 6: The reverse current equals the field current. The field current is proportional to n_i^2 , and n_i^2 is orders of magnitude larger in Ge because of its smaller bandgap as compared to Si.
- Answer 7: Avalanche breakthrough and Zener breakthrough.
- Answer 8: $D/\mu = kT/e$.

Chapter 10

Solar Cells



Abstract The energy demand of our earth could easily be provided by solar energy. In a solar cell, the energy of the sunlight is converted into electrical energy due to the photovoltaic effect. A solar photon is absorbed in, e.g. Silicon producing an electron in the conduction band and a hole in the valence band, which are separated in the electrical field of a pn-junction, thus producing a current. Given the spectrum of the sun, there is an optimum bandgap for a solar cell material: If the bandgap is larger than the photon energy, the photon is (almost) not absorbed, if it is smaller, part of the photon energy is wasted by thermalization. The bandgap of Si is close to that optimum. For this reason and for reasons of price and availability, Si is the most widely used solar cell material. Because Si is an indirect semiconductor, a layer of about 200 μm is required in order to quantitatively absorb the sunlight, resulting in a “thick film solar cell”. Direct semiconductors like GaAs, CdTe, and a-Si:H (see Sect. 8.2) are used for “thin film solar cells” of a few nanometers thickness. For the Si solar cell, high-quality material is required to avoid loss of efficiency due to recombination of the electrons and holes produced by the sunlight. To reduce costs, multi-crystalline material is used, which is produced on a large scale by block casting and subsequent “directional solidification”. Research is in progress to avoid the expensive Siemens process for purification. Recorded solar cell efficiency for multi-crystalline Si is now near 23%, and for monocrystalline Si, it is near 27%. In 2019, 7.9% of the electrical power demand in Germany was provided by solar energy.

10.1 The Sun, a Gigantic Energy Source

Among the possible sources of regenerative energies, solar energy is considered the one with the highest potential. Some numbers support this: the total power hitting our earth from the sun is about 120.000 Terawatt $= 1.2 \cdot 10^{17}$ W. The world's total energy consumption at present is about 13 TW. So only a tiny fraction of about 10^{-4} would have to be harvested to provide the world's energy demand.

Physics comment: Total energy production of the sun

It might be interesting to consider at this point the incredible energy production of the sun as a whole: the distance between sun-earth is about $R = 150 \text{ Mio km} = 1.5 \times 10^{11} \text{ m}$, so the area of a sphere with that radius is $4\pi R^2 = 28.3 \times 10^{22} \text{ m}^2$. We shall see later (Table 10.1) that 1353 W/m^2 of solar power is received at the earth outside the atmosphere. So the total amount of emitted power is

$$P_{\text{Sun}} = 1353 \times 28.3 \times 10^{22}, W = 3.83 \times 10^{26} \text{ W}$$

This power is produced by burning hydrogen (H) to Helium (He) accompanied by a mass loss according to the energy/mass relation of Einstein

$$M = P_{\text{sun}}/c^2 = 0.43 \times 10^{10} \text{ kg/s or } 4.3 \text{ Mio tons per s.}$$

Isn't it remarkable that the sun, with that mass loss per second, holds on for billions and billions of years!

A few more numbers with regard to the solar energy potential (a rough estimate): 1 m^2 in central Europe is hit by solar energy of about 0.85 kW/m^2 when the sun shines (Fig. 10.4, table therein). In Germany, we have about 1800 h of sunshine per year. This gives total energy of 1530 kWh/m^2 per year. With an efficiency of 20%, this results in electrical energy of about 300 kWh/m^2 per year.

Let us compare this to a big power plant producing $1\text{GW} = 10^6 \text{ kW}$ electrical energy, running 8000 h per year. It will thus produce $8 \times 10^9 \text{ kWh}$ per year. How much area A covered by solar cells would be required to produce the same amount?

$$300 \text{ kWh/m}^2 \times A = 8000 \times 10^6 \text{ kWh}$$

$$A = 8000/300 \times 10^6 \text{ m}^2 = 27 \text{ km}^2$$

Evidently it is a considerable piece of land required to replace a large power plant.

10.2 The Solar Cell, a pn-Junction Device

The way to transform directly solar energy into electricity is via solar cells, a device based on the **photovoltaic effect** (Fig. 10.1a).

When a pn-junction (without external voltage) is hit by a light beam, a current is observed in the circuit and a voltage at the load.

Figure 10.1b explains the working principle by showing basically the band level scheme of a pn-junction as given in Chap. 9, Fig. 9.3b.

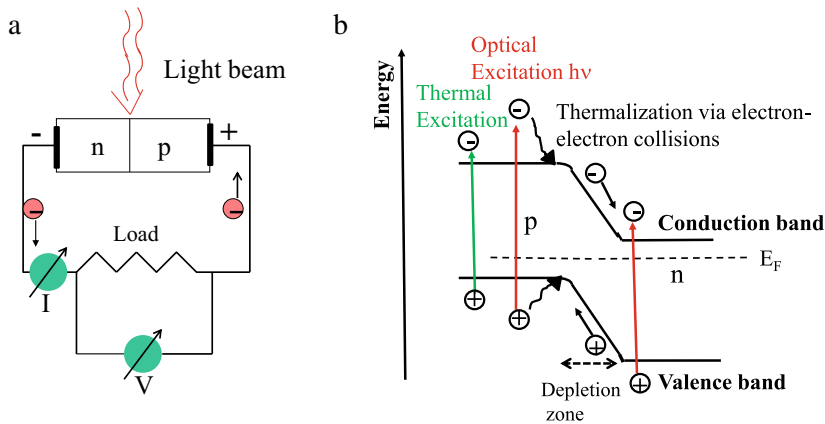


Fig. 10.1 **a** Schematic set-up of a solar cell, **b** Working principle of the solar cell

As discussed in the context of (9.3) in Sect. 9.1, the charge carriers providing the field current are created by thermal excitation, either being created in the depletion zone or arriving there by diffusion from the neutral regions. In the presence of light, another process enters the scene, the process of optical excitation, as included in Fig. 10.1b by the red arrows, which requires that $h\nu > E_G$.

The charge carriers being excited in the depletion zone, or arriving there by diffusion, just fall “down the potential hill”, i.e. the electrons from left to right and the holes from right to left. The crucial point to understand is that the origin of these charge carriers is basically the same as the one forming the field current (Fig. 9.4b), with the only difference that the latter are excited by thermal excitation, while the former by optical excitation. This means that the effect of optical excitation just consists in increasing the field current by a contribution due to the light. It follows from this consideration that the n-doped side of the solar cell is the cathode, while the p-doped side is the anode.

So the $I(U)$ characteristics in the presence of light are given by

$$I(U) = I_0(\exp^{eU/kT} - 1) - I_{\text{light}} \quad (10.1)$$

(to be compared with (9.12) valid in the dark).

$I(U)$ from (10.1) is shown in Fig. 10.2a.

Figure 10.2b shows the $I(U)$ characteristics for three different light intensities.

Only the lower right quadrant of Fig. 10.2a is shown, and reflected, as is usually done, at the U axis. Figure 10.3 shows this in more detail.

When the circuit is “open”, i.e. the two electrodes are not connected, no current flows ($I = 0$) and the voltage is U_{oc} (“open circuit voltage”). When the two electrodes are shortened, the maximum current flows, the so-called “short circuit current I_{sc} ”, while U is zero. With a load mounted between the two electrodes (see load in Fig. 10.1a), current and voltage correspond to some point on the $I(V)$ characteristics

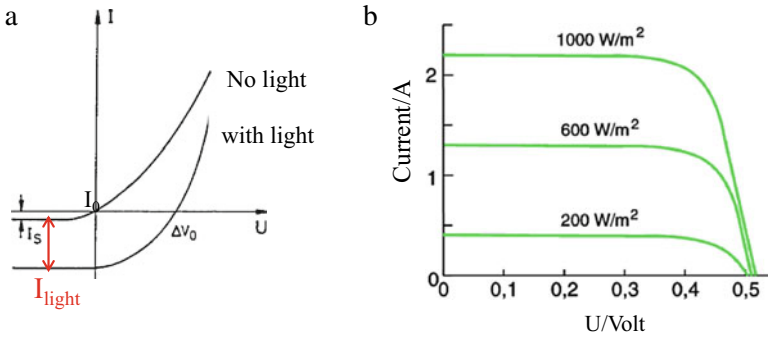
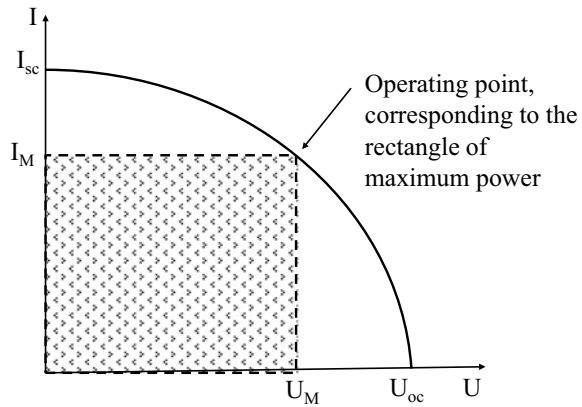


Fig. 10.2 **a** $I(U)$ characteristic of a pn-junction in the absence and presence of light, **b** $I(U)$ of a solar cell for three different light intensities (from www.solarserver.de/wissen/basiswissen/photovoltaik-typen-und-eigenschaften-von-solarzellen/)

Fig. 10.3 $I(U)$ characteristics of a solar cell, details

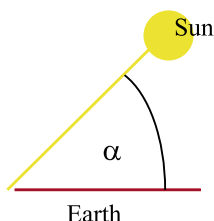


(the operating point in Fig. 10.3), and the power delivered by the solar cell is $I(V) \cdot V$. Figure 10.3 shows the optimum “operating point” where the product $I_M \cdot U_M$, is maximum, i.e. where the power extracted from the solar cell is maximum, represented by the “maximum power rectangle”. In order to adjust the operation to this point of maximum power, the load has to be properly chosen, which, by the way, is the case for any type of current source. The product of maximum power divided by the product $I_{sc} \cdot U_{oc}$, (i.e. the product of the maximum possible current times the maximum possible voltage) is called the “filling factor” FF of the solar cell

$$(I_M U_M) / (I_{sc} U_{oc}) = FF \quad (10.2a)$$

Often the “filling factor FF ” is given to characterize the performance of a solar cell. The maximum power to be harvested from the solar cell at a given light intensity is given by

Fig. 10.4 Power of sunlight per m^2 . AM means “Air mass”



α	Air Mass	Power (W/m^2)
		1353
90°	1	990
45°	1,5	930
30°	2	840

$$P_M = I_M U_M = FF * I_{sc} U_{oc} \quad (10.2b)$$

By applying (10.2a) to the $I(V)$ characteristic shown in Fig. 10.3, one obtains approximately a filling factor $FF = 0.75$ for this case.

The table in Fig. 10.4 gives a few more numbers to calculate the actual performance of a photovoltaic device.

It shows the power of the sunlight arriving at a solar module depending on the angle of incidence α . Each angle of incidence α corresponds to a certain amount of air, the power arriving on the cell for a given α is given in column 2. AM 1 means that the sunlight passes just one layer of the atmosphere vertically.

Numerical considerations concerning the performance of a solar cell

The characteristics of the solar cell (10.1)

$$i(U) = i_0(\exp^{eU/kT} - 1) - i_{\text{light}}$$

is easily understood (as explained) by realizing that the current produced by the light is based on the same process as the field current, i.e. on the excitation of electrons from the valence band to the conduction band, with the only difference that this excitation is done by light in the solar cell, and done thermally in the pn-junction in the dark. Just for the interest, we examine the question of what the ratio between the two processes is. The field current is given by i_0 , and the current induced by light excitation is i_{light} , so we are asking for the ratio $r = i_0/i_{\text{light}}$.

Let us consider a typical solar cell under illumination by the sun under AM 1,5 at room temperature.

From a list of commercial solar cell, we take as an example a cell with a filling factor $FF = 0.75$, for which $i_{sc} = 0.038 \text{ A}/\text{cm}^2$ and $U_{oc} = 0.6 \text{ V}$ is given at AM = 1.5.

From (10.2b), we see that such a cell produces (at AM = 1.5)

$$P = 0.038 \text{ A}/\text{cm}^2 * 0,6 \text{ V} * 0.75 = 0.0171 \text{ W}/\text{cm}^2.$$

What, by the way, is the efficiency of such a solar cell? As seen from the table in Fig. 10.4, the sun provides $0,093 \text{ W/cm}^2$ at AM = 1.5. So the efficiency is $\eta = 0.0171/0.093 = 18\%$, which is a reasonable value.

With these data for a typical Si solar cell, we now come back to our genuine question, the ratio $\eta = i_0/i_{\text{light}}$. Under open circuit conditions ($i = 0$, $V = V_{\text{oc}}$), 10.1 gives

$$i_0 \exp(eU_{\text{oc}}/kT) - i_0 - i_{\text{light}} = 0$$

so

$$i_{\text{light}} = i_0 [\exp(eU_{\text{oc}}/kT) - 1] \quad (10.3a)$$

or

$$i_{\text{light}}/i_0 = [\exp(eU_{\text{oc}}/kT) - 1] \quad (10.3b)$$

Because of $kT = 0.026 \text{ eV}$ at room temperature, this amounts to $eU_{\text{oc}}/kT = 0.6 \text{ V}/0.026 \text{ V} = 23$, so, by neglecting the 1 in the brackets, one gets

$$I_{\text{light}}/i_0 = e^{23} \approx 10^{10} \quad (10.3c)$$

So under typical sunlight exposure, the photocurrent exceeds the current i_0 produced by thermal excitation at room temperature by 10 orders of magnitude.

In the following an estimate of the absolute value of i_0 will be given:

For short circuit conditions, i.e. for $U = 0$, $i = i_{\text{sc}}$, we obtain from (10.1)

$$i_{\text{sc}} = -i_{\text{light}} \quad (10.3d)$$

Because i_{sc} is given to be $0,038 \text{ A/cm}^2$, we thus know from (10.3c)

$i_0 = 0.038 \text{ A/cm}^2 * 10^{-10} = 3.8 * 10^{-12} \text{ A/cm}^2$ in absolute numbers. So this is the current density produced by thermal excitation at room temperature in a “typical” pn-junction.

10.3 Operating Principle of the Solar Cell in Detail

Figure 10.5a shows in more detail the basic physical processes in a-Si solar cell.

Figure 10.5b shows the technical set-up of a-Si solar cell corresponding to the operating scheme shown in Fig. 10.5a

The sunlight passes through the finger electrodes at the upper surface. When a photon excites an electron from the valence band to the conduction band (see

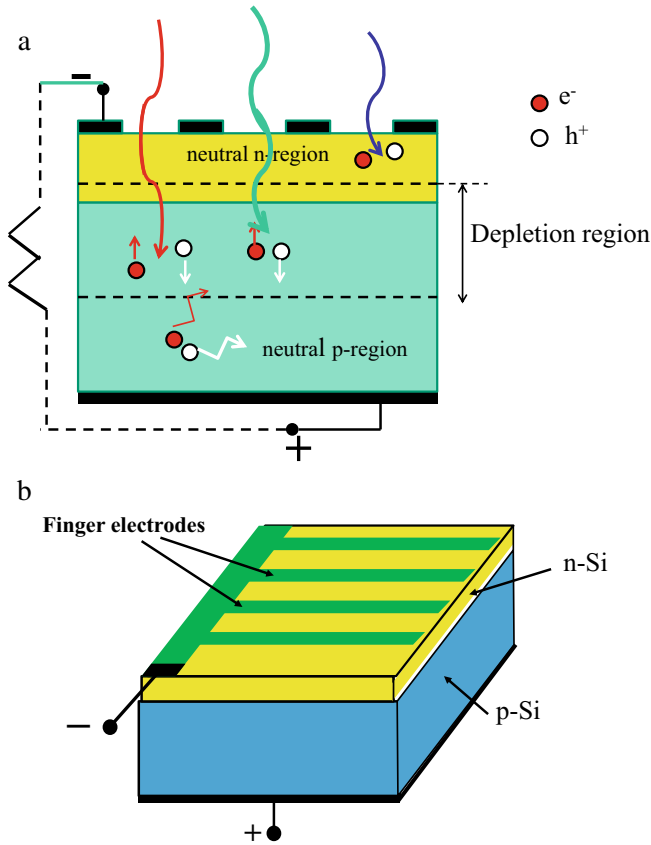


Fig. 10.5 **a** Basic physical processes in a-Si solar cell, **b** Technical set-up of a-Si solar cell

Fig. 10.1b), a free electron and a free hole are produced. When the absorption occurs in the depletion region, the charged particles follow the electric field of the depletion region, the electrons toward the cathode, and the holes toward the anode. The electrons and holes produced in the neutral regions will diffuse through the material and might eventually reach the depletion region; here they will be exposed to the electric field which attracts them toward cathode and anode, respectively. If the electrodes are connected via a load (Fig. 10.1a), electrons will flow from cathode to anode in the outer circuit.

The process competing with the production of free carriers via optical excitation from the valence to the conduction band is called recombination. Recombination is the enemy of efficiency! When electrons at some point on their way to their electrode (the cathode) recombine with a hole migrating toward its electrode (the anode), they are lost for the current production. So it is obvious that the mean free path length in the depletion region and the diffusion length (in the neutral zone) must be large enough such that the carriers reach their electrodes. This is a matter of material quality.

Lattice defects and impurities may act as recombination centers (see Sect. 11.3 on LEDs), and thus shorten diffusion length and free path lengths. This is the reason why the quality requirements for Si in Si photovoltaic are such that the expensive Siemens process (Sect. 5.2) still has to be applied for the purification of the material.

Physics comment: The binding energy of excitons

When an electron is excited from the valence band to the conduction band, resulting in an electron in the conduction band and a hole in the valence band, the electron and hole tend to form a bound system due to Coulomb attraction, the so-called exciton. So one would expect that after the optical excitation process in the neutral zone of the solar cell, excitons will diffuse through the lattice until they are eventually separated in the electrical field of the pn-junction. So far the notion of “excitons” had not shown up in this section, the discussion was on the diffusion of free electrons and holes in Si. The justification for this is the fact that the binding energy of excitons in Si is so low that at room temperature and above the excitons dissociate, resulting in the formation of free electrons and holes. This binding energy is about 15 meV, which is smaller than (or in the range of) thermal energies $kT_{300K} = 0.026$ eV at room temperature. So, at room temperature and above, thermal energies are sufficient to decompose the excitons into free electrons and holes.

The issue of the binding energy of a positive charge $+e$ and a negative charge $-e$ due to Coulomb interaction has been discussed in Chap. 3, Sect. 3.2 in the context of the hydrogen model of donors and acceptors.

In order to be precise (and as a little exercise to the reader), the “hydrogen model” has to be slightly modified to be applied to excitons. In the case of the exciton, our starting point is not the H atom, but the hydrogen-like system Positronium, which is the bound system formed by an electron (charge $-e$, mass m_0) and a positron (charge $+e$, mass m_0). The ionization energy of this typical two body system is not any longer given by

$$R_H = m_0 e^4 / 8h^2 \epsilon_0^2 = 13.6 \text{ eV} \quad (10.4a)$$

(the one of Hydrogen), but by $\mu e^4 / 8h^2 \epsilon_0^2$, where μ is the “reduced mass” of the electron–positron system.

$$\mu = m_0 * m_0 / (m_0 + m_0) = m_0 / 2.$$

So the ionization energy of Positronium is

$$R_P = \mu e^4 / 8h^2 \epsilon_0^2 = (m_0 / 2) e^4 / 8h^2 \epsilon_0^2 = (1/2) R_H \quad (10.4b)$$

In order to discuss the exciton, which in a way is a “positronium inside matter”, we have to proceed analogously to the case of the hydrogen model for donors and acceptors, i.e. we have to replace ϵ_0^2 in (10.4b) by $(\epsilon_0 \epsilon_r)^2$, where $\epsilon_r = 11.9$ is the dielectric constant of Si, and m_0 by the reduced mass $m_e m_h / (m_e + m_h) = \mu_{exc}$,

where m_e and m_h are the effective masses of electrons and holes in Si. With numbers of Table 1.2 chapter 1 : $m_e = 0.26 m_0$, $m_h = 0.32 m_0$ (mean value of m_{hh} and m_{lh}). the reduced mass is

$$\mu_{exc} = 0.14m_0$$

so, instead of (10.4b), one gets for the ionization energy $E_I(\text{exciton})$ of the exciton in Si

$$E_I(\text{exciton}) = 1/2 * R_H * 0.14 * 1/e_f^2 = 6.7\text{meV}$$

which evidently is considerably smaller than $kT_{300K} = 26 \text{ meV}$.

So excitons will essentially be dissociated in Si at room temperature resulting in free electrons and holes.

This, by the way, is totally different in organic semiconductors, the binding energy of excitons being considerably larger, i.e. usually between 0.3 and 0.5 eV, much larger than the thermal energies at room temperature. So what is relevant in organic electronics is not the diffusion length of free carriers, but one of excitons and the mechanism of charge separation in the electric field at the interface between hole- and electron-transporting layers (see sect.15.8).

10.4 The Optimum Bandgap for a Solar Cell Material

In the following, it will be shown that there is an optimum band gap for the material of a solar cell, in other words: The bandgap of a semiconductor material should neither be too large nor too small in order to be useful as a material for photovoltaics.

Figure 10.6 shows the emission spectrum of the sun.

(The unit Watts/m²/nm refers to the power irradiated on one square meter per spectral width of one nm).

Because in a solar cell, the solar photons have to excite electrons from the valence band to the conduction band in order to produce a current, it is evident that the bandgap should not be too large (Fig. 10.7a left) because all photons with $h\nu < E_G$ are not absorbed and would be lost. But, as indicated in Fig. 10.7a (right), the band gap should not be too small either relative to the photon energy for the following reason:

When the electron is excited into a state of the conduction band high above the band edge, it will, in a very short time, dissipate its energy to the other electrons of the conduction band via collisions and thus drop down the conduction band edge, a process called “thermalization” (see also Fig. 10.1b). Correspondingly, a hole left deep in the valence band will quickly be filled by electrons from higher states of the valence band, and thus end up at the valence band edge. The thermalization energies

Fig. 10.6 Solar spectral irradiance in Watts per m² and nm

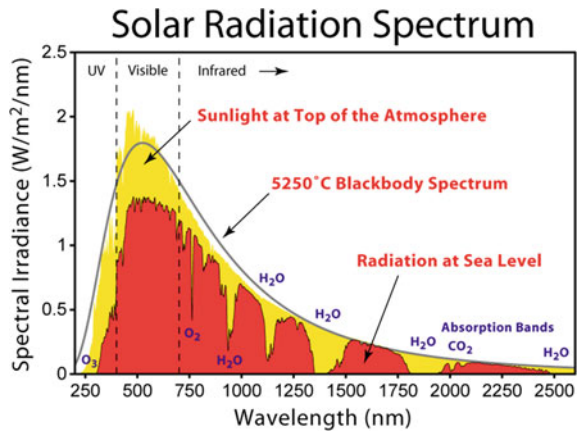
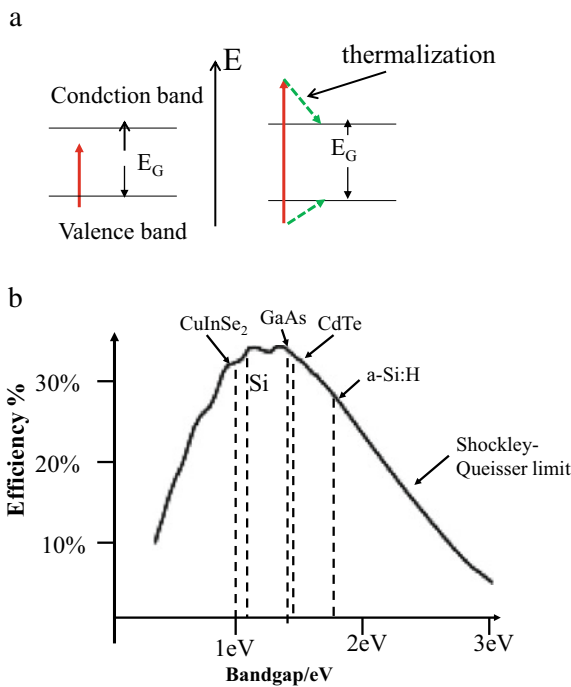


Fig. 10.7 a The case of “photon energy too small”, and “photon energy too large” relative to the band gap, **b** Maximum efficiency as function of bandgap



of electrons and holes, the sum of which amounts to $h\nu - E_G$, i.e. the “excess energy” of the photon, will thus be transformed into heat, only the bandgap energy E_G remains available for conversion into electrical energy.

When the bandgap should neither be too small relative to the energy of the solar photons, nor too large, there must be an optimum bandgap. Figure 10.7b shows the

attainable efficiency of a solar cell as a function of the bandgap for various solar cell materials. The line in Fig. 10.7b marks the “Shockley-Queisser limit”.¹

The maximum is at $E_G = 1.34$ eV with an efficiency of 33.7% for the solar spectrum at AM 1.5. Fortunately, the maximum is fairly broad, so for Si with its bandgap of 1.12 eV, it amounts to about 32%. GaAs and CdTe are slightly better, but Si is the dominating material in photovoltaics due to low cost, availability, and existing technology. In applications where costs are not relevant, as in space technology, GaAs is also used. A further advantage of GaAs in space applications is that GaAs, as a direct semiconductor, provides thin film solar cells resulting in low weight of the modules.

An approach to achieve efficiencies beyond the Shockley-Queisser limit is the realization of tandem solar cells. In tandem solar cells, two or more solar cells are stacked one on top of the other, the one with the higher bandgap $E_{\text{gap}}(1)$ on top, the one with the lower bandgap $E_{\text{gap}}(2)$ at the bottom. The one on top, with bandgap $E_{\text{gap}}(1)$ will harvest the photons with energies larger than $E_{\text{gap}}(1)$ and let pass through the ones with lower energies. The one at the bottom with bandgap $E_{\text{gap}}(2)$ then will harvest the photons with energies $E_{\text{gap}}(2) < h\nu < E_{\text{gap}}(1)$.

An analysis along the lines leading to the Shockley-Queisser limit leads to the following result:

The efficiency limit of a tandem cell with two semiconductor materials is 42% with $E_{\text{gap}}(1) = 1.9$ eV and $E_{\text{gap}}(2) = 1.0$ eV, and the efficiency limit for three semiconductors would be raised to 49% with the three bandgaps 2.3, 1.4, and 0.8 eV, respectively.²

10.5 Thin Film Solar Cells

Si is an indirect semiconductor (Sect. 1.7). Because of the small absorption coefficient for light, a solar cell has to be some hundred μm thick (typically 200 μm) in order to absorb a large fraction of the sunlight. Direct semiconductors would absorb the sunlight quantitatively already as thin films of less than 1 μm , thus providing “thin film solar cells”.

It is remarkable that in spite of the toxicity problems of Cd, CdTe modules are legally approved. A very large solar park with 52 MW peak power has been installed in Waldpolenz (Saxonia, Germany). CdTe is appreciated as a solar cell material because it is a direct semiconductor providing thin film solar cells. Special attention should also be given to CuInSe₂ (CIS) (see Fig. 10.7b) and modifications like CuIn(SGa)₂ which are considered as promising candidates for thin film solar cells with reported efficiencies near 20%. The third promising material for a thin film solar cell technology is amorphous Si (a-Si), as discussed in chapter 8.5, based on

¹ William Shockley [1].

² Ameri et al. [2].

the fact presented in Sect. 8.2 that amorphous Si, contrary to crystalline Si, shows direct semiconductor properties. So, presumably, CIS, $\text{CuIn}(\text{SeGa})_2$, CdTe, and a-Si will share the growing thin film solar cell market. In 2017, the global market share of thin film solar cell was 4.5%, 1.9% of which was provided by $\text{CuIn}(\text{SeGa})_2$, 2.3% by CdTe, and 0.3% by a-Si.

10.6 Cheaper Silicon for Photovoltaics

At present, the vast majority of solar cells is made of Si. Electronic Si is produced from SiO_2 by the Siemens process (Sect. 5.1) which is quite expensive in terms of energy consumption and manufacturing costs. Thus, the question arises whether alternative methods might be used to reduce those costs, given the fact that the purity requirements for photovoltaics are high, but nevertheless lower than for high end electronic applications. To give some numbers: For solar grade Si, an impurity content as low as 10^{-6} (atomic fraction), depending on the impurity species, is considered to be sufficient, while the one for electronic grade Si is 10^{-9} to 10^{-10} .

The Si obtained by reduction of SiO_2 , the “metallurgical Si” is still of very poor quality (impurity atomic fraction about 1–2%). The Siemens process then consists of transforming the Si into trichlorosilane (SiHCl_3), then purifying the SiHCl_3 to an extremely high degree by distillation, and transforming it back into Si. The methods for “cheaper Si” avoid these chemical reactions and aim at purifying directly the metallurgical Si. For that reason, we speak of the “metallurgical route” to solar grade Si.

Most impurities in metallurgical Si (like Fe, Al, C, Ca, Mg, and others, depending on the SiO_2 source) can be taken out by the float zone process (Sect. 5.4 and Fig. 5.3) based on the fact that, in a system with liquid Si in contact with solid Si, the impurity accumulates in the liquid to a degree given by the segregation coefficient $\kappa = c_{\text{solid}}/c_{\text{liquid}}$. The impurities P and B, however, cannot be removed by segregation because the segregation coefficients are too large (κ is 0.35 for Si-P, and 0.7 for Si-B). It is evident that these impurities are particularly harmful by providing an uncontrolled conductivity, because P acts as a shallow donor, and B acts as a shallow acceptor (Tables 3.2 and 3.3 in Chap. 3). So P and B have to be removed by a different procedure.

Removal of Phosphorus

P can be removed by heating under vacuum. The Si is heated up to the melting point, usually electromagnetically, which can be done because the Si melt is metallic. When the melt is stirred, and under low pressure, the P tends to evaporate.

Removal of Boron

One way to remove Boron impurities from Si is by exposing the melt either to a stream of H_2 and H_2O , with Ar as a carrier gas, or to a plasma of Ar, H_2 , and O_2 . O

and H then will react with the B forming in a first reaction step SiO and HBO. The B will finally end up as an impurity in SiO_2 .

Figure 10.8 summarizes schematically the route from metallurgical Si to solar grade Si.

The figure indicates that the process will be repeated several times in order to achieve the desired degree of purity. At present, the route via the Siemens process is still strongly dominant but the direct way starting from metallurgical Si and avoiding “Trichlorsilane chemistry” is gaining ground because of lower costs and lower energy consumption.

Polycrystalline Si solar cells

A way to reduce costs and energy consumption is by use of polycrystalline instead of monocrystalline material. In this way, the Czochralski process can be avoided and is—preferentially—replaced by the “casting technology”. In this process, a melt of Si is poured into a preheated vessel (some m^3 size on an industrial scale!) with a temperature gradient, the lower temperature at the bottom. Then the temperature field is pulled upwards. In Fig. 10.9a, the situation is shown where the melting temperature lies somewhere in the middle of the vessel, below Si is in the solid state (blue), and above in the liquid state (green).

Because of this “directional solidification”, the grain boundaries show a strong texture running from top to bottom. When the whole block has been cooled down in this way, it is sawed vertically with respect to these grain boundaries. Figure 10.9b

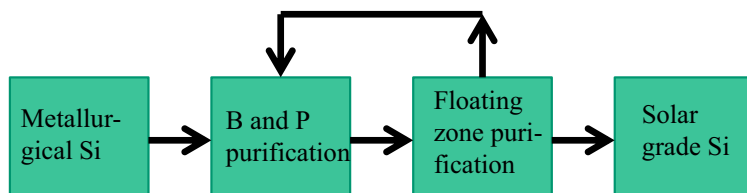


Fig. 10.8 Route from metallurgical Si to solar grade Si

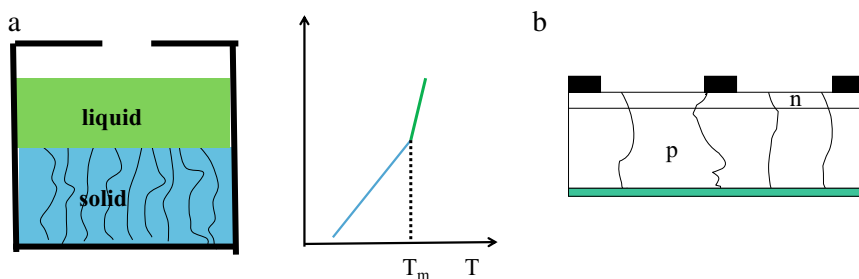


Fig. 10.9 a Directional solidification of polycrystalline Si, b Side view cross section of a polycrystalline Si solar cell

shows that, as a consequence of the texture of the grain boundaries, the electrons and holes in the solar cell only rarely have to cross grain boundaries on their way to the electrodes, because the diffusion path towards the electrodes is parallel to the grain boundaries. Thus reduction of efficiency by recombination at grain boundaries is avoided.

It should be noted in passing that directional solidification is a common industrial process of great importance in materials technology in general, in particular applied in metallurgical casting. In this way mechanical properties can be achieved coming close to the ones of a monocrystalline material.

10.7 A short summary of the state of the art in photovoltaics³

(Keep in mind that some of the following numbers are subjected to rapid changes). Record lab solar cell efficiency for monocrystalline Si is 26.7%, for multi-crystalline Si, it is 22.3%.

The efficiency of modules is slightly lower, e.g. 24.4% have been realized for monocrystalline Si.

The PV capacity installed in Germany in 2018 was 45.4 GWp, about 9% of the 515 GWp installed worldwide. Remark: The capacity of a photovoltaic installation is usually given in kW or MW peak, abbreviated kWp or MWp. “Peak” refers to the maximum power achieved under standard conditions, i.e. a solar irradiance of 1000 W/m², a modul temperature of 25 °C at air mass 1.5, conditions which are rarely fulfilled altogether.

The annual increase of PV installation in the world was 24% between 2010 and 2017. The leading producer nowadays is China with a share of 70%.

In 2017, the share of the global market was 4.5% for thin film solar cells, 60.8% for multi-crystalline, and 32.2% for monocrystalline solar cells.

A question often discussed is the one of the Energy Payback Time (EPBT) of modules, i.e. the time after which the energy invested for production of the module has paid back. The EPBT is between 0.7 and 2 years at present depending on technology. Because of the developments in the technology (e.g. material quality, improvements in power electronics) and fabrication process, it improved from about 3.3 years in 1990 to about 1 year at present for multi-crystalline Si cells installed in Southern Europe.

In order to increase the efficiency of Si solar cells beyond the Shockley-Queisser limit, the so-called tandem and multijunction solar cells are a major topic of photovoltaic research, as has been discussed already to some extent in Sect. 10.4. Optimum values for the two bandgaps involved have been given at the end of section 10.4. It can be shown that quite a good combination of bandgaps would be 1.70 and 0.77 eV for a tandem solar cell. The use of the ternary III-V-compound GaInP for the top sub-cell (E_G between 1.8 and 1.9 eV) and Ge ($E_G = 0.66$ eV) for the lower layer comes close to that optimum.

Another combination under intense investigation is Perovskites on Si, a system favored due to the large progress in Perovskite solar cells (the latter being built from

³ Data taken from “Photovoltaics Report”, Fraunhofer Institute of Solar Energy ISE March 2019.

thin films of $(\text{CH}_3\text{NH}_3\text{PbX}_3)$ (X stands for a halide). A problem with Perovskites is the use of Pb because of the toxicity problem, nevertheless special permissions for use in solar cells might be granted, as in the case of CdTe. Another problem to be dealt with is the long-term stability of the cells.⁴

10.7 Questions and Answers

- Question 1: What is the photovoltaic effect?
- Question 2: How does the photovoltaic effect influence the $I(V)$ characteristics of a pn-diode?
- Question 3: What do field current and photocurrent have in common?
- Question 4: The bandgap of a semiconductor material for a solar cell should neither be too large nor too small, resulting in an optimum bandgap. Why?
- Question 5: What is the optimum bandgap? Name a few materials coming close to the optimum.
- Question 6: Name a few thin film- and thick film solar cell materials.
- Question 7: Why is high quality material required for solar cells?
- Question 8: What is the record efficiency of a-Si monocrystalline solar cell?
- Question 9: What is the favorite process in the fabrication of multi-crystalline solar cells (replacing the crystal growth processes applied for monocrystalline cells)?
- Question 10: How large is the energy payback time of a modern Si solar cell?

Answers

- Answer 1: When a pn-junction is exposed to light, a current is produced, as well as a voltage at an external load. This is the photovoltaic effect.
- Answer 2: It adds a term to the field current. This term (the photocurrent) is (at not too high intensities) proportional to the light intensity.
- Answer 3: They both consist of charge carriers produced by excitation of electrons from the valence band to the conduction band, thermally in case of the field current, optically in case of the photocurrent. In the electric field of the depletion zone, a current is formed out of these free carriers.
- Answer 4: If the bandgap is too large, only a fraction of the sunlight is absorbed, just the ones with photon energies larger than the band gap. If the bandgap is too small, part of the energies of the absorbed photons are lost for the current production due to thermalization.
- Answer 5: The optimum is at 1.34 eV. Materials close to the optimum are GaAs, CdTe, a-Si, Si, and CIS.
- Answer 6: Thick film solar cells are made from indirect semiconductors, thin film solar cells from direct semiconductors. Thick film solar cell: Si. Thin film solar cells: GaAs, CIS, CdTe, a-Si.

⁴ For a survey see Zhao et al. [3].

- Answer 7: Because defects and many impurities act as recombination centers for electrons and holes, thus reducing efficiency.
- Answer 8: About 27%.
- Answer 9: Block casting.
- Answer 10: Around 1–1.5 years.

References

1. W. Shockley, H.J. Queisser, Detailed balance limit of efficiency of p–n junction solar cells. *J. Appl. Phys. Band* **32**(3), 510–519 (1961) <https://doi.org/10.1063/1.1736034>.
2. T. Ameri, G. Dennier, C. Lungenschmied, C. Brabec, Organic solar cells: a review. *Energy Envir. Sci.* **2**, 347–363 (2009). <https://doi.org/10.1039/B817952B>.
3. Y. Zhao, C. Brabec et al., A bilayer conducting polymer structure for planar perovskite solar cells with 1400 h operational stability at elevated temperatures. *Nat. Energy*. <https://doi.org/10.1038/s41560-021-00953-z>.

Chapter 11

Light Emitting Diodes (LEDs)



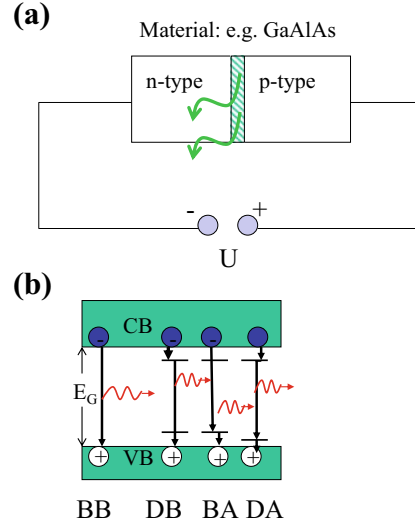
Abstract Traditional light sources like light bulbs, energy saving lamps, halogen lamps, etc. are more and more replaced by light emitting diodes (LEDs), which might be considered as the “light source of the future”. LEDs are based on the light emitting process in a pn-junction. In the depletion zone around the junction, electrons and holes recombine under emission of photons with an energy close to the bandgap energy. A detailed discussion results in linewidths of a few 10 nm. The light emitting processes have to compete with non radiative recombination processes which reduce the internal quantum efficiency. Internal efficiency is strongly a matter of material quality. The efficiency is further reduced by total reflection at the surface of the diode or at internal interfaces, resulting in the “external quantum efficiency”. LEDs are now available in the spectral regions from UV to IR. An important step in optoelectronics was the availability of the blue LED, based on the nitride semiconductors GaN and InN. The major problem to be overcome was the one of p-doping of GaN. Considerable progress has still to be achieved in the UV region, in particular in the far UV. Here the semiconductor AlN might play a role. As far as the nitride semiconductors are concerned there is still room for further progress in providing proper substrates for the active layers. Since the blue LED is available at sufficient efficiency, white LEDs can be produced based on the principle of “Luminescence conversion”. Fast progress has been achieved for all spectral regions in terms of efficiency, measured in Lumen/Watt. In modern white LEDs near 50% of the electrical energy is converted into light energy. The economically important improvement of the efficiency and of the color temperature of white LEDs goes along with further progress in the field of phosphors. So the white LED is a driving force in the field of phosphor materials.

