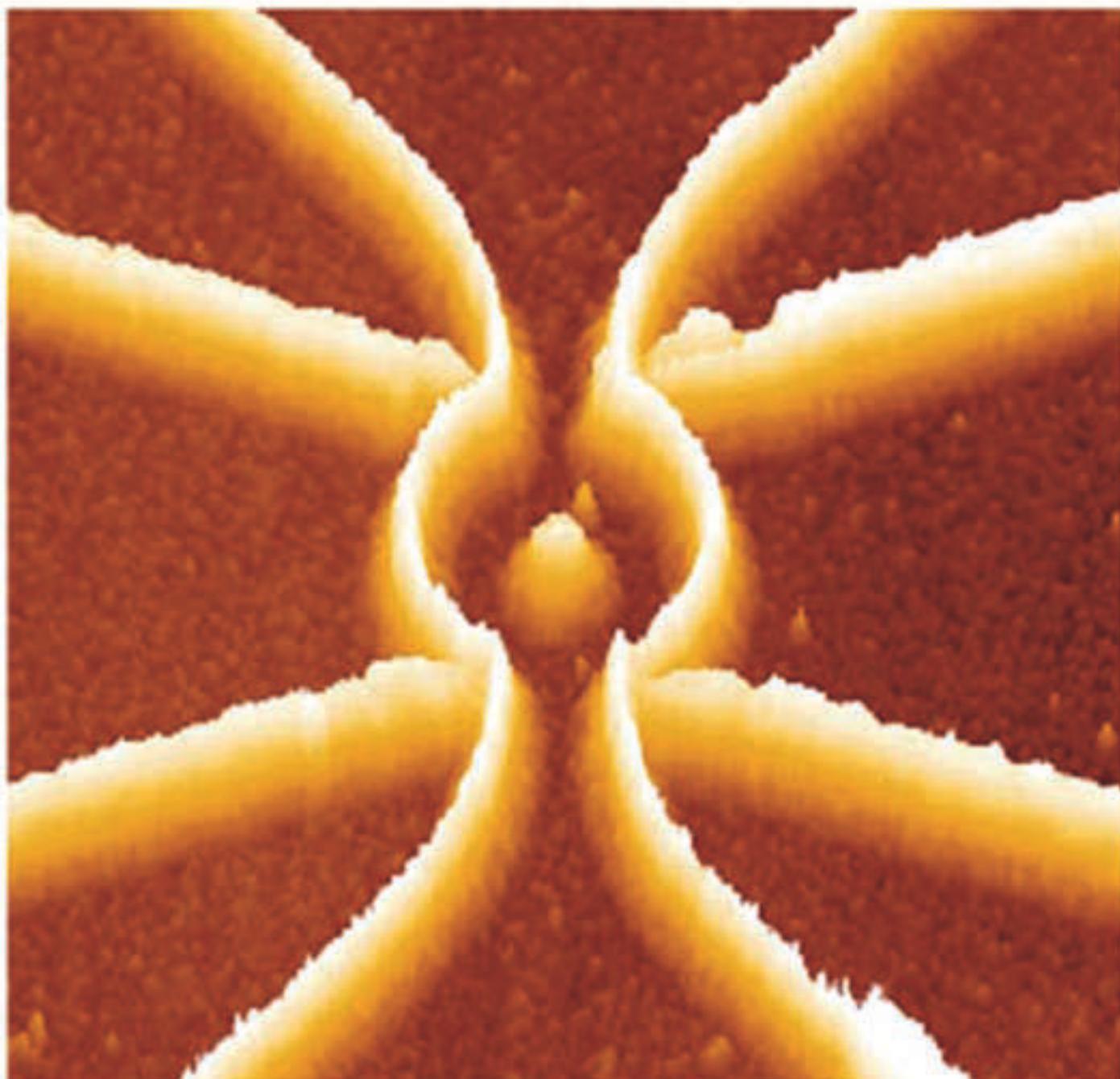


Thomas Heinzel

Mesoscopic Electronics in Solid State Nanostructures

Second, Completely Revised and Enlarged Edition



Thomas Heinzel

Mesoscopic Electronics in Solid State Nanostructures

Second, Completely Revised and Enlarged Edition



WILEY-VCH Verlag GmbH & Co. KGaA

This Page Intentionally Left Blank

Thomas Heinzel

Mesoscopic Electronics in Solid State Nanostructures

Second, Completely Revised and Enlarged Edition



WILEY-VCH Verlag GmbH & Co. KGaA

1807–2007 Knowledge for Generations

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!



William J. Pesce
President and Chief Executive Officer



Peter Booth Wiley
Chairman of the Board

Thomas Heinzl
**Mesoscopic Electronics in Solid
State Nanostructures**

The Author

Prof. Dr. Thomas Heinzel
Dept. of Solid State Physics
Heinrich-Heine-Universität Düsseldorf
thomas.heinzel@uni-duesseldorf.de

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.:
applied for**British Library Cataloguing-in-Publication Data**
A catalogue record for this book is available from
the British Library.

**Bibliographic information published by
the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this
publication in the Deutsche Nationalbibliographie;
detailed bibliographic data are available in the
Internet at <<http://dnb.d-nb.de>>.

©2007 WILEY-VCH Verlag GmbH & Co. KGaA,
Weinheim

All rights reserved (including those of translation
into other languages). No part of this book may be
reproduced in any form – photopyting, microfilm,
or any other means – transmitted or translated into
a machine language without written permission
from the publishers. Registered names,
trademarks, etc. used in this book, even when not
specially marked as such, are not to be
considered unprotected by law.

Typesetting Da-TeX Gerd Blumenstein, Leipzig
Printing Strauss GmbH, Mörlenbach
Binding Litges & Dopf GmbH, Heppenheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN: 978-3-527-40638-8

Dedication

To Carola and Alexander

This Page Intentionally Left Blank

Contents

Preface *XIII*

1	Introduction	1
1.1	Preliminary remarks	1
1.2	Mesoscopic transport	2
1.2.1	Ballistic transport	3
1.2.2	The quantum Hall effect and Shubnikov–de Haas oscillations	5
1.2.3	Size quantization	7
1.2.4	Phase coherence	8
1.2.5	Single-electron tunneling and quantum dots	9
1.2.6	Superlattices	10
1.2.7	Spintronics	11
1.2.8	Samples, experimental techniques, and technological relevance	11
2	An update of solid state physics	15
2.1	Crystal structures	16
2.2	Electronic energy bands	18
2.3	Occupation of energy bands	27
2.3.1	The electronic density of states	27
2.3.2	Occupation probability and chemical potential	29
2.3.3	Intrinsic carrier concentration	29
2.3.4	Bloch waves and localized electrons	31
2.4	Envelope wave functions	32
2.5	Doping	36
2.6	Diffusive transport and the Boltzmann equation	40
2.6.1	The Boltzmann equation	41
2.6.2	The conductance predicted by the simplified Boltzmann equation	44
2.6.3	The magneto-resistivity tensor	46
2.6.4	Diffusion currents	47

2.7	Scattering mechanisms	48
2.8	Screening	50
3	Surfaces, interfaces, and layered devices	57
3.1	Electronic surface states	59
3.1.1	Surface states in one dimension	59
3.1.2	Surfaces of three-dimensional crystals	65
3.1.3	Band bending and Fermi level pinning	67
3.2	Semiconductor–metal interfaces	68
3.2.1	Band alignment and Schottky barriers	69
3.2.1.1	The Schottky model	72
3.2.1.2	The Schottky diode	73
3.2.2	Ohmic contacts	73
3.3	Semiconductor heterointerfaces	74
3.4	Field effect transistors and quantum wells	77
3.4.1	The silicon metal–oxide–semiconductor field effect transistor	77
3.4.1.1	The MOSFET and digital electronics	81
3.4.2	The Ga[Al]As high electron mobility transistor	84
3.4.3	Other types of layered devices	87
3.4.3.1	The AlSb–InAs–AlSb quantum well	87
3.4.3.2	Hole gas in Si–Si _{1-x} Ge _x –Si quantum wells	89
3.4.3.3	Organic FETs	89
3.4.4	Quantum confined carriers in comparison to bulk carriers	91
4	Experimental techniques	97
4.1	Sample preparation	97
4.1.1	Single crystal growth	98
4.1.2	Growth of layered structures	100
4.1.2.1	Metal organic chemical vapor deposition (MOCVD)	101
4.1.2.2	Molecular beam epitaxy (MBE)	101
4.1.3	Lateral patterning	107
4.1.3.1	Defining patterns in resists	107
4.1.3.2	Direct writing methods	110
4.1.3.3	Etching	112
4.1.4	Metallization	113
4.1.5	Bonding	115
4.2	Elements of cryogenics	116
4.2.1	Properties of liquid helium	117
4.2.1.1	Some properties of pure ⁴ He	117
4.2.1.2	Some properties of pure ³ He	120
4.2.1.3	The ³ He/ ⁴ He mixture	121
4.2.2	Helium cryostats	122

4.2.2.1	^4He cryostats	122
4.2.2.2	^3He cryostats	125
4.2.2.3	$^3\text{He}/^4\text{He}$ dilution refrigerators	125
4.3	Electronic measurements on nanostructures	127
4.3.1	Sample holders	128
4.3.2	Application and detection of electronic signals	128
4.3.2.1	General considerations	128
4.3.2.2	Voltage and current sources	129
4.3.2.3	Signal detectors	130
4.3.2.4	Some important measurement setups	133
5	Important quantities in mesoscopic transport	139
5.1	Fermi wavelength	139
5.2	Elastic scattering times and lengths	139
5.3	Diffusion constant	140
5.4	Dephasing time and phase coherence length	143
5.5	Electron-electron scattering time	144
5.6	Thermal length	144
5.7	Localization length	145
5.8	Interaction parameter (or gas parameter)	145
5.9	Magnetic length and magnetic time	145
6	Magneto-transport properties of quantum films	147
6.1	Landau quantization	148
6.1.1	Two-dimensional electron gases in perpendicular magnetic fields	148
6.1.2	The chemical potential in strong magnetic fields	151
6.2	The quantum Hall effect	154
6.2.1	Phenomenology	154
6.2.2	Toward an explanation of the integer quantum Hall effect	156
6.2.3	The quantum Hall effect and three dimensions	161
6.3	Elementary analysis of Shubnikov–de Haas oscillations	162
6.4	Some examples of magneto-transport experiments	165
6.4.1	Quasi-two-dimensional electron gases	165
6.4.2	Mapping of the probability density	167
6.4.3	Displacement of the quantum Hall plateaux	167
6.5	Parallel magnetic fields	169
7	Quantum wires and quantum point contacts	177
7.1	Diffusive quantum wires	179
7.1.1	Basic properties	179
7.1.2	Boundary scattering	181

7.2	Ballistic quantum wires	182
7.2.1	Phenomenology	182
7.2.2	Conductance quantization in QPCs	184
7.2.3	Magnetic field effects	191
7.2.4	The “0.7 structure”	195
7.2.5	Four-probe measurements on ballistic quantum wires	195
7.3	The Landauer–Büttiker formalism	198
7.3.1	Edge states	199
7.3.2	Edge channels	202
7.4	Further examples of quantum wires	204
7.4.1	Conductance quantization in conventional metals	204
7.4.2	Molecular wires	206
7.4.2.1	Carbon nanotubes	206
7.5	Quantum point contact circuits	210
7.5.1	Non-Ohmic behavior of QPCs in series	210
7.5.2	QPCs in parallel	212
7.6	Semiclassical limit: conductance of ballistic 2D systems	214
7.7	Concluding remarks	218
8	Electronic phase coherence	223
8.1	The Aharonov–Bohm effect in mesoscopic conductors	223
8.2	Weak localization	226
8.3	Universal conductance fluctuations	229
8.4	Phase coherence in ballistic 2DEGs	234
8.5	Resonant tunneling and s-matrices	236
9	Single-electron tunneling	247
9.1	The principle of Coulomb blockade	247
9.2	Basic single-electron tunneling circuits	250
9.2.1	Coulomb blockade at the double barrier	252
9.2.2	Current–voltage characteristics: The Coulomb staircase	255
9.2.3	The SET transistor	259
9.3	SET circuits with many islands: The single-electron pump	265
10	Quantum dots	273
10.1	Phenomenology of quantum dots	274
10.2	The constant interaction model	279
10.2.1	Quantum dots in intermediate magnetic fields	283
10.2.2	Quantum rings	285
10.3	Beyond the constant interaction model	287
10.3.1	Hund’s rules in quantum dots	287
10.3.2	Quantum dots in strong magnetic fields	287

10.3.3	The distribution of nearest-neighbor spacings	290
10.4	Shape of conductance resonances and I–V characteristics	294
10.5	Other types of quantum dots	297
10.5.1	Metal grains	298
10.5.2	Molecular quantum dots	299
10.6	Quantum dots and quantum computation	301
11	Mesoscopic superlattices	309
11.1	One-dimensional superlattices	310
11.2	Two-dimensional superlattices	312
11.2.1	Semiclassical effects	312
11.2.2	Quantum effects	318
12	Spintronics	323
12.1	Ferromagnetic sandwich structures	324
12.1.1	Tunneling magneto-resistance (TMR) and giant magneto-resistance (GMR)	324
12.1.2	Spin injection into a non-magnetic conductor	328
12.2	The Datta–Das spin field effect transistor	332
12.2.1	Concept of the Datta–Das transistor	332
12.2.2	Spin injection in semiconductors	333
12.2.2.1	Interface tunnel barriers	333
12.2.2.2	Ferromagnetic semiconductors	335
12.2.3	Gate-induced spin rotation: The Rashba effect	336
12.2.4	Spin relaxation and spin dephasing	339
A	SI and cgs units	343
B	Correlation and convolution	345
B.1	Fourier transformation	345
B.2	Convolutions	345
B.3	Correlation functions	347
C	Capacitance matrix and electrostatic energy	349
D	The transfer Hamiltonian	353
E	Solutions to selected exercises	355
	References	383
	Index	393

This Page Intentionally Left Blank

Preface

Exploring new orders of magnitude seems to be a part of human nature. This tendency is reflected even in the language. If a particular “megadeal” is really stunning, we feel more and more obliged to call it a “gigadeal”. It will not take long before the first “terastar” pops up! Each new generation of particle colliders, telescopes, or lasers, in fact of almost any scientific device or technique you can think of, extends the accessible interval of a physical quantity. It is rather the rule than the exception that novel phenomena are discovered in such a process, some of which have been anticipated, others of which have not. Such an evolution seems to gain speed as soon as it gets classified as useful, besides pure scientific interest. It is of course debatable what exactly should be considered as “useful”.

In any case, the miniaturization of electronic circuits during the past 50 years has been both scientifically rewarding and useful in our daily life. It is certainly not necessary to support this statement by examples. Scientifically, however, the expression “microstructure” no longer fuels our imagination. Meanwhile, the really exciting electronic circuits are *nanostructures*. As this name already suggests, nanostructures are objects with structures in the nanometer (nm) regime, which can mean just 1 nm, but also just a little less (in fact, in some cases, even somewhat more) than 1000 nm. The point of nanostructure science is that, within the last two decades, tremendous progress has been made in fabricating, controlling and understanding structures in this size regime. This is true for a wide variety of fields, including, for example, gene technology, crystal growth or microchip – excuse me, nanochip – fabrication. The resulting novel possibilities at hand are really breathtaking and get heavily explored by a significant fraction of the scientific community. In many cases, having control over the size and shape of an object in the nanometer regime means being able to control its chemical and/or physical properties. For example, the size of a semiconductor nanocluster determines its optical emission spectrum via size quantization; while from introductory solid state physics lectures, we have learned that this property is related to the bandgap, an intrinsic feature of the material. By now, the size reduction has actually

reached dimensions that are of interest for chemistry and biology, and there is a rapidly growing overlap. For example, you can think of an ionic channel (they reside in cell membranes and control the electrochemical potential of cells by selectively transferring certain types of ions across the membrane) as a molecular transistor. On the other hand, nanostructure physicists have started to use DNA strands as wires and templates for nanocircuit fabrication.

One branch of nanoscience deals with the electronic transport properties of solid state nanostructures. This field is often referred to as “mesoscopic transport”, an expression which indicates that the explanations for the observed transport phenomena must be sought somewhere in between microscopic and macroscopic models. The purpose of the present book is to introduce the reader to this topic from an experimental point of view. “The reader” is hereby assumed to be a student of physics or a related field, who has just finished introductory courses, in particular those on solid state physics and quantum mechanics, and plans to study nanoscience more closely. The reader is picked up at the knowledge he/she is likely to have, and a ride is given to ongoing research activities in the field of mesoscopic transport. Along the way, the elementary concepts and nanostructures are introduced.

Selecting illustrative experiments for such a purpose is of course a highly subjective matter. The author has tried to pick particularly instructive examples well known to him, which can furthermore be explained within the scope of this book. These examples have thus not necessarily been of high relevance for the evolution of the field, and I apologize for this shortcoming. It should be remarked that, in some of the figures, the original data have been redrawn for better reproduction quality and for a consistent presentation.

The text contains a somewhat unusual feature, namely “papers” in the exercises sections. Their purpose is to encourage the student to go through selected, usually quite recent, original publications. They are referred to by [P chapter.number] in the text. The student should be able to summarize the beef of such a paper in a 15 minute talk. The reader is strongly encouraged actually to do this. Besides collecting complementary information and getting exposed to different styles of presentation, the experience of being able to understand the stuff written down not in a textbook but in an original paper can be highly motivating.

I hope that after going through the text, the reader will not only be able to join with some confidence an experimental research group working in this field, but also feel well prepared for more advanced theoretical lectures on mesoscopic physics.

This book, and with it the author, has enjoyed a lot of encouragement and support from many sides. I would like to thank particularly Hermann Grabert and Wolfgang Häusler, who read through parts of the manuscript and made many valuable comments. I am grateful to my colleagues Klaus Ensslin,

Andreas Fuhrer, Miha Furlan, Ryan Held, Thomas Ihn, Silvia Lüscher, Jörg Rychen, and Volkmar Senz for countless fruitful discussions and stimulating ideas. Special thanks go to those who supplied figures for this book, namely Günther Bauer, Mildred Dresselhaus, Andreas Fuhrer, Theo Geisel, Adam Hansen, Roland Ketzmerick, Anupam Madhukar, Andy Sachrajda, Elke Scheer, Jürgen Smet, and Horst Stormer. Furthermore, I thank my students, whose critical but always constructive comments have shaped and improved the presentation of the material. Last, but not least, I thank my wife Ulrike. With her tremendous energy and selfless support, she managed to supply the refuge I needed to transform my disorganized lecture notes into a book.

Thomas Heinzel, Düsseldorf, 2006

This Page Intentionally Left Blank

1

Introduction

1.1

Preliminary remarks

Over the past 30 years, the miniaturization of electronic devices has strongly influenced the technological evolution. Just think of the progress made in communication technology, or of the improvements in personal computers. For the money spent on a pocket calculator (which was barely able to carry out the four basic arithmetical operations) 30 years ago, you can buy today a desktop computer able to solve quite sophisticated numerical tasks, which in the 1970s could be tackled only by supercomputers. *Moore's law* states that roughly every three years the number of transistors per microchip doubles. This law has been valid remarkably well in the past three decades, and it is expected to hold for some more years to come, although probably with a slightly reduced rate. This process, however, requires an ongoing reduction of the sizes of features, which up to now has essentially been achieved by using smaller wavelengths for optical lithography (the wavelength determines the resolution limit via diffraction). This is much more challenging than it may sound, for several reasons.

First of all, a quick glance in an optics textbook reveals that the index of refraction of all common glasses diverges rapidly as the wavelength gets reduced to about 200 nm. In addition, metals become transparent at their plasma frequencies, which typically fall in the same range of wavelengths. Hence, constructing both lenses and mirrors for the 100 nm regime is not that easy. Currently, the wavelengths used for lithography are of the order of 250 nm. Alternative lithographic techniques are able to pattern significantly smaller feature sizes. Although electron beam lithography is used in industry for some fabrication steps, it is too expensive for mass production of microchips. Novel patterning schemes, such as self-assembly or lithography with scanning probe microscopes, are presently the subject of extensive studies in research labs all around the world. It is, however, very unlikely that these techniques will replace optical lithography within the foreseeable future.

Second, the patterns illuminated in an optical photoresist have to be transferred into a structured device. Processes like developing the photoresist,

semiconductor etching, metal evaporation, alloying or selective doping must be carried out without losing the resolution. Furthermore, the devices must be connected to wires, and the inevitable heating generated during operation must be kept under control.

Suppose that nanoscientists find adequate solutions to all these technological problems – there is in fact little doubt that they will. Then, however, another issue will become more and more important: all the above considerations implicitly assume that the components of a microchip can be scaled down arbitrarily without changes in their performance. This is not the case! Conventional transport theory makes presumptions about certain length and energy scales. For example, it is assumed that the electron mean free path is small compared to the feature size of the device, like the gate length of a transistor. The concept of resistivity is based on this assumption. Within the Boltzmann theory of electronic transport, it is assumed that the acceleration of the Fermi sphere by external electric fields gets compensated by many kinds of relaxation processes, which generate friction. In a stationary state, these friction forces balance the effects of the external field, and the resistivity of the sample can be defined.

What happens for device sizes comparable to the mean free path, or to other relevant length scales? Well, we then enter the regime of *mesoscopic transport*. Novel effects occur, which may profoundly change the device performance. Introducing these effects is the major goal of this book. In the following section, we will look at the specific length and energy scales somewhat more closely and give examples for typical transport properties of samples in the mesoscopic regime.

1.2

Mesoscopic transport

What characterizes the mesoscopic regime? The answer depends on the particular quantity under study. For the above example, the criterion would be that the device size must be comparable to or smaller than the electronic mean free path ℓ_e . Other length scales are the de Broglie wavelength of the electrons that carry the current, which in all cases studied in this book are those electrons at the Fermi edge. Their de Broglie wavelength is the Fermi wavelength $\lambda_F = h / \sqrt{2m^* E_F}$, where m^* denotes the effective electron mass (see Chapter 2 for details), and E_F is the Fermi energy. If the feature sizes of the sample are comparable to λ_F , the wave character of the electrons will become essential, and their kinetic energies will quantize. This fact is often referred to as *size quantization*, which is nothing but elementary quantum mechanics. If size quantization takes place in one spatial direction only, the electron system is

confined to two dimensions, and we speak of *quantum films*, which are the topic of Chapters 3 and 6. Suppose we confine our electrons in a second spatial direction. Their motion then becomes one-dimensional, and we have a *quantum wire*. The basic properties of quantum wires are discussed in Chapter 7. Finally, we can confine the electron in all directions, like in an atom. The resulting objects are known as *quantum dots* or *artificial atoms* (see Chapter 10). Another important length scale is the *phase coherence length*. Most of us are aware of the diffraction pattern that electrons produce as they traverse a double slit setup in a vacuum tube. However, we usually do not think of electronic interference effects in solid state devices. Nevertheless, these effects do occur and become particularly important in devices with dimensions of the order of the phase coherence length. Phase coherent electrons are the topic of Chapter 8. Furthermore, it has turned out that the granular character of the electrons, which even in macroscopic samples plays an important role since it is responsible for shot noise, becomes increasingly important in nanostructures. The point here is that the energy needed to charge a small island with a single electron may become significant. The resulting effects are summarized by *single-electron tunneling*, which is the topic of Chapter 9.

Fig. 1.1 gives an overview of the most important mesoscopic regimes, and we continue with a brief survey of the phenomena to be discussed.

1.2.1

Ballistic transport

In order to enter the ballistic regime, the mean free path ℓ_e , which roughly speaking is the average distance an electron travels before getting scattered,¹ must be small compared to the relevant sample length L . At room temperature, a major source of scattering is electron–phonon interaction, with a mean free path of the order of 20 nm. How are we supposed to describe electron transport through a wire with $L < \ell_e$? Elementary solid state physics tells us that Bloch electrons in a perfect crystal lattice experience no resistance at all. We are therefore tempted to expect an infinite conductance. This would mean that in such small circuits, there is no dissipation, no heat generation and no energy loss as the electronic signal is transferred, like in a superconductor! Surprisingly, we cannot avoid resistances as we transfer electrons across ballistic wires, although, strictly speaking, the wire itself does have an infinite conductance. It should surprise you even more that the conductance we measure is in fact quantized in multiple integers of $2e^2/h$ (see Fig. 1.2). We do not worry too much about the sample details for now. After going through Chapters 3 and 4, we will know that below the sample surface shown in the picture to the left, a quantum film of electrons resides that has been removed

¹) A more accurate definition will be given in Chapter 2.

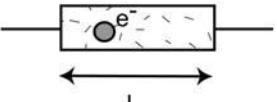
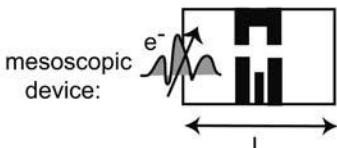
conventional device:		mesoscopic device:	
	$L \gg l_e$		$L \lesssim l_e$
diffusive		ballistic	
$\ell_\phi \gg l_e$	incoherent	$\ell_\phi \lesssim l_e$	phase coherent
$L \gg \lambda_F$	no size quantization	$L \lesssim \lambda_F$	size quantization
$e^2/C < k_B\Theta$	no single electron charging	$e^2/C \gtrsim k_B\Theta$	single electron charging effects
$L \gg l_s$	no spin effects	$L \lesssim l_s$	spin effects

Fig. 1.1 The left column summarizes what we mean by a “conventional device”, like the resistor sketched at the top. Electrons can be thought of as strongly localized wave packets without spin, which move through a disordered device with the drift velocity. Their mean free path is much smaller than the device size L . Transport is *diffusive*. Since the phase coherence length ℓ_ϕ is also small compared to L , the transport is *incoherent*. Furthermore, the Fermi wavelength is much smaller than L , and consequently size quantization is absent. Moreover, the capacitance of the device is so large that the energy needed to charge it with a single electron is negligibly small. In the right column, the conditions necessary to enter the mesoscopic

regime are shown. The cartoon at the top indicates a sample free of scatterers, except for a manmade non-conducting structure (black). Transport through the sample is therefore ballistic. If $\ell_\phi \geq L$, the electrons pass through the sample coherently, and we can expect interference effects. Furthermore, the feature sizes may be comparable to λ_F , such that size quantization occurs. The capacitances may be sufficiently small, such that single-electron charging effects may become observable. Finally, the electron spin (denoted by the arrow) can have implications for the transport properties in mesoscopics, provided the spin dephasing length ℓ_s is larger than the device.

underneath the bright lines, which are oxide lines on a semiconductor (GaAs) surface. You will then, hopefully, accept the fact that the whole area shown here is free of scatterers, at least at low temperatures. The structure can thus be thought of as a *three-terminal device*. If a voltage is applied between the *source* and the *drain* terminals, a current will pass through the narrow constriction defined by the two oxide lines. Such ballistic constrictions with size quantization in two directions are called *quantum point contacts*. The width of this constriction is of the order of the Fermi wavelength and can be tuned by applying an additional voltage to the third terminal, labeled as *planar gate*. This works because of the *field effect*, which should again have become clear after reading Chapters 3 and 4. The conductance of this device as a function of the planar gate voltage is shown to the right. At temperatures of a few kelvins, the conductance shows steps in units of $2e^2/h$, which vanish at more

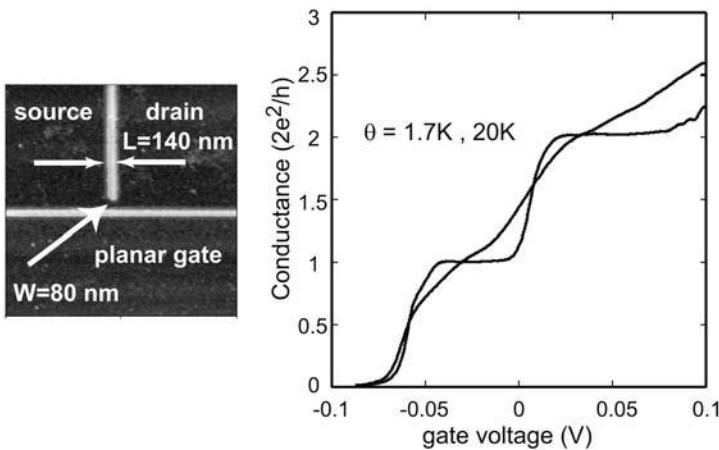


Fig. 1.2 Ballistic transport through a quantum point contact. To the left, the surface topography of a GaAs microchip is shown. The picture has been taken with an atomic force microscope. The chip hosts a quantum film about 30 nm below its surface, which is removed underneath the bright lines. A small and short wire of length 140 nm and

width 80 nm connects source and drain. By applying voltages to the planar gate electrode, the width of the wire can be tuned. The measurement to the right shows the conductance of the wire as a function of the gate voltage. At low temperatures, a conductance quantization in units of $2e^2/h$ is visible, which vanishes around 20 K.

elevated temperatures. This effect is one of the most fundamental observations [317, 326] in mesoscopic transport.

Where is the resistance and where does the voltage drop? After all, there are no scatterers. What determines the energy scale of thermal smearing? How do we model transport through ballistic samples in the first place? These issues are discussed in Chapter 7.

1.2.2

The quantum Hall effect and Shubnikov–de Haas oscillations

Fig. 1.3 shows the resistance of a homogeneous electronic quantum film along the direction of the current flow (the *longitudinal resistance* R_{xx}), as well as perpendicular to it (the *Hall resistance* R_{xy}). Apparently, the Hall resistance quantizes in units of $h/(je^2)$ in strong magnetic fields, and in units of $h/(2je^2)$ at smaller magnetic fields (j is an integer). This is the *quantum Hall effect*, to be discussed in Chapter 6. It was discovered by von Klitzing and coworkers [176]. Soon afterwards, it became clear that this quantization of the Hall resistance is independent of the material system, as long as the electron gas is two-dimensional. In 1982, these observations were supplemented by the discovery of the fractional quantum Hall effect by Tsui and coworkers [305]. This variation is observed only in samples with very high electron mobilities,

and has its origin in strong electron–electron interactions. We will not discuss the fractional quantum Hall effect in this book, though. It is tempting to suspect that the quantum Hall effect is somehow related to the conductance steps in quantum point contacts, which quantize in the same units. But how can this be? The sample size here is hundreds of micrometers, which is certainly larger than the mean free path. Second, the sample is two-dimensional. Also, we are now looking at the Hall resistance, while in the previous example, we looked at the two-terminal conductance ($G_{xx} + G_{xy}$ strictly speaking), and the magnetic field was zero. As we shall see in Chapter 7, there is in fact a close, although by no means obvious, relation between these two effects.

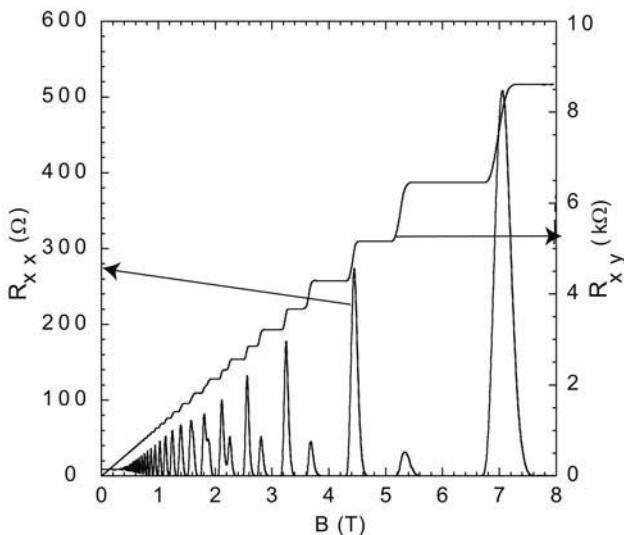


Fig. 1.3 Shubnikov–de Haas oscillations and the quantum Hall effect. We look at a measurement of the longitudinal and the Hall resistance (R_{xx} and R_{xy} , respectively) of a two-dimensional electron gas as a function of a magnetic field applied perpendicular to the plane of the electron gas. The experiment has been performed at a temperature of 100 mK.

Note that the behavior of R_{xx} is strongly correlated to the quantum Hall effect. We observe longitudinal resistance peaks at the steps in R_{xy} , while R_{xx} becomes zero in the regions of quantized Hall resistances. These oscillations are known as *Shubnikov–de Haas oscillations*. Any explanation for the quantum Hall effect should therefore also explain these oscillations, in particular the remarkable fact that the resistance vanishes! It should be remarked that the quantum film does *not* become superconductive. You may further wonder why the resistance of a diffusive two-dimensional electron gas can vanish, while that of a ballistic one-dimensional electron gas remains non-zero; in fact, it remains surprisingly large! It is an essential part of this book to answer

these questions, and to reveal their interconnections. For now, we leave it at the statement that, in quantum films placed in strong magnetic fields, the scattering of electrons is strongly suppressed, and the transport develops a one-dimensional character.

1.2.3

Size quantization

Particularly in modern semiconductor heterostructures (see Chapter 3 for more on this), the Fermi wavelength can become as large as 100 nm, and may thus be comparable to the size of the device. This strongly modifies the electronic density of states and changes the dimensionality of the electron system. Size quantization plays an essential role for the phenomena presented above. We will see that, in these semiconductor structures, many of the model potentials treated in elementary quantum mechanics can be tailored, such that we have some sort of a quantum mechanics construction kit at hand. We will meet, for example, parabolic quantum wells, square wells, or triangular potentials.

If you think about this, a non-trivial question probably springs to mind: the electrons are in a crystal, after all. The wave functions must obey Bloch's theorem. How is it that we can we speak of simple potentials and wave functions as encountered in elementary quantum mechanics? The answer is actually well established in elementary solid state theory and is most frequently used in relation to the potential and energies of doping atoms in semiconductors. It consists of the envelope function approximation and the concept of effective masses. The effects of the crystal are thereby taken into account by a dielectric constant, and by assigning an effective mass to the electron, which then moves in the superimposed potential. This approximation is used throughout this book, after its introduction in Chapter 2.

Size quantization and the corresponding change of the dimensionality (see Table 1.1) are already sufficient to change the properties of an electron gas profoundly. For example, the quantum Hall effect is absent in three-dimensional electron gases.

Tab. 1.1 Effect of size quantization on the electronic properties.

Dimension	Energy dependence of density of states	Unit of resistivity
3	$\propto \sqrt{E}$	$\Omega \text{ m}$
2	constant	Ω
1	$\propto 1/\sqrt{E}$	Ω/m
0	δ functions	n.a.

1.2.4

Phase coherence

When we speak of an electron, we mean a wave packet which certainly has some phase coherence length ℓ_ϕ . We expect interference effects of the electronic waves to play a role on length scales smaller than the phase coherence length. The phase coherence is destroyed by inelastic scattering events, such as electron–phonon scattering and electron–electron scattering, both of which depend strongly on temperature. At low temperatures, ℓ_ϕ may actually become as large as 100 μm . A prominent example of electronic interference is the Aharonov–Bohm effect in small quantum rings (Fig. 1.4).

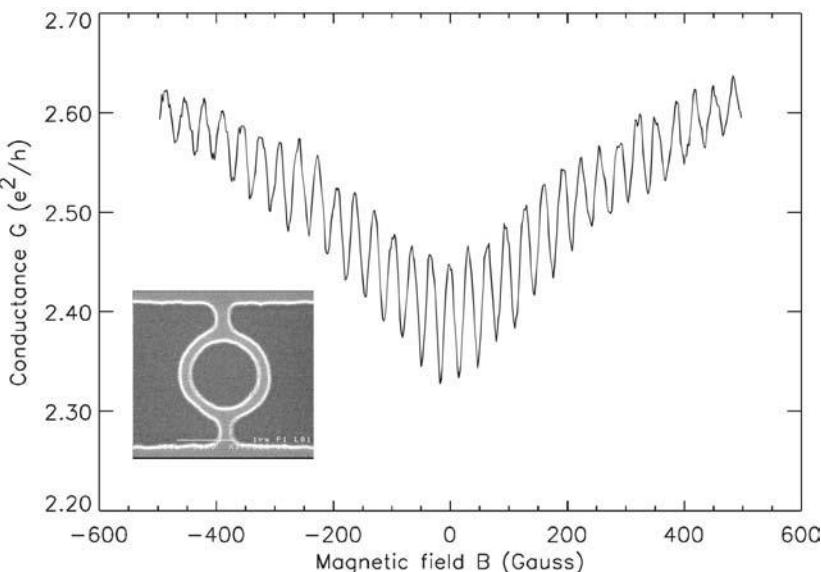


Fig. 1.4 The resistance of a small ring with a diameter of about 1 μm (the light gray areas in the inset) as a function of a magnetic field applied perpendicular to the ring plane shows periodic oscillations, known as Aharonov–Bohm oscillations. They indicate that a significant fraction of the electrons traverse the ring phase coherently. Taken from [236].

This will be explained in more detail in Chapter 8. An important consequence of phase coherence is that the resistance becomes a non-local quantity. Suppose we apply a current to a sample and measure the longitudinal voltage drop. This setup yields the longitudinal resistance R_{xx} . In conventional devices, it would just be the resistivity of the sample, multiplied by a geometrical factor. In a phase coherent sample, however, scattering events outside the probed region may influence the local electron density between the voltage probes. Just think of the increase of complexity in the operation of a circuit of transistors within the phase coherence length, which then mutually influ-

ence each other. It should be remarked that, on the other hand, the option of building electronic circuits in a phase coherent electron gas offers fascinating possibilities, which are outside the scope of this book, even though we outline the road from quantum dots to quantum bits and quantum computation in Chapter 10.

A phase coherent electron gas is not necessarily ballistic, since elastic scatterers do not cause dephasing. Situations can be established where the electronic dephasing is governed by electron–electron scattering, which does not show up (or only marginally so) in the resistance. This is the case because electron–electron scattering events do not modify the total momentum of the electron gas. We can therefore ask how phase coherent, diffusive systems behave. Such systems show in fact some interesting phenomena, which are presented in Chapter 8 as well.

1.2.5

Single-electron tunneling and quantum dots

Size reduction goes along with a reduction of capacitances. Consider a parallel plate capacitor of area L^2 at a separation L . Its capacitance C scales with L . At small sizes, the energy required to store an additional electron on it, $E = e^2/2C$, may become larger than the thermal energy. As a consequence, the quantization of charge can dominate the behavior of suitable circuits, in which tunneling of single electrons across leaky capacitors carries the current. This so-called single-electron tunneling can be used to design new types of devices, in particular the single-electron tunneling transistor. Probably, it is Gorter who deserves the credit for giving birth to the field of mesoscopic physics [123]. In 1951, he suggested that experiments by van Itterbeek and coworkers [164], who measured the current through metal grains embedded in an isolated matrix, could be explained by single-electron charging. The first transistor that exploited this effect was built by Fulton and Dolan in 1987 [109]. Fig. 1.5 shows an experimental realization of such a transistor in a semiconductor structure. We can call it a “transistor” since the gate voltage controls the current flowing between two further contacts. Single electron tunneling is a very important member of the family of mesoscopic effects and will be presented in Chapter 9. The structure shown in Fig. 1.5 actually represents also an example of a quantum dot. The electrons in the island are confined in all spatial directions, while their Fermi wavelength is comparable to the dot size. Quantum dots were “discovered” by chance during the investigations of disordered quantum wires, which segregated into small islands [269]. Soon afterwards, they were fabricated on purpose [209, 210]. The particular properties of such quantum dots are discussed in Chapter 10, where we will also explain the data shown here in somewhat more detail.

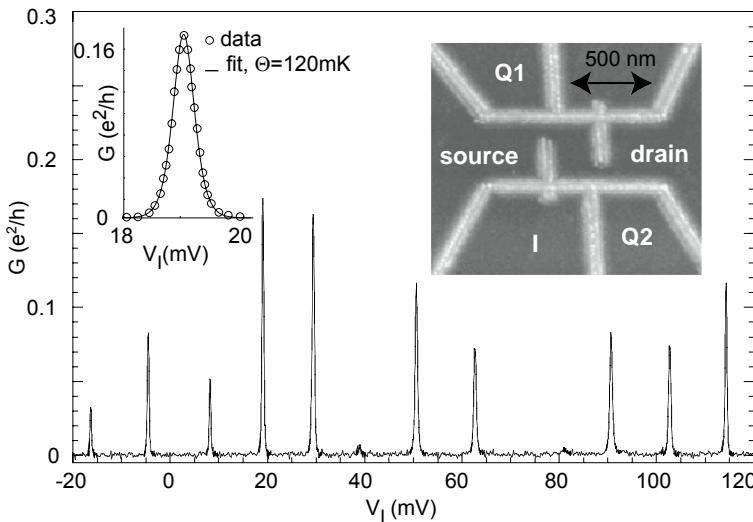


Fig. 1.5 The right inset shows again the surface topography of a semiconductor with a two-dimensional electron gas underneath. Here, the bright lines enclose a small island. It is coupled to source and drain via two quantum point contacts, which in this case are closed, i.e. they form tunnel barriers for the electrons. This can be achieved by adjusting the voltages applied to gates Q1 and Q2 accordingly. The main figure shows the

conductance through the island as a function of the gate voltage V_1 applied to region 1. Voltage V_1 tunes the potential of the island. The conductance peaks indicate that only for a particular island potential can electrons be transferred between the island and the leads. The left inset shows a fit to a function one would expect for peaks that are governed by thermal smearing of the Fermi function.

1.2.6 Superlattices

An interesting type of superpotentials are artificial crystals. They can be manufactured by patterning periodic structures on top of a semiconductor, followed by a transfer of the pattern into the electron gas. The resulting artificial lattices are either one- or two-dimensional and have been investigated in many experiments, following [315] and [32]. An alternative route is to grow layers of different semiconductor materials on top of each other [85]. In contrast to laterally patterned samples, these lattices are almost exclusively one-dimensional.

We can not only build model potentials this way, but also study the effects that occur in principle in periodic potentials, but remain inaccessible in natural crystals. Some prominent examples of such effects are treated in Chapter 11.

1.2.7

Spintronics

According to Fig. 1.1, the spin is irrelevant in conventional devices. The spin of course plays a decisive role in many properties of the electron system. Most importantly, it is the fermionic spin of $1/2$ that determines the electron statistics. We can only guess how our world would look if the charge carriers in solids had integer spin! Also, ferromagnetism is basically spin physics. Moreover, important semiconductor properties like the valence band structure depend in a very straightforward way on the spin. What we mean here, however, is that, once these basic implications of the electron spin have been taken into account, our device performance no longer depends on sample-specific spin properties. For example, the resistance is assumed to be independent of the spin coherence length or the spin polarization of the current.

Spintronics (spin electronics) summarizes all effects for which this is no longer true. For example, the resistance of a nanostructure that is smaller than the spin dephasing length ℓ_s can depend on the spin polarization of the injected current. The relevance of spin effects in commercial devices is increasing. Giant magneto-resistance read heads are standard in hard disks, while the concept of magnetic storage devices based on the tunneling magneto-resistance is a very elegant one. We will become familiar with the most important concepts of spintronics in Chapter 12.

1.2.8

Samples, experimental techniques, and technological relevance

You will almost certainly have noticed that the temperature has been quite low in all the examples. The highest temperature encountered so far was 20 K , at which the remarkable conductance quantization in Fig. 1.2 was no longer visible. Also, all the samples have been patterned semiconductors, so-called Ga[Al]As *heterostructures*, to be more precise. This seems to be a very narrow range of materials and temperatures. We live at room temperature, and the semiconductor industry makes its living from silicon. There is no doubt that these material systems and effects are fascinating from a purely scientific point of view. But are they really relevant for applications?

As far as the material is concerned, it is true that the Ga[Al]As system is sort of a workhorse for research in mesoscopic transport. Many groups work exclusively with Ga[Al]As heterostructures. This material is very versatile, and the electron gases can reach an almost incredible quality. For example, the electronic mean free path at low temperatures can exceed $100\text{ }\mu\text{m}$ at low temperatures. The foundation for achieving the corresponding ultra-high electron mobilities was laid by Dingle et al. in 1978, who invented a technique called *modulation doping* [72]. As we shall see, this technique allows spatial

separation of the doping ions from the mobile carriers in semiconductor heterostructures, and scattering is therefore greatly reduced. The details will be presented in Chapter 3. Silicon, however, is the material of choice for fabricating microprocessors. First of all, silicon is readily available in large quantities. A major advantage of Si is that it has a natural oxide with excellent mechanical and electronic properties. It is therefore easy to fabricate high-quality insulators on-chip. Also, the advantages of Ga[Al]As are particularly striking at low temperatures; they are smaller at room temperature, although still quite relevant. In fact, Ga[Al]As systems fill certain niches in the market. They are used in optoelectronics (which is outside the scope of this book), since they have a direct bandgap, in contrast to silicon. Also, Ga[Al]As is used in certain applications where high speed and low noise are essential. You sometimes hear that Ga[Al]As is, and will always remain, the material of the future. On the other hand, Si has played, and still plays, an important role in mesoscopic research as well. It is probably fair to state that transistor structures entered the field of mesoscopic transport in 1966, when it was observed that the electron gas in a *Si MOSFET* (metal–oxide–semiconductor field effect transistor) has in fact a two-dimensional character, due to size quantization at the interface between the silicon and its oxide [89]. The quantum Hall effect, for example, was discovered in a Si MOSFET [176]. In Chapters 2 and 3, we will therefore predominantly discuss these two material systems. However, other systems are by no means irrelevant in mesoscopic research! Each material has its particular strengths and weaknesses, and sometimes more exotic systems are best. For example, InAs offers an extremely high effective g -factor and a very small effective mass. Hole gases in SiGe, on the other hand, have large effective masses. The choice of the material thus often depends on the particular experiment one has in mind. It should furthermore be stressed that metallic nanostructures play a very important role in the field as well. Several seminal mesoscopic experiments have actually been performed in small metallic structures. For example, Aharonov–Bohm oscillations were observed in metal loops several years before they were seen in semiconductor rings. Also, the first single-electron tunneling transistor was made from aluminum. We will frequently meet metallic samples throughout the book.