11.1 The Light Emitting Process in Semiconductors

Another straight forward application of the pn-junction is the light emitting diode (LED). Figure 11.1a shows the basic set-up.

When current flows in forward direction across a pn-junction in a semiconductor, electrons and holes meet in the depletion region and may emit light by recombination. The light emitting processes are shown in Fig. 11.1b.

Fig. 11.1 **a** Basic set-up of an LED. **b** Light emitting processes in an LED



The figure shows from left to right the different transitions in a self explaining notation: Band-band transition (BB), donor-band-transition (DB), band-acceptor transition (BA), and donor-acceptor (DA) transition.

Because the energy levels of the donors and acceptors are close to the band edges (order kT , which is 26 meV at room temperature), the energy of the emitted photons approximately equals the bandgap energy (which is of order eV). So

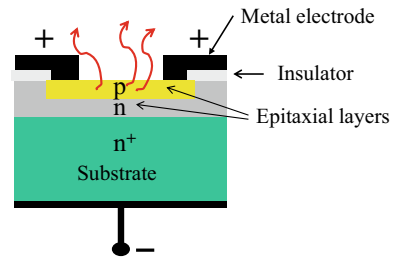
$$h\nu \cong E_G \quad (11.1)$$

Because $\lambda = c/\nu$ this gives $\lambda = hc/E_G$, and by expressing λ in nm and E_G in eV, one easily obtains

$$\lambda(\text{nm}) = hc/E_G = 1240/E_G(\text{eV}) \quad (11.2)$$

a formula everybody interested in the field of optoelectronics should memorize! Fig. 11.2 shows the structure of a simple LED.

Fig. 11.2 Structure of a simple LED



It essentially shows a pn-junction formed by two epitaxial layers grown on a substrate, plus the electrodes. The substrate is heavily doped to assure a high conductivity. The LED evidently is poled in the forward direction, as it should be.

11.2 The Linewidth of LEDs

As shown in Fig. 11.1b, the emitted light is a superposition of various transitions which differ by something of the order of 10 meV in energy. To be more specific, Table 11.1 gives some energy levels (in meV) of shallow donors and acceptors in GaAs as an example.

The transitions including donor or acceptor states thus have photon energies lower than the band-band transitions by something between 6 and 35 meV, to be compared to a bandgap energy of 1.42 eV in GaAs. So, when all the transitions shown in Fig. 11.1b contribute to the emission of the LED, a linewidth of about 30–40 meV might result.

A comparable contribution to the linewidth of the LED arises from the fact that a spectral linewidth of that size is inherent in the band-band transition alone, as will be discussed in the following.

The photon energy of the band-band transition is not precisely equal to $h\nu = E_G$ because, as schematically shown in Fig. 11.3, the electrons and holes connected by the radiative transition have a certain thermal energy distribution which will result in a shift to a slightly higher energy and a certain spectral width of this transition.

More precisely, the maximum of the transition will be shifted to a photon energy of

$$E_{\max} = h\nu_{\max} = E_G + \frac{1}{2}kT \quad (11.3a)$$

From looking at Fig. 11.3, one would be tempted to assume that the maximum of the emission would connect the maxima of the electron and hole distribution in the conduction band and the valence band resulting in a maximum of the emission of the LED at $E_G + kT$. The situation, however, is more sophisticated because in calculating the emission spectrum, one has to take into account that not any electron in the conduction band can recombine with any hole in the valence band. On the contrary, because of momentum conservation, electrons with a given momentum, i.e. a given value of k , can only recombine with a hole of the same k . This results in

Table 11.1 Positions of shallow donor and acceptor levels in GaAs relative to the band edge

Donors	S	Se	Si
$E_{\text{Donor}}/\text{meV}$	6	6	5,8
Acceptors	Be	Mg	Si
$E_{\text{Acceptor}}/\text{meV}$	28	28	35

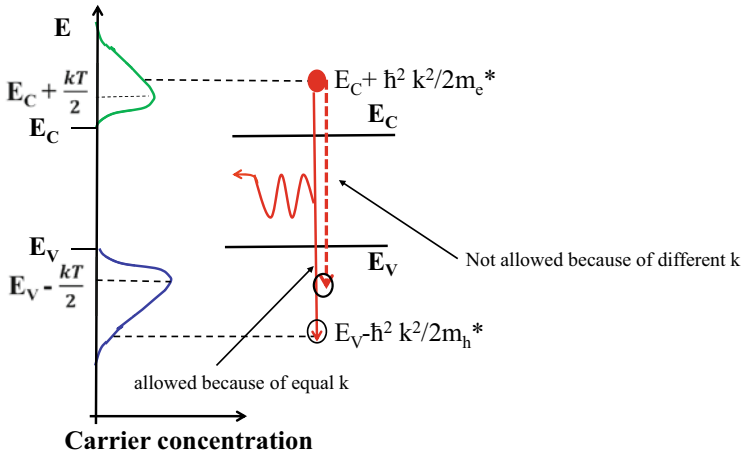


Fig. 11.3 Origin of the linewidth of the band-band transition

11.3a, and leads to a full width of half maximum of the optical transition¹

$$\Delta(h\nu) = 1.8kT \quad (11.3b)$$

The spectral width of an optical transition is usually not given in terms of a spread $\Delta(h\nu)$ of photon energies, but a spread of the wavelength $\Delta\lambda$. By making use of $h\nu = hc/\lambda$, $\Delta(h\nu)$ can be easily converted into $\Delta\lambda$ $d(h\nu)/d\lambda = -(1/\lambda^2)hc$, or, by switching from $d(h\nu)$ and $d\lambda$ to the differentials $\Delta h\nu$ and $\Delta\lambda$:

$$\Delta(h\nu)/\Delta\lambda = -(1/\lambda^2)hc, \text{ or } \Delta\lambda = \Delta(h\nu)\lambda^2/hc \quad (11.3c)$$

(the minus sign is omitted in the context of linewidths).

So (11.3b) can as well be written as

$$\Delta\lambda = 1.8kT * \lambda^2/hc \quad (11.3d)$$

for the full width at half maximum in terms of the wavelength spread.

Numerical example for the wavelength shift and width of the band-band transition as observed in an LED

We make use of (11.3a) and (11.3b).

At room temperature $kT \cong 0.026$ eV. The bandgap of GaAs is 1.42 eV at room temperature, which would correspond to a wavelength of (11.2)

$$\lambda/\text{nm} = 1240/1.42 (\text{eV}) = 873.2 \text{ nm.}$$

¹ For a detailed treatment see e.g. E. Fred Schubert, Jaehae Cho, Jong Kuy Kim. Light emitting diodes, <https://doi.org/10.1016/B978-0-12-803581-8.01081-X> December 2016.

According to (11.3a), the band-band transition is shifted by $1/2 kT = 0,013 \text{ eV}$, so its photon energy is now $1.42 \text{ eV} + 0.013 \text{ eV} = 1.433 \text{ eV}$, corresponding to a wavelength of $1240/1.433 \text{ nm} = 865.3 \text{ nm}$. So we have a shift of about $\Delta\lambda = -8 \text{ nm}$ relative to the band gap energy.

Evidently, the shift relative to wavelength corresponding to the bandgap is small due to the relatively small size of kT . So the bandgap E_G still is a good approximation to the real emission energy (or wavelength).

The spectral width of the transition according to 11.3b, then will be

$$\Delta(h\nu) = 1.8 * kT = 1.8 * 0.026 = 0.0468 \text{ eV} \quad (\text{full width half maximum})$$

corresponding, according to 11.3d, to a linewidth of ca. 30 nm.

Equation 11.3d shows that the linewidth, given in terms of wavelength, is strongly increasing with wavelength, due to the conversion from energy to wavelength. So the spectral width in terms of wavelength of an IR LED is much larger than the one of an LED emitting in the visible.

Because the linewidth of the LED (11.3d) is due to the thermal distribution of electrons and holes in the bands, the way to reduce the linewidth of an LED is by going to lower temperatures.

11.3 Radiative and Nonradiative Processes

The light emitting processes in a LED compete with “non radiative” or “radiationless” processes where the recombination energy is released via phonons instead of photons, thus resulting in heating of the material instead of light emission. The ratio of the number of radiative transitions to the total number of transitions (radiative plus non radiative) is called “(internal) quantum efficiency” of the LED.

Let the probability of a radiative transition of an electron from the conduction band to the valence band per time be $W_{\text{rad}} = 1/\tau_{\text{rad}}$, where τ_{rad} is the lifetime related to this process, and $W_{\text{nonrad}} = 1/\tau_{\text{nonrad}}$ be the probability for a nonradiative transition (see Fig. 11.4a), then the total transition probability is

$$1/\tau_{\text{total}} = (1/\tau_{\text{rad}}) + (1/\tau_{\text{nonrad}}) \quad (11.4)$$

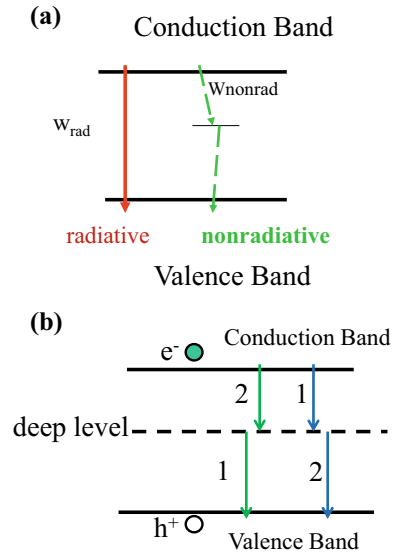
Then, according to the definition, the quantum efficiency is given by

$$\eta = (1/\tau_{\text{rad}})/(1/\tau_{\text{total}}) = (1/\tau_{\text{rad}})/(1/\tau_{\text{rad}} + 1/\tau_{\text{nonrad}})$$

or

$$\eta = 1/(1 + \tau_{\text{rad}}/\tau_{\text{nonrad}}) \quad (11.5)$$

Fig. 11.4 **a** Radiative band-band transition and nonradiative transitions via recombination centers. **b** Nonradiative recombination, partial steps



So the lifetime for the radiative transition has to be short as compared to the one for the nonradiative transition in order to achieve a high quantum yield. From a practical point of view, this is a matter of material quality, because lattice defects as well as certain impurities tend to act as recombination centers for nonradiative decay.

As indicated in Fig. 11.4a, nonradiative decays stem from defect centers giving rise to new energy levels in the bandgap. In case of such an additional energy level in the bandgap, nonradiative recombination can occur in two different ways, depending on whether the deep level is occupied or unoccupied by an electron; see Fig. 11.4b

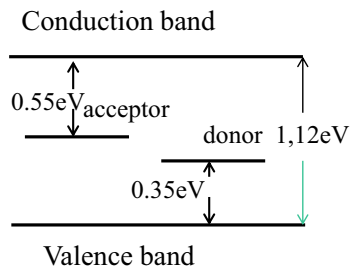
First case: The deep level is occupied by an electron. In this case (left side of Fig. 11.4b, green arrows), first the electron from the deep level will jump to the hole in the valence band (green arrow 1), then the electron from the conduction band will jump to the empty free level (green arrow 2).

Second case: Deep level empty. In this case (right side of Fig. 11.4b, blue arrows), first the electron will jump down from the conduction band to the deep level (blue arrow 1), and subsequently it will jump down to the hole in the valence band (blue arrow 2).

In both cases, the final result will be that an electron of the conduction band has recombined with a hole in the valence band. We have described the two cases in such detail in order to make clear, that the recombination via a deep level works in both ways: When the deep level is occupied (case 1), or when it is unoccupied (case 2). This is an essential point in calculating the lifetime of the nonradiative recombination process! This is why the so-called “Shockley–Read–Hall process (SRH process)” works independently of whether the deep level is occupied or empty.

The Shockley–Read–Hall process is a radiationless process, the recombination energy is carried away by phonons. Phonon energies are of the order 10 meV, while

Fig. 11.5 Deep levels of Au in Si



bandgaps are of order eV. This means that some 100 phonons are involved in a radiationless recombination transition. The matrix elements for creation or annihilation of phonons are relatively small, this makes—qualitatively spoken—the process so slow that in direct semiconductors, radiationless processes are generally much slower than the radiative recombination process, making direct semiconductors suitable for optoelectronic applications. In indirect semiconductors, however, the radiationless processes dominate due to momentum conservation requirement (Chap. 1). The reason why energy levels in the bandgap give rise to strong radiative transitions is that they split the recombination process into two subsequent steps, for instance, processes 1 and 2 in Fig. 11.3b. A closer analysis shows that the recombination time gets shortest when the energy level of the recombination center is close to the middle of the bandgap (which means that approximately the same number of phonons has to be created for both partial steps 1 and 2). So “deep levels” near the middle of the bandgap are the most efficient recombination centers.

In Silicon, Au, Ag, and Cu give rise to energy levels near the middle of the bandgap, thus strongly affecting the carrier lifetimes, and even very low impurity levels around $10^{10}/\text{cm}^3$ have to be avoided.

Figure 11.5 shows the energy levels of Au in Si, to give this important case as an example.

Au gives rise to a donor and acceptor level in the bandgap, the acceptor level being perfectly in the middle of the bandgap. So Au in Si is an extremely efficient nonradiative recombination center.

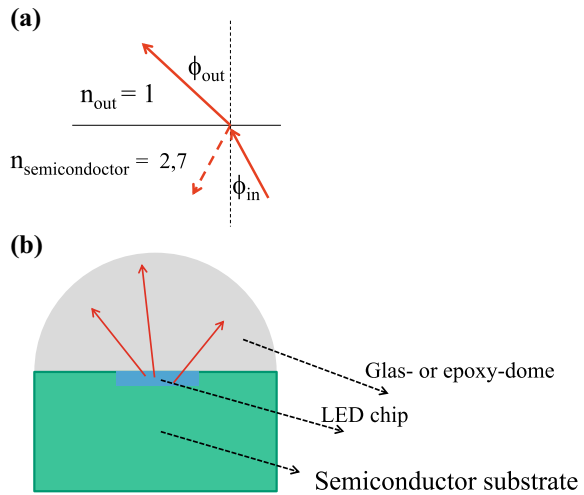
11.4 External and Internal Quantum Efficiency

Radiationless transitions are not the only source of loss. Another one is due to the fact that some of the light produced in the pn-junction cannot escape from the diode. A major reason for this is total reflection at the interface semiconductor/air, as shown in Fig. 11.6a.

According to Snellius’ law of reflection

$$\sin \phi_{\text{out}} / \sin \phi_{\text{in}} = n_{\text{in}} / n_{\text{out}} \quad (11.6)$$

Fig. 11.6 **a** Reflection and refraction at the interface semiconductor/air. **b** LED, encapsulated with epoxy dome



Total reflection sets in when $\phi_{\text{out}} = 90^\circ$. With $n_{\text{out}}(\text{air}) = 1$ and a typical value for inorganic semiconductors of $n_{\text{in}} = 2.7$, total reflection sets in at $\sin \phi_{\text{in}} = 0.37$ or $\phi_{\text{in}} = 21.7^\circ$. So only light within the cone of $\phi < 21.7^\circ$ can leave the semiconductor. A choice of several measures is available to increase the fraction of light which can escape the semiconductor chip. One way is to put a dome of epoxy on the LED chip as shown in Fig. 11.6b.

The reflection at the interface semiconductor-dome is reduced as compared to air because the difference in the refractive indices semiconductor—epoxy is smaller, and because the light hits the interface dome-air at an angle near 90° . Another way is the incorporation of scattering centers in the form of nanoparticles into an epoxy encapsulation. A photon with $\phi_{\text{in}} > 21.7^\circ$ might thus be scattered into a direction $\phi_{\text{in}} < 21.7^\circ$ not undergoing total reflection. A third method would be roughening of the surface of the semiconductor resulting in scattering processes avoiding total reflection. All in all a fraction of more than 50% of the photons might leave the LED, such that, if the internal quantum efficiency is 90%, the external quantum efficiency would reach 45%.

11.5 LEDs for All Colors

According to [11.2](#)

$$\lambda(\text{nm}) = 1240/E_G(\text{eV})$$

The bandgap determines the wavelength and thus the color of the LED.

LEDs are available now for essentially all wavelength regions. The following table defines these regions more specifically:

The principle of obtaining a large variety of bandgaps by mixing of compound semiconductors has already been explained in Sect. 7.2. The bandgap $E_G(\text{GaN}) = 3.42 \text{ eV}$ of GaN corresponds to a wavelength 363 nm, the one of InN $E_G(\text{InN}) = 0.7 \text{ eV}$ to a wavelength of $1.77 \mu\text{m}$, so one would expect that, by a proper admixture x of In to GaN, $\text{Ga}_{1-x}\text{In}_x\text{N}$ -LEDs between those wavelength limits (i.e. from the UV to the IR region) could be provided.

It turns out, however, for too large an admixture of In ($x \geq 0.3$) the quantum efficiency starts to decrease significantly. This is related to two problems:

- With increasing In content the crystal quality decreases, partly due to the formation of In conglomerates.
- The difference in electronegativity is larger for In/N as compared to Ga/N, which results in formation of internal electric fields.

Numerical example

The bandgap energy of GaInN (like the one of other ternary semiconductors) can well be described by a quadratic equation

$$E_G(\text{Ga}_{1-x}\text{In}_x\text{N}) = (1 - x)E_G(\text{GaN}) + xE_G(\text{InN}) - x(1 - x)b \quad (11.7)$$

where b is called the “bowing parameter” of the ternary compound. For GaInN, the bowing parameter has been determined to be $2.8\text{--}2.96 \text{ eV}^2$. With $E_G(\text{GaN}) = 3.42 \text{ eV}$ (Table 7.2) and $E_G(\text{InN}) = 0.7 \text{ eV}$ and with $b = 2.88 \text{ eV}$, one obtains for $x = 0.3$: $E_G(\text{Ga}_{0.7}\text{In}_{0.3}\text{N}) = 2.3 \text{ eV}$, which, according to (11.2), corresponds to a wavelength well in the green: $\lambda = 539 \text{ nm}$.

So, as a result, GaInN-LEDs are used in the spectral regions between the near UV and the green (Table 11.2).

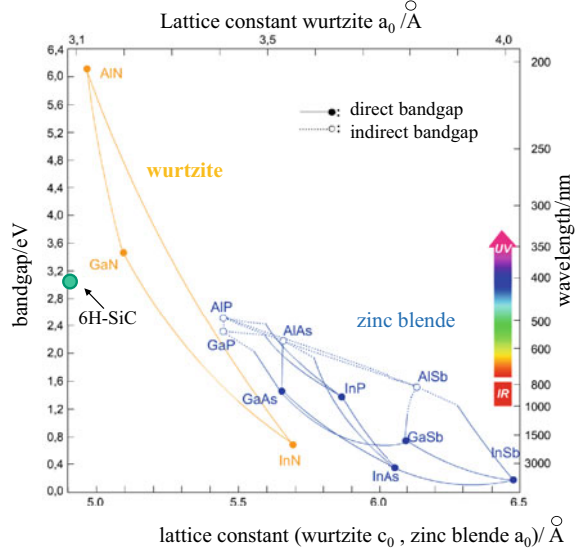
A major problem in applying mixed compound semiconductors is the one of finding an appropriate substrate for the epitaxial growth of the material, as has been presented in Sect. 7.2. As an example of good “lattice match” between substrate and epitaxial layer, the case of $\text{Al}_{1-x}\text{Ga}_x\text{As}$ has been discussed there. The issue of lattice match between semiconductor materials is generally discussed on the basis of diagrams like the one shown in Fig. 11.7.

The figure shows band gaps and lattice constants for a number of semiconductor materials. From the figure, it is immediately obvious that, e.g. $\text{As}_{1-x}\text{Ga}_x\text{As}$ can well be grown on a GaAs substrate because GaAs and AlAs (and consequently all mixtures of them) have almost the same lattice constant (see also Sect. 7.2). The diagram in Fig. 11.7, or rather a section of it, will be used to find out the appropriate material systems to be used in semiconductor lasers for optical data transmission (see Sect. 12.5). The substrate problem connected with the nitride semiconductors will be discussed at the end of the next section.

Table 11.2 Wavelengths regions and corresponding semiconductor materials (for a deeper understanding of the table; see also Fig. 11.10, e.g.: $\text{Al}_{1-x-y}\text{Ga}_x\text{In}_y\text{P}$ is the more recently used material)

Wavelength/nm	Color	E_G semiconductor	Material
230–400	Ultraviolet (UVC-UVB-UVA)	$5.4 > E_G > 3.1 \text{ eV}$	$\text{Al}_{1-x}\text{Ga}_x\text{N}$ $\text{Al}_{1-x-y}\text{Ga}_x\text{In}_y\text{N}$
400–450	Violet	$3.1 \text{ eV} > E_G > 2.8 \text{ eV}$	$\text{Ga}_{1-x}\text{In}_x\text{N}$
450–490	Blue	$2.8 \text{ eV} > E_G > 2.5 \text{ eV}$	$\text{Ga}_{1-x}\text{In}_x\text{N}$
490–560	Green	$2.5 \text{ eV} > E_G > 2.2 \text{ eV}$	$\text{Ga}_{1-x}\text{In}_x\text{N}$
560–590	Yellow	$2.2 \text{ eV} > E_G > 2.1 \text{ eV}$	$\text{GaAs}_{1-x}\text{P}_x$ $\text{Al}_{1-x-y}\text{Ga}_x\text{In}_y\text{P}$
590–630	Orange	$2.1 \text{ eV} > E_G > 2.0 \text{ eV}$	$\text{GaAs}_{1-x}\text{P}_x$ $\text{Al}_{1-x-y}\text{Ga}_x\text{In}_y\text{P}$
630–700	Red	$2.0 \text{ eV} > E_G > 1.8 \text{ eV}$	$\text{Al}_{1-x}\text{Ga}_x\text{As}$ $\text{Al}_{1-x-y}\text{Ga}_x\text{In}_y\text{P}$
>700	Infrared (IR)	$E_G < 1.8 \text{ eV}$	$\text{Al}_{1-x}\text{Ga}_x\text{As}$
1.54 μm	Wavelength of opt data transmission	$E_G = 0.805 \text{ eV}$	$\text{In}_{0.58}\text{Ga}_{0.42}\text{As}_{0.90}\text{P}_{0.10}$

Fig. 11.7 Bandgap versus lattice constant of compound semiconductors



11.6 The Story of the Blue LED

A crucial step in making LEDs available for general lighting was the invention of a sufficiently efficient blue light emitting diode. At the time of this invention (in the 1990s), the other colors were available already. So the coming up of the blue LED

did not just mean the availability of another color, but the completion of the whole visible spectrum, thus providing the possibility of producing any color by mixing of red, green, and blue, as done, e.g. in the color TV, including white.

As seen from Fig. 11.7 (and as discussed in the previous section), blue light emitting diodes can be produced by mixing GaN and InN to form $\text{Ga}_{1-x}\text{In}_x\text{N}$. The reason for the blue LED coming to the market decades later than the other colors was for a long time the difficulty of providing p-type GaN, and, directly related to it, p-type $\text{Ga}_{1-x}\text{In}_x\text{N}$. The history of p-type GaN is quite instructive and will be presented in short. Let us remark before that a blue light emitting diode was offered commercially already in the 1980s by the Siemens corporation based on 6H-SiC (bandgap 3.0 eV), a pioneering achievement at the time, but it had a very low internal quantum efficiency due to the fact that SiC is an indirect semiconductor.

The obvious way to produce p-type GaN would be by doping with an element of the second group like Mg. Mg would occupy substitutionally a Ga site and thus provide a shallow acceptor. But all attempts to do so failed for decades. The breakthrough came when it was discovered that this failure was due to the presence of hydrogen in the material. Hydrogen was present in GaN because the latter was generally produced with Ammonia NH_3 as a N source, resulting in an incorporation of hydrogen into the GaN. In the attempt to dope the material with Mg, instead of occupying substitutionally a Ga site, Mg formed a complex with hydrogen. It was the ingenious discovery of Akasaki and Amano and Nakamura,^{2,3} to notice that GaN:Mg turned p-type after it had been investigated in an electron microscope! The electron irradiation had broken the H-Mg-bond and thus “activated” the Mg acceptor. It turned out later that this activation by breaking the H-Mg-bond can also be achieved thermally which is more widely used now.

Another serious problem with the blue LED was the lack of a proper substrate for the epitaxial layers of GaInN. As discussed below, the growth of bulk GaN is not established as a technology at present, so “homoepitaxy”, the growth of GaN or GaInN on GaN wafers, is not possible at present industrially.

As a matter of fact, the tremendous development of nitride-based semiconductor technology (in optoelectronics as well as, e.g. in high frequency and power electronics) is based on the progress in methods of heteroepitaxy, a good example how the advances in crystal growth precede the progress in semiconductor technology. A first step into this direction was the use of 6H-SiC wafers as a substrate for GaN epitaxial films. This development was still close to conventional homoepitaxy because 6H-SiC and GaN both have the Wurtzite structure with $a = 3.073 \text{ \AA}$ for 6H-SiC and $a = 3.16 \text{ \AA}$ for GaN, corresponding to a fairly small.

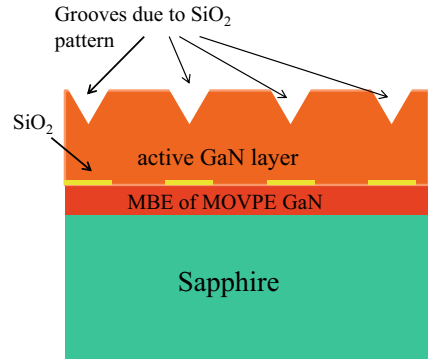
mismatch of about 2.8%. Actually, the use of 6H-SiC wafers as substrates for the blue nitride LED was the first large scale industrial application of SiC.

Completely new methods of heteroepitaxy were required, however, to make use of a comparably cheap material like sapphire as substrates for GaN and other nitride

² Amano et al [1].

³ I. Akasaki, H. Amano and S. Nakamura were awarded the Nobel Prize in Physics for “the invention of efficient blue light emitting diodes”.

Fig. 11.8 Principle of Epitaxial Lateral Overgrowth (ELO)



devices. The mismatch of sapphire versus of GaN is 15.8% measured in the most favorable crystallographic direction.⁴ To overcome this problem, completely new technologies had to be developed following the invention of the blue light emitting diode. A widely used method of that kind is the one of epitaxial lateral overgrowth (ELO) described in short in the following⁵ (Fig. 11.8).

First, a layer of GaN is grown on the sapphire substrate by MBE or preferentially MOVPE a few μm thick. Because of the huge mismatch, this layer may have as much as 10^{10} to 10^{11} cm^{-2} dislocations (see Sect. 6.3), far too many to provide a material for efficient LEDs. Instead a “mask” of, e.g. SiO_2 - or SiN_x - stripes (yellow in the figure) is produced followed by the epitaxial growth of the active layer. This layer growth first vertically above the regions left free between the SiO_2 stripes, then the regions above the SiO_2 stripes will be closed by lateral overgrowth. In these regions, the vertical growth is somewhat retarded leaving the grooves on top, which then have to be polished away.

Apart from these heteroepitaxy methods, developments are under way to provide GaN wafers for homopitaxy, i.d. for epitaxy of GaN or GaInN on GaN substrates, not so much for LEDs but rather for high-end electronic devices like transistors. If a fast epitaxial growth method like HVPE (Sect. 7.4) is applied, thick wafer-like samples with a dislocation density as low as 10^5 cm^{-2} can be achieved, well suited for electronic devices of different kind.

11.7 White LEDs

Given the fact that in the European Union, the traditional light bulbs are no longer permitted because of their low power efficiency, the white LED is generally considered to be the “light bulb of the future”.

⁴ Tassev [2].

⁵ Yu et al. [3].

A white light lamp could in principle consist of a chip carrying a red, a green, and a blue LED in the right intensity ratio. This, however, would not be expedient for some reasons: A relatively expensive production process would have to be applied, in combination with the installation of three separate power supplies for the three colors. Moreover, the color temperature of that lamp would not be stable over time, because the three LEDs would degrade differently with time. For these and other reasons, a much simpler construction is now in use; see Fig. 11.9a.

The left side of Fig. 11.9a shows the basic design. An GaInN-LED produces light in the near UV-blue spectral region as shown in Fig. 11.9a on the right side. The chip is encapsulated in an epoxy dome which contains in granular form some phosphor, generally $\text{Y}_3\text{Al}_5\text{O}_{12}$ (Yttrium aluminum garnet YAG) doped with Ce^{3+} . We will not enter into structure and physics of phosphors in general, nor that particular one. Figure 11.9b just shows very schematically the levels of $\text{YAG}:\text{Ce}^{3+}$ and the absorption and emission lines.

The blue light of the LED is partly absorbed, followed by an emission in the spectral region yellow–red, i.e. an orange emission. The superposition of the escaping fraction of the blue light and the orange emission gives an impression of white light in the eye of the observer, because blue and orange are complementary colors. This type of white light emitting diode which totally dominates the market nowadays

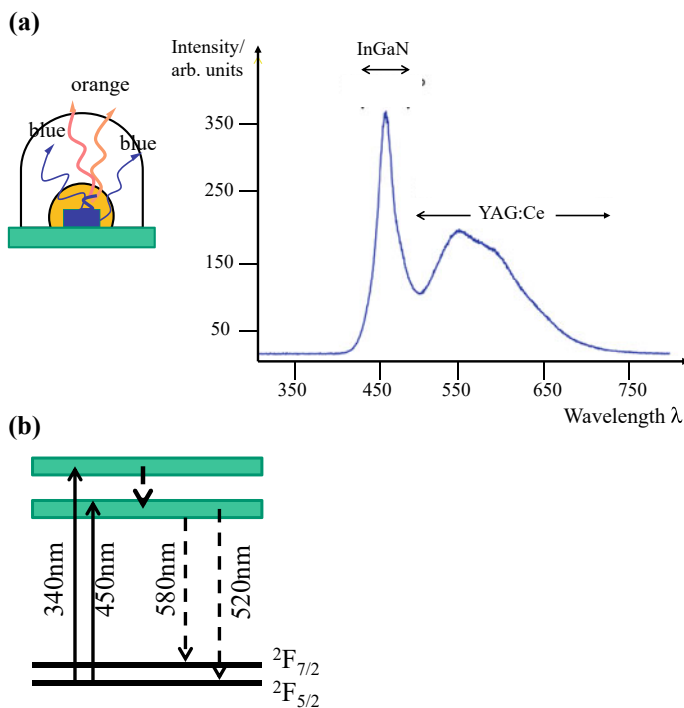


Fig. 11.9 (a) White LED. Design and emission spectrum. (b) Level scheme of Ce^{3+} in YAG

is referred to as “LUCOLED”, which stands for “Luminescence Converting Light Emitting Diode”.

Different from many claims found in literature, the inventor of the white LED was Juergen Schneider from the IAF Freiburg (Fraunhofer Institute of Applied solid State Physics). In first realization, he used an organic phosphor, but quickly switched to the inorganic YAG:Ce.

Nowadays, certain variations of YAG:Ce are in use in order to provide different color temperatures. YAG:Ce results in a fairly “cold” white, addition of Gd, for example, shifts the emission toward the red giving the white a “warmer” appearance, while, example given, the partial replacement of Al by Ga shifts it more to the blue.

Visible and IR LEDs are highly developed now, a major issue of today's research is the one of UV LEDs.

11.8 UV LEDs

The technical development of UV LEDs in terms of intensity and efficiency is far behind the one of LEDs in the visible and IR region, in spite of the fact that there would be a multitude of interesting applications. The UV covers by definition the wavelength regions from about 100–380 nm. Following the absorption wavelengths of certain trace gases in the atmosphere, this spectral range has been separated into three regions: UVA from 380 nm down to 315 nm, UVB from 315 to 280 nm, and UVC from 280 to 100 nm. An important application would be, e.g. curing of polymers, typically in the UVA and UVB range. A big issue in the UVC region is disinfection of water as well as materials and instruments, as required, e.g. in hospitals. By radiation with wavelengths around 265 nm and shorter, damage to the DNA of the bacteria is initiated. For many of these latter applications, mercury pressure lamps have been used or are being in use, but they are bulky, produce much heat, always provide a small risk because of the high pressure, and—last but not least—the mercury requires special precautions in the disposal.

Efficiencies of UV LEDs are far behind the ones in the visible. While blue light emitting diodes may have external quantum efficiencies as high as 70%, UV LEDs, depending on the classifications UVA, UVB or UVC, are limited to a few percent, with fast progress, however, being made.⁶

As far as materials are concerned, it is obvious from Fig. 11.7 that mixed crystals of AlN and GaN, i.e. $\text{Al}_{1-x}\text{Ga}_x\text{N}$ can cover almost the full UV range, from $x = 0$ (pure AlN) with a bandgap of 6.2 eV corresponding to $\lambda = 200$ nm (11.2), to $x = 1$ (pure GaN) with its bandgap of 3.4 eV corresponding to $\lambda = 365$ nm. The problem is the lack of appropriate substrates for these epitaxial layers. The actual technical development strongly relies on techniques like the ones described in Sect. 11.6,

⁶ Kneissl et al. [4].

e.g. ELO, which permit the use of “cheap” materials for the substrates like Si and sapphire overcoming in one way or the other the large lattice mismatch. Parallel to these efforts endeavors are under way to make GaN- or AlN- monocrystalline wafers available to UV technology, based on the idea that for high-end applications like UV-semiconductor-lasers homoepitaxial processes might be more promising. For the very short wavelength, which would require Al-rich $\text{Al}_{1-x}\text{Ga}_x\text{N}$ (x close to 0) this would amount to a growth on AlN wafers, for longer wavelengths in the regions of UVB and UBA, corresponding to a small Al-fraction, GaN might be appropriate as a substrate. Growth of bulk GaN and AlN is described in Sect. 7.4. In both cases, quality and diameter are still too unsatisfactory to decide on the potential and significance of homoepitaxy as compared to the heteroepitaxial methods.

11.9 Brightness of LEDs

Figure 11.10 shows the dramatic development of the brightness of LEDs of different colors with time for the period of the most dramatic development (mind the logarithmic scale!).

Remarkable, among other interesting facts, is the speed of catching up of the blue LED once the problem of p-doping of GaN had been understood (Sect. 11.6), starting from the SiC LED in the 1980s. It is evident that this development cannot go on forever for the simple reason that the efficiency cannot be higher than 100% in the sense that the power of light output cannot be higher than the electrical power fed into the LED. What is the limit in terms of lumens/Watt?

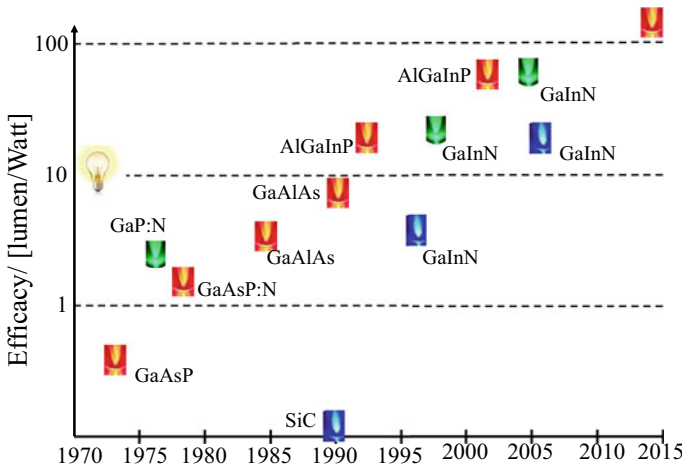
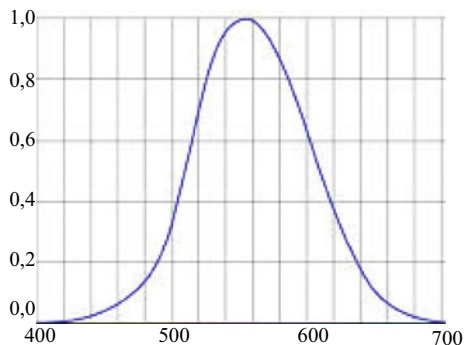


Fig. 11.10 Brightness development of LEDs in different spectral regions

Fig. 11.11 Spectral sensitivity $\lambda S(\lambda)$ of the human eye as function of wavelength (wavelength in nm)



Physics comment. The Lumen as unit of “brightness” of a light source

The “lumen” is a widely used unit to describe the brightness of a lightsource. The definition of the lumen runs as follows:

A light source of a “luminous flux” (or “brightness”) of 1 lm emits a power of 1/683 W in total space at a wavelength of 555 nm.

This definition of the unit “lumen” sounds somewhat special and even queer. The background is the following: The intention is to define a unit for the brightness of a lightsource. What does this mean? It means that we want to connect the fairly subjective impression of “brightness” of the light source with the power emitted by the light source in form of light. This, evidently, is done in the definition given above.

The definition refers to a certain wavelength of the light. This is because the impression of brightness in our eyes is strongly depending on the wavelength. In order to produce the impression of a bright light source much more power in terms of watt is required in the blue than, e.g. in the green, because the human is much more sensitive in the green than in the blue. This means that the connection between the impression of “brightness” and the power of the light source given in Watts has to be given for a defined, specific wavelength. This specific wavelength is chosen as $\lambda = 555 \text{ nm}$, because this corresponds to the maximum of the sensitivity of the human eye.

But why this number 1/683? The reason is the following: If it is done that way, **the new unit lumen, as defined above, is directly related to the traditional unit candela in the way that a totally isotropic light source of 1 cd emits a total luminous flux (i.e. has a brightness) of 4π lumen.** It is the general procedure in defining a new unit that it is done in a way which produces a direct connection with the old unit, which makes it easy to compare statements given in the new unit with those given in the old one. To go a further step backwards: The “classical” unit candela was chosen in such a way that it corresponds roughly to the brightness of a candle.

“Brightness” or “luminous flux” for wavelengths other than 555 nm:

The definition given above is equivalent to the following statement: **A light source with $\lambda = 555$ nm, which transforms 1 W of electrical power 100% into light, has an efficacy of 683 lm/W.**⁷

This concerns the efficacy at $\lambda = 555$ nm. For any other wavelength, the sensitivity curve $S(\lambda)$ shown in Fig. 11.11 of the human eye comes into play.

$S(\lambda)$ equals 1 at the wavelength 555 nm of maximum sensitivity, and tends toward 0 at the limits of the spectral sensitivity of the human eye in the UV and the IR. The relevant definition for arbitrary wavelength then reads: **A light which transforms 1 W of electrical power 100% into light, has an efficacy**

$$\eta(\lambda) = S(\lambda) * 683 \text{ lm/Watt} \quad (11.8a)$$

Let the fraction of electrical power which is transformed into light be $\varepsilon(\lambda)$, the electrical power $P(\lambda)$ and the brightness $H(\lambda)$, then an electrical power $P(\lambda)$ will produce a brightness $H(\lambda)$ of

$$H(\lambda) / \text{lumen} = \varepsilon(\lambda) \eta(\lambda) P(\lambda) = \varepsilon(\lambda) * S(\lambda) * 683 \text{ lm} * P(\lambda) / \text{Watt} \quad (11.8b)$$

Numerical example

How much electrical power of a 80% efficacy light source is required to produce 500 lm (a) in the green at 555 nm and (b) in the blue at 480 nm?

From 11.8b, one obtains

$$P(\lambda) / \text{Watt} = H(\lambda) / \text{lm} / (\varepsilon(\lambda) S(\lambda) 683 \text{ lm})$$

(a) $H(\lambda) = 500 \text{ lm}$, $\varepsilon(\lambda) = 0.8$, $S(\lambda = 555 \text{ nm}) = 1$, $S(\lambda = 555 \text{ nm}) = 1$, so

$$P(555 \text{ nm}) = 0.92 \text{ W}$$

(b) $S(\lambda) = 0.15$ (Fig. 11.11), the other parameters being equal to a), so

$$P(480 \text{ nm}) = 6.13 \text{ W}$$

The little example shows that it is considerably harder to produce a bright blue light source than a bright green light source. This is so because, the definition of

⁷ The reader might expect here the word “efficiency” instead of the rarely used “efficacy”. However, in connection with light sources the technical term “efficacy” is used when the brightness of the “pure” light source is discussed in terms of lumen/Watt, while the term “efficiency” includes the fixture of the light source. Any light source, be it an incandescent bulb or an LED, has to be mounted somehow resulting in a loss of usable light. So in the following we shall use the technical term “efficacy” characterizing the “pure” light source.

brightness (lumen) involves the sensitivity of the human eye, which is much smaller in the blue than in the green.