Within the past few years, novel materials have moved to the focus of attention. One example is *carbon nanotubes* (see Fig. 1.6). They were discovered in 1991 by Iijima [160]. These rolled-up graphite sheets can be thought of as extremely small quantum wires. We will study some of their properties in Chapters 7 and 11. Furthermore, electronic quantum films can also be generated in organic polymers. These systems, which are among the potential “materials of the future”, are briefly presented in Chapter 3 as well. Finally, it should be said that, along with the advance of nanotechnology, transport experiments on single molecules have become possible. Conductance quan-

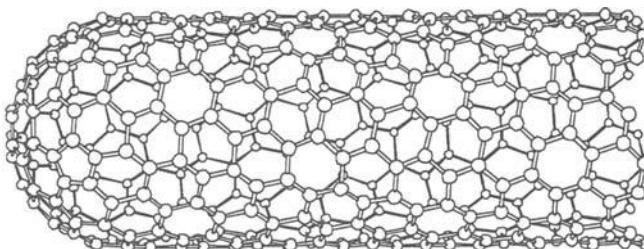


Fig. 1.6 Structure of a carbon nanotube. The circles denote carbon atoms in a graphite sheet, which is rolled up and forms a tube with a diameter of a few nanometers. The ends are supposedly capped by a carbon hemisphere. After [257].

tization through a single hydrogen molecule has been reported by Smit and coworkers [281], for example. We will occasionally mention such examples of *molecular electronics*. It should be pointed out that the concepts exemplified on semiconductor nanostructures can be transferred to molecules in a straightforward way.

Now, how about the relevance of the mesoscopic transport effects for applications? Well, there are already applications. For example, resistance quantization in the quantum Hall effect is so accurate that it is used as the resistance standard in many countries. Second, there is no fundamental reason why mesoscopic effects should not occur at room temperature. This is in contrast to superconductivity, for example, since the critical temperature is well below room temperature for all the superconductors known up to now. The temperature at which a mesoscopic effect vanishes, on the other hand, is essentially determined by the feature size. Just scale down your structure below, say, the phase coherence length or the mean free path, and you will see the mesoscopic behavior at room temperature. There are in fact several examples for mesoscopic behavior at room temperature, some of which we will meet later on. In many (but not all) cases, the samples are cooled down just because we are not yet able to pattern them at sufficiently small length scales. Phonons are major obstacles for the electronic motion, and often limit the mean free path at room temperature. Optical phonons have typical energies around 10 meV, and are thus frozen out below about 30 K. Also, the density of acoustic phonons is greatly reduced by cooling the samples.

Smaller feature sizes also mean stronger size quantization and larger energy separations between adjacent discrete energy levels. We can resolve the quantized structure as soon as the thermal smearing of the Fermi function becomes small compared to this energy level spacing. This is certainly the case for atoms at room temperature, but not for the artificial atom of Fig. 1.5, for example. Cooling the samples can therefore be regarded as a convenient way to look to the future, i.e. how the devices to come will behave at room tem-

perature once their size has been sufficiently reduced. Table 1.2 gives some typical length scales.

Tab. 1.2 The typical length scale at which the mesoscopic regime is reached depends on the temperature. The numbers just give an order of magnitude.

Temperature (K)	L (nm)
4.2 (liquid helium)	<5000
77 (liquid nitrogen)	<100
300 (room temperature)	<10

Even if we leave aside such extrapolations to the future, one important aspect of technological relevance remains. Joining the field of mesoscopic transport not only means that the student is going to work in an extremely rich and exciting field of research with many surprises and findings of fundamental relevance to be discovered. He/she will moreover learn a lot about the materials, processing techniques, and measurement concepts that are of utmost relevance in the present-day semiconductor industry with its world gross product of hundreds of billions of dollars. This combination is highly appreciated by the researchers in the field, since their range of career choices is unusually large.

Both the technology of patterning nanostructures as well as performing transport experiments at very low temperatures are very important issues. It is furthermore of great help to have an idea of experimental and technological boundary conditions to appreciate the measurements and the conditions under which they have been performed. Chapter 4, which deals with such issues, is therefore one of the central chapters.

A very limited selection had to be made for this book, but two missing topics should probably be singled out. The large and extremely active field of interacting electron gases (besides elementary screening and single-electron tunneling) has been left out. It concerns issues such as the fractional quantum Hall effect, coupled double layers of two-dimensional carrier gases, the metal-insulator transition in two dimensions, Luttinger liquids or Kondo correlations. Also, the fascinating topic of mesoscopic noise is not included.

2**An Update of Solid State Physics**

Mesoscopic systems are prepared from various materials, which are often, but not always, semiconductors. Some basic knowledge of their bulk properties is important and represents the major part of this chapter. Although this is in many respects just a polishing up of solid state physics at an introductory level, we introduce relevant specific properties of the materials of interest along the way, in particular of Si and GaAs. Occasionally, conventional metals and carbon crystals are mentioned as well.

We begin with a brief recapitulation of the most relevant crystal structures in Section 2.1, and proceed by looking at the corresponding electronic band structures of the materials in Section 2.2. Here, it is of particular importance to model the valence and conduction bands around their maximum and minimum, respectively. As always, we can approximate the energy dispersions near the band extremal points by parabolas, which leads to the concept of *effective masses*. We shall see that, within this approximation, the crystal properties can be “put aside” in many cases. Instead, the charge carriers behave like free electrons with a modified mass. The properties of electrons and holes within the effective mass approximation are looked at in Section 2.3. Also, the effective mass approximation allows us to work with *envelope wave functions*. Within this approach, superpotentials, like those frequently met in nanostructures, can be treated with a Schrödinger equation for just this superpotential. The crystal potential enters only via the effective masses and its dielectric constant. This is a very elegant concept, which simplifies our life substantially in subsequent chapters. This approximation is the topic of Section 2.4.

Doping is the standard way to fill the bands of a semiconductor with a significant and temperature-independent carrier density. The important issues concerning doping are reviewed in Section 2.5. In the subsequent section, we look at the transport properties of electron gases within the simplest version of the Boltzmann model. We will occasionally use these results when looking at diffusive samples. Furthermore, it is of help to know the approximations that enter this model, in order to appreciate the deviations we will look at later on. A non-vanishing resistance indicates that some type of scattering mech-

anism must be present, which is the topic of Section 2.7. Finally, we spend a few words on screening in Section 2.8.

Readers who discover that parts of this chapter are white spots on their map of solid state physics knowledge are encouraged to consult one of several excellent introductory textbooks for further information, like [12, 346]. If everything sounds familiar, please consider this chapter as a warm-up exercise!

2.1

Crystal structures

Many elements and compounds crystallize in a face centered cubic (fcc) lattice. This is not surprising, since this crystal structure represents one of the two possible realizations of close packings, which one might naively expect to occur when identical or very similar spheres are piled up. Both Si and GaAs have this lattice structure. The lattice constant a is the length of one edge of a unit cell. Si is composed of two fcc lattices shifted relative to each other by $(a/4, a/4, a/4)$. This crystal structure is also known as the diamond structure. GaAs also has a two-atom base, except that here one base atom is Ga and the other is As. This is the zinc blende lattice. The lattice constants are 0.565 nm for Si and 0.543 nm for GaAs (both numbers hold for room temperature). Fig. 2.1 shows the Si and the GaAs structures.

The reciprocal lattice of an fcc lattice is a body centered cubic (bcc) lattice. Since the crystal momentum is invariant under translations by reciprocal lat-

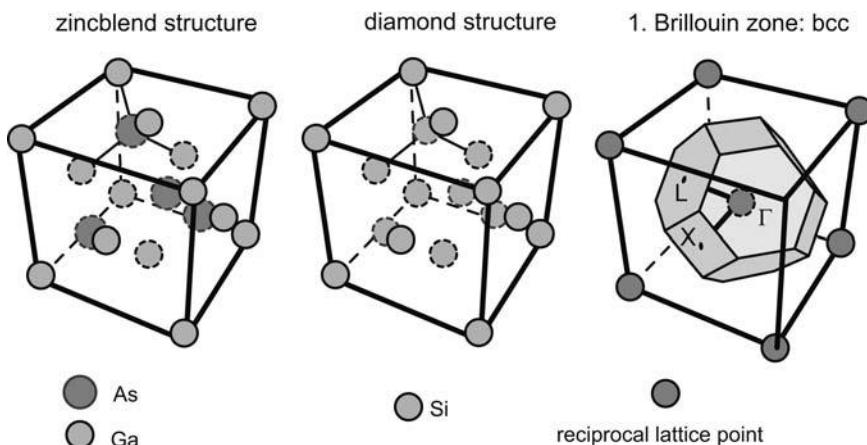


Fig. 2.1 Crystal structures of GaAs (left) and Si (center), as well as their first Brillouin zone (right), a truncated octahedron. Points of high symmetry are labeled as K, Γ and L; see text.

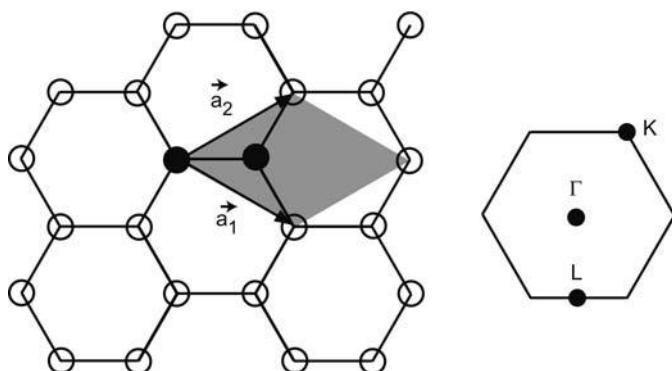


Fig. 2.2 Structure of a graphite sheet. Left: The unit cell (gray) of this hexagonal lattice is spanned by the lattice vectors $|\vec{a}_1|$ and $|\vec{a}_2|$, with a lattice constant of $|\vec{a}_1| = |\vec{a}_2| = 0.246 \text{ nm}$. It contains a basis of two carbon atoms (full circles) occupying non-equivalent sites. The distance between two neighboring atoms is 0.142 nm . Right: The first Brillouin zone of the graphite sheet with points of high symmetry.

lattice vectors, we can represent the behavior of electrons and phonons within one elementary cell of the reciprocal lattice, which is always chosen as the first Brillouin zone. For an fcc lattice, this is a truncated octahedron, composed of six squares and eight hexagons (see Fig. 2.1). The center of the first Brillouin zone is labeled the Γ -point, while the centers of the hexagons and squares are referred to as the L- and X-points, respectively. Occasionally, one hits upon more exotic directions of lower symmetry, such as K, U, and W, which are located at the center of the edges and at the corners of the first Brillouin zone.

Germanium crystallizes in a diamond structure like silicon. Any binary combination of Al, Ga, or In with As, Sb or P (the so-called III–V compounds) will result in a zinc blende lattice. Combining these group III elements with nitrogen can lead to either an fcc lattice or a hexagonal lattice, depending on the crystallization process and the subsequent treatment. This is also the case for most II–VI compounds, such as CdSe or ZnS (which gave the zinc blende structure its name, after all). Thus, when working with semiconductors, you will barely ever meet any further crystal structures. To finish this section, let us have a look at a particularly simple lattice, namely a sheet of graphite, the second crystal structure that carbon forms besides diamond. It consists of a hexagonal lattice of sp^2 -hybridized carbon atoms (Fig. 2.2).

Question 2.1: Calculate the reciprocal lattice of the graphite sheet and construct its first Brillouin zone.

The reciprocal lattice is again hexagonal. The center of the first Brillouin zone is denoted by Γ , the corners by K , and the centers of the edges are labeled L (Fig. 2.2).

2.2

Electronic energy bands

An electronic energy band is an energy interval in which electronic states exist in the crystal. The bands are separated by *bandgaps*. This energy structure is obtained by solving the Schrödinger equation for electrons in the crystal

$$\left[-\frac{\hbar^2}{2m}\Delta + V_{\text{crystal}}(\vec{r}) \right] |\phi(\vec{k}, \vec{r})\rangle = \epsilon |\phi(\vec{k}, \vec{r})\rangle \quad (2.1)$$

Here, the electronic wave functions depend on both the wave vector \vec{k} and the spatial coordinates \vec{r} . They are denoted by $|\phi(\vec{k}, \vec{r})\rangle$, while $V_{\text{crystal}}(\vec{r})$ is the crystal potential. Elementary solid state physics tells us that the wave functions have to obey Bloch's theorem, which states that they are of the form

$$|\phi(\vec{k}, \vec{r})\rangle = |u_{\vec{k}}(\vec{r})e^{i\vec{k}\vec{r}}\rangle \quad (2.2)$$

where $u_{\vec{k}}(\vec{r})$ has the periodicity of the crystal lattice. Such wave functions are *Bloch functions*. The task is to determine the eigenvalues $\epsilon(\vec{k})$ and eigenvectors, which is usually done by Fourier-transforming the differential equation into an algebraic equation. An exact solution, though, is only possible for some special cases. Some reasonable approximation is therefore called for. How Eq. (2.2) is then solved in detail depends on the model. The *nearly free electron model* starts from a free electron gas and treats a weak periodic crystal potential within perturbation theory. Here, the bandgaps emerge from interferences of the electronic waves that get scattered from the crystal potential, which results in standing waves at the edges of the Brillouin zones. The reader is referred to the extensive literature on solid state physics for details. Here, we look at a different approach, which constructs the electronic eigenstates from those of the individual atoms that form the crystal. This approach is known as the *tight binding model*. Within this picture, the energy bands and the bandgaps are remainders of the discrete energy spectrum of the atoms.

The tight binding model is based on the assumption that the atomic orbitals $|\phi_{a,n}(\vec{r})\rangle$ belonging to an energy eigenvalue E_n are a good starting point for constructing Bloch waves $|\xi_n(\vec{k}, \vec{r})\rangle$. Let us assume that there is only one atom per unit cell. We can define Bloch functions via

$$|\xi_n(\vec{k}, \vec{r})\rangle \equiv \frac{1}{\sqrt{N}} \sum_{\vec{R}_j} e^{i\vec{k}\vec{R}_j} |\phi_{a,n}(\vec{r} - \vec{R}_j)\rangle \quad (2.3)$$

Here, the lattice vectors are denoted by \vec{R}_j . The crystal wave functions can be expanded in these Bloch functions, such that

$$|\phi(\vec{k}, \vec{r})\rangle = \sum_n d_n(\vec{k}) |\xi_n(\vec{k}, \vec{r})\rangle \quad (2.4)$$

Suppose the energy level E_n is non-degenerate and there is no other energy level nearby. In that case, it is reasonable to assume that $|\phi(\vec{k}, \vec{r})\rangle \equiv |\phi\rangle = |\xi_n(\vec{k}, \vec{r})\rangle$ represents a very good approximation of the Bloch waves that emerge from the atomic wave functions $|\phi_{a,n}(\vec{r})\rangle$. The Schrödinger equation of the crystal for this scenario is now multiplied from the left by $\langle\phi|$, with the result

$$\langle\phi|H|\phi\rangle = \epsilon\langle\phi|\phi\rangle \quad (2.5)$$

Two integrals have to be evaluated: (i) $\langle\phi|\phi\rangle$ and (ii) $\langle\phi|H|\phi\rangle$. We write down (i) in the more explicit form

$$\langle\phi|\phi\rangle = \frac{1}{N} \sum_{l,j} \langle\phi_{a,n}(\vec{r} - \vec{R}_l)|\phi_{a,n}(\vec{r} - \vec{R}_j)\rangle e^{i\vec{k}(\vec{R}_j - \vec{R}_l)} \quad (2.6)$$

The integrals occurring in this sum are known as *overlap integrals*, since they measure the overlap of wave functions centered at the lattice points \vec{R}_l and \vec{R}_j . The atomic wave functions decay exponentially. It is therefore a reasonable approximation to neglect overlap integrals for two different sites, i.e.

$$\langle\phi_{a,n}(\vec{r} - \vec{R}_l)|\phi_{a,n}(\vec{r} - \vec{R}_j)\rangle \approx \delta_{lj} \quad (2.7)$$

It is actually here where the “tight” from the tight binding model enters, because Eq. (2.7) implies that the spatial extent of the atomic wave functions is small compared to the lattice constant. Thus, in integral (i), we are left with N identical terms equal to 1, since the atomic wave functions are normalized, and we end up with

$$\langle\phi|\phi\rangle = 1 \quad (2.8)$$

In integral (ii), let us first note that

$$\langle\phi_{a,n}(\vec{r} - \vec{R}_l)|H|\phi_{a,n}(\vec{r} - \vec{R}_j)\rangle$$

depends only on $\vec{R}_l - \vec{R}_j$. Therefore, the summation over l and j can be written as a summation over N terms only, where each term corresponds to the total contribution to one difference vector. We can thus choose $l \equiv 0$ and obtain

$$\langle\phi|H|\phi\rangle = \sum_j \langle\phi_{a,n}(\vec{r})|H|\phi_{a,n}(\vec{r} - \vec{R}_j)\rangle e^{i\vec{k}\vec{R}_j} \quad (2.9)$$

This equation can be simplified by splitting up the crystal Hamiltonian into the atomic Hamiltonian $H_a(\vec{r})$ (at $\vec{R}_l = 0$) and a contribution of all other atomic potentials

$$\delta V(\vec{r}) \equiv \sum_{m \neq 0} V(\vec{r} - \vec{R}_m) \quad (2.10)$$

In addition, we treat the terms with $j = 0$ separately, such that integral (ii) can be expressed as

$$\begin{aligned} \langle \phi | H | \phi \rangle &= E_n + \underbrace{\sum_{j \neq 0} \langle \phi_{a,n}(\vec{r}) | \phi_{a,n}(\vec{r} - \vec{R}_j) \rangle e^{i\vec{k}\vec{R}_j}}_{=0 \forall j, \text{Eq. (2.7)}} + \dots \\ &\quad + \underbrace{\langle \phi_{a,n}(\vec{r}) | \delta V(\vec{r}) | \phi_{a,n}(\vec{r}) \rangle}_{\equiv \beta} + \sum_{j \neq 0} \langle \phi_{a,n}(\vec{r}) | \delta V(\vec{r}) | \phi_{a,n}(\vec{r} - \vec{R}_j) \rangle e^{i\vec{k}\vec{R}_j} \end{aligned} \quad (2.11)$$

where we denote the value of the first integral in the second line as β , while the second integral is referred to as the *transfer integral*, which remains to be discussed. Note that each term in the transfer integral is actually a sum that runs over the contribution of all atomic potentials to that close to site $l = 0$:

$$\langle \phi_{a,n}(\vec{r}) | \delta V(\vec{r}) | \phi_{a,n}(\vec{r} - \vec{R}_j) \rangle = \sum_{m \neq 0} \langle \phi_{a,n}(\vec{r}) | V(\vec{r} - \vec{R}_m) | \phi_{a,n}(\vec{r} - \vec{R}_j) \rangle$$

Two final approximations are frequently made here. First of all, the influence of potentials from non-nearest neighbors is neglected. Second, terms with $m \neq j$ can be expected to be small compared to those with $m = j$, since they contain functions centered at three locations instead of two. Under these assumptions, the transfer integral reads

$$\langle \phi_{a,n}(\vec{r}) | V(\vec{r} - \vec{R}_j) | \phi_{a,n}(\vec{r} - \vec{R}_j) \rangle = \begin{cases} \gamma & j \text{ nearest neighbor to } 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.12)$$

The tight binding band emerging from the energy level E_n finally reads

$$E_n(\vec{k}) = E_n + \beta + \gamma \sum_{j \in \text{n.n.}} e^{i\vec{k}\vec{R}_j} \quad (2.13)$$

The energy for the possible wave vectors has developed a dispersion, which originates in the requirement that the wave functions have to obey the Bloch theorem. It is worth emphasizing that, besides the approximations already stated above, we have also assumed here that the atomic wave function has (at least) the symmetry of the lattice. Otherwise, γ would be anisotropic.

The tight binding approach is rather general in nature and still works in more complicated cases, which can be included by straightforward extensions. Moreover, in situations where the atomic wave functions are not tightly

bound, one can still construct Bloch waves from appropriately defined functions (known as Wannier functions) localized at each lattice site.

Question 2.2: Determine the energy dispersion for the simplest case, namely for a single band in one dimension, with a constant (and negative) transfer integral γ , and a vanishing overlap integral. Show that the energy dispersion in that case reads $E(k) = E_0 + 2\gamma \cos(ka)$.

As a more elaborate example, we mention the graphite sheet, in which atomic s and p orbitals generate the bands of relevance. Treating this system in terms of the tight binding model requires several extensions to the basic scheme outlined above. We have two atoms per lattice site and four atomic wave functions of different symmetry. The result of this calculation [313] is shown in Fig. 2.3. For the p_z orbitals of the carbon atoms arranged in a honeycomb configuration, a bonding and an antibonding π band results. To a first approximation, its tight binding energy dispersion is

$$E_\pi(\vec{k}) = \pm T \sqrt{1 + 4 \cos\left(\frac{1}{2}\sqrt{3}k_x a\right) \cos\left(\frac{1}{2}k_y a\right) + 4 \cos^2\left(\frac{1}{2}k_y a\right)} \quad (2.14)$$

Solids are usually classified as metals, semiconductors, or insulators. In a metal, at least one of the bands is partly occupied with electrons. These bands are called conduction bands in metals. In semiconductors and insulators, all bands are either full or empty at zero temperature. Here, the full band with the highest energy is the valence band, while the conduction band is the empty band with the lowest energy. In a semiconductor, a significant density of electrons can be transferred from the valence band into the conduction band by thermal excitation, which requires a bandgap of less than, say, 4 eV. Consequently, semiconductors are just small-bandgap insulators.

It turns out that the graphite sheet is a very special case in this classification scheme. The bonding π band is in fact the valence band, while its antibonding counterpart is the conduction band. As can be seen from Eq. (2.14), the valence band can be mapped onto the conduction band by a reflection at the planes defined by the K-points. The conduction band and the valence band of a graphite sheet, represented by bold lines in Fig. 2.3, touch each other at the K-points. It can thus be regarded as a semiconductor with zero bandgap.¹

By adopting the tight binding method appropriately, the band structure of other materials, like Si and GaAs, can be calculated. Naively, one might assume that, due to the similar crystal structures, the band structures of the two

1) In bulk graphite, the interaction between adjacent graphite sheets causes small energy shifts of both π bands, such that they overlap somewhat around the K-points. It is therefore a metal with an extremely small carrier density.

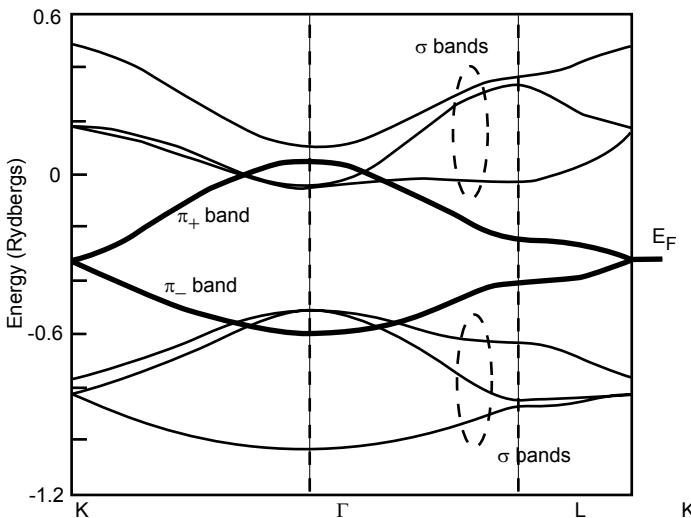


Fig. 2.3 Band structure of a graphite sheet. The valence band and conduction band touch each other at the K-points. After [228].

semiconductors should be very similar as well. However, this is not the case, mainly because the Ga–As base is polar, while the Si base is covalent. Fig. 2.4 shows the structures of the valence and conduction bands of both crystals. The extremal points of the bands shown here dominate both the electronic and optical properties. The number of electrons in the conduction band, as well as the number of holes in the valence band, is small compared to the number of available electronic states in all cases of relevance, and the few carriers will find themselves in close proximity to the band extremal points. Around these extremal points, we can expand the energy dispersion in a Taylor series up to second order:

$$E(\vec{k}) = E_0 + \frac{1}{2} \vec{k} \cdot \left(\frac{\partial^2 E}{\partial k_i \partial k_j} \right) \cdot \vec{k} \quad (2.15)$$

By comparing this expression with the energy dispersion of the free electron gas, $E(\vec{k}) = \hbar^2 \vec{k}^2 / 2m$, we see that the tensor of second derivatives of the energy can be identified with the effective masses

$$\frac{1}{\hbar^2} \left(\frac{\partial^2 E}{\partial k_i \partial k_j} \right) = \left(\frac{1}{m^*} \right)_{ij} \quad (2.16)$$

which is therefore also known as the *effective mass tensor*. It can be diagonalized, such that the extremal points of the energy bands can be characterized by three effective masses along the principal axes. Carriers in semiconductors therefore usually behave free-electron-like, except that their masses have been

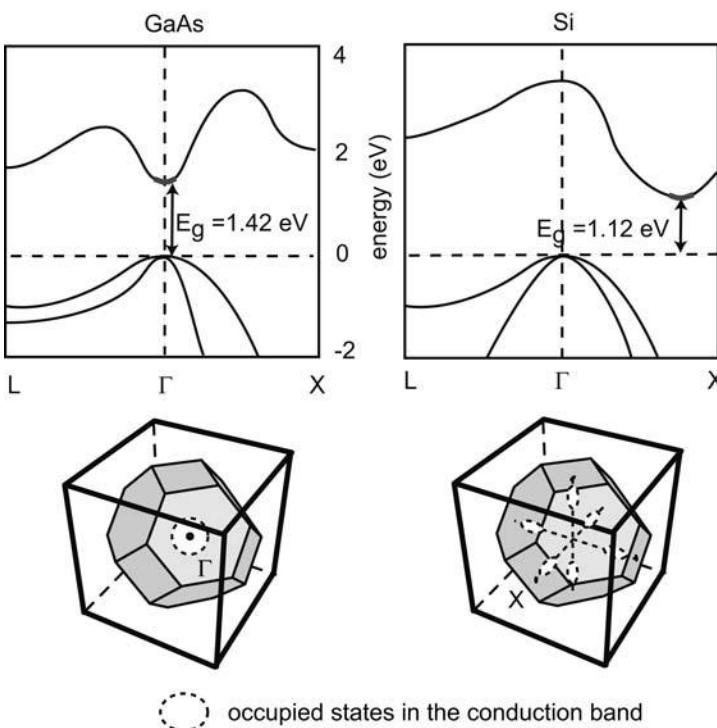


Fig. 2.4 Top: Electronic band structures of GaAs and Si close to the bandgap. Bottom: Schematic location and shape of the regions the electrons occupy at typical electron densities.

changed by the crystal structure. Throughout the rest of the book, we will use effective masses to describe the behavior of carriers.

Question 2.3: What is the effective mass around the minimum of the energy band obtained in Question 2.2?

Let us have a somewhat closer look at these band structures. Si has a conduction band minimum at $0.85\overrightarrow{\Gamma X}$. Around the conduction band minimum, two different effective masses exist, a transverse mass in all directions perpendicular to the $\overrightarrow{\Gamma X}$ direction, $m_{e,t} = 0.19m$, and a longitudinal mass along the $\overrightarrow{\Gamma X}$ direction, $m_{e,l} = 0.92m$. Since there are six X-points, the conduction band minimum in Si shows a sixfold degeneracy known as “valley degeneracy”. In GaAs, the conduction band minimum is located at the Γ -point. Here, the three effective electron masses are identical: $m_{e,1}^*(\text{GaAs}) = m_{e,2}^*(\text{GaAs}) = m_{e,3}^*(\text{GaAs}) = 0.067m$. In both materials, there are two (nearly) degenerate va-

lence bands at the Γ -point. As in most semiconductors of interest, the valence band emerges from atomic p states, which have a threefold orbital degeneracy and a spin degeneracy of 2. Typically, the corresponding σ band formed by the atomic s orbitals has its maximum well below the maximum of the p bands and does not have to be taken into account for transport considerations. In the crystal, the degeneracy of the p orbitals is removed, and three different, spin-degenerate bands are obtained. Two of them are shown in Fig. 2.4, while the third one is split off and shifted to lower energies. This splitting has its origin in the spin-orbit interaction. The spin-orbit Hamiltonian is given by

$$H_{\text{so}} = \frac{\hbar}{4m^2c^2} \underline{\sigma} \cdot \vec{\nabla} V \times \vec{p} \quad (2.17)$$

where V is the electrostatic potential, and $\underline{\sigma}$ are the Pauli spin matrices. This is a relativistic term, which means we have to replace the Schrödinger equation by the Dirac equation, and the wave function becomes a two-component spinor. In a spherical symmetric potential, the spin-orbit Hamiltonian becomes proportional to the scalar product of the angular momentum and the spin $\vec{L} \cdot \vec{S}$. To get an idea what the spin-orbit Hamiltonian does to the energies, we assume that the interaction in the solid can be approximated by that in the individual atoms. It is then clear from atomic physics that this term separates the fourfold degenerate $j = 3/2$ states from the twofold degenerate $j = 1/2$ states, where j denotes the total angular momentum quantum number. The $j = 1/2$ state is lowered in energy, by an amount that essentially depends on the strength of the atomic Coulomb potential. The heavier the nucleus, the stronger is this spin-orbit splitting Δ_{so} . This tendency can be seen experimentally: $\Delta_{\text{so}}(\text{graphite}) \approx 6 \text{ meV}$, $\Delta_{\text{so}}(\text{Si}) \approx 45 \text{ meV}$, and $\Delta_{\text{so}}(\text{GaAs}) \approx 340 \text{ meV}$.

The energy dispersions of the remaining four bands with $j = 3/2$ can be conveniently described within the $\vec{k} \cdot \vec{p}$ approximation, a method to model the dispersion around the extremal points of an energy band. We consider a semiconductor with a band maximum at $\vec{k} = 0$, as is the case for the valence bands under study. Within the $\vec{k} \cdot \vec{p}$ model, the spatial derivatives in the Schrödinger equation of the crystal, Eq. (2.2), are carried out only for the plane wave component of the Bloch function of the type (2.2). The equation

$$\left\{ \frac{p^2}{2m} + \frac{\hbar \vec{k} \cdot \vec{p}}{m} + \frac{\hbar \vec{k}^2}{2m} + V(\vec{r}) \right\} u_{n,\vec{k}}(\vec{r}) = E_n(\vec{k}) u_{n,\vec{k}}(\vec{r}) \quad (2.18)$$

emerges. Here, n denotes the band index. For $\vec{k} = 0$, it simplifies significantly, and we assume that an approximate solution can be found for all bands involved. A non-vanishing but small wave vector can then be treated as a perturbation.

First of all, the term $\propto \vec{k}^2$ produces an energy shift that depends on \vec{k} , but does not couple the bands. Technically, it can just be added to the crystal

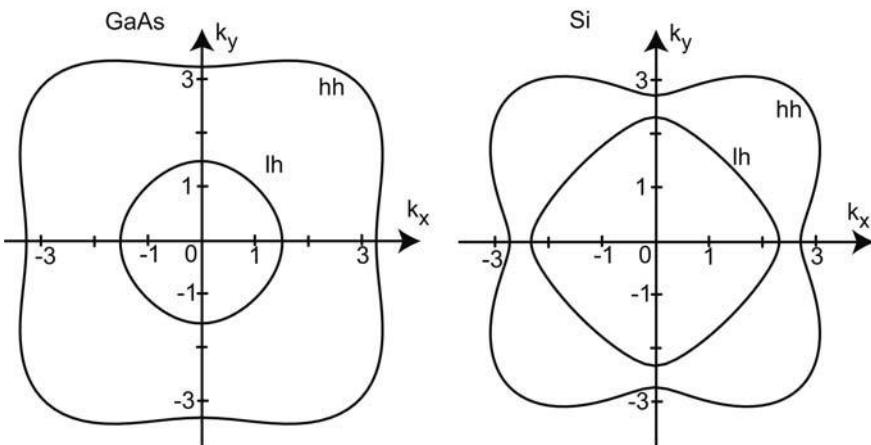


Fig. 2.5 Warped surfaces of constant energy for heavy and light hole bands in GaAs (left) and Si (right). A cross section through the plane $k_z = 0$ is shown for a typical Fermi energy of 10 meV. The wave numbers are measured in units of 10^8 m^{-1} .

potential. The term containing $\vec{k} \cdot \vec{p}$, however, must be treated with degenerate perturbation theory. It turns out that the second-order term is the leading one, since the first-order term is linear in \vec{k} and must thus vanish at a maximum. The matrix elements are given by

$$h_{ij}(\vec{k}) = \frac{\hbar^2}{m^2} \sum_{q=1; q \neq i,j}^4 \frac{\langle n_i, 0 | \vec{k} \cdot \vec{p} | n_q, 0 \rangle \langle n_q, 0 | \vec{k} \cdot \vec{p} | n_j, 0 \rangle}{\epsilon_{n_i, 0} - \epsilon_{n_q, 0}} \quad (2.19)$$

The bands are labeled by $n_{i,j,q}$ here. This 4×4 matrix equation gives energy eigenvalues of the type

$$E_{\text{lh},\text{hh}} = \frac{\hbar^2}{2m} \left[\gamma_1 k^2 \pm \sqrt{4\gamma_2^2 k^4 + 12(\gamma_3^2 - \gamma_2^2)(k_x^2 k_y^2 + k_y^2 k_z^2 + k_z^2 k_x^2)} \right] \quad (2.20)$$

The γ_i are the Luttinger parameters, which depend on the material. For GaAs, $\gamma_1 = 6.95$, $\gamma_2 = 2.25$, and $\gamma_3 = 2.86$. In Si, $\gamma_1 = 4.29$, $\gamma_2 = 0.34$, and $\gamma_3 = 1.42$. The “+” energy dispersion corresponds to a lighter effective mass for all directions in \vec{k} -space. The band is therefore referred to as the *light hole (lh) band*. Correspondingly, the “−” sign represents the energy dispersion for the *heavy hole (hh) band*. To get an idea of the shape of the hole bands, consider a surface of constant energy. The first term on the right-hand side in Eq. (2.20) describes a sphere, which is warped by the second term. The warping is \vec{k} -dependent and of opposite sign in the two bands for all directions. These surfaces are therefore known as *warped spheres* (see Fig. 2.5).

Note that both bands remain twofold degenerate in this treatment at $\vec{k} = 0$. This is known as the Kramers degeneracy. It is removed in polar crystals, such

as GaAs or InP, due to the absence of an inversion center. The corresponding correction to the Hamiltonian is known as the Dresselhaus term. The resulting energy splitting, however, is small, i.e. in the range of μeV , although it causes measurable effects to occur at very low temperatures.

As per the definition, the properties of a hole (characterized by energy E_h and wave vector \vec{k}_h) are those of a fully occupied band with the corresponding electron, i.e. the electron of energy $-E_h$ and wave vector $-\vec{k}_h$, removed. A hole is thus a quasi-particle with a positive effective mass and a positive charge of $q = +e$. For the warped structure of the valence bands discussed above, it is common to specify effective masses for the hh and the lh bands by evaluating the band structure for very small wave vectors and averaging them over all directions in reciprocal space. The literature values vary somewhat; typical values are $m_{hh}^*(\text{Si}) = 0.54m$, $m_{lh}^*(\text{Si}) = 0.15m$, $m_{hh}^*(\text{GaAs}) = 0.51m$, and $m_{lh}^*(\text{GaAs}) = 0.08m$.

GaAs has a direct bandgap, meaning that the minimum in the conduction band is at the same location in k -space as the maximum of the valence band. The bandgap of Si is indirect. A large momentum transfer is necessary for exciting electrons from the valence band maximum into the conduction band minimum.

Question 2.4: Compare the momentum of a photon with the energy of the Si bandgap with the momentum difference between the Γ -point and the conduction band minimum in Si.

Therefore, Si can absorb photons with an energy close to the bandgap only if phonons are absorbed/emitted simultaneously. This is a rather unlikely process, which makes crystalline Si a poor material for optoelectronics.

Owing to anharmonic contributions to the lattice vibrations, the crystals shrink as they get cooled down. As a consequence, the bandgap increases with decreasing temperature. Empirically, one finds (see Fig. 2.6)

$$E_{g,\text{Si}}(\Theta) = 1.17 \text{ eV} - \frac{4.73 \times 10^{-4} \Theta^2 \text{ K}^{-1}}{\Theta + 636 \text{ K}} \text{ eV}$$

$$E_{g,\text{GaAs}}(\Theta) = 1.52 \text{ eV} - \frac{5.4 \times 10^{-4} \Theta^2 \text{ K}^{-1}}{\Theta + 204 \text{ K}} \text{ eV}$$

For many applications, a fraction x of the Ga atoms in GaAs is replaced by Al atoms, and the ternary $\text{Al}_x\text{Ga}_{1-x}\text{As}$ results. For $x \leq 0.38$, the bandgap increases linearly with x , with a maximum at $E_g(\Gamma, x=0.38) = 1.92 \text{ eV}$, and can be tailored for a specific application. For $x \geq 0.38$, however, the local minimum close to the X-point becomes the global minimum of the conduction band, and the material becomes an indirect semiconductor. Pure AlAs

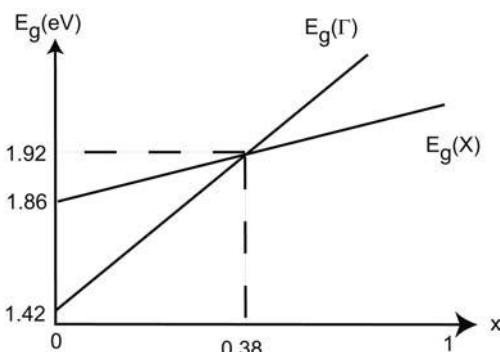


Fig. 2.6 Sketch of the bandgap in $\text{Al}_x\text{Ga}_{1-x}\text{As}$ as a function of the Al concentration x .

has a bandgap of $E_g = 2.16$ eV. Note that the positions of the Al atoms are random, which means that the ternary compound is *not* a crystal. Nevertheless, we can speak of band structures and effective masses, since such crystals can be treated within an averaging procedure known as the “virtual crystal approximation” (see [21]).

2.3

Occupation of energy bands

In this section, we study how the electrons occupy the valence and conduction bands. The electron density n in a band is obtained by integrating over the spectral electron density $n(E)$, i.e. the density of electrons in the interval $[E, E + dE]$. The spectral electron density is given by the spectral density of electronic states $D_d(E)$ available, multiplied by their occupation probability. Here, the index d denotes the dimensionality of the system. We will briefly discuss these two quantities.

2.3.1

The electronic density of states

The electronic density of states $D_d(E)$ is the number of electronic states in $[E, E + dE]$ and per unit volume. It depends on the dimensionality d of the system and the energy dispersion $E(\vec{k})$ of the electronic band under consideration. The usual way to calculate $D_d(E)$ is to determine the electronic mode density in k -space $D_d(\vec{k})$ of a cavity of size L^d and transform it into the energy space via $E(\vec{k})$. We carry out this calculation for a two-dimensional system with a parabolic energy dispersion, since this is what we will encounter most frequently in the following.

Consider a crystal square with a base length L , oriented along the x - and y -axes. We assume periodic boundary conditions, and use plane waves as base functions.² An electronic state Ψ exists at wave vector \vec{k} if

$$\Psi(\vec{r} + (L, L)) = \Psi(\vec{r}) \quad \Rightarrow \quad \vec{k} = \frac{2\pi}{L}(n_x, n_y)$$

with n_i being an integer. The allowed wave vectors form a simple square lattice in k -space with a lattice constant of $2\pi/L$. Each state is g -fold degenerate due to spin and valley degeneracies. Hence, there are g states in the volume $(2\pi/L)^2$. States of equal $|\vec{k}|$ are located on a circle. The number of states dN_2 in an annulus of radius k and width dk is given by

$$dN_2 = g \frac{2\pi k}{(2\pi/L)^2} dk$$

with $k = |\vec{k}|$. This gives a density of states in k -space of

$$D_2(k) = \frac{1}{L^2} \frac{dN}{dk} = \frac{gk}{2\pi}$$

$D_2(E)$ is obtained from $D_2(k)$ by a coordinate transformation

$$D_2(E) = D_2(k) \frac{dk}{dE} = \frac{gm^*}{2\pi\hbar} \quad (2.21)$$

Here, we have used the energy dispersion for electrons with an isotropic effective mass m^* ,

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

The density of states in one, two, and three dimensions are shown in Fig. 2.7.

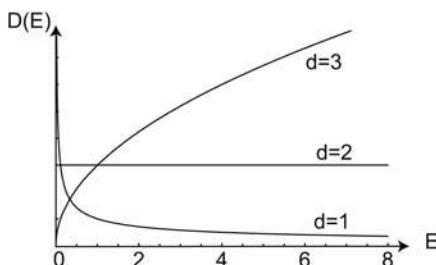


Fig. 2.7 The electronic density of states within the effective mass approximation as a function of energy, in one, two, and three dimensions.

- 2) It can be shown that the results do not depend on the boundary conditions.

Question 2.5: Calculate $D_3(E)$ and $D_1(E)$. Show that

$$D_3(E) = g \frac{(2m^*)^{3/2}}{4\pi^2 \hbar^3} \sqrt{E} \quad (2.22)$$

and

$$D_1(E) = g \frac{\sqrt{2m^*}}{2\pi\hbar} \frac{1}{\sqrt{E}} \quad (2.23)$$

What does the density of states look like for a zero-dimensional system?

2.3.2

Occupation probability and chemical potential

In equilibrium, fermions occupy states of energy E with a probability given by the Fermi–Dirac distribution function

$$f(E, \Theta) = \frac{1}{e^{(E-\mu)/k_B\Theta} + 1} \quad (2.24)$$

Here, μ denotes the chemical potential, i.e. the energy for which the density of occupied states with larger energies equals the density of empty states with lower energies. This definition, by the way, also holds if the occupation probability is not a Fermi–Dirac distribution. Furthermore, Θ is the temperature. The *Fermi energy* E_F is the energy at which $f(E, \Theta=0)$ jumps from 1 to 0. Clearly, $\mu = E_F$ at $\Theta = 0$. For $\Theta > 0$, μ may differ from E_F , depending on the energy dependence of the density of states.

In a metal, at least one band is by definition partly occupied at $\Theta = 0$. Therefore, E_F is located within an energy band. Semiconductors, on the other hand, are defined as crystals where the conduction band is empty, and E_F thus resides in the bandgap. The same is of course true for insulators. The electron density in a band ranging from E_{bottom} to E_{top} is obtained from

$$n = \int_{E_{\text{bottom}}}^{E_{\text{top}}} n(E) dE = \int_{E_{\text{bottom}}}^{E_{\text{top}}} D_d(E) f(E, \Theta) dE \quad (2.25)$$

Note that the dimensionality d of $D_d(E)$ also determines the dimensionality of n . The Fermi function and the spectral electron density are sketched in Fig. 2.8.