Physics comment: Candela and Lumen: The candela as a “fundamental unit”

It has been mentioned above that the definitions of lumen and candela are directly related to each other. An interesting point is that the candela (and directly with it the lumen) shows directly up in the list of fundamental units of physics. As a short comment to the main components of that list, let us notice that it includes the seven units: meter, kilogram, second, Ampere, Kelvin, mole, and candela. Most recently, a new definition has been given to the kilogram, relating it to the value of Planck’s constant h by fixing the value of the latter by definition. As a little demonstration how these fundamental units are interrelated, let us consider in which way the definition of the kg is related to Planck’s constant and to the fundamental constants meter and second.

Planck constants has the dimension $[h] = [\text{power}] \times [\text{time}] = (\text{kgm/s}^2)\text{ms} = \text{kgm}^2/\text{s}$.

So, by the fixation of the second via a hyperfine frequency of Cs¹³³:

$\Delta\nu_{\text{Cs}} = 9192631770 \text{ s}^{-1}$, the velocity of light to.

$c = 299792\,458 \text{ m/s}$, and the recent fixation of h to the numerical value.

$h = 6.62607015 \times 10^{-34} \text{ kgm}^2/\text{s}$, we can write

$$1 \text{ kg} = (h\Delta\nu_{\text{Cs}}/c^2)(299792458)^2 / [(6.62607015 \times 10^{-34})(9192631770)]$$

which demonstrates how the kg is related to the fundamental units s and m after a fixation of h , showing how closely the fundamental units are interrelated.

So it is natural for the reader to ask what is so fundamental about the candela that it shows up on an equal footing with Planck’s constant or the velocity of light. The answer is simply: Nothing. Nothing is fundamental about the candela. It was just that in the nineteenth century, when the systematics of fundamental units was installed, the illumination industry was an important branch of industry, and they thought it useful to have a definition of the luminous flux.

Brightness (or luminous flux) limit of a white LED, present state of the art

It is a fact that the white LEDs now approach a brightness of about 100 lm/W. What is the physical limit?

In Sect. 11.7, it is described that the commercial white LED consists of a blue LED and a surrounding epoxy layer with an admixed phosphor which transforms part of the blue light into orange light. For a simple estimate of the brightness limit, let us take $\lambda = 480 \text{ nm}$ for the blue light and $\lambda = 590 \text{ nm}$ for the orange light. From Fig. 11.11, we read $S(480 \text{ nm}) \approx 0.15$ and $S(590 \text{ nm}) \approx 0.75$. It turns out from practical experience that an impression of white light is obtained if (roughly) 40% of

the total light energy are emitted in the blue and 60% in the orange. (Variations of a few percent result in different perceptions of white like “cold white”, corresponding to more blue, or “warm white”, corresponding to more orange),

In order to calculate the maximum luminous flux, we start from the assumption that 1 W of electrical power is converted into 1 W of light power. That means the luminous flux $F(\text{blue})$ emitted in the blue is, according to 11.7

$$F(\text{blue}) = 0.15 * 0.4 \text{ W} * 683 \text{ lm/W} = 41 \text{ lm}, \text{ — and in the orange}$$

$$F(\text{orange}) = 0.75 * 0.6 \text{ W} * 683 \text{ lm/W} = 307 \text{ lm}$$

So the total luminous flux of such a white LED with 100% efficacy would be 348 lm/W.

For a precise calculation, we would have to average $S(\lambda)$ over the emission spectrum of the blue LED to obtain $S(\text{blue})$, and average $S(\lambda)$ over the emission spectrum of the phosphor to obtain $S(\text{orange})$. We would also vary slightly the 40% blue light power depending on the intention to obtain a “warmer” or a “colder” light. But in any case, this upper limit of 350 lm/W would remain a good guess. At present, about 100 lm/Watt are state of the art,⁸ but something like half of the upper limit, i.e. 150–190 lm/W seem technically feasible as luminous flux of white light emitting diodes.

Today white LEDs are so efficient that they are widely used already in high power applications like front lights of automobiles, street illuminations, and many others!

Let us finally note that not only efficacy, but also lifetime is a most impressive advantage of the LED as compared to the classical incandescent light bulb. The latter shows an efficacy of about 15 lm/Watt with a lifetime of roughly 1000 h, while the corresponding numbers for a modern white LED would be 100 lm/Watt with a lifetime of roughly 60,000 h!

11.10 Questions and Answers

Questions

- Question 1: What is the relationship between emission wavelength in nm and bandgap of an LED-material in eV? Give the formula!
- Question 2: Which kind of defects are particularly efficient as nonradiative recombination centers?
- Question 3: Give the main reason for the difference between internal and external quantum efficiency?

⁸ Detailed numbers for efficacy and “usable lifetime” are given in Paul Morgan Pattison, Monica Hanson, Jeffrey Y. Tsao: LED lighting efficacy: Status and directions, *Comptes Rendus Physique* **19** (3) 2018.

- Question 4: What determines basically the linewidth of the band-band transition?
- Question 5: What is the size of kT at room temperature?
- Question 6: What was the problem resulting in the retarded development of the blue LED?
- Question 7: What caused the problem of p-doping of GaN?
- Question 8: What is the basic set-up of a white LED?
- Question 9: Which semiconductor material is suitable for UVC LEDs?
- Question 10: What is the unit of “luminous flux”.

or “brightness” of a light source? Give the definition!

- Question 11: Why is it harder to make a bright light source in the blue than in the green?

Answers

- Answer 1: $\lambda/\text{nm} = 1240/(E_G/\text{eV})$.
- Answer 2: Those which give rise to a level close to the middle of the bandgap.
- Answer 3: Total internal reflection at the interface between semiconductor and air.
- Answer 4: The density of states at the band edges of CB and VB and their occupancy at the respective temperature, resulting in a spectral width of ca. 1.8 kT .
- Answer 5: 26 meV .
- Answer 6: The problem of p-doping of GaN.
- Answer 7: Presence of H leading to a passivation of the Mg acceptor by hydrogen.
- Answer 8: The one of a “LUCOLED”: A blue LED encapsulated in an epoxy which contains a phosphor transforming blue light partly into orange light.
- Answer 9: $\text{Al}_{1-x}\text{Ga}_x\text{N}$.
- Answer 10: The lumen (lm). Its definition: A light source with $\lambda = 555\text{nm}$, which transforms 1 W of electrical power 100% into light, has an efficacy of 683 lm/Watt . At another wavelength λ its efficacy is $S(\lambda) \cdot 683\text{ lm/Watt}$, where $S(\lambda)$ is the spectral sensitivity of the human eye.
- Answer 11: Because the spectral sensitivity of the human eye is much smaller in the blue than in the green.

References

1. H. Amano, M. Kito, K. Hiramatsu, I. Akasaki, *Jpn. J. Appl. Phys.* **28**, L2112–L2114 (1989)
2. V.L. Tassev, *Crystals* **7**(6), 178 (2017). <https://doi.org/10.3390/cryst7060178>
3. Z. Yu, M. Johnson, J. Brown et al., Epitaxial lateral overgrowth of GaN on SiC and sapphire substrates. *MRS Internet J. Nitride Semicond. Res.* **4**, 447–452 (1999). <https://doi.org/10.1557/S1092578300002878>
4. M.Kneissl, T. Seong, J. Han, H. Amano, The emergence and prospects of deep-ultraviolet light emitting diode technologies. *Nat. Photon.* **13**, 233–244 (2019). <https://doi.org/10.1038/S41566-019-359-9>

Chapter 12

Semiconductor Lasers



Abstract Semiconductor lasers play an important role in technical applications like data recording, data reading, laser printing, optical devices like pointers and microscopes, and all kind of fabrication processes like polymer hardening, 3d-printing, structuring, and many others. Another important field of application is medicine (surgery, in particular ophthalmology). Laser action sets in a situation far away from thermal equilibrium, characterized by different Fermi energies of valence band and conduction band. This difference should be larger than the band gap in order to fulfill the “laser condition”, a situation which can be achieved in a pn-junction formed by highly doped p- and n-type material and operated at sufficiently high forward current. At a certain forward current, the “laser threshold current”, laser action sets in. Beyond the laser threshold, the emitted light is characterized by extreme monochromaticity and directionality as a consequence of the fact that the photons essentially originate from stimulated emission. The “threshold current” can be considerably lowered in more sophisticated arrangements like double heterojunctions, due to “charge confinement” and “light confinement” in the active region, or multi-quantumwell lasers. Blue lasers based on GaInN are at the basis of “Blue ray” data storage. Long distance optical data transmission via optical fibers makes use of IR-lasers operating at $1.54\text{ }\mu\text{m}$ on the basis of the quaternary compound GaInAsP.

12.1 The Laser Principle

“Laser” stands for “**L**ight **a**mplification by **s**timulated **e**mission of **r**adiation”. Semiconductor lasers play an important role in modern data and communication technology. Important applications are CD, DVD, and Blue ray data storage and optical data transmission via optical fibers. Many other applications could be listed, like laser printing, range finding, the laser microscope, and other optical devices of all kind. The laser nowadays is an important tool in surgery, in particular, in surgery of the eye.

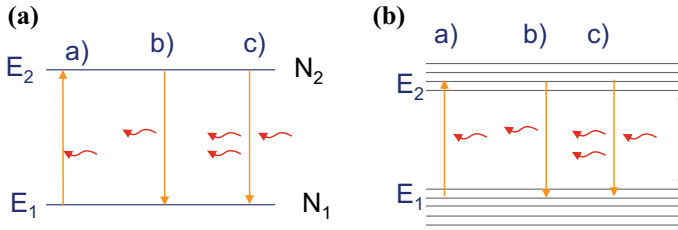


Fig. 12.1 **a** The basic interactions between light and matter in an atom or molecule. **b** The basic interactions between light and matter in a solid

The physics of the semiconductor laser is basically the same as the one of any other laser system^{1,2} but includes some essential differences. So it might be instructive to discuss the semiconductor laser in a direct comparison to atomic or molecular lasers. In this way, similarities as well as differences will get visible.

Figure 12.1a gives the three basic interactions between light and matter in an atom or molecule according to Einstein's historical treatment. Figure 12.1b shows the optical transitions for the case of a solid with its energy levels arranged in energy bands.

Einstein interpreted the famous formula of Max Planck for the black body radiation as a manifestation of an equilibrium between *three* processes:

- (a) Absorption,
- (b) Spontaneous emission,
- (c) Stimulated emission.

where the third one (stimulated emission) had not been considered before. Only by including stimulated emission Planck's formula could be reproduced.

Let us first consider the case of atoms and molecules (Fig. 12.1a). Absorption can take place if the energy difference $E_2 - E_1$ between two electronic states of an atom or molecule equals $h\nu$, ν being the light frequency. Let the energy density of the light field at frequency ν be $\rho(\nu)$. N_1 is the number of atoms in the interaction zone between light and matter in the ground state with energy E_1 , N_2 the number of atoms in the excited state with energy E_2 .

The three processes (a), (b), and (c) then can be quantified in the following way:

The number Z_{12}^{absorb} of absorption processes per unit time is proportional to N_1 and $\rho(\nu)$, so

$$Z_{12}^{\text{absorb}} = B_{12}N_1\rho(\nu) \quad (12.1a)$$

The constant of proportionality B_{12} is named "Einstein coefficient of absorption".

The number Z_{21}^{spontan} of spontaneous emission processes per unit time is proportional to the number N_2 of atoms in the excited state with energy E_2

¹ Svelto, Orazio [1].

² Winnacker, Albrecht [2].

Fig. 12.2 Amplification of light in a medium



$$Z_{21}^{\text{spontan}} = A_{21}N_2 \quad (12.1b)$$

A_{21} is the “Einstein coefficient of spontaneous emission”.

The new process postulated by Einstein in order to reproduce Planck’s formula is “stimulated emission”. Here a photon of energy $h\nu$ stimulates the transition of an excited atom to the ground state under emission of a photon identical to the stimulating one. In a strict discussion the notion of an “identical” photon would have to be replaced by “a photon belonging to the the same mode of the light field” (for details see the physics comment “On the high directionality and monochromaticity of laser light” in Sect. 12.3). So the number of stimulated emissions per unit time is proportional to N_2 and the energy density of the stimulating light

$$Z_{21}^{\text{stimulated}} = B_{21}N_2\rho(\nu) \quad (12.1c)$$

B_{21} is called the “Einstein coefficient of stimulated emission”. It turns out—without being derived here—that the requirement of an equilibrium between these three processes results in Planck’s formula only under the condition $B_{12} = B_{21}$. In today’s understanding, by the way, this equality is a consequence of the basic quantum mechanical rule that the squares of the transition matrix elements for forward- and reverse processes are equal, which, in turn, is an expression of the principle of time reversal symmetry. The principle applies here because of the fact that stimulated emission is the reverse process of absorption.

In the next step, we consider the case where a light beam crosses a medium of atoms or molecules which absorb and emit this light, and ask the question: **Under which condition is the light amplified when crossing the medium** as symbolized in Fig. 12.2?

Evidently the light is amplified when the number of emission processes exceeds the number of absorption processes, i.e.

$$Z_{21}^{\text{spontan}} + Z_{21}^{\text{stimulated}} > Z_{12}^{\text{absorb}} \quad (12.2a)$$

Or, by inserting (12.1a)–(12.1c)

$$A_{21}N_2 + B_{21}N_2\rho(\nu) > B_{12}N_1\rho(\nu) \quad (12.2b)$$

At this point a simplification is made: Eq. (12.2b) is certainly fulfilled if one of the summands on the left side is simply omitted, i.e. if $B_{21}N_2\rho(\nu)$ alone is already larger than $B_{12}N_1\rho(\nu)$. This is trivial, on one hand. On the other hand, there is an important physical insight behind the justification of that step. What we do in this way is to omit the spontaneous emission relative to the stimulated emission. The point is that this

approximation does not affect significantly the quantitative answer to our question, because it turns out that under condition of light amplification (which, as will be seen in a moment, amounts to the condition of laser action) stimulated emission is always the strongly dominating effect as compared to spontaneous emission.

So our simplified condition now reduces to $B_{21} N_2 \rho(\nu) > B_{12} N_1 \rho(\nu)$, or, because of $B_{12} = B_{21}$

$$N_2 > N_1 \quad (12.3)$$

This is the condition for light amplification. As will be shown below, the condition for light amplification is identical to the one of laser action. So (12.3) is also called the “laser condition”. Let us summarize: **In order to achieve a light amplification (or laser action) a situation of “inversion” is required, more atoms or molecules must be in the excited state than in the ground state of the optical transition.** We shall come back to that point in the next “physics comment”.

Next the case of a solid is considered. In a solid in general and a semiconductor in particular, the energy states are ordered in “energy bands” (see Chap. 1), where each band consists of innumerable (but distinct!) energy states as indicated very schematically in Fig. 12.1b. Let E_2 be the energy of one of those states in the conduction band, and E_1 be one of the state in the valence band.

The difference in treating this case as compared to the one of the molecular or atomic state laser (Fig. 12.1a) is that we replace the number of atoms N_2 in the excited state by the probability $f(E_2)$ that the conduction band state under consideration is occupied, and N_1 by the probability $f(E_1)$ that E_1 is occupied. Then the number of absorption processes per time will be given by

$$Z_{12}^{\text{absorp}} = K_{12} f(E_1) [1 - f(E_2)] \rho(\nu) \quad (12.4a)$$

Here it has been taken into account that the absorption process into E_2 can only take place if the final state in the conduction band is unoccupied, the probability of which is $1 - f(E_2)$. K_{12} is the “Einstein coefficient of absorption”.

The number of spontaneous decays per unit time is given by

$$Z_{21}^{\text{spontan}} = L_{21} f(E_2) [1 - f(E_1)] \quad (12.4b)$$

because the initial state must be occupied (factor $f(E_2)$) and the final state must be empty (factor $[1 - f(E_1)]$). L_{21} is the “Einstein coefficient of spontaneous emission”.

And, finally, by the analogous consideration, the number of stimulated emission process per unit time is

$$Z_{21}^{\text{stim}} = K_{21} f(E_2) [1 - f(E_1)] \rho(\nu) \quad (12.4c)$$

K_{21} is the “Einstein coefficient of stimulated emission”.

For the same arguments as given above for the relation $B_{12} = B_{21}$, we can state that $K_{21} = K_{12}$.

Again, as above, we ask under which condition light, which crosses the solid (see Fig. 12.2), is amplified. The condition that the number of emission processes should be larger than the one of the absorption processes leads to

$$Z_{21}^{\text{spontan}} + Z_{21}^{\text{stimulated}} > Z_{12}^{\text{abs}}$$

which by use of (12.4a)–(12.4c) gives

$$L_{21}f(E_2)[1 - f(E_1)] + K_{21}f(E_2)[1 - f(E_1)]\rho(\nu) > K_{12}f(E_1)[1 - f(E_2)]\rho(\nu)$$

As above, we neglect the spontaneous emission relative to the stimulated emission

$$K_{21}f(E_2)[1 - f(E_1)]\rho(\nu) > K_{12}f(E_1)[1 - f(E_2)]\rho(\nu)$$

Because of $K_{12} = K_{21}$, this simply results in

$$f(E_2) > f(E_1) \quad (12.5)$$

This is the “laser condition” for a solid state system with its band structure. The occupation probabilities of the upper states (in the conduction band) must be larger than the ones in the lower states (in the valence band).

Physics Comment: The Laser, A System Far Away from Thermal Equilibrium.

The conditions (12.3) and (12.5) for laser action correspond to a situation very far away from thermal equilibrium. The ratio of occupation numbers or occupation probabilities in thermal equilibrium is given by the Boltzmann factor $e^{-\frac{E_2 - E_1}{kT}}$. (as long as $E_2 - E_1 \gg kT$, otherwise Fermi–Dirac would have to be applied). If $E_2 - E_1 = 2 \text{ eV}$, then, at room temperature, $(E_2 - E_1)/kT \approx 80$, resulting in a ratio of 10^{-35} . Given the fact that one mole corresponds to roughly 10^{24} atoms or molecules, it is obvious that in thermal equilibrium not even a single atom (or molecule) is in the excited electronic state in thermal equilibrium at room temperature. As a matter of fact, the system must be brought into that state by “pumping” energy into it, a laser must be “pumped” in order to realize the laser condition. In the case of a Helium–Neon laser, this pumping mechanism is provided by a He–Ne gas discharge, pumping the necessary energy into the system. In the case of the semiconductor laser, the energy source for the pumping process is the power supply for the current sent through the pn-junction.

So (12.3) and (12.5) describe a situation of extreme disequilibrium. In the case of a semiconductor, generally a special case of disequilibrium is encountered, which will be considered next. This situation is the following: A thermal equilibrium exists among the electrons in the conduction band, and also among the electrons in the valence band, but no equilibrium is established between the bands. If, for instance, electrons are excited by a strong light source from the valence band to the conduction band, at a rate much larger than the rate of recombination, the ratio of the number of electrons in the conduction band to those in the valence band might strongly deviate from equilibrium, i.e. from the Fermi–Dirac distribution. However, because of a high scattering rate between electrons within the conduction band, thermal equilibrium might be established within the conduction band, that means a Fermi–Dirac distribution is established among the conduction band electrons, given by

$$f_{CB}(E_2) = 1 / [\exp(E_2 - F_{CB}) / kT + 1] \quad (12.6a)$$

where F_{CB} is the Fermi energy of the conduction band.

If in that same material there is a sufficiently high scattering rate among the electrons in the valence band, a thermal equilibrium might be established between the electrons given by

$$f_{VB}(E_1) = 1 / [\exp(E_1 - F_{VB}) / kT + 1] \quad (12.6b)$$

where F_{VB} is the Fermi energy in the valence band.

At first sight, it might seem surprising that here, in the context of the valence band, we speak of energy distribution of electrons, not of holes. But, of course, the “energy distribution of the holes” is the result of scattering processes among the electrons of the valence band and their resulting energy distribution.

The fact that thermal equilibrium exists within the bands, but not between the bands, is reflected by the fact that the energy distribution can be described by a Fermi–Dirac distribution in both bands, the Fermi energies of the two bands being different, however. In case of thermal equilibrium of the material as a whole they would be identical.

Equations (12.6a) and (12.6b) will now be inserted into the laser condition (12.5).

$$1 / [\exp(E_2 - F_{CB}) / kT + 1] > 1 / [\exp(E_1 - F_{VB}) + 1]$$

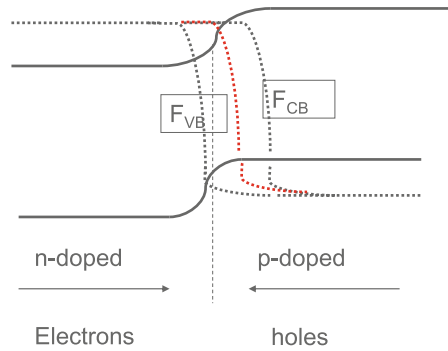
This is the case if $E_1 - F_{VB} > E_2 - F_{CB}$

$$E_2 - E_1 < F_{CB} - F_{VB}$$

Because E_2 is the energy of a state in the conduction band, and E_1 is the one of a state in the valence band, $E_2 - E_1 \geq E_G$ for all E_1 and E_2 , this amounts to

$$E_G < F_{CB} - F_{VB} \quad (12.7)$$

Fig. 12.3 Realization of the laser condition (12.7)



To put it into words: **The laser condition in a semiconductor is fulfilled if the difference between the Fermi energies of the conduction band and the valence band is larger than the bandgap.** This situation can be realized in the depletion zone between highly doped n- and p-semiconductors. Figure 12.3 shows the band scheme and the Fermi level around the pn-junction in the presence of an external voltage (corresponding to the difference in the Fermi levels at the very right and left of the figure).

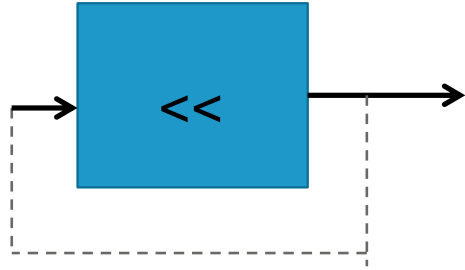
Without external voltage, the Fermi levels would be equal. The dotted red line represents the behavior of the Fermi level in case of thermal equilibrium between the two bands. If, however, a strong current in the forward direction is present, the Fermi levels of CB and VB get separated, expressing the fact that there is no longer an equilibrium situation between the bands.

It is obvious from Fig. 12.3 that the situation described by the laser condition (12.7) can only arise if the material is so highly doped on the n- and the p-side that the respective Fermi levels are positioned within the conduction band and the valence band. This usually happens at donor or acceptor concentrations above ca. 10^{20} cm^{-3} . That the Fermi level is positioned within the band, not in the bandgap, is exceptional for a semiconductor, but characteristic for a metal (Fig. 1.4 in Chap. 1). In that case, we speak of a “degenerate semiconductor”. **So the laser condition can be realized in a pn-junction of degenerate semiconductor materials at sufficiently high currents in the forward direction.**

12.2 Light Amplification and Lasing

In the previous section, the conditions for light amplification were derived, see (12.3) and (12.5). The usual application of a laser system, however, is not the one of a light amplifier (in spite of the “amplification” in the name!), but of a light source. It was already mentioned in the context of (12.3) that the two applications are closely related in the sense that the conditions, (12.3) and (12.5), apply to the case of a light emitter as well. This is due to a very general rule of communication technology: **Any**

Fig. 12.4 Principle of feedback



amplifier can be transformed into an oscillator by the principle of feedback. So if the condition for light amplification is fulfilled, the one for laser action is fulfilled as well if proper feedback is provided.

This important point is explained in Fig. 12.4.

Imagine an ac amplifier, where a fraction of the output is coupled back into the input. If this happens with the right phase, oscillation will set in. An example may illustrate this: Let us assume that the amplifier amplifies the input power by a factor of 5. If one fifth of the output is coupled back into the input with the right phase, a stable oscillation will set in.

So, in order to transform a light amplifier (Fig. 12.2) into an oscillator, somehow the idea of feedback has to be applied, i.e. a fraction of the output light has to be fed back into the input. This is done by putting the laser medium into an optical resonator; see Fig. 12.5.

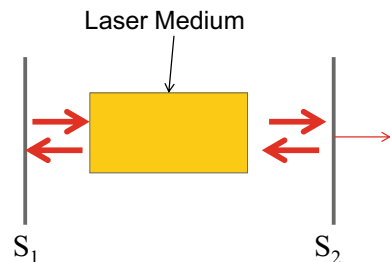
By the mirrors, a high fraction of the amplified light is sent back into the amplifying medium. This is just what we called “feedback” previously.

Mirror S_2 has a reflectivity clearly below 100% (but as high as 98-99%), so a fraction of the light leaves the resonator and thus provides the emitted laser beam.

The mirrors of the resonator in case of a semiconductor laser can just be provided by cleavage surfaces of the semiconductor crystal. According to Fresnel’s formulas, the reflectivity of such a plain under vertical incidence is given by

$$R = [(n - 1)/(n + 1)]^2$$

Fig. 12.5 Laser resonator



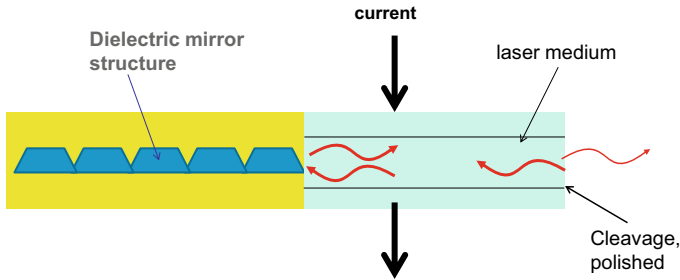


Fig. 12.6 Dielectric mirror as part of the resonator

where n is the index of refraction of the crystal. With e.g. $n = 3.8$ for GaAs this gives $R = 0.35$, a fairly small value, but sufficient for many practical cases. Much better is the widely used structure Fig. 12.6.

One of the mirrors (the highly reflecting mirror S_1 of Fig. 12.5) is realized as a dielectric mirror, i.e. a system of many layers with a periodicity of the index of refraction of $\lambda/2$. Such a system can achieve a reflectivity near 100% for a sufficient number of layers. There are different ways to achieve the periodic variation in the index of refraction. One would be a variation of the material, e.g. in the case of AlGaAs a periodic change in the composition of $\text{Al}_{1-x}\text{Ga}_x\text{As}$. Another one would be a periodic variation of the doping level, because the index of refraction slightly depends on that parameter. This kind of laser structure is called “distributed feedback laser”, in which one of the resonator mirrors being “distributed” in space.

The light emitted by a laser differs in a dramatic way from one of the conventional light source. Fig. 12.7a shows the light intensity as a function of the driving current below and above the threshold.

Below the threshold current, the system behaves essentially like a LED, for not too high intensities, the light power increases approximately linearly with the current (red line in fig. 12.7a). However, in a laser diode, once the laser threshold is exceeded,

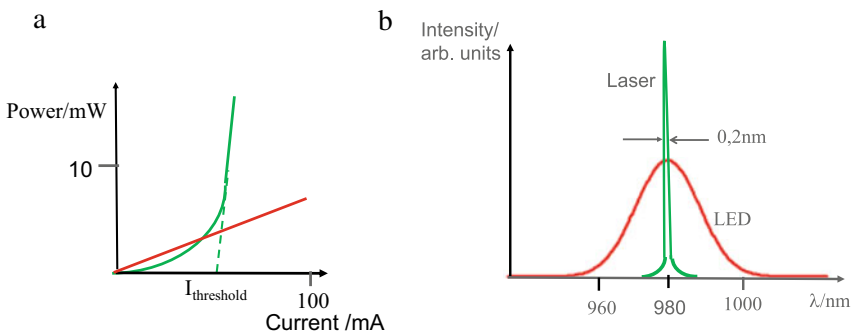


Fig. 12.7 **a** Comparison LED-laser: Output versus driving power. **b** Comparison LED-laser: Emission spectrum

the light power increases rapidly in a highly nonlinear way (green line). This is due to the fact that beyond the laser threshold rapidly the stimulated emission takes over being a nonlinear selfenhancing process.

Fig. 12.7b shows the development of the spectrum of the emitted light beyond the threshold. The data, by the way, refer to a InGaAs laser with wavelength 980 nm which can be used to optically pump the Erbium repeater as described in Sect. 12.5.

While the linewidth is of the order of some tens of nanometers below the threshold, comparable to an LED (Sect. 11.2), it shrinks toward a few tenths of nm above the threshold. This extremely narrow linewidth is a general feature of laser light. It has to do with a property of stimulated emission not mentioned so far: The additional photon emitted under stimulated emission has almost the same direction and wavelength as the stimulating photon. “The same direction” results in the extreme directionality of laser light which surprises everybody when she sees a laser in action for the first time. “Almost the same wavelength” results in the extreme monochromaticity of the light shown in Fig. 12.7b.

Physics Comment: On the High Directionality and Monochromaticity of Laser Light

From the point of view of basic physics, the statement “the stimulated photon has almost the same direction and wavelength as the stimulating one” has to be reformulated in a more precise way based on basic laws of statistical mechanics. The light field in the resonator exists in form of many “modes” or “Eigenschwingungen” of the resonator resulting from the boundary conditions of the electromagnetic field on the mirrors. Below the laser threshold, a large number of these modes are “filled” by the photons through the spontaneous emission. Above the threshold, stimulated emission takes over. Photons are Boson particles. According to the general rules of Bose statistics, the probability of creating a Boson is proportional to $\sqrt{n(n+1)}$ where n is the number of photons already present in the final quantum mechanical state of the emission process. The “final quantum mechanical state” in this case is a mode of the resonator. So stimulated emission goes preferentially into the resonator mode with the highest occupation number of photons. As a consequence, the photons accumulate just in one mode, the strongest one (i.e. the one with the lowest optical losses). So they have the high directionality and narrow spectral width corresponding to that one single mode. In that sense, a laser is often described as a highly ordered system: Essentially all the photons in the resonator are accumulated in one mode, different from the “messy” situation below the laser threshold with its many different active modes.

Physics Comment: Some General Remarks Concerning Lasers

Often the laser is described as an analogy to the situation encountered much more generally in highly ordered systems, e.g. the living cell. This idea of the laser as an analogy to ordered systems in general and biological systems, in particular, has been pioneered by Hermann Haken.³ The analogy between the laser and living organisms can be summarized as follows:

1. It is a system far away from thermal equilibrium.
2. To maintain that situation far away from thermal equilibrium, a continuous energy supply is required (corresponding to the pumping process of the laser mentioned above) in a living cell provided by the metabolism.
3. A high degree of order has developed as a result of feedback processes (ideally all photons are in one resonator mode). Correspondingly, the highly cooperative chemistry of the living cell is characterized by many of these “self reproducing” processes.

12.3 Double Heterojunction Laser

The simplest semiconductor laser structure would just be a straight pn-junction (Fig. 12.8a) plus the resonator mirrors **Comment:** In my manuscript the 4 figures 8a,b,c,d are clearly distinct and each accompanied by their own figure caption. This clarity is lost here! Please mark clearly which figure is 12.8c and 12.8d, as given in the manuscript.

This simple design, however, would require a very high threshold current density of typically 10^5 A/cm^2 . The corresponding high Joule heat would not be tolerated by any material at least under dc-operation. The most common design to provide much higher efficiency and, consequently, a much lower threshold current is the “double heterojunction laser”. Figure 12.8(b) shows the set-up with the AlGaAs system as an example.

A heterojunction is a junction of two different semiconductors. The figure shows a device consisting of two such heterojunctions (hence “double heterojunction laser”), one between p-AlGaAs and p-GaAs, the other one between p-GaAs and n-AlGaAs. The characteristic feature of this design is that a material with a smaller bandgap (i.e. GaAs with $E_G = 1.42 \text{ eV}$ in the example) is sandwiched between two layers with a higher bandgap (here AlGaAs with E_G between 1.42 and 2.2 eV depending on the Al content). As seen in Fig. 12.8c, in this case, a barrier for the electrons develops between the GaAs and the p-AlGaAs, and a barrier for the holes between the GaAs and the n-AlGaAs.

³ See e.g. Haken [3].

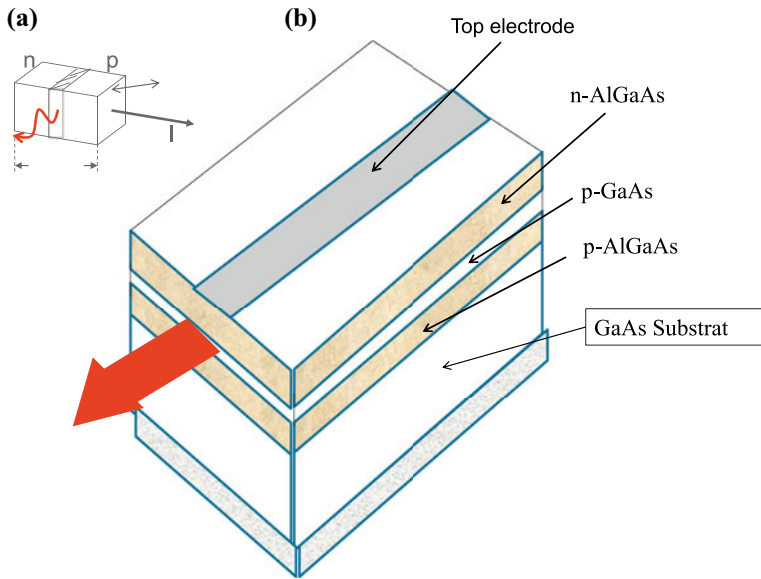
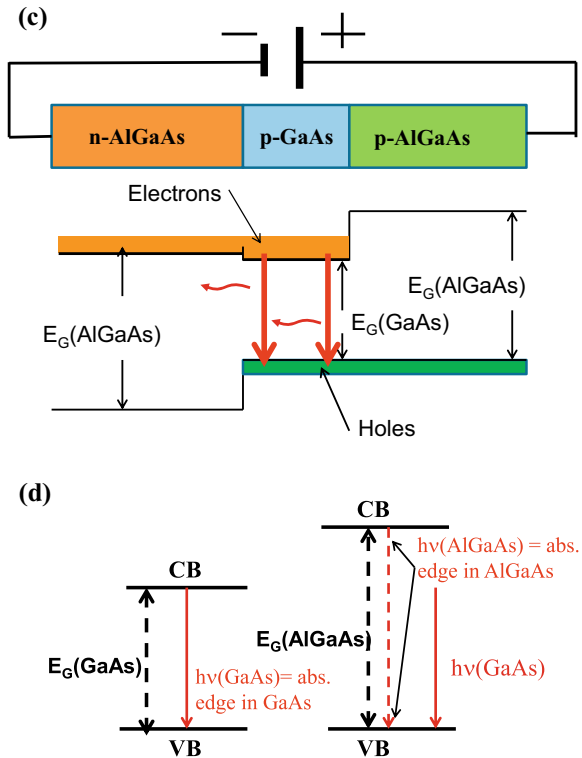


Fig. 12.8 (a) Semiconductor laser as a simple pn-structure (b) Double heterostructure laser (c) Double hetero junction: Design (above and energy band scheme (below), (d) On the difference of indices of refraction in GaAs and GaAlAs

As a result, electrons and holes accumulate in the sandwiched GaAs-layer leading to high electron- and hole concentrations in this layer, and consequently to a high recombination rate (the latter being $\sim np$). In this way, threshold current densities as low as $10^2\text{--}10^3\text{ A/cm}^2$ can be achieved, making efficient laser action and, in particular, dc-operation possible (i.e. a cw-light beam). The capability of the double heterostructure to accumulate charge carriers in the active region is often described as “**charge confinement**”. In this context, let us remark that another factor also contributes to the low laser threshold of the double heterostructure, the “**optical confinement**”. The refractive index of the active layer (e.g. GaAs in Fig. 12.8b) is higher than the one of the enveloping layers with a larger bandgap, $n(\text{GaAs}) = 3.4$ ($\lambda = 873\text{ nm}$) as compared to $\text{Al}_{0.4}\text{Ga}_{0.6}\text{As}$ with $n = 3.3$ at this wavelength. So a light-guiding effect by total internal reflection occurs confining the light in the active layer. The laser light mainly propagates in the active layer where amplification by stimulated emission occurs, not in the surrounding layers where absorption dominates.

Physics Comment: Light Confinement

The “optical confinement” is based on the fact (as just mentioned), that the high bandgap material that “sandwiches” the active layer in double heterojunction lasers has a lower index of refraction than the active layer material. Is this just

Fig. 12.8 (continued)

a lucky coincidence for some semiconductor material combinations, or is it a general necessity based on fundamental laws of optics? Does a higher bandgap of a III-V-semiconductor necessarily mean a lower index of refraction?

This is the case! The argument runs as follows (for clarity of expression, it is given in terms of the example GaAs/AlGaAs; see Fig. 12.8d. Let us first remember two things:

- the absorption edge in a semiconductor equals the bandgap. So the absorption edge of GaAs is $E_G(\text{GaAs})$, and that of AlGaAs is $E_G(\text{AlGaAs})$.
- The index of refraction of a medium increases toward the absorption edge (normal dispersion).

Because the photon energy of the laser is close to the bandgap energy of the active laser medium GaAs, evidently the photon energy of the laser light is further away from the absorption edge of AlGaAs than from the absorption edge of GaAs. So the index of refraction of the laser light in the active medium GaAs tends to be larger than that of the laser light in AlGaAs. This, by the

way, is even more the case because the index of refraction at the band edge of AlGaAs is already smaller than that of GaAs at its band edge.

Let us note that the double heterojunction structure, as efficient as it is, is quite a challenge from the point of view of materials science. **What is needed for a double heterojunction laser is not just one material providing the desired wavelength, but a complete “system” of materials: The “active” semiconductor with the right bandgap, producing the desired wavelength, then a material with a larger bandgap to form the double heterostructure, and last but not least, the right substrate as a template for the growth of these epitaxial layers.** Fortunately, the compound semiconductors with the extensive miscibility of the III and V group elements offer an enormous range of bandgaps and wavelengths, covering the spectral region from the UV far into the infrared. An example of how to select the right materials system in a given case will be presented in the context of optical data transmission in Sect. 12.5.

Let us remark that more sophisticated laser structures than the double heterojunction laser are being used now, like the multi-quantum well laser or the superlattice multi-quantum well laser. We refer to special literature on semiconductor lasers.

12.4 Lasers for Optical Data Storage and Optical Data Transmission

Mass markets for semiconductor lasers are optical data storage and optical data transmission.

Data storage: The first system on the market was the CD player. Here the information, (stored as a scheme of spots on the rotating disc which either reflect the laser light onto a photodiode, or not, corresponding to logical one and zero, respectively), is written and read by a semiconductor laser in the near-infrared (780 nm). The “minimal resolvable distance” d of a light beam with wavelength λ according to Ernst Abbé is given by $d = \lambda/2NA$, where NA is the “numerical aperture”. This means that in order to increase the storage capacity of a disk the wavelength of the laser has to be reduced. So in the next generation (DVD systems) red laser light at $\lambda = 650$ nm was used. The present state of the art is “Blue ray” with $\lambda = 405$ nm. The laser being used is a GaInN laser diode.

Optical data transmission: An application of the highest technical importance in our days is optical data transmission, providing the huge transmission capacities of the Internet. More than 99% of intercontinental data transmission goes via optical fibers (quartz glass). At present (in the year 2022), about 950,000 km of deep-sea cables are in operation, the longest of them SEAME-3 with 40,000 km length. The transmission, of course, should be performed at a wavelength of very low attenuation of the fiber. Figure 12.9 shows the attenuation of a quartz fiber as a function of

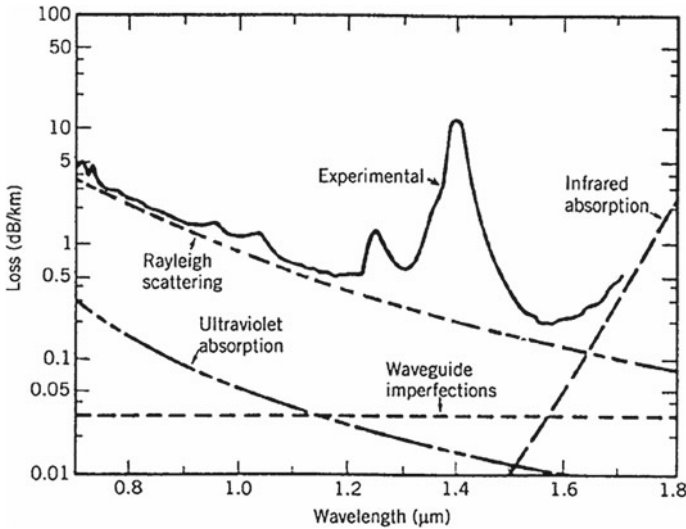


Fig. 12.9 Attenuation as a function of wavelength in a quartz fiber

wavelength. It is given in units of decibel/km, the decibel being defined by

$$\text{Attenuation in dB} = 10 \log_{10}(I_{\text{in}} / I_{\text{out}}) \quad (12.8)$$

I_{in} is the ingoing intensity, I_{out} the outgoing one. An attenuation of, e.g. 20 dB means that the light is damped by a factor of 100. It should be mentioned at this point that the low attenuation of an actual optical fiber for data transmission is the result of enormous development in glass technology achieved essentially by Corning Glass in the 1970s, mainly by reducing the OH^- contamination (i.e. the water contamination) of the glass. While the loss was about 1000 dB/km around 1965, it could be reduced to 20 dB/km around 1970, 1 dB/km in 1975, reaching finally a value slightly below 0.2 dB/km. 0.2 dB/km amounts to 20 dB at a cable length of 100 km, so in a modern quartz fiber the light intensity is attenuated by about a factor of 100 after 100 km fiber length. As a matter of fact, after 100 km, the signal has to be reamplified by the so-called repeater to its original strength.

The decrease of the fiber losses toward the infrared is a consequence of Rayleigh scattering caused by inhomogeneities of the glass. As known from elementary optics, Rayleigh scattering goes proportional to $1/\lambda^4$, which is indeed observed below $\lambda = 1.6 \mu\text{m}$ (see Fig. 12.9). For $\lambda > 1.6 \mu\text{m}$, an increase is observed due to absorption via the Si–O-bond vibration. In between, there is an absolute minimum of about $1.54 \mu\text{m}$. The peak around $1.4 \mu\text{m}$ is the remaining OH^- absorption which used to be responsible for the high attenuation of the material in the “old days”. This peak gives rise to a second (relative) minimum at about $1.3 \mu\text{m}$.

Physics Comment: Can the Attenuation of Quartz Fibers Be Further Improved?

As small as the minimum attenuation around $1.54\text{ }\mu\text{m}$ might be, it is still inconvenient, resulting in the requirement of a repeater every 100 km as mentioned above. This raises the question of whether the attenuation of these fibers in the minimum can be significantly further improved beyond $0.1\text{--}0.2\text{ dB/km}$ seen in Fig. 12.9. The answer is “no” for the following basic reason: As mentioned above, the remaining attenuation is due to Rayleigh scattering, i.e. to scattering by materials inhomogeneities on a scale considerably smaller than the wavelength of the light. These inhomogeneities cannot be deliberately small in an amorphous material like quartz glass because the amorphous state is characterized by fluctuations in bonding lengths and bonding angles resulting necessarily in local fluctuations of the index of refraction (see Sect. 8.1). Those give rise to the Rayleigh scattering. **So—by principle—Rayleigh scattering can never be completely suppressed in amorphous materials.** This limit is approximately reached in today’s quartz fibers.

12.5 Laser Materials for Optical Data Transmission, an Example of Bandgap Engineering

As discussed at the end of Sect. 12.3, for a semiconductor laser of the desired wavelength, not just one material, but a “material system” is required, consisting of the active layer material itself, the material with a higher bandgap for the double heterojunction, and both of them lattice-matched to a substrate.

The optimum wavelength for optical data transmission is around $1.54\text{ }\mu\text{m}$ (Fig. 12.9), corresponding to a bandgap of 0.805 eV . To “engineer” the proper materials system, we use Fig. 12.10.

Looking at the line which connects GaAs and InAs, it is seen that at a certain mixing ratio x in $\text{Ga}_x\text{In}_{1-x}\text{As}$, the lattice constant is identical to the one of InP, which is the case for $x = 0.47$, as marked in the figure. So any mixture of this “ternary compound” $\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ with InP can be grown on an InP substrate with a perfect lattice match. $\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ has a bandgap of about 0.7 eV , as the figure shows, InP one of 1.32 eV . So a mix of the two materials $(\text{Ga}_{0.47}\text{In}_{0.53}\text{As})_y(\text{InP})_{1-y}$ will result in a bandgap between these two values. The larger the InP content, the larger the bandgap. It turns out, as roughly seen from Fig. 12.10, that the desired bandgap of $E_G = 0.805\text{ eV}$ is achieved for $y = 0.9$.

So a “quaternary” compound $(\text{Ga}_{0.47}\text{In}_{0.53}\text{As})_{0.9}(\text{InP})_{0.1}$ with

a Ga content of $0.47y = 0.47 \cdot 0.9 = 0.42$,

an In content of $0.53 \cdot 0.9 + 0.1 = 0.48 + 0.1 = 0.58$,

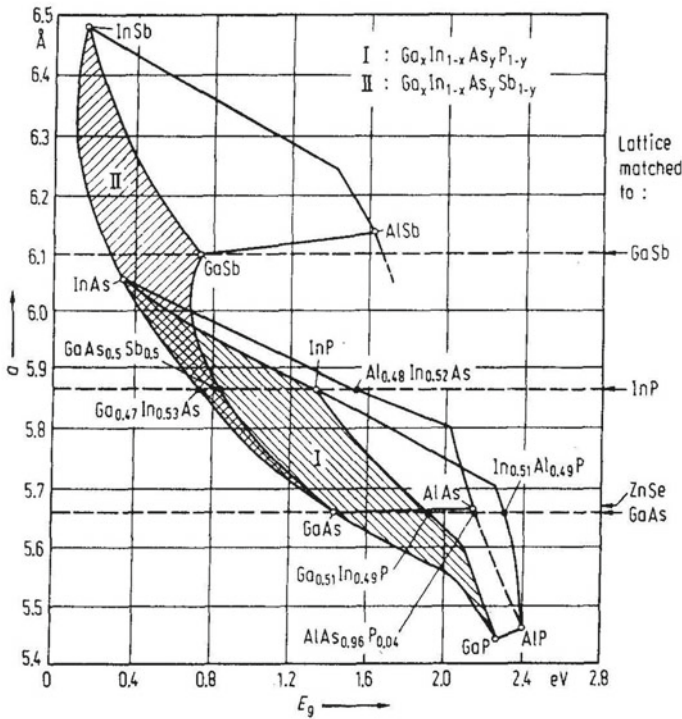


Fig. 12.10 Lattice constant versus band gap for mixed compound semiconductors (taken from Data in Science and Technology, Semiductors, Group IV Elements and III-V-Compounds. Springer-Verlag Berlin Heidelberg 1991)

an As content of $y = 0.9$,
and a P content of 0.1

in short: the compound $\text{Ga}_{0.42}\text{In}_{0.58}\text{As}_{0.9}\text{P}_{0.1}$ will have the desired bandgap 0.805 eV (corresponding to $\lambda = 1.54 \mu\text{m}$) being perfectly lattice-matched to the InP substrate. For the layers with larger bandgap, as required for the double heterojunction, a material $(\text{Ga}_{0.47}\text{In}_{0.53}\text{As})_y(\text{InP})_{1-y}$ with a y smaller than 0.9 can be used, resulting in a larger bandgap than 0.8 eV and—again—perfect lattice match to InP. Actually, optical data transmission is one of the most important applications of InP substrates!

Along these lines, “bandgap engineering” can be carried through for all kinds of applications of compound semiconductors.

A final remark in this chapter is dedicated to the “repeater” which is required every 100 km in the optical fibers for intercontinental data transmission. In the “old days”, this was realized by a photodetector receiving the incoming light, which controlled the next laser providing the light for the next 100 km. Nowadays, this is replaced by a system that amplifies the incoming signal via stimulated emission. The system is schematically shown in Figs. 12.11 and 12.12

Fig. 12.11 Energy levels and optical transitions of Er^{3+}

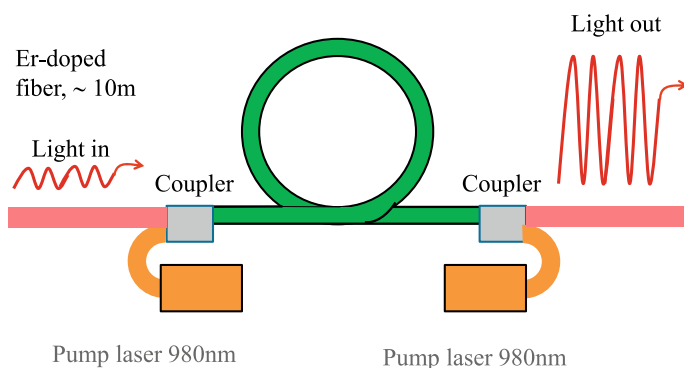
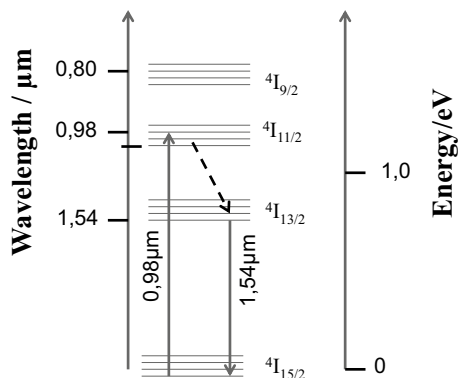


Fig. 12.12 Erbium amplifier as a repeater

Figure 12.11 shows the energy levels of the 4f-states of Er^{3+} in solids (because of the shielding of the 4f electrons by the 5s, 5p, and 6s shell, the energy scheme is only weakly dependent on the specific solid).

The essential point is that Er^{3+} has an optical transition from $4I_{3/2}$ to $4I_{15/2}$ which coincides exactly with the 1.54 μm wavelength as applied in optical data transmission. As indicated in the figure, the upper level of this transition can be occupied by a light of 980 nm with subsequent relaxation from $4I_{11/2}$ to $4I_{13/2}$. In this way, the laser condition (12.3) can be realized, with more ions being in the $4I_{13/2}$ excited state than in the $4I_{15/2}$ ground state (the situation of “inversion”). Under this condition, light of $\lambda = 1.54 \mu\text{m}$ will be amplified.

Figure 12.12 shows how this method is put into operation in a repeater.

A fiber of 10–20 m length doped with Er^{3+} ions (plotted in green) is inserted into the transmission fiber. Light of a 980 nm laser is coupled into this piece of fiber. At sufficient intensity of this pumping light, “inversion” occurs and the incoming signal at 1.54 μm is amplified by a factor of about 100. This process is repeated for about every 100 km.

Let us mention at this point that the “Er-repeater” is one of the relatively few technical applications where the laser principle is used for light amplification, not for providing an oscillator as usual. This is noteworthy given the fact that “amplification” is part of the acronym “Laser”.

12.6 Questions and Answers

- Question 1: Name the three basic interactions between light and matter?
- Question 2: What are the laser conditions for (a) an atomic or molecular laser and (b) a semiconductor laser?
- Question 3: How is the laser condition (12.5) realized in a semiconductor laser?
- Question 4: How is an amplifier transformed into an oscillator?
- Question 5: What is the role of the laser resonator?
- Question 6: What is the reason for the high monochromaticity and directionality of laser light?
- Question 7: Explain the notion of “charge confinement” and “light confinement” in the context of the double heterojunction laser?
- Question 8: Which laser and which wavelength is used for “Blue ray” systems?
- Question 9: What is the optimum wavelength for optical data transmission?
- Question 10: A “repeater” has to be installed every ca. 100 km in a long-distance transmission line. What kind of repeater is used nowadays?
- Answer 1: Absorption, spontaneous emission, and stimulated emission.
- Answer 2: (a) $N_2 > N_1$, (b) $f(E_2) > f(E_1)$.
- Answer 3: By injecting a current in the forward direction into the pn-junction with a strength above the “threshold current”. The semiconductor has to be highly doped (“degenerate semiconductor”).
- Answer 4: Via feedback. Part of the amplified signal has to be coupled back into the input.
- Answer 5: The laser resonator, realized by e.g. the end-faces of the semiconductor crystal and/or a distributed feedback system, provides the feedback system required for laser action.
- Answer 6: The reason is that the photons, being Bose particles, tend to accumulate in just one mode, the “strongest” one, of the resonator.
- Answer 7: Because of the barriers that form at the interface between the active layer and the surrounding layers of the larger bandgap, electrons and holes accumulate in the active layer, they are “confined” in the active layer. So the recombination probability is increased. The light is confined in the active layer because the surrounding material with a larger bandgap has a lower index of refraction.
- Answer 8: GaInN laser with $\lambda = 405$ nm.
- Answer 9: $1.54 \mu\text{m}$, because at this wavelength the attenuation of quartz fibers has its minimum.

Answer 10: The repeater is based on the amplification of the light by a piece of Er-doped glass fiber, making use of an Er^{3+} -transition precisely at 1.54 μm .

References

1. O. Svelto, *Principles of Lasers*, 4th edn, Trans. by D. Hanna (Springer, 1998). ISBN 0-306-45748-2
2. A. Winnacker, *Physik von Maser und Laser* (Bibliographisches Institut – Mannheim, Wien, Zuerich, 1984). ISBN 3-411-01694-9
3. H. Haken, *Synergetics—An Introduction*, 3rd revised and enlarged edn. (Springer, Berlin, Heidelberg, Tokyo, 1983)

Chapter 13

Transistors



Abstract This chapter is dedicated to the transistor in general. But concepts and features will mostly be discussed for the case and in terms of the Si-transistor, which is by far the most important one in modern electronics. The understanding achieved along this line may quite generally be extended to other semiconductor materials, however, occasionally with characteristic modifications.

The transistor is the most important device for the control of currents and voltages in electronic circuits. There are basically two types of transistors: the bipolar transistor and the field effect transistor, based on completely different working principles. The bipolar transistor, either in its pnp- or its npn-version, is based on the “cooperation” of two serial pn-junctions; in the field effect transistor the current between a “source”- and “drain”-electrode is controlled by an electric field, produced by a third electrode, the “gate”, creating a conductive channel connecting source and drain. Understanding the bipolar transistor requires an insight into the different currents in the pn-junctions; understanding the field effect transistor basically requires a closer look at the capacitor formed by the gate electrode, the Si-oxide insulator, and the Si-semiconductor. The special feature of this system is the penetration of the electric field into the semiconductor, the “field effect”.

The Si-field-effect transistor, the “MOSFET”, is one of the basic components of integrated circuits. Its historical development followed the trend of miniaturization in this field as well as the requirements of mass production. Problems in that context where, among others, purity of materials, the high quality demands on the SiO₂-insulator and its interface with Si, and the replacement of SiO₂ by new “high k-dielectrics”.

The MOSFET can be used to create, under appropriate conditions, a two-dimensional electron gas in the channel. This possibility gave rise to an important discovery in modern physics, the “quantized Hall effect” (Sect. 13.4).

A transistor is a device for the control of currents and voltages. There are basically two types of transistors, the bipolar transistor and the field effect transistor, treated in the following Sects. 13.1 and 13.2, respectively.

13.1 The Bipolar Transistor

In a way, the bipolar transistor is an application of the pn-junction (Chap. 9) in the sense that it consists of two pn-junctions in series. Figure 13.1a shows the basic design.

The “pnp transistor” shown in the figure consists of three differently doped layers of Silicon, an n-type layer sandwiched between a highly doped and a less doped p-type layer. An “npn” transistor would consist of a p-type layer sandwiched between two n-type layers. In the following, we shall focus on the pnp transistor, as shown in Fig. 13.1; the npn transistor could be treated in an analogue way.

The three layers are called emitter E, basis B, and collector C. The pn-transition between emitter E and basis B is driven in the forward direction, and the one between basis B and collector C in the backward (reverse) direction.

In order to understand the pnp-transistor, we have to remember some of the basic features of the pn-junction (Sects. 9.1 and 9.2) being summarized in the following:

- The current in the forward direction in a pn-junction is essentially the diffusion current and strongly depends on the applied voltage (it corresponds to the exponential term in the $I(U)$ characteristic Eq. 9.12), because it has to overcome the potential barrier between p- and n-sides.
- For sufficiently large reverse voltages, the current approaches the constant field current, independently of the applied external voltage. The field current may be supplied by different processes, like thermal excitation in the case of the “pure” pn-junction (Sect. 9.2), by optical excitation as in the solar cell, or by injection,

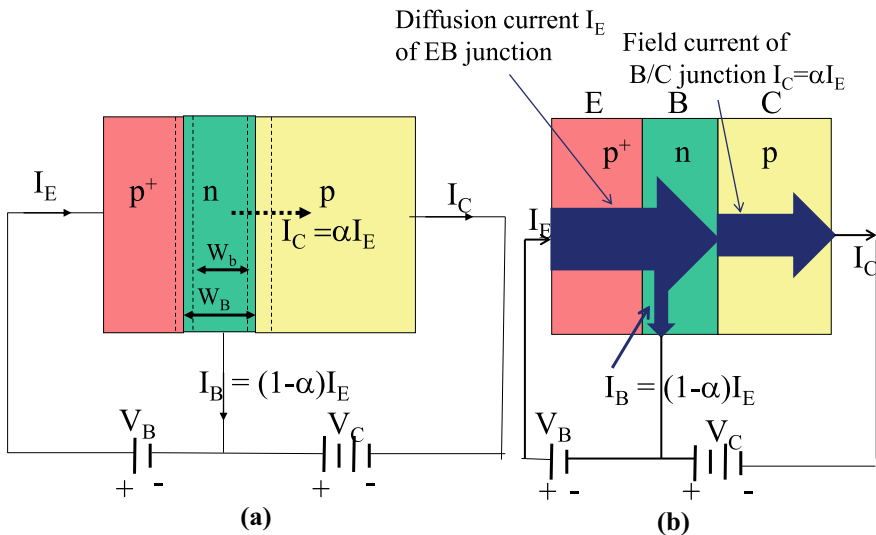


Fig. 13.1 **a** Basic design of the pnp transistor. **b** The currents in the pnp transistor

as discussed in the following. It is independent of the applied voltage because the carriers just have to “slide down” the potential wall (Sect. 9.2).

The idea of the pnp-transistor now is the following; see Fig. 13.1b.

The purpose of the transistor is to provide a controllable current I_C to an external “output” circuit. This current is produced by the pn-junction B/C run at the voltage V_C . Because this diode is run in reverse direction, I_C corresponds essentially to the field current of that junction (at sufficiently high V_C). The field current, as remarked above under (b), is essentially determined by the free carriers present in the depletion zone. So the current I_C can be controlled by injecting free carriers into the depletion zone.

This is done by providing these carriers via the current I_E flowing in the other pn-junction E/B. So, the higher the I_E , the higher the I_C , which we want to control. It is essentially a hole current, because the p-area is very strongly doped.

Because the junction E/B is run in the forward direction, I_E depends strongly on the applied voltage V_B of that junction. So the logic is V_B controls I_E , and I_E controls the output current I_C because it supplies the free charge carriers to the depletion zone of B/C. In this way, it produces the field current I_C in the reverse biased junction B/C.

Evidently, the bipolar transistor is a very clever combination of a forward biased pn-junction with a reversed biased one!

Physics Comment: Different Origins of the Field Current in a Reverse Biased pn-Junction

It might seem kind of strange at first sight that the current I_C , which is intended to be controlled and subsequently be used in an external circuit, is just the field current of a reverse biased pn-junction, because we are accustomed to imagining the current in a reverse biased pn-junction to be small, given the application of a pn diode as a rectifier diode. This, however, is a misunderstanding. The reason why the field current in a rectifier diode is so small is that it is provided by the charge carriers which are produced by thermal excitation across the bandgap. This, indeed, is a very weak source of free carriers, as quantitatively discussed already in the context of Eq. 10.3d at the end of Sect. 10.2. If, on the other hand, the free carriers providing the field current are produced in high concentrations, as in the case of the solar cell, or here, by injection from the forward direction the biased junction E/B, the field current can be correspondingly large!

The current I_E entering the basis is divided, according to Kirchhoff’s law, into the collector current

$$I_C = \alpha I_E \quad (13.1a)$$

and the basis current

$$I_B = (1 - \alpha)I_E \quad (13.1b)$$

The collector current I_C is supplied by the holes which diffuse through the basis into the depletion zone between B and C. So α represents the fraction of the holes reaching the depletion layer and supplies the field current which flows into the reverse biased pn-junction between B and C.

It should be noted for completeness that there is a small flow of electrons going from C to B, produced by the electrons which are thermally excited in the depletion layer between B and C, an undesired “leakage current”.

I_C/I_B is called “amplification factor A” of the transistor. From Eqs. (13.1a) and (13.1b) follows

$$A = \alpha/(1 - \alpha) \quad (13.1c)$$

So the “amplification factor” A is large when α is close to 1.

The term “amplification factor” for the ratio I_C/I_B refers directly to the basic idea of the bipolar transistor. The purpose of the transistor is to supply a controllable current I_C to an external circuit at a given voltage V_C . The control is provided via the current I_B , which in turn is adjusted by V_B , the basis-emitter voltage. The aim is to control a large current I_C via a small current I_B , i.e. to produce a large effect by a small effort. After all, I_C is the current to be used, while I_B is lost. In that sense, the term “amplification factor” for I_C/I_B is properly chosen: It gives the ratio of the current to be used relative to the one being lost for the control of the used current.

From this basic idea it is evident that, in order to get a high amplification factor A (i.e. an α close to 1; see Eq. 13.1c), one has to assure that a fraction as large as possible of the holes which carry I_E arrive in the depletion region of junction B/C. To put it differently, one has to make sure that only few holes are lost after entering the basis B.

In the following, we want to understand the factors which determine α . The holes supplied by I_E reach the collector by diffusion through the basis with thickness W_B ; see Fig. 13.1a. (To be precise, the layer to be traversed by diffusion is not the total basis, but the part W_b outside the field-free depletion zone, as shown in the figure, but we assume $W_B \approx W_b$.) According to Einstein’s general law of diffusion, the time required to overcome a distance W_B by diffusion, called the transit time τ_T , is given by

$$\tau_T = W_B^2/2D_h \quad (13.2)$$

D_h is the diffusion constant of the holes in the basis.

Not all the holes injected into the basis by I_E will reach the collector because on their way they might go lost by recombination with electrons. Let the recombination time of the holes in the basis be τ_h . Then the probability per unit of time of a hole to going lost by recombination is $1/\tau_h$. Hence, the probability to go lost during the transition time τ_T is $w = (1/\tau_h) \cdot \tau_T$, and consequently the probability not going lost is $1 - w = 1 - \tau_T/\tau_h$. Because the current I_C is just the one supplied by the holes not

lost on their way through the basis, one gets

$$I_C = (1 - w)I_E = (1 - \tau_T/\tau_h)I_E \quad (13.3)$$

By comparing (13.3) with (13.1a), one gets

$$\alpha = (1 - \tau_T/\tau_h) \quad (13.4)$$

So α is close to 1 when $\tau_T \ll \tau_h$. This is intuitively clear: When the transit time through the basis is very short compared to the recombination time, most of the holes will reach the collector before they are lost by recombination. So α is determined by the ratio τ_T/τ_h . What in turn does determine τ_T and τ_h ?

The recombination time τ_h is strongly determined by the concentration of lattice defects, doping level, and by impurities giving rise to recombination centers. For a doping concentration of about 10^{16} cm^{-3} in the basis and high-quality Si, a value $\tau_h = 500 \text{ ns}$ is a good estimate.

The transit time τ_T is obtained from Eq. 13.2. Qualitatively, it is obvious from this equation what the relevant parameters are: For the transition time to be short, the basis has to be thin (i.e. W_B small!) and the carriers have to move fast through the basis, corresponding to a large D_h .

D_h can be obtained from the Einstein relation already used in Sect. 9.1:

$$D_h = (kT/e)\mu_h \quad (13.5)$$

Numbers

We take $W_B = 2 \mu\text{m} = 2 \cdot 10^{-4} \text{ cm}$, $\tau_h = 500 \text{ ns}$ (see above), and $\mu_h = 0.0475 \text{ m}^2/\text{Vs} = 475 \text{ cm}^2/\text{Vs}$, from Table 2.2 in Chap. 2.

First, D_h is calculated from (13.5). Because kT at room temperature is 0.026 eV , $kT/e = 0.026 \text{ V}$. This gives

$$D_h = 0.026 \text{ V} \cdot 475 \text{ cm}^2/\text{Vs} = 12.4 \text{ cm}^2/\text{s}$$

This is inserted in (13.2):

$$\tau_T = (2 \cdot 10^{-4} \text{ cm})^2 / (2 \cdot 12.4 \text{ cm}^2/\text{s}) = 1.6 \cdot 10^{-9} \text{ s} = 1.6 \text{ ns}$$

So from (13.4) α is calculated to be

$$\alpha = (1 - 1.6\text{ns}/500\text{ns}) = 0.997$$

This means (Eq. 13.1a) that 99.7% of the holes entering the basis via I_E reach the collector forming I_C . This gives an amplification factor (Eq. 13.1c)

$$A = 0.997/(1 - 0.997) = 332$$

The fact that so many (99,7%) of the holes reach the collector is evidently due to two factors: The thickness of the basis is small ($W_B = 2 \mu\text{m}$), and the recombination time τ_h is so much larger than the transit time τ_T .

The Essential Role of Material Quality and Purity

As mentioned already above, to achieve a small τ_h is a matter of material perfection and purity. This is the reason why high-quality Si is required for the bipolar transistor.

Importance of a Small Width W_B of the Basis

The width W_B enters crucially into the transit time τ_T (Eq. 13.2) and thus into the amplification factor. But it is also important for another relevant performance parameter of the transistor: The frequency limit! The current I_C is determined by the holes diffusing through the basis. So in order to change I_C , this diffusion current has to be changed. From this basic idea, it follows that the time required to switch the transistor is essentially given by the transit time. For the frequency limit, this means that $\nu_{\max} = 1/\tau_T$. For our numerical example, this implies

$$\nu_{\max} = 1/(1,6 * 10^{-9}\text{s}) = 10^9/1,6 \text{ Hz} = 625 \text{ MHz}$$

Considerably, larger frequency limits can consequently be obtained by a smaller width of the basis.

13.2 Field Effect Transistor

The field effect transistor is the transistor type most widely used in modern integrated circuits. Figure 13.2a shows the basic scheme.

The current to be controlled flows from the source electrode to the drain electrode; it is controlled via the gate electrode. The transistor shown in Fig. 13.2a is called Metal Oxide Semiconductor Field Effect Transistor, in short MOSFET. The part MOS of this technical term refers to the materials of the stack gate-insulation-substrate. The “field effect” giving the device its name will be discussed in some detail below.

Obviously, the current from source to drain has to pass two pn-junctions (between the p-substrate and the two highly doped n^+ regions under the source- and drain contacts, respectively). No matter what the polarity of the voltage V_{DS} between source and drain is, one of these pn-junctions will be poled in the reverse direction, so the source-drain current will be almost zero. This is the situation when the gate voltage is “below threshold”, a phrasing that will get clear in a moment. In Fig. 13.3, the curve $I_{DS}(V_{DS})$ marked by (a) corresponds to this situation: A very small current, only very slightly increasing with the source-drain Voltage V_{DS} , as is typical for a reverse current in a pn-junction.

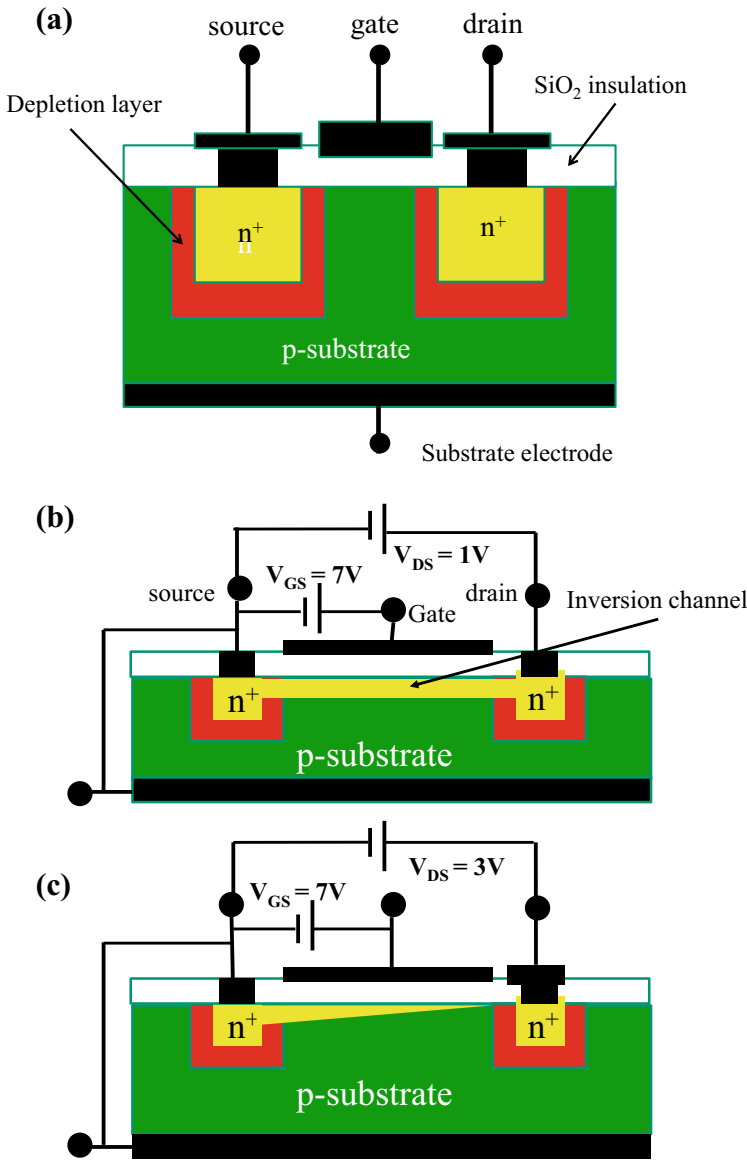
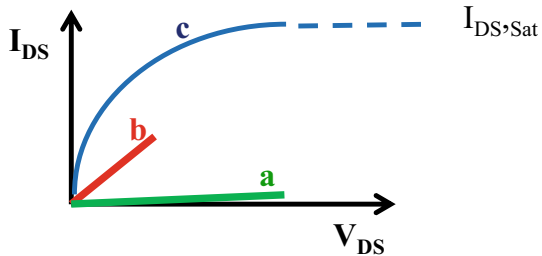


Fig. 13.2 **a** Basic scheme of the MOS-field effect transistor. **b** Formation of an inversion channel. **c** "Pinch-off" of inversion channel

Fig. 13.3

Drain source current I_{DS} as a function of drain-source voltage V_{DS} for different gate voltages



We next consider the case when a positive voltage is applied to the gate, as shown by curve b in Fig. 13.3.

The positive potential on the gate electrode will push away the (positive) holes from the near-surface region of the semiconductor into the bulk, at the same time (negative) electrons will be pulled from the bulk into the region near the surface. If the voltage applied to the gate is large enough, i.e. beyond a certain threshold voltage V_{th} , an n-type channel will thus be formed right below the surface; so a current will flow between the source and drain through this channel. The source and drain are connected by a conducting medium. The situation is just like the one where a wire connects two electrodes, so Ohm's law applies, and the current I_{DS} is proportional to the voltage V_{DS} . The channel formed in this way is called an "inversion channel", because it is an n-channel in a p-type semiconductor.

The proportionality between I_{DS} and V_{DS} only holds for sufficiently small V_{DS} .

The situation developing for larger V_{DS} is explained in Fig. 13.2c).

Let us assume, for example, that the "threshold voltage" V_{th} is 4V (note that V_{GS} in Fig. 13.2b is 7V, so $V_{GS} > V_{th}$, and an inversion channel is formed as correctly shown in the figure). Let the voltage V_{DS} between source and drain in Fig. 13.2c be 3V. These 3V drop along the channel. So at the end of the channel, the voltage difference between substrate and gate is only $7V - 3V = 4V$, so it just corresponds to the threshold voltage and consequently the thickness of the channel approaches 0. The inversion channel "pinches off" at its end as shown in the figure. This narrowing of the channel is the reason why the current I_{DS} no longer grows linearly with V_{DS} (Fig. 13.3c) and reaches a "saturation current" $I_{DS,Sat}$.

A closer analysis shows that beyond the "linear regime" (Fig. 13.3b), the drain-source current I_{DS} can be approximated by

$$I_{DS} = (W/L)\mu C_{ox}[(V_{GS} - V_{th})V_{DS} - V_{DS}^2/2] \quad (13.6a)$$

L is the length of the channel, W its width (i.e. the dimension perpendicular to the drawing's plane). C_{ox} is the capacitance of the system gate electrode-insulator-substrate. The first term in this sum corresponds to Ohm's law ($I_{DS} \sim V_{DS}$); the second one is due to the fact that the channel depth d gets narrower toward the end of the channel because the difference between the potential of the substrate and the threshold voltage gets smaller, the effect that finally ends in the pinch-off.

As explained above, the pinch-off of the channel occurs when $V_{GS} - V_{DS} = V_{th}$
or

$$V_{DS} = V_{GS} - V_{th} \quad (13.6b)$$

If we call the current at the pinch-off point the “saturation current” and insert (13.6b) into (13.6a), we obtain for the saturation current

$$I_{DS,Sat} = (W/2L)\mu C_{ox}(V_{GS} - V_{th})^2 \quad (13.6c)$$

For larger V_{DS} (>3 V in the example), this pinch-off happens already at a point before the gate’s end. The current now has to flow through the substrate.

In summary, it can be stated that the current in a field effect transistor can be controlled by the creation and shaping of an inversion channel between source and drain via the gate voltage V_{GS} .

The field effect transistor FET was developed later than the bipolar transistor. This was due to some difficulties specific to the FET.

The insulating layer under the gate: Problems and solutions

The insulating layer under the gate commonly consists of SiO_2 . In a listing of properties which make Si a favorite in semiconductor technology, the oxide is commonly mentioned in a leading position. As a matter of fact, quite a number of commonly used semiconductors do not develop a closed oxide layer, like GaAs or GaN. Others, however, do, like SiC. So the SiC MOSFET is an important component now in high-power electronics. The problem in the development of the Si-MOSFET was that an oxide of outstanding quality and purity is required which first had to be developed in this context.

As far as purity is concerned, a crucial point is the avoidance of Na^+ - and K^+ -contaminations of the Oxide. Na^+ -ions or K^+ -ions in the SiO_2 layer will push holes away from the insulator-semiconductor interface and thus may produce an inversion channel even without the application of a positive potential to the gate. In any case, the threshold voltage would be poorly defined.

Another difficulty which considerably slowed down the development of the MOSFET is the “quality” of the interface between the oxide and the semiconductor. The structures of SiO_2 and Si are totally different, so dangling bonds will develop in the interface which could act as charge traps. These charges will also affect the formation of the channel and thus result in missing or poor definition of a threshold voltage. In Chap. 14, “integrated circuits”, it will be mentioned that two oxidation processes are commonly in use, wet oxidation and dry oxidation. The latter results in much slower growth of the SiO_2 providing a sufficiently “passivated” interface between Si and SiO_2 . This, by the way, is just an example of a general rule in semiconductor technology: Slow growth generally results in higher perfection.

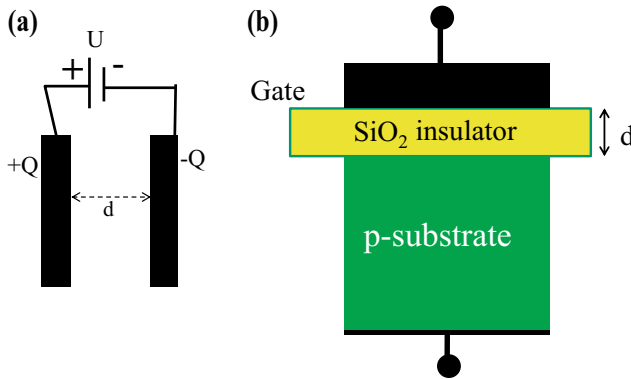


Fig. 13.4 **a** Capacitor with two metallic electrodes. **b** Capacitor formed by gate, insulator, and substrate of a MOSFET

Physics Comment. The Field Effect And Its Dependence On Carrier Concentration

The electric field produced by supplying a potential to the gate penetrates the semiconductor, attracts electrons, and pushes away holes, resulting in the formation of an inversion channel. This raises an interesting general question: Under which conditions and to what extent does an electric field penetrate a solid? The standard answer to be found in the textbooks is: The electric field does not penetrate a conductor, however, it does enter an insulator. This answer, obviously, is not of a quantitative nature. Where is the borderline that separates “conductors” and “insulators” given the fact that the conductivities of materials differ by about 25 orders of magnitude (Chap. 2, Table 2.1)?

There are two aspects to be considered:

Whether an electric field enters a material is first a matter of time. Let us consider a piece of a very good conductor, e.g. a metal like Cu. When this is suddenly exposed to an electrical field $\mathbf{E}$, the free electrons will feel a force $\mathbf{F} = e\mathbf{E}$ and will move around freely until the distribution of the charges in the metal is such that the internal field is zero. Only then the free carriers will no longer feel a force and a stable, stationary distribution of charge is achieved. This field-free state is approached exponentially with a time constant called the “dielectric relaxation time” τ_R .

From Maxwell’s equations, it is deduced in a straightforward way that τ_R is given by.

$$\tau_R = \varepsilon \varepsilon_0 \rho \quad (13.7)$$

where ε_0 is the dielectric permittivity of the vacuum, ε the one of the material considered, and ρ the resistivity.

Let us consider this first for the case of highly doped Si with an electron concentration of about 10^{17} cm^{-3} with a conductivity of $\sigma = 10^3 \text{ S/m}$ (Table 2.1), i.e. $\rho = 10^{-3} \Omega\text{m}$. (The value given for Si in Table 2.1 refers to a low doping concentration of 10^{14} cm^{-3} .)

One gets

$$\tau_R = 11.7 * 8.85 * 10^{-12} \text{ C/Vm} * 10^{-3} \Omega\text{m} = 0.1 * 10^{-12} \text{ s} \quad (13.8a)$$

an extremely short, almost immeasurable time.

For a standard Si material with, e.g. a dopant concentration of 10^{14} cm^{-3} and $\sigma = 1 \text{ S/m}$, $\rho = 1 \Omega\text{m}$, one obtains $\tau_R = 0.1 * 10^{-9} \text{ s} = 0.1 \text{ ns}$.

Next, let us for comparison consider a very good insulator like quartz with. $\sigma \approx 10^{-18} \text{ S/m}$ (Table 2.1), i.e. $\rho = 10^{18} \Omega\text{m}$, $\epsilon = 4.5$:

$$\tau_R = 4.5 * 8.85 * 10^{-12} * 10^{18} \text{ s} \approx 4 * 10^7 \text{ s} \quad (13.8b)$$

which is more than a year.

At this point, one might be tempted to conclude that the reason for the electric field not penetrating an insulator is time. It just takes too long, and perturbing factors would make the process essentially unobservable. This however is a misconception which leads us to the second aspect to be considered when we discuss the problem of the field effect: The reason for penetrating the electric field into a solid might be that there are just not enough charge carriers available for compensating the external field by local rearrangement. Actually, this may lead us back to the example of the MOSFET.

Figure 13.4a shows a simple capacitor with two metallic electrodes.

Figure 13.4b shows for comparison a section of a MOSFET consisting of the gate, the gate oxide, and the p-substrate, which can be considered as a capacitor with a metallic gate electrode, a dielectric (SiO_2), and the counter electrode, the p-Si substrate.

In the following, it will be shown that an electric field produced by, e.g. a few volts at the capacitor will barely penetrate the metallic counter electrode (Fig. 13.4a), but will have a considerable penetration depth into the counter electrode made of p-Si (Fig. 13.4b). First, the case of Fig. 13.4a:

The capacitance C of a capacitor is given by

$$C = \epsilon \epsilon_0 A/d \quad (13.9a)$$

ϵ = dielectric permittivity of the dielectric (no dielectric in Fig. 13.4a, so $\epsilon = 1$)

$$\epsilon_0 = 8.85 * 10^{-12} \text{ C/Vm} = 8.85 * 10^{-14} \text{ C/Vcm}$$

A = area of electrodes and d = separation of electrodes.