2.3.3

Intrinsic carrier concentration

The carrier concentration in a perfect, impurity-free semiconductor crystal is called “intrinsic”. Here, the carriers are exclusively generated by thermal excitation of electrons from the valence band into the conduction band, which

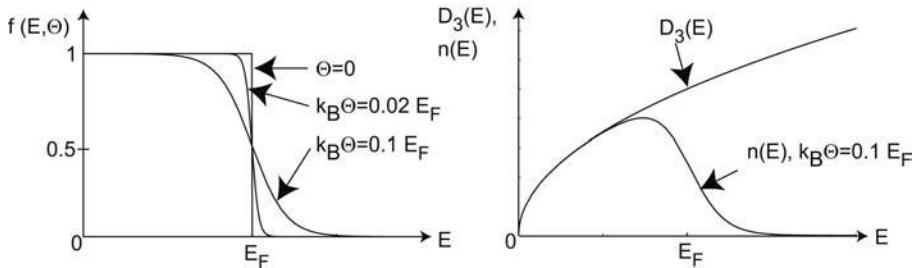


Fig. 2.8 Thermal smearing of the Fermi function (left), and the density of states ($d = 3$) as well as the spectral carrier density $n(E)$ (right).

means that $n = p$ (p denotes the hole density). Within the effective mass approximation and for a spin-degenerate system ($g = 2$), the carrier densities are given by

$$\begin{aligned} n &= \frac{\sqrt{2}m_e^{*3/2}}{\pi^2\hbar^3} \int_{E_C}^{\infty} \sqrt{(E - E_C)} f(E, \Theta) dE \\ p &= \frac{\sqrt{2}m_h^{*3/2}}{\pi^2\hbar^3} \int_{-\infty}^{E_V} \sqrt{(E_V - E)} [1 - f(E, \Theta)] dE \end{aligned} \quad (2.26)$$

The chemical potential is close to the center of the bandgap, slightly shifted toward the band with the lighter effective mass.³ Therefore, it is safe to assume that $|E_{C,V} - \mu| \gg k_B\Theta$. This tells us that only the tails of the Fermi function, far away from the chemical potential, lie inside the bands, and can be well approximated by a Boltzmann distribution, i.e.

$$f(E, \Theta) = \exp[-(E - \mu)/k_B\Theta]$$

A brief calculation gives

$$n = N_C e^{-(E_C - \mu)/k_B\Theta}, \quad p = P_V e^{(E_V - \mu)/k_B\Theta} \quad (2.27)$$

N_C and P_V are known as “effective densities of state”, and are given by

$$N_C = \frac{1}{4} \left(\frac{2m_e^* k_B \Theta}{\pi \hbar^2} \right)^{3/2}, \quad P_V = \frac{1}{4} \left(\frac{2m_h^* k_B \Theta}{\pi \hbar^2} \right)^{3/2}$$

An immediate consequence is the “law of mass action for charge carriers”

$$np = N_C P_V e^{-E_g/k_B\Theta} \implies n = p = \sqrt{N_C P_V} e^{-E_g/2k_B\Theta}$$

- 3) This is qualitatively clear as μ is given by the condition $n = p$, and the density of states increases with increasing effective mass.

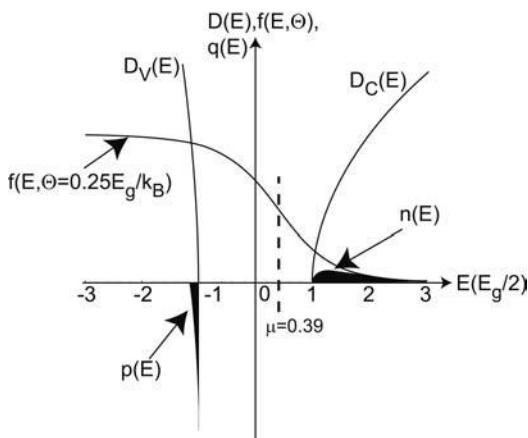


Fig. 2.9 Fermi function, density of states, and spectral carrier densities $q(E)$ ($q = n, p$) in an intrinsic semiconductor with $m_h^* = 2m_e^*$. The temperature is $\Theta = 0.25E_g/k_B$, which corresponds to a chemical potential $\mu = 0.39E_g/2$.

Inserting p in Eq. (2.19) leads to

$$\mu = E_V + \frac{1}{2}E_g + \frac{3}{4}k_B\Theta \ln(m_h^*/m_e^*) \quad (2.28)$$

Fig. 2.9 summarizes the relations between density of states, Fermi function and intrinsic carrier densities in a semiconductor.

The exponential dependence of n and p on temperature causes *carrier freeze-out* as the temperature is reduced. At room temperature, we have an intrinsic electron density of $n_{Si} = 1.45 \times 10^{16} \text{ m}^{-3}$ for silicon and $n_{GaAs} = 1.8 \times 10^{12} \text{ m}^{-3}$ for GaAs (see Exercise E2.7).

At these typical, small carrier densities, the electron Fermi surface consists of six rotational ellipsoids in Si, of spheres in GaAs, as indicated in Fig. 2.4. In the valence band, the warped surfaces in the previous section represent Fermi spheres.

2.3.4

Bloch waves and localized electrons

By this time, you may be wondering how Bloch waves, which are extended over the whole crystal, relate to the conventional picture of an electron of momentum \vec{k}_e , that at the time t can be found at position \vec{r}_e , moving with a velocity \vec{v}_e through the crystal. In order to localize a particle, one has to build a localized wave packet from the Bloch waves. The time dependence enters via the solutions of the time-dependent version of the Schrödinger equation (2.1), i.e.

$$|\phi(\vec{k}, \vec{r}, t)\rangle = |\phi(\vec{k}, \vec{r})\rangle e^{-iE(\vec{k})t/\hbar} \quad (2.29)$$

The electronic wave packet is constructed by

$$|\Phi_e(\vec{k}_e, \vec{r}_e, t)\rangle = \int_{-\infty}^{\infty} w(\vec{k} - \vec{k}_e) |\phi(\vec{k}, \vec{r}, t)\rangle d\vec{k} \quad (2.30)$$

where $w(\vec{k} - \vec{k}_e) \equiv \delta\vec{k}$ is a weight function sharply peaked – on the scale of the extension of the Brillouin zone – at \vec{k}_e . An expansion of this integral in $\delta\vec{k}$ around \vec{k}_e to first order shows that Eq. (2.30) represents the time-independent Bloch wave for \vec{k}_e , strongly modulated by a window function located at (\vec{r}_e, t) . Furthermore, the window function moves through the crystal with velocity

$$\vec{v}_e \equiv \frac{1}{\hbar} \vec{\nabla}_{\vec{k}} E(\vec{k})|_{\vec{k}_e} \quad (2.31)$$

This construction implies that we can only speak of an electron as a localized particle if there are sufficiently many Bloch waves available. This is not necessarily the case in nanostructures. Here, the localized electron picture has to be used with care. Furthermore, a consequence of the sharply peaked character of the weight function is that the spatial extension of such an electron wave packet, i.e. the de Broglie wavelength, is always larger than the lattice constant. The details related to this picture are the topic of Exercise E2.4.

2.4

Envelope wave functions

So far, the materials have been homogeneous. Real crystals are certainly not perfect. Their translational symmetry can be perturbed, either by e.g. unwanted lattice imperfections, or by intentionally built-in superpotentials. We will frequently see such superpotentials later on. How do the wave functions and energy levels in such a perturbed crystal look?

Consider a lattice imperfection with the perturbation potential $V_p(\vec{r})$. For simplicity, we take only one electronic band into account. The Schrödinger equation for the imperfect crystal reads

$$\left[-\frac{\hbar^2}{2m} \Delta + V_{lattice}(\vec{r}) + V_p(\vec{r}) \right] \Phi(\vec{r}) = E \Phi(\vec{r}) \quad (2.32)$$

The solution $\Phi(\vec{r})$ is no longer a Bloch function, but it can be expanded in the Bloch wave functions of the unperturbed band

$$\Phi(\vec{r}) = \sum_{\vec{k}'} c_{\vec{k}'} \xi(\vec{k}', \vec{r}) \quad (2.33)$$

Inserting this expansion into Eq. (2.21), multiplying by $\xi^*(\vec{k}, \vec{r})$, and integrating over the whole crystal gives

$$\epsilon(\vec{k})c_{\vec{k}} + \sum_{\vec{k}'} c_{\vec{k}'} a(\vec{k}, \vec{k}') = Ec_{\vec{k}} \quad (2.34)$$

with the matrix elements

$$a(\vec{k}, \vec{k}') = \langle \xi(\vec{k}, \vec{r}) | V_p(\vec{r}) | \xi(\vec{k}', \vec{r}) \rangle \quad (2.35)$$

We plan to rewrite Eq. (2.34) in the form of a Schrödinger equation with a newly defined wave function, which will be the envelope function. This can be done by making two approximations, namely (i) $V_p(\vec{r})$ varies slowly on the scale of the lattice constant, and (ii) the effective mass approximation.

Our first task is finding an appropriate expression for $a(\vec{k}, \vec{k}')$. We have assumed that $V_p(\vec{r})$ varies slowly on the scale of individual unit cells. This means that we can keep $V_p(\vec{r})$ constant within each cell, which is referred to by the corresponding lattice vector \vec{R} . In order to use this in Eq. (2.35), we split the integral, which runs over the whole crystal, into integrals running over unit cells, and sum them up:

$$\begin{aligned} a(\vec{k}, \vec{k}') &= \sum_{\vec{R}} \int_{\text{cell } \vec{R}} \xi^*(\vec{k}, \vec{r}) V_p(\vec{r}) \xi(\vec{k}', \vec{r}) d\vec{r} \\ &= \sum_{\vec{R}} V_p(\vec{R}) \int_{\text{cell } \vec{R}} \xi^*(\vec{k}, \vec{r}) \xi(\vec{k}', \vec{r}) d\vec{r} \end{aligned} \quad (2.36)$$

Since $\xi(\vec{k}, \vec{r})$ is of the form given by Eq. (2.2), i.e. $\xi(\vec{k}, \vec{r}) = u_{\vec{k}}(\vec{r}) e^{i\vec{k}\vec{r}}$, the cell integral can be written as

$$\int_{\text{cell } \vec{R}} u_{\vec{k}}^*(\vec{R} + \vec{r}) u_{\vec{k}'}(\vec{R} + \vec{r}) e^{i(\vec{k}' - \vec{k})(\vec{R} + \vec{r})} d\vec{r} \quad (2.37)$$

The function $u_{\vec{k}}^*(\vec{r}) u_{\vec{k}'}(\vec{r})$ has the periodicity of the lattice and can thus, according to the Fourier theorem, be expanded in harmonic functions with the same periodicity:

$$u_{\vec{k}}^*(\vec{r}) u_{\vec{k}'}(\vec{r}) = \sum_{\vec{G}} \alpha(\vec{G}) e^{i\vec{G}\vec{r}}, \quad \alpha(\vec{G}) = \frac{1}{V} \int_V u_{\vec{k}}^*(\vec{r}) u_{\vec{k}'}(\vec{r}) e^{-i\vec{G}\vec{r}} d\vec{r} \quad (2.38)$$

where \vec{G} is a reciprocal lattice vector. With the Fourier expansion inserted in Eq. (2.36), we obtain

$$a(\vec{k}, \vec{k}') = \sum_{\vec{G}, \vec{R}} \alpha(\vec{G}) V_p(\vec{R}) \int_{\text{cell } \vec{R}} e^{i(\vec{k}' - \vec{k})(\vec{R} + \vec{r})} e^{i\vec{G}(\vec{R} + \vec{r})} d\vec{r} \quad (2.39)$$

which can be simplified considerably. First of all, $e^{i\vec{G}\vec{R}} = 1$. Second, since $V_p(\vec{r})$ varies smoothly, only Bloch waves within a narrow interval of \vec{k} -vectors will contribute to $a(\vec{k}, \vec{k}')$, and we can assume that $\vec{k}' - \vec{k}$ is small on the scale of the smallest reciprocal lattice vector. Therefore, it is justified to approximate $e^{i(\vec{k}' - \vec{k})(\vec{R} + \vec{r})} \approx e^{i(\vec{k}' - \vec{k})\vec{R}}$. After taking these considerations into account, Eq. (2.39) reads

$$a(\vec{k}, \vec{k}') \approx \sum_{\vec{G}, \vec{R}} \alpha(\vec{G}) V_p(\vec{R}) e^{i(\vec{k}' - \vec{k})\vec{R}} \int_{\text{cell } \vec{R}} e^{i\vec{G}\vec{r}} d\vec{r}$$

In addition, Green's theorem for functions with the periodicity of the lattice [12] tells us that

$$\int_{\text{cell } \vec{R}} e^{i\vec{G}\vec{r}} d\vec{r} = V_{\text{cell}} \delta_{\vec{G}, 0} \quad (2.40)$$

where V_{cell} is the volume of the unit cell.

Question 2.6: Prove Eq. (2.40) for a one-dimensional crystal.

With Eq. (2.40), we obtain

$$a(\vec{k}, \vec{k}') = \alpha(0) \sum_{\vec{R}} V_p(\vec{R}) V_{\text{cell}} e^{i(\vec{k}' - \vec{k})\vec{R}}$$

Summing up the contributions of all cells can now be replaced by an integration over the whole crystal, such that

$$a(\vec{k}, \vec{k}') = \alpha(0) \int_V V_p(\vec{r}) e^{i(\vec{k}' - \vec{k})\vec{r}} d\vec{r}$$

It remains to determine $\alpha(0)$, which we approximate by

$$\alpha(0) = \frac{1}{V} \int_V u_{\vec{k}}^*(\vec{r}) u_{\vec{k}'}(\vec{r}) d\vec{r} \approx \frac{1}{V}$$

This is justified since $\vec{k}' \approx \vec{k}$, and the integral in the definition of $\alpha(0)$ should give a value very close to 1 (recall that the functions $\{u_{\vec{k}}\}$ are orthonormal). This finally leads to

$$a(\vec{k}, \vec{k}') = \frac{1}{V} \int_V V_p(\vec{r}) e^{i(\vec{k}' - \vec{k})\vec{r}} d\vec{r} \quad (2.41)$$

Inserting Eq. (2.41) in Eq. (2.34) and using the effective mass approximation for the unperturbed crystal,

$$\epsilon_{\vec{k}} = E_C + \hbar^2 \vec{k}^2 / 2m^* \quad (2.42)$$

Eq. (2.34) changes to

$$\frac{\hbar^2 \vec{k}^2}{2m^*} c_{\vec{k}} + [E_C - E] c_{\vec{k}} + \frac{1}{V} \sum_{\vec{k}'} c_{\vec{k}'} \int V_p(\vec{r}) e^{i(\vec{k}' - \vec{k})} d\vec{r} = 0 \quad (2.43)$$

We proceed by defining the *envelope wave function* as

$$\psi(\vec{r}) = \frac{1}{\sqrt{V}} \sum_{\vec{k}'} c_{\vec{k}'} e^{i\vec{k}'\vec{r}} \quad (2.44)$$

We plan to insert the envelope wave function in Eq. (2.43) by substituting $c_{\vec{k}}$ and $\vec{k}^2 c_{\vec{k}}$. This can be done via the relations

$$c_{\vec{k}} = \sum_{\vec{k}'} c_{\vec{k}'} \delta(\vec{k} - \vec{k}') = \frac{1}{V} \int \sum_{\vec{k}'} c_{\vec{k}'} e^{i\vec{k}'\vec{r}} e^{-i\vec{k}\vec{r}} d\vec{r} = \frac{1}{\sqrt{V}} \int \psi(\vec{r}) e^{-i\vec{k}\vec{r}} d\vec{r}$$

and

$$\begin{aligned} \vec{k}^2 c_{\vec{k}} &= \sum_{\vec{k}'} \vec{k}'^2 c_{\vec{k}'} \delta(\vec{k} - \vec{k}') = \frac{1}{V} \int \sum_{\vec{k}'} \vec{k}'^2 c_{\vec{k}'} e^{i\vec{k}'\vec{r}} e^{-i\vec{k}\vec{r}} d\vec{r} \\ &= \frac{1}{\sqrt{V}} \int (-\Delta \psi(\vec{r})) e^{-i\vec{k}\vec{r}} d\vec{r} \end{aligned}$$

The equation

$$\int e^{-i\vec{k}\vec{r}} \left[-\frac{\hbar^2 \Delta}{2m^*} + E_C - E + V_p \right] \psi(\vec{r}) d\vec{r} = 0$$

is obtained, which is fulfilled for all \vec{k} only if

$$\left[-\frac{\hbar^2 \Delta}{2m^*} + V_p(\vec{r}) \right] \psi(\vec{r}) = [E - E_C] \psi(\vec{r}) \quad (2.45)$$

This is the *envelope wave equation*. For perturbation potentials that vary slowly on the scale of the crystal unit cell, the energy eigenvalues of $V_p(\vec{r})$ in the crystal correspond to the energy eigenvalues of $V_p(\vec{r})$ in a homogeneous medium with the dielectric constant of the crystal, and for particles that have the effective mass of the corresponding electronic band. The energy eigenvalues obtained from the envelope wave equation are relative to E_C , the conduction band bottom, in our case. The envelope wave functions are thus just regular wave functions that solve Eq. (2.45).

2.5

Doping

In many cases, it is desirable to have predominantly one type of mobile carrier, or to have a carrier density independent of temperature within a certain range. This can be achieved by implanting suitable impurities, also known as dopants, into the crystal. As an example, consider a Si atom replacing a Ga atom in a GaAs crystal (Fig. 2.10). Only three of the four valence electrons of Si can be placed in the covalent bonds with adjacent As atoms. The remaining electron will be bound to the attractive potential of the Si ion in the GaAs environment. This model will resemble a Coulomb potential in a medium with the dielectric constant of GaAs. It is straightforward to estimate the energy levels and the wave functions of this potential by using the effective mass approximation.

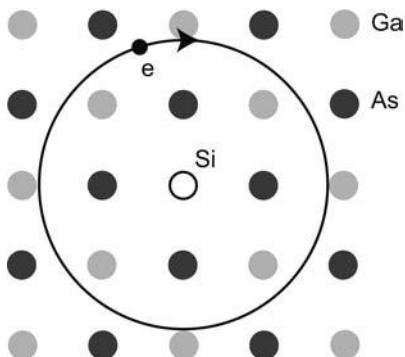


Fig. 2.10 Schematic example of a donor atom in a semiconductor (left). To the right, the energy levels of the donor and acceptor ground states with respect to the band edges are sketched.

The envelope wave equation for the electron in the donor potential reads

$$\left[-\frac{\hbar^2 \Delta}{2m_e^*} - \frac{e^2}{4\pi\epsilon\epsilon_0 r} \right] \psi(\vec{r}) = [E - E_C] \psi(\vec{r}) \quad (2.46)$$

The hydrogen-like energy levels of the doping atom are given with respect to the conduction band bottom. Compared to hydrogen, the energy spectrum is compressed by the factor $(1/\epsilon^2)(m_e^*/m)$,

$$E_{D,j} = E_C + \frac{1}{\epsilon^2} \frac{m_e^*}{m} E_H = E_C - 13.6 \text{ eV} \times \frac{1}{j^2} \frac{1}{\epsilon^2} \frac{m_e^*}{m}.$$

Since, for semiconductors, $\epsilon \approx 10$ and $m_e^* \approx 0.1m$, the binding energy of a typical donor is reduced by a factor of ≈ 1000 as compared to the hydrogen atom, and is just a few meV. The effective Bohr radius becomes very large. For

the ground state of the dopant ($j = 1$), it is found that

$$a_B^* = \epsilon \frac{m}{m_e^*} a_B \approx 5 \text{ nm},$$

which is much larger than the lattice constant. This in retrospect justifies our assumption that the doping electrons actually see the average dielectric constant of the host crystal.

Such weakly bound electrons can be easily thermally excited into the conduction band. Impurities that generate such levels are called donors. Simultaneously, states just above the valence band can be occupied by electrons from the valence band by thermal excitation (suppose an Si atom replaces an As atom). Impurities that generate this kind of state are called acceptors. Equivalently, we can rephrase this process and say “the acceptor donates a hole to the valence band”. In reality, the doping atoms usually do not replace the crystal atoms. Rather, they are placed at interstitial sites, and it depends on the local potential whether the atom acts as a donor or as an acceptor. Typical n-dopants for Si are Sb and P, while B and Al are common p-dopants. In both cases, the binding energy of the electrons (holes) is in the range of 50 meV. Si is predominantly an n-dopant for GaAs, with a binding energy of about 6 meV, while Be or Zn can be used for p-doping. Here, the hole binding energy is of the order of 30 meV. Some dopants, such as oxygen or chromium, have deep doping levels, which mean that they lie somewhere around the center of the bandgap. This cannot be explained with the envelope function model, where only the parameters of the semiconductor host enter. It remains to mention that there are also excited dopant levels, which are of no further interest to us.

How do the carrier densities change due to the doping process? The law of mass action still holds, but all doping atoms have to be included in the effective density of states. This, together with the charge neutrality condition, determines the carrier densities. We denote by n_D and p_A the total density of donors and acceptors, by n_D^0 and p_A^0 the density of neutral donors and acceptors, and by n_D^+ and p_A^- the density of ionized donors and acceptors. These quantities are related via $n_D = n_D^0 + n_D^+$ and $p_A = p_A^0 + p_A^-$. In addition, charge neutrality requires $n + p_A^- = p + n_D^+$.

Let us take n-doping as an example and calculate n . Now, the assumption made in the intrinsic case, $k_B\Theta \ll E_C - \mu$ is no longer justified. The full solution of this problem is beyond the scope of this book. Instead, we look at a simplifying approximation, which captures the main points. Suppose that $p_A = 0$ and intrinsic carriers can be neglected. We further assume that $E_C - \mu > k_B\Theta$, but $E_D - \mu \approx k_B\Theta$. This means that the doping is so high that it pulls the chemical potential very close to the energy level of the dopant. It is important to note that, for typical doping energies close to the valence or the conduction band, the occupation probability is no longer given by a Fermi–Dirac distribution. We discuss the origin qualitatively for a donor level. The derivation of

Eq. (2.24) is based on the assumption that each energy level can be occupied twice without an additional energy associated with double occupancy. This is only true in the non-interacting case. In atoms, the Coulomb energy to be paid for sticking two electrons in the same orbital state typically exceeds the binding energy of the doping atom. If the donor ground state were filled with two electrons, its energy would increase above the conduction band edge, such that this state is unstable. Therefore, only three occupations have to be included in the quantum statistics leading to the probability distribution: the donor state is either empty, or occupied with an electron with spin up or spin down. The probability distribution

$$f(E_D, \Theta) = \frac{1}{\frac{1}{2}e^{(E_D-\mu)/k_B\Theta} + 1}$$

results. By a similar argument, it can be shown that for acceptor levels, the corresponding probability distribution reads

$$f(E_A, \Theta) = \frac{1}{\frac{1}{2}e^{(\mu-E_A)/k_B\Theta} + 1}$$

For a detailed discussion of this issue, see Exercise E2.3. Therefore, the density of occupied donor levels is given by

$$n_D^0 = n_D (1 + \frac{1}{2}e^{(E_D-\mu)/k_B\Theta})^{-1}$$

such that

$$n = n_D^+ = n_D - n_D^0 = \frac{n_D}{1 + 2e^{(\mu-E_D)/k_B\Theta}} \quad (2.47)$$

Also, within our approximation, we can write $n = N_C e^{-(E_C-\mu)/k_B\Theta}$, similar to the intrinsic case. Inserting

$$e^{\mu/k_B\Theta} = \frac{n}{N_C} e^{E_C/k_B\Theta}$$

in Eq. (2.47) results in a quadratic equation for n , the positive solution of which reads

$$n = \frac{N_C}{4} e^{(E_D-E_C)/k_B\Theta} \left[-1 + \sqrt{1 + \frac{8n_D}{N_C} e^{(E_C-E_D)/k_B\Theta}} \right] \quad (2.48)$$

Three regimes can be distinguished, which are summarized in Fig. 2.11.

- $k_B\Theta \ll E_C - E_D$ (freezeout regime)
In this limit, Eq. (2.48) gives

$$n = \sqrt{\frac{N_C n_D}{2}} e^{-(E_C-E_D)/2k_B\Theta}$$

By comparing this expression with the intrinsic case, one finds that the energy levels of the donors play the role of the valence band edge. In effect, the doping has reduced the apparent bandgap by three orders of magnitude. Owing to the modified statistics, the effective donor density of states is half the doping density.

- $k_B\Theta \gg E_C - E_D$, but $k_B\Theta < E_g$ (saturation regime)

Expanding Eq. (2.48) with respect to $(E_C - E_D)/k_B\Theta$ to first order gives $n \approx n_D$. In this regime, the carrier density is constant, as long as intrinsic carriers can be neglected.

- $k_B\Theta \approx E_g$ (intrinsic regime)

This case is not included in Eq. (2.48), but it is clear that now the carrier concentration depends exponentially on the temperature, and the doping electrons can be neglected, since the doping density is by definition much lower than the density of crystal atoms.

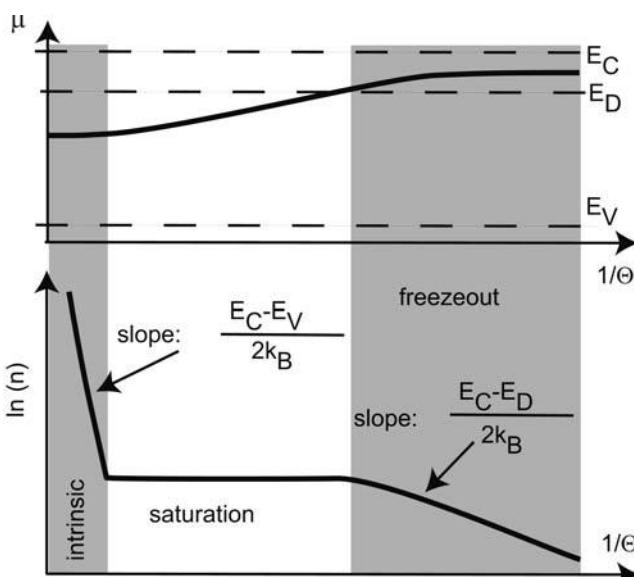


Fig. 2.11 Electron density of a doped semiconductor as a function of the inverse temperature.

The chemical potential reacts accordingly to the temperature and can be easily calculated. For low temperatures, it resides close to the donor level, while at high temperatures, it approaches the middle of the bandgap.

We have already mentioned that not all impurities generate shallow doping levels. Impurities with *deep levels* can be used for “undoping” samples. In some cases, it is desirable to have a semiconductor of extremely high resistivity at room temperature. Owing to unavoidable residual impurities which act

as dopants, the resistivity of ultra-pure GaAs, for example, is no larger than about $1\Omega\text{ m}$. Chromium acts as an acceptor in GaAs with an energy level close to mid-gap. Hence, the residual doping electrons can be removed from the conduction band by a rather small density of Cr doping, which is, however, much higher than the residual n-doping (typical doping densities are of the order of $n_{\text{Cr}} \approx 2 \times 10^{23} \text{ m}^{-3}$). This way, the resistivities can be increased by more than three orders of magnitude. Therefore, such semiconductor materials are called *semi-insulating*. As a consequence, the Fermi level is typically fixed at the energy of the deep dopant, and one speaks of *pinning of the Fermi level*.

2.6

Diffusive Transport and the Boltzmann Equation

Before we discuss the conventional theory of diffusive transport, we briefly summarize some important facts regarding electrons in solids.

- Neither full nor empty bands carry current.
- The resistance of a perfect, static crystal with at least one partially filled electronic band vanishes. The electrons obey the semiclassical equations of motion

$$\vec{v}(\vec{k}) = \frac{1}{\hbar} \vec{\nabla}_{\vec{k}} E(\vec{k})$$

$$\frac{d\vec{k}}{dt} = -\frac{e}{\hbar} (\vec{E} + \vec{v}(\vec{k}) \times \vec{B})$$

Resistance is generated by deviations from the perfect lattice, such as phonons, impurities, lattice dislocations, but also by surfaces and interfaces.

- Electron-electron scattering changes the total momentum of the electron gas only in exceptional cases, and therefore, to a good approximation, does not contribute to the resistance.
- For small applied electric fields, only a tiny fraction of the electrons with energies close to the Fermi level contribute to the current.

In an introductory solid state physics course, transport usually means diffusive transport: a steady state is established between the external electromagnetic fields and the friction inside the solid, which on a microscopic scale is generated by various scattering events. The sample size investigated is much larger than the mean free path, which is the distance an electron travels before it is scattered. This means we observe a homogeneous friction which stems from averaging over all microscopic scattering events.

The Boltzmann equation plays a central role in the theory of diffusive electronic transport. Even though electron-electron interactions and phase coherence are neglected, the general version of the Boltzmann equation is a non-trivial integro-differential equation. Only after its linearization, the relaxation time approximation and some further assumptions does the equation give us a simple picture of how an electric field acts on the carriers: essentially, the Fermi sphere is displaced in k -space without changing its shape. The relaxation time approximation introduces a phenomenological parameter known as “momentum relaxation time”, frequently also referred to as the “Drude scattering time”, τ . All important scattering mechanisms are contained in this parameter. We use it in the Drude model to include magnetic field effects.

Anything that disturbs the perfect lattice will lead to scattering of electrons. Lattice imperfections, which we describe by a perturbation Hamiltonian V_p , will scatter electronic waves from the initial state $|\vec{k}\rangle$ into a final state $|\vec{k}'\rangle$. The scattering matrix elements $W_{\vec{k}, \vec{k}'}$ have to be calculated from

$$W_{\vec{k}, \vec{k}'} \propto |\langle \vec{k}' | V_p | \vec{k} \rangle|^2 \quad (2.49)$$

A large subfield of transport theory is to calculate such matrix elements for all kinds of scatterers. We will mention some important scattering mechanisms below.

2.6.1

The Boltzmann equation

In general, both external fields as well as scattering will modify the Fermi distribution, which we write here as $f(\vec{k}) = [1 + e^{(E(\vec{k}) - \mu)/k_B\Theta}]^{-1}$. The electron distribution function $\phi(\vec{k}, \vec{r}, t)$ is, in the most general case, not a Fermi function. It may depend on \vec{r} and on the time t . Note that the points $\{\vec{k}, \vec{r}\}$ constitute the phase space, with

$$\frac{2}{(2\pi)^3} \phi(\vec{k}, \vec{r}, t) d\vec{k} d\vec{r}$$

being the number of electrons in $d\vec{k} d\vec{r}$ for systems with a spin degeneracy of 2.

We consider the evolution of $\phi(\vec{k}, \vec{r}, t)$ in the time interval dt after time t due to an external, static electric field \vec{E} . We could add the effect of a magnetic field, which is dealt with in a similar way, although this is somewhat more elaborate [270]. Within dt , an electron located at (\vec{k}, \vec{r}) in phase space at time t moves to $(\vec{k} + \delta\vec{k}, \vec{r} + \delta\vec{r})$, which, according to the semiclassical equations of motion, equals $(\vec{k} - (e/\hbar)\vec{E} dt, \vec{r} + \vec{v}(\vec{k}) dt)$. This only holds if the electron is not scattered into a different region of the phase space. Also, not all electrons in $(\vec{k} + \delta\vec{k}, \vec{r} + \delta\vec{r})$ at time $t + dt$ were at (\vec{k}, \vec{r}) at time t : they could have been scattered into this volume within dt . These scattering events change $\delta\phi$, which

we write as

$$\delta\phi = \left[\frac{\partial\phi(\vec{k}, \vec{r}, t)}{\partial t} \right]_{\text{scatter}} dt$$

This results in

$$\begin{aligned} & \phi\left(\vec{k} - \frac{e}{\hbar}\vec{E} dt, \vec{r} + \vec{v}(\vec{k}) dt, t + dt\right) d\vec{k} d\vec{r} \\ &= \phi(\vec{k}, \vec{r}, t) d\vec{k} d\vec{r} + \left[\frac{\partial\phi(\vec{k}, \vec{r}, t)}{\partial t} \right]_{\text{scatter}} dt d\vec{k} d\vec{r} \end{aligned}$$

The size of the volume element $d\vec{k} d\vec{r}$ cannot change, which is the statement of Liouville's theorem on the evolution of semiclassical systems in phase space. Now, the *general Boltzmann equation* is obtained by expanding the left-hand side in a Taylor series in dt up to first order:

$$\vec{v}(\vec{k}) \cdot \vec{\nabla}_{\vec{k}}\phi(\vec{k}, \vec{r}, t) - \frac{e\vec{E}}{\hbar} \cdot \vec{\nabla}_{\vec{k}}\phi(\vec{k}, \vec{r}, t) + \frac{\partial\phi(\vec{k}, \vec{r}, t)}{\partial t} = \left[\frac{\partial\phi(\vec{k}, \vec{r}, t)}{\partial t} \right]_{\text{scatter}} \quad (2.50)$$

In principle, Eq. (2.49) can be calculated from the scattering matrix elements for all scattering mechanisms of relevance (like e.g. electron–phonon scattering or impurity scattering; see [270] for a detailed discussion), each weighted by the corresponding occupation probability of the initial state and the probabilities for finding the final state empty. These probabilities, however, are just the distribution functions $\phi(\vec{k}, \vec{r}, t)$, and $1 - \phi(\vec{k}, \vec{r}, t)$, respectively. Therefore, the general Boltzmann equation is in fact a complicated integro-differential equation, and models as well as approximations are needed to evaluate the scattering term.

A rather crude approximation consists of putting all these scattering mechanisms together and assuming that they generate an average “relaxation time” τ , which we further assume to be independent of \vec{k} and \vec{r} . This is based on the following picture. Provided the system is homogeneous in real space, we can drop the spatial coordinates. If we switch off the external field at time t_0 , the distribution function will exponentially relax to $f(\vec{k})$ with a decay time τ :

$$\phi(\vec{k}, t) = f(\vec{k}) + (\phi(\vec{k}, t_0) - f(\vec{k}))e^{-t/\tau}$$

Since $\vec{E} = 0$, this relaxation will take place exclusively via scattering, and hence

$$\frac{\partial\phi(\vec{k}, t)}{\partial t} = \left[\frac{\partial\phi(\vec{k}, t)}{\partial t} \right]_{\text{scatter}} = -\frac{\phi(\vec{k}, t) - f(\vec{k})}{\tau}$$

which simplifies the general Boltzmann equation considerably. In a stationary state (no time dependence), this now reads

$$-\frac{e\vec{E}}{\hbar} \cdot \vec{\nabla}_{\vec{k}}\phi(\vec{k}) = -\frac{\phi(\vec{k}) - f(\vec{k})}{\tau} \quad (2.51)$$

Eq. (2.51) can be further evaluated by considering small electric fields only. In this regime, the deviation of ϕ from the Fermi function should be roughly linear in \vec{E} , and we can thus write $\vec{\nabla}_{\vec{k}}\phi(\vec{k}) \approx \vec{\nabla}_{\vec{k}}f(\vec{k})$. Now, Eq. (2.43) represents a Taylor expansion of $\phi(\vec{k})$ in $(e\tau\vec{E}/\hbar)$ up to first order:

$$\phi(\vec{k}) = f(\vec{k}) + \vec{\nabla}_{\vec{k}}f(\vec{k}) \frac{e\tau\vec{E}}{\hbar}$$

The right-hand side is a good approximation for $f(\vec{k} + e\tau\vec{E}/\hbar)$, provided that $e\tau\vec{E}/\hbar \ll \vec{k}$. We thus finally find a *simplified Boltzmann equation*, which states that, under all the approximations made, small electric fields displace the Fermi surface in k -space by $e\tau\vec{E}/\hbar$ (Fig. 2.12):

$$\phi(\vec{k}) = f\left(\vec{k} + \frac{e\tau\vec{E}}{\hbar}\right) \quad (2.52)$$

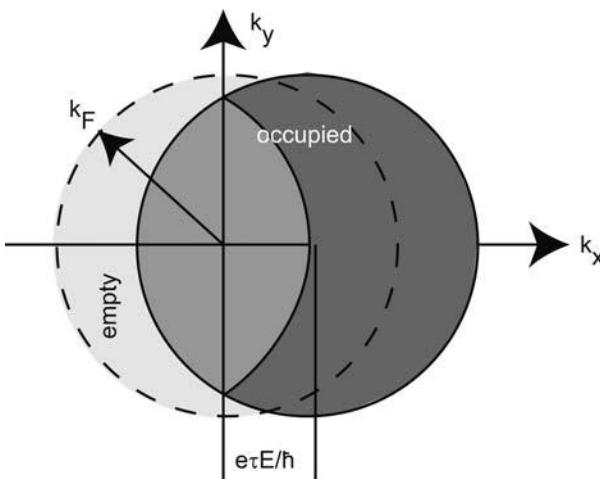


Fig. 2.12 The displaced Fermi sphere as obtained from the Boltzmann equation. The electrons in the dark gray region carry the net current.

Electrons get accelerated and scatter into empty states via elastic or inelastic processes, which emphasizes again the diffusive and dissipative character of the Boltzmann model. As a consequence, this displacement is quasi-static. In addition, we see that only electrons close to the surface of the Fermi sphere contribute to the current. For states deep inside the Fermi sphere, the partial current generated by an electron with momentum $\hbar\vec{k}$ is canceled by the electron with momentum $-\hbar\vec{k}$.

2.6.2

The conductance predicted by the simplified Boltzmann equation

It remains to calculate the conductance σ predicted by the assumptions leading to Eq. (2.52). In general, σ is a tensor defined by

$$\vec{j} = \sigma \vec{E}$$

However, it makes sense to assume $\vec{j} \parallel \vec{E}$, such that σ is actually a scalar. It is obtained from the current density via

$$\vec{j} = \sigma \vec{E} \quad \Rightarrow \quad \sigma = \frac{\vec{j} \cdot \vec{E}}{\vec{E}^2} \quad (2.53)$$

In order to calculate \vec{j} , we have to integrate over the \vec{k} -space, weighting each state by its occupation probability. State \vec{k} contributes a partial current of

$$\vec{j}(\vec{k}) = -e\phi(\vec{k})\vec{v}(\vec{k}) = -\frac{e\hbar}{m^*}\vec{k}\phi(\vec{k})$$

The total current density is obtained by summing up the contributions of all states. Since, for a spin degeneracy of 2, each state occupies a volume of $4\pi^3$ in \vec{k} -space, this summation can be written as the integral

$$\vec{j} = \int \vec{j}(\vec{k}) d\vec{k} = -\frac{e\hbar}{4\pi^3 m^*} \underbrace{\int \vec{k}f(\vec{k}) d\vec{k}}_{=0} + \int \vec{k}\vec{\nabla}_{\vec{k}}f(\vec{k}) \frac{e\tau\vec{E}}{\hbar} d\vec{k}$$

Since

$$\vec{\nabla}_{\vec{k}}f(\vec{k}) = \frac{\partial f(\vec{k})}{\partial E}\vec{\nabla}_{\vec{k}}E(\vec{k}) = \frac{\partial f(\vec{k})}{\partial E}\frac{\hbar^2\vec{k}}{m^*}$$

the current density equals

$$\vec{j} = -\frac{e^2\tau\hbar^2}{4\pi^3 m^{*2}} \int \vec{k} \frac{\partial f(\vec{k})}{\partial E} [\vec{k}\vec{E}] d\vec{k}$$

With Eq. (2.53), we can write

$$\sigma = -\frac{e^2\tau\hbar^2}{4\pi^3 m^{*2}} \int \frac{(\vec{k}\vec{E})^2}{\vec{E}^2} \frac{\partial f(\vec{k})}{\partial E} d\vec{k}$$

For sufficiently low temperatures,

$$-\frac{\partial f(E)}{\partial E} = \delta(E - E_F) = \delta(k - k_F) \frac{m^*}{\hbar^2 k}$$

which results in the surface integral

$$\begin{aligned}\sigma &= \frac{e^2\tau}{4\pi^3 m^*} \int \frac{(\vec{k}\vec{E})^2}{\vec{E}^2} \delta(k - k_F) \frac{1}{k} d\vec{k} \\ &= \frac{e^2\tau}{4\pi^3 m^*} \int_{\theta=0}^{2\pi} \int_{\varphi=0}^{\pi} k_F^3 \cos^2 \varphi \sin \varphi d\varphi d\theta = \frac{e^2\tau k_F^3}{3\pi^2 m^*}\end{aligned}$$

Since the electron density n is given by $n = 3\pi^2 k_F^3$, we find

$$\sigma = \frac{ne^2\tau}{m^*} = ne\mu \quad (2.54)$$

Here, we have defined the electron mobility by $\mu \equiv e\tau/m^*$.

Question 2.7: Prove that $\sigma = ne\mu$ also holds in two dimensions.

Result (2.54) is at first sight quite strange: the conductivity is proportional to the total electron density, and it seems like all electrons would contribute equally to the current. However, we know that only the electrons at the Fermi surface carry current. The explanation is that a higher electron density increases the number of electrons and the electron velocity at the Fermi surface, which turns out to give a conductivity proportional to n .

We can use Eq. (2.54) to define a useful quantity, the drift velocity \vec{v}_d as

$$\vec{v}_d = -\frac{\vec{J}}{en} = -\mu \vec{E} \quad (2.55)$$

The drift velocity is thus an effective average velocity, which leads to an equation for the current density that is formally identical to the Drude expression, which was derived by assuming that all electrons contribute equally to the current and move through the crystal with an average drift velocity.

Along similar lines, it can be shown that, in the additional presence of magnetic fields, the current density can be written as

$$\vec{J} = \sigma(\vec{E} + \vec{v}_d \times \vec{B}) \quad (2.56)$$

This current density corresponds to the stationary solution of the classical equation of motion

$$m^* \frac{d^2 \vec{r}}{dt^2} + \frac{m^*}{\tau} \vec{v}_d = -e(\vec{E} + \vec{v}_d \times \vec{B}) \quad (2.57)$$

Thus, electrons are moving at velocity \vec{v}_d through the crystal and experience a Stokes-type friction term given by $m^* \vec{v}_d / \tau$.

2.6.3

The magneto-resistivity tensor

We proceed by studying the electron transport according to Eq. (2.57) in weak magnetic fields, with “weak” being specified by

$$\omega_c = e|\vec{B}|/m_e^* \ll 1/\tau \quad (2.58)$$

This condition means that the distance the electrons travel before getting scattered (the *mean free path* $\ell_e \equiv v_F\tau$) is small compared to the cyclotron circumference $2\pi r_c$. We will see in Chapters 6 and 7 what happens when the electrons can complete the cyclotron orbits without getting scattered. Suppose a magnetic field is applied in the z -direction, $\vec{B} = (0, 0, B)$. In such a case, we obtain

$$\begin{aligned} j_x &= \sigma E_x + \sigma v_y B = \sigma E_x + \frac{ne^2\tau}{m_e^*} v_y B = \sigma E_x - j_y \omega_c \tau \\ j_y &= \sigma E_y - \sigma v_x B = \sigma E_y - \frac{ne^2\tau}{m_e^*} v_x B = \sigma E_y + j_x \omega_c \tau \\ j_z &= \sigma E_z \end{aligned}$$

where v_i are the components of the drift velocity vector. Solving this system of equations for \vec{j} gives $\vec{j} = \underline{\sigma} \vec{E}$ with

$$\underline{\sigma} = \frac{\sigma}{1 + \omega_c^2 \tau^2} \begin{pmatrix} 1 & -\omega_c \tau & 0 \\ \omega_c \tau & 1 & 0 \\ 0 & 0 & 1 + \omega_c^2 \tau^2 \end{pmatrix} \quad (2.59)$$

Here $\underline{\sigma}$ is known as the magneto-conductivity tensor. Its components can be experimentally determined by measuring four-probe resistances using “Hall bar” shaped samples (Fig. 2.13). Voltage probes are attached to a rectangular thin film of the material, aligned parallel to the x - and y -directions, and perpendicular to the magnetic field direction. The transport in the z -direction remains unaffected by \vec{B} and is of no further interest to us. We can determine the components ρ_{xx} and ρ_{xy} of the resistivity tensor by applying a current in the x -direction and measuring the voltage drops V_x and V_y . Since

$$\begin{pmatrix} V_x \\ V_y \end{pmatrix} = \begin{pmatrix} \rho_{xx} & \rho_{xy} \\ -\rho_{xy} & \rho_{xx} \end{pmatrix} \cdot \begin{pmatrix} I_x \\ I_y \end{pmatrix} \cdot S$$

and

$$\begin{pmatrix} I_x \\ I_y \end{pmatrix} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ -\sigma_{xy} & \sigma_{xx} \end{pmatrix} \cdot \begin{pmatrix} V_x \\ V_y \end{pmatrix} \cdot \frac{1}{S}$$

where S is a geometry factor (see [235]), we can establish the relation between the components of the resistivity and the conductivity tensors:

$$\rho_{xx} = \frac{\sigma_{xx}}{\sigma_{xx}^2 + \sigma_{xy}^2}, \quad \rho_{xy} = \frac{-\sigma_{xy}}{\sigma_{xx}^2 + \sigma_{xy}^2} \quad (2.60)$$

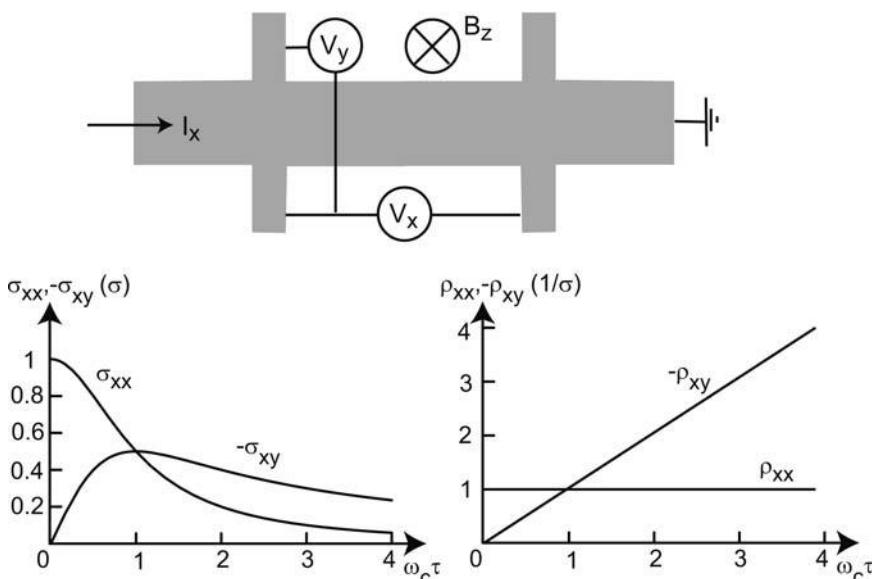


Fig. 2.13 Top: Top view of a Hall geometry. The magnetic field is applied perpendicular to the sheet. Bottom: The components of the conductivity and the resistivity tensors are shown to the left and to the right, respectively.