With, e.g. $A = 1 \text{ cm}^2$ and $d = 1 \text{ }\mu\text{m}$, one gets

$$\begin{aligned} C &= 8.85 \times 10^{-14} \text{ C/(Vcm)} \times 1 \text{ cm}^2 / 10^{-4} \text{ cm} \\ &= 8.85 \times 10^{-10} \text{ C/V} \end{aligned}$$

With, e.g. $U = 2 \text{ V}$, one gets for the charge Q on the electrodes

$$Q = CU = 17.7 \times 10^{-10} \text{ C}$$

Because of $e = 1.6 \times 10^{-19} \text{ C}$, this corresponds to 1.1×10^{10} elementary charges. Let us consider Cu as the material for the electrodes. Because in Cu each atom provides about 1 free electron to the lattice, this implies that ca. 1.1×10^{10} Cu atoms can provide the charge on the plates. How much material is this?

The density of Cu is $\rho(\text{Cu}) = 8.96 \text{ g/cm}^3$, the atomic weight $M(\text{Cu}) = 63.5 \text{ g/Mol}$, and the Avogadro number $N_A = 6 \times 10^{23} \text{ /Mol}$. With these numbers, we get for the number of atoms per cm^3 :

$$\begin{aligned} N(\text{Cu}) &= [\rho(\text{Cu})/M(\text{Cu})]N_A = 8.96 \times 6 \times 10^{23} / 63.5 \\ &= 8.5 \times 10^{22} \text{ cm}^{-3} \end{aligned}$$

From this, one gets for the number of atoms per cm^2 of one atomic layer

$$N(1 \text{ layer Cu}) = (8.5 \times 10^{22})^{2/3} = 1.9 \times 10^{15} \text{ cm}^{-2}$$

Because, as stated above, the charge on the electrode corresponds to 1.1×10^{10} elementary charges, this implies that a fraction of 10^{-5} of one atomic layer can provide the charge on the capacitor's electrode, and the electric field does not have to penetrate the bulk of the metallic electrode to collect the necessary charge.

The situation looks completely different if one of the electrodes consists of a semiconductor, e.g. of p-Si with a doping concentration of 10^{15} cm^{-3} (Fig. 13.4b). In the same way as shown above, for Cu we obtain for the number of atoms per volume for Si (with $\rho(\text{Si}) = 2.32 \text{ g/cm}^3$, $M(\text{Si}) = 28 \text{ g/Mol}$)

$$N(\text{Si}) = 5 \times 10^{22} \text{ cm}^{-3}$$

Hence, a doping concentration of 10^{15} cm^{-3} corresponds to a fraction of

$$f = 10^{15} / (5 \times 10^{22}) = 2 \times 10^{-8}$$

So the number of Si atoms of 1 layer per area is $(5 \times 10^{22} \text{ cm}^{-3})^{2/3} = N(\text{one layer Si}) = 13.5 \times 10^{14} \text{ cm}^{-2}$, which contains

$$N(\text{one layer Si}) * f = 2.7 \times 10^7 \text{ acceptor atoms}$$

(In case of monocrystalline Si the number of atoms per layer would slightly depend on the crystallographic orientation, but this is irrelevant for this order of magnitude calculation). Hence, in order to compensate the charge of 1.1×10^{10} elementary charges on the metallic electrode, $1.1 \times 10^{10} / 2.7 \times 10^7 \approx 0.41 \times 10^3 = 410$ layers of the semiconductor are required to provide the charge on the semiconductor electrode of the capacitor. Because the separation of atomic layers in Si is about 0.3 nm, this means that the field has to penetrate $410 \times 0.3 \text{ nm} \approx 120 \text{ nm}$ into the Silicon.

At a dopant concentration of 10^{14} cm^{-3} , the field would have to penetrate $1.2 \text{ }\mu\text{m}$ to collect the necessary charges.

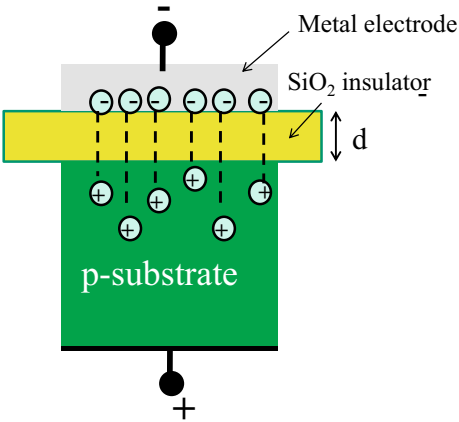
In this way, it is seen that the penetration depth in a material is determined by the concentration of charges in the material. **The field has to enter the material up to the depth required to collect the necessary charges for the compensation of the field.**

A final remark concerns the capacitance of a capacitor consisting of a metal, the dielectric, and a semiconductor electrode: Because the field lines connecting the charges in both electrodes may be much longer than the thickness d of the dielectric, the “effective separation” of the two electrodes would be considerably larger than d (Fig. 13.5); the capacitance then would be considerably reduced relative to the one given by Eq. 13.9a of a capacitor with two metallic electrodes. Because, as follows from the previous arguments, the penetration depth of the field would become smaller with lower voltage U , the capacitance would become voltage-dependent and would get larger with decreasing U . This is, to give an example of applications of this effect, the principle of the determination of carrier concentrations and distributions via capacitance measurements.

13.3 A MOSFET Problem Related to Miniaturization of Modern Electronics: High k -dielectrics

Gate electrode, insulator, and substrate as part of the MOSFET form a capacitor. In the course of continuous miniaturization of integrated circuits (Chap. 14), the area available for this capacitor gets smaller and smaller. On the other hand, there is a demand to keep the capacity of this capacitor as high as possible. The main reason is seen from Eqs. 13.6a and 13.6c: the current in the linear regime as well as the

Fig. 13.5 E-field lines connecting negative and positive charges in a capacitor with a metal and a semiconducting electrode and SiO₂ as dielectric



saturation current is proportional to C_{ox} . So in order to keep the currents sufficiently high (for reasons of noise and reproducibility!), a sufficiently high C_{ox} is required.

The capacitance of the system gate-insulator-substrate C_{ox} is given by Eq. (13.9a). In order to keep the capacitance as high as possible while continuously decreasing the area A , the thickness d of the SiO₂ insulator had to be decreased more and more, down to about one nanometer at present, corresponding to a few atomic layers. At this level, the problem arises that, due to quantum mechanical tunneling of charges through the layer, a leakage current develops: the insulator no longer insulates! Several steps have been taken to counteract this problem.

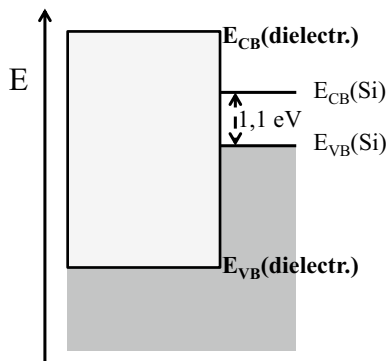
One is to add nitrogen to the oxide, replacing the SiO₂ with SiON. This reduces the leakage current at a given thickness by about a factor of 2.

A more general approach is to replace SiO₂ with a gate oxide with significantly higher permittivity. The dielectric permittivity of SiO₂ is relatively small, only $\epsilon = 3.9$. There are materials with much higher ϵ , as given in Table 13.1.

Table 13.1 Permittivity of potential insulator materials (data from H. Huff, D. Gilmer: High dielectric constant materials: VLSI MOSFETApplications, Springer, Berlin 2004)

Material	Permittivity ϵ
SiO ₂	3,9
Si ₃ N ₄	7
Al ₂ O ₃	9
Y ₂ O ₃	15
ZrO ₂	25
HfO ₂	25
La ₂ O ₃	30
Ta ₂ O ₅	26
TiO ₂	80

Fig. 13.6 “Band-offset”
between Si and gate oxide



By choosing a dielectric with considerably higher ϵ than the one of SiO_2 , the same capacity C_{ox} can be achieved at a considerably higher thickness d of the dielectric. So leakage currents due to tunneling would be essentially eliminated.

A new dielectric, however, will in general raise new problems in terms of dangling bonds between dielectric and semiconductor; see Sect. 13.2 “The insulating layer under the gate”. One way to deal with this problem is to introduce a buffer of just one layer of SiO_2 between the Si and the new dielectric which may considerably reduce the problem by bringing the system closer to the established Si- SiO_2 combination.

Technicians in charge of semiconductor process technology are always very reluctant to introduce a new material into the process. Given the fact that already extremely small impurity levels may result in changes of electronic properties, the new dielectric might result in severe “side effects” when being used together with Si. The following questions might come up:

Which elements are “electrically active” in one sense or the other, e.g. in affecting carrier lifetime, mobility, or conductivity, and which ones are not? Moreover, technological issues are involved like process temperature, vapor pressure, etc.

Another issue in choosing a gate dielectric is “band-offset”, as explained in Fig. 13.6.

In order to keep leakage currents low, there must be an energy barrier for electrons for crossing from the conduction band of Si into the conduction band of the gate oxide, and for the crossing of holes from the valence band of Si into the valence band of the gate oxide. This requirement is fulfilled in the schematic drawing, Fig. 13.6. The band-offsets must have sufficient height and the right sign.

So in the end, the choice of suitable materials turns out to be very limited. The new “high k-dielectric” used in present Si-technology is HfO_2 (Note: “k” is often used instead of ϵ for the dielectric constant in American literature.). The ratio $\epsilon(\text{HfO}_2)/\epsilon(\text{SiO}_2) = 25/3.9$ is 6.4; so the HfO_2 -gate dielectric can be 6.4 times thicker than the SiO_2 dielectric, providing the same capacity, other things being equal. Leakage currents then no longer play a significant role.

The MOSFET is a crucial component of modern electronics, in particular in integrated circuits and data storage. The state of the art, as far as miniaturization is concerned, will be treated in Chap. 14.

13.4 The Two-Dimensional Electron Gas, Realized in a MOSFET

Let us first consider the problem of a free electron gas in a three-dimensional potential well. The Schrodinger equation with infinitely high potential walls and dimensions a , b , c in x -, y - and z -directions results in the energy level scheme

$$E(n_x, n_y, n_z) = (\hbar^2/8m_{\text{eff}}) \left[n_x^2/a^2 + n_y^2/b^2 + n_z^2/d^2 \right] \quad (13.10a)$$

m_{eff} is the effective mass of the electron. n_x , n_y , and n_z are integers starting from zero. Let d be much smaller than a and b : $d \ll a$, $d \ll b$; let the energies E of the electrons be sufficiently small such that

$$E \ll (\hbar^2/8m_{\text{eff}})/d^2 \quad (13.10b)$$

then all the electrons are in the state with $n_z = 0$, so we can write for the energy states

$$E_{n_x, n_z} = (\hbar^2/8m_{\text{eff}}) \left[n_x^2/a^2 + n_y^2/b^2 \right] \quad (13.11)$$

The system can be described as “two-dimensional”.

Physics Comment: Two-Dimensional Systems and the de Broglie Relation

The fact that the system can be described as “two-dimensional” can be expressed in the following intuitive way:

If one dimension of the potential well is small compared to the de Broglie wavelength of the electrons, the system can be considered as “two-dimensional”.

This can be shown by writing the condition Eq. (13.10b) in terms of the de Broglie wavelength λ_B , which is defined by the de Broglie relation

$$p = h/\lambda_B \quad (13.12)$$

where p is the momentum of the particle. Because of (13.12), the condition (13.10b) can be written as

$$E = (1/2m_{\text{eff}})p^2 = (1/2m_{\text{eff}})h^2/\lambda_B^2 << (1/8m_{\text{eff}})h^2/d^2$$

which amounts to $1/\lambda_B^2 << 1/4 d^2$

$$\text{which is the case if } d << \lambda_B \quad (13.13)$$

This result is quite intuitive in the sense that the wave character of the particle somehow defines a spatial extension of the particle given by its de Broglie wavelength. So, if Eq. 13.13 holds, the electron does not “fit” into the sample forming the potential well.

This idea can, of course, be considered in terms of the uncertainty relation. Applied to the z-direction it says $\Delta z \Delta p \geq h$. So, if the electrons are confined to a volume with a very small extension in the z-direction (called d here), Δp must be sufficiently large to fulfill the uncertainty relation:

$$\Delta z \Delta p = d \Delta p \geq h$$

Because certainly p has to be larger than Δp , this amounts to $d \cdot p \geq h$ or $h/p = \lambda_B \leq d$.

This is obviously violated if (13.13) holds, so in this case the movement of the electrons can't have a component in the z-direction; it is “two-dimensional”.

In numbers: We consider a gas of free electrons at a temperature T. Then the energy of the electrons is of the order of kT, and the momentum $p = (2 mE)^{1/2}$ of the order $(2m_e kT)^{1/2}$. The de Broglie wavelength is given by $\lambda_B = h/p = h/(2m_{\text{eff}} kT)^{1/2}$.

By introducing the numbers for h, m_{eff} (for m_{eff} we take the free electron mass for our estimate), and k, we obtain

$$\lambda_B = 1,32 * 10^{-7} * (1/T)^{1/2} \text{ with } \lambda_B \text{ in meters and } T \text{ in K}$$

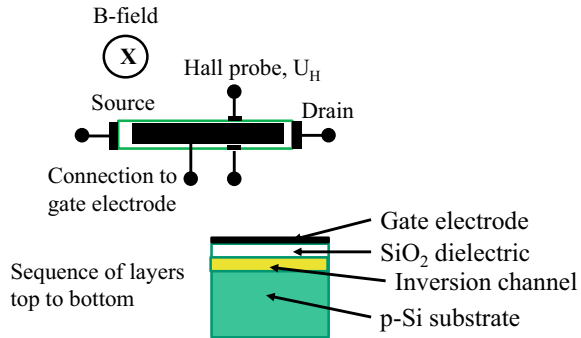
At $T = 1 \text{ K}$ one gets $\lambda_B = 132 \text{ nm}$.

So at a temperature of 1 K, a layer thickness d of considerably less than 100 nm will result in a two-dimensional electron gas.

Realizations of two-dimensional electron gases

A straightforward way to produce a two-dimensional electron gas is by confining it into a sufficiently thin layer of the semiconductor material, a “quantum well”. This can, e.g. be done by providing a double heterostructure as described in Sect. 12.3. A similar way would be the formation of a semiconductor superlattice. A superlattice is a system of layers of two alternating materials, like GaAlAs/GaAs. If the one with the smaller bandgap is sufficiently thin, the electrons would be confined in these layers. Such superlayers are used, e.g. in modern types of semiconductor lasers.

Fig. 13.7 Vertical view on the MOSFET for observation of the quantum Hall effect



Another way to set up a two-dimensional electron gas relates to the MOSFET. The MOSFET, as explained above, is opened by creating an inversion channel via the gate voltage. In the case of a p-substrate, the current is carried by the electrons flowing through the channel. The thickness of the channel is controlled by the gate voltage. If this voltage is larger than the threshold voltage, required to create the channel, but small enough to keep the channel sufficiently thin, the MOSFET acts as a two-dimensional n-type conductor.

The “two-dimensional” MOSFET played a crucial role in one of the famous experiments of modern physics, the discovery of the quantum Hall effect.¹ Klaus von Klitzing received the Nobel Prize for Physics for this discovery in 1985. In this experiment, a MOSFET (Fig. 13.2a) was driven under conditions where the n-inversion channel was so thin that it acted as a two-dimensional electron gas. A low temperature (1.5 K) was applied such that essentially no thermal excitation between the valence band and conduction band occurred. At the same time, the system was exposed to a very high magnetic field perpendicularly to the channel. Under these conditions, the Hall resistance was measured (see Chap. 2 and Fig. 2.3a therein).

Figure 13.7 shows the MOSFET shown in Fig. 13.2b seen from above.

The current I flows from source to drain, and the Hall voltage U_H is measured perpendicularly to the current. The “Hall resistance” is defined as

$$R_H = U_H/I \quad (13.13a)$$

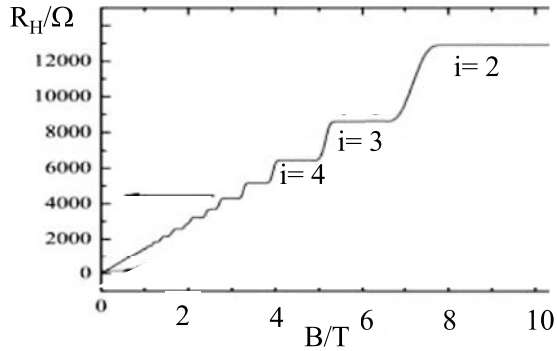
The remarkable result of the pioneering experiment of von Klitzing described above was that the Hall resistance as a function of B shows a stepwise behavior as shown in Fig. 13.8 with the steps situated at

$$R_H = h/e^2 * (1/i) \quad (13.13b)$$

In other words: R_H plotted as a function of B just proceeds in steps from h/e^2 to $\frac{1}{2} h/e^2$ to $\frac{1}{3} h/e^2$ $\frac{1}{i} h/e^2$... where i is an integer number.

¹ Klitzing et al. [1].

Fig. 13.8 Stepwise behavior of R_H as observed in the quantum Hall effect



In that sense, the values for R_H are “quantized”, and R_H comes along in discrete steps.

Physics Comment: The Quantum Hall Effect as a Consequence of Magnetic Flux Quantization

To understand this quantization, let us first relate to the fact discussed in the context of the Hall effect in general (in the context of Eq. (2.18)). There it was pointed out that the Hall voltage U_H depends on the areal density n_s of the charge carriers in the device under consideration via

$$U_H = (1/n_s e)IB \quad (13.14)$$

The reason for the quantization of the Hall effect, as we shall show, is given by the fact that under the conditions of the experiment n_s , the areal density of the electrons in the two-dimensional systems is quantized, i.e. can only take discrete values. We will not give the complete theory of the quantum Hall effect, but will point out its general line of argument, because it is an outstanding example of the effects showing up in electronics on the nanoscale.

The fact that n_s is quantized has in turn to do with the quantization of the magnetic flux B . Let us go step by step:

The magnetic flux density (usually simply called “magnetic flux B ” with the SI unit $1 \text{ T} = 1 \text{ Vs/m}$ always is the result of the combined action of an integer number of flux quanta. Each of those flux quanta produces a magnetic flux $\Phi = h/e$. When we have Z flux quanta per area, they produce a B -field of $B = Z\Phi = Zh/e$. The area F_1 taken by 1 flux quantum is

$$1/Z = h/eB = F_1$$

It now can be shown, by considering the cyclotron motion of electrons in a strong magnetic field at low temperature, that this area F_1 taken by a flux quantum always corresponds to an integer number i of those cyclotron orbits, that means the number n_s of electrons per area equals

$$i/F_1 = ieB/h = n_s.$$

By inserting this into Eq. (13.14), we obtain

$$U_H = \frac{1}{\frac{ieB}{h}} \frac{1}{e} IB = (h/ie^2)I$$

or, for the Hall resistance

$$U_H/I = R_H = h/ie^2$$

in agreement with Eq. 13.13b.

Evidently, the quantum Hall effect is a consequence of the quantization of the magnetic flux, the flux quanta crossing a two-dimensional electron gas as realized in a MOSFET. The famous experiment is shortly discussed here as an example of the kind of quantization effects which occur on the scale of nanometers (“nanometers” here being the thickness of the MOSFET channel) and at low temperatures (the low temperatures around 1 K making the cyclotron orbits the dominant motion of the electrons).

It is left as a little exercise to the reader to show that h/e^2 has the dimension of a resistance (to be given in Ohm). Numerically, by inserting h and e , we get

$$h/e^2 = 25.812,807 \, \Omega \quad (13.15)$$

This equation, by the way, amounts to the modern definition of the resistance unit Ohm. That means: One takes the values of h and e according to their definitions within the actual system of fundamental units, and then the Ohm is defined such that Eq. (13.15) is fulfilled.

One could certainly have taken the value $25.812,807 \, \Omega$ taken from Eq. 13.15 as the new unit of the resistance and might, e.g. have called it 1 Klitzing, but that unit would have been somewhat too large for practical use.²

² The author remembers a talk of von Klitzing where the latter expressed his uneasiness with the idea that in many practical applications, the authors would quote resistances in terms of “Milliklitzing”.

13.5 Questions and Answers

- Question 1: What are the two basic types of transistors?
- Question 2: What is the basic structure of a bipolar transistor (take the pnp-transistor as example)?
- Question 3: Why is it important to have a small width of the basis in the bipolar pnp-transistor?
- Question 4: What is the meaning of “MOSFET”?
- Question 5: What determines the penetration depth of the electric field into the semiconductor under the gate of a MOSFET?
- Question 6: What is the “dielectric relaxation time”?
- Question 7: What makes a system a “two-dimensional electron gas”?
- Question 8: What is the “quantum Hall effect”?
- Question 9: In which way can the leakage currents in a MOSFET be kept small if the area of the gate, and consequently the thickness of the dielectric, is getting smaller and smaller in the course of miniaturization?
- Question 10: What is the modern definition of the unit “Ohm”?
- Answer 1: The bipolar transistor and the field effect transistor.
- Answer 2: A pn-junction poled in forward direction plus a pn-junction in reverse direction: Emitter, basis, and collector.
- Answer 3: Because the transit time across the basis should be small: (a) in order to keep the losses by recombination small, the ratio of transit time to recombination time has to be small. (b) A small transit time means high frequency limit of the transistor.
- Answer 4: Metal–Oxide–Semiconductor Field-Effect-Transistor.
- Answer 5: The dopant concentration in the Si-semiconductor.
- Answer 6: The time it takes for approaching (exponentially) the value 0 of the electric field inside the material after exposing a material to the electric field. Its value is $\tau = \epsilon\epsilon_0\rho$.
- Answer 7: One of the dimensions of the sample has to be considerably smaller than the de Broglie wavelength of the free electrons.
- Answer 8: Under appropriate conditions, the Hall resistance in a two-dimensional electron gas varies in steps of $(h/e^2)/i$, where i is an integer.
- Answer 9: By replacing the SiO_2 with a material with higher permittivity ϵ (a “high k material”). Actually, HfO_2 turns out to be “CMOS-compatible”.
- Answer 10: The “Ohm” is defined by giving the ratio h/e^2 a certain value, i.e. $h/e^2 = 25.812,807 \, \Omega$.

Reference

- 1 K. V. Klitzing, G. Dorda, M. Pepper, New method for high-accuracy determination of the fine-structure constant based on quantized hall resistance. *Phys. Rev. Lett.* **45**(6), 494 (1980)

Chapter 14

Integrated Circuits



Abstract Progress in modern technology is strongly based on the dramatic progress in integrated circuits. Components and basic elements of the fabrication process are photoresists, ion implantation for doping, oxidation furnaces, etching processes, formation of conductor tracks, light sources for photolithography, clean room technology, circuit design, and others, brought together in a “roadmap” defining so-called “technology nodes”. The impressive achievements of integrated circuit technology followed a “top-down” route described over decades by “Moore’s law”. This “top-down route” implies that the basic principles of the technology were carried through to smaller and smaller scales while maintaining their general features. The basic element along this route, being a common feature of each technology step, is photolithography. Proceeding to smaller and smaller structures via photolithography means the application of smaller and smaller wavelengths. So the corresponding light sources started out from spectral lines of the mercury lamp and proceeded via excimer lasers to soft X-ray sources. This development had to go along with a corresponding progress in optics, starting from glass lenses to silica optics, proceeding further to CaF_2 lenses, and finally to grazing reflection systems. In addition, resolution enhancement techniques were developed pushing forward, to some extent, the limits defined by Abbé’s law. The other features present throughout the miniaturization process concern the basic devices, i.e. the use of capacitors and transistors as basic elements. Capacitors store the information represented by their charge state, and transistors act as switches to let charges flow in and out of the capacitor. Miniaturization has reached atomic dimensions. It is approaching limits often described as the “end of Moore’s law”, with the coming up of new “bottom-up techniques” described as processes “beyond Moore”.

14.1 Moore’s Law

The keyword in this development is miniaturization. The dimensions of the devices combined in an integrated circuit have shrunk by many orders of magnitude over the decades, resulting in a corresponding increase in the number of devices in a given area. This development follows, to an amazing degree, a rule known as “Moore’s

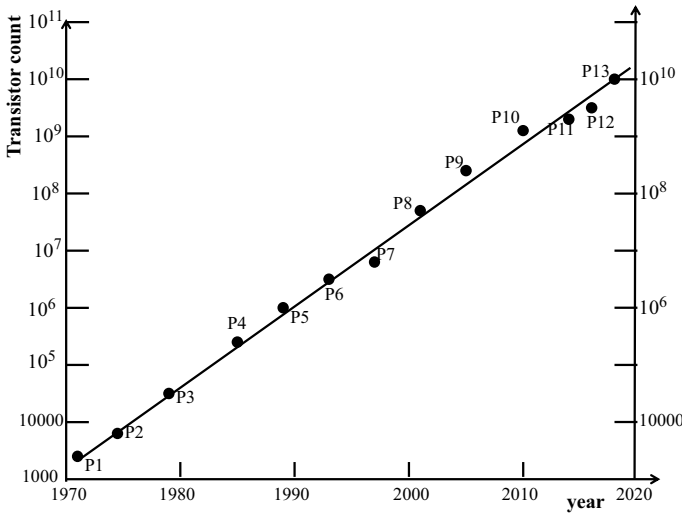


Fig. 14.1 Development of transistor counts in comparison to Moore’s law

law”, saying that the number of individual components on a chip increases by a factor of about 2 every 2 years. Because the transistor (generally a MOSFET) is a central element in modern circuitry, the number of components is often represented by the “transistor count”, the number of transistors on a chip. Figure 14.1 gives the temporal development of the transistor count of processors,¹ the latter being the central processing units (CPUs) organizing operations in a computer.

The solid line gives the best linear fit and corresponds to an increase of the transistor count by a factor of 2 every 2.2 years, thus coming very close to the general phrasing of Moore’s law.

Gordon Moore, born in 1929, is one of the pioneers of Silicon Valley. He is one of the founding fathers of the Fairchild Company. Later, in 1968, he founded together with Andy Grove and Robert Noyce the Intel company, together with Samsung the largest microelectronic manufacturer in the world. He later was CEO of Intel. The “law” named after him was formulated in the present generally accepted form in 1975.

The law as such is not strictly defined. Most often, it refers to the doubling time of transistor counts, sometimes to the doubling time of the transistor counts per cm², which is not the same, because the area of chips differs to some relatively small degree, sometimes processors are taken into account, sometimes Random Access Memories (RAMs), sometimes computing power is included in the discussion of progress, and sometimes required power per operation. So it might be appropriate to

¹ The data are taken from the list of processors given, e.g. in Wikipedia under “transistor counts”. P1–P10 correspond to typical INTEL processors, P11–P13 to typical Apple processors given in the list, representative for the state of the art in the corresponding generation.

Table 14.1 Development of technology nodes

Technology node/nm	1000	180	90	45	32	7
Year	1985	1999	2002	2007/2008	2010	2019

talk, in a slightly more general way, about the doubling of “complexity” of integrated circuits in a time period of about two years.

It looks like remarkable anticipation that Moore after only about one decade made this prediction which turned out to be valid now for 4–5 decades, and it really is. It should be taken into account, however, that we are dealing to some extent with a self-fulfilling prophecy. In the course of time, Moore’s law developed into the basis of the roadmap of a gigantic industry ranging from the big players in chip fabrication to hundreds of suppliers who adjusted their planning along the lines given by Moore’s law.

Actually, the milestones of that roadmap are defined in terms of so-called “technology nodes” which describe in some way the smallest structure which can be produced in the corresponding state of technology. Since 1997, these “technology nodes” are defined in the “International Technology Roadmap for Semiconductors (ITRS)”, produced by the Semiconductor Industries Associations (SIAs). Often, that smallest structure was considered to be represented by the gate length of the MOSFET; in our days, it corresponds to half the distance between neighboring conductor tracks (actually the gate length may be slightly smaller than that distance). A few examples of when those ITRS technology nodes were achieved are given in Table 14.1.

That Moore’s law has come to its end is underlined by the fact that the SIA of the United States in 2017 stopped to contribute to that roadmap² obviously rating its idea as outdated. Nevertheless, technology is proceeding towards even smaller and smaller structure sizes, based on the use of soft X-ray sources for photolithography and corresponding progress in optics. For the present status of development, see the end of Sect. 14.5.

A completely different approach toward smaller and smaller dimensions, finally resulting in molecular scales, may be opened by the “bottom-up” route of “Molecular Electronics” as will be shortly discussed in Chap. 16.

14.2 Photolithography: Components and Steps of Fabrication

Photolithography is the basic process in the fabrication of integrated circuits. Before entering this topic in more detail, we shall discuss some of the crucial components of the process. One of them is the so-called photoresists.

² Sandia National Laboratories SAND2017-dddd J, published in [1].

Photoresists

A photoresist is a polymer the solubility of which in an organic solvent can be changed by exposure to light. This effect is due to a change in the polymer chain length. It can generally be stated that a shortening of the chain length will result in an increase in solubility in an organic solvent. Two kinds of photoresists are in use: “Positive” photoresists and “negative” photoresists.

In a positive photoresist, the solubility is increased by light exposure, due to the fact that the light cuts bonds between the monomers resulting in a shorter chain length. In a negative photoresist, the opposite effect is observed: The solubility is reduced by light exposure. In this case, the light enhances polymerization resulting in a larger chain length and, correspondingly, reduced solubility.

Figures 14.2 and 14.3 give examples of how photolithography is being used to provide the basic steps of integrated circuit fabrication technology. Figure 14.2 shows the case of a positive photoresist.

We consider a Si wafer completely covered with SiO_2 , which in the course of the fabrication process has to be opened at certain spots in order, e.g. to deposit a metallic contact on these spots. The procedure is the following:

1. The wafer (with the SiO_2 layer on its surface) is covered by a positive photoresist.
2. The wafer is exposed [see (a) in the figure] to light on the spots that subsequently have to be opened. This is done by the use of a mask.
3. Because the solubility of the photoresist has been increased at the illuminated spots, the photoresist can be removed at these spots by an appropriate organic solvent [see (b) in the figure].
4. Subsequently, the SiO_2 can be taken off by etching at the spots which are no longer protected by the resist [see (c) in the figure]. The etching of SiO_2 is done through hydrofluoric acid (HF) either via the reaction

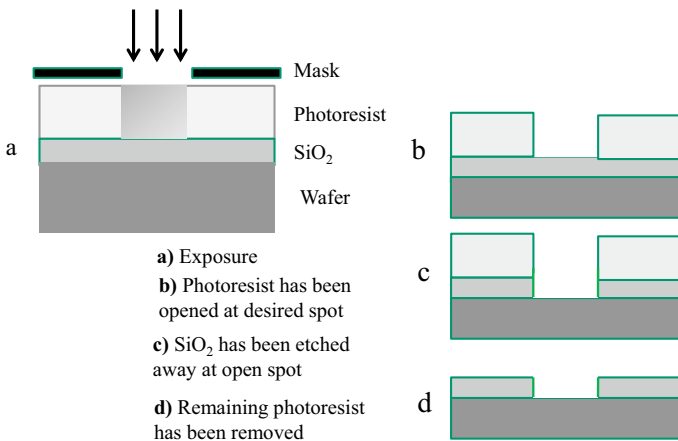
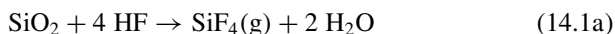
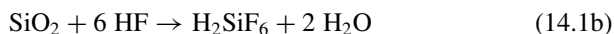


Fig. 14.2 Scheme of photolithography, positive photoresist



or via the reaction



H_2SiF_6 (Hexafluoridokieselsaeure) is a liquid.

5. Finally, all of the remaining photoresist is removed from the wafer via an organic solvent [see (d) in the figure].

So the Oxide has been removed from all the spots that have been exposed to light.

Figure 14.3 shows how, by means of a negative photoresist, a wafer can be covered with metal except on certain points left open.

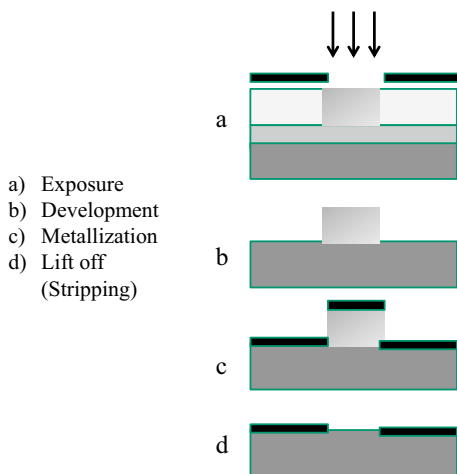
After light exposure through a mask and development of the photoresist, the photoresist just is left over at the spots which were exposed. This is because the solubility of the resist has been reduced in the area where it was exposed. Then the whole wafer is covered with metal (metallization). Then, via the “lift-off process”, the rest of the photoresist is etched away, taking with it the metal on top. The wafer now is covered with metal except at the spots which have been photo-exposed.

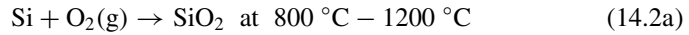
Oxidation

As mentioned already in the context of the MOSFET (Sect. 13.2), the existence of an oxide (SiO_2) forming a closed, insulating layer is one of the most useful and important features of Si, contributing strongly to the success of Si technology. Oxidation can be done in two ways:

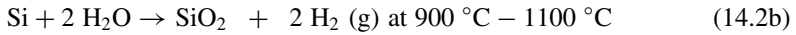
Either by “dry oxidation” exposing the Si to Oxygen:

Fig. 14.3 A negative photoresist in use





or by “wet oxidation” exposing the Si to water vapor:



The wet process is considerably faster than the dry one, but, as already mentioned in Sect. 13.2, dry oxidation results in a more perfect oxide, so it is, e.g. applied to produce the oxide under the gate of the MOSFET which requires high quality.

Conductor tracks

An important feature of integrated circuits are, of course, the conductor tracks connecting the components of the integrated circuit. Conventionally, they were made from Al. There are, however, a couple of arguments which make copper preferable:

- (a) Cu has a smaller resistivity as compared to Al ($1.56 \mu\Omega\text{cm}$ versus $2.45 \mu\Omega\text{cm}$).
- (b) Cu has a higher melting temperature (1084°C versus 660°C). The point here is that after setting up the conductor tracks, no further production steps with process temperature higher than the melting temperature can be applied. Given the relatively low melting temperature of Al, this is a severe restriction in fabrication processes using Al tracks.
- (c) Electromigration

Electromigration occurs at high current densities in a conductor track. Under this condition, the moving conductor electrons exert a force on the ion cores sufficient to remove them from their regular sites. Those metal ions then might migrate into the semiconductor and affect its electronic properties. Electromigration of Cu is about ten times slower as compared to Al. For Al tracks on Si, this has been a major problem in Si technology.

Cu, however, goes along with a very serious problem. It easily diffuses into Si via interstitial diffusion (Sect. 6.2, Fig. 6.3). After entering Si, it forms a deep level acting as a strong recombination center resulting in a considerable shortening of the carrier lifetime. This phenomenon has to be suppressed. This can be done by inserting a Tantalum layer between the Cu conductor track and the Silicon. This layer acts as a diffusion barrier and so prevents the Cu from entering the Si. Actually, it turned out that the combination of a layer of Ta_2N (ca. 50 nm) together with a Ta liner of a few nm thickness, as shown in Fig. 14.4, provides a more effective barrier.³

Ion implantation

A most important tool in integrated circuit technology is ion implantation, being generally used for doping of materials. The ions to be applied for doping purposes (Boron, Phosphorus...) are produced in an ion source, then accelerated in an ion accelerator by voltages of usually some 100 keV, up to even a few MeV. They are

³ Tantalum as a diffusion barrier between copper and silicon: Failure mechanism and effect of nitrogen additions: K. Holloway, et al. [2].

Fig. 14.4 Cu conductor track with Tantal and Tantal-N diffusion barriers

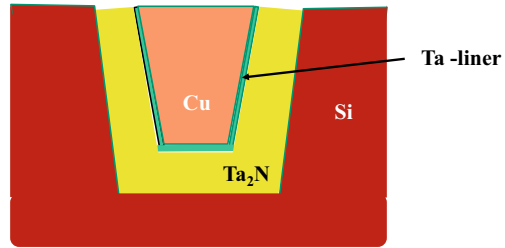
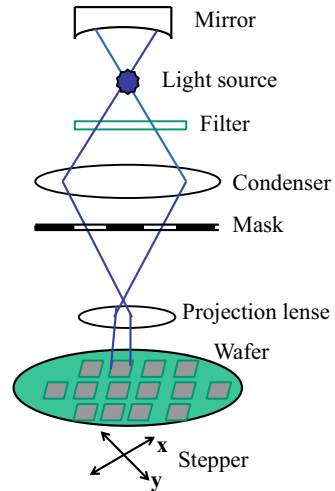


Fig. 14.5 Mask aligner



focused to a high degree, and then deposited on the chip at the desired spot. The energy of the ions defines the penetration depth of the ions and thus the layer inside the chip where the dopants are required, ranging from 10 nm at low energies up to a few microns at high energies. A major issue is the depth distribution of the dopants which can be modeled by Monte Carlo simulations. Another big topic is annealing. The process of ion implantation generally results in severe radiation damage, so the crystal structure has to be restored by thermal annealing processes which heavily affect the resulting electronic properties. All in all, doping by ion implantation is a sophisticated and essential part of integrated circuits technology.⁴

14.3 Exposure in the Photolithography

Figure 14.5 shows the optical scheme for the exposure **process** via a “mask aligner”.

⁴ Ion implantation Science and Technology. J. F. Ziegler [3].

The figure shows the way the image of the structure contained on the mask is transferred optically to the wafer in order to illuminate the photoresist. The optics, by the way, are the same as those in the classical slide projector. Only a few technical details shall be given.

The opaque areas of the **photomask** usually consist of a layer of chromium on a glass substrate.

The manufacturing of the masks, certainly a crucial step in circuit technology, is done basically in the same way as one of the chips. An important difference is the use of electron beam lithography instead of optical lithography when very high resolution is required. Because electron beam lithography is a serial process, the image being produced point by point, not by imaging as a whole, the process is much slower, which, however, does not matter too much because only one mask is needed for the exposure of a large number of wafers in a production line.

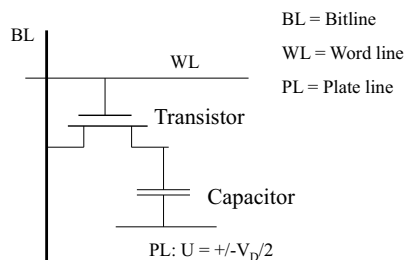
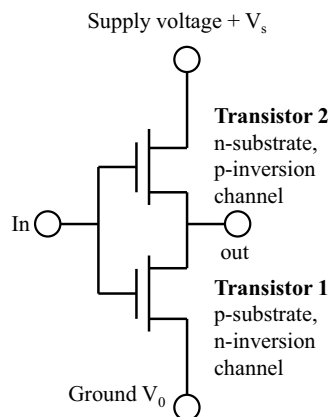
The mask and the optical system are part of a “**mask aligner**” which has to guarantee the correct positioning of the wafer relative to the mask. Because the wafer is not exposed as a whole, but “step by step”, a “**stepper**” (see Fig. 14.5) moves the wafer with extreme precision. It is positioned by means of markers on the wafer. It should be mentioned that in modern technology, the traditional mask aligner system may be replaced by a **laser writer**. The use of a laser writer is, as electron lithography, a serial process. For that reason, mainly masks are produced in that way. But ultra fast systems make laser writers more and more interesting for the exposure of wafers as well, in particular in cases where only a limited number of chips is required. The big advantage of laser writers is that they can be programmed individually for each desired structure. Some experts believe that the laser writer is the “mask aligner of the future”.

14.4 Some Important Constituents of Integrated Circuits

Dynamic Random Access Memories (DRAMs) have often been the groundbreaking element in the further miniaturization of integrated circuits. In a DRAM, each memory cell can be addressed individually.⁵ Figure 14.6 shows a DRAM.

The capacitor contains the bit-information, logic 1 corresponding to voltage $+V_D/2$, logic 0 to $-V_D/2$ at the capacitor. Via the word line, connected to the gate, the transistor switch is opened or closed in order to charge or discharge, via the bit line, the memory capacitor to $\pm V_D/2$. Via the plate line, the voltage level of the capacitor is defined.

⁵ A good illustration is given by Wikipedia: “A DRAM would correspond to a book. Each page can be opened individually, as opposed to a parchment scroll, which can only be opened “sequentially” by scrolling.

Fig. 14.6 DRAM cell**Fig. 14.7** Inverter as an example for CMOS

The inverter as an example of CMOS technology

The most important constituents of computers are the so-called “logical gates”. Logical gates are electronic circuits which represent basic logical functions, the so-called “boolean functions”. The most simple one will be discussed in the following. It will at the same time give an introduction to the notion of “CMOS technology”. Figure 14.7 shows an inverter.

Its function is the following: If “In” is 1, “Out” is 0, and if “In” is 0, Out is 1. It consists of two MOSFETs, one (no.1) with a p-substrate and thus opened via creating an n-inversion channel, the other one with an n-substrate, thus opened via the creation of a p-inversion channel (see Sect. 13.2). Let the ground V_0 be 0 V and the supply voltage V_s be positive. The “In”- electrode is connected to both gates. We first consider the case that “In” is + 1, represented by a positive voltage larger than the threshold voltage of transistor 1. In this case, transistor 1 is open, because an n-inversion channel is formed, and transistor 2, at an appropriate value of its threshold voltage, is closed. So “out” is on the ground level $V_0 = 0$ corresponding to logical 0; the input signal 1 is inverted.

It can be easily seen that in the case of “In” being on ground level $V_0 = 0$, transistor 1 is closed and transistor 2 is open. So “Out” is on the positive voltage level $+V_s$, corresponding to logical 1.

In a similar way, circuits for the other logical gates AND, NAND, OR, NOR, XOR, and XNOR can be formed.

The point is that, as shown through the example of the inverter, these circuits can be built in a particularly simple way by using both types of MOSFETs, one with a p-substrate, opened by forming an n-inversion channel, and the “complementary” one with n-substrate, opened by forming a p-inversion channel. It was explained in Sect. 13.2 that from a purely functional point of view with regard to an individual MOSFET, it is better to have a p-substrate opened by an n-inversion channel because of the considerably higher carrier mobility of the electrons in Si (as compared to the holes). However, the advantage of a simpler circuit design via the use of both types of MOSFETs is a sufficient argument to make use of the “complementary” technology. For this reason, the technology of integrated Si circuit technology is given the general name **CMOS technology (for “Complementary MOSFET technology”)**. It is a widely used technical term. When, e.g. a new material or a new feature is introduced into Si-technology, often the question is asked whether this innovation is “CMOS-compatible”.

14.5 Light Sources for Photolithography

According to Abbé’s law, light can be focused only on a spot size of the order of the light wavelength, more precisely the limit of structure size d is given by

$$d = k_1 * \lambda / \text{NA} \quad (14.3)$$

where λ is the wavelength of the light and NA the numerical aperture. k_1 is a process-related factor determined, e.g. by the actual lens system, photoresist and other factors, usually around 0.4–0.5.

So in the course of structure miniaturization, shorter and shorter wavelengths had to be applied. For a long time, the development proceeded along the use of shorter wavelengths of the mercury discharge lamp. To give an example: The 10 μm technology node was achieved by making use of the so-called G-line at 435.8 nm. At the 1 μm technology node, lithography switched to the “i-line” at 365.01 nm. Big progress was achieved—and a big technological challenge was mastered—when the mercury lamp was replaced by the excimer laser. While for the 250 nm node the KrF laser ($\lambda = 248$ nm) had been used, the ArF laser ($\lambda = 193$ nm) came into play for the 180 nm node and the following one.

It should be mentioned at this point that the most critical step in this context was the development of a suitable material for the lenses. While for 248 nm (i.e. for the KrF laser) fused silica, a synthetic quartz produced from the gas phase could still be used as a lens material, for 193 nm (the ArF-laser) and 157 nm (the F_2 -laser) a completely new material had to be provided for the lenses, monocrystalline CaF_2 . CaF_2 is still sufficiently transparent down to 193 and 157 nm, but, by the

way, not beyond 125 nm. It must be emphasized that the provision of large lenses of this material proved to be a great challenge to crystal growth technology, given the high demand for crystal quality and purity. Purity and crystal quality are paramount requirements to assure sufficient optical quality. An important issue in this context is also “radiation resistance”, the stability of the material toward damage under heavy UV radiation. Radiation defects could act, e.g. as scattering centers, giving rise to a background of scattered light disturbing the fabrication process. By producing defect centers of lower symmetry, radiation damage could result in an optical asymmetry in the otherwise cubic CaF_2 .

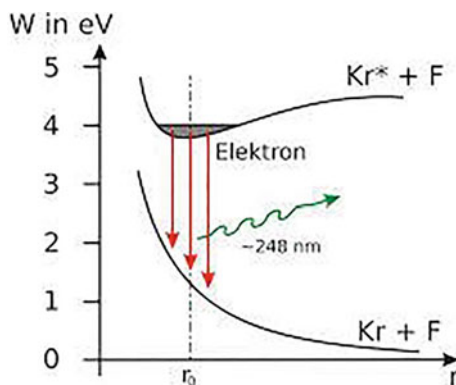
Physics Comment: The Excimer- and the Exciplex-laser

Given the fact that excimer lasers played a central role along the road to shorter and shorter wavelengths for photolithography, we shall give a short introduction to this type of laser here.

The laser principle has been extensively treated in Chap. 12. The “laser condition” for an atomic or molecular laser is (12.3 and Fig. 12.1a) $N_2 > N_1$, saying that the occupation N_2 of the upper level of the laser transition should be larger than the one of the lower level (N_1). “Excimer” stands for “excited dimer”. The dimer in this context consists of two noble gas atoms like Ar_2 or Kr_2 . “Exciplex” stands for “excited complex” like ArF or XeF . At first sight, the notion of a molecular system like Ar_2 or ArF seems most surprising, because, as generally known, noble gases do not form molecules. Talking about lasers, however, use is made of an amazing fact: Ar_2 or KrF do not form molecules in their ground state, but form a molecular compound in their electronically excited state. The situation is explained in Fig. 14.8 for the case of KrF .

The figure shows in its lower part (“Kr + F”) the interaction energy W of a Kr- and a F-atom as a function of their separation r , when both atoms are in their ground state. Because the atoms repel each other, the energy increases

Fig. 14.8 Exciplex level scheme and optical transitions



continuously with smaller separation r , no energetic minimum exists which would correspond to a bound state. The situation looks different, however, if the Kr is in an electronically excited state Kr^* (see curve $\text{Kr}^* + \text{F}$). In this case, a minimum of the interaction energy W exists for a separation r_0 , corresponding to a bound state with atomic distance r_0 . This fact seems most surprising given the general fact that molecules are more strongly bound in their ground states than in their excited states. So why does a bound state exist when the Kr is in an excited electronic state, and not in the ground state? The answer is the following.

After electronic excitation of the Kr, the excited electron undergoes a (“partial”) transition to the F-atom such that the system resembles a molecule (Kr^{*+}F^-). In such a system, the chemical bond is no longer based on the van der Waals interaction as in the ground state, but on the much stronger ionic interaction. The ionic interaction results in a bound state at distance r_0 . A corresponding situation is encountered in other excimer- and exciplex systems.

The decay of the electronically excited state (Kr^{*+}F^-) to the ground state ($\text{Kr} + \text{F}$) under photon emission can be used for a laser system. In an excimer- or an exciplex-laser, the excitation (the “pumping process”, i.e. the excitation of the Kr-atom) is done by an electron beam. Wavelengths are in the UV; see the numbers given above for Ar_2 and KrF .

A special feature of the excimer- and exciplex-laser favoring high intensities is given by the fact that the lower state of the optical transition is unstable—it decays within roughly 10^{-12} s—so that the laser condition $N_2 > N_1$ (12.3 and Fig. 12.1a) can be easily upheld.

Recent development in lithography light sources: Soft X-rays

For the most recent technology node in 2019 (7 nm!), a completely new concept had to be applied, the one of EUV lithography (Extreme Ultraviolet). Here, a soft X-ray source of 13.5 nm is applied. The biggest challenge in setting up the EUV lithography is given by the fact that all materials, even air, strongly absorb this wavelength, so completely new concepts for the optics have to be set up. Instead of lenses mirror optics have to be used, usually with grazing incidents, where high reflection can be achieved for soft X-rays. To keep Rayleigh scattering low (Rayleigh scattering is the scattering process at objects much smaller than the wavelength), the surface roughness of all components involved in the exposure process has to be kept extremely low (<0.25 nm). Because of the high absorption coefficient in air, the whole system has to be operated in vacuum. Nevertheless, the first results with 45 nm structure size were already reported in 2008; EUV with 13.5 nm is on the market since 2019, a major technology player, by the way, being the Zeiss company.

It should be mentioned at this point that grazing incident reflection is not the only way to achieve reasonable reflectivities in EUV optics. Multilayer systems are being

used to achieve high reflectivity even at angles close to 90° .⁶ Multilayer systems are not only used for mirror optics in the EUV, but also are the basic elements of the masks required for lithography. It was mentioned in Sect. 14.3 that the conventional photomasks consist of a layer of Chromium on a glass substrate, the chromium layer defining the opaque areas of the mask. Contrary to this concept, the photomasks in EUV optics are based not on the contrast between absorbing and transmitting areas, but between absorbing and reflecting areas. The reflecting areas consist of typically 40 layers alternating between silicon and molybdenum, providing a system of Bragg diffraction (see Sect. 12.2, context of Fig. 12.6, for multilayer reflectors).

For the wavelength of $\lambda = 13.5$ nm so far no laser sources are available, instead plasma discharges are being applied; at present, preferentially Sn is in use as a medium. To give an idea of the sophisticated technology: In such a plasma discharge source, some 10^4 Sn droplets per second are falling into a vacuum chamber. After the actual position has been monitored, the droplet is hit by an intense laser pulse producing the plasma which then emits the 13.5 nm. Intensity, of course, is a great issue; some 100 W in the spectral region of 13.5 nm are required.

Physics Comment: On the Production of Soft X-rays

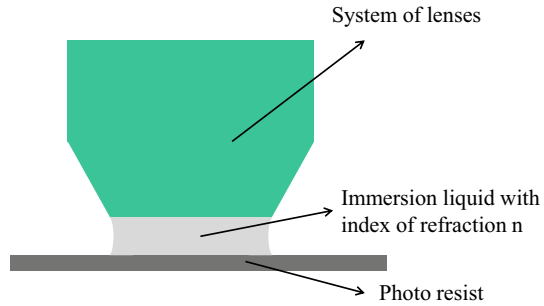
While the emission lines of atoms in their neutral or singly ionized states are in the region of a few 100 nm, considerably shorter wavelengths are encountered in more highly ionized atoms. This means that in order to achieve radiation in the region of, e.g. 13.5 nm a plasma has to be produced by providing sufficient energy for heating up the atomic ensemble. In the case of the EUV system (13.5 nm) considered here, this is done by a CO₂ laser, producing emissions at 9.6 μm and 10.6 μm corresponding to transitions between vibrational states of the CO₂ molecule. The CO₂ laser is suitable for extremely high power applications; actually, the one discussed here is one of the few applications of the CO₂-laser in our days. The laser produces the Sn-Plasma with atoms in highly ionized states which then emits the 13.5 nm.

It is most remarkable that this kind of system could be developed “in time”, i.e. in agreement with the “schedule” set by Moore’s law! Actually, in 2017 IBM presented a 5 nm test chip; actually, in 2020 Samsung and Taiwan Semiconductor Manufacturing Company (TSMC) have applied EUV technology in production, and most recently IBM presented its first 2 nm test chip!⁷ About 4 years from now commercialization is expected.

⁶ A system of that kind (used for imaging a line in EUV lithography) is described in Wikipedia under “Extreme ultraviolet lithography”.

⁷ <https://newsroom.ibm.com/2021-05-06-IBM-IBM-Unveils-World's-First-2-nm-Chip-Technology>, Opening a New Frontier for Semiconductors.

Fig. 14.9 Scheme of immersion lithography



14.6 Resolution Enhancement Techniques

The large gap between the 193 nm- and 13,5 nm-lithography was partially bridged by certain “resolution enhancement techniques” which helped to achieve structure sizes considerably below 193 nm even by exposure at this wavelength. Two of them will be shortly presented in the following.

Double patterning

Double patterning is⁸ a fascinating way of lithography capable of pushing down the k_1 -value (14.3) below 0,25. Without going into details the procedure is the following: In order to achieve a structure considerably smaller than the wavelength, the lithography process is decomposed into two exposure steps. In each step, a structure is produced within the Abbé resolution limit, but the superposition of the two structures results in the desired pattern below the resolution limit. The process can be extended to more than two such exposures.

Immersion lithography

Another way to reduce the Abbé limit of the structure size d (14.3) is the introduction of an immersion liquid with an index of refraction n between the lens system and the photoresist; see Fig. 14.9.

This procedure is well established in microscopy and reduces d by a factor of n . $n = 1.33$ in case of water.

Phase shift masks

An interesting scheme of resolution enhancement is the one of phase shift masks. It is explained in Fig. 14.10 for the example of a periodic structure.

The left half of the figure shows the “conventional” case of a mask with a periodic on–off structure with a periodicity a . The mask is exposed to a light beam of wavelength λ . In the plain of the mask, the E-field is zero where the mask is closed

⁸ Chang-Moon Lim, Seo-Min Kim, Young-Sun Hwang, Jae-Seung Choi, Keun-Do Ban, Sung-Yoon Cho, Jin-Ki Jung, Eung-Kil Kang, Hee-Youl Lim, Hyeong-Soo Kim, and Seung-chan Moon “Positive and negative tone double patterning lithography for 50 nm flash memory”, Proc. SPIE 6154, Optical Microlithography XIX, 615,410 (15 March 2006); <https://doi.org/10.1117/12.656187>.

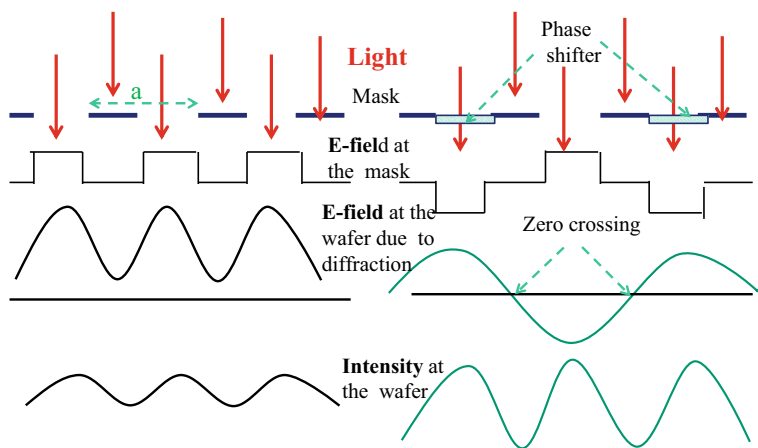


Fig. 14.10 The phase shifter method. Left side: no phase shifter in the mask; right side: With phase shifter

and nonzero where the mask is open. This clear structure is softened in the plain of the wafer due to diffraction. The intensity of the light, i.e. the square of the E-field, at the wafer is shown in the lowest line.

On the right side of Fig. 14.10, the same sequence is shown, however, a phase shifter of 180° has been installed in every other opening of the mask. The E-field right below the mask is opposite in sign under the phase shifter as compared to the field under an opening. Again, the structure in the plain of the wafer is softened due to diffraction. As seen in the lowest line, the intensity, i.e. the square of the E-field, now has a much sharper structure than the one obtained without a phase shifter due to the fact that the E-field here shows a zero crossing!

The general idea of the phase shifter method is that the mask is structured in such a way that by diffraction the desired structure, as sharp as intended, is produced on the wafer. The complexity has been transferred into the mask! A tricky idea is to use diffraction for reduction of the negative effects of diffraction!

14.7 Some Concluding Remarks on the Present Situation of Chip Technology

The technology node 7 nm was achieved in 2019 (Table 14.1). As mentioned at the end of Sect. 14.5, 5 nm technology has been used in production since 2020 by Samsung and Taiwan Semiconductor Manufacturing Company (TSMC), and IBM in 2021 presented its first 2 nm test chip.

It is evident that miniaturization cannot proceed much longer; the technology node 7 nm corresponds to approximately 30 atomic diameters, not to mention the 5 nm and the 2 nm already realized in test chips, as mentioned above. So evidently

transistors can't get much smaller. Development will rather go toward even closer packaging of components and, last but not the least, use of the vertical dimension in Si wafers.

The fabrication of an integrated circuit is a process of enormous complexity, requiring a sequence of 300–500 steps, each of them with an extremely high reliability in order to assure a sufficient yield of the production line. Only very large companies can afford the huge investments required to set up a chip factory. The big ones at present are INTEL in the Silicon Valley, the pioneering company founded by Gerald Moore (see Sect. 14.1), and Samsung from South Korea, each of them with a turnover of about \$60 billion on a chip market of together about 440 billions in 2019. IBM and TSCM have to be included in this list of leading companies.

This number of 440 billion dollars, impressive as it might be, does by no means reflect the true importance of chip technology. Chips are the key technology of our time, determining progress in almost all sections of industry in particular and society in general, be it mobility, health, environment, artificial intelligence, and what have you! It should be mentioned at this point that access to chips and availability is a dominant strategic issue⁹ in the competition between the world powers, the US and China. The fact that TSCM, a technologically and commercially leading protagonist in the field, a Taiwan company, gives an additional degree of complexity to the US-China issue.

14.8 Questions and Answers

- Question 1: What does Moore's law say?
 Question 2: What is a "technology node"?
 Question 3: Which technology node was achieved in 2019?
 Question 4: What is a photoresist?
 Question 5: What is a positive, and what is a negative photoresist?
 Question 6: What is the dominant insulating material in integrated circuits?
 Question 7: How is SiO₂ produced in the fabrication of integrated circuits?
 Question 8: Which material is used for the opaque areas of a photomask?
 Question 9: Name the light sources that followed each other in the history of integrated circuits!
 Question 10: Which material for the lenses is used in the extreme UV (193 nm, 157 nm)?
 Question 11: Name a few resolution enhancement techniques.
 Answer 1: The transistor count on a chip increases by a factor of about 2 every 2 years. Or, the complexity of integrated circuits doubles about every 2 years.

⁹ At the time when this book was being written (end of 2021, beginning of 2022), shortage of chips heavily slowed down the recovery of world economy after the Corona crisis.

- Answer 2: A technology node describes the smallest structure which can be produced in the corresponding state of technology. At present, “half the distance between neighboring conductor tracks” is taken as this smallest structure, before, it was the gate length of the MOSFET.
- Answer 3: Technology node 7 nm.
- Answer 4: A photoresist is a polymer the solubility of which can be changed by exposure to light. There are positive and negative photoresists.
- Answer 5: The solubility of a positive photoresist is increased by light exposure, and the one of a negative photoresist is decreased by light exposure.
- Answer 6: SiO_2 . It forms a closed insulating layer on Si.
- Answer 7: By wet or dry oxidation. Dry oxidation is the slower process resulting in an oxide of higher quality.
- Answer 8: Chromium on glass usually.
- Answer 9: Mercury lamp, different spectral lines of shorter and shorter wavelengths following each other. After that, excimer lasers and exciplex-lasers came in, followed by soft X-ray sources, now at 13,5 nm.
- Answer 10: Monocrystalline CaF_2 .
- Answer 11: Double patterning, immersion lithography, and phase shift masks.

References

1. Erik P. (2017). Sandia national laboratories SAND2017-xxxx J, published in DeBenedictis, “Its time to redefine Moore’s law again”. *Computer* **50**, (2), 72–75
2. Holloway K, Fryer PM, Cabral C Jr., Harper JM, Bailey PJ Kelleher KH (1992) *J. Appl. Phys.* **71**, 5433; <https://doi.org/10.1063/1.350566>
3. Ziegler JF (1988) *Ion implantation science and technology*. Academic Press Boston, ISBN 0-12-780621-0

Chapter 15

Organic Electronics



Abstract It came as a big surprise to the electronic community once it turned out that organic materials, small molecules as well as polymers, could be used for active electronic materials. Indeed, conductivity can be given to polymers by a special type of “doping”. More important, however, is the approach to supply free charge carriers into insulators by injection. In this way, “electron conducting materials” and “hole conducting materials”, corresponding to n-type and p-type materials in inorganic electronics, can be obtained. On this basis, pn-junction devices like light-emitting diodes (OLEDs) and solar cells are being developed and have reached or approached market maturity. A very promising aspect of organic electronic materials is the one of the “printing technology” for electronic devices and circuits.

15.1 Organic Materials for Electronics, a Totally Different Approach

In the previous chapters, we were dealing with “classical” semiconductors like Si, GaAs, and SiC, all of them inorganic materials. More recently, beginning in the last decade of the past century, a totally new class of electronic materials has attracted much attention and is the focus of many activities in the field of electronic devices like light-emitting diodes, solar cells, sensors, and others: Organic semiconductors.

The fact that organic materials like polymers could be used as “active” elements in electronic devices came as a surprise to most of the community because they had traditionally been considered and applied as insulators for cable insulation and other purposes. Could it be that an excellent insulator with essentially no free charge carriers can act as an active element in electrical devices and circuits? A convincing simple answer to this question is given by the old vacuum tube, which dominated electronics in the first half of the last century. It is certainly a demonstrator that a perfect insulator can carry electronic current because there is no better insulator than the vacuum. The point is that the charge carriers in the vacuum tube are not intrinsically present, but they are injected via the glow electrode. An analogous thing happens in organic electronics.

15.2 Conjugated Double Bonds

We start with some basics of organic chemistry. The carbon atom has two $2s$ electrons and two $2p$ electrons in its outer shell (the shell number 2). They hybridize into 4 identical sp^3 -orbitals pointing toward the edges of a tetraeder.

This results in the diamond structure shown in Fig. 15.1.

The sp^3 -orbitals in diamond do not overlap; they are “localized”, so diamond is an insulator.

The situation is different in the case of graphite, the other well-known modification of Carbon; see Fig. 15.2.

The $4s^2 2p^2$ electrons hybridize in this case to 3 sp^2 -orbitals, lying in the plain of the carbon ring forming an angle of 120° relative to each other, and one π -orbital, sticking vertically out of the plain (green color). The π -orbitals overlap, so these electrons are delocalized and thus give rise to the metallic conductivity of graphite.

Figure 15.3 shows the case of benzene.

Fig. 15.1 The diamond structure

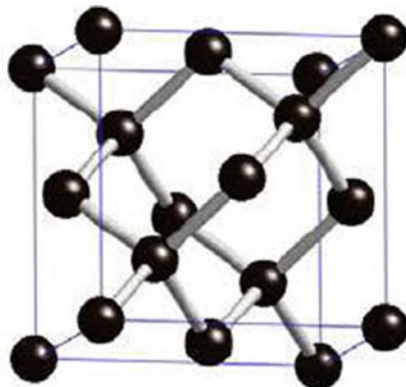


Fig. 15.2 The chemical bonding situation in graphite

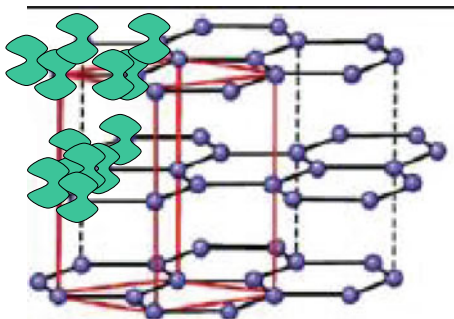
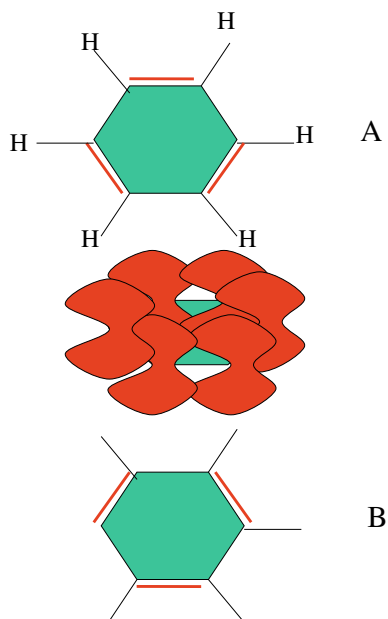


Fig. 15.3 sp^2 -orbitals (black lines) and π -orbitals (in red) in benzene



Again, like in the case of graphite, three of the four valence electrons are hybridized to sp^2 -orbitals. They are drawn as black lines in the figure. One of them binds the hydrogen atom.

The red lines in part A and part B of the figure mark the fourth electron, which is a so-called π -state (the “ π ” refers to the quantum number of angular momentum which is 1 here). The π -electrons can be included in the model of the benzene ring either the way shown in part A or in part B; both representations are equivalent. Such a situation of alternating single and double bonds, resulting in the two equivalent representations in parts A and B of the figure, are called “conjugated double bonds”. The important general aspect of these “conjugated double bonds” in our context is that the corresponding orbitals (red in Fig. 15.3), which are vertical on the plain of the ring, overlap; they can no longer be attributed to a particular atom; they are delocalized on the ring.

15.3 Band Structure and Conductivity in Polymers

In the following, we will consider charge transport on a chain of carbon atoms that form the backbone of a polymer. The most straightforward example for this case is Polyacetylene, shown in Fig. 15.4.

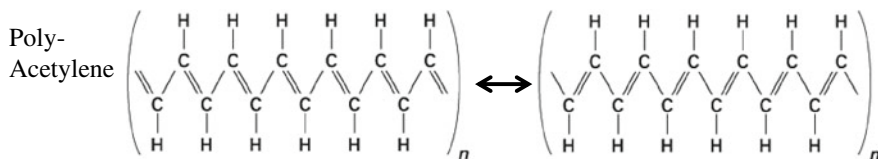


Fig. 15.4 The conjugated double bonds in polymers shown, for example, Poly-Acetylene in the two equivalent representations

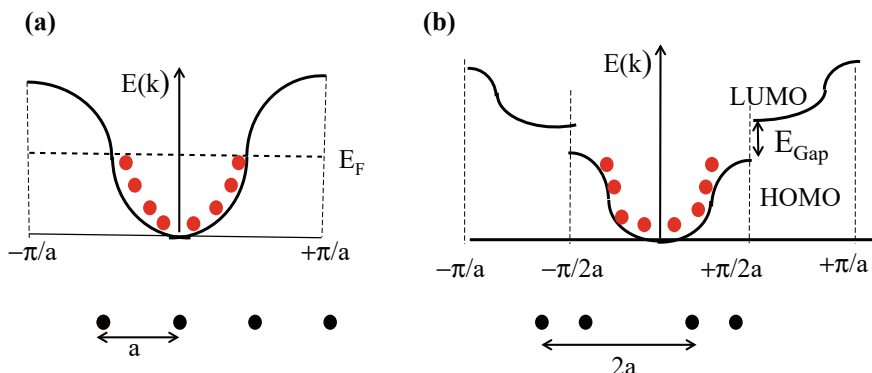


Fig. 15.5 **a:** Bandstructure of a polymer with separation of the atoms of the linear carbon chain **b:** Bandstructure of a polymer with alternating separations of the atoms in the carbon chain (Peierls' instability)

We shall limit our discussion on this case in order to present some basic concepts on charge transport in polymers, as first outlined in ref.¹ For a more general discussion including other polymer structures, as well as issues like transport in amorphous and crystalline organic materials, we refer to ref.²

Like in graphite, three sp^2 -orbitals forming an angle of 120° with each other are situated in a plane, with the fourth electron, the π -electron, vertical on that plane. The spatial extension of the π -electrons is such that they partly overlap; in that way they are delocalized and give rise to an electrical conductivity along the carbon chain. Analogous to the situation outlined for the benzene ring, the double bonds between the carbons can be visualized in the two equivalent representations shown in Fig. 15.4.

The band model states for the case of a polymer chain with separation a of the carbon atoms that the lowest not fully occupied band (i.e. the conduction band) would be half filled (Fig. 15.5a).

The figure shows the first Brillouin zone for this case, extending from $-\pi/a$ to $+\pi/a$. So far, the situation is like the one in a metal (Chap.1, Fig. 1.4); the polymer

¹ Su et al. [1].

² Baessler and Koehler [2].

would show a metallic conductivity. This, however, is not found in reality. This is the result of an effect shown in Fig. 15.5b, the so-called Peierls’ instability.

“Peierls’ instability” refers to the fact that a carbon chain with equidistant atoms separated by distance a turns out to be unstable. Instead, a situation will develop where longer and shorter distances alternate, the double bonds corresponding to the shorter separation. Energetically this is more favorable. As a result, the periodicity of the system is no longer a , but $2a$. The conduction band, which was half full in the case of periodicity a , now splits into two bands, the lower one now being completely full, the higher one being empty, and with a bandgap in between. This corresponds to the situation of a semiconductor (or an insulator, depending on the size of the bandgap); see Sect. 1.3. So the Peierls instability transforms a polymer chain from a metallic system into a semiconductor.

Let us point out here that the Peierls instability is a nice example of the phenomenon that a change in the periodicity of a system can result in a drastic change in the band structure. We shall encounter an equivalent case in Sect. 16.3 when dealing with metallic and semiconducting carbon nanotubes.

The upper band in Fig. 15.5b, the empty one, is called the “lowest unoccupied molecular orbital”, the “LUMO”. It corresponds to the conduction band in inorganic semiconductors. The lower band is called the “highest occupied molecular orbital”, the “HOMO”, corresponding to the valence band in inorganic semiconductors.

15.4 Polymer Conductivity by “Doping”

As long as the LUMO is empty and the HOMO is full, no conductivity exists (Sect. 1.3). Only if electrons are brought into the LUMO, the polymer achieves an n-type conductivity along the direction of the carbon chain, or if electrons are removed out of the HOMO, leaving holes behind, a p-type conductivity is created.

These processes are represented in Fig. 15.6a, b.

The upper line of Fig. 15.6a shows the creation of a negative “polaron”. An electron jumps from the reducing agent Na on the nearby CH-orbital, leaving behind a Na^+ ion. From there, the electron jumps along the carbon chain from CH-orbital to CH-orbital (see the second and third lines of Fig. 15.6a), following an external electric field and thus giving rise to a current. The Na^+ ion, being much less mobile than the electron, stays where it is. An important notion has to be introduced at this point: The (negatively charged) electron on the CH-orbital causes a deformation of its environment, so what moves along the chain is not just the electron, but this negatively charged deformation region which is called a polaron, in this case a negative polaron.

By adding an oxidizing agent like Iodine to the polymer, an electron can be taken out of the HOMO, and a hole conductor results. This is shown in Fig. 15.6b.

The electron jumps from a CH-orbital to the Iodine (the electron then is bound to a I_3^- complex). The remaining positive charge on the orbital gives rise to a local deformation. What moves along the chain is this positively charged deformation region, the “positive polaron” (second and third lines of Fig. 15.6b).

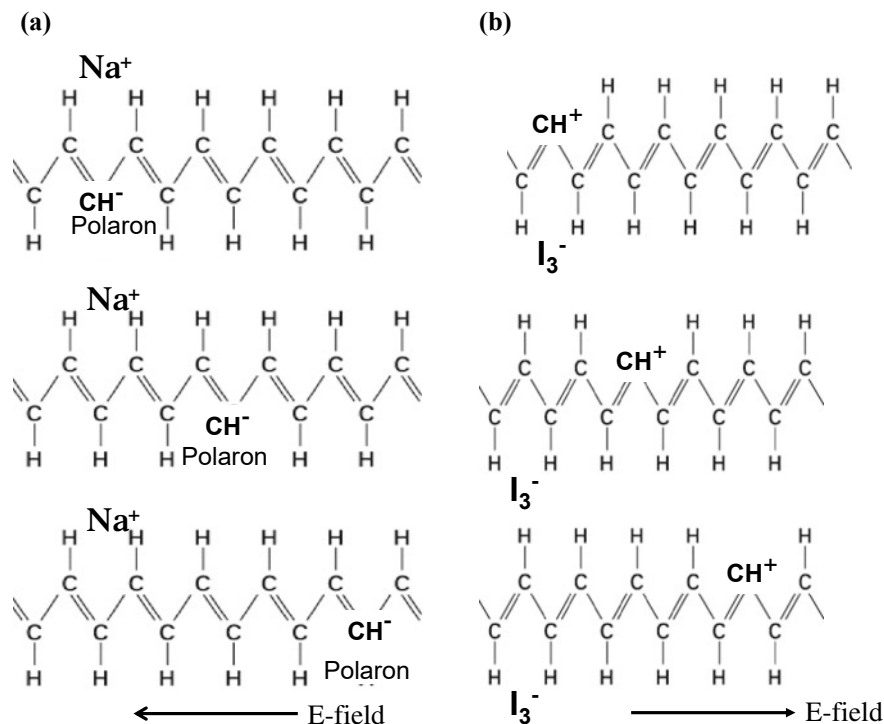


Fig. 15.6 **a:** Migration of a negative polaron on the carbon chain of a polymer **b** Migration of a positive polaron on the carbon chain of a polymer

In this way, by adding a few percent of an oxidizing or a reducing agent to a polymer, electrical conductivities can be achieved even close to the ones of metals. The high conductivity of polyacetylene upon doping with iodine was first reported³ in 1977. Alan J. Heeger, Alan Mac Diarmid, and Hideki Shirakawa were rewarded the Nobel Prize in Chemistry in 2000 for the discovery. In the “old days”, high conductivities in polymers could only be reached by adding soot to the organic material. An undesired side effect of that procedure was that conductive polymers always looked black, a property not desired in many circumstances and applications.

It should be noted that the notion of “doping” used in this context has a meaning quite different from the one used in “classical” semiconductor technology (Chap. 3). The difference is already obvious from the concentrations of the dopants applied in order to provide a decent conductivity. The “dopant” concentration of a conjugate polymer is of the order of percent, while in Si it would (roughly) be some ppm. While the doping of, e.g. Si directly produces additional free charge carriers in the bulk of the material, providing a conductivity proportional to the doping concentration, the “doping” of the polymer only provides a conductivity to the individual polymer

³ Shirakawa et al. [3].

chains. The transport of the charges then consists in “running along the chain”, then “jumping to the next chain”, and so on many times from electrode to electrode. So the total resistivity of the polymer consists of two contributions, the one resulting from the movement of the charge along the polymer backbone, and the other one from the jumping process from polymer chain to polymer chain. The latter contribution tends to be the dominating one. It is obvious that for this process to be effective, a considerable fraction of the polymer molecules has to be made conductive, at least a few percent.

The situation is visualized in Figs. 15.7a, b.

Fig. 15.7 **a** Low concentration of highly conducting polymer chains (shown in yellow) **b** High concentration of highly conducting polymer chains

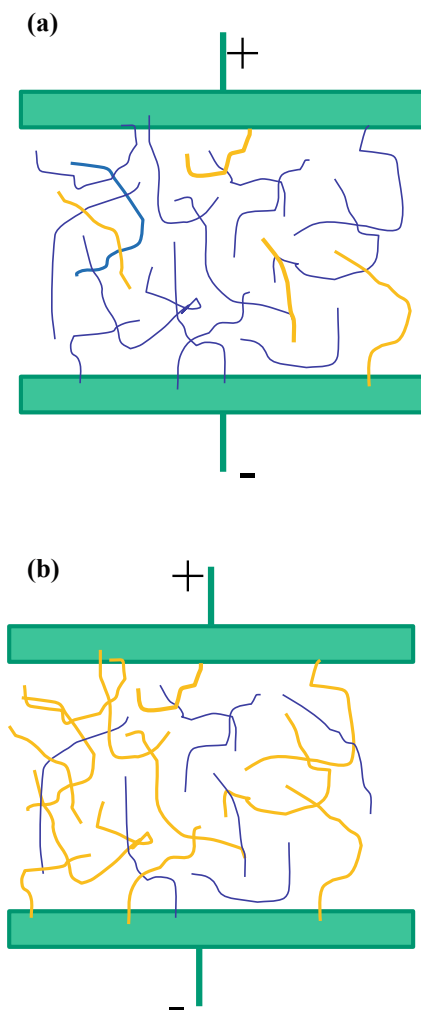


Figure 15.7a shows the case of low dopant concentration, when the concentration of highly conducting polymer chains is small. No connected line of highly conducting polymers is connecting the electrodes. So the resistivity is high. In Fig. 15.7b, the dopant concentration is so high that lines of conducting polymer chains connect the electrodes, and the current is high. This happens, depending on polymer structure and chain lengths, at doping concentrations of a few percent.

15.5 Conductivity of Polymers by Charge Carrier Injection

In the previous Sect. 15.4, an intrinsic conductivity has been given to the polymer by introducing permanent charge carriers into the HOMO (holes) or LUMO (electrons) of the polymer. If a voltage is applied, these carriers will deliver a current from electrode to electrode. So an intrinsically conductive material is produced, with conductivities which in some cases come close to the ones of a metal.

There is, however, a completely different approach to providing a charge transport, i.e. a current, in a polymer. The way to induce conductivity in this case is to inject the carriers via the electrodes. The process is shown in Figs. 15.8a, b, c for the case of PEDOT (Poly-3,4-ethylenedioxythiophen), a polymer widely used in organic electronics. Its molecular structure is seen in Fig. 15.8a.

In Fig. 15.8b, a layer of PEDOT is mounted between an Au- and an Al-electrode, the Au-electrode poled as the anode. Holes are injected into PEDOT via the Au-electrode, and electrons via the Al-electrode.

The process is explained in Fig. 15.8c, which schematically presents the HOMO and LUMO of PEDOT and the Fermi energies of the Au- and the Al-electrode.

Au is a noble metal. As such, it has a large work function, resulting in a relatively small energy difference between the Fermi energy $E_F(\text{Au})$ and the HOMO. So electrons can be excited thermally from the HOMO to the Au-electrode, in other words holes are injected from the Au-electrode into the HOMO of PEDOT. This happens at the anode.

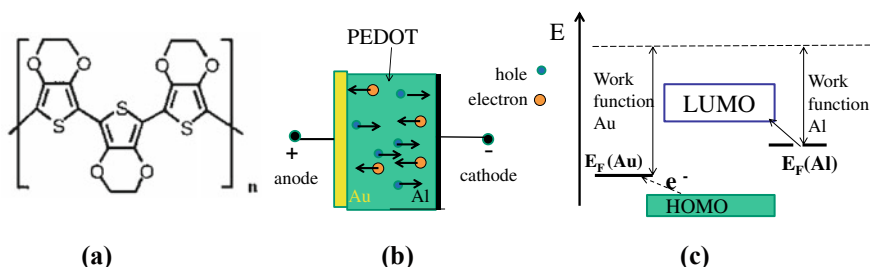


Fig. 15.8 a Molecular structure of PEDOT b Injection of charge carriers into PEDOT c Energies of HOMO and LUMO of PEDOT, and Fermi energies of Au- and Al-electrode

Al is a much less noble metal. As such it has a much smaller work function. So its Fermi energy is fairly close to the LUMO of PEDOT, and electrons can thermally be excited to enter the LUMO from the Al-electrode. This happens at the cathode.

As a result of the injection processes at the anode and the cathode, a current will flow. That current is not based on the presence of intrinsic electrons and holes, but relies on the injection of the carriers. In a way, the situation is similar to the historic vacuum tube, which has been referred to already in Sect. 2.2, Fig. 2.2. Here, the current goes through the vacuum, i.e. the “medium” is a perfect insulator; the charge carriers result from the injection. There is another parallel between vacuum tube and charge transport through a polymer via injection: When the voltage is increased, the current is limited by the build-up of a space charge. The cloud of electrons which builds up in front of the cathode blocks further injection of electrons, correspondingly for the holes at the anode. We are dealing with “space charge limited currents”.

As mentioned above, the cathode has to be a fairly non-noble metal in order to keep the energy barrier between the metal and the LUMO as small as possible. However, even Al is often not very satisfactory in that respect; its work function being still too large relative to the LUMO of many organic materials. One widely used method to lower the work function (i.e. to lower the energy barrier between electrode and LUMO) is to cover the Al-electrode with a subatomic layer of a less noble element like Ca. This method, however, goes along with a severe technical problem. This layer, made of a very non-noble element, is very sensitive to oxygen and humidity. So the method raises the problem of an extremely good encapsulation of the device. The best solution is an extremely thin foil of glass. Glass, however, is an obstacle on the way to flexible electronic devices like LEDs or solar cells, in particular to the application of roll-to-roll sputtering or printing of electronic devices, as is being aimed at in organic electronics.