We thus find that ρ_{xx} does not depend on \vec{B} , and $\rho_{xy} = -B/en = R_H B$. Here ρ_{xy} is the Hall resistivity, and $R_H = -1/en$ is known as the Hall coefficient. Hall measurements are actually a standard tool to determine carrier densities. It may be counter-intuitive at first sight that, for $\rho_{xx} = 0$, σ_{xx} becomes zero as well. Furthermore, the Onsager–Casimir symmetry relation should be mentioned, which states that the result of a measurement is exactly the same when all current and voltage sources are exchanged, and the polarity of the magnetic field is reversed. One consequence is that two-probe measurements, in which the voltage drop is measured between the source and drain contacts, must be symmetric with respect to $B = 0$.

Question 2.8: Write down Eqs. (2.60) for an anisotropic sample.

2.6.4

Diffusion currents

In close analogy to the treatment of drift currents, the Boltzmann model can be applied to diffusion currents, i.e. currents as a consequence of a position-dependent varying chemical potential $\mu(\vec{r})$, which can have its origin in a gradient of the temperature or of the carrier density. Assuming constant tem-

perature and no external electric fields, the Boltzmann treatment results in a diffusion current density

$$\vec{j}_{\text{diff}} = \frac{n\sigma_0}{e} \vec{\nabla} \mu(\vec{r}) \quad (2.61)$$

The diffusion current is frequently expressed in terms of the carrier density gradient and the diffusion constant D ,

$$\vec{j}_{\text{diff}} = eD\vec{\nabla}n(\vec{r}) \quad (2.62)$$

Question 2.9: Show that for a quasi-free electron gas, the relation between D and the mobility μ equals

$$\begin{aligned} D &= \frac{2E_F}{3e}\mu && \text{in } d = 3 \\ D &= \frac{E_F}{e}\mu && \text{in } d = 2 \\ D &= \frac{2E_F}{e}\mu && \text{in } d = 1 \end{aligned} \quad (2.63)$$

Such relations are known as Einstein relations.⁴

2.7

Scattering mechanisms

As mentioned in Section 2.6, many scattering mechanisms contribute to the average momentum relaxation time τ . Each process has its characteristic matrix element $W_{\vec{k},\vec{k}'}$, Eq. (2.49). The relevance of a particular kind of scattering varies greatly and depends on the carrier density as well as on the temperature. How in detail the matrix elements are calculated is treated in several excellent books, e.g. [254, 270]. Each scattering mechanism can be characterized by its contribution to the carrier mobility μ_i , which sum up to the total mobility according to the Matthiesen rule, $1/\mu = \sum_i 1/\mu_i$. In pure crystals, the sole source of scattering is lattice vibrations. Electron–phonon scattering has several facets. In crystals with valley degeneracy, electrons may be scattered between valleys, which requires absorption or emission of a phonon. In polar and/or piezoelectric crystals, on the other hand, lattice vibrations go along with strong oscillating electric fields. In real crystals, charged impurities may dominate the scattering rates. We briefly present the most important scattering mechanisms below.

4) The Einstein relation for particles obeying the Boltzmann statistics is Eq. (5.10).

An impurity breaks the symmetry of the lattice and causes scattering. If the impurity is neutral, the scattering rates are usually negligible. Charged impurities, however, represent screened Coulomb scatterers, with peak potentials that can become comparable to the Fermi energy.⁵ Clearly, an electron with a larger kinetic energy will get deflected by a smaller angle as it gets scattered, and we can expect that the mobility increases as the temperature, and with it the average electron kinetic energy, increases. In fact, an evaluation of the corresponding matrix element shows that, for weak Coulomb potentials and within the Born approximation, the resulting mobility is $\propto \Theta^{3/2}$, multiplied by a logarithmic correction, i.e. a factor that depends logarithmically on Θ .

Electron–phonon scattering can be divided into deformation potential scattering and scattering of electrons by the corresponding electric fields. By deformation potential scattering, we mean scattering at the lattice deformations caused by the phonons. Here, scattering at acoustic phonons is the most important mechanism. Since the energy transfers are small in electron–acoustic phonon scattering, it can be treated as quasi-elastic. A simple argument gives the correct temperature dependence. The density of acoustic phonons n_{ac} is proportional to the Bose–Einstein distribution, which, for large temperatures compared to the phonon energy, varies as $1/\Theta$. Since the mobility is proportional to n_{ac}/\bar{v} (\bar{v} is the average electron velocity, which is $\propto \sqrt{\Theta}$), we expect that the mobility due to electron–acoustic phonon scattering is $\propto \Theta^{-3/2}$. This is in fact observed experimentally.

Furthermore, both optical and acoustic phonons can assist the electron in scattering between the valleys in a crystal with valley degeneracy, such as Si. The corresponding momentum transfers are quite large, since the separation of the valley in reciprocal space is of the order of the size of the Brillouin zone.

This completes the list of the scattering mechanisms relevant in Si. In this material, ionized impurities dominate the mobility at low temperatures, while quasi-elastic acoustic phonon scattering is the most important mechanism at intermediate temperatures. For $\Theta > 200$ K, inter-valley scattering becomes significant as well. Consequently, the mobility in Si shows a maximum as a function of temperature. Its position depends on both the impurity density and the carrier density. Electron mobilities up to $1 \text{ m}^2/\text{Vs}$ have been achieved in Si.

GaAs is a polar material, and consequently lattice vibrations are always accompanied by oscillating electric fields. They are particularly strong for optical phonons. The resulting scattering mechanism is called polar scattering. Optical phonons vanish for temperatures below ≈ 60 K, and consequently polar scattering is relevant only above this temperature. In the limit $k_B\Theta \gg \hbar\omega_{\text{op}}$ (ω_{op} denotes the optical phonon frequency, which for GaAs is of the order of 5 meV; see Fig. 2.14), it can be shown that the resulting mobility varies as

5) The screened Coulomb potential is studied in Exercise E2.5.

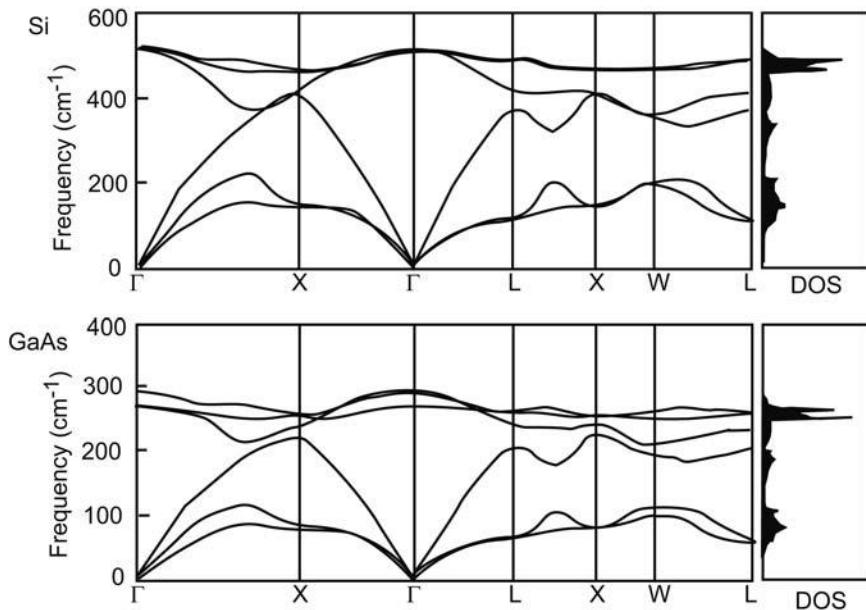


Fig. 2.14 Phonon dispersions for Si (top) and GaAs (bottom). After [118].

$\Theta^{-1/2}$. If the crystal is piezoelectric like GaAs, a crystal deformation generates a polarization field as well, which is another source of scattering, called piezoelectric scattering. As for the polar scattering, the mobility due to piezoelectric scattering is $\propto \Theta^{-1/2}$, although this temperature dependence holds for a larger range of temperatures.

Fig. 2.15 summarizes the contributions of different scattering mechanisms to the electron total mobility of GaAs. A comparison with measurements reveals that, at low temperatures, ionized impurity scattering dominates, while, at higher temperatures, the mobility is entirely determined by polar scattering. In a small temperature range around the emerging maximum of the mobility, piezoelectric scattering is significant. Furthermore, it is seen that acoustic phonon scattering plays no role, in contrast to the scattering in Si.

2.8 Screening

The conduction electrons react to perturbations. They collect in the potential valleys and avoid the peaks. As a consequence, the external potential is reduced to an effective potential in the crystal; the electrons “screen” the perturbation. The goal of this section is to present a qualitative picture of how

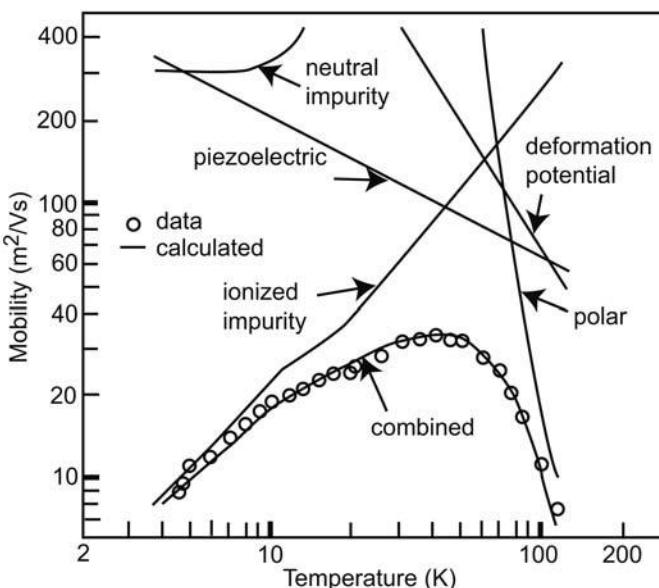


Fig. 2.15 Measured electron mobility in GaAs (circles) as a function of temperature, including the theoretical contributions of relevant scattering mechanisms (full lines). The sample contained a donor density of $n_D = 4.8 \times 10^{19} \text{ m}^{-3}$ and an acceptor density of $n_A = 2.1 \times 10^{19} \text{ m}^{-3}$. After [289].

the electron density is modified by perturbations. For a more detailed discussion of screening and electron–electron interactions, the reader is referred to textbooks on solid state physics (see e.g. [12, 127, 346]).

By time-dependent perturbation theory, it can be shown that the screening in a free electron gas depends on the wave vector \vec{q} and the frequency ω of the perturbation. It can be expressed by a dielectric function $\epsilon(\vec{q}, \omega)$ of the type

$$\epsilon(\vec{q}, \omega) = 1 + \epsilon_{\text{lattice}} + \frac{e^2}{\epsilon_0 q^2} \sum_{\vec{k}} \frac{f(\vec{k}) - f(\vec{k} + \vec{q})}{E(\vec{k} + \vec{q}) - E(\vec{k}) + i\alpha} \quad (2.64)$$

Here, $\epsilon_{\text{lattice}}$ means the dielectric function of the lattice [12], and $\alpha \rightarrow 0$ is the (small) convergence parameter, which can be related to a scattering time. The dielectric function describes how the Fourier components of the external potential energy $V_{\text{ext}}(\vec{q}, \omega)$ are screened and result in an effective potential energy $V_{\text{eff}}(\vec{q}, \omega)$, namely

$$V_{\text{eff}}(\vec{q}, \omega) = \frac{V_{\text{ext}}(\vec{q}, \omega)}{\epsilon(\vec{q}, \omega)} \quad (2.65)$$

In the static limit ($\omega \rightarrow 0$), within the effective mass approximation, and for low temperatures $k_B\Theta \ll E_F$, the sum in Eq. (2.64) can be calculated analyti-

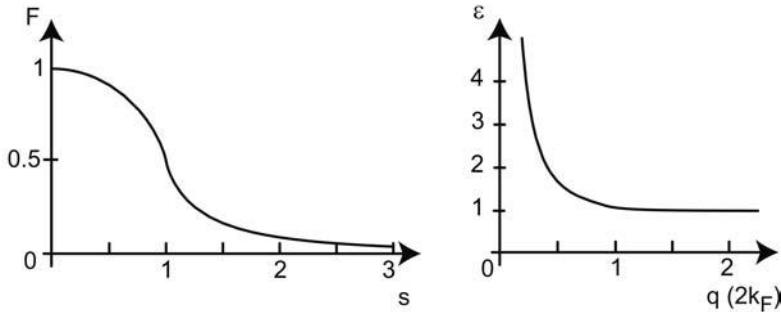


Fig. 2.16 The function $F(s)$ (left) and the static dielectric function $\epsilon(\vec{q})$ (right) for free electrons with typical experimental parameters in doped semiconductors ($k_F = 10^8 \text{ m}^{-1}$).

cally, and one finds

$$\begin{aligned}\epsilon(\vec{q}) &= 1 + \frac{k_{\text{TF}}^2}{q^2} F\left(\frac{q}{2k_F}\right) \\ F(s) &= \frac{1}{2} + \frac{1-s^2}{4s} \ln \left| \frac{1+s}{1-s} \right|\end{aligned}\quad (2.66)$$

Here, we have defined the Thomas–Fermi screening vector

$$k_{\text{TF}} = e \sqrt{\frac{D(E_F)}{\epsilon_0}} \quad (2.67)$$

The functions $F(s)$ and $\epsilon(\vec{q})$ are shown in Fig. 2.16. Most notably, $\epsilon(\vec{q})$ drops significantly as \vec{q} increases. Above $2\vec{k}_F$, it rapidly approaches 1. The reason is simply that, for low temperatures, the wave vector of occupied states differs by no more than $2\vec{k}_F$. The term $f(\vec{k}) - f(\vec{k} + \vec{q})$ means that only states contribute where the occupation of the two states characterized by \vec{k} and $\vec{k} + \vec{q}$ is different. The number of contributing states thus increases as $0 \leq \vec{q} \leq 2\vec{k}_F$, but remains constant for $\vec{q} > 2\vec{k}_F$ in Eq. (2.52), which means that $\epsilon(\vec{q})$ drops significantly at $\vec{q} = 2\vec{k}_F$. This point in fact represents a logarithmic singularity, which has important consequences for the screening. As an example, consider the potential of a point charge with an external potential energy given by

$$V_{\text{ext}}(r) = -\frac{Ze}{r} = -\frac{Ze^2}{(2\pi)^3} \int \frac{4\pi}{q^2} e^{i\vec{q}\vec{r}} d\vec{q}$$

Correspondingly, the effective potential energy can be written as

$$V_{\text{eff}}(r) = -\frac{Ze}{r} = -\frac{Ze^2}{(2\pi)^3} \int \frac{1}{\epsilon(\vec{q})} \frac{4\pi}{q^2} e^{i\vec{q}\vec{r}} d\vec{q}$$

The induced charge density $\rho_{\text{ind}}(\vec{r})$ is then given by the Poisson equation

$$\Delta V_{\text{ind}}(\vec{r}) = \Delta[V_{\text{eff}}(\vec{r}) - V_{\text{ext}}(\vec{r})] = -e \frac{\rho_{\text{ind}}(\vec{r})}{\epsilon_0}$$

such that

$$\rho_{\text{ind}}(\vec{r}) = \frac{Ze}{(2\pi)^3} \int \frac{1}{\epsilon(\vec{q}) - 1} e^{i\vec{q}\vec{r}} d\vec{q}$$

Evaluating this integral shows that only terms with the argument $\vec{q} \approx 2\vec{k}_F$ contribute significantly, and that, for large r ,

$$\rho_{\text{ind}}(\vec{r}) = \frac{Ze}{\pi} \frac{k_{\text{TF}}^2}{k_F^2(4 + k_{\text{TF}}^2/2k_F^2)} \frac{\cos(2k_F r)}{r^3} \quad (2.68)$$

The charge density thus develops a periodic component, with a period of half the Fermi wavelength. This can be understood in terms of a standing wave due to a superposition of the incoming waves and the waves reflected at the perturbation potential. These oscillations are known as Friedel oscillations.⁶

Papers and Exercises

P2.1 Transport in strongly disordered media has a very different character as compared to transport in crystalline conductors, since all electronic states are localized. It is frequently analyzed in terms of the *variable range hopping* model, which predicts a characteristic temperature dependence of the resistivity. Work out this expression using reference [4].

E2.1 A microchip factory processes “3-inch wafers”, i.e. monocrystalline, cylindrically shaped semiconductor disks, with a diameter of 3 inches and a thickness of 0.5 mm. Unfortunately, the silicon wafer badge and the GaAs wafer badge have not been labeled. Someone suggests determining the material by weighing the wafers. Is this realistic?

Si has an atomic mass of 28.09 amu and crystallizes with a lattice constant of 0.543 nm. For Ga and As, the atomic mass is 69.72 amu and 74.92 amu, respectively. GaAs has a lattice constant of 0.565 nm. Calculate the weight of the two wafer types.

6) Friedel oscillations are not a special property of screened Coulomb potentials. It can be shown [129] that, for large distances from the perturbing potential,

$$\rho_{\text{ind}}(\vec{r}) = A \frac{\cos(2k_F r + \phi)}{r^3}$$

where the phase ϕ and the constant A depend on the potential.

E2.2 Suppose certain macromolecules form two-dimensional crystals. For each unit cell, $2\pi/3$ electrons are available for electronic bands. The unit cell is defined by the lattice vectors $\vec{a}_1 = (4 \text{ nm}, 0)$ and $\vec{a}_2 = (1 \text{ nm}, 3 \text{ nm})$.

- (a) Calculate the reciprocal lattice vectors \vec{b}_1, \vec{b}_2 . Draw the reciprocal lattice and construct the first and the second Brillouin zones.
- (b) Suppose the energy dispersion can be approximated by that of free electrons. What is the radius of the Fermi sphere? Draw the Fermi surface in the reciprocal lattice.
- (c) Which period(s) would you expect in a de Haas–van Alphen experiment? The magnetic field is applied perpendicular to the crystal plane.

E2.3 As noted in Section 2.5, double occupation of a typical doping level is forbidden by the Coulomb repulsion. This causes the distribution function to deviate from a Fermi function.

- (a) Reassure yourself that the average occupation number $\langle n_j \rangle$ of a state j with energy E_j is given by the Fermi function if the Coulomb interaction is neglected. Recall from statistical mechanics that $\langle n_j \rangle$ is given by

$$\langle n_j \rangle = \frac{\sum_n n e^{-n(E_j - \mu)/k_B\Theta}}{\sum_n e^{-n(E_j - \mu)/k_B\Theta}}$$

where n is the number of particles in state j .

- (b) Derive the modified distribution function for donor states.
- (c) What does the average hole occupation number for an acceptor level look like?

E2.4 Carry out the calculation that leads to the interpretation of a localized electron in terms of a Bloch function wave packet (Section 2.3.4). Show that the wave packet is extended over several lattice constants.

E2.5 Study the dielectric function in the limit $\vec{q} \ll 2\vec{k}_F$. Show that the potential of a screened point charge can be written as

$$V_{\text{eff}}(\vec{r}) = -\frac{Ze^2}{r}e^{-k_{\text{TF}}r}$$

Screening in this limit is known as Thomas–Fermi screening.

E2.6 Consider a periodic potential composed of δ functions

$$V(x) = -V_0 \sum_{n=-\infty}^{\infty} \delta(x + na)$$

with $V_0 > 0$ and n integer.

- (a) Determine the eigenvalue E_0 and the eigenfunction $\Phi_0(x)$ of a single δ function, $V_{\text{single}}(x) = -V_0\delta(x)$.
- (b) Show that the tight binding wave functions of the crystal

$$\Psi_k(x) = \sum_{j=-\infty}^{\infty} \Phi_0(x - ja) e^{ikja}$$

satisfy Bloch's theorem.

- (c) Use the wave function and show by the method sketched in the text that the dispersion relation takes the form

$$E(k) = E_0 + \frac{\beta + \sum_{n=1}^{\infty} \gamma_n \cos(kna)}{1 + \sum_{n=1}^{\infty} \alpha_n \cos(kna)}$$

Note that life gets easier if the term with $j = 0$ is treated separately.

- (d) What is the effective mass around $k = 0$?

- E2.7** Use the bandgap energies and the effective masses given in the text to calculate the effective densities of states N_C and P_V , as well as the intrinsic carrier concentrations, for Si and GaAs at room temperature. What happens to the carrier concentrations as the materials are cooled to liquid nitrogen temperature?

Further Reading

There are several excellent books on solid state physics available, e.g. [12, 127, 346]. For particular properties of semiconductors, see [270, 342]. The material parameters of the important semiconductors are listed in a condensed, yet informative, way in [126].

This Page Intentionally Left Blank

3

Surfaces, Interfaces, and Layered Devices

Infinite crystals are convenient for introducing solid state physics. Quite often, surface effects can be neglected in real crystals, as the fraction of surface atoms is vanishingly small. Surfaces play a very important role, though. First of all, they are the interface between the crystal and the outside world. Across surfaces, the energy bands relate to the vacuum level, which is the energy of an electron at rest outside the crystal. It takes the energy Φ_A , also known as the *work function*, to transfer an electron at the chemical potential in the crystal into the vacuum. In pure semiconductors and insulators, this is an impossible process, since there are no states at the chemical potential. The *electron affinity* ξ_e is therefore introduced in addition. It measures the energy difference between the vacuum level and the bottom of the conduction band. Their numerical values depend on both the bulk band structure and surface-specific properties [12].

The regions of interest in most mesoscopic samples, as well as in the majority of commercial microchips, are very close to surfaces and/or interfaces, the influence of which on the active region is usually highly relevant. In fact, crystal interfaces are frequently tailored to provide useful properties. The most elementary interface is that between a crystal and vacuum, which is the topic of Section 3.1. We will see that, at a surface, electronic states can exist that are absent in the bulk, with typical energies in the bandgap of the bulk material, as sketched in Fig. 3.1. These states are not additional states. Rather, they emerge from valence and conduction band states. This is evident if we recall that the number of electronic states in the crystal equals the number of states in all the atoms the crystal contains. In order to appreciate the mechanism leading to surface states, a model calculation for a one-dimensional crystal within the tight binding model is presented in Section 3.1.1. Generalizing the results to three dimensions is conceptually simple, and results in surface bands that are two-dimensional in character (see Section 3.1.2). Typically, surface bands are partly filled. The chemical potential at the surface will usually be located somewhere inside the surface band, which lies inside the bandgap of the bulk. In equilibrium, the surface chemical potential will align with that in the bulk, which in general requires a charge transfer between bulk states and surface

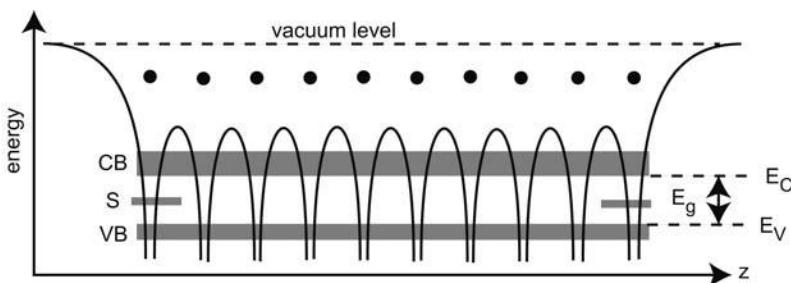


Fig. 3.1 Schematic representation of the potential landscape in a finite crystal, which gets modified close to the surface. Surface states (S) may result, with typical energies inside the gap between the valence band (VB) and the conduction band (CB).

states. The consequences of this mechanism are *band bending* and *Fermi level pinning*, introduced in Section 3.1.3. Both effects are of utmost importance in semiconductor nanostructures.

Generalizing the properties of crystal surfaces to other types of interfaces is straightforward, once the concepts are at hand. Similar to surfaces, charge rearrangements will align the two chemical potentials. Nevertheless, interface states have a somewhat different character. Two types of interfaces are relevant for us. Metal–semiconductor interfaces are studied in Section 3.2. They come in two “flavors”, Schottky contacts and ohmic contacts. Equally important is the semiconductor heterointerface, the topic of Section 3.3. As we shall see, it is quite common to combine layers of different semiconductors and take advantage of the band alignment.

After these preparations, we are ready to look at examples of devices that rely on interface effects. The most important structures for our purposes are the Si MOSFET (metal–oxide–semiconductor field effect transistor) and the Ga[Al]As HEMT (high electron mobility transistor), which will be introduced in Section 3.4. It will become clear that, due to band bending at interfaces, carrier gases are formed which can be *two-dimensional*. Such systems will be the workhorse for most of the experiments discussed in subsequent chapters. There are, however, many more interesting ways to combine semiconductors layers, and we will briefly present some examples in that section as well.

3.1

Electronic surface states

3.1.1

Surface states in one dimension

A surface breaks the translational symmetry of the crystal. How this fact modifies the electronic structure can be studied in various models. Their lines of arguing, however, are similar: Bloch's theorem allows solutions with imaginary wave vectors. However, they correspond to evanescent waves, which is unphysical in an infinite crystal. This is no longer necessarily true at surfaces, where two exponentially decaying wave functions, one for each material, may match to form a localized state.

We use the tight binding model to study how surface states emerge from a σ energy band of the bulk, i.e. a band formed from atomic s orbitals. This model goes back to [121, 122] and [276]. We start from the Schrödinger equation of a finite one-dimensional crystal composed of N atoms with a lattice constant a :

$$\frac{\hbar^2}{2m} \Delta \Psi(z) + [E - V(z)] \Psi(z) = 0 \quad (3.1)$$

where $V(z)$ is the periodic lattice potential. The individual atoms are described by

$$\frac{\hbar^2}{2m} \Delta \Phi(z - z_n) + [E_0 - U(z - z_n)] \Phi(z - z_n) = 0 \quad (3.2)$$

Here, E_0 is the energy eigenvalue of the atom for the s state under consideration. We insert the ansatz

$$\Psi = \sum_n c_n \Phi(z - z_n) \quad (3.3)$$

in Eq. (3.1), multiply by $\Phi^*(z - z_m)$ and integrate over space. If only nearest-neighbor coupling is considered, the matrix elements

$$\langle \Phi_m | \Phi_n \rangle = \delta_{mn} + \beta \delta_{m,n\pm 1}$$

$$\langle \Phi_m | V - U | \Phi_n \rangle = \begin{cases} -\alpha & n = m \notin \{1, N\} \\ -\alpha' & n = m \in \{1, N\} \\ -\gamma & m = n \pm 1 \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$$

are obtained. Since both $(E - E_0)$ and β are small, their product can be neglected. With the definitions $E - E_0 + \alpha = \epsilon$ and $\alpha - \alpha' = \epsilon_0$, the coefficients

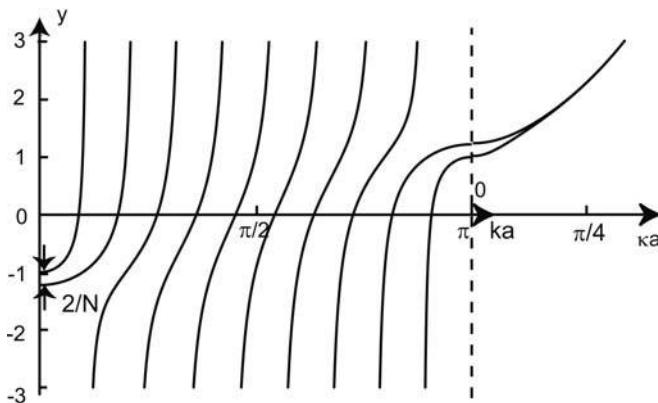


Fig. 3.2 Plot of $y(k)$ for a crystal consisting of $N = 10$ atoms. For $-1 \leq y \leq 1$, ten real wavenumbers k are obtained. For $y > 1$, two values of k have non-vanishing imaginary components, $k = \pi/a + ik$. In this regime, eight real and two complex solutions are obtained.

c_n obey the relations

$$\gamma c_{n-1} + \epsilon c_n + \gamma c_{n+1} = 0, \quad n \in \{2, N-1\} \quad (3.5)$$

$$(\epsilon - \epsilon_0)c_1 + \gamma c_2 = 0 \quad (3.6)$$

$$\gamma c_{N-1} + (\epsilon - \epsilon_0)c_N = 0 \quad (3.7)$$

We write c_n in the form

$$c_n = A e^{ikna} + B e^{-ikna} \quad (3.8)$$

For the bulk, Eq. (3.5) gives the well known dispersion relation

$$\epsilon = 2\gamma \cos(ka) \quad (3.9)$$

which we use in Eqs. (3.6) and (3.7) to calculate the allowed wave numbers in the finite crystal. After some algebra and by using the relations $\sin x \sin y = \frac{1}{2} \cos(x-y) - \frac{1}{2} \cos(x+y)$ as well as $\cos(2x) = \cos^2 x - \sin^2 x$, one finds the condition

$$y \equiv -\frac{\epsilon_0}{\gamma} = \frac{-\sin(Nka) \pm \sin(ka)}{\sin[(N-1)ka]} \quad (3.10)$$

This function is plotted in Fig. 3.2 for $N = 10$. Clearly, an N -atom crystal contains N electronic states that emerge from the atomic s states. In fact, for $-1 \leq y \leq 1$, condition (3.10) delivers N real wave numbers. For $|y| > 1$, however, only $N - 2$ real solutions are found!¹ It can be shown that $\gamma <$

1) There is a small interval $1 \leq |y| \leq 1 + 2/N$ for which $N - 1$ real solutions are obtained. As N is usually large, this region is irrelevant.

0 for s bands [122]. We thus focus on the region $y > 1$. Apparently, two states have formed with an imaginary component of the wave number, which corresponds to states outside the σ band. Inserting $k = \pi/a + ik$ into (3.10) gives

$$y = \frac{\sinh(Nka) \pm \sinh(ka)}{\sinh[(N-1)ka]} \approx e^{\kappa a} \quad (3.11)$$

where the approximation holds in the limit of large N . In this case, the states have the energy dispersion

$$E = E_0 - \alpha + 2\gamma \cosh(\kappa a) \quad (3.12)$$

Since $y > 1$ and $\cosh[\ln(y)] > 1$, the energies of these states are *larger* than those of the bulk states (see Fig. 3.2).

How do the wave functions of these split-off states look? The complex wave numbers describe evanescent wave functions, localized at the crystal surface. This is easily established by calculating the coefficients c_n . Inserting the dispersion relation (3.12) in Eqs. (3.6) and (3.7) gives

$$c_2 = -c_1 e^{-\kappa a}, \quad c_{N-1} = -c_N e^{-\kappa a} \quad (3.13)$$

The remaining coefficients are obtained by recursively applying Eq. (3.5). Since for symmetry reasons $c_1 = c_N$, one obtains

$$c_{n+1} = c_1 (-1)^n [e^{-n\kappa a} + (-1)^N e^{-(N-n)\kappa a}] \quad (3.14)$$

Such a wave function is sketched in Fig. 3.3 for our model crystal. Even for this tiny 10-atom crystal, it is strongly localized at the crystal surfaces and extends only a few lattice constants into the bulk! These are the surface states, which have emerged from the bulk states of the σ band.

Question 3.1: Determine the wave functions of the bulk states using the dispersion given by Eq. (3.9).

Shockley included the effect of band crossings in an extended version of this model [276], and studied the energy of the surface states as a function of the lattice constant. His famous results are summarized in Fig. 3.4. He found that for lattice constants where the bands have crossed, a common case in real crystals, both the upper band and the lower band contribute surface states inside the bandgap. The energies of the surface states are typically close to mid-gap.

Surface states of one-dimensional crystals are sometimes categorized into “Maue–Shockley” states (following the models developed in [205] and [276]),

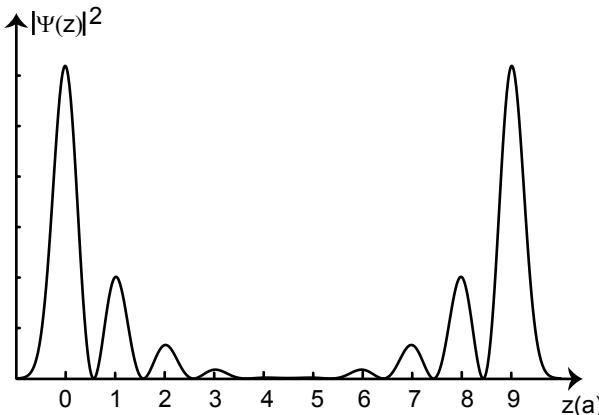


Fig. 3.3 Model wave function for a surface state in a 10-atom crystal. The atomic wave functions $\Phi_n(z) = e^{-4(z-na)^2}$, and $\kappa a = 0.5$, corresponding to $y \approx 1.65$, have been used.

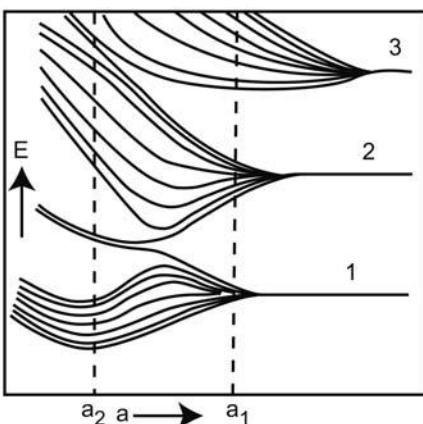


Fig. 3.4 Energy of surface states in the one-dimensional Shockley model, shown as a function of the lattice constant a . At a_2 , for example, both donor-like and acceptor-like surface states are present. After [276].

where the potential at the surface is not modified, except that the periodic potential is interrupted, and “Tamm–Goodwin” states [122, 295], where the surface states occur due to modifications of the surface potential as compared to the bulk. From a more general point of view, these different types of surface states are special cases of a symmetry requirement [343].

It is furthermore instructive to see how surface states are formed within the nearly free electron model. Within this standard model of solid state physics, the periodic potential $V(z)$ is assumed to be weak compared to the kinetic

energy of the electrons. It can thus be well approximated by its Fourier expansion up to first order,

$$V(z) = V_0 + V_g e^{igz} \quad (3.15)$$

Here, V_0 is the constant part of the potential inside the crystal with respect to the vacuum level, and g denotes the smallest reciprocal lattice vector, $g = 2\pi/a$. Because $V(z)$ is weak, it can be treated as a perturbation. The states belonging to wave numbers k and $k - 2\pi/a$ with k very close to the edge of the first Brillouin zone are nearly degenerate.² The effect of the weak periodic potential on the energies of such states can be studied by degenerate perturbation theory. The result of such a calculation for $q = (\pi/a) - k \ll \pi/a$ is the energy dispersion

$$E(q) = V_0 + \epsilon_{\pi/a} + \epsilon_q \pm \sqrt{4\epsilon_{\pi/a}\epsilon_q + |V_g|^2} \quad (3.16)$$

with

$$\epsilon_p = \frac{\hbar^2 p^2}{2m}, \quad p = \left\{ \frac{\pi}{a}, q \right\}$$

At $q = 0$, a bandgap of $2|V_g|$ results [12, 346], which corresponds to the range of energies where the wave number has an imaginary part. Owing to the non-vanishing imaginary part of the wave vector, the corresponding wave functions decay exponentially in space, which is unphysical in a crystal with perfect periodicity. Therefore, no states exist within the bandgap.

Note that $E(q)$ is a continuous function in the complex plane. For energies in $[\epsilon_{\pi/a} - |V_g|, \epsilon_{\pi/a} + |V_g|]$, q becomes imaginary, and by substituting $q = -i\kappa_c$ (κ_c is a real number), Eq. (3.16) can be rewritten as

$$E(\kappa_c) = V_0 + \epsilon_{\pi/a} - \epsilon_{\kappa_c} \pm \sqrt{|V_g|^2 - 4\epsilon_{\pi/a}\epsilon_{\kappa_c}} \quad (3.17)$$

This energy dispersion is shown in Fig. 3.5. It can be measured across the whole bandgap in certain materials, for example in InAs, by tunneling experiments [232].

The plot of $E(\kappa)$ resembles a semicircle that connects E_V and E_C . How do the surface states emerge from these considerations? Well, an eigenstate close to the surface can exist, provided its wave functions can be properly matched to a wave function that decays exponentially into the vacuum. This scenario is schematically depicted in Fig. 3.6.

2) Of course, similar degeneracies exist at the boundaries of higher Brillouin zones.

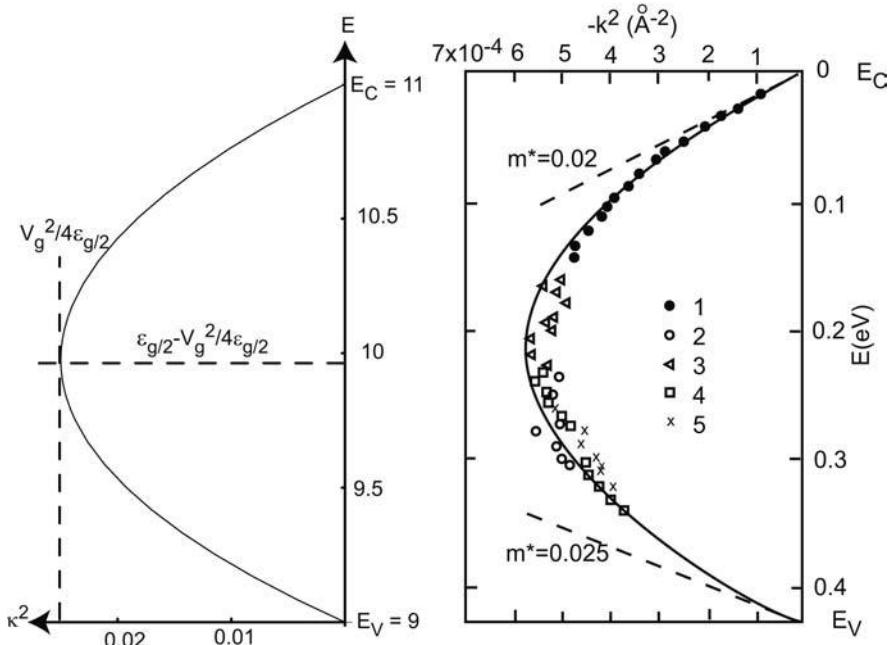


Fig. 3.5 Left: Imaginary energy dispersion inside the bandgap as obtained within the approximation of a weak periodic potential, according to Eq. (3.1). Here $\epsilon_{g/2} = 10$ and $V_g = 1$ were used as model parameters. Right: Measurement of the imaginary energy dispersion in InAs by surface barrier tunneling experiments. After [232]. The symbols correspond to different samples. Note that the energy is plotted vs. κ^2 .

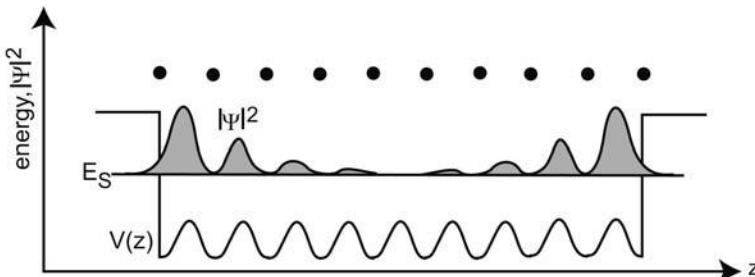


Fig. 3.6 Sketch of a surface state probability amplitude $|\Psi(z)|^2$ in a 10-atom crystal within the nearly free electron model. The potential is approximated by a harmonic function. The wave functions decay exponentially into the vacuum. Inside the crystal, the wave

function oscillates with the period of the lattice, while the amplitude drops exponentially as the distance to the surface increases. Note that there may be a phase shift of $\Psi(z)$ with respect to the position of the atoms.

Consider a crystal with a surface at $z = 0$. For weak periodic potentials, the total potential is given by

$$V(z) = \begin{cases} 0 & z \leq 0 \\ V_0 + 2V_g \cos(gz) & z > 0 \end{cases} \quad (3.18)$$

The wave function $\Phi_s(z)$ of a surface state is composed of two components,

$$\Phi_s(z) = \begin{cases} \Phi_v(z) \propto e^{\kappa_v z} & z \leq 0 \\ \Phi_c(z) \propto e^{-\kappa_c z} \cos [2(\pi/a)z + \phi_0] & z > 0 \end{cases}$$

where ϕ_0 denotes a possible phase factor. From the continuity conditions for $\Phi_s(z)$ and its first derivative with respect to z at $z = 0$, the condition

$$\cos^{-2} \phi = \frac{V_0 + V_g}{\epsilon_{\pi/a}} \quad (3.19)$$

can be derived after some lengthy algebra. If a solution for ϕ exists, the finite crystal has a surface state at energy

$$E_s = -\frac{\hbar^2 \kappa_v^2}{2m} \quad (3.20)$$

with

$$\kappa_v = \frac{\pi}{a} \tan \phi - \kappa_c, \quad \kappa_c = \frac{m V_g a}{\pi \hbar^2 \sin^2 \phi}$$

The details of these calculations can be found in the further reading at the end of this chapter.

Thus, the model of nearly free electrons gives qualitatively the same result as the tight binding model, namely that localized states may exist at the surface, with typical energies inside the bandgaps.

3.1.2

Surfaces of three-dimensional crystals

Extending the previous results to three dimensions is straightforward, although a quantitative treatment can be a formidable task. It requires not only symmetry considerations, but also inclusion of surface recombinations. The periodic pattern of chemical bonds is interrupted at the surface, and the unsaturated bonds (so-called “dangling bonds”) rearrange themselves to form a new electronic structure. This usually goes along with a change of the surface crystal structure. A famous example for such a surface recombination is the 7×7 reconstruction of Si(111) [38]. Another common scenario is chemical binding to a monolayer of adatoms which saturate the dangling bonds. In either case, it is clear that the electronic surface structure has little to do with the bulk structure.

A perfect crystal surface is certainly periodic within the surface plane, and it is reasonable to set up a tight binding model for the two-dimensional surface layer. Each surface state obtained within a one-dimensional model now corresponds to a two-dimensional band, with a bandwidth given by the transfer

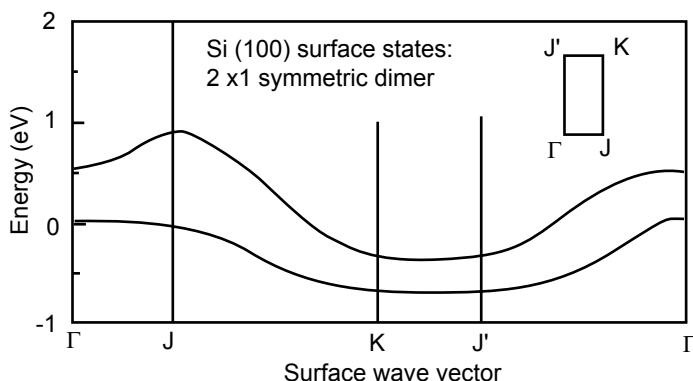


Fig. 3.7 Calculated energy dispersion of the two-dimensional bands of a Si(100) surface. After [51].

and overlap integrals between surface atoms. The number of states per unit area in a surface band corresponds to the density of surface atoms, n_s , which is of the order of $5 \times 10^{18} \text{ m}^{-2}$ [194]. Measurements of the surface density of states give typical results of the order of $4 \times 10^{18} \text{ m}^{-2} \text{ eV}^{-1}$, such that the width of the surface bands is of the order of the bandgap of the bulk material. More quantitative calculations of surface band energy dispersions (an example is given in Fig. 3.7) essentially confirm these simple considerations. We will represent surface bands in graphical representations as shown later in Fig. 3.8.

Independently of such issues, charge neutrality must be maintained at the surface. For a semiconductor with a neutral surface, this means that the number of occupied surface states must equal the number of states that have been removed from the valence band due to the formation of the surface band. Since the surface band with valence character can overlap with the surface band that has emerged from conduction band states, both types of surface bands can be partially filled (see Fig. 3.8). Surface states with valence band character can be regarded as donor-like. Likewise, we can call surface states with predominantly conduction band character acceptor-like.³

An important quantity is the “charge neutrality level” μ_{CN} . At this energy, the character of the surface states changes from predominantly donor-like to predominantly acceptor-like. Typically, μ_{CN} has an energy close to the center of the bandgap. For a neutral surface, the surface states are filled up to μ_{CN} . In general, the surface can be charged, however, and the chemical potential at the surface, μ_s , may differ from μ_{CN} .