Very thin glass below 100 μm thickness, however, is flexible enough to be applied in role-to-role printing. So, since a few years, there is considerable progress in the role-to-role fabrication of organic electronic devices on glass substrates.⁴

15.6 Charge Carrier Transport

Once a charge carrier has been injected into the material, it will move across it driven by an electric field. There are various modes of transport in an organic material. A very rare one is the “band transport” as it occurs regularly in the inorganic semiconductors, with freely moving “Bloch electrons” (see Chap. 1). This is only being observed in some very pure crystalline organic materials like naphthalene or perylene. Another one is “polaronic transport”. When the charge carrier moves through the material, at every regular site the interaction with the environment causes a deformation of the environment accompanied by a reduction of the site energy; the carrier is trapped at

⁴ For instance, role-to-role printing of ITO-electrodes on 100 μm and 50 μm glass substrates has been demonstrated by Hiroshi et al. [4], For “ITO-electrode” see Sect. 15.7a.

the site in an energetic pit which it has created by itself. This self-trapping energy has to be overcome by thermal excitation for the next jump in the transport process.

Because the organic materials usually are a highly disordered system, with a high density of states of traps, the **most common transport is the one by hopping from trap to trap**, a process that could be either thermally activated or goes on by tunneling. Even the tunneling would in general be thermally assisted, so the hopping process is characterized by its specific temperature dependence, as generally observed for amorphous materials; see (8.2), (Sect. 8.3).

Due to the fact that charge transport in organic materials is not via band transport as in the traditional semiconductors Si or GaAs, the mobility is much smaller. While the electron mobility is $1500 \text{ cm}^2/\text{Vs}$ in Si and $8500 \text{ cm}^2/\text{Vs}$ in GaAs at room temperature (Table 2.2), it may be as low as $10^{-4} \text{ cm}^2/\text{Vs}$ – $1 \text{ cm}^2/\text{Vs}$ in organic materials. This is certainly one of the major limitations of organic electronics, given the fact that “speed” (i.e. frequency limit) of devices is directly related to mobility (Sect. 2.5).

Nevertheless, considerable progress has been achieved in terms of devices based on organic materials. Two examples will be given referring to the most advanced applications: Organic light-emitting diodes (OLEDs) and solar cells.

15.7 Organic Light-Emitting Diodes (OLEDs)

Figure 15.9a shows the underlying energy level scheme of an OLED, consisting of the polymer PPV (Poly(-p-phenylene-vinylene) structure; see Fig. 15.9b) as a good electron conducting medium, and PEDOT as a good hole conducting medium.

The electrons are injected via an Al-electrode. The latter is covered with a sub-monolayer of Ca to reduce further the work function of the electrode and thus facilitate the injection into the LUMO of the PPV (see Sect. 15.5). The holes are injected from the ITO electrode (see below section “Transparent electrodes”). Electrons and holes meet near the interface of PPV and PEDOT (but rather within the PPV) where they recombine after forming a system of electron and hole, bound by Coulomb

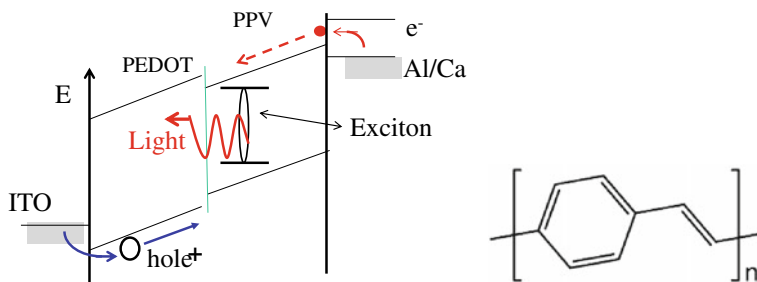


Fig. 15.9 a Energy level scheme of an OLED b Molecular structure of PPV

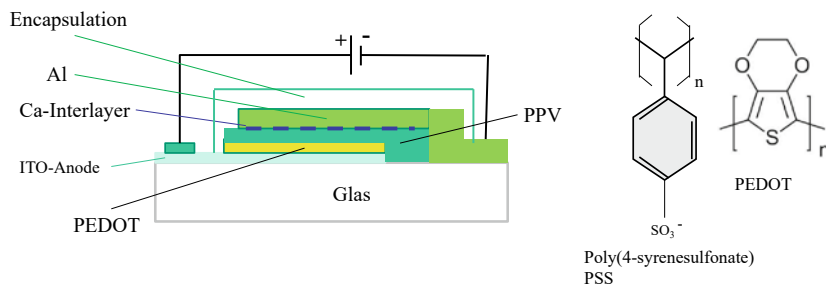


Fig. 15.10 **a** OLED construction scheme **b** Molecular structure of PEDOT:PSS

attraction, i.e. an exciton. Upon recombination, light is emitted with a photon energy corresponding to the binding energy of the exciton.

Figure 15.10a shows the basic construction scheme of such an OLED

The figure shows the relevant elements ITO electrode, PEDOT layer, PPV layer, and the Al-electrode covered with the sub-monolayer of Ca. The whole device is mounted into some encapsulation in order to avoid degradation through incoming humidity. The light leaves the device through the transparent ITO anode.

Some technical aspects

(a) Transparent electrodes

ITO (Indium Tin Oxide)

Organic light-emitting diodes (OLEDs) as the one shown in Fig. 15.10a require a transparent electrode, the anode. Electrodes in electronic devices are usually made of metals. Metals, however, are not transparent in the visible and the near-infrared, because the light excites collective motions of the free electrons (“plasmonic excitations”) giving rise to the opaqueness as well as the high reflectivity of metals. As an alternative, a semiconductor with a sufficiently large bandgap could be used with a high conductivity provided by heavy doping. The commonly used semiconductor used for that purpose is ITO, Indium Tin Oxide.

Indium Tin Oxide consists of ca. 90% In_2O_3 mixed with about 10% (in weight) of SnO_2 , with a high work function making the injection of holes into the LUMO of PEDOT possible.

Sn, a fourth group element, on the In-site acts as a donor, resulting in n-type conductivity. A high Sn-concentration of 10^{20} cm^{-3} to 10^{21} cm^{-3} makes it a “degenerate” semiconductor (see end of Sect. 12.1) with sufficiently high conductivity. ITO, due to its large bandgap, is a transparent electrode in the visible and near UV used in a large variety of devices like flat panel displays, liquid crystal displays, organic solar cells, antistatic coatings, and many others. The high demand for ITO for these applications has contributed to the large increase of the price of In in the last years. Apart from the price, there are other limitations to the use of ITO. One of them is that for many of these applications flexible electrodes are desired, providing many

new possibilities in terms of design for lamps, displays, solar cells, and many more fancy applications. ITO can be deposited on flexible substrates like organic foils, or very thin, lightly flexible glass. There are, however, limitations to this type of flexible electrodes. In case of repeated bending, tiny cracks will develop resulting in a decline in the optical quality of the transparent electrode. Because of these limitations as well as in view of the high costs of In, other types of transparent electrodes are under investigation, none of them, however, so far achieving the conductivity of ITO.

PEDOT:PSS

One promising type of a transparent electrode is the purely polymeric material PEDOT:PSS. PEDOT has already been mentioned in the context of OLEDs as a hole conducting polymer; see figs. 15.8, 15.9, and 15.10a. A purely transparent film of good conductivity can be achieved by a complex of the two polymers PEDOT and PSS (see Fig. 15.10b) usually quoted as PEDOT:PSS.

The conductivity of that complex results from the transport of holes along the carbon chain of PEDOT. As explained above in Sect. 15.4, these holes are produced by taking out electrons from the primarily full HOMO of the polymer PEDOT by an oxidizing agent which in this case is the PSS which gets negatively charged. The PSS at the same time makes the PEDOT chains dispersible in water, thus making it possible to fabricate PEDOT:PSS films out of the liquid phase, as required in printable electronics.

Silver nanowires

Transparent flexible electrodes can be produced by composites of Ag nanowires with some polymer. The high conductivity of the Ag nanowires can provide high conductivity; the proper choice of the polymer and a well-chosen mixing ratio will assure good transparency.

Carbon nanotubes (CNT)

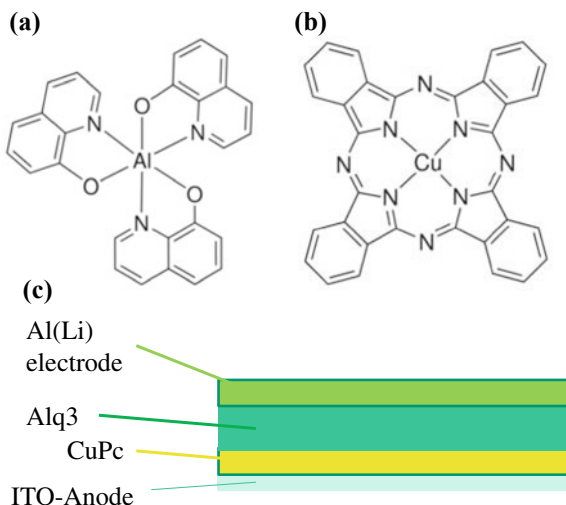
Instead of silver nanowires, also single wall carbon nanotubes (see Sect. 16.3) inserted in some polymer film have been used for replacing ITO in OLEDs. Evidently for this application, metallic nanotubes (as opposed to semiconducting ones) have selectively been used (for details, see Sect. 16.3). Such films of nanowires in a polymer film will, of course, show a certain roughness which can be smoothed, e.g. by combining it with PEDOT:PSS. It should be mentioned in this context that another carbon allotrope, graphene, is also a candidate for a transparent electrode material.

Finally, it should be stated that the development of a really satisfying transparent flexible electrode is still a major challenge to materials science.

Small molecule OLEDs

In Sect. 15.3, conjugated polymers were introduced as basic materials for organic electronic devices. There is, however, an alternative technology for organic electronic devices, based not on polymers, but on small organic molecules. As an example, a

Fig. 15.11 **a** Structure of Alq3, a small molecule **b** CuPc, an established small molecule “hole transporter” **c** Simplified scheme of a small molecule OLED



“small molecule” widely used for OLEDs and solar cells is shown in Fig. 15.11a, shortly named Alq3.

As the figure shows, it is a “ π -conjugated” molecule with alternating single- and double bonds in the rings. The overlapping π -electrons give rise to the conductivity of Alq3 films, which turn out to be a good electron transporting material.

By combining such a good electron transporting material like Alq3 with a good hole-transporting material like CuPc, for example, organic light-emitting diodes can be produced; see Fig. 15.11b. c.

Holes are injected into the HOMO of the hole-transporting material, electrons into the LUMO of the electron transporting material. Where they meet, they form excitons resulting in the emission of light by subsequent recombination. This is the basic principle of OLEDs, whether they are based on polymers or small molecule materials.

A point to be underlined in the context of organic LEDs is the fact that the light-emitting process is not the recombination between free charge carriers, or carriers weakly bound in shallow donors or acceptors, as in the case of inorganic LEDs (see Sect. 11.1 and Fig. 11.1b), but, as stated above, recombination within excitons.

A basic difference between inorganic and organic electronics, which plays an important role not only in the context of OLEDs, results from the property that Coulomb forces are shielded much stronger in inorganic materials as compared to organic materials. This manifests itself in the fact that the permittivity ϵ_r in an inorganic material typically is much larger than in an organic one (example $\epsilon_r = 11.9$ in Si and even 16 in Ge, but only $\epsilon_r = 2.3$ in, e.g. Polyethylene).

One important consequence of this fact concerns “excitons”. It has been mentioned in Sect. 10.3 that the binding energy of excitons in Si (as well as in other inorganic semiconductors) is only a few meV, comparable to or smaller than thermal energies at room temperatures, as a consequence of the large permittivity. So excitons in Si

dissociate at room temperature. The radius of this hydrogen-like system is of the order of a few lattice constant (for the Bohr radius of a hydrogen-like system in Si, see Sect. 3.3). **The binding energy of excitons in organic materials, due to the weak shielding, is much larger (typically 0.3–0.5 eV); they do not dissociate at room temperature and their radius may be of the order of the size of a molecule.** Because of that strong localization these excitons may be bound to a single molecule in their host material. Excitons of the first kind (like the ones in Si) are called Wannier-Mott excitons, and the ones of the second kind Frenkel excitons. The different nature of the excitons strongly affects the behavior of various electronic devices. What diffuses and recombines in inorganic semiconductor devices like LEDs and solar cells are essentially free electrons and holes, while in organic LEDs and solar cells one is dealing with the bound “quasi-particles”, the Frenkel excitons. The most relevant consequence of this fact is that in OLEDs, recombination occurs within Frenkel excitons. A consequence of this is the famous, or rather “infamous”, singlet–triplet problem.

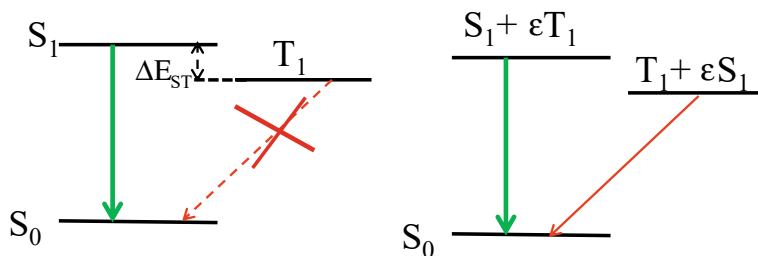
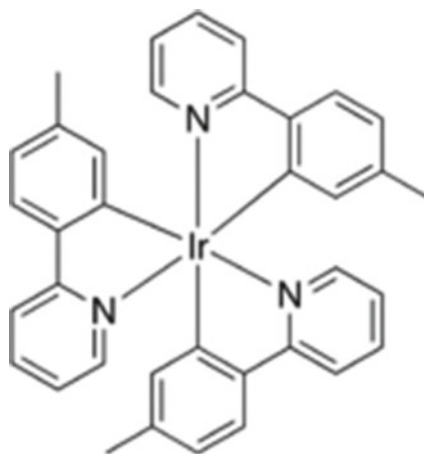


Fig. 15.12 **a** Level scheme of “pure” singlet–triplet system **b** Level scheme of mixed singlet–triplet system

Fig. 15.13 The phosphorescent emitter $\text{Ir}(\text{mppy})_3$



Physics comment: The singlet–triplet problem in OLEDs

The electrons and holes forming the Frenkel exciton are spin $\frac{1}{2}$ “quasi-particles”. The term “quasi-particles” refers to the fact that we are not dealing with free electrons and holes, but units consisting of the free particle plus the molecular environment deformed by the Coulomb forces; see, for example, the discussion of “polarons” in Sect. 15.4). The two spins of $\frac{1}{2}$ each may couple either to a state with total spin with quantum number $S = (\frac{1}{2} - \frac{1}{2}) = 0$, called a singlet state, or a state with total spin quantum number $S = (\frac{1}{2} + \frac{1}{2}) = 1$, called a triplet state. The notion “triplet state” refers to the fact that a state with spin quantum number $S = 1$ shows a threefold degeneracy related to the three magnetic quantum numbers $m_s = 0, \pm 1$. The corresponding level scheme is shown in Fig. 15.12a.⁵

When the electron and the hole come together to form the Frenkel exciton, they may end up either in the S_1 state, the lowest excited state of the singlet system, before they recombine by emission of a photon, or in the lowest triplet state T_1 . The important point to consider now is **that optical transitions between triplet states and singlet states are forbidden** due to symmetry reasons. So all the electrons and holes that end up in a triplet state in the formation process of the exciton are lost for the production of photons; they will somehow go to the lowest state S_0 by radiationless transitions (normally by phonon emission). Because a triplet state with total spin 1 can be either in a substate with magnetic quantum number $m_s = 1$, or $m_s = 0$, or $m_s = -1$, while a singlet state exists only in a state with magnetic quantum number $m_s = 0$, the statistical weight of a triplet state is three times the one of a singlet state, in other words: Three-quarters of the excitons will be formed in the triplet state and thus will be lost for production of light, while only one quarter will end up in a singlet state, followed by emission of a photon. It is a very important issue in the physics and technology of OLEDs to overcome this fundamental loss mechanism.

One solution applicable in the case of small molecules is to “dope” the system with molecules which contain a very heavy atom, i.e. an atom with a high atomic number, acting as an emitter after combining with the exciton. For reasons explained in a moment, in such a system the singlet and triplet states will be mixed, i.e. the singlet state has an admixture of a triplet state, and the triplet state has an admixture of a singlet state. This is visualized in Fig. 15.12b by the notation $S_1 + \epsilon T_1$, replacing the former lowest excited singlet state S_1 , and the notation $T_1 + \epsilon S_1$.

In the following, the state $S_1 + \epsilon T_1$ will be called the “mixed singlet state”, the state $T_1 + \epsilon S_1$ the “mixed triplet state”, as compared to the “pure” singlet and triplet states shown in Fig. 15.12a.

It is seen in Fig. 15.12b that there are two optical transitions now, one with higher intensity and shorter lifetime, because it implies a large contribution

from the allowed $S_1 \rightarrow S_0$ transition, marked in green, and one with lower intensity and longer lifetime, because it is only based on the small contribution ϵS_1 to the excited state, marked in red.

An emitter which shows this latter emission, i.e. an emission out of the “mixed triplet state”, thanks to its admixture of some singlet state, is called a “phosphorescent emitter”. So a phosphorescent emitter would be one that harvests a considerable fraction of the 75% triplet excitons and thus can provide a much higher quantum efficiency as compared to a pure fluorescent emitter.

A well-known phosphorescent emitter is the Iridium(III) shown in Fig. 15.13, with the characteristic heavy atom in its center.

Apart from neutral Ir(III) complexes of that kind, also Pt(II) complexes can provide efficient phosphorescent emitters.

The role of the heavy atom is often described by saying that “spin–orbit coupling” is responsible for this increase in radiation efficiency. This is true, but has to be explained: In atoms with a low atomic number the orbital momenta of the electrons add up to a total orbital momentum L , and the spins add up to a total spin S . In atoms with a high atomic number this is no longer the case, L and S are no longer “good quantum numbers”. This is due to spin–orbit coupling, which means the orbital momentum l_i and the spin s_i of each electron first add up to an angular momentum j_i , before all the angular momenta j_i of the electrons add up to the total J :

As a Frenkel exciton the exciton is so strongly localized, as mentioned before in the context of “small molecule OLEDs” that it may be bound to a single molecule. So, if the exciton is bound to a “heavy” molecule like the one shown in Fig. 15.13 it assumes a different coupling scheme of the spins. Associated with the heavy molecule, electron and hole of the exciton, due to spin orbit coupling, no longer form either a pure $S=1$ state or a pure $S=0$ state, but mixtures of those as in Fig. 15.12. Spin–orbit coupling in this way produces some “intersystem crossing” such that the problem of 75% loss due to the formation of the triplet excitons can be reduced.

The notation “phosphorescent emitter” is, by the way, based on a misunderstanding. Generally, we speak of “phosphorescence” when an excited electron is trapped in some electronic trap and emits only after being released into an emission state, usually by thermal excitation. This results generally in a long optical decay time. Probably, the longer emission time is the reason why emission out of the mixed triplet state is called “phosphorescence” in the context of organic LEDs. But not every slow emission is a “phosphorescent” emission!

Phosphorescent emitters have a higher radiative efficiency as compared to the fluorescent ones, because part of the triplet state excitons is “harvested” for the light emission. For the development of highly efficient OLEDs for TV and large area

⁵ (The physics of systems of two spin $\frac{1}{2}$ particles is often treated in the textbooks on atomic physics in the context of the He-atom with its 2 electrons).

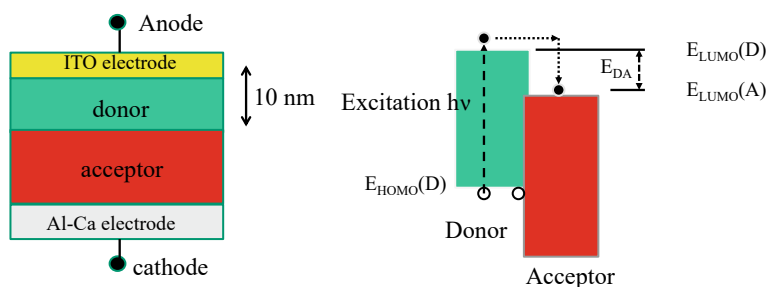


Fig. 15.14a,b: Bilayer organic solar cell b charge separation at the interface between donor- and acceptor layers

screens, efficient emitters are required in the three colors red, green, and blue. There are good, efficient, and stable phosphorescent emitters now in the red and the green, while the situation is much less favorable in the “deep blue”.

Here, some progress has been achieved by the coming up of “thermally activated delayed fluorescent” (TADF) materials.⁶ Those are materials where the separation ΔE_{ST} between the S_1 and the T_1 state is so small that thermal excitation from T_1 to S_1 is possible (see Fig. 15.12a).

Nevertheless, as far as application is concerned, optoelectronics is the most advanced branch of organic electronics. Large area TV screens and displays based on OLEDs are on the market, being considered as “high-end” products. Solar cells might be the next mass market for organic electronics.

15.8 Organic Solar Cells

A device close to market maturity could be the organic solar cell.

An organic solar cell could in principle be built in the same way as an inorganic one (see Fig. 10.5a, b), consisting of two electrodes, an n-type and a p-type layer, as shown in Fig. 15.14a.

The light, entering through the transparent ITO electrode, will excite electrons, which immediately form Frenkel excitons with the remaining hole. As stated above (Sect. 15.7), those excitons are stable and will diffuse toward the donor–acceptor interface where they are split into electrons and holes (or, to be precise, negative and positive polarons; see Sect. 15.4).

The polarity of the organic solar cell, like the one of an inorganic solar cell, is the one of a pn-junction without external voltage. So from the argument given at the end of Sect. 9.1, the “acceptor side” is the cathode, the donor side the anode.

The charge separation mechanism in organic solar cells and its consequences

⁶ Michael Yin Wong et al.: [5].

As stated in the previous section in the context of the singlet–triplet problem, excitons in organic materials are Frenkel excitons and as such more strongly bound than the ones in inorganic semiconductors. As a consequence, they do not split thermally at room temperatures or the common temperatures of a solar cell. But in a solar cell, they have to split and be separated into free charge carriers in order to provide the photocurrent. So the question is: How and where does this happen?

As already mentioned, it happens at the interface between the n-type and the p-type layer, i.e. donor- and the acceptor material. Fig. 15.14b shows this in an energy band representation.

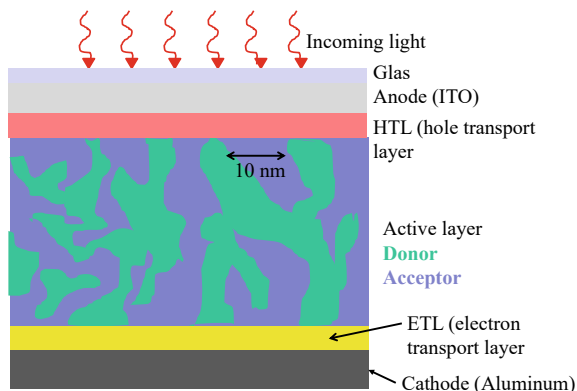
The basic difference between an inorganic solar cell and an organic one, relevant for the process of charge separation, is that at the interface between the donor and acceptor layers two different materials meet (called materials D and A, respectively, in the following), while in the inorganic solar cell the same materials meet, just being different dopes. Let us consider the case where, as is generally true, the polarons are produced by light absorption in the donor material D, i.e. D is the “active” material. After optical excitation, i.e. after providing an energy $h\nu$ to the system, we have an n-polaron in the LUMO of D, and a hole-polaron in the HOMO of D. After separation, the hole is still on the D-side, the electron (i.e. the n-polaron) on the A-side. So it is easily seen from the figure that the energy required for the separation of the Frenkel exciton is provided by the energy difference between the LUMO-energy in material A and the one in material B. So the energy difference $E_{\text{LUMO}}(\text{D}) - E_{\text{LUMO}}(\text{A}) = E_{\text{DA}}$ must be equal to or larger than the binding energy of the Frenkel exciton in the active material D. Because these binding energies are in the region of 0,3 eV upwards, this kind of difference between the LUMOs of the two materials is required.

While in an inorganic solar cell the question of whether a solar photon can produce free charges, which then provide the photocurrent, is just a question of the bandgap ($h\nu > E_{\text{G}}!$), in organic solar cells an additional condition comes into play: The one of the right Lumen-Lumen energy difference E_{DA} , making a splitting of the exciton into free particles possible. **So the question of the optimum material choice for a solar cell is not just one of a proper bandgap, as answered by Fig. 10.7a, but one of the right material combination!** For this reason, there is no such thing as the “Shockley-Queisser limit” of an organic solar cell. Only the right combination will provide an optimum of efficiency.

The diffusion length of the Frenkel excitons is only of the order of 10 nm, so, in order to let the excitons reach the interface, the active layer should have this thickness. The problem is that for a quantitative absorption of the sunlight, the thickness should rather be 100 nm. A solution specifically offered by polymer materials is the one of the widely studied “bulk heterojunction solar cell (BHJ)” (Fig. 15.15).

In a BHJ cell, the two polymers (donor and acceptor) are first deposited on the glass/ITO electrode, e.g. by spin coating. After that, a phase separation between the two polymers is initiated which results in a system of interpenetrating domains with diameters of the order 10 nm (see, for example, the process of spinodal segregation, a process resulting in phase separation). In this way, a diffusion path of only this length is required for the excitons to reach the donor–acceptor interface, while the thickness of the whole system can be sufficiently large for quantitative absorption of

Fig. 15.15 Bulk heterojunction organic solar cell



the sunlight. At the interface, the excitons are split. Of course, as the figure shows, the domain structure has to be such that a percolation takes place, so that the positive and negative polarons find a path going all the way through to the electrodes.

Cathode and anode of such a BHJ solar cell are defined by the electrode material. Electrons will leave at the Al-Ca electrode, holes at the ITO electrode (see Fig. 15.15), so the latter one is the anode.

Organic solar cells are approaching efficiencies of 20% in so-called tandem configurations. It should be pointed out once again, however, that “efficiency is not everything in photovoltaics”. Costs and possible design, like flexible devices, are an issue as well. One idea, for example, is to cover glass facades of large buildings and skyscrapers with semitransparent solar cell foils, protecting, on the one hand, the interior from solar radiation, producing, on the other hand, solar energy.⁷

Let us remark in general that a fascinating feature of organic electronic in general is the possibility for structuring the devices and circuits by printing processes. For that reason, fabrication studies focus on the fabrication from the liquid phase instead of the evaporation processes preferred in the early development of organic electronics.

The field of organic electronics is not the focus of this book, only an outlook has been given. In this outlook, it should finally be mentioned that many other applications of organic electronic materials are under study, like field effect transistors and all kinds of sensors, in particular bio-compatible sensors for medical applications.

15.9 Questions and Answers

Questions

⁷ An example: On Thursday, April 29, 2021, the German Newspaper Frankfurter Allgemeine Zeitung reported on page 19: “Windows develop towards powerplants. The Heliatek Company will produce on a large scale organic solar foils”.

- Question 1: What type of chemical bonds plays a crucial role in organic electronic materials?
- Question 2: Which special feature of conjugated double bonds is responsible for their special role in organic electronics?
- Question 3: How can a conductivity along polymer chains be achieved?
- Question 4: How can free carriers be brought into an otherwise insulating organic material?
- Question 5: Which band in organic materials corresponds to the conduction band, which one to the valence band?
- Question 6: Give examples for transparent electrodes, as required for OLEDs and organic solar cells.
- Question 7: What is a Wannier-Mott exciton, and what is a Frenkel exciton?
- Question 8: How large is the diffusion length of the exciton in an organic semiconductor?
- Question 9: What is the structure of a bulk heterojunction (BHJ-) solar cell?
- Question 10: Which efficiencies are achieved in organic solar cells? What are the problems?

Answers

- Answer 1: Conjugated double bonds.
- Answer 2: Their spatial extension resulting in an overlap which gives them a delocalized character.
- Answer 3: By extracting an electron out of the chain via an acceptor like iodine, or by injecting an electron via a donor, e.g. an alkali metal.
- Answer 4: By injection via a suitable electrode: Via a sufficiently noble metal (i.e. a high work function material) for injection of holes into the HOMO, or via a low work function material (like an alkali metal) for injection of electrons into the LUMO.
- Answer 5: Conduction band: The LUMO, i.e. the lowest unoccupied molecular orbital. Valence band: The HOMO, i.e. the highest occupied molecular orbital.
- Answer 6: ITO (the most widely used one), or a composite of some polymer with Ag nanowires or Carbon nanotubes.
- Answer 7: Excitons are bound systems of a positive and a negative elementary charge. In a Wannier-Mott exciton, the Coulomb field between them is strongly shielded, resulting in a small binding energy and large separation of the two charges, a separation which includes many atoms or molecules of the host lattice. In the Frenkel exciton, the Coulomb field is only weakly shielded, resulting in a large binding energy and a small separation of positive and negative charges. They are typically sitting on the same molecule.
- Answer 8: About 10 nm.

- Answer 9: In a BHJ solar cell, donor and acceptor polymers form a system of interpenetrating domains with dimensions of about 10 nm, such that the diffusion length of the exciton formed by absorption of light is larger than or comparable to the distance of the next interface.
- Answer 10: About 20%. Problems concern chemical as well as radiation stability.

References

1. W.P. Su, J.R. Schrieffer, A.J. Heeger, Solitons in polyacetylene. *Phys. Rev. Lett.* **42**, 1698 (1979)
2. H. Baessler, A. Koehler, Charge transport in organic semiconductors. *Top. Curr. Chem.* **312**, 1–66 (2012). https://doi.org/10.1007/128_2011_218
3. H. Shirakawa, E.J. Louis, A.G. MacDiarmid, C.K. Chiang, A.J. Heeger, Synthesis of electrically conducting organic polymers: Halogen derivatives of polyacetylene, (CH) *x*. *J. Chem. Soc. Chem. Commun.* **16**, 578 (1977)
4. H. Tamagaki, Y. Ikari, N. Ohba, Role-to-role sputter deposition on flexible glass substrates. *Febr. Surf. Coat. Technol.* **241**, 138–141 (2014). <https://doi.org/10.1016/j.surfcoat.2013.10.056>
5. Michael Yin Wong et al.: Deep blue oxadirole-containing thermally activated delayed fluorescence emitters for Organic light emitting diodes, <https://doi.org/10.1021/acsami.8b.8b11136>, 07.Sept. (2018).

Chapter 16

Towards Molecular Electronics



Abstract In this chapter, a new approach to miniaturization of electronic devices and circuits is described. While the present technology has proceeded to smaller and smaller dimensions along the roadmap set by Moore's law and is based essentially on the same technology, with photolithography as the basic fabrication element and MOSFET and capacitor as the central devices for data storage and switching, and with Moore's law coming to an end, completely new concepts are required. "Molecular electronics" opens new paths in that direction, starting from nano-sized and molecular building blocks with appropriate electronic properties and functions, and combining those with electronic devices and circuits with completely new fabrication concepts. While the "conventional" approach often is described as "top-down development", the latter could rather be called a "bottom-up development", which actually is the technical term used in this context. Some of those nano-sized building blocks will be presented in this chapter like fullerenes, carbon nanotubes, graphene, and single molecule interconnects. First applications are in sight. A common feature of these building blocks is that they are based on carbon allotropes and carbon chemistry. This points toward an evolution of a future "carbon-electronics" complementary to present Si-technology.

16.1 From "Top-Down" to "Bottom-Up"

In Chap. 14, the development toward smaller and smaller structures of electronic devices was considered. It is evident that this development cannot go on for ever and ever. At present, the present technology node 7 nm is achieved, test chips down to even 2 nm structure size have been presented (Sects. 14.5 and 14.7), so, by simple extrapolation of Moore's law atomic dimensions of some nanometers will be reached in a few years. This will certainly require a completely new technology. So far, miniaturization proceeded in the way that essentially the same technology (centered around photolithography) advanced to a smaller and smaller dimension. This procedure is often described as "top-down development". Another way toward devices of atomic dimensions would be to go the opposite way, i.e. to start from components of atomic and molecular dimensions and combine those to electronic devices and circuits, a

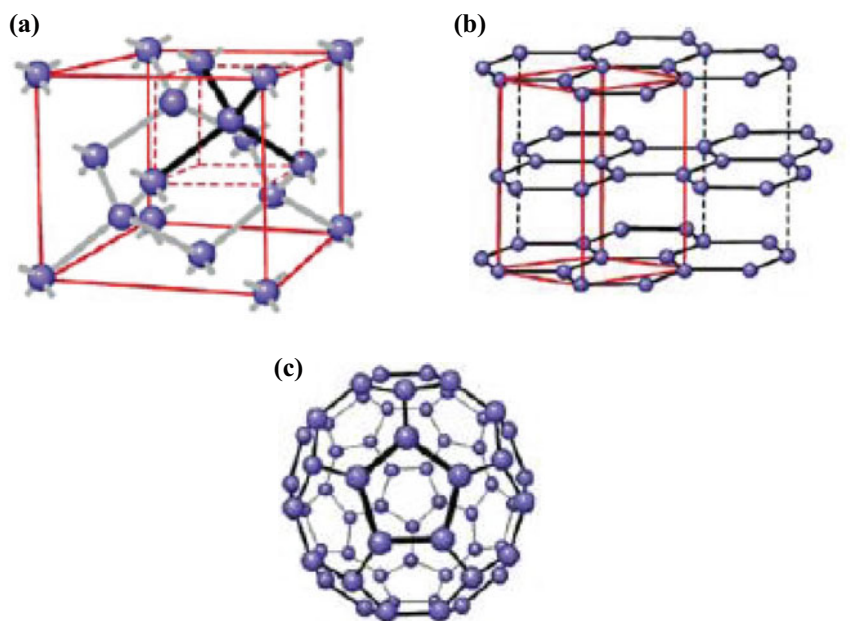


Fig. 16.1 (a) Carbon allotropes: diamond structure (b) Carbon allotropes: graphite structure (c) Carbon allotropes: fullerene C 60

procedure which then could be called “bottom-up development”. Some steps in that direction will be described in this chapter.

16.2 Carbon Allotropes, Fullerenes

Allotropes are structural modifications of an element. Two allotropes of the element carbon have been known for a long time: Diamond and graphite; see Fig. 16.1a, b.

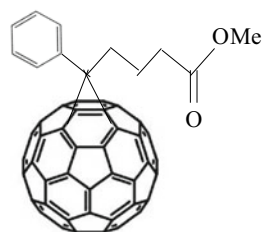
More recently, quite a number of additional allotropes have been discovered like the fullerenes.

Fullerenes are cage-like closed carbon structures. Figure 16.1c shows one of them, the “Buckminster Fullerene” C 60 consisting of 60 carbon atoms.

C 60 has been named by the discoverers⁽¹⁾ after the architect Richard Buckminster Fuller, who constructed domes of that shape. A C 60 molecule, as shown in Fig. 16.1c, consists of 20 hexagons and 12 pentagons, like a real soccer ball. The diameter of a C 60 molecule is about 0,72 nm. There exists a number of other fullerenes like C 70, C 72, C 76, C 80, and many others. C 60 fullerenes are, e.g. used in organic electronics because of their hole-transporting and electron-blocking properties.

¹ Kroto et al.(1).

Fig. 16.2 Fullerene as constituent of an n-semiconducting material



Many fullerene derivatives are used as n-type semiconductors. Figure 16.2, for illustration of this point, just shows an example among many others. The n-type conductivity in this case is provided by the side-chain, not the fullerene.

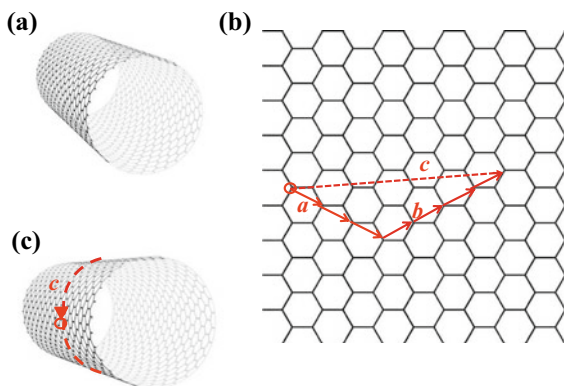
16.3 Carbon Nanotubes

Another carbon allotrope discussed in the following are the “carbon nanotubes” (CNT), shown in Fig. 16.3a being of great interest in the context of nanoelectronics.

Carbon nanotubes can be considered as the result of rolling up just one layer of graphite (or, what means the same, a sheet of “graphene”; see Sect. 16.4). Carbon nanotubes typically have a diameter between 1 and 10 nm; they may be as long as some tenths of μm , even close to a mm. An interesting point in view of electronic applications is that nanotubes may be metallic or semiconducting; see, for discussion of this point, Fig. 16.3b, c.

If we consider the nanotube as the result of rolling up a film of graphene, the different types of nanotubes can be described and distinguished by giving the axis c along which the graphene has been rolled up:

Fig. 16.3 **a** Carbon allotropes, single wall carbon nanotube **b** Definition of the roll-up axis c (indices “n, m”) of a nanotube **c** Nanotube from rolling up a graphene plain along axis c



$$\mathbf{c} = n\mathbf{a} + m\mathbf{b} \quad (16.1)$$

where $\mathbf{a}$ and $\mathbf{b}$ are the unit vectors in the hexagonal two-dimensional lattice as shown in Fig. 16.3b. This axis is defined by the two integer numbers n and m . So the axis given in the figure would be the one with $n = 3$, and $m = 4$. It can be shown that a nanotube is metallic if $n = m$, or if $m - n$ can be divided by 3, otherwise the nanotube is semiconducting. It follows from this fact that, if the nanotubes are produced “statistically”, 1/3 will be metallic and 2/3 will be semiconducting.

Side remark: To make this latter point clear, let us state that this rule amounts to saying that the nanotube is metallic if $|n - m| = 3i$, i being an integer, and it is semiconducting if $|n - m| = 3i \pm 1$. Evidently, because of the ± 1 , there are twice as many combinations (n, m) resulting in a metal as compared to combinations resulting in a semiconducting nanotube.

In Sect. 13.4, a two-dimensional electron gas has been discussed. It was shown that in a situation where one dimension is sufficiently small (i.e. smaller than the de Broglie wavelength), the energy levels related to that dimension are so widely separated that the higher ones are not occupied by electrons under normal thermal conditions. In that sense, the system is “two-dimensional”. The carbon nanotube, now, is a system which in two dimensions (i.e. the circumference) has molecular dimensions, while in the third dimension (the length) it can be considered “macroscopic”. So it is plausible that the energy levels related to the transverse dimensions are not accessible under realistic thermal conditions, but the ones related to longitudinal modes of the electronic waves are. In that sense, nanotubes can be treated as one-dimensional electronic systems.

Physics comment. The Carbon nanotube as a one-dimensional system

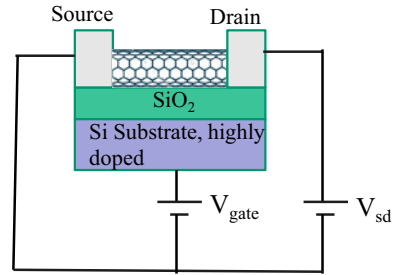
Only a short comment will be given here concerning the statement given above, that, depending on n and m , the nanotube may be either metallic or semiconducting. When, as explained, the nanotube can be considered as the result of rolling up a graphene layer along a certain axis $\mathbf{c}$, which is defined by the indices m and n (Fig. 16.3c), it is evident that the resulting nanotube will show a certain periodicity in $\mathbf{c}$ -direction (the direction of its length). So this periodicity, i.e. the unit cell of this one-dimensional system, is determined by n and m .

In summary: n and m decide on the periodicity, and the periodicity in turn decides on whether all bands are full or empty (semiconductor) or whether the highest occupied band is partly full (metal); see as an example the discussion of Peierls’ instability in Sect. 15.3.

Applications of carbon nanotubes

Metallic nanotubes may be used as **interconnects** in microelectronics, metallic or semiconducting ones as components of molecular-sized electronic devices.