3) In general, surface states do not have pure valence band (or conduction band, respectively) character. Rather, they are an admixture of both types of band states. Surface states will be more valence-like the closer their energies are to the valence band, and vice versa.

3.1.3

Band bending and Fermi level pinning

So far, we have considered intrinsic materials, where the energy of the electrons in the donor-like surface states increases as compared to their bulk value, but the surface remains neutral. This situation changes in doped semiconductors. Let us take an n-doped semiconductor as an example (Fig. 3.8).

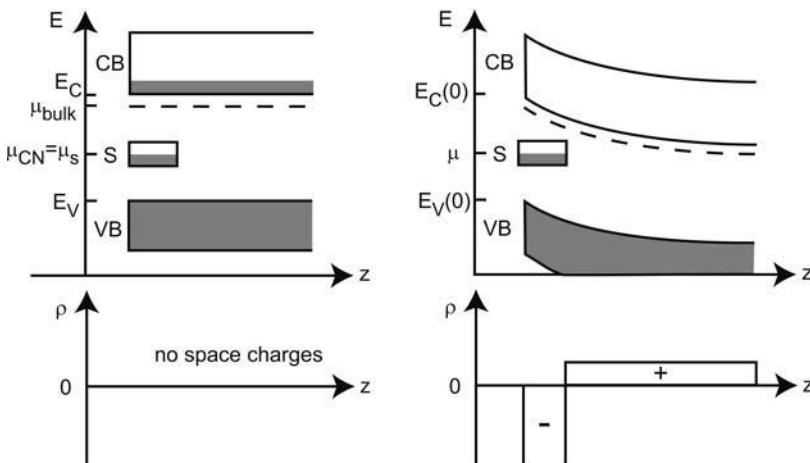


Fig. 3.8 Band bending at the surface of an n-doped semiconductor before equilibration. Gray areas indicate occupied states. Left: Schematic representation of the band structure close to the surface before equilibration. The sketched surface bands have a width of roughly $0.2E_g$. Right: After equilibration, the surface gets charged, an upward band bending results, and the Fermi level gets pinned close to the charge neutrality level μ_{CN} .

If only donor-like surface states existed, nothing would happen. However, usually both donor- and acceptor-like surface states are present. In that case, some donor electrons in the conduction band will reduce their energy by occupying the acceptor-like surface states. A negative surface charge is generated, counterbalanced by a positive space charge that originates from ionized donors within a depletion length z_{dep} away from the surface, such that overall charge neutrality is maintained. Consequently, an electric field and the corresponding electrostatic potential will build up, and the energy bands bend upwards as they approach the surface. It is self-evident that, in a p-doped semiconductor, the band bending will occur toward lower electron energies, since holes accumulate at the surface. In equilibrium, the chemical potential is constant throughout the crystal. To be somewhat more quantitative, consider an n-doped semiconductor with a band bending extending a distance z_{dep} from the surface into the bulk. In this region, the complete ionization of the donors leads to a space charge density of $\rho = en_D$. The missing electrons

occupy the surface states, with a surface electron density of $n_s = n_D z_{\text{dep}}$. The Poisson equation gives us the z -dependence of the potential $V(z)$ via

$$\frac{d^2V}{dz^2} = -\frac{en_D}{\epsilon\epsilon_0} \quad \Rightarrow \quad V(z) = -\frac{en_D}{\epsilon\epsilon_0}(z - z_{\text{dep}})^2 \quad (3.21)$$

for $z \in [0, z_{\text{dep}}]$. Note that this is the potential an electron feels as it approaches the surface from $z \geq z_{\text{dep}}$ *without* exiting the crystal. Therefore, the potential maximum of the conduction band is reached at $z = 0$ and equals

$$V(0) = -\frac{eN_D z_{\text{dep}}^2}{2\epsilon\epsilon_0}$$

Assuming that the surface states that get filled lie close to the center of the bandgap, we can estimate the depletion length for typical parameters, say $E_g = 1.4 \text{ eV} \Rightarrow V(0) \approx 0.7 \text{ eV}$, $n_D = 10^{24} \text{ m}^{-3}$, and $\epsilon = 12$, and find a depletion length of $z_{\text{dep}} \approx 30 \text{ nm}$. The band bending thus extends across many lattice constants, and the depleted region is much larger than the spatial extension of the surface states.

For our model parameters, a surface charge density of $n_s \approx 3 \times 10^{16} \text{ m}^{-2}$ results, which is much smaller than the integrated density of surface states. Since $D_s(E) \approx 5 \times 10^{18} \text{ m}^{-2}$, this means that the chemical potential at the surface μ_s changes only by a few meV due to the charge transfer. Hence, to a good approximation, μ_s does not depend on the doping density. This property is often coined by the statement that the Fermi level is *pinned* by the surface states at μ_{CN} . In that respect, surface states act similarly to deep dopants used to generate semi-insulating materials – see Chapter 2.

3.2

Semiconductor–metal interfaces

The ideas of wave function matching and modified transfer integrals at surfaces are also applicable to the important metal–semiconductor interface. Before the metal and the semiconductor get in contact, their common energy scale is the vacuum level, and the relative position of the bands in both materials is trivial. As the interface is formed, however,⁴ the local lattice structure at the interface changes, which can give rise to interface states. We distinguish

4) We assume that the surfaces are clean, and contain no oxides. This is in fact more or less the case in real devices, as possible oxide layers can be etched away, and the metal is usually deposited on top of the semiconductor in a high-vacuum environment. Furthermore, the interfaces considered in theory are usually atomically flat. This is quite hard to achieve experimentally.

between *Schottky barriers*, where charge carriers have to tunnel through a barrier as they move across the interface, and *ohmic contacts*, where such a barrier is absent or highly penetrable.

3.2.1

Band alignment and Schottky barriers

The character of the previously mentioned interface states depends on the energy. At energies in the gap between the full bands of both materials, localized states may form, with a character very similar to that of surface states. For energies inside a full band of both materials, the semi-extended wave functions have to be matched. Both situations have no further consequences, as all the states involved are filled anyway. A new scenario is obtained for energies that lie in an energy band of one component, but in a gap of the second one. The most relevant case is a semiconductor bandgap with energies inside the conduction band of the metal, and we focus on this scenario.

Heine [147] has shown that the metallic wave function can be matched to the evanescent wave functions in the semiconductor for all energies. Hence the metal induces a continuum of interface states in the semiconductor bandgap, so-called *induced gap states* (IGSs). Close to the interface, the semiconductor thus develops a non-zero density of states at energies inside the metallic bands (Fig. 3.9). Suppose that, at the beginning, both materials are well separated and do interact. Clearly, their common energy scale is the vacuum level, and the energy difference between the chemical potential in the metal μ_M and the electron affinity in the semiconductor $\xi_{e,SC}$ equals the energy difference between the semiconductor conduction band bottom and μ_M (Fig. 3.7). In many, but not all, cases this means that μ_M is somewhere in between the top of the valence band and the bottom of the conduction band of the semiconductor. Next, we assume that the surfaces are so close to each other that IGSs are formed in the semiconductor bandgap. Fig. 3.10 shows a calculation of the density of states across a GaAs–Al interface. Within the first two GaAs double layers away from the interface, a significant density of IGSs is generated, while the Al density of states remains essentially unchanged even for the Al layer at the interface.

Since the IGSs inside the semiconductor are built from the virtual gap states of the semiconductor, their properties are semiconductor-specific. Like a surface state, such an interface state is predominantly acceptor-like or donor-like. A charge neutrality level and the interface work function of the semiconductor Φ_S can be defined, and again the integrated density of states must remain constant. How do the band structures of the metal and the semiconductor align with respect to each other? In general, the charge neutrality level in the

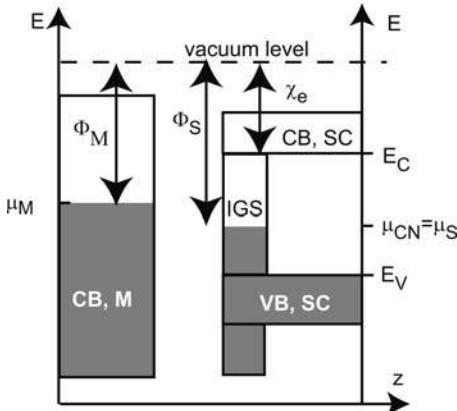


Fig. 3.9 Typical energy band alignment between a metal (left) and a semiconductor (right) before charge transfer across the interface is allowed. Extended states in the metal induce gap states in the semiconductor at all energies inside the bandgaps. The induced gap states (IGSs) are filled up to μ_{CN} , and the common energy scale of both band structures is the vacuum level.

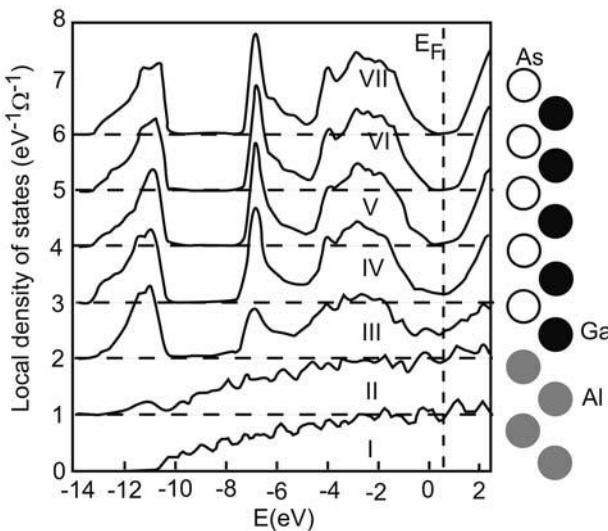


Fig. 3.10 Calculated density of states across an Al–GaAs interface.
After [33].

semiconductor μ_{CN} will be different from μ_M , and a charge transfer will take place.

We discuss the alignment using the interface between a metal and an n-doped semiconductor as an example (Fig. 3.11(a)). In a gedanken experiment, we assume that, for now, the donor electrons are not allowed to occupy surface

states. Charge transfer across the interface forms a dipole and aligns, via the dipole potential that obeys the Poisson equation, μ_M with μ_S . This dipole is strongly localized, since the IGSs only extend over a few lattice constants into the semiconductor (Fig. 3.11(b)). The semiconductor bands have been bent upwards in the case drawn here, and this energy difference adds to the difference between Φ_M and ξ_e . The barrier at the interface is known as the Schottky barrier V_S . In an n-doped semiconductor, the IGSs will get occupied by donor electrons as well. As in the previous section, a space charge layer builds up in the semiconductor close to the interface and generates the band bending sketched in Fig. 3.11(c). Since the region of depleted donor electrons z_{dep} is much larger than the width of the IGSs, the band bending due to the interface dipole is often drawn as a step function, resulting in a band diagram as shown in Fig. 3.11(d).

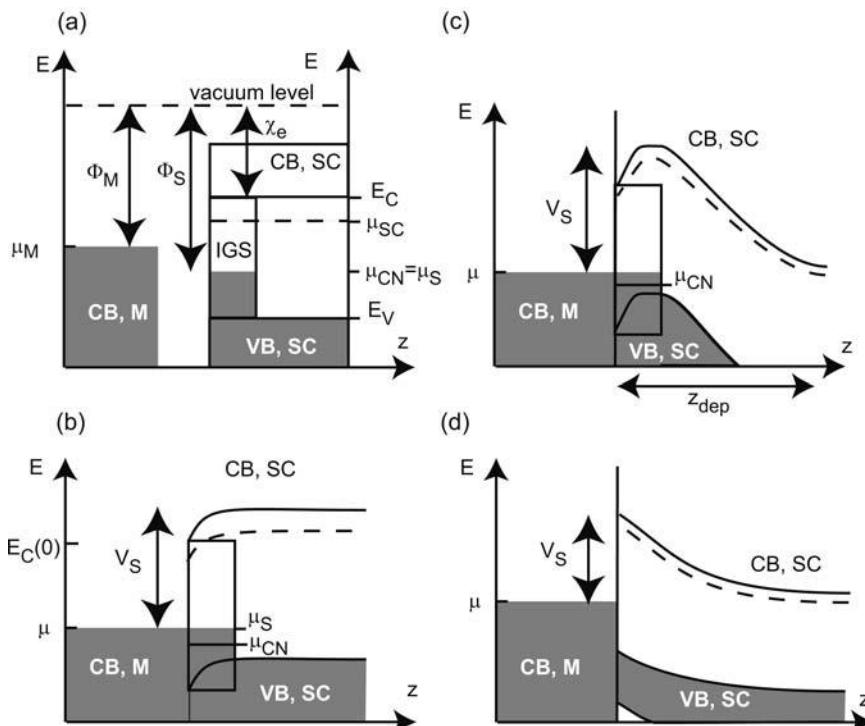


Fig. 3.11 Metal-induced gap states and band alignment at a metal–semiconductor interface. (a) Relative band energies of a metal and an n-doped semiconductor after formation of IGSs, but before charge is transferred. (b) Charge transfer across the interface generates an interfacial dipole, aligns μ_M with μ_S , and creates a Schottky barrier V_S . (c) The donor electrons will occupy the IGSs as well, generating a depletion layer of width z_{dep} and an additional band bending. (d) Since z_{dep} is much larger than the width of the IGSs, the interface band bending is usually drawn as a sharp step function, as indicated.

and generates a highly localized band bending. (c) The donor electrons will occupy the IGSs as well, generating a depletion layer of width z_{dep} and an additional band bending. (d) Since z_{dep} is much larger than the width of the IGSs, the interface band bending is usually drawn as a sharp step, as indicated.

Similarly, of course, an opposite bending can occur for a p-doped semiconductor.

Within this picture, we would expect V_S to depend upon Φ_M , and this is in fact the case. Fig. 3.12(b) shows that an approximately linear relation exists between V_S and Φ_M , with a slope characteristic for each semiconductor ($d(eV_S)/d\Phi_M \approx 0.05$ for n-GaAs and ≈ 0.25 for Si).

3.2.1.1 The Schottky model

The Schottky model neglects the consequences of interface states and models the Schottky barrier formation entirely by using bulk parameters [266]. It is frequently used, and we thus briefly sketch it here, using again an n-doped semiconductor as an example. In a consideration similar to the previous one, we start with the two materials separated by an impenetrable tunnel barrier, such that charge transfer is impossible. As pointed out above, the difference between the metal Fermi level and the conduction band bottom of the semiconductor is $\Phi_M - \xi_{e,SC}$. Bringing the materials closer together will at some point allow for charge transfer. In the case depicted here, the donor electrons get transferred into the metal until the chemical potentials μ_M and μ_{SC} are aligned (Fig. 3.12(a)). The Schottky model thus predicts that $V_S = \Phi_M - \xi_{e,SC}$, which can be taken as a coarse approximation to the observed barrier heights shown in Fig. 3.12(b).

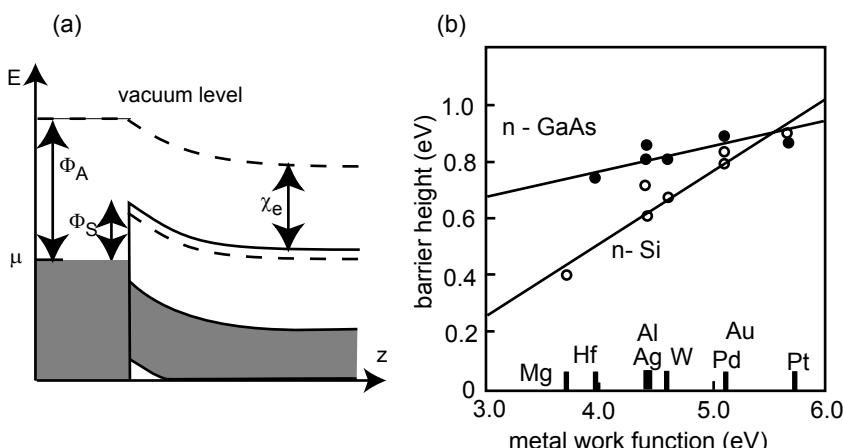


Fig. 3.12 (a) Positions of the Fermi levels of a metal and an n-doped semiconductor in equilibrium as obtained within the Schottky model. (b) Schottky barriers of Si and GaAs in contact with different metals, plotted as a function of the metal work function. After [294].

3.2.1.2 The Schottky diode

Semiconductor–metal interfaces with a Schottky barrier act as a diode. The resistance of such a barrier is dominated by the depletion zone in the semiconductor, as well as by the Schottky barrier height. Applying a positive voltage to the metal with respect to the grounded semiconductor pulls the electrons in the semiconductor toward the interface, thus reducing the width of the space charge layer, as well as the height of the barrier. The current is typically dominated by electrons with a sufficiently large thermal energy to diffuse across this barrier. Since the distribution function in this high-energy tail can be approximated by a Boltzmann distribution, the density of such electrons increases exponentially with applied voltage, and therefore an exponential (diode-like) I – V characteristic results (Fig. 3.13). Although we will not make use of Schottky barriers as diodes in subsequent chapters,⁵ it is important to keep in mind how the band bending in a semiconductor close to a metal can be modified by a voltage applied across the Schottky barrier.

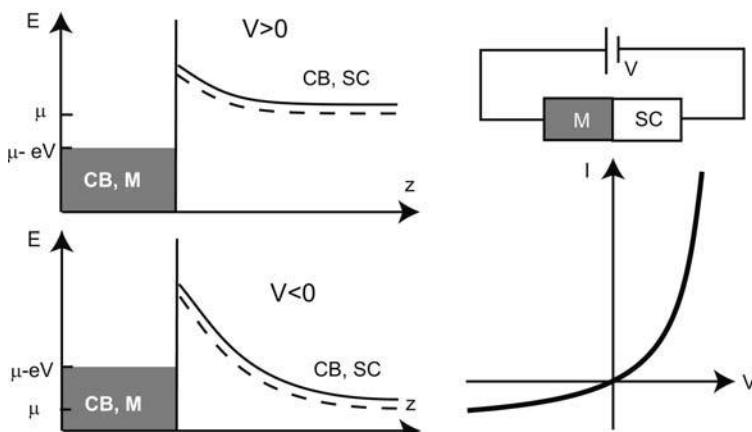


Fig. 3.13 The tunnel barrier formed by the depletion layer at a Schottky contact depends on the bias voltage (left), which results in a diode-like current–voltage characteristic (lower right). The upper right scheme shows the direction of the applied voltage.

3.2.2

Ohmic contacts

So far, we have assumed that the chemical potential at the interface in equilibrium is at some energy inside the semiconductor bandgap. This is not always what we need. Operating a semiconductor device requires current carried

5) In some setups, the currents across Schottky contacts actually disturb the experiment. Such currents are referred to as gate leakage, for reasons that will soon become clear.

by a metal wire to be fed into the semiconductor at some point. This should be done with the lowest resistance possible, and a diode-like current–voltage characteristic would certainly be an annoyance. The interface should behave according to Ohm's law.

Such ohmic contacts can be formed naturally at some metal–semiconductor interfaces where the equilibrium Fermi level lies above $E_{C,SC}$; see Fig. 3.14(a). No tunnel barrier is formed in that case, and charge can flow freely across the interface. An example is the interface between InAs and a metal. Most material combinations of importance, however, form Schottky barriers, and we have to design some sort of ohmic contact. This is done by reducing the Schottky barrier to insignificant heights and widths by two means. First of all, a metal is used for ohmic contact formation, such that V_S is reduced as much as possible. Second, the semiconductor is heavily doped, which reduces the width of the tunnel barrier, according to Eq. (3.21); see Fig. 3.14(b). Since the resistance across the Schottky barrier depends exponentially on both the width and the height of the tunnel barrier, a current–voltage characteristic results that is ohmic for practical purposes.

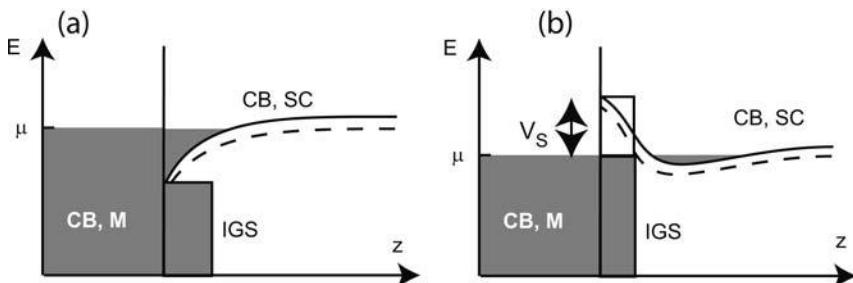


Fig. 3.14 (a) An ohmic contact without a Schottky barrier between the metal and the semiconductor. (b) Scheme of a metal–semiconductor interface with a Schottky barrier that works as an ohmic contact.

3.3

Semiconductor heterointerfaces

The probably best known example of a semiconductor interface is the p–n junction, where a p-doped region meets an n-doped region of the same semiconductor host crystal. Close to the junction, donor electrons recombine with acceptors and generate a space charge region, which can be tuned by applying a voltage across the junction. The p–n junction is the basis of bipolar devices, but plays no role in our subsequent considerations; the interested reader is referred to standard textbooks on solid state physics and semiconductor physics [12, 270].

Instead, the semiconductor devices we will study frequently contain *heterointerfaces*, i.e. interfaces between two different semiconductor crystals. As in the metal–semiconductor interface, an important question is how the band structures align with respect to each other. We can get a qualitative understanding of the alignment mechanism by modifying the above considerations accordingly, and consider an interface between semiconductor 1 (n-doped) and semiconductor 2 (p-doped) as an example (Fig. 3.15).⁶ Again we begin with the relative band energies of the two semiconductors before charge transfer has taken place (Fig. 3.15(a)). For energies inside the bandgap of one semiconductor and inside an energy band of the second semiconductor, IGSs are generated. Additional states may arise for energies in the bandgap of both materials, which we neglect for simplicity. In any case, surface chemical potentials are defined for both materials, which may differ from the chemical potentials of the bulk as well as from each other. Correspondingly, charge transfer will take place across the interface. Fig. 3.15(b) shows the resulting band alignment after electrons have been transferred from the valence band in semiconductor 2 into the IGSs of semiconductor 1. Finally, the bulk chemical potentials will be aligned by electron transfer from the n-doped semiconductor into the p-doped semiconductor, and a space charge dipole layer, with a typical extension of several tens of nanometers into both materials, is obtained. This is shown in Fig. 3.15(c), where the dipole potential due to the occupation of IGSs has been included already in the band offsets, similar to Fig. 3.9(d).

Question 3.2: Discuss the band alignment between two semiconductors when surface effects are neglected, i.e. a model similar to the Schottky model for the metal–semiconductor interface.

Phenomenologically, one distinguishes between several types of alignment (see Fig. 3.16):

Type I The smaller bandgap lies completely inside the larger bandgap. An important example is the $\text{GaAs}-\text{Al}_x\text{Ga}_{1-x}\text{As}$ interface discussed below in detail.

Type II Here, both bands of crystal 1 lie above the corresponding bands of crystal 2. The $\text{InAs}-\text{AlSb}$ interface has such a structure. In the “staggered” alignment, one of the band edges of material 1 resides inside the

6) Doping is not necessary for the alignment mechanisms to work, but simplifies the discussion somewhat. In intrinsic materials, polarization charges contribute significantly to the dipole at the interface.

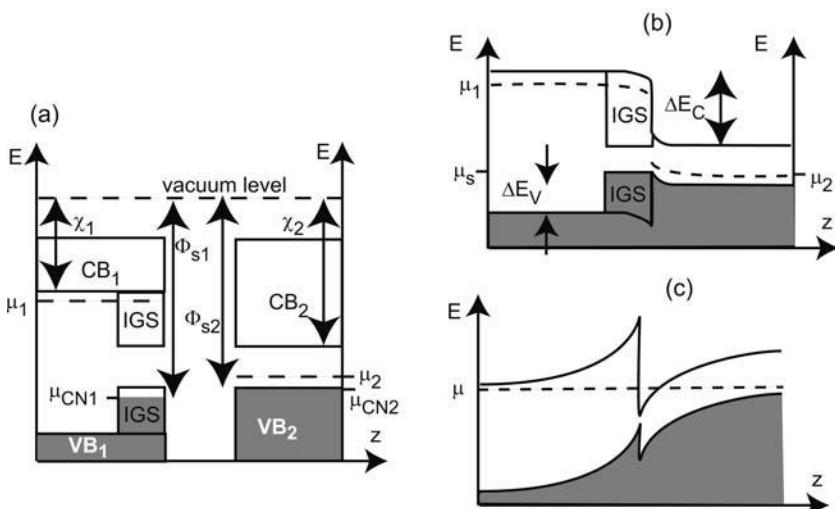


Fig. 3.15 Band alignment at a semiconductor heterointerface. (a) Both band structures are drawn with respect to the vacuum level before charge is transferred. Gap states are induced, along with possible localized states in the bandgap of both materials (not shown). (b) The band alignment is modified by a charge dipole, which aligns the surface chemical potentials. Here, we have kept the doping charges immobile, such that the re-

sulting band alignment is the same as for the intrinsic case. (c) In the case of doped semiconductors, the bulk chemical potentials will align by transfer of doping charges across the heterointerface and the corresponding formation of depletion layers, which are much larger than the spatial extension of the IGSs. Here, the band offsets are drawn as sharp steps, neglecting the spatial extension of the IGSs.

bandgap of material 2; while in the “misaligned” type, the top of the valence band of material 1 lies above the bottom of the conduction band of material 2. The most prominent example for this type of alignment is the InAs–GaSb interface.

To finish this section, it should be mentioned that very similar considerations give the band alignment between a semiconductor and an insulator.

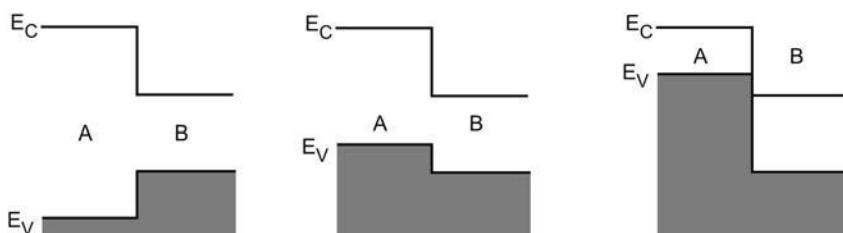


Fig. 3.16 Different types of band alignments at semiconductor heterointerfaces: type I (left); type II staggered (center); and type II misaligned (right).

Several theoretical treatments of band alignments have been developed. The concept of electronegativity, introduced by Pauling [234], has proved highly useful in molecular physics and can be applied to crystal interfaces as well. An instructive evaluation of this approach is given in Paper P3.1. Furthermore, various extensive tight binding models for interfaces have turned out to be highly successful. These concepts are developed in [142], and references therein.

It should be remarked that agreement between theory and experiment is often hampered by imperfect interfaces containing defects or impurity atoms. A problem we have excluded is interface strain due to different lattice constants. With state-of-the art technology, however, close-to-perfect interfaces can be grown, and the measured band alignments agree well with more sophisticated theoretical considerations – see e.g. [213].

3.4

Field effect transistors and quantum wells

The properties of interfaces can be used to construct useful devices as well as fascinating nanostructures. Field effect transistors are very important in both respects. Many mesoscopic samples comprise some sort of field effect transistor, which are frequently denoted by the acronym FET. These devices rely heavily on interface effects. The two most important FETs in our context are the Si MOSFET and the GaAs HEMT. These are by no means the only systems though. Particularly in research, a wide variety of heterostructure devices is used. Some examples are given at the end of this section.

3.4.1

The silicon metal–oxide–semiconductor field effect transistor

This type of FET is the basic building block of the vast majority of present-day integrated circuits. A scheme of the Si MOSFET is shown in Fig. 3.17(a). A silicon chip is, say, p-doped and electrically contacted with two ohmic contacts that act as source and drain. A metal electrode resides in between the ohmic contacts, separated by a SiO_2 layer from the Si. This M–O–S layer sequence can be thought of a Schottky diode with an insulator inserted at the M–S interface, in order to increase the resistance strongly. Currents between the metal electrode and the semiconductor are neglected in the following. With no voltage applied, the resulting band structure across the interface is shown in Fig. 3.17(b). The p-doping is typically rather weak, say $N_A \approx 10^{21} \text{ m}^{-3}$, such that the resistivity of the Si is high. By applying a voltage to the metal electrode with respect to drain, a band bending is induced in the Si, and a corresponding charge accumulation at the Si– SiO_2 interface is generated, as

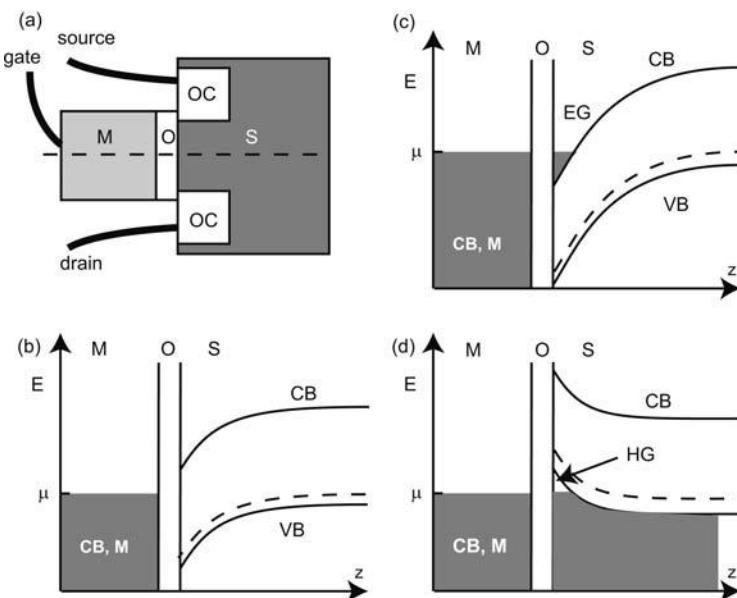


Fig. 3.17 (a) Schematic illustration of a silicon MOSFET. A source–drain voltage is applied to a p-doped silicon wafer at two ohmic contacts (OC). A metal electrode M (“gate”) in between the ohmic contacts is separated from the silicon by a SiO_2 layer. (b) Band alignment across the M–O–S interface (dashed line in (a) for $V_g = 0$). (c) Applying a positive voltage to the gate increases

the band bending. Above a threshold gate voltage, the conduction band bottom drops below μ at the O–S interface, and an electron gas (EG) is induced (inversion). (d) A sufficiently large negative gate voltage pulls the top of the valence band above μ , and a hole gas (HG) is generated at the surface (accumulation).

depicted for the case of a positive voltage in Fig. 3.17(c). Here, E_C of the Si has dropped below the Fermi level, and electrons collect at the interface *in the conduction band*. Hence, an electron gas is generated which is confined in the z -direction, but free in the directions parallel to the surface. For sufficiently high electron densities in this free electron gas, its conductance is much higher as compared to the p-doped bulk. We speak of *inversion* if the free carrier gas has the opposite sign than the carriers in the bulk due to doping. For appropriate doping densities, we can generate a free hole gas at the O–S interface by applying negative voltages to the metal electrode. This situation is referred to as *accumulation*. Devices that offer the possibility of generating both electron and hole gases are known as *ambipolar*.

The current that flows between source and drain can thus be controlled by the voltage applied to the metal electrode, which is therefore known as the *gate*. The oxide prevents a current flowing between the gate and the silicon, which would reduce the performance of the switch. This three-terminal device thus represents a transistor that relies on the electrostatic field effect.

However, we are not so much interested in the technological applications of MOSFETs in our context. Readers interested in the origin of the current-voltage characteristics of MOSFETs or in other characterizations of technological relevance are referred to [294]. Rather, we focus on the electron gas that can be formed at the O–S interface in Fig. 3.17(c). Apparently, its spatial extension in the z -direction is very small, as we have seen already above. Typically one finds that E_C is below the Fermi level for about 20 nm. Furthermore, the electron densities in such interface layers are much smaller than metallic densities, and the Fermi wavelength is larger. A crude estimation gives $\lambda_F \approx 20$ nm. Therefore, we expect size quantization effects in the electron gas. Fig. 3.18 shows a zoom-in of the conduction band structure at the oxide–semiconductor interface. The potential is roughly triangular. By applying an appropriate gate voltage, a situation can be established in which only the energy of the first quantized state is below the Fermi level. Since the electrons are not confined parallel to the interface, a *two-dimensional electron gas* (2DEG) results. The conduction band bottom of this 2DEG is at E_0 in Fig. 3.18. We sometimes speak of a two-dimensional subband. If more than one subband is occupied, the electron gas is said to be quasi-two-dimensional.

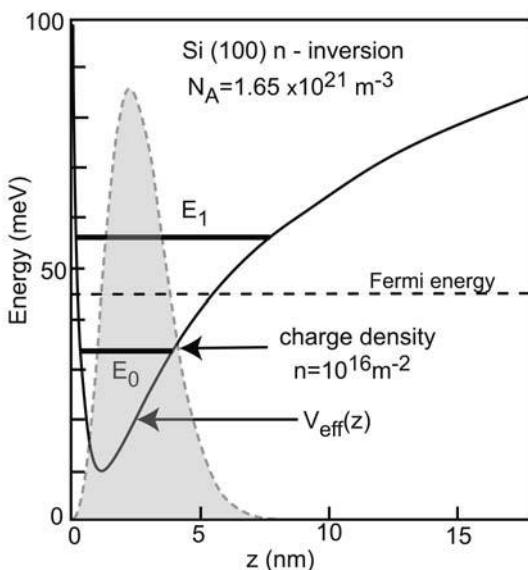


Fig. 3.18 Energy diagram of the conduction band in a Si MOSFET close to the O–S interface, as obtained from a self-consistent calculation that includes electron–electron interactions and screening. In a potential well, quantized states are formed. The resulting electron gas is effectively two-dimensional as long as only the first quantized state lies below the Fermi level. Also indicated is the electronic wave function. After [9].

Like in three dimensions, this electron gas can be described by an effective mass and by the two-dimensional density of states. However, some care is required in adopting the bulk parameters to a two-dimensional carrier gas at an interface. We will meet some of the related issues later on. For now, we just look at the effective electron mass of the 2DEG in the Si MOSFET. Suppose the Si crystal plane at the interface is a (100) plane, a very common case. The electrons move freely parallel to this plane only. Therefore, it is self-evident to project the valley-degenerate Fermi ellipsoids into that plane (see Fig. 3.19), which results in four spin-degenerate ellipses and a twofold valley-degenerate and spin-degenerate circle at the center. Owing to interface effects, however, the degeneracy between the ellipses and the circles gets removed, and the conduction band at the circle is about 20 meV below the conduction band minimum in the ellipses [10]. At room temperature, both types of minima are occupied. At low temperatures, however, the electrons have a single effective mass of $m^* = 0.19m_e$ parallel to the surface, and the valley degeneracy is reduced to 2.

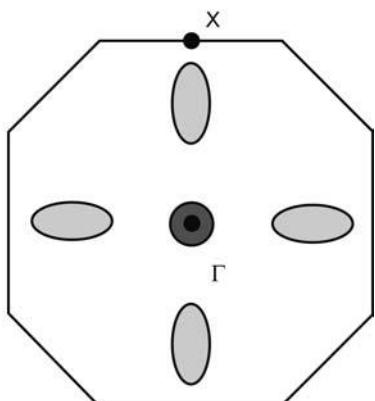


Fig. 3.19 Projection of the Si Fermi surface for typical electron densities onto a (100) plane. Two ellipsoids get projected onto the Γ -point. Their energy is reduced as compared to the four projected ellipses due to interface effects.

The two-dimensional character of this interface electron gas has some most surprising consequences, as will be seen in Chapter 6. But this is not the only interesting property of such electron gases. Furthermore, the electron densities are much smaller than in conventional metals, and can be tuned. The Fermi wavelength is comparatively large, and size quantization effects can be expected laterally also, provided the MOSFET is patterned accordingly. In addition, low density means that electron-electron interactions are more important, due to reduced screening.

Since the electrons are to some degree spatially separated from the ionized donors, impurity scattering is reduced and the electron mobility increases. In fact, the mobility of an electron gas at a O-S interface can be two orders of magnitude larger than the mobility of bulk Si. The mobility is typically dominated by scattering at impurities embedded in the oxide. Furthermore, the oxide is amorphous. The oxide atoms are by no means periodically arranged, which will cause additional electron scattering. However, due to size quantization, the probability of finding electrons right at the O-S interface is reduced (see the wave function in Fig. 3.18). The maximum of the probability density is several nanometers away from the interface.

3.4.1.1 The MOSFET and digital electronics

Microprocessors are essentially Si MOSFET circuits which can perform calculations in the binary system. A second major component of digital electronics are random access memories (RAMs). They are also built from MOSFETs. The Si MOS material system is the system of choice due to its vastly superior properties. First of all, extremely pure and large Si crystals can be grown, while the material resource is abundantly available. Second, Si has a natural oxide with excellent characteristics, such as a high breakdown electric field, good mechanical stability, and inertness to most chemicals.

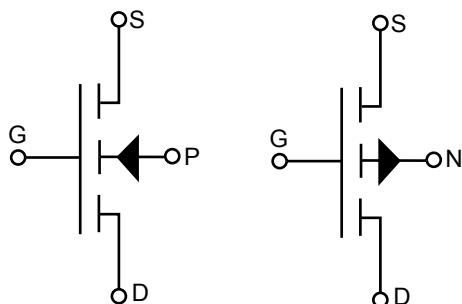


Fig. 3.20 Symbols for n-channel (left) and p-channel (right) MOSFETs.

Within digital electronics MOS technology, the complementary MOS (CMOS) concept has become dominant. *Complementary* means that both p-doped and n-doped transistors – termed n-channel and p-channel transistors, respectively (see Fig. 3.20), due to the carrier type in the interface channel under inversion – are used in one circuit. It sounds much more expensive to define two kinds of transistors on one chip, so we should expect to have a good reason for doing this. In fact, CMOS circuits have a much lower power consumption. This can be seen by looking at the realizations of a binary inverter, the *NOT* gate, in NMOS (i.e. in a circuit that uses n-channel MOSFETs

exclusively) and in CMOS technology (see Fig. 3.21). This circuit is supposed to deliver a zero as output for input 1 and vice versa. In CMOS technology, the logic levels are defined as voltages $V \in [-0.2 \text{ V}, 1.6 \text{ V}] \equiv 0$ and $V \in [3.5 \text{ V}, 5.5 \text{ V}] \equiv 1$. In NMOS, this task can be implemented as shown in Fig. 3.21(a). The input (A) is applied to the gate of an n-channel MOSFET, which is biased via a supply voltage ($V = 5 \text{ V}$) that drops across the MOSFET channel in series with a properly selected resistor R . The output (\bar{A}) is picked up between the resistor and the MOSFET entrance. For $A = 0$, the MOSFET channel is closed and the output equals approximately the supply voltage. For $A = 1$, the channel is open, which sets the output to ground. Suppose the channel resistance is 100Ω in the open and $10 \text{ M}\Omega$ in the closed configuration. Then, a reasonable choice of R would be $100 \text{ k}\Omega$, which means a power consumption in the open channel state of $P \approx V^2/R = 250 \mu\text{W}$. This power can be greatly reduced by using the corresponding CMOS circuit in Fig. 3.21(b): there, the resistor is replaced by a p-channel MOSFET, with the gate connected to the input as well. As can be easily seen, this setup also acts as an inverter, but the power consumption is determined by a closed MOSFET channel in both states. In our above example, P drops by two orders of magnitude.

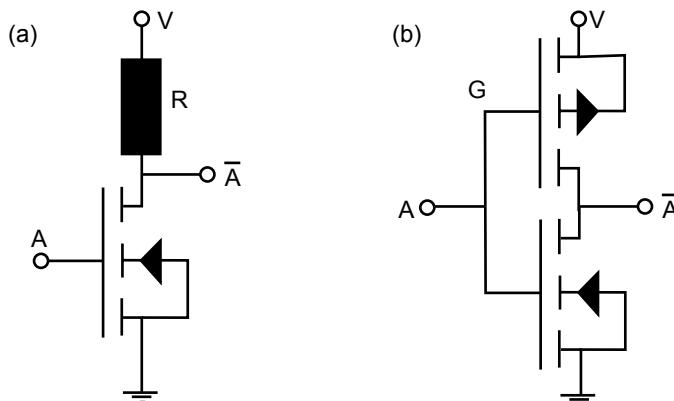


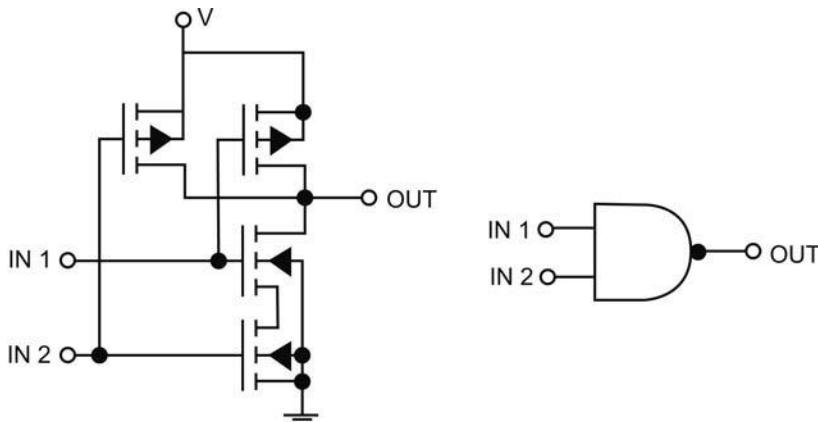
Fig. 3.21 The NOT gate (binary inverter), realized in (a) NMOS and (b) CMOS technology.

It is straightforward to implement more complex circuit elements. As an example, we consider a possible CMOS realization of a memory cell, which is based on the *flip-flop* and finds an application in *static random access memory* (SRAM). Such a cell can be designed by establishing a mutual feedback between two NAND gates that perform the NOT AND operation on two inputs according to the truth table shown in Table 3.1.

A possible circuit layout for realizing a NAND gate is given in Fig. 3.22: two p-channel MOSFETs are arranged in parallel between the supply voltage and the output connection, while two n-channel transistors are located in series

Tab. 3.1 Truth table for the NAND operation.

Input 1	Input 2	Output
0	0	1
0	1	1
1	0	1
1	1	0

**Fig. 3.22** Layout of a CMOS NAND gate (left) and its symbolic representation (right).

between the drain and the output. Each of the two input signals is applied to the gates of one n-channel and one p-channel MOSFET. The output voltage is on “0” if both n-channel MOSFETs are open, and on “1” if just one of the p-channel MOSFETs is in the conducting state.

Two NAND gates can be the building blocks of a storage cell, when coupled as shown in Fig. 3.23. The output of each NAND gate is coupled to one of the inputs of the other gate, while the remaining inputs are the two external input signals. The output of the upper gate is defined as the stored bit M. In the “hold” configuration, both inputs are held at “1”, which allows two stable configurations:

- M is at “1” and \bar{M} is at “0”.
- M is at “0” and \bar{M} is at “1”.

The stored bit M is fixed and can be read any time by, for example, applying it to a gate of a MOSFET.

Suppose we would like to write $M = 0$. This is achieved by putting IN1 to “0” for a short time, while keeping IN2 at the “1” level. Independently of the previously stored bit, M goes to “1” and \bar{M} to “0”. After the bit has been

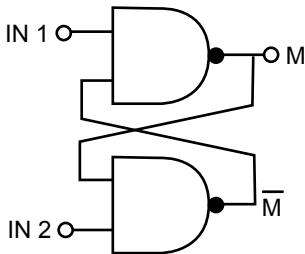


Fig. 3.23 Realization of a flip-flop with two NAND gates.

written in the cell, IN 1 is set to “1” again. Similarly, $M = 0$ is deposited in the cell by giving a “0” pulse to IN 2. Note that “0” at both inputs has to be avoided, since an unstable situation arises.

This example concludes our brief excursion into the field of digital electronics. How the numerous elements required in a microprocessor are implemented in CMOS technology is not only fascinating, but also of utmost technological relevance. For more information, the reader is referred to the specialized literature, like [325].

3.4.2

The Ga[Al]As high electron mobility transistor

In this system, the two-dimensional electron gas is generated inside the GaAs, at the interface formed between $\text{Al}_x\text{Ga}_{1-x}\text{As}$ and GaAs. The band alignment of this interface is of type I. The band offsets depend on x (see Fig. 2.6). A typical choice is $x = 0.3$. In that case, the conduction band of $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ is 300 meV above that of GaAs. The top of the $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ valence band is located about 160 meV below that of GaAs. This is of no further interest here, as we are going to consider an electron gas again.

In contrast to Si, the GaAs remains undoped. Instead, the electrons are provided by a doping layer inside the $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$. Usually, Si is used as a donor. The doping layer can be spatially separated from the $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ by several tens of nanometers (see Fig. 3.24(a)). While most of the doping electrons that get thermally excited into the conduction band occupy the nearby surface states, some of them (typically about 10%) reduce their energy by falling across the interface into the GaAs conduction band. This doping technique is called *modulation doping*; it was first demonstrated by Dingle [72]. An accurate doping density is essential in designing a good HEMT structure. Only a few percent deviation from the correct doping density can have either of two effects: mobile electrons are produced in the doping layer (a “bypass”), or the triangular potential at the heterointerface remains empty. While the dop-

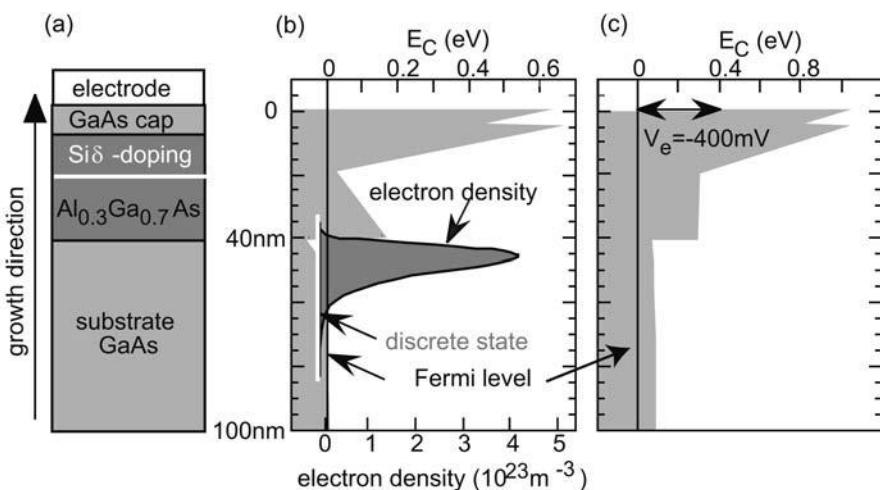


Fig. 3.24 (a) Band alignment at a modulation-doped GaAs-Al_xGa_{1-x}As interface. (b) Schematic structure of a GaAs HEMT with the gate electrode grounded. (c) For gate voltages below -400 mV , E_C at the interface moves above the chemical potential, and the electron gas is depleted.

ing density and the thickness of the spacer layer determine the density of the 2DEG, it can be tuned with a top gate over wide ranges.

Consequently, two charge dipoles build up, one between the surface and the doping layer, and a second one between the GaAs-Al_{0.3}Ga_{0.7}As heterointerface.⁷ This results in the band structure sketched in Fig. 3.24(b). As in the Si MOSFET, the resulting electron gas can be two-dimensional, and its carrier density can be tuned by applying voltages to a gate on top of the heterostructure (see Fig. 3.24(c)). Thus, the electron gas is present in this structure if no gate voltage is applied, or if there is no gate at all. Modulation doping of GaAs heterostructures caused great progress in the electron mobilities (Fig. 3.25). The reason is twofold. First of all, Al_xGa_{1-x}As is quasi-crystalline, in contrast to the SiO₂ layer in a Si MOSFET. Although the Al atoms replace the Ga atoms at random sites, this ternary compound is a somewhat distorted zinc blende crystal structure with a well defined lattice constant. The lattice mismatch between GaAs and Al_{0.3}Ga_{0.7}As is only 0.4%. Hence, the electrons in the 2DEG see an almost perfectly periodic environment, and the interface causes much less scattering as compared to the O-S interface in a Si MOSFET. Second, the ionized donors, which are a strong source of scattering, are spatially separated from the electron gas. Consequently, the screened Coulomb potentials that the electrons see are much weaker and generate predominantly small-angle scat-

7) Note the thin GaAs cap layer at the surface. Its purpose is to avoid oxidation of the Al_{0.3}Ga_{0.7}As layer when exposed to air.

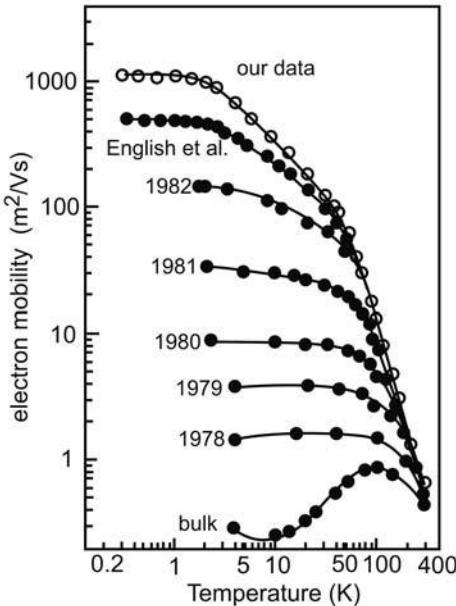


Fig. 3.25 Evolution of electron mobilities over time, after modulation doping was introduced. After [240]. Please compare this graph to Figs. 2.15 and 3.31.

tering. In the years 1978 to 1985, the layer sequences and the compositions of Ga[Al]As HEMTs were improved, and the increase in low-temperature electron mobilities achieved in this period was truly remarkable (see Fig. 3.25). For example, an extremely high mobility of $\mu = 1440 \text{ m}^2/\text{Vs}$ has been reported [307]. This corresponds to a mean free path of $120 \mu\text{m}$. Although very similar devices can be built of several materials, like Ga[Al]N for example, the Ga[Al]As heterostructure has remained unsurpassed in terms of electron mobility.

Another advantage of the Ga[Al]As system is the possibility of designing the spatial variation of the band structure by controlling the Al content during sample growth. For example, quantum wells can be grown by embedding a thin layer of GaAs in two $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ layers. Varying the Al content parabolically during growth, i.e. $x \propto (z - z_0)^2$, results in a parabolic quantum well in the growth direction (see Fig. 3.26). Hence, quantum mechanical model potentials can be experimentally realized this way, as long as the envelope function approximation is reasonable. We will occasionally meet such structures later on.

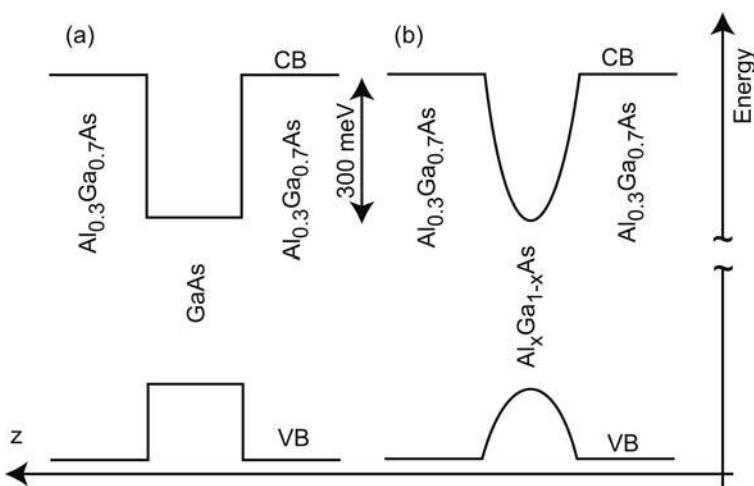


Fig. 3.26 Sketches of (a) square and (b) parabolic quantum wells, fabricated in the Ga[Al]As system. In (b), the Al concentration x is varied $\propto z^2$ around the center of the well, up to $x = 0.3$.

3.4.3

Other types of layered devices

We conclude this section with a selection of further interesting heterostructures. In particular, the Si[Ge] and the InAs–AlSb quantum wells are presented. Also, we will have a look at organic FETs. The materials cannot be combined arbitrarily, though. The lattice constants of the two components that form the interface should differ as little as possible. Differences in the lattice constants will inevitably lead to strained layers, which generates lattice dislocations and thus additional scattering. If the strain gets larger than $\approx 1\%$ homogeneous film growth is no longer possible, and strained islands of one material form instead. While these islands have fascinating properties (see the following chapter), they are of course unacceptable when a clean and homogeneous interface is required. A plot of the bandgap of different semiconductors vs. their lattice constant is known as the bandgap engineer's map. It reveals what kind of materials can possibly be combined (see Fig. 3.27).

3.4.3.1 The AlSb–InAs–AlSb quantum well

The band alignment of this material system is type II misaligned. Fig. 3.28 shows the band structure of an InAs quantum well, sandwiched in between two layers of AlSb. The surface is again capped with a binary crystal that does not contain Al, GaSb in this case. Modulation doping of this system is a tricky business. The standard donor for III–V systems is Si, which unfortunately acts

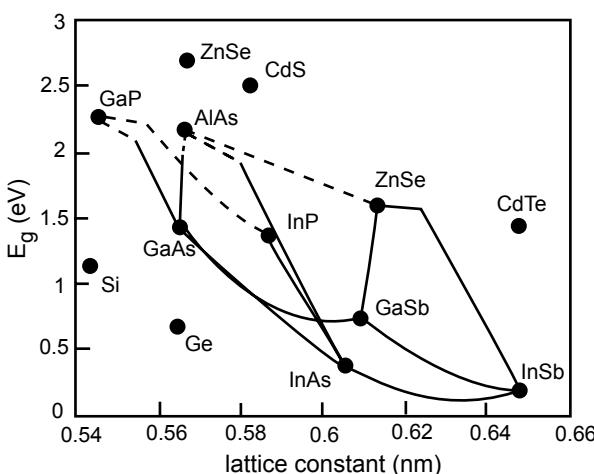


Fig. 3.27 The bandgap engineer's map shows what materials can be combined such that the lattice mismatch remains tolerable. Apparently, GaAs and AlAs match very well, while the combination of Si with Ge will be accompanied by strain effects.

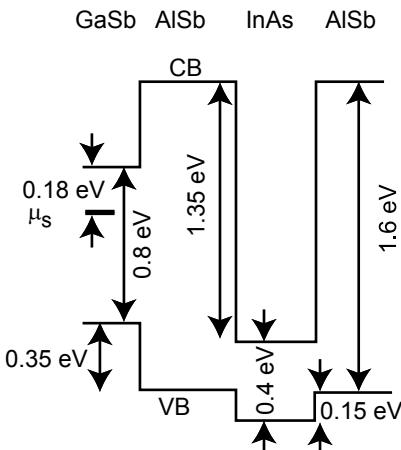


Fig. 3.28 Schematic sketch of the band structure in an InAs–AlSb quantum well. The surface is capped by a GaSb layer with a surface chemical potential 180 meV below the bottom of the conduction band.

as an acceptor in GaSb and AlSb. Te is known to generate n-doping in AlSb, but this element causes technological problems during sample growth.

Although there are ways to circumvent this problem (see e.g. [28]), doping is not necessary to fill the quantum well with electrons. The reason is that, in this system, the surface chemical potential in the GaSb layer is just 180 meV below

the GaSb conduction band, well above the conduction band bottom of InAs. Therefore, electrons get transferred from the surface band into the quantum well. Rather large electron densities up to $n \approx 1.2 \times 10^{16} \text{ m}^{-2}$ can be achieved in such undoped quantum wells, by keeping the distance of the well from the surface sufficiently small, e.g. of the order of 30 nm. In intentionally doped systems, electron densities up to $n \approx 5.6 \times 10^{16} \text{ m}^{-2}$ have been reported [28]. At room temperature, mobilities can be of the order of $3 \text{ m}^2/\text{Vs}$ [43], which increase to $\mu \geq 30 \text{ m}^2/\text{Vs}$ at liquid helium temperatures. This material is also interesting because of its large effective electronic g-factor of $g^*(\text{InAs}) = -14$, which is about 32 times larger than in GaAs ($g^*(\text{GaAs}) = -0.44$). The Zeeman splitting of the energy levels in external magnetic fields is thus very strong.

3.4.3.2 Hole gas in Si–Si_{1-x}Ge_x–Si quantum wells

As an example of a hole gas, we consider the Si–Si_{0.85}Ge_{0.15}–Si quantum well depicted in Fig. 3.29. The interface is type II staggered. The band offset depends on x and occurs to a large fraction in the valence band. In our example, a boron-doped layer was grown in between the quantum well and the surface. Similar to the Ga[Al]As heterostructure discussed previously, the holes partly fill the surface band, and are partly transferred into the quantum well. Although hole mobilities in this system are rather low, namely of the order of $\mu \approx 1 \text{ m}^2/\text{Vs}$ even at reduced temperatures, this system represents a way to generate modulation doping in a Si-based material, which is of technological importance. From a fundamental point of view, this material is interesting, for example, because hole gases with greatly reduced screening properties can be generated. Hence, it is a system suited to study effects based on strong electron–electron interactions.⁸

3.4.3.3 Organic FETs

As a final example, we mention field effect transistors that use an organic semiconductor, such as pentacene or polythiophene-based materials (see Fig. 3.30), as host for the electron gas. These materials have several advantages over their inorganic colleagues. They are soluble in organic solvents and can thus be spin-coated over large areas, which makes the deposition inexpensive. Organic light-emitting diodes are already commercially available, so that more complex, all-organic optoelectronic circuits may become possible soon. Also, the properties of organic semiconductors are tunable to some extent via their chemical synthesis. For example, alkyl side chains of various lengths can be chemically attached to a polythiophene backbone, which re-

8) Two-dimensional electron gases can be generated in this material system as well. Here, the electrons collect in a Si quantum well that forms between two layers of Si_{1-x}Ge_x. The mobility of such electron gases can reach $\mu \approx 50 \text{ m}^2/\text{Vs}$ [163].

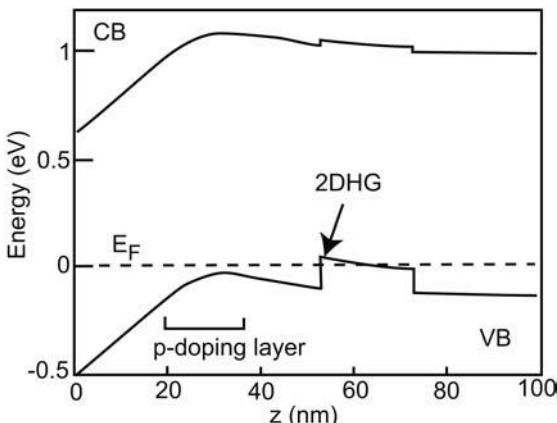


Fig. 3.29 Band alignment of a modulation-doped Si–Si_{0.85}Ge_{0.15}–Si quantum well. A two-dimensional hole gas (2DHG) is formed in the Si_{0.85}Ge_{0.15} valence band. The horizontal “square bracket” denotes the Si layer that was doped with B, which acts as acceptor. The doping density was $N_A = 6.5 \times 10^{25} \text{ m}^{-3}$ over a height of 15 nm in the z -direction. Adapted from [271].

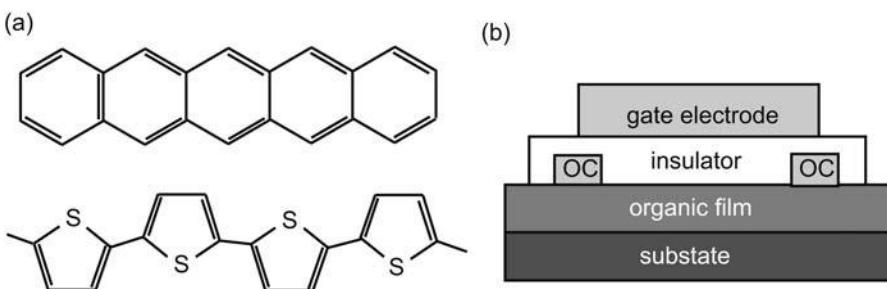


Fig. 3.30 (a) Organic oligomers and polymers, like pentacene (top) or polythiophene (bottom), are examples of plastic FET materials. (b) A typical schematic layout of such a transistor.

sults in the family of poly(3-alkyl)thiophenes. By varying the length of the side chains, the lattice constant, the bandgap, the resistivity, and many other properties can be tuned. Moreover, the material is mechanically soft and can be operated on flexible substrates.

The common characteristic feature of such molecules is a π -conjugated electron system, in which single and double π bonds alternate along the chain. Novel types of applications can be thought of with such plastic transistors, such as disposable electronics, as well as large-area devices. So far, however, their electronic performance does not match those of established transistors. Room-temperature mobilities of $\mu = 7 \times 10^{-6} \text{ m}^2/\text{Vs}$ have been reported [153], for example. A major present problem is the degradation of these

materials under ambient conditions. Also, the understanding of the charge transport in organic semiconductors is far from complete. Owing to the combination of small dielectric constants, and therefore large screening lengths, with the low electron densities and the inherently existing strong disorder, the modeling of the transport in these systems poses a theoretical challenge.

3.4.4

Quantum confined carriers in comparison to bulk carriers

It has already been mentioned above that the effective mass of electrons in a Si MOSFET differs from their effective mass in bulk silicon. This raises the question as to what extent the bulk properties are relevant for mobile carriers in heterostructures.

First of all, quantum confinement changes the band structure and the effective masses. Consider, for example, a 2DEG in a quantum well of finite height. Clearly, the tails of the wave function extend into the barrier material, where the electrons have a different effective mass. Nevertheless, the conditions of the wave function and its derivative being continuous at the interface remain valid. This implies that the energy dispersion of the electrons is changed; see [83] for a discussion of such boundary conditions. Holes experience more dramatic modifications. Since the quantized energies of a quantum well depend on the effective mass, it is intuitively clear that the degeneracy of heavy holes and light holes at the Γ -point is removed by the quantum confinement. It can be shown that in fact the effective masses are reversed close to $\vec{k} = 0$, i.e. the light holes in the bulk material become the heavy holes in the quantum well. Furthermore, the confinement causes a mixing of the two bands, which leads to further strong modifications of the hole energy dispersion. As a result, it is often quite misleading to speak of heavy and light holes in quantum confined structures. For a quantitative discussion of these issues, the reader is referred to [21].

All these descriptions implicitly assume that the envelope function approximation and the concept of effective masses remain valid in heterostructures. This requires that the superposed potential varies slowly on the scale of the lattice constant, which is clearly not the case at a heterointerface or a narrow quantum well. It has been shown, though, that the envelope wave equation can also be derived for abrupt interfaces, as long as the envelope *function* still varies slowly. This issue is addressed in Paper P3.2.

Second, the screening properties are modified in two dimensions, which is intuitively easy to see, since the scattering potential is still three-dimensional, but can be screened only in two dimensions by the electrons, while only polarization charges can screen in the third direction. It can be shown that, for a strictly two-dimensional carrier system, the static dielectric constant in the

limit of low temperatures (see Eq. (2.66) for the three-dimensional case) is given by

$$\epsilon(\vec{q}) = \begin{cases} 1 + (k_{\text{TF}}/q) & q \leq 2k_{\text{F}} \\ 1 + (k_{\text{TF}}/q) \sqrt{1 - (2k_{\text{F}}/q)^2} & q > 2k_{\text{F}} \end{cases} \quad (3.22)$$

The resulting charge density induced by a Coulomb potential reads, at large distances from the scattering center,

$$V_{\text{eff}}(\vec{r}) = \frac{Ze}{\epsilon\epsilon_0} \frac{4k_{\text{TF}}k_{\text{F}}^2}{(2k_{\text{F}} + k_{\text{TF}})^2} \frac{\sin(2k_{\text{F}}r)}{(2k_{\text{F}}r)^2} \quad (3.23)$$

Thus, the screened potential drops with r^{-2} as compared to r^{-3} in three dimensions.

Furthermore, additional scattering mechanisms, which are absent in bulk materials, are possible in quantum confined systems. The scattering of electrons on ionized impurities has a somewhat different character in modulation-doped systems as compared to bulk materials, since they are spatially separated from the electrons by a spacer layer. The residual and usually small density of ionized impurities inside the electron gas is comparatively small in high-quality systems. One may be tempted to guess that the broader the spacer layer, the higher the mobility. This is not the case, though, since as the spacer thickness becomes larger, the carrier density gets smaller, and screening becomes less effective. Hence, a maximum in the mobility as a function of the spacer thickness is observed. In Ga[Al]As HEMTs, the optimum spacer thickness depends on the cleanliness of the material and the doping density. It varies between ≈ 20 nm and ≈ 60 nm. Another scattering mechanism in FET structures is interface roughness scattering. The interface clearly constitutes a deviation from perfect periodicity and consequently generates scattering. In the case of a Ga[Al]As HEMT, this is of minor importance. In narrow quantum wells, however, where fluctuations at both interfaces are important, this mechanism may become important. In Si MOSFETs, on the other hand, the oxide is amorphous, and interface roughness scattering is not negligible.

Alloy scattering occurs in compound materials such as $\text{Al}_x\text{Ga}_{1-x}\text{As}$. The replacement of Ga atoms by Al atoms takes place at random positions, and a non-periodic potential results. This kind of scattering usually plays no significant role, as long as the carriers reside in a crystalline material, such as GaAs, with a barrier made of a ternary compound, since only the evanescent tails of the wave function feel this kind of disorder.

In Fig. 3.31 a model calculation adopted to some typical data is shown, which surveys the relevance of various scattering mechanisms in a Ga[Al]As HEMT. While alloy scattering and interface roughness scattering are irrelevant except at very low temperatures and in extremely clean samples, the ionized impurities are split into two components, namely a density of homogeneously

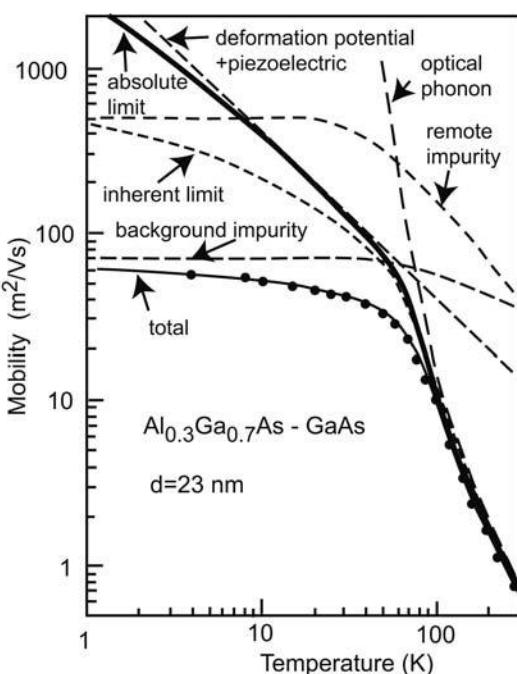


Fig. 3.31 Significance of various scattering processes in a Ga[Al]As HEMT. Black dots denote experimental results for a typical structure, with an electron density of $n = 2.2 \times 10^{15} \text{ m}^{-2}$, and a spacer thickness of $d = 23 \text{ nm}$. The density of the modulation doping was $8.6 \times 10^{22} \text{ m}^{-3}$. This doping,

which causes the remote impurities, was present within a 20 nm layer between the surface and the spacer. In addition, a homogeneous density of background impurities of $9 \times 10^{19} \text{ m}^{-3}$ was assumed, which is a typical number for high-quality GaAs. After [314].

distributed background impurity, which can be reduced in principle by fabricating cleaner samples, and a density of remote impurities, which is necessary, since they are the donors that provide the electrons. The *inherent limit* in this figure indicates the mobility that would be obtained in a sample in the absence of background impurities, but with the remote donors still in the sample. The *absolute limit* represents a situation where the remote donors do not influence the mobility, which is then given by processes intrinsic to a perfect $\text{GaAs}-\text{Al}_x\text{Ga}_{1-x}\text{As}$ interface. It is worth comparing this behavior with the temperature dependence of bulk GaAs (Fig. 2.15) and Fig. 3.25. Most strikingly, the reduction of the mobility as the temperature is lowered in bulk GaAs is absent. This is the effect of the modulation doping, which separates the conduction electrons from the ionized donors, and thus breaks the $\mu \propto \Theta^{3/2}$ law. Second, it becomes apparent that the best samples shown in Fig. 3.25 hit the absolute mobility limit at intermediate and high temperatures ($4 \text{ K} \leq \Theta \leq 400 \text{ K}$) set by deformation potential scattering, piezoelectric scattering, and optical phonon

scattering. The saturation at very low temperatures is due the residual impurities, but it occurs not very much below the absolute limit. Thus we cannot expect to see another huge increase in electron mobilities in this material.

Finally, inter-subband scattering should be mentioned (see Fig. 3.32), which denotes scattering events for which the incoming carrier is scattered into another subband of the confined potential. For elastic scattering events, this is only possible if at least two subbands are already occupied, and the system is thus not strictly two-dimensional.

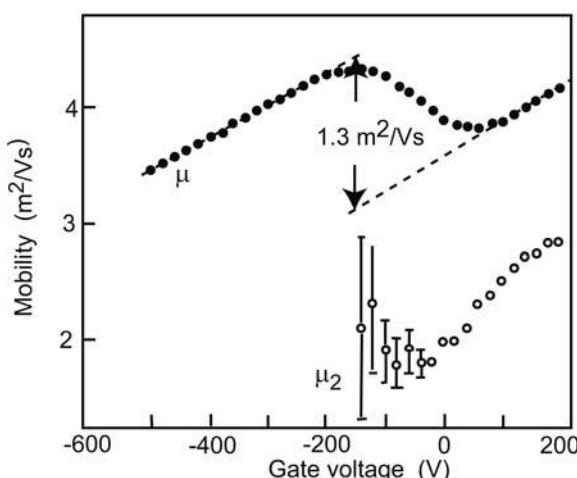


Fig. 3.32 Electron mobilities in a GaAs HEMT as a function of the gate voltage. Above a threshold electron density, the second two-dimensional subband gets occupied, and the mobility drops due to additional inter-subband scattering. After [291].

Papers and Exercises

- P3.1** In [104], band alignments between semiconductor heterostructures are predicted within the concept of electronegativities. Discuss the basic idea of this approach.
- P3.2** A highly instructive article on the validity of the envelope function approximation in semiconductor heterostructures is [44]. Work out the author's line of arguing.

E3.1 The wave functions of the ground state in a triangular potential can be approximated by the Fang–Howard function

$$\Phi(z) = \sqrt{b^3/2} ze^{-bz/2}$$

Calculate the expectation value of the electron location. Discuss the consequences of this result for a Si MOSFET, for example.

E3.2 Construct a NOR gate in CMOS technology with four MOSFETs.

Further Reading

Extensive and comprehensive reviews on the physics of surface states are [69] and [213]. An introduction to the theory of crystal surfaces can be found in [68]. For the quantum mechanical properties of layered devices, [21] is a valuable source of information.

This Page Intentionally Left Blank

4 Experimental Techniques

This chapter introduces the experimental tools and techniques involved in preparing nanostructures and measuring their transport properties. The majority of mesoscopic devices are made of semiconductor heterostructures. Section 4.1 describes how they are fabricated. This process usually includes single-crystal growth, followed by lateral patterning of the crystal slices. The lateral patterning is done by various lithographic techniques, while self-assembly is used occasionally. Already in the Introduction, it has become clear that many of the experiments are performed at liquid helium temperatures, which is the regime below $\Theta = 4.2\text{ K}$, the boiling temperature of ${}^4\text{He}$ at 1 bar. Therefore, the concepts and techniques of generating such a low-temperature environment are discussed in Section 4.2. This includes the relevant properties of liquid helium as well as the essentials of helium cryostats. Finally, some basic understanding of electronics is very helpful for the discussion of the transport experiments. This is the topic of Section 4.3.

The present chapter cannot replace a thorough treatment of these issues, which would require a bulky textbook for each section. Rather, our goal is to provide the knowledge needed to appreciate the constraints of the experiments set by the technology.

4.1 Sample preparation

Fabricating nanostructures for mesoscopic transport experiments is a major technological challenge. The requirements concerning material purity, lithographic resolution, and process control are at the edge of present-day technology. We will exemplify the technology using Si and GaAs as examples, and occasionally mention some special properties of other materials. As a rule, we refrain from specifying process parameters and experimental recipes. For details, the reader is referred to the specialized literature at the end of this chapter.

Silicon dominates in industry, while GaAs and other materials are essentially only used where silicon devices are significantly less useful, like in op-

toelectronics or in ultra-high-speed, ultra-low-noise applications. From a technological point of view, Si has several major advantages. First of all, the raw material is quartz sand, easily available and cheap. Second, its high mechanical stability simplifies all process steps. A very important point is the fact that Si can be easily oxidized into SiO_2 , which has excellent mechanical and electronic properties, like high breakdown electric fields and large resistivities. GaAs oxides, on the other hand, have poor electronic properties, and are more or less useless for electronic applications, like in capacitors.

During processing, we have to prevent dust particles hitting this crucial area, since they may cause defects in the lithographic patterns, which get transferred into the nanostructure in subsequent process steps. Therefore, processing is usually done in a *clean room* in which the air is heavily filtered. Clean rooms are classified by the number N of dust particles larger than 500 nm per cubic foot. For example, a class 100 clean room contains $N = 100$ such particles. It is true that feature sizes have been scaled down to well below 500 nm, and, along with this, the size of the relevant dust particles has decreased. Nevertheless, the clean room class as defined above is a useful quantity, as the particle size distributions in air are well known.

Mesoscopic researchers typically prepare a small number of individual samples, with an active area below a square millimeter. Of course, it is much harder to keep a 4-inch wafer free of dust particles, and the demands posed on an industrial clean room are much higher. Furthermore, although a high sample yield is desirable in research as well, a research lab has no problems living with yields below 99%.

Fig. 4.1 shows a processed Ga[Al]As HEMT structure. It becomes immediately apparent that a whole bunch of fabrication steps is required. To begin with, the semiconductor heterostructure must be grown. This involves single-crystal growth, plus some sort of epitaxy to add the layers of different materials. Next, a piece of the wafer has to be processed by lateral lithography. The two-dimensional electron gas is contained in the *mesa*, which should have a suitable geometry. Ohmic contacts and gate electrodes have to be patterned. Finally, wires must be connected to the electrodes.

We will discuss these fabrication steps below. Although the details of the processes depend on the material used, the technological concepts are almost universal.

4.1.1

Single crystal growth

A standard method to grow silicon single crystals is the so-called *Czochralski method* (Fig. 4.2(a)). A small single crystal (the *seed*, which also determines the crystal direction) is immersed in a purified silicon melt (Si has a melting point of 1412°C at atmospheric pressure). The atmosphere should be inert; an argon

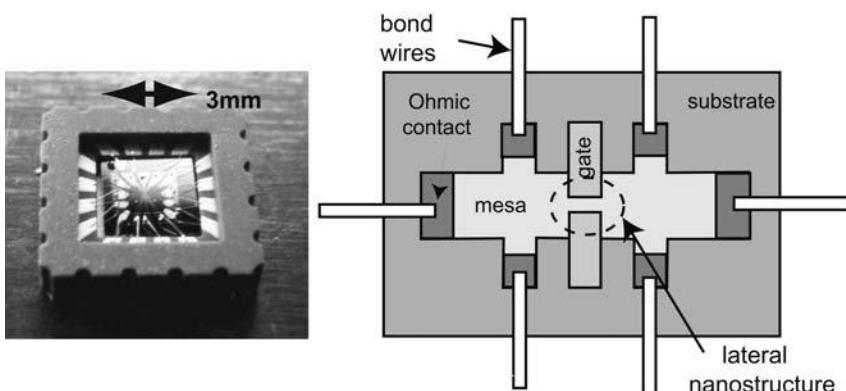


Fig. 4.1 Left: Photograph of an assembled Ga[Al]As HEMT structure. The chip resides inside a ceramic chip carrier. Electrical contact between the carrier and the chip is made by bonded wires. The chip carrier has a size of $5\text{ mm} \times 5\text{ mm}$. Right: Schematic components of a typical microchip designed for mesoscopic transport measurements.

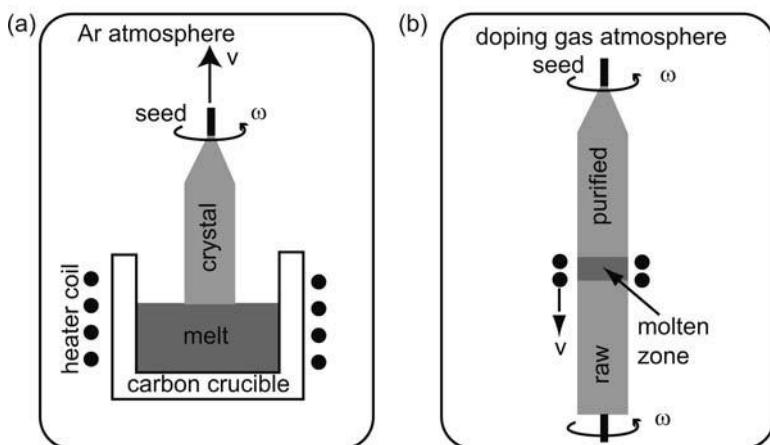


Fig. 4.2 (a) The Czochralski scheme for crystal growth used to grow Si single crystals. The angular velocity and the velocity are ω and v , respectively. (b) The zone pulling technique.

atmosphere is often used. The seed is rotated ($\omega \approx 2\pi\text{s}^{-1}$) and slowly (with a speed of, say, 1 mm/s) pulled out of the melt. The pulling speed determines the diameter of the crystal cylinder, which may be as large as 10 inches (25 cm). A typical length is 1 m.

Another widely used technique is zone pulling (Fig. 4.2(b)). Here, the raw crystal, which may be a Czochralski grown crystal, is molten locally via eddy current heating with an RF coil. The temperature is about 1450°C . The setup resides inside a high vacuum chamber. For impurity atoms, it is energetically

favorable to be in the melt. They collect in the molten zone, evaporate there and can be pumped away. With this technique, impurity concentrations below 10^{13} cm^{-3} can be obtained. Undoped Si crystals can have resistivities above $10 \Omega \text{ m}$, which reflect their high purity. If the crystal has to be doped, a doping gas atmosphere is established in the growth chamber, e.g. a B_2H_6 atmosphere for boron doping, or a PH_3 atmosphere for phosphorus doping.

GaAs is a binary material and as such more difficult to grow. A general problem with multi-component melts is the different vapor pressures of the components. In order to avoid compositional changes of the melt over time, the vapor pressures must be controlled. In the case of GaAs, As has an overpressure of 0.9 bar; while in InP, the P overpressure is 60 bar. Two methods are common for compensating the overpressures. In the *liquid encapsulated Czochralski* (LEC) technique, the melt is covered with a fluid that does not intermix (Fig. 4.3(a)). As a result, no gas can escape from the melt. In the Bridgeman technique, the melt resides in a closed quartz tube, in which the correct As overpressure is established by heating solid As outside the melt to the corresponding temperature (Fig. 4.3(b)). Here, the crystal is grown by moving the melt along a suitable temperature gradient and at an appropriate speed.

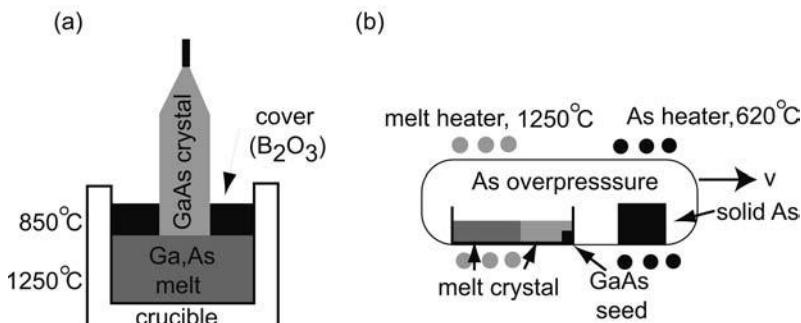


Fig. 4.3 Schemes for growing of two-component crystals, such as GaAs. (a) In the LEC technique, the melt is covered by an impenetrable liquid. The crystal is pulled through this cover layer. (b) In the Bridgeman technique, a closed tube contains a crucible that hosts the melt, a seed, and the freshly grown crystal, as well as a piece of solid As.

The temperature of the solid As is kept at a temperature that corresponds to an As overpressure of 0.9 bar. The melt solidifies at the location where the spatially varying temperature reaches 1238°C. Crystal growth is established by moving the tube along the temperature profile.

4.1.2

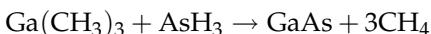
Growth of layered structures

Pulling a crystal out of a melt is perfect for fabricating substrates, which are usually obtained by cutting the crystal cylinder into thin disks called *wafers*. A typical substrate thickness is 300 μm , while surfaces roughnesses

of a few nanometers can be achieved by mechanical polishing. These techniques are not suited to growing layered structures, such as Ga[Al]As heterostructures. Here, we need something that provides ultra-clean growth of individual monolayers, and the material composition must be controllable. Several techniques are established for the growth of layered semiconductor structures, and we will briefly discuss two of them, metal organic chemical vapor deposition (MOCVD) and molecular beam epitaxy (MBE).

4.1.2.1 Metal organic chemical vapor deposition (MOCVD)

In this technique, the substrate is mounted in a vacuum chamber (Fig. 4.4(a)). The atoms of the semiconductor components to be grown are introduced via suitable molecular gas flows. The gas molecules crack at the surface and deposit the semiconductor atom on the substrate. In the case of GaAs, a possible chemical reaction with $\text{Ga}(\text{CH}_3)_3$ and AsH_3 as input gases is



taking place around a temperature of 1120°C . The advantage of MOCVD is its relatively low cost. A disadvantage is the high toxicity of the gases involved. In addition, the material grown is not as clean as that obtained with the second technique we shall now look at.

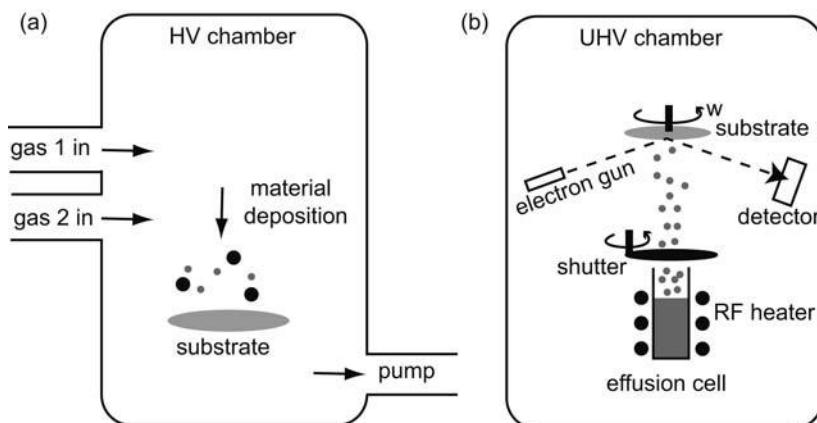


Fig. 4.4 Schematic view into (a) an HV chamber for MOCVD, and (b) a UHV chamber for MBE.

4.1.2.2 Molecular beam epitaxy (MBE)

Here, atomic layers are grown in an ultra-high-vacuum chamber, with pressures of the order of 10^{-11} mbar. A substrate is inserted in the UHV chamber, heated, and slowly rotated (Fig. 4.4(b)). The components of the semiconduc-

tor are supplied by effusion cells, which can be individually heated to provide the flux needed, as well as opened and closed. For growing standard GaAs HEMTs, for example, Ga, As, Al, and Si (for n-doping) effusion cells are needed. In this way, the crystal can be grown monolayer by monolayer, and can be selectively doped: the modulation doping encountered in Chapter 3 can be easily implemented. Typical growth rates are 0.1 nm/s. The growth can be calibrated and monitored with *reflection high-energy electron diffraction* (RHEED). Here, an electron beam with an energy of about 10 keV hits the surface under a very small angle (a degree or so), and gets reflected at the surface. Its penetration depth is a few monolayers only, such that the reflected interference pattern is highly sensitive to the roughness and the crystal structure of the surface. The reflected intensity shows a minimum if the coverage of the monolayer is 50%, which corresponds to maximum roughness. When a monolayer has just been completed, the scattering has the highest specularity, and the reflected intensity shows a maximum.

Although MBE is very expensive and time-consuming, it is widely used to grow heterostructures for mesoscopic transport experiments, since the quality of the samples is unsurpassed by any other method. The high pressure is needed to make the residual gas monolayer formation time sufficiently large (see Exercise E4.1).

Ga[Al]As heterostructures are frequently grown by MBE. After some surface cleaning, the growth begins with a buffer layer, consisting of a series of GaAs–AlAs superlattices with a short period (*a short-period superlattice*, SPS). The purpose of the SPS is twofold. First of all, the mechanically polished GaAs substrate is not atomically flat. It has been found that an SPS reduces the roughness due to polishing to nearly atomic flatness [237]. Second, the superlattice tends to trap impurities that may diffuse from the substrate into the electronically active layers grown on top.

MBE can also be used to prepare more complicated structures than sequences of layers with translation invariance in two dimensions. These technologies are in the focus of present-day research. As a first example, we consider a technique called *cleaved edge overgrowth* (CEO) (Fig. 4.5). Here, the layer growth is interrupted at the right point, and the wafer is cleaved inside the MBE chamber, such that the grown sequence of layers appears on an atomically flat surface. *Cleaving* includes scratching the wafer at its edge and subsequently breaking it by mechanical pressure. The GaAs wafer breaks at the scratched position along a single crystal plane. Subsequently, MBE growth is continued on top of this freshly cleaved surface.

Extremely small nanostructures of effectively one- and even zero-dimensional character, with atomically flat interfaces, have been produced in this way. For a review, see [319]. We will discuss some properties of such nanostructures in subsequent chapters.

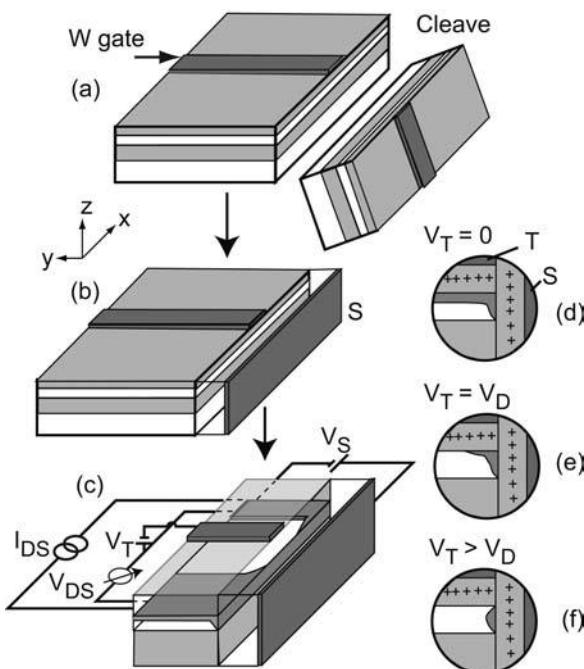


Fig. 4.5 Quantum wire production by CEO. (a) A thin GaAs layer (white), embedded in Al_xGa_{1-x}As layers (gray), is shown. The sample is partly covered with a tungsten gate stripe. This structure has been obtained by MBE growth in the z -direction. The Al_xGa_{1-x}As layer on top of the GaAs well is modulation-doped with Si. After the growth has been completed, the wafer is cleaved perpendicular to the stripe. (b) MBE growth is continued on the freshly cleaved surface, i.e. in the y -direction. A modulation-doped

Al_xGa_{1-x}As layer is covered by another tungsten layer. (c) The electrical connections to the different elements. Voltages can be applied to the two gates T and S, and a current is applied between the two areas separated by gate T. (d)–(f) Sketches of how different gate voltage combinations shape the electron gas. In particular, a one-dimensional wire is formed in (f), which extends along the quantum well at the cleaved and regrown interface. After [337].

Self-assembled quantum dots are a second example of ongoing MBE research activities. Growing a semiconductor (B) on top of an appropriate substrate (A) does not necessarily lead to atomically flat films that build up monolayer by monolayer. Rather, one distinguishes three growth modes: the one we have just encountered is known as *Frank-van der Merwe* growth [102]; alternatively, growth of new material can take place in terms of isolated islands, called the *Volmer-Weber* mode [311]; or in terms of islands connected via a thin layer of the same material, the so-called wetting layer, called the *Stranski-Krastanov* mode [293] (see Fig. 4.6).

Which kind of growth takes place depends on an interplay of different energy scales. In a strongly simplified view, we can assume that there are surface energies per unit area E_A and E_B related to the surfaces of material A and B,

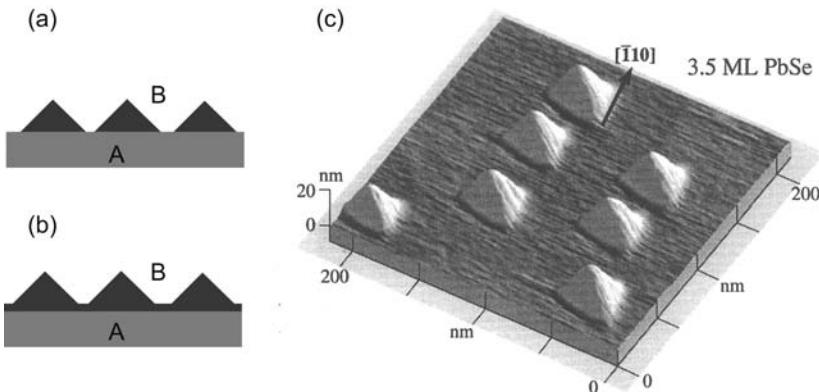


Fig. 4.6 Growth of material B on top of material A in (a) the Volmer-Weber and (b) the Stranski-Krastanov modes. (c) Atomic force microscope picture of PbSe islands on a PbTe(111) substrate. Note the homogeneous size distribution and the orientation of the pyramids. Part (c) has been reproduced from [242].

respectively. In addition, there is an interface energy per unit area E_{AB} . Let S be the fraction of the surface of A covered by B. The total energy is then given by

$$E = (1 - S)E_A + SE_B + SE_{AB}$$

Here, we have assumed that the surface of B remains flat, and that none of the materials get strained, i.e. that A and B have identical lattice constants. This energy will get minimized. It follows that for $E_{AB} + E_B < E_A$, S will get maximized, and Frank-van der Merve growth takes place. On the other hand, for $E_{AB} + E_B > E_A$, a minimized S minimizes the energy, and we have Volmer-Weber growth. In order to establish Stranski-Krastanov growth, we need a lattice mismatch between A and B, which gives an elastic strain energy per unit area in B in addition, given by $E_{str}d$, where d denotes the thickness of the layer. Minimizing the total energy again predicts homogeneous film formation for

$$d < \frac{E_A - E_B}{E_{str}}$$

provided we can neglect the interface energy and material A suffers no strain. This simple picture shows that after the wetting layer of thickness d (typically two to four monolayers) is completed, it is energetically more favorable to continue growth with island formation.