Fig. 16.4 Scheme of carbon nanotube FET, after³



As long as the scattering length of electrons in the one-dimensional metal is long compared to the length of the nanotube, it can be shown, as a consequence of the quantum mechanical nature of the system, that the conductance G of the metallic nanotubes is independent of their length and is given by

$$G = 2 G_0 = 4 e^2/h \quad (16.2)$$

where $G_0 = 2e^2/h$ is the so-called “quantum conductance”. From (16.2), one obtains a maximum conductance of $1.55 \cdot 10^{-4}$ S, or a minimum resistance of $R = 1/G = 6.5$ k Ω for a metallic carbon nanotube.

Let us compare this with a copper wire of dimensions comparable to a nanotube, e.g. a diameter of 10 nm and a length of 1 μ m. Our motivation for this question is the idea of whether a carbon nanotube might replace a copper wire as an interconnect in a nanoelectronic device.

Copper has a resistivity of $1.7 \cdot 10^{-8}$ Ω m. If inserted into the expression $R = \rho l / \pi r^2$ for the resistance of a wire of radius r and length l , this results in $R = 220$ Ω , to be compared with 6.5 k Ω . So at first sight, the copper nanowire seems a much better interconnect. This approach however is incorrect. One has to take into account that the resistivity is strongly increased by scattering at the surface when the wire’s diameter gets down to some ten nanometers or lower. Copper wires that thin are not available, but calculations taking into account these scattering effects² show that copper wires of approximately 10 nm diameter and 2.5 μ m length would have comparable resistance as single wall nanotubes of that dimensions. **So carbon nanotubes might be conceived as interconnects in a “bottom-up” context of microelectronics replacing Cu nanowires.**

The use of carbon nanotubes as a **channel of a MOSFET** (“Carbon Nanotube Field effect Transistor—CNFET”) has also been demonstrated. An example is shown in Fig. 16.4.

A nanotube connects two electrodes acting as source and drain. The dielectric is SiO₂, like in the traditional MOSFET. The gate electrode is the highly doped Si electrode (a material often being used as an electrode instead of a metal, permitting

² Kreupl et al.(2).

³ Bagavathi et al.(3)

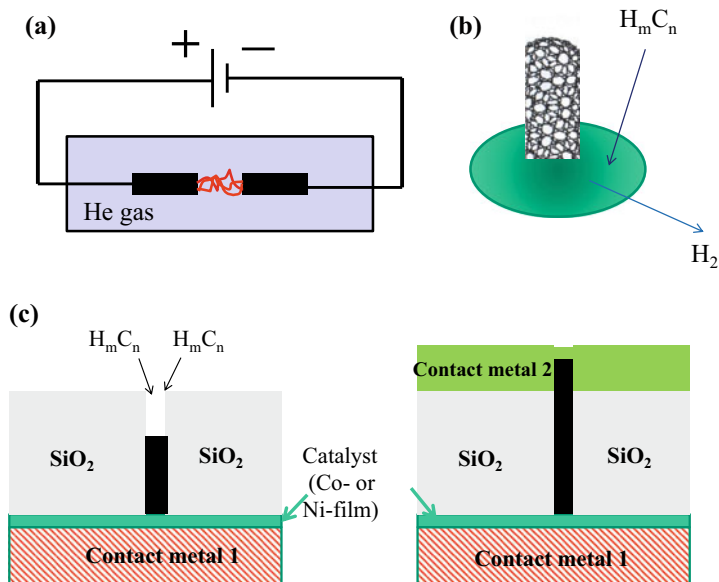


Fig. 16.5 **a** Production of nanotubes in an arc discharge **b** Carbon nanotube growing out of a Ni (or Co) droplet **c** Bottom-up growth of a CNT interconnect

an extended use of Si-technology). So the technical term for the device shown in Fig. 16.4 is not MOSFET, but—appropriately—CNTFET (carbon nanotube field effect transistor). The performance, however, is comparable to the one of a MOSFET.

A pioneering example is given in⁴ where a carbon nanotube was placed randomly from a solution of carbon nanotubes on the two platinum electrodes forming source and drain (as in the scheme given in Fig. 16.4).

It has been mentioned in Sect. 15.7 that a network of metallic carbon nanotubes in a polymer composite is a promising candidate for transparent electrodes in OLEDs.

Quite a number of methods are existing to separate metallic and semiconducting single wall nanotubes when they have come as a mixture in a solution. This can either be done based on different electronic properties of the two kinds of particles, like the different electronic polarizability, or based on a different “chemistry” of both.

Production of Carbon Nanotubes

Figure 16.5a and b presents two “classical” ways to produce nanotubes. In Fig. 16a, an arc discharge between two graphite electrodes is shown. Carbon particles as well as a mixture of multiwall carbon nanotubes will form on the anode. If Ni or Co is dissolved in the graphite, it will catalyze the formation of single wall nanotubes on the anode.

⁴ Verschuieren et al. (4).

In Fig. 16.5b, a Ni or Co droplet on a (heated) substrate is exposed to some hydrocarbon.

Carbon is dissolved in the Ni or Co until an oversaturation is achieved and carbon is segregating at the surface in the form of nanotubes. In the cap of the growing nanotube, some Ni- or Co- “pollution” is dissolved resulting in further growth of the nanotube.

Figure 16.5c demonstrates how a single wall nanotube can be grown at a defined spot on a substrate in order to form, e.g. a CNT interconnect connecting two metallic electrodes.

The electrode (contact metal 1) is covered with a thin film of the catalyst Ni or Co. The SiO_2 on top is opened by etching at the spot where the CNT has to be formed (left side of the figure). Subsequently, the open channel is infiltrated by the hydrocarbon H_mC_n . By contact with the catalyst, a CNT grows according to the scheme of Fig. 16.5b. Finally (Fig. 16.5c right side), the contact metal 2 is deposited on top. So an interconnect between the two electrodes is formed in a “bottom-up” approach.

16.4 Graphene

A carbon allotrope which is attracting much attention at present is graphene. Graphene is a macromolecule consisting of a monolayer of carbon rings (Fig. 16.6), or, to put it differently, just one layer of graphite.

So graphene can be described as a two-dimensional material, which, indeed, shows interesting properties. The chemical bonding situation has already been described in the context of graphite; see Sect. 15.2. Three of the 4 valence electrons of the carbon atom hybridize into 3 sp^2 orbitals in the plain, forming the ring structure; the fourth, the π -orbital, sticks vertically out of the plain (see Fig. 15.2 for graphite). Because of the overlap of these π -electrons, a high conductivity is achieved. At the same time,

Fig. 16.6 Graphene, a monolayer of carbon rings

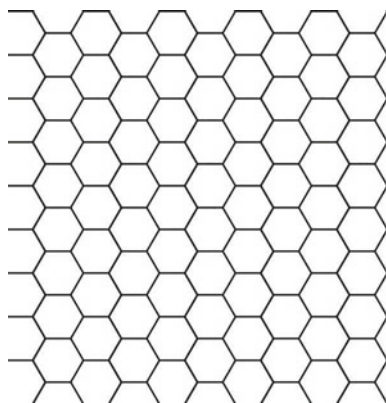
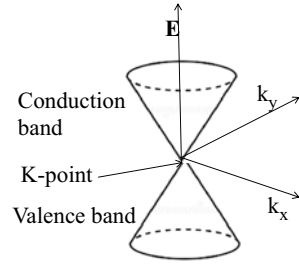


Fig. 16.7 Linear dispersion relation of graphene near the “K-points” where conduction band and valence band touch each other



high mobilities up to 200,000 cm²/Vs have been observed at room temperature⁵ (compared to the one in Si at room temperature of 1500 cm²/Vs; Sect. 2.2, Table 2.2).

Before we turn to some of the possible applications of graphene, we shall discuss two of its extraordinary physical properties.

Physics comment: Band structure of graphene

The most remarkable features of the band structure of graphene are its zero bandgap and linear dispersion relation. At six points of the first Brillouin zone, the so-called K-points, the conduction band and the valence band touch each other; the dispersion relation $E(k)$ around these so called K-points is shown in Fig. 16.7 (in Fig. 16.7, one of the 6 K-points in the first Brillouin zone has been given the coordinates $k = 0$ for simplicity of the description).

The dispersion relation around this point turns out to be a linear one:

$$E = \pm \hbar v_F k \quad (16.3)$$

where

$$v_F = \sqrt{3} * ta/2\hbar \quad (16.4)$$

and t is the “tunneling matrix element” to be calculated in the framework of the tight binding model, and $a = 0.246$ nm is the length of the lattice vector of graphene (Fig. 16.3b). If both quantities are inserted into (16.4), one obtains for v_F (which evidently has the dimension of a velocity; see 16.3) the value

$$v_F \cong c/300 \quad (16.5)$$

⁵ Mozorov et al. (5).

It is interesting to compare this with the situation encountered in relativistic mechanics. Here, the relation between energy and momentum, i.e. the dispersion relation, is given by

$$E^2 = p^2 c^2 + (m_0 c^2)^2 \quad (16.6a)$$

For a massless particle like the photon (or approximately like the neutrino) with $m_0 = 0$, one obtains

$E = \pm pc$ or, by writing $p = \hbar k$:

$$E = \pm \hbar ck \quad (16.6b)$$

By comparing (16.6b) with (16.3), one sees that in a way the electrons in graphene behave like massless relativistic particles having an “effective velocity of light” $v_F \cong c/300$. This comparison can be extended in various respects like by stating that the two “branches” with $+$ and $-$ signs of the energy correspond to particle and antiparticle in the case of massless particles, and they correspond to electrons and holes in the conduction band and valence band of graphene, respectively. Without deepening this topic, let us remark that the analogy can be further extended by stating that electrons in graphene, like massless relativistic particles, can be described by the Dirac-Weyl equation (the Dirac-Weyl equation is the Dirac equation for massless particles). This analogy has attracted much interest in theoretical studies.

Production of graphene

Historically, graphene has first been produced by the method of adhesive tape, for the first time successfully applied by Andre Geim and Konstantin Novoselov⁶ and coworkers. They managed to pull very thin carbon layers from graphite by means of such a tape. By repeating this a few times, they finally managed to obtain single layers which they deposited on SiO_2 . A certain challenge consisted in identifying those flakes which really consisted of just one layer, i.e. of graphene. This could be done optically. The true graphene flakes could be made visible in a microscope by properly adjusting the thickness of the SiO_2 substrate. For their pioneering work on graphene, Novoselov and Geim received the Nobel prize for Physics in 2010. Quite a number of other methods have been developed since then. One consists in epitaxial growth on a metallic substrate⁷ by Chemical Vapor Deposition (CVD). In this process, the carbon source is a gas like methane (CH_4) which by a catalytic process is decomposed, and the carbon is subsequently deposited as a monoatomic layer (i.e. graphene) on a metal surface like Cu or Nb. Another production method, to mention only a few, is based on the thermal decomposition of SiC. At high temperature, there is a tendency

⁶ Novoselov et al. (6).

⁷ Berger et al. (7).

for SiC to “carbonize”. Si escapes from the surface leaving carbon behind. So, by heating 4H-SiC to about 1100°C under low pressure in a very controlled process, a graphene layer can be produced at the surface. The latter two methods have the advantage of providing a wafer-scale technique, not only tiny flakes as other methods do.

Let’s come back to the thermal decomposition technique, which results in a graphene layer on a SiC substrate. It cannot be excluded that the SiC substrate will affect the observed properties. An important step to keep the effect of the substrate on electronic properties small is by starting out from a semi-insulating SiC substrate, thus avoiding any flow of currents in the substrate. (For semi-insulating semiconductors, see Sect. 4.5. SiC can be made semi-insulating by doping with Vanadium which provides deep electronic levels in SiC.)

Applications of graphene

A high potential for important applications is attributed to graphene, being still in a stage of research and development.

Graphene can be considered as an interesting material for transparent electrodes. The conductivity is high, and only about 2% of visible and IR-light are absorbed. Flexible conductive films may be produced which could be desirable for a number of applications.

Graphene has a high potential for biochemical sensor applications. Biomolecules are easily adsorbed to the graphene film. (“Graphene is a surface, nothing else”!) Charge transfer between molecule and graphene film will result in a change of conductivity, which would act as the sensor principle.

Graphene might be used as basic electronic material for nanostructured devices and circuits. Field effect transistors (FETs) have been demonstrated with a graphene film providing the channel.

Bandgap engineering of graphene

In spite of being a unique and fascinating property of graphene, the complete lack of a bandgap constitutes a problem for many electronic applications like the use of graphene as a channel material for field effect transistors. As long as the bandgap is zero, a FET cannot be completely switched off because the unavoidable intrinsic defects will always provide free electrons or holes to the channel, given the fact that no thermal activation is required. One widely used way to open a bandgap is by forming a narrow strip of graphene. The zero bandgap of graphene exists, to be precise, only in samples of infinite dimensions. In a stripe, the bandgap is inversely proportional to the width. In order to achieve a useful bandgap of, e.g. 100 meV, a width as low as 10 nm is required. To achieve this via an etching process in connection with a protective mask is quite a challenge, but can be done for instance by use of electron beam lithography (EBL) for the fabrication of the mask.

Another strategy for opening the bandgap in graphene is by adding certain covalently bound groups to the graphene layer, i.e. by “functionalization” of the graphene. It has been demonstrated that by coupling phenyl groups to the graphene, the on/off ratio of a FET can be considerably increased.

16.5 Single Molecule Interconnects and Anchor Groups

In Sect. 16.3, CNTs were discussed as possible interconnects in electronic devices. The final step of a bottom-up approach to molecular electronics would, however, be the use of single molecules as interconnects in devices. In this way, single molecules would be the building blocks of electronics, corresponding to an ultimate definition of “molecular electronics”.

When considering the use of single molecules as interconnects, one has to keep in mind that the transport of an electron through a single molecule, due to the wave nature of the electron, is a quantum mechanical phenomenon. Quantum mechanics tells that a single molecule would have a minimum resistance of

$$1/G_0 = 1/(2e^2/h) = 12,95 \text{ k}\Omega \quad (16.7a)$$

or a maximum conductance of

$$G_0 = 77\mu\text{S} \quad (16.7b)$$

where G_0 is called the “conductance quantum”^{8,9} In reality, it could be considerably larger due to perturbations of the electronic transport.

(In Sect. 16.3, a maximum conductance for metallic carbon nanotubes $2 G_0$ has been given in 16.2. The factor 2 in comparison to 16.7a, b has to do with the fact, not discussed further here, that the conductance of a quantized one-dimensional system, like a molecule or a nanotube, is given by G_0 times the number of “channels” contributing to the conductance. In metallic nanotubes, two energy bands, i.e. channels, contribute to the charge transport.)

In recent years, a couple of experiments concerning the transport properties of individual molecules have been described. A crucial point in such experiments is the realization of a good and reproducible junction between the electrodes and the molecule. A presentation of the issue is, e.g. in a paper by A. V. Rudnev et al.¹⁰ In the paper, the electrical conductivity of an organic molecule is measured in a scanning tunneling microscope (STM), the tip of the STM forming one of the electrodes (Fig. 16.8).

A well-defined contact between the gold tip of the STM and the organic molecule is provided via the amine group NH_2 as an “anchor group”.

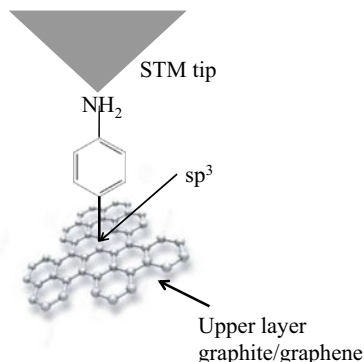
A crucial problem to be investigated in the paper was the one in producing a good and reproducible contact between the graphite electrode and the molecule. It turned out, as expected, that the quality of that contact strongly depends on the type of bonding between the graphite and the organic molecule. If the organic molecule is adsorbed on the graphite only by the van der Waals force, the contact is only poorly

⁸ Agrait et al. (8).

⁹ Cuevas et al. (9).

¹⁰ Rudnev et al. (10).

Fig. 16.8 Measurement of the electrical conductivity of an organic molecule via an STM



defined. If, however, the molecule forms a C–C-bond, i.e. a covalent bond with the graphite, stable and reproducible contacts can be obtained. While the graphite surface, which can essentially be considered as a graphene film, consists of sp^2 -orbitals, the molecule is attached to the surface by a sp^3 -node which provides the desired stability to the contact. In the experiment, the STM is operated in the so-called break-junction mode, a kind of use of the STM which helps to clearly identify the situation where just one molecule is providing the contact between the two electrodes.

16.6 Molecular Switches

The typical switching device in classical electronics is the transistor, e.g. the field effect transistor (Sect. 13.2). The current between the source and drain electrode is switched between on and off via the gate voltage. In the course of advancing miniaturization, however, it gets more and more difficult to make enough space available for three electrodes. One way to deal with this problem would be to find a photochromic molecule that can be optically switched between two states of strongly different conductance. A molecule of that kind is shown in Fig. 16.9.

As the figure shows, the central ring is opened via visible light and closed again by UV light, resulting in two states differing strongly in conductance. Evidently, the effect can be used for switching as well as for information storage, the latter

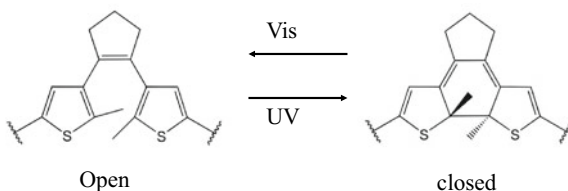


Fig. 16.9 Diarylethene molecule

application made possible by the fact that both states are stable at room temperature¹¹. The idea in the quoted paper is to open or close the junction between two graphene layers, provided by the diarylethene molecule, via a light pulse.

16.7 C-Based Electronics, Complementary to Si-Technology

As a summary this Chapter, we might state that nanoscale electronics is one the ways toward “molecular electronics”, the latter, however, being still being in the laboratory stage. The examples of fullerene- and carbon nanotube-based devices, the achievement of stable contacts on a graphite (or graphene!) electrode based on covalent C–C bonding as well as switchable molecular interconnects show that C-based electronics might offer great perspectives for an electronic technology on molecular scales.

Let us note as a final remark that in this way, the development could follow the path of biological evolution which evidently has preferred carbon chemistry on its path to higher and higher complexity. While biological systems generally are based on highly sophisticated chemical processes, our nervous system, i.e. our “data processing apparatus”, is relying on charge transfer processes, i.e. electrical processes. The “devices” being active in these systems, the nerves and their interconnects are slow and relatively inefficient as compared to modern Si-devices, but evidently a superior “circuit technology” making use of complex three-dimensional structures makes all the difference.

This, of course, is an issue far remote of the reach of this book.

16.8 Questions and Answers

Questions

- Question 1: Name the most widely used carbon allotropes.
- Question 2: Give a few applications of fullerenes.
- Question 3: Are carbon nanotubes metallic, semiconducting, or insulators?
- Question 4: What are the potential applications of CNTs?
- Question 5: What is graphene?
- Question 6: What are the most remarkable features of the band structure of graphene?
- Question 7: How was graphene produced for the first time?
- Question 8: For certain applications of graphene, like a channel of a FET, the zero bandgap has to be opened. How can that be done?

¹¹ Jia et al.(11).

- Question 9: What are the outstanding electronic properties of graphene?
- Question 10: Give some potential applications of graphene.
- Question 11: In most production methods, graphene is produced in relatively small flakes. Can it also be produced “wafer-sized”? Give two examples!
- Question 12: How can a good electrical contact be formed between a single organic molecule (to serve as an interconnect) and a graphite or graphene electrode?

Answers

- Answer 1: Diamond, graphite, fullerenes, carbon nanotubes, and graphene.
- Answer 2: Hole-transporting and electron-blocking layers. Derivatives of fullerenes are used as n-type semiconductors.
- Answer 3: They can be either metallic or semiconducting, depending on the axis *c* along which the graphene has been rolled up to form the CNT.
- Answer 4: Interconnects, a channel of a MOSFET, and transparent electrodes.
- Answer 5: A monoatomic plain of carbon rings forming a two-dimension solid.
- Answer 6: The zero bandgap and the linear dispersion relation around the minima of the conduction band and the maxima of the valence band, i.e. around the so-called K-points.
- Answer 7: By the method of adhesive tape, first found by Geim and Novoselov.
- Answer 8: (a) By forming a narrow strip of graphene, e.g. around 10 nm wide. (b) By functionalization of graphene, e.g. by phenyl groups.
- Answer 9: Extremely high electron mobility and high conductivity.
- Answer 10: Transparent electrodes, biochemical sensors, and electronic material for nanostructured devices and circuits, like for the channel of a FET.
- Answer 11: Yes. (a) By controlled thermal decomposition of a 4H SiC wafer, preferentially a semi-insulating one. (b) By CVD of a carbon source like methane on a metallic carrier.
- Answer 12: The molecule must form a covalent bond with the graphite or graphene, not a van der Waals bond.

References

1. H.W. Kroto, J.R. Heath, S.C. O'Brien, R.F. Curt, R.E. Smalley, Buckminster Fullerene. *Nature* **318**, 162–163 (1985)
2. F. Kreupl, A. P. Graham, M. Liebau, G. S. Duesberg, R. Seidel, E. Unger. Carbon nanotubes for interconnect applications, *Microelectronic Engineering* **64**(1–4), 399–408, 2002; see there Table I
3. C. Bagavathi, K. A. Narayanankutty: *Int. J. of nanotechnology and applications*, ISSN 0973-G31X **6**(1) 23–34(2012)
4. S. A. R. M. Verschuere, C. Dekker, *Nature* **393**, discussed in Rainer Waser (Ed.): *Nanoelectronics and technology* Wiley-VCH 2003 **49** (1998)
5. S. V. Morozov, K. S. Novoselov, M. I. Katsnelson, F. Schedin, D. C. Elias, J. A. Jaszczak, and A. K. Geim. Giant intrinsic carrier mobilities in graphene and its bilayer. *Phys. Rev. Lett.* **100**, 016602—Published 7 January (2008)

6. K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Gregorieva and A. A. Firsov, *Science* **306**, 666 (2004).
7. C. Berger, Z. Song, T. Li, X. Li, A.Y. Ogbazghi, R. Feng, Z. Dai, A.N. Marchenkov, E.H. Conrad, P.N. First, W.A. de Heer, Ultrathin epitaxial graphite: 2D electron gas properties and a route toward graphene-based nanoelectronics. *J. Phys Chem B* **108**, 19912 (2004)
8. N. Agrait, A. Levy Yeyatt, J. M. van Ruitenbeck, Quantum properties of atomic-sized conductors. *Phys. Rep.* **377**, 81–279 (2003)
9. J.C. Cuevas, E. Scheer, *Molecular electronics: An introduction to theory and experiment* (World Scientific Publishing, Singapore, 2010)
10. A.V. Rudnev, V. Kaliginedi, A. Droghetti, H. Ozawa, A. Kuzume, M.-A. Haga, P. Broekmann, I. Rungger, Stable anchoring chemistry for room temperature charge transport through graphite-molecule contacts. *Sci. Adv.* **3**, e1602297 (2017)
11. C. Jia et al., Covalently bonded single-molecule junctions with stable and reversible photo-switched conductivity. *Science* **352**, 1443–1445 (2016)

List of Physics Comments

Chapter 1

- The splitting of energy levels in a periodic lattice
- Dispersion relations of phonons and electrons in comparison
- The role of the momentum of the photon in the recombination process

Chapter 2

- Current flow as a dissipative process
- On the validity of Ohm's law
- The independence of the conductivity of the magnetic field

Chapter 3

- Why Nitrogen is not a donor in Si
- Spatial extension of a shallow donor in Si
- Electronegativity determines the site of an impurity

Chapter 4

- Boltzmann- or Fermi–Dirac distribution in a semiconductor
- A remark concerning the definition of a semiconductor
- Semiinsulating semiconductors, an “imitation” of the intrinsic case
- General features of the law of mass action

Chapter 6

- Lattice defects are always present in thermal equilibrium

Chapter 7

- Similarity of the band structure of direct and indirect semiconductors
- On the notion of “epitaxy”
- Why is ultra high vacuum (UHV) required for molecular beam epitaxy (MBE)?

Chapter 8

The mobility edge

Optical absorption and radiative recombination in an amorphous semiconductor

Passivation of donors and acceptors by hydrogen

Chapter 9

The Einstein relation

Chapter 10

Total energy production of the sun

The binding energy of excitons

Chapter 11

The Lumen as a unit of “brightness” of a light source

Candela and Lumen: The candela as a “fundamental unit”

Chapter 12

The laser, a system far away from thermal equilibrium

On the high directionality and monochromaticity of laser light

Some general remarks concerning lasers

Light confinement.

Can the attenuation of quartz fibers be further improved?

Chapter 13

Different origins of the field current in a reverse biased pn-junction

The field effect and its dependence on carrier concentration

Two-dimensional systems and the de Broglie relation

The quantized Hall effect as a consequence of magnetic flux quantization

Chapter 14

The excimer- and the exciplex-laser

On the production of soft X-rays

Chapter 15

The singlet–triplet problem in OLEDs

Chapter 16

Carbon nanotube as a one-dimensional system

Band structure of graphene

Name Index

A

Abbé, Ernst, [192](#)
Akasaki, Isamu, [169](#)
Amani, [169](#)
Amano, Hiroshi, [169](#), [172](#)

B

Bardeen, John, [96](#)
Bloch, Felix, [10](#)
Boltzmann, Ludwig, [91](#), [93](#)
Brattain, Walter Houser, [96](#)

D

Dash, W.C., [78](#)
Diarmid, Alan M., [244](#)

E

Einstein, Albert, [133](#), [180](#)

G

Geim, Andre, [269](#)
Grove, Andy, [222](#)

H

Heeger, Alan J., [244](#)

L

Lely, Jan Anthony, [105](#)

M

Matare, Herbert F., [96](#)

Moore, Gerald, [236](#)

Moore, Gordon, [222](#)

Mott, Nevill Francis, [118](#), [122](#)

N

Novoselov, Konstantin, [269](#)

Noyce, Robert, [222](#)

O

Ohm, Georg Simon, [25](#), [26](#)

P

Pauling, Linus, [48](#), [96](#)

Pauli, Wolfgang, [81](#)

Pfann, Gardner, [75](#), [79](#)

S

Schneider, Juergen, [172](#)

Schottky, Walter, [70](#)

Shirakawa, Hideki, [244](#)

Shockley, William Bradford, [96](#)

Sommerfeld, Arnold, [96](#)

Spence, Eberhard, [70](#)

V

von Klitzing, Klaus, [216](#), [218](#)

W

Welker, Heinrich, [96](#), [97](#), [112](#)

Subject Index

A

Abbé limit, 234
Abbé resolution limit, 234
AlN, 104, 106, 159, 172
Amorphous semiconductor, 115, 119
Amorphous state, 118
Amplification factor, 203
Anchor group, 271
Annular saw, 75
Arrhenius behavior, 85
A-Si, 115, 120, 121
 a-Si:H, 115, 122–124, 143
 doping-, 122
As precipitates, 89

B

Ballistic transport, 25, 36, 37
Bandgaps, 32
Bandstructure, 12, 19, 20, 241
 -of GaAs, 16
 -of Si, 16
Bipolar transistor, 10, 64, 96, 129, 140,
 199, 200
 amplification factor-, 202
Bloch function, 11, 15
Boltzmann, 54
 -distribution, 55, 133, 134
 -factor, 183
 -statistics, 55
Boolean functions, 229
Boron
 -nitride, pyrolytic, 102
 -oxide, 102
Bottom-up development, 262
Bottom-up techniques, 221
Bravais lattices, 4

Brightness, 176
Brightness of LED, 173
Brillouin zone, 11, 16
Bulk crystals of GaN, 107
Burger's vector, 86, 90, 106

C

Candela, 176
Carbon allotropes, 262
Carbon inclusion, 89
Carbon Nanotubes (CNT), 250, 261, 263,
 264, 273, 274
 applications of-, 264
 -field effect transistor, 265, 266
 metallic-, 264–266, 271
 production of-, 266
 semiconducting-, 243, 264
 -transparent electrode, 250
Carrier concentration
 intrinsic-, 58, 60, 64
 -temperature dependence, 64
CD player, 192
CdS, 120
CdTe, 120, 153
Cell, 148
Charge carrier, 247
CMOS, 229
CMOS complementary MOSFET
 technology, 230
Compound semiconductors, 101, 167, 192,
 195
 history-, 95
 mixed-, 167, 195
 new-, 104
Conductance quantum, 271
Conduction band, 3, 7, 16, 40, 47, 52, 57

Conductivity, 244, 258
 Conductor tracks, 226
 Al-, 226
 Cu-, 226
 Conjugated, 250
 -double bonds, 240, 258
 -polymers, 250
 Czochralski, 69, 74, 75, 77, 78, 105
 Czochralski growth, 73, 74, 79, 89, 102, 105

D
 Density of state, 23, 51, 52, 67, 115, 118, 178, 248
 effective-, 53, 54
 -mass, 23, 53
 Density of state masses, 23
 Diamond, 45, 46
 Diamond structure, 6, 240, 262
 Dielectric mirror, 187
 Dielectric relaxation time, 208
 Diffusion, 81
 -barrier, 226
 -constants, 84, 85
 -current, 84, 129, 131, 133, 135, 200
 interstitial-, 226
 Dislocation, 33, 78, 86, 100
 -as recombination centre, 86
 edge-, 85, 86
 -electronic properties, 94
 screw-, 81, 85–87
 Dislocation free, 78
 Dislocation free Si, 78
 Dispersion relation, 3, 12, 16, 23
 linear –, 274
 Dissipative process, 28
 Donors
 - shallow, 39, 46, 59, 63, 77
 Double heterostructure laser, 190
 Double patterning, 234
 Dynamic Random Access Memory (DRAM), 228
 electromigration, 226

E
 Efficiency, 124, 144, 148, 153, 156, 157
 Einstein
 -coefficient, 180–182
 -relation, 133, 134, 203
 Einstein coefficient of absorption, 180, 182
 Einstein coefficient of spontaneous emission, 181, 182

Einstein coefficient of stimulated emission, 181, 182
 Electromigration, 226
 Electronegativity, 39, 46, 47
 -list of, 48
 Energy band, 7, 10, 11, 21, 41, 49, 130
 Energy bands, 3, 7, 10–12, 21, 40, 41, 49, 130, 180, 182
 Energy band structure, 116
 Energy Payback Time (EPBT), 156
 Epitaxial Lateral Overgrowth (ELO), 170
 Epitaxy, 108
 Er-repeater, 197
 Excimer, 231, 232
 Excimer laser, 230, 237
 Exciplex, 232
 Exciplex laser, 231, 237
 Exciplexlasers, 231
 Exciton, 150
 binding energy-, 150

F
 Fermi-Dirac, 184
 Fermi-Dirac distribution, 55, 133
 Fermi-Dirac function, 63
 Fermi energy, 57
 -in intrinsic semiconductor, 59
 -in semiinsulating material, 62
 Fermi energy E_F , 63
 Fermi level, 63, 131, 185
 Field current, 129, 131, 133–135, 147, 200
 Field effect, 207, 208
 Field effect transistor, 36, 64, 96, 123, 199, 204, 257
 -frequency limit, 35
 inversion channel-, 206
 Filling factor, 146, 147
 Flat, 75
 Flat panel X-ray detector, 124
 Float zone, 71, 73–75, 78
 Float zone process, 69
 Fluorescent emitter, 254
 Flux quantization, 217
 Frenkel defect, 82
 Frenkel exciton, 252, 258
 Fresnel's formulas, 186
 Fullerenes, 261, 262, 273, 274

G
 GaAs, 102, 153
 GaN, 104
 GaN light emitting diodes, 112

Graphene, 263, 267, 273
 -application of, 270
 -band structure of, 268
 -production of, 269

Graphite, 262
 -structure, 262

H

Hall

-constants, 31
 -effect, 25, 30
 quantum-, 216–219
 -resistance, 218

Heavy holes, 20

Helmholtz free energy, 91, 92

Highest Occupied Molecular Orbital
 (HOMO), 243, 246, 250, 251, 258

High k-dielectrics, 211

Hole, 9

-concentration of holes, 51, 67, 132

Hopping conduction, 119

Hydride Vapour Phase Epitaxy (HVPE),
 107, 170

Hydrogen model, 39, 43–45, 150

I

Immersion lithography, 234

Indirect semiconductors, 17, 18, 119, 120,
 165

Indium Tin Oxide (ITO), 249, 257

InP, 102

growth, 102

Insulators, 8, 40

Integrated circuits, 77, 78, 104, 199, 204,
 221, 228, 236

International Technology Roadmap for
 Semiconductors (ITRS), 223

Interstitials, 82

Intrinsic case, 57, 59

Inverter, 229

Ion implantation, 221, 226

Isoelectronic Impurities, 47

K

Kirchhoff's law, 201

L

Laser, 179

-condition, 182, 184, 185, 196
 double heterojunction-, 189, 194

semiconductor-, 17, 56, 87, 100, 167,
 179, 192, 194
 writer, 228

Laser principle, 179

Lattice defects

extrinsic-, 81
 -in thermal equilibrium, 90
 intrinsic-, 81
 vectors, 4

Lattice mismatch, 85

Law of mass action, 57, 65–67, 132

LEC, 102, 103, 105

LEC process, 102, 103

LED, 17, 159

blue-, 105, 168, 169

-color, 166

GaNN-, 167

-bowing parameter, 167

-linewidth, 161, 163

UV-, 172

white-, 159, 170, 172, 176, 178

Lely process, 105

Light holes, 20

Light sources, 230

Line defects, 81, 85

Linewidths, 159

Longitudinal effective mass, 22

Lowest Unoccupied Molecular Orbital
 (LUMO), 243, 246, 248, 258

Lumen, 176, 178

Lumen as unit of “brightness”, 174

Luminescence Converting Light Emitting
 Diode (LUCOLED), 172, 178

M

Macrostate, 92

Maragani convection, 74

Mask aligner, 227, 228

Mass

conductivity-, 21
 effective-, 19, 21
 longitudinal-, 22
 transverse-, 22

Metal, 8

Metallurgical, 156

Metallurgical route, 154

Metal Organic Vapour Phase Epitaxy
 (MOVPE), 95, 108, 111, 170

Metals, 8, 25, 244, 249

Metallurgical Silicon, 69, 71, 77, 154, 155

Microgravity, 74

Micropipe, 90, 106

- Microstate, 92
- Microtubes, 90
- Mismatch, 101, 112, 169, 170, 172
- Mobility, 25, 30–33, 86, 96
 - and frequency limit, 35
 - and Hall effect, 30
 - doping dependence, 33
 - in GaAs, 33
 - in III-V-compounds, 95, 96
 - temperature dependence, 33
- Mobility edge, 118
- Molecular Beam Epitaxy (MBE), 95, 108
- Molecular electronics, 261
- Molecular switches, 272
- Monoatomic chain, 14
- Moore's law, 221, 223, 233
- MOSFET, 199, 204, 207, 214, 218, 222, 226, 229, 265
 - threshold voltage, 216
- Multiwire saws, 76, 79

- N**
- Nanotube is metallic, 264
- "Negative" photoresists, 224
- Nitrogen, 212
 - donor, 42
 - in SiON, 212
- Nonradiative, 164
- Nonradiative processes, 87, 163
- Notch, 75
- Number of holes, 54

- O**
- Ohm's law, 25, 27–29, 206
- Open_{circuit} voltage, 145
- Optical data transmission, 179, 192, 194
- Optimum bandgap, 157
- Organic electronics, 239, 248, 250, 255, 257, 258
- Organic Light-Emitting Diodes (OLEDs), 248–250, 252, 253, 258
 - singlet-triplet problem, 253
 - small molecule-, 250
- Organic solar cells, 249
- Oxidation, 207, 221, 225, 237
 - dry-, 225
 - wet-, 226

- P**
- Passivation, 123, 178
- PEDOT, 246, 250

- Peierls instability, 243, 264
- Phase shift masks, 234, 237
- Phosphor, 33
- Phosphorescent emitter, 254
- Phosphorus, 41
- Photolithography, 223, 227, 230, 231
 - exposure, 227
 - Light sources, 230
- Photomask, 228, 236
- Photoresists, 221, 224, 228, 234, 236
 - negative, 224, 225, 236, 237
 - positive, 224
- Photovoltaic effect, 143, 144, 157
- Physical Vapour Transport (PVT), 105, 106
- Pn-junction, 129, 144, 159, 165, 183, 185, 189, 199, 200, 204, 239
 - diffusion current, 129, 131, 133, 135, 200
 - field current, 129, 131, 133
 - i(U) characteristic, 146
 - reverse current, 204
- pn-junction, 129
- Point defects, 81, 82
- Point lattice, 3, 23
- Polaron, 243, 253, 255, 257
- Polymer, 179, 224, 239, 241
 - as photoresist, 224
 - carrier injection in-, 246
 - conductivity-, 243
 - curing, 172
 - hardening, 179
- Poly(-p-phenylene-vinylene) (PPV), 248
- PVT-process, 90

- Q**
- Quantized Hall effect, 199
- Quantum efficiency, 163
 - external-, 159, 165, 166, 177
 - internal-, 159, 163, 165, 166, 169, 177
- Quantum Hall effect, 216, 217
- Quartz fiber, 193
 - attenuation, 193
- Quaternary compounds, 95, 179

- R**
- Radiative processes, 163
- Radiative recombination, 87
- Rayleigh scattering, 193, 232
- Rectifying diode, 129
- Resolution enhancement, 236
- Resolution enhancement techniques, 234
- Reverse current, 204

S

Saturation velocity, 35
 Sawing damage, 76
 Schottky defect, 83
 Schroedinger equation, 3, 10, 15, 51, 214
 Screw dislocations and crystal growth, 87
 Segregation coefficient, 75, 154
 Self-trapping energy, 248
 Semiconducting nanotube, 264
 Semiconductor
 amorphous-, 115, 119
 definition-, 58
 direct-, 17
 indirect-, 18, 19, 95, 143, 153
 Semiconductor Industries Associations (SIAs), 223
 Semiconductor lasers, 179
 Semiinsulating, 63, 64, 104
 Semiinsulating GaAs, 89
 Semiinsulating semiconductors, 61
 Semiinsulating, 104
 Shallow acceptors, 62, 122
 Shallow donors and acceptors, 46
 Shallow donors and acceptors in GaAs, 161
 Shallow donors or acceptors, as e.g. P or B in Si, 122
 Shockley-Queisser-limit, 153
 Shockley-Read-Hall process, 164
 Si, 119
 Au energy levels in, 165
 SiC, 104
 4H-, 18, 67, 89, 90, 98, 104–106, 112, 270
 6H-, 18, 64, 65, 67, 104–106, 112, 169
 growth of-, 90, 105
 -MOSFET, 104, 207
 -polytypes, 104
 Siemens process, 69–71, 77, 143, 150, 155
 SiHCl₃, 71, 77, 154
 Silver nanowire, 250
 Single molecule interconnects, 271
 SiON, 212
 Small molecule, 251
 Small molecules, 253
 Soft X-rays, 232, 237
 Solar cell, 10, 18, 124, 129, 140, 143, 144, 148
 a-Si-, 120
 -efficiency, 143
 GaAs-, 18
 operating principle-, 148
 optimum bandgap-, 151
 organic-, 255, 258

Perovskite, 157

tandem-, 153
 thick layer-, 18, 143
 thin layer-, 143, 153

Space charge limited current, 29

Stacking faults, 89, 94, 103

Stepper, 228

T

Technology node, 221, 223, 230, 232, 235, 236

Technology node 7nm, 237

Temperature dependence of mobility, 33

Ternary, 95

Ternary compound, 194

Thermalization, 143, 151

Three-dimensional defects, 81

Threshold voltage, 206, 229

Top-down development, 261

Top-down route, 221

Transistor count, 222

Transparent electrodes, 249, 250, 266

Trichlorosilane SiHCl₃, 69

Two-dimensional defects, 81

Two-dimensional electron gases, 214, 215

U

Ultra High Vacuum (UHV), 109

Unit cell, 6, 119

Unit cell of Silicon, 6

UVC LED, 178

V

Vacancies, 82

Vacancy diffusion and interstitial diffusion, 83

Valence band, 3, 7, 16, 40, 47, 57

Variable range hopping, 121

Vertical Gradient Freeze (VGF), 102, 103, 112

W

Wafering, 69, 75

Wannier-Mott exciton, 252

X

X-ray flat panel detector, 115

Y

Yttrium aluminum garnet YAG
($\text{Y}_3\text{Al}_5\text{O}_{12}$), [171](#)

Z

Zone refining, [75](#)