Experimental studies as well as theoretical considerations have demonstrated that strain does not cause, as one might assume, dislocations inside the dots; rather, the substrate gets elastically strained as well [81]. Such islands, e.g. InAs grown on GaAs, are typically of pyramidal shape and have

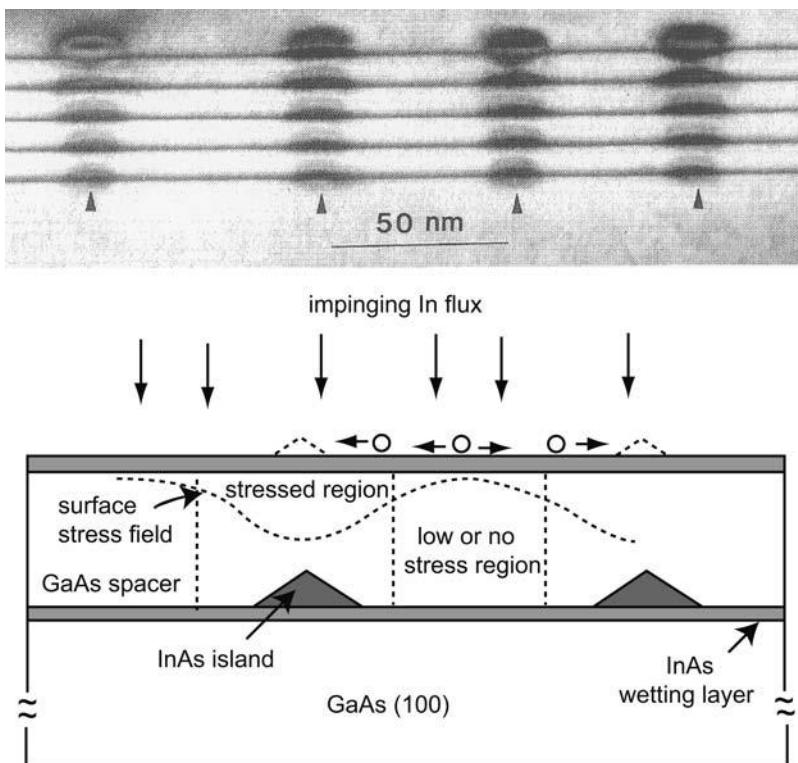


Fig. 4.7 Top: Transmission electron microscope picture of InAs layers, separated by 36 monolayers of GaAs. The vertical direction is the growth direction. The InAs SAQDs, seen as dark spots, align on top of each other. The bottom figure shows a schematic sketch of the SAQDs and the wetting layer. Adapted from [334].

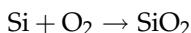
very homogeneous size distributions with variances around 10% only [190]. Since their sizes are in the range of 20 nm in width and a few nanometers in height,¹ strong quantization effects can be expected inside, and have in fact been observed in many experiments, some of which we will discuss later on. Therefore, they are referred to as *self-assembled quantum dots* (SAQDs). In some systems, the SAQDs even align with each other and form lattices of various dimensions. A three-dimensional SAQD superlattice is the topic of Paper P4.2. Here, we look at growth of linear chains of InAs SAQD islands, embedded in GaAs (Fig. 4.7).

In this example, the strain in the GaAs induced by the buried InAs islands modulates the GaAs surface energy, and thus the freshly offered InAs will preferentially form dots at locations where the lattice mismatch is minimum,

1) This is below the resolution limit of lithographic techniques, as we will see shortly.

which is above the locations of the SAQDs next to the surface. Clearly, the spacing between adjacent InAs layers can be optimized for maximum probability of SAQD alignment. For large distances, the strain modulation at the surface becomes too weak; while for very small spacings, neighboring points of extremal strain begin to overlap.

A very different process for preparing a layered structure is thermal oxidation of Si. For growing oxides used in electronic applications, the technique of choice is usually *dry oxidation*. The Si wafer is placed in a furnace at a temperature of about 1000°C and exposed to oxygen. The wafer oxidizes via the reaction



The oxygen diffuses through the already grown oxide layer and reacts with the Si at the Si-SiO₂ interface. The oxide growth rate therefore drops as its thickness increases. Furthermore, the oxide penetrates into the Si; about 50% of the oxide layer is located below the original wafer surface. Breakdown electric fields for oxides grown with this technique can be of the order of 5×10^8 V/m, and are thus well suited for electronic applications.

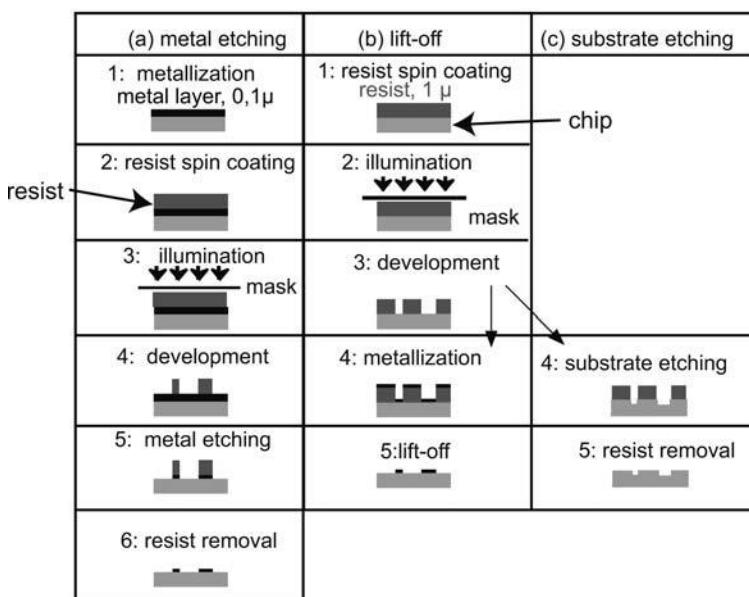


Fig. 4.8 Comparison of different lateral patterning schemes for semiconductors.

4.1.3

Lateral patterning

The special MBE techniques CEO and SAQD growth are by no means standard technology. It is more usual to grow a heterostructure with two-dimensional translation invariance parallel to the surface, and pattern the nanostructure subsequently by some sort of lateral processing. Examples for typical sequences of process steps are given in Fig. 4.8. Column (a) is typical for many Si fabrication steps. The substrate is covered with a homogeneous metal layer, which is subsequently coated with a suitable resist. Illumination and development of the resist through a mask exposes some areas of the metal layer, while others are protected by the resist. The illumination is usually carried out with ultraviolet (UV) light or with electrons. An etch step follows, which selectively removes the free metal surfaces. Here, the resist acts as an etch mask. Finally, the resist gets removed, and a patterned metal layer on the substrate results. This technique is rarely ever used in GaAs processing, since essentially all suitable metal etchants attack GaAs as well. Therefore, fabrication scheme (b) is typically used. Here, the substrate is first covered by resist, which gets illuminated and developed. Now, the metal is evaporated on the substrate, with the patterned resist acting as evaporation mask. The *lift-off* step follows, i.e. the resist is removed with the metal film on top. The final result is identical to that of scheme (a). For selective etching of the substrate (c), steps 1 to 3 are identical to (b). Then, the patterned resist is used as an etch mask for the substrate. We now discuss these process steps in further detail.

4.1.3.1 Defining patterns in resists

The two standard techniques for imposing a pattern into a resist are optical lithography and electron beam lithography.

Optical lithography By this we mean illumination of a photoresist by visible or ultraviolet light. The sample is coated with a thin and homogeneous photosensitive resist. This is done by casting some resist solution onto the sample, which is then rotated for about one minute at high speed, typically a few thousand rpm. The spinning speed and the viscosity of the solution determine the thickness of the resist layer, which is of the order of 1 μm .

After baking the resist, the sample is mounted into a mask aligner, a device designed for adjusting the sample with respect to a mask that contains the structure to be illuminated. The mask aligner is equipped with a light source of high power that illuminates the resist film through the mask (see Fig. 4.9). The pattern sizes are Doppler-limited, which means that the smallest feature sizes are about half the wavelength ($\approx 150 \text{ nm}$) divided by the index of refraction of the resist (≈ 1.5), which limits the resolution to roughly 100 nm. The

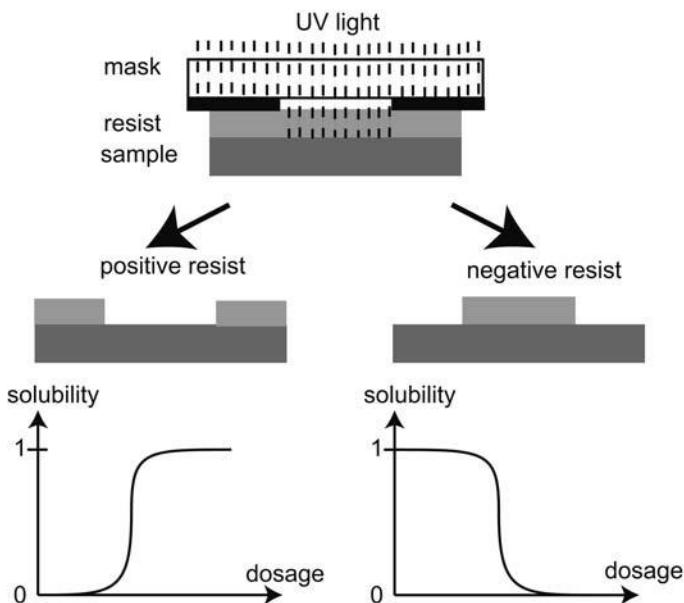


Fig. 4.9 Top: Contact illumination. The mask pattern is transferred to the resist via illumination and subsequent development. Center: Resulting resist cross section for positive (left) and negative (right) resist. Bottom: Solubility characteristics for the two resist types.

mask can be a quartz plate coated with a thin chromium film, which contains the pattern to be illuminated. In the contact illumination scheme, the Cr film is in mechanical contact with the resist and blocks the light, such that the resist underneath the Cr remains unexposed.

During contact illumination, the mask suffers contamination due to dust particles on top of the resist, as well as by resist adhesion. This can be avoided by projection illumination, where the mask pattern is transferred to the resist via lenses. This technique is widely used in industry, but somewhat unusual in research labs.

Photoresists can be classified as *positive* and *negative*. The solubility of the exposed areas increases for a positive resist, while it decreases for a negative resist (see Fig. 4.9). Immersing the sample into a suitable developer removes the corresponding sections of the resist film. Both types of resists have in common that their solubility as a function of the illumination dosage is a step-like function. This ensures high resolution and sharp edge profiles. It may seem irrelevant at first what kind of resist is used in a particular process. There may, however, be some process-specific requirements that favor one type or the other. Most importantly, a negative resist predominantly produces an *undercut profile*, which means that after development the resist area in contact with the sample is smaller than the area at the resist surface (Fig. 4.10). This is

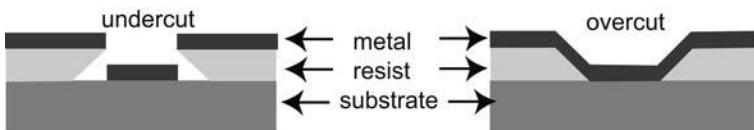


Fig. 4.10 Undercut (left) and overcut (right) resist profiles after illumination, development, and surface metallization.

a consequence of the approximately exponentially decreasing intensity of the illuminating light as it penetrates into the resist. An undercut profile is highly desirable for subsequent metallization steps, in which the resist itself serves as mask. After the metallization, the resist including the metal film on top usually has to be removed in a *lift-off* step, which is bound to fail for resists with an *overcut profile*, since the metal on the sample and that on top of the resist are connected. An undercut profile avoids this problem, provided the thicknesses of the metal layer and resist are properly selected.²

In principle, the resolution can be increased by using shorter wavelengths. In X-ray lithography, resists are illuminated with wavelengths in the 10 nm regime. While significant progress has been achieved over the past decade, severe technological obstacles have to be overcome before this version of optical lithography can be widely used. Photoelectrons limit the resolution to several tens of nanometers, and optical components as well as masks are difficult to pattern, since metals get transparent in the ultraviolet. The ultimate limit of such lithographic techniques is set by the resolution of the resists, which contain organic polymers. The crosslinking of the polymers is enhanced or reduced by the light, which modifies their solubility accordingly. Thus, the resolution cannot become better than the size of the corresponding monomers, which is of the order of 0.5 nm.

Electron beam lithography For feature sizes below ≈ 150 nm, electron beam lithography is currently the technique of choice. Instead of light, electrons may be used as well for illuminating resists, which are in this case polymers like PMMA (polymethylmethacrylate) with a well defined molecular weight. In a positive resist, the electron beam breaks the bonds between the monomers, and an increased solubility results. In negative resists, on the other hand, the electron beam generates inter-chain crosslinking, which decreases the solubility. In that respect, electrons have a very similar effect to ultraviolet light on the resist. A typical experimental setup is shown in Fig. 4.11. A focused electron beam is scanned in a predefined pattern across the sample using deflection coils in the electron optics. In contrast to optical lithography, this is a serial and therefore a slow process. However, structure sizes of 50 nm

2) It should be mentioned that techniques exist for achieving undercut profiles with positive resist as well.

and even below can be fabricated. Many research groups use electron beam lithography in the lab for all feature sizes below 2 μm or so, because the technique gives very good and reproducible results. One type of electron beam lithography uses a high-energy beam of electrons (about 30 keV or larger), which produces extremely small spot sizes of about 1 nm only. However, the illumination resolution is worse than this, since the spatial distribution of secondary electrons backscattered from the substrate actually illuminates the resist (Fig. 4.11). Since the intensity of those electrons drops from the substrate toward the surface of the resist, an undercut profile is intrinsic to this process. The undercut is often enhanced by a two-layer electron beam resist with different dosages.

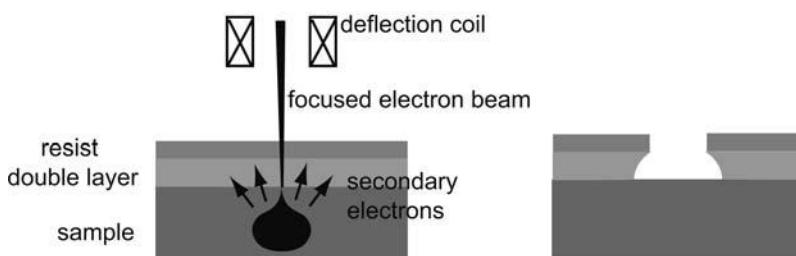


Fig. 4.11 Left: A focused electron beam is scanned across the sample surface with a pattern generator that drives the deflection coils, which are part of the electron optics of the electron microscope. The electrons get scattered both elastically and inelastically in the substrate, and secondary electrons are generated, which have a large cross section for resist illumination. Right: The resulting profile of a two-layer electron-sensitive resist after illumination.

4.1.3.2 Direct writing methods

By definition, such methods do not require resists. Rather, the sample is patterned directly by the exposure. The number of process steps (see Fig. 4.8) is reduced from five or six to just one. We briefly discuss two methods.

Focused ion beam writing The experimental setup resembles the electron writing system, with the electron source replaced by an ion source (e.g. gallium). The ions are implanted in the substrate and localize the electrons in the exposed areas. Highly resistive regions can be defined this way. However, the lateral depletion is rather large, typically above 100 nm. Suitable ion beams can also be used to dope the sample locally.

Scanning probe lithography As an example of current research activities, we briefly discuss lithography techniques based on scanning probe microscopes (SPMs) [36, 37]. Recently, tremendous progress has been made in this respect.

Since SPMs achieve atomic resolution, they are highly promising tools for achieving a further, significant, size reduction. Meanwhile, SPMs have been used in a wide variety of operational modes in order to modify surfaces (for a review see [202]). Moving single atoms with an SPM tip [82] and material deposition from the tip on the substrate [201] have been demonstrated experimentally, for example. Amazing nanodevices can also be fabricated by scratching [162]. Another widely investigated technique is local oxidation of the substrate. In [63], a variety of substrates were oxidized locally by applying a negative voltage to the tip of an SPM with respect to the grounded substrate. Local oxidation with an atomic force microscope has also been used to pattern the electron gas in Ga[Al]As heterostructures directly [149]. Anodic oxidation is a standard process to oxidize surfaces of metals and semiconductors. The setup for local oxidation with an AFM is essentially identical (Fig. 4.12). Here, the water film forming under ambient conditions on top of the substrate provides the electrolyte. A conductive AFM tip acts as cathode, while the chip to be nanostructured is grounded. As a result, the sample surface oxidizes in close vicinity to the AFM tip. The 2DEG is depleted underneath the oxide lines in shallow HEMT structures. The underlying mechanism can be understood in a simple picture: As the cap layer is oxidized, the semiconductor surface gets closer to the 2DEG, while the surface area, and thus the number of surface states, is slightly increased. In samples with the 2DEG so close to the surface, only $\approx 10\%$ of the donor electrons from the doping layer go into the 2DEG, while the remaining 90% fill the surface states. A small reduction of the distance between surface and the 2DEG changes the internal electric fields and can lead to depletion.

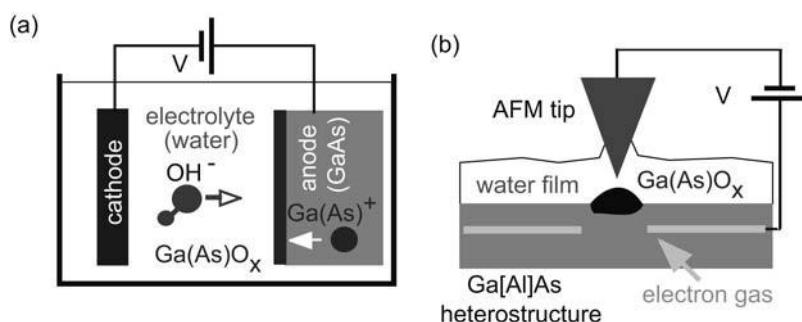


Fig. 4.12 (a) Scheme for conventional anodic oxidation of GaAs.
 (b) Downscaled version of anodic oxidation, with the conductive tip of an atomic force microscope as the cathode, and the water film on top of the sample as electrolyte.

4.1.3.3 Etching

An important technique of transferring the resist pattern into the sample is etching. Patterned resists can be used as etch masks, provided the etchant is sufficiently selective. We distinguish between dry etching and wet chemical etching.

Dry etching The setup for dry etching techniques consists of a vacuum chamber with two electrodes at the top and the bottom. The sample is placed at the bottom, which may be the anode or the cathode, depending on the process. A gas discharge is ignited, and the ions of the etch gas hit the sample (Fig. 4.13). One speaks of plasma etching if the reaction is purely chemical. Oxygen plasma etching is often used to remove resist layers. The low-energy ions avoid damage of the semiconductor and metal components of the sample. A purely physical technique, on the other hand, is ion etching. Here, suitably selected ions are generated and strongly accelerated toward the sample. The physical impact removes sample atoms. Here, resists may serve as masks for a limited time. Radiation damage in the sample, combined with the required high vacuum and the large rate of material deposition at the walls, make this a rather unusual technique. Widely used, however, is *reactive ion etching*. Here, both the physical and the chemical aspects of the ionic bombardment are important. A very convenient side product in this kind of etching is polymer formation at the etched walls, which prevents lateral removal of material. As a consequence, very steep and deep grooves can be etched.

Wet chemical etching Wet chemical etching means immersing the sample in a suitable etchant solution. In contrast to metals, the majority of the common semiconductors are not attacked by pure acids. Therefore, the etch typically

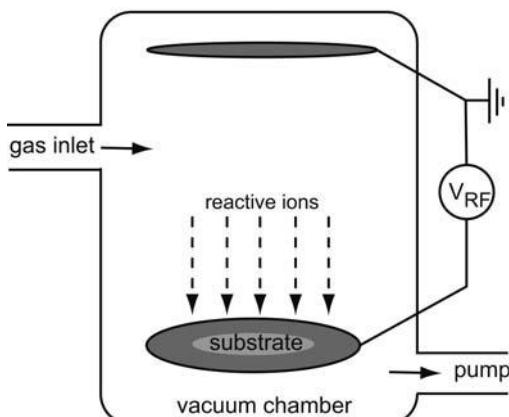


Fig. 4.13 Scheme of a vacuum chamber for reactive ion etching.

consists of a mixture of an oxidizer (such as H_2O_2), an acid (like HCl), and water. H_2O_2 oxidizes the semiconductor, while the acid removes the freshly formed oxide. The oxidation and etch rates depend on the etch composition as well as on the crystal direction. The resulting edge profile can thus be tuned accurately. For many purposes, an overcut edge profile is desirable, since often thin metal layers have to be deposited later on the surface. A metal layer thinner than the etched depth may get disconnected across an etched step with undercut profile.

4.1.4

Metallization

By "metallization", we mean the deposition of metal films on the semiconductor surface. This is usually done by evaporation of the metal in a vacuum chamber. The metals are molten (or sublimed, respectively) in a crucible made of tungsten or carbon, which can be done by heating the crucible with a current, or by focusing an electron beam onto the metal (see Fig. 4.14). At sufficiently high temperature, the metal vapor pressure is so high that a metal film grows at the exposed surfaces with a rate of the order of a nanometer per second. The film thickness is monitored by an oscillating quartz plate. As the metal gets deposited on the quartz, its resonance frequency gets smaller. This effect can be calibrated, and the film thickness can be measured with high accuracy. For lift-off processes, the film thickness should be smaller than the thickness of the resist, for obvious reasons. Typical metallization layers measure thicknesses between 20 nm and a few micrometers.

Of particular importance for the fabrication of nanostructures is the so-called *angle evaporation technique* [74], because feature sizes below the litho-

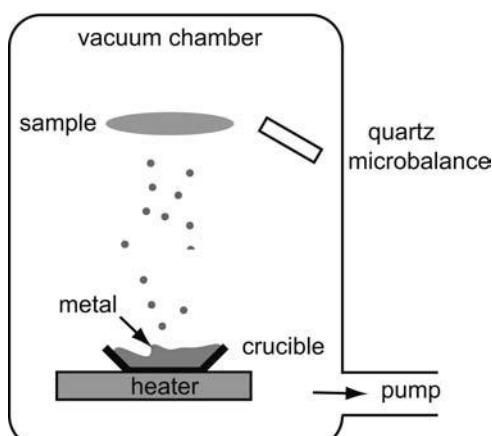


Fig. 4.14 Scheme of an evaporation system for metallizations.

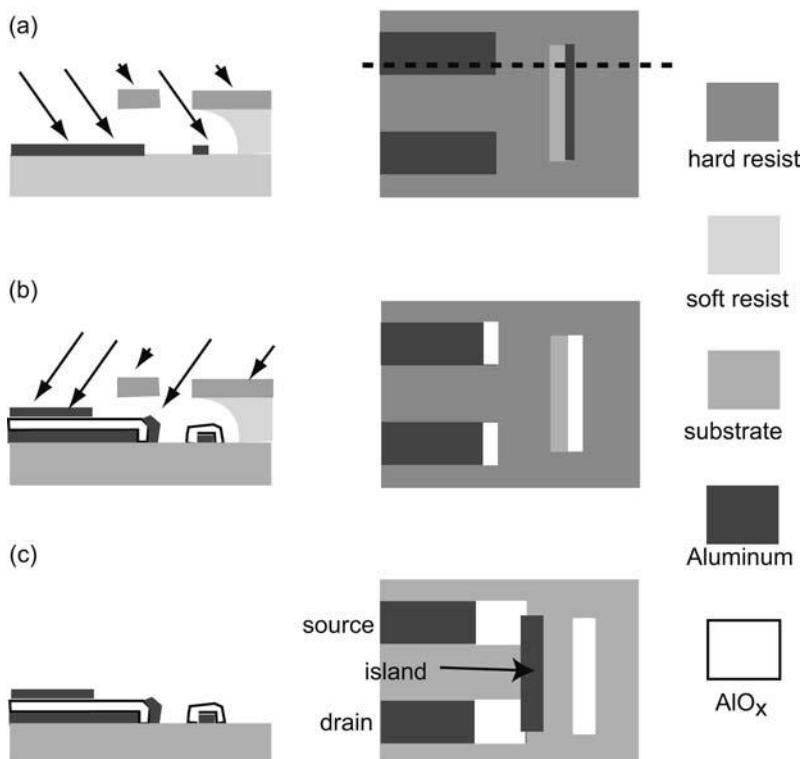


Fig. 4.15 Angle evaporation. The right column shows a top view of a sample section after illumination by electron beam lithography, development, and subsequent evaporation of aluminum under a certain angle. To the left, a cross section of the layers along the dashed line in the right figure is shown. (a) A layer of electron beam resist with low dosage is covered by a resist with a higher dosage. This leads to cage formation as an extreme version of an undercut profile. The upper re-

sist layer is free-standing over a certain area. (b) The Al gets oxidized, and a second Al layer is evaporated on top at a different angle, as indicated by the arrows. A sandwich structure with small overlap areas results, which can be below the pattern sizes in the resist mask. (c) The resulting structure, a small Al island coupled to two leads via small-area tunnel barriers, is shown after the resist layers have been removed. Such islands will be investigated further in Chapter 9.

graphic resolution can be made this way. The trick is to evaporate successive layers of metals from different angles and use the resist as a shadow mask. The technique is illustrated in Fig. 4.15. Overlap areas as small as $30\text{ nm} \times 30\text{ nm}$ can be prepared routinely by angle evaporation.

As pointed out in the previous chapter, metal–semiconductor interfaces form Schottky barriers for the vast majority of material combinations. In order to obtain an ohmic contact, a suitable metal film is evaporated and afterwards alloyed into the semiconductor. “Suitable” means that the Schottky barrier should be small, and the metal should have a low melting point and should act

as a dopant in the semiconductor. For GaAs, the $\text{Au}_{0.88}\text{Ge}_{0.12}$ eutectic alloy³ is a standard ohmic contact material. The Schottky barrier of the $\text{Au}_{0.88}\text{Ge}_{0.12}$ -GaAs system is only 0.3 eV. In addition, eutectic AuGe has a melting point of 360°C, and already at 420°C it begins to alloy into GaAs. Ge atoms diffuse into the Ga and act as donors. This diffusion can be enhanced by adding a small fraction of Ni to the alloy. The resistivity of such a contact is of the order of $10^{-6} \Omega \text{ m}$. The low process temperatures are important, since they ensure ohmic contact formation well below critical temperatures for other processes, such as Si dopant migration in GaAs, which would damage the modulation doping profile.

Finally, one small practical note should be made here. Since, in many cases, mesoscopic transport experiments involve application of strong magnetic fields, it is very important that the ohmic contacts extend across the mesa edge. Otherwise, the contact resistance increases sharply in strong magnetic fields, since the electrons move in cyclotron orbits in the electron gas, and localize within a small area around the contact (see Fig. 4.16). This problem arises in particular in two-dimensional electron gases and at cyclotron radii below the mean free path.

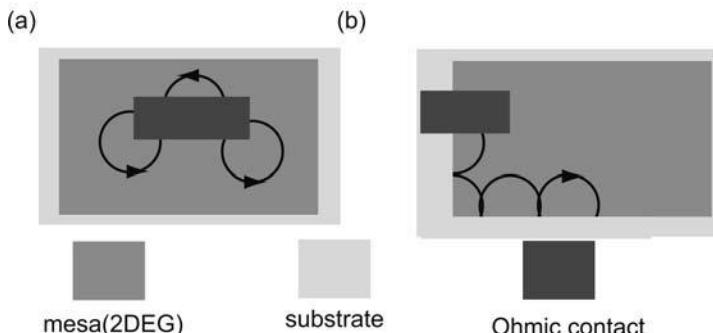


Fig. 4.16 In strong magnetic fields, electrons move in cyclotron orbits and thus remain localized close to the ohmic contact (a), unless the contact crosses the mesa edge (b).

4.1.5 Bonding

Once the sample is patterned and everything looks good, the last step in the fabrication process is to mount the sample into a chip carrier and to connect wires to the ohmic contacts and to the gate electrodes. Two versions of this so-called *bonding* are widely used. In ball bonding, the tip of a gold wire is molten locally by a discharge or a flame, and is pressed against a bond pad defined

3) The numbers here give the weight fraction.

on the sample surface. The sample is heated to a moderate temperature, say 200°C, and a connection forms via thermo-compression. The second scheme is known as wedge bonding (see Fig. 4.17). Here, the wire is pressed against the bond pad and rubbed across it with an ultrasonic frequency. The friction force is sufficient to locally melt the materials, and an alloy is formed that holds the wire in place. After the second bond, the wire is pulled and breaks at the weakest point, which is right after the position of the wedge.

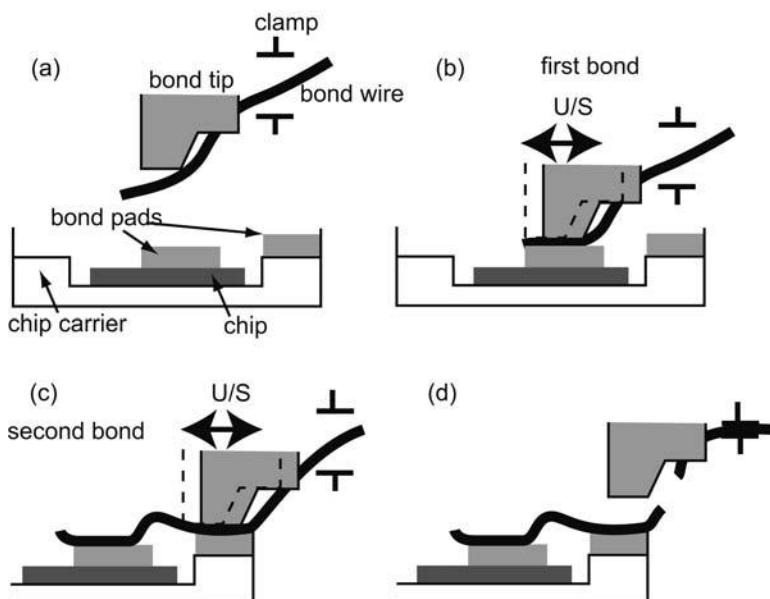


Fig. 4.17 Wedge bonding. (a) The bond tip containing the wire is positioned on top of the bond pad on the sample. (b) The wire is wedged onto the bond pad, and the tip is retracted with the wire clamp open. (c) The second bond is performed on the pad integrated in the chip carrier. (d) Here, the tip retracts with the clamp closed, and the wire breaks behind the second wedge.

4.2

Elements of cryogenics

Helium is the only element that remains liquid when cooled to the lowest possible temperatures (well below 1 mK) at atmospheric pressure. It is therefore the prime candidate as a refrigeration medium for temperatures below the condensation temperature of nitrogen (77 K). The vast majority of mesoscopic transport experiments are performed in this temperature range. The latent heat that has to be paid when liquid helium is evaporated generates the

cooling power made use of in helium cryogenics. Continuous evaporation of liquid is possible by pumping off the vapor pressure. Therefore, we will look at the properties of liquid helium (LHe), as well as cryostats, the devices used to establish low temperatures.

4.2.1

Properties of liquid helium

The physics of LHe is extremely interesting and rich, and experimentalists working on transport in nanostructures will almost inevitably come into contact with its unusual properties.

Helium comes in two isotopes, the boson ${}^4\text{He}$ and the fermion ${}^3\text{He}$. The mono-isotopic liquids therefore have very few properties in common. As a liquid is cooled, kinetic energy is taken away from the atoms. At the condensation temperature, the attractive interatomic van der Waals forces start to dominate in any liquid other than LHe, and crystallization sets in. Helium is the only element for which the van der Waals force is smaller than the kinetic energy of the atoms due to zero-point fluctuations. The van der Waals forces in He are particularly weak since the atoms have no dipole moment. On the other hand, the zero-point fluctuation energy is particularly large, due to the small atomic mass. Only by applying a pressure above ≈ 30 bar are the atoms squeezed sufficiently close together such that crystallization sets in.

So much for the common properties of ${}^3\text{He}$ and ${}^4\text{He}$. We now look at some properties of the pure isotopic liquids, before we turn to the interesting issue of ${}^3\text{He}/{}^4\text{He}$ mixtures.

4.2.1.1 Some properties of pure ${}^4\text{He}$

Fig. 4.18 shows the phase diagram of ${}^4\text{He}$. Under atmospheric pressure, it liquefies at $\Theta = 4.2\text{ K}$. The density of the liquid is $\rho(\text{L}{}^4\text{He}) = 125\text{ kg/m}^3$. The vapor pressure drops approximately exponentially as LHe gets colder, and reaches 1 mbar at $\Theta = 1.2\text{ K}$. As we cool the liquid, we cross the λ line at some temperature, which for atmospheric pressure happens at $\Theta_\lambda = 2.17\text{ K}$, also known as the λ point. The λ point got its name from the specific heat as a function of Θ around this transition, a function that looks like this omnipresent Greek letter. For $\Theta > 2.17\text{ K}$, ${}^4\text{He}$ behaves just like any ordinary liquid. As we lower Θ and cross the λ point, ${}^4\text{He}$ undergoes a phase transition and develops highly remarkable properties. ${}^4\text{He}$ in this phase is often referred to as He II. In fact, the phase transition at the λ point can be modeled as a Bose–Einstein condensation, i.e. the condensation of a boson gas. Within such a model, ${}^4\text{He}$ above the λ point is described as a gas, which is not a bad approximation, considering the weak interactions. At $\Theta = 0$, on the other hand, all atoms of He II are in the ground state. At higher temperatures, the

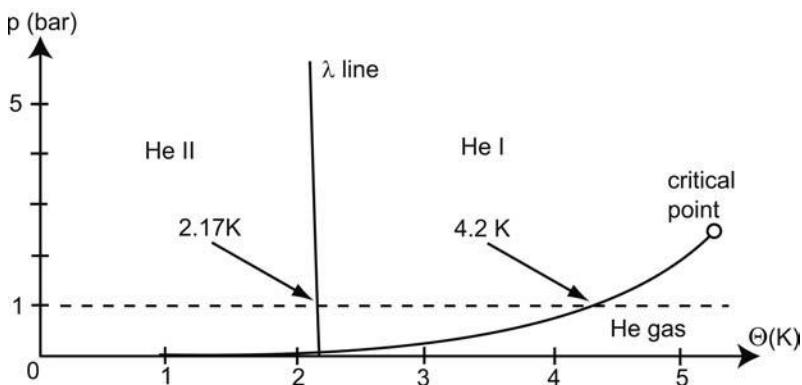


Fig. 4.18 Phase diagram of ${}^4\text{He}$.

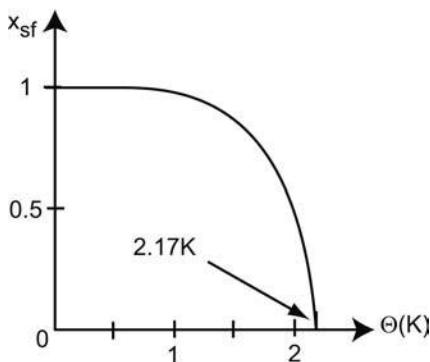


Fig. 4.19 Superfluid fraction x_{sf} of He II as a function of temperature.
After [11].

energy levels in a Bose–Einstein condensate (BEC) are occupied according to the Bose–Einstein distribution function

$$f_{\text{BE}}(E, \Theta) = \frac{1}{e^{(E-\mu)/k_B\Theta} - 1}$$

A pure BEC, however, cannot explain the observed behavior of He II. Rather, in [192], a two-liquid model has been proposed, which treats He II as a mixture of a normal fluid and a superfluid, which interpenetrate on a microscopic length scale, similar to the electronic state in a type II superconductor. The normal fluid behaves just like ${}^4\text{He}$ above the λ point. In particular, it has a non-vanishing entropy and viscosity. The superfluid, on the other hand, has zero entropy and viscosity, which means, for example, that there is no flow resistivity. Furthermore, the thermal conductivity of the superfluid is infinitely large. How the composition of He II changes with temperature has been measured in a seminal experiment [11]: A torsion pendulum made of a stack of thin disks was immersed in He II, and the damping of the oscillation was mea-

sured as a function of temperature. Since the normal fluid is viscous, it adds to the moment of inertia of the system via the Hagen–Poiseuille law, while the superfluid does not. The measured composition of He II is shown schematically in Fig. 4.19. As the temperature is lowered, the normal fluid fraction rapidly vanishes and an almost pure superfluid remains for $\Theta < 0.7\text{ K}$. This two-component mixture has some unique properties that we should know, in order to appreciate its behavior in cryogenic equipment.

Absence of bubbling If we heat a conventional liquid, it starts bubbling, since the liquid evaporates at some random spot, and the gas bubble rises to the surface. In He II, the thermal conductivity is very large, and evaporation takes place at the surface only. Hence, He II is perfectly quiet, even if it boils off. In a simple picture, we can understand the extremely high thermal conductivity as follows. Imagine we connect heat reservoirs to both ends of a tube filled with He II. At the end with higher temperature, superfluid is transformed into normal fluid, with a final ratio in accordance with Fig. 4.19. The heat is transferred to the low-temperature end by normal fluid convection. Here, the normal fluid is re-transformed into superfluid. Since the superfluid carries no heat (its entropy is zero), all the heat is thereby absorbed by the heat sink. The heat transfer is therefore very efficient.

He II osmosis Consider two chambers filled with He II, connected to each other by a *superleak*, i.e. a connection only permeable for superfluid helium (see Fig. 4.20(a)). Such connections can be made by extremely fine capillaries, or by tubes stuffed with powder. This setup immediately reminds us of an osmotic pressure cell, with the semipermeable membrane being the superleak, the solvent being the superfluid, and the normal fluid component starring as the solute. Recall that, in osmosis, the solute can be thought of as a gas, and that the osmotic pressure evolves due to the tendency of the solvent to equalize the concentrations in both chambers. As we heat He II in one chamber, the fraction of normal fluid increases, and superfluid will enter this chamber, in order to dilute it. Consequently, a pressure difference is built up. In equilibrium, the hydrostatic pressure will compensate the osmotic pressure, and the surface positions in the two chambers will differ by Δh .

Superfluid film creeping He II tends to creep over any wall of reasonable height, as long as its temperature stays below the λ point. Therefore, containers filled with He II to different heights will equilibrate their surface levels (see Fig. 4.20(b)). This effect has its origin in the extreme adhesion of He II to surfaces. Within the framework of liquid–solid interfaces, this is known as “complete wetting”. Since the shape of the liquid surface is determined by the

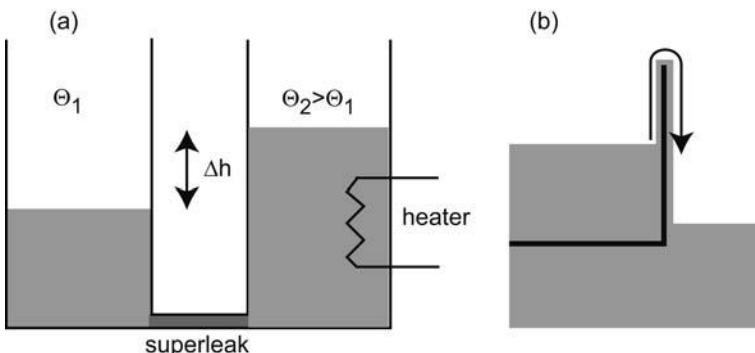


Fig. 4.20 (a) Sketch of a He II osmotic cell. (b) Superfluid film creeping across a wall.

condition that the tangential force vanishes, this effect occurs for $-\sigma_{ls} > \sigma_{gl}$, where σ_{ls} (σ_{gl}) denote the liquid–solid (gas–liquid) interface tension.

4.2.1.2 Some properties of pure ^3He

The phase diagram of ^3He is sketched in Fig. 4.21. For our purposes, the additional phase occurring at extremely low temperatures below 2 mK is irrelevant.⁴ The density of L^3He is $\rho(^3\text{He}) = 59 \text{ kg/m}^3$. Under atmospheric pressure, it liquefies at $\Theta = 3.19 \text{ K}$. This boiling point is about 1 K below that of ^4He , which can be easily understood, since its mass is smaller, and thus the atoms have a larger average velocity at the same temperature. Consequently, the vapor pressure is also higher at identical temperatures (see Fig. 4.24). It drops to 10^{-3} mbar at about $\Theta = 270 \text{ mK}$. ^3He atoms are fermions, and the liquid can be approximated by a Fermi gas, with many analogies to an electron gas.

Question 4.1: Calculate the Fermi energy of ^3He .

Within the Fermi liquid picture, we can imagine that each ^3He is surrounded by a screening cloud, which results in quasi-particles with an effective mass given by the interactions. At atmospheric pressure, $m^*(^3\text{He}) \approx 3m(^3\text{He})$. For practical cryogenic purposes, ^3He behaves as an ordinary liquid.

- 4) For $\Theta < 2 \text{ mK}$, the ^3He atoms form Cooper pairs and undergo a Bose–Einstein condensation into superfluid ^3He . Further phases exist at high pressures.

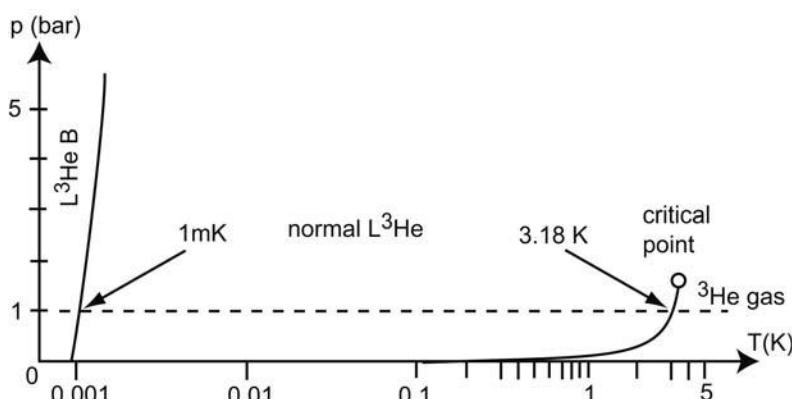


Fig. 4.21 Phase diagram of ${}^3\text{He}$.

A further important point concerning ${}^3\text{He}$ is its near-to-complete natural absence on Earth. It can be produced by nuclear reactions and is extremely expensive. Therefore, all ${}^3\text{He}$ cryostats keep it in a closed cycle.

4.2.1.3 The ${}^3\text{He}/{}^4\text{He}$ mixture

Let us first look at the phase diagram of this mixture (Fig. 4.22(a)). For $\Theta > 860 \text{ mK}$, nothing spectacular happens. The main effect of the ${}^3\text{He}$ is to reduce the λ point of the homogeneous mixture. Below the λ line, ${}^3\text{He}$ dissolved in He II can just be thought of as an additional fraction of the normal fluid component. For temperatures below 860 mK , a remarkable phase separation into a ${}^3\text{He}$ -poor phase (called the *dilute phase*, D) and a ${}^3\text{He}$ -rich phase (called the *concentrated phase*, C) takes place. At these temperatures, the pure He II is almost completely superfluid, and the dissolved ${}^3\text{He}$ forms a normal fluid component.

A qualitative understanding of the phase separation can be obtained by recalling that ${}^3\text{He}$ is a Fermi liquid, while ${}^4\text{He}$ in this regime is a Bose–Einstein condensate. The ${}^3\text{He}$ dissolved in ${}^4\text{He}$ can be thought of as a dilute Fermi gas with an effective mass given by the interaction between the ${}^3\text{He}$ atoms and the surrounding ${}^4\text{He}$, which is $m^*({}^3\text{He in } {}^4\text{He}) \approx 2.4m({}^3\text{He})$. Since superfluid ${}^4\text{He}$ has zero viscosity, the ${}^3\text{He}$ atoms can move around without friction, once the ${}^3\text{He}$ – ${}^4\text{He}$ interaction is included in the effective mass. L^3He can be regarded as a Fermi gas as well. We just have to establish the conditions for which the chemical potentials of the C phase and the D phase are identical. Here, the superfluid ${}^4\text{He}$ plays no role, as all these atoms are in the ground state. The problem somewhat resembles the alignment of chemical potentials at interfaces discussed in the previous chapter. Here, the common energy level is again the vacuum level, i.e. the energy of a ${}^3\text{He}$ atom at rest in the vacuum.

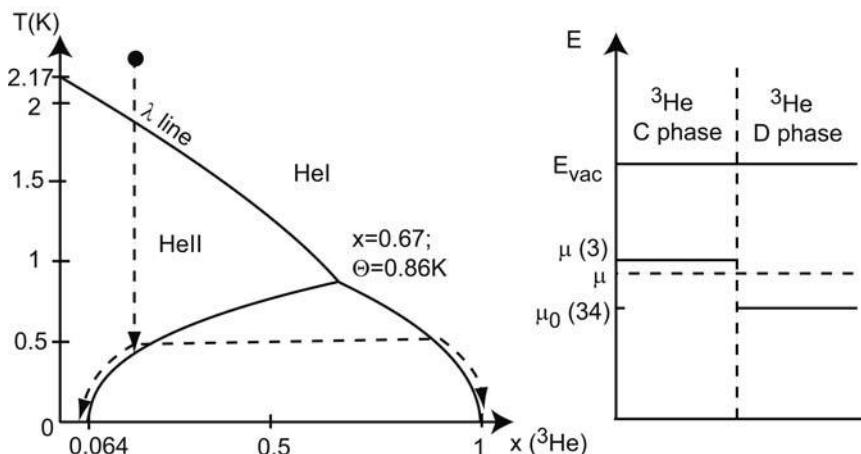


Fig. 4.22 Left: Phase diagram of the $^3\text{He}/^4\text{He}$ mixture vs. ^3He concentration x and temperature Θ . The tricritical point is at $x = 0.67$ and $\Theta = 860\text{ mK}$. At lower temperatures, the mixture segregates into a ^3He -rich concentrated (C) phase and a ^3He -poor dilute (D) phase. Right: Sketch of the chemical potential of the two phases at $\Theta = 0$.

The chemical potential, $\mu(3)$, of the C phase is somewhat higher than that, $\mu_0(34)$, of a single ^3He atom in ^4He , which can be understood by the fact that the (attractive) van der Waals forces are slightly larger in ^4He , since the average separation of the atoms is smaller. Hence, ^3He atoms will go into ^4He until the chemical potentials have aligned. This is the reason why even at $\Theta = 0$, the D phase still contains 6.4% of ^3He atoms. Note that it is energetically unfavorable for ^4He atoms to reside in the C phase.

4.2.2

Helium cryostats

Helium cryostats can be classified according to the kind of helium mixture for which they are designed. Occasionally, liquid nitrogen cryostats are used as well, for temperatures above 77 K. However, from our discussion of the ^4He cryostat, their design should be pretty obvious. We begin with the “high-temperature” helium cryostats.

4.2.2.1 ^4He cryostats

Helium has a small latent heat, which means it boils off easily. Therefore, the LHe cryostat has to be thermally decoupled from the environment. This is achieved by several means. Separating the He vessel from the outer world by a vacuum avoids heating via convection. Second, the LHe container is made of a material with a poor thermal conductivity, such as glass or stainless steel.

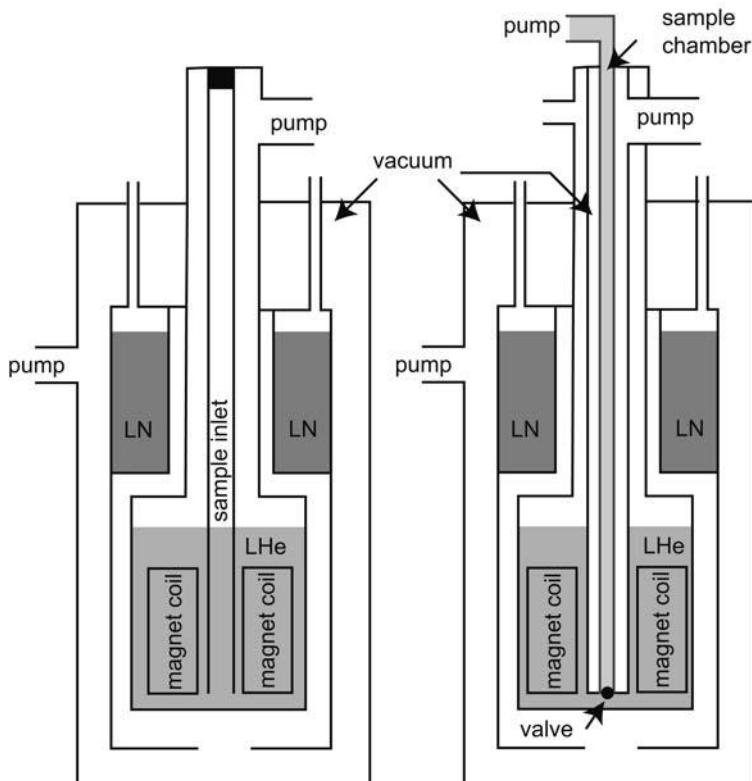


Fig. 4.23 Sketches of a ${}^4\text{He}$ bath cryostat (left) and a ${}^4\text{He}$ gas flow cryostat (right).

Finally, the thermal radiation from the environment is shielded by surrounding the LHe vessel with liquid nitrogen, in order to reduce the temperature of the blackbody radiation that hits the He dewar. Alternatively, it is possible to wrap the dewar in “super-insulating” foil, which is a multilayer of insulating foil, where each layer is coated with a metal on one side. Examples of L^4He cryostats are shown in Fig. 4.23.

In a *bath cryostat*, the sample is simply immersed in the LHe. The liquid, and with it the sample, can be cooled by pumping away the He vapor. This causes LHe to evaporate, which costs the latent heat and thus cools the liquid. The pumping speed and the incoming heat flux essentially determine the lowest possible temperature. To be somewhat more quantitative, recall the Clausius–Clapeyron equation, which gives the slope of the liquid–gas transition line as a function of temperature as

$$\frac{dp}{d\Theta} = \frac{L}{\Theta(V_{\text{gas}} - V_{\text{liquid}})} \quad (4.1)$$

Here, the latent heat per atom is given by $L = \Theta(S_{\text{gas}} - S_{\text{liquid}})$, where S_{gas} and S_{liquid} denote the atomic entropy of the gas and the liquid, respectively. We have further assumed here that L does not depend on temperature, which is a reasonable approximation for LHe. If we neglect the volume of the liquid (for LHe at 4.2 K, it is a factor of 750 smaller than the volume of the vapor), and model the gas as an ideal gas, $pV = nk_B\Theta$, it is found by integration of Eq. (4.1) that the vapor pressure p drops exponentially as Θ decreases, i.e.

$$p(\Theta) = p_0 \exp\left(-\frac{L}{k_B\Theta}\right) \quad (4.2)$$

The cooling power P is simply the latent heat taken from the liquid per evaporated atom, multiplied by the number of atoms evaporated per unit time,

$$P = \frac{dn}{dt}L \quad (4.3)$$

Since dn/dt is determined by the pumping speed dV/dT of the pump used via

$$\frac{dn}{dt} = \frac{1}{m_{\text{He}}} \frac{dM}{dt} = \frac{1}{m_{\text{He}}} \rho \frac{dV}{dt} = \frac{p(\Theta)}{k_B\Theta} \frac{dV}{dt}$$

the cooling power drops exponentially as Θ decreases.

Question 4.2: The latent heat of ${}^4\text{He}$ is 88 J/mol. What is the cooling power at $\Theta = 1.2\text{ K}$ when a pump with a pumping speed of $200\text{ m}^3/\text{h}$ is used?

The steady state is reached when the cooling power equals the heat load of the LHe. With a conventional pump with a pumping speed of, say, $10\text{ m}^2/\text{h}$, a temperature of about 1.2 K can be reached. Lower temperatures somewhat below 1 K are possible, but require very powerful pumps. Therefore, if this temperature range is needed, people usually prefer a ${}^3\text{He}$ cryostat or a dilution refrigerator.

Sometimes, temperatures above 4.2 K are required. The device of choice then is a *gas flow cryostat*. Here, the sample sits in a flow of cold helium gas, which enters the sample chamber via a needle valve. The sample chamber itself is thermally decoupled from the LHe by an additional vacuum chamber. The sample temperature can be adjusted by controlling the power applied to a heater for the gas, in combination with the gas flow rate. Continuous variation of the temperature between 1.2 K and room temperature is possible in gas flow cryostats. Many cryostats are equipped with a superconductive magnet, which is cooled below the critical temperature by the LHe. Most of these magnets are made from Nb alloys, since they have very large critical magnetic

fields. A typical magnet is able to generate magnetic fields of the order of 10 T, although magnets with maximum magnetic fields of 20 T and more are commercially available. Experiments at higher magnetic fields can be carried out at some national and international high magnetic field laboratories.

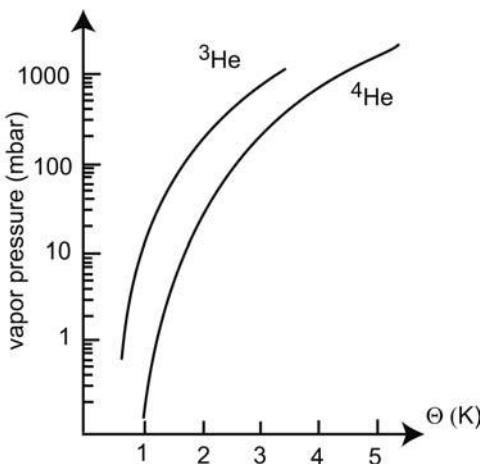


Fig. 4.24 Vapor pressure of ${}^3\text{He}$ and ${}^4\text{He}$.

4.2.2.2 ${}^3\text{He}$ cryostats

Below 1 K, the vapor pressure of ${}^3\text{He}$ is much higher than that of ${}^4\text{He}$ (Fig. 4.24). Therefore, temperatures down to about 270 mK can be reached easily by pumping L ${}^3\text{He}$. In a ${}^3\text{He}$ cryostat, the ${}^3\text{He}$ is isolated from the ${}^4\text{He}$ precooling stage by an inner vacuum chamber (Fig. 4.25). As mentioned already, the ${}^3\text{He}$ is kept in a closed cycle. The pumped ${}^3\text{He}$ gas is collected in a storage vessel. Measurements can be performed until all the L ${}^3\text{He}$ has been pumped, which results in measurement intervals up to one day. The ${}^3\text{He}$ gas can be condensed by a small, pumped ${}^4\text{He}$ pot, which is connected to the ${}^4\text{He}$ bath via a needle valve, such that its temperature stays well below 3.2 K, the condensation temperature of ${}^3\text{He}$. Some cryostats are equipped with a continuous flow option. Here, the pumped ${}^3\text{He}$ is immediately recondensed. For the price of a somewhat higher base temperature due to the additional heat load, the measurement period becomes unlimited this way.

4.2.2.3 ${}^3\text{He}/{}^4\text{He}$ dilution refrigerators

This type of cryostat uses the special properties of ${}^3\text{He}/{}^4\text{He}$ mixtures in a clever way, and makes possible temperatures as low as 1 mK and even lower. Since the D phase of the mixture is approximately a dilute Fermi gas, it can be thought of as the ${}^3\text{He}$ vapor of the C phase, with a significant vapor pressure

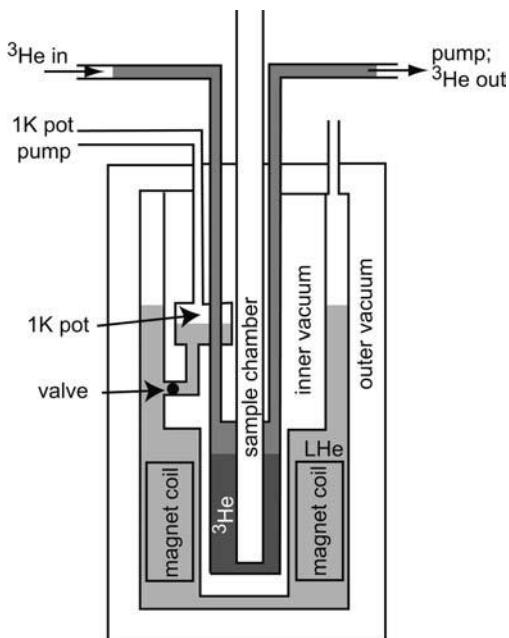


Fig. 4.25 Schematic sketch of a ${}^3\text{He}$ cryostat.

even at $\Theta = 0$. Since the C phase has a smaller density, the “liquid” will float on top of the “gas”, though. Pumping the ${}^3\text{He}$ atoms out of the D phase will surely cause ${}^3\text{He}$ from the C phase to evaporate, which pulls the corresponding effective latent heat out of the mixture. This is the cooling mechanism in a dilution refrigerator as sketched in Fig. 4.26. The mixture rests in the *mixing chamber*. The D phase is connected through a tube with the *still*, a pot that gets heated to about 600 mK. At this temperature, the vapor pressure of ${}^3\text{He}$ is significant, while that of ${}^4\text{He}$ is negligible. The still therefore effectively distills ${}^3\text{He}$ from the D phase. The missing ${}^3\text{He}$ in the D phase gets delivered by “evaporation” across the C–D phase boundary, and the mixture in the mixing chamber gets colder. Usually, the evaporated ${}^3\text{He}$ is recondensed into the mixing chamber by a pot filled with ${}^4\text{He}$, which gets pumped to temperatures below the condensation temperature of ${}^3\text{He}$. This is the “1 K pot”. The freshly condensed ${}^3\text{He}$, of course, still has a much higher temperature than the mixture. The heat flow in the mixing chamber is therefore optimized by a flow impedance in the condenser line. In addition, the outgoing gas at the still temperature is used to further precool the condensed ${}^3\text{He}$ via heat exchangers. Virtually all mesoscopic transport experiments below 270 mK have been carried out by thermally coupling the sample to the mixing chamber, either by immersing it directly, or by mounting it in the vacuum at the outside wall of the mixing chamber.

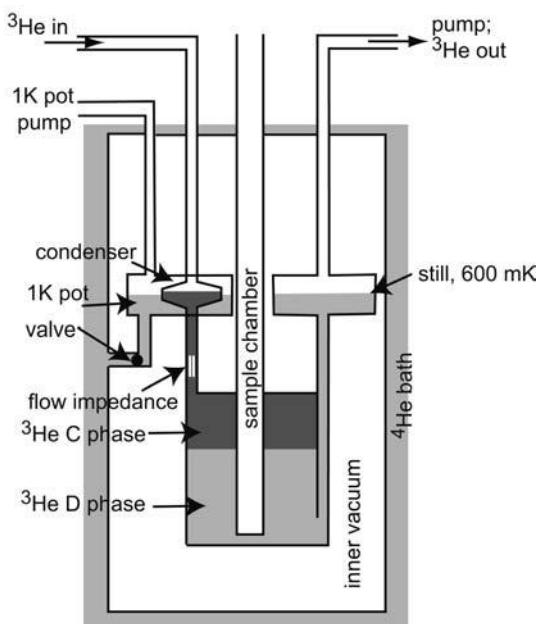


Fig. 4.26 Essential components of a ^3He / ^4He dilution refrigerator.

4.3

Electronic measurements on nanostructures

Measuring the resistance and conductance of a sample requires the application of currents and/or voltages, as well as the detection of voltage drops and/or currents, respectively. Conceptually, these measurements are very simple. The greatest efforts in practice are usually related to the reduction of the electronic noise level. This is done by avoiding ground loops, filtering, and choosing the right cables, among other important issues. As in the previous section, we are not that much interested in these technical details. This topic has been dealt with in great detail in excellent books (see the further reading section at the end of this chapter). Our goal here is to present in brief some basic setups, just enough for the reader to know what type of setup has been used in the experiments to be discussed. In the previous section, we have seen that the cryostats available set some limitations to the temperature range. Likewise, the measurement setup limits the physical quantities, as well as their ranges, that can be measured. The present section, together with the previous one, should put us in a position to judge why a particular experimental setup has been used, and how it affects the parameter ranges. We begin by showing how the samples are actually mounted in the low-temperature environment, before we discuss the most important electronic measurement setups.

4.3.1

Sample holders

A sample holder contains the sample in an appropriate way, and is mounted in the sample space of the cryostat. Its basic components are sketched in Fig. 4.27. The sample is mounted in some kind of carrier, which is placed inside the cryostat, in the center of the magnetic field. Cables are brought into the sample space via a vacuum feedthrough at the top end. Typically, the wires run in twisted pairs, which reduces the currents induced by the magnetic field due to vibrations, since the magnetic flux through adjacent loops points in opposite directions. Furthermore, the sample holder contains baffles, i.e. polished metal plates which reflect the thermal radiation from the top. Some sample holders are equipped with a rotator, which permits the sample to be tilted with respect to the magnetic field (which points in the vertical direction in most cryostats).

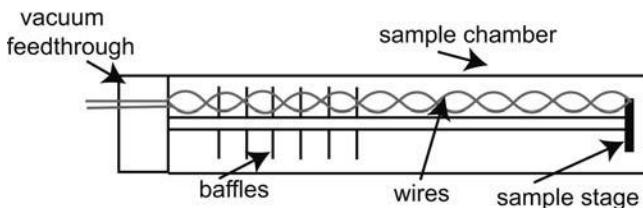


Fig. 4.27 Sketch of a sample holder used to place a specimen in the sample space of a cryostat.

4.3.2

Application and detection of electronic signals4.3.2.1 **General considerations**

For many experiments, measuring in a low-temperature environment only makes sense when the electrical signals are kept sufficiently small. Suppose, for example, that we plan to investigate the transmission properties of a tunnel barrier. The low temperature reduces the thermal smearing of the Fermi function, which corresponds to an energy scale of $\delta E = 3.52k_B\Theta \approx 300 \mu\text{eV}/\text{K} \cdot \Theta$. Therefore, the voltage drop across the barrier at, for example, $\Theta = 1 \text{ K}$ should be small compared to $300 \mu\text{V}$. For larger voltage drops, the temperature no longer determines the energy resolution.

Measurements can be performed AC or DC. AC measurements have the advantage that a lock-in amplifier can be used, a device that selectively detects signals with the source frequency, within a narrow bandwidth. In addition, phase-sensitive measurements are possible, such that, for example, capaci-

tance measurements can be performed by measuring the voltage drop with a phase shift of $\pi/2$ with respect to the source signal. Although the frequency selection greatly reduces the noise, it is not always best to use an AC signal. For example, imagine the sample has a very large resistance, such that the capacitances, which are always present in the leads, cause significant phase shifts. This makes it hard to determine the resistive part of the impedance. Also, theoretical results are often obtained for DC transport.

Furthermore, the distinction between resistance and resistivity (conductance and conductivity, respectively) has to be clearly understood. The plain result of, say, applying a current I and measuring the voltage drop ΔV is the resistance, $R = \Delta V/I$. If the sample is macroscopic, we can assume that the voltage drops homogeneously in between the voltage probes, and we can translate the resistance into a resistivity, an intrinsic property of the sample, by taking the sample geometry into account. This is no longer true in mesoscopic samples. Here, the measurement does not average over a large volume of randomly distributed scatterers, and the sample simply does not have a resistivity.

4.3.2.2 Voltage and current sources

High-quality commercial voltage sources typically provide voltages in the range of volts, with an accuracy of, say, 10^{-6} . Hence, some conversion to smaller voltages, or to a small current, is often necessary. This is done by a voltage divider, or a voltage-to-current conversion, respectively (see Fig. 4.28). The voltage divider simply consists of two resistors in series connected to the output voltage V_S of the commercial voltage source. The potential in between the two resistors is applied to the sample with respect to ground. This voltage is given by

$$V_{\text{out}} = V_S \frac{R_2}{R_1 + R_2}$$

In order to divide V_S by a few orders of magnitude, R_1 must be much larger than R_2 . This immediately implies an experimental limitation, since R_1 adds to the effective internal resistance of the voltage source. Connecting a sample with a resistance of R_S causes the applied voltage to drop to

$$V_{\text{out}} = \frac{R_2}{R_1 + R_2 + (R_1 R_2 / R_S)}$$

The circuit in Fig. 4.28(a) is only a good voltage source for $R_S \gg R_1$. So R_1 should be chosen as small as possible. The required output voltage then determines R_2 . These resistors cannot be arbitrarily small, however, since a minimum current of $I = V_S / (R_1 + R_2)$ must be provided by the voltage source. Hence, only samples of high resistance should be voltage-biased with such a setup.

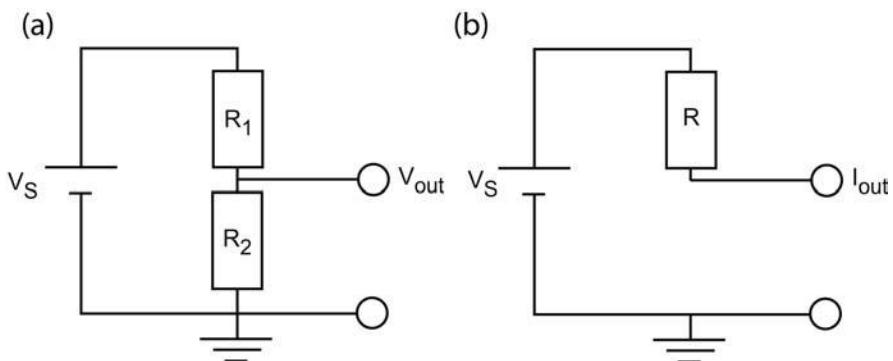


Fig. 4.28 A voltage divider (left) and a voltage-to-current conversion (right).

An analog consideration holds for current biasing the sample – see the circuit of Fig. 4.28(b). The current I_{out} is simply given by $I_{out} = V_S/R$. The setup is a good current source only if the sample resistance is small compared to the conversion resistance. On the other hand, the available voltage sources, the noise, and the minimum currents needed limit R to $R < 100 \text{ m}\Omega$. Consequently, samples with low resistance should be current-biased.

Question 4.3: Calculate how the sample resistance modifies the current in the setup of Fig. 4.28(b).

4.3.2.3 Signal detectors

A signal detector should not modify the measurement, which implies that the input resistance of a voltage detector should be large compared to the sample resistance, while that of a current detector should be small. The simplified setup shown in Fig. 4.29(a) shows the principle of voltage amplification with a transistor, which we suppose to be a Si MOSFET or a Ga[Al]As HEMT. Properly designed, they will operate at low temperatures as well, and can be integrated into the chip that hosts the experiment, which is useful in some cases. The advantage is an enhanced sensitivity and reduced thermal noise. The voltage to be amplified is superimposed onto the gate voltage, which defines the operating point of the transistor. The output voltage is the voltage drop between source and drain, which is highly sensitive to the gate voltage. A supply voltage is applied between source and drain, in series with a resistor R at the source side. Hence,

$$V_{out} = V_{\text{supply}} - RI_{SD}$$

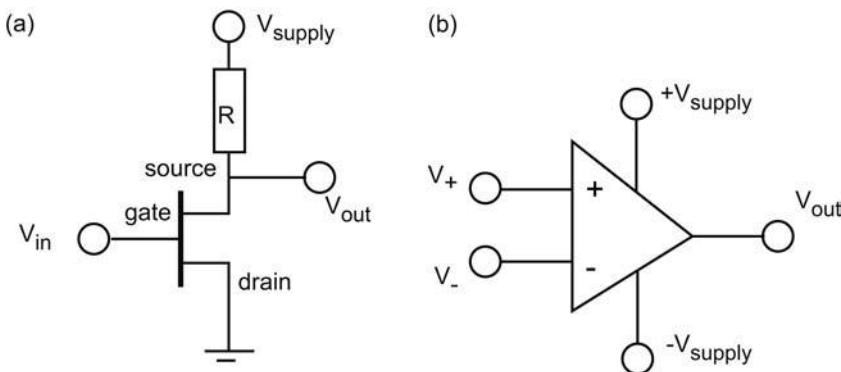


Fig. 4.29 Left: Scheme of a field effect transistor circuit used to amplify the input voltage V_{in} . Right: Symbol of an operational amplifier. The supply voltage is typically $\pm 15 \text{ V}$.

Here, I_{SD} is the current that flows from the supply to drain. We have assumed that the current that flows between the gate and source or drain is small compared to the current provided by the supply, which is reasonable in field effect transistors. A small⁵ input voltage V_{in} changes the current by

$$\Delta I_{\text{SD}} = t V_{\text{in}}$$

where $t = \partial I_{\text{SD}} / \partial V_{\text{in}}$ denotes the *transconductance* of the transistor at the operating point. Consequently, the output voltage changes according to

$$\Delta V_{\text{out}} = -Rt V_{\text{in}}$$

The amplification is thus $a = -Rt$. Since t is typically of the order of 10^{-3} A/V , an amplification by $\approx 10^3$ can be obtained in this way.

More common, however, are detector circuits that rely on operational amplifiers. A scheme is shown in Fig. 4.29(b). Operational amplifiers are three-terminal devices with two inputs and one output. In addition, they require a bipolar supply voltage of typically $\pm 15 \text{ V}$. Their internal structure is of no further interest to us here. We consider them as a black box with the following features.

- The input resistance is very high, e.g. $10^{12} \Omega$.
- The output resistance is small, typically of the order of 100Ω .
- For the circuit of Fig. 4.29(b), the output voltage is proportional to the difference of the two input voltages: $V_{\text{out}} = a_0(V_{\text{in},+} - V_{\text{in},-})$. The amplification $a_0 \approx 10^6$ to 10^8 is the *open loop gain*. Note, however, that the output voltage cannot exceed the supply voltages.

5) By "small", we mean voltages in the microvolt regime.

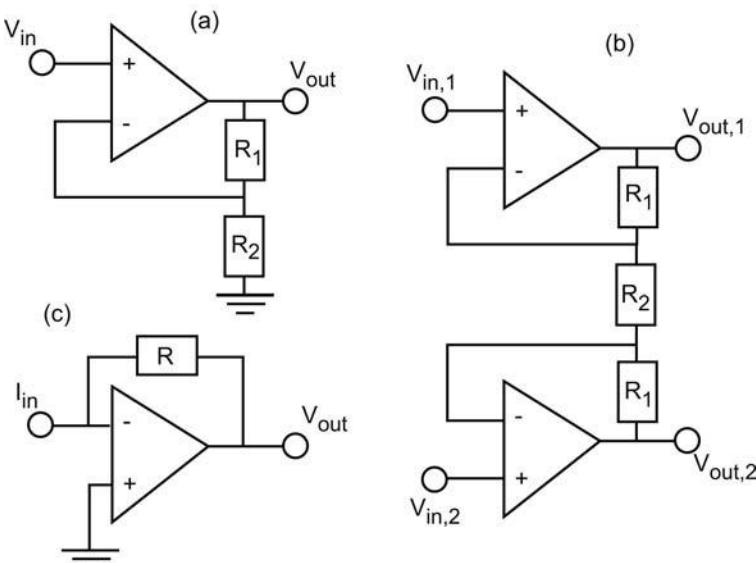


Fig. 4.30 (a) A voltage amplifier, (b) a differential amplifier, and (c) a current-to-voltage converter.

- If a fraction of the output is, via some circuit elements, fed back into one of the inputs, the operational amplifier adjusts its output V_{out} such that $(V_{\text{in},+} - V_{\text{in},-}) = 0$.

These properties make operational amplifiers extremely useful. We study how operational amplifiers can be used conceptually to measure voltages and currents. The circuit of Fig. 4.30(a) is a voltage amplifier. Why? Suppose $R_1 \gg R_2$. The voltage to be amplified is connected to the “+” input, and the output assumes the value $V_{\text{out}} = a_0(V_{\text{in}} - V_-)$. The voltage divider connected to the output determines $V_- = V_{\text{out}}R_2/(R_1 + R_2)$. Hence,

$$V_{\text{out}} = \left(V_{\text{in}} - V_{\text{out}} \frac{R_2}{R_1 + R_2} \right) a_0 \implies V_{\text{out}} \left(\frac{1}{a_0} + \frac{R_2}{R_1 + R_2} \right) \approx V_{\text{out}} \frac{R_2}{R_1} = V_{\text{in}}$$

The approximation is valid for $a_0 \gg R_1/R_2$. Hence, we see that the feedback reduces the open loop gain to the amplification R_1/R_2 , which can be chosen within wide ranges. Note that, in this circuit, the input resistance of the voltmeter is very high, namely that of the “+” input. Note further that this example implies that $a_0 = \infty$ for an ideal operational amplifier, which leads to the condition $(V_{\text{in},+} - V_{\text{in},-}) = 0$.

In many experiments, voltage differences are to be measured, either to exclude contact resistances (see below) from the measurements, or to measure without a direct reference to ground. Also, differential measurements limit the pickup noise, since a large fraction of this noise will be identical in both

measurement wires. Differential measurements can be made with a differential amplifier as shown in Fig. 4.30(b). It should now be easy for you to work out the amplification of this circuit.

Question 4.4: Verify that the amplification of the circuit in Fig. 4.30(b) is

$$\frac{V_{\text{out},1} - V_{\text{out},2}}{V_{\text{in},1} - V_{\text{in},2}} = 1 + 2 \frac{R_1}{R_2}$$

To conclude this brief section on operational amplifiers, let us have a look at the current meter depicted in Fig. 4.30(c). Here, the “+” input is grounded, and an input *current* is applied to the “−” input. There is no place for the current to go, and it thus charges up the input capacitance C_- , which is inevitably present. The total current that arrives at the “−” input is

$$I_-(t) = I_{\text{in}}(t) + \frac{1}{R} V_{\text{out}}(t)$$

On the other hand,

$$V_{\text{out}}(t) = a_0(V_{\text{in}}^+ - V_{\text{in}}^-) = -\frac{a_0}{C_-} \int_{t'=0}^t I_-(t') dt'$$

We differentiate this expression with respect to t and substitute $I_-(t)$ with the previous equation. This gives

$$-\frac{C_-}{a_0} \frac{dV_{\text{out}}}{dt} = I_{\text{in}}(t) + \frac{1}{R} V_{\text{out}}(t)$$

The left-hand side is approximately zero, due to the large open loop gain. Consequently,

$$V_{\text{out}}(t) = -RI_{\text{in}}(t)$$

The input current is converted into a voltage with a conversion ratio determined by the resistor. Its resistance can be very high, like $R \approx 1 \text{ G}\Omega$, since the condition is that R must be small compared to the input impedance. Thus, the output voltage adjusts in such a way that there is no charge buildup at the input. The current is effectively drained at the “−” input. For this reason, the “−” input is sometimes referred to as *virtual ground*. In cryogenic experiments, the conversion resistor R is sometimes mounted inside the cryostat, in order to reduce the thermal noise.

4.3.2.4 Some important measurement setups

As we have just seen, low-resistance samples should be investigated by applying a current and detecting the voltage drop (Fig. 4.31). This can be done in a

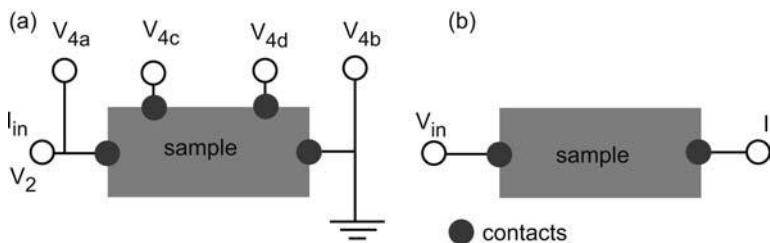


Fig. 4.31 (a) Two-point and four-point resistance measurements.
(b) Setup for a conductance measurement.

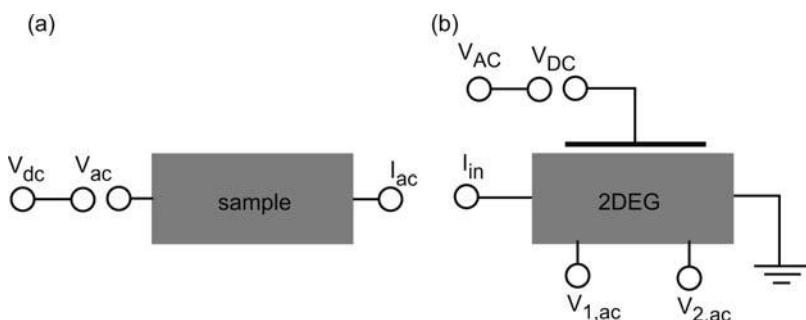


Fig. 4.32 Setups for measuring (a) the differential conductance, and (b) the transconductance of a field effect transistor.

two-probe configuration, where the voltage drop is measured at the connections used to apply the current and ground the sample. This has the disadvantage that not only the sample is measured, but also the leads and the contacts, which in case of a 2DEG are the ohmic contacts between the sample surface and the electron gas. In a quasi-four-probe setup, two wires are connected to both contacts used. Applying a current I_{in} in Fig. 4.31(a) and measuring the voltage between 4a and 4b eliminates the wire resistance, but not the contact resistances. Therefore, a true four-probe configuration is preferable, where the contacts used for measuring the voltage are different from those used to pass the current through the sample. In Fig. 4.31(a), this setup corresponds to measuring $V_{4c} - V_{4d}$. True four-probe setups are not always possible, though. It is then difficult or even impossible to discriminate between the contact resistances, which may be quite high, and the resistance of the sample. Note that, in a two-probe measurement, we measure $R_{xx} + R_{xy}$. The individual components of the resistivity tensor can be measured only with the corresponding four-probe configurations.

A conductance measurement, on the other hand, is usually a two-terminal experiment. Here, the current meter is in series with the sample. This setup is preferable for samples with a high resistance above $\approx 1 \text{ M}\Omega$.

Sometimes, it is convenient to be able to measure a differential quantity, such as the differential conductance dI/dV , or the transconductance $t = dI/dV_G$, of a transistor. Here, I is the source–drain current, V the source–drain voltage, and V_G denotes the gate voltage. This can be done by superposing a small AC voltage onto the DC voltage that is tuned, and detecting only those signals with the superimposed frequency with a lock-in amplifier. Schematic setups are shown in Fig. 4.32. Such differential measurements often give a higher resolution and a lower noise level, since the lock-in technique can be used where absolute measurements must be performed DC. Of course, we obtain the current as a function of the gate voltage, or of the source–drain voltage, by simple numerical integration of the measured differential trace.

Papers and Exercises

- P4.1** The paper [320] reports the preparation of nanostructures of dimension two, one and zero by cleaved edge overgrowth. Work out how the authors managed to do this, and discuss their way of detecting the dimensionality.
- P4.2** A three-dimensional fcc lattice of self-assembled quantum dots has been grown by Springholz and coworkers [286]. Focus on the mechanism behind the ordering. How is the lattice constant tuned?
- P4.3** A clever way of electron beam-assisted pattern transfer is presented in [108]. Describe how it works and what advantages this technique may offer.
- P4.4** The *van der Pauw* technique, named after the author of the original proposal [235], is an important concept for measuring semiconducting samples. Use [312] to discuss this technique.
- E4.1** *Pressure considerations for molecular beam epitaxy.* Use the kinetic gas theory to show that gas molecules hit a unit area with a rate F , given by

$$F = \frac{p}{\sqrt{2\pi mk_B\Theta}}$$

where p is the partial pressure of molecules with mass m . How long does it take until a monolayer of oxygen has been formed on the surface? Assume an O_2 partial pressure of $p = 10^{-10}$ mbar. Assume further that all molecules that hit the surface remain adsorbed (i.e. the sticking coefficient is unity).

E4.2 Some considerations concerning clean rooms. Dust particles with a size above $> 500 \text{ nm}$ frequently cause trouble in microchip fabrication. They rest on the resist and generate defects, like interrupted connections, or short circuits. They are therefore also known as “killer defects”. The yield Y is the fraction of working microchips, which can be written as

$$Y = \frac{1}{(1 + AD)^n}$$

where D denotes the density of killer defects, A is the chip area, and n is the number of relevant process steps (steps that involve resist illuminations).

- (a) Suppose that $n = 12$ for fabricating a certain microchip. What is the maximum D when a yield of at least 0.5 is needed for a chip of size $A = 2 \text{ cm}^2$?
 - (b) Under these conditions, how many defects are acceptable on an 8-inch wafer?
 - (c) For a rough estimate, assume that, during a process step, 1/6th of all dust particles inside a volume of (8 in^3) get deposited on the wafer. What clean room class is needed in order to obtain a yield above 90%? [The class of a clean room is the number of particles with size above 500 nm in a volume of one cubic foot.]
- E4.3** An electron beam resist has a sensitivity of $S = 2 \text{ C/m}^2$. The pattern generator places the focus of the electron beam at positions in a grid of $2^{13} \times 2^{13}$ points (a “13-bit resolution”). The writing field A is chosen via the magnification of the electron microscope, and selected to an area of $100 \mu\text{m} \times 100 \mu\text{m}$. Calculate the *dwell time*, the time the electron beam has to rest at each position. What is the minimum size of a single illuminated spot that guarantees homogeneous illumination?

- E4.4** Analyze the operational amplifier circuit of Fig. 4.33. How is the output voltage related to a *time-dependent* input voltage? Discuss the response of the output voltage to a step-like input voltage $V_{\text{in}}(t) = V_0 \theta(t - t_0)$. Can you imagine a possible application of this circuit?

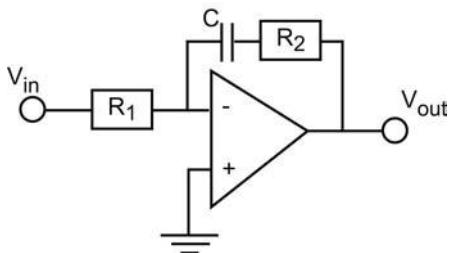


Fig. 4.33 Circuit for Exercise E4.4.

Further Reading

An overview of both the technology and the applications of semiconductor devices is given in [294]. A review of silicon processing technology has been given in [143]. For details of GaAs processing, see [331]. For both Si and GaAs processing, see e.g. [116].

A very nice review of many aspects of matter at low temperatures can be found in [206]. The book [329] treats the amazing properties of helium in a more rigorous way. In [195], you will find an extensive discussion of dilution refrigerators.

If you need recipes and practical tips for measurements in a cryogenic environment, you will find almost certainly what you need to know in [253] or in [255].

You are encouraged to read through an introductory textbook on electronics, like [154] or [101].

This Page Intentionally Left Blank

5

Important Quantities in Mesoscopic Transport

As pointed out already in the Introduction, the mesoscopic regime is characterized by certain scales in space, time, and energy. They will be introduced in this short chapter. We will frequently refer to these definitions later on.

5.1

Fermi wavelength

The *Fermi wavelength* $\lambda_F = 2\pi/k_F$ is the de Broglie wavelength of the electrons at the Fermi edge. Size quantization thus takes place at length scales comparable to λ_F , although we will see systems with characteristic sizes of $10\lambda_F$ that still show size quantization. The Fermi wavelength decreases as the electron density n_d (d denotes the dimensionality of the electron gas) increases, while the exact relation depends on d . For a spin degeneracy of 2 and within the effective mass approximation, one finds

$$\begin{aligned} \text{in } d = 3: \quad \lambda_F &= 2^{3/2} \left(\frac{\pi}{3n_3} \right)^{1/3} \\ \text{in } d = 2: \quad \lambda_F &= \frac{\sqrt{2\pi}}{n_2} \\ \text{in } d = 1: \quad \lambda_F &= \frac{4}{n_1} \end{aligned} \tag{5.1}$$

Thus, the Fermi wavelength is directly obtained from the electron density, which can be determined via Hall measurements. Note that λ_F does not depend on the effective mass.

5.2

Elastic scattering times and lengths

The *quantum scattering time* τ_q is the average time between successive elastic scattering events of arbitrary strength. It is related to the *quantum scattering length* ℓ_q via $\ell_q = v_F \tau_q$. Here, v_F denotes the Fermi velocity of the electrons at

the Fermi edge, i.e. $v_F = \sqrt{2E_F/m^*}$. Hence, ℓ_q is just the average distance the electrons at the Fermi energy travel without being elastically scattered. The quantum scattering length does not determine the resistivity, though. For momentum relaxation, the scattering angle is important as well. In fact, weighting each scattering event with scattering angle ϕ by the factor $(1 - \cos \phi)$ leads to the momentum relaxation time (which we will call the Drude scattering time) τ as introduced in Section 2.6.

The *elastic mean free path* ℓ_e is defined as $\ell_e = v_F \tau$ and represents the average distance an electron moves in between two subsequent, strong scattering events, also referred to as large-angle scattering events. In the case of a two-dimensional electron gas,

$$\ell_e = \frac{\hbar}{e} \mu \sqrt{2\pi n_2}$$

Question 5.1: Write down the expressions for ℓ_e as a function of μ and n for one and three dimensions.

Inserting typical numbers for a 2DEG in a GaAs HEMT at low temperatures, one finds $\ell_e \approx 8 \mu\text{m}$. The ratio τ_q/τ is determined by the relevance of various scattering mechanisms. Of course, $\tau_q/\tau \leq 1$ must hold. If this ratio is small compared to 1, the variance of the disorder potential is weak on the scale of the Fermi energy. For GaAs HEMTs at low temperatures, for example, one typically finds $\tau_q/\tau \approx 0.1$.

The Drude scattering time follows directly from the resistivity, once the electron density is known. The quantum scattering time can be extracted from $\rho_{xx}(B)$, provided that magneto-resistivity oscillations can be observed. This, by the way, is also a method to determine the effective mass. This analysis is discussed in Section 6.3.

5.3

Diffusion constant

The *diffusion constant* D originates from the diffusion equation

$$\frac{dn}{dt} = (\vec{\nabla} n) D (\vec{\nabla} n) \quad (5.2)$$

which tells us that gradients in the electron density cause diffusion (see e.g. [252]). We discuss the diffusion constant in a one-dimensional model; the extension to higher dimensions is straightforward. Owing to the Brownian motion, the electrons experience a fluctuating, Brownian force $\vec{b}(t)$, which

averages to zero in large time intervals. This can be included in the equation of motion for the electrons (Eq. (2.57)) by simply adding $\vec{b}(t)$ to the forces $\vec{F}(t)$ exerted by the external fields. The result is the Langevin equation

$$m^* \left(\frac{d\vec{v}}{dt} + \frac{\vec{v}}{\tau} \right) = \vec{F}(t) + \vec{b}(t) \quad (5.3)$$

Suppose the external forces are zero. From statistical physics [252], it is well known that the diffusion constant is obtained from the correlation function¹ $C_{v_i v_j}(t)$ of the electron velocities

$$D_{ij} = \int_{t=0}^{\infty} C_{v_i v_j}(t) dt, \quad C_{v_i v_j}(t) \equiv \langle v_i(0)v_j(t) \rangle \quad (5.4)$$

where $\langle \dots \rangle$ denotes an average over many electron trajectories, and i, j are the coordinates x and y . We take the derivative of $C_{v_i v_j}(t)$ with respect to time and replace $d\vec{v}/dt$ using the Langevin equation, which gives

$$\frac{dC_{v_i v_j}(t)}{dt} = \frac{1}{m^*} C_{v_i b_j}(t) - \frac{1}{\tau} C_{v_i v_j}(t) \quad (5.5)$$

Here, the first term on the right-hand side vanishes, as there is no correlation between the velocity at time $t = 0$ and the Brownian force at time t . One therefore obtains a differential equation for the velocity autocorrelation function with the solution

$$C_{v_i v_j}(t) = C_{v_i v_j}(0)e^{-t/\tau} \quad (5.6)$$

Consequently, within the approximations made, the diffusion constant equals

$$D_{ij} = C_{v_i v_j}(0)\tau \quad (5.7)$$

Since all electrons move with the Fermi velocity, but in arbitrary directions in a plane, $C_{v_i v_j}(0)$ for a two-dimensional system is obtained from

$$C_{v_i v_j}(0) = \frac{1}{2\pi} \int_0^{2\pi} v_i(0)v_j(0) d\phi \implies \begin{cases} C_{v_i v_i}(0) = \frac{1}{2}v_F^2 \\ C_{v_i v_j}(0) = 0 \quad (i \neq j) \end{cases} \quad (5.8)$$

which gives the diffusion tensor

$$D = \frac{1}{2}v_F^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (5.9)$$

¹) Correlation functions are introduced in Appendix B.

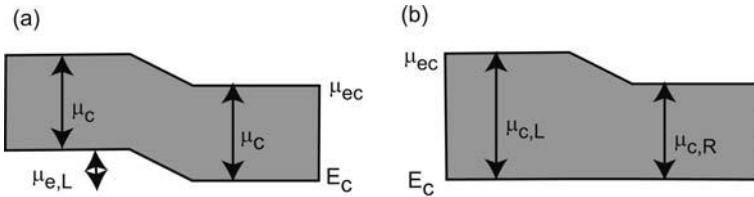


Fig. 5.1 (a) An electrostatic potential gradient and (b) a chemical potential gradient can generate identical gradients in the electrochemical potential.

Note that we can make use of the equipartition theorem of statistical mechanics and write for two dimensions

$$\frac{1}{2}mv_F^2 = k_B\Theta$$

which gives the *classical Einstein relation*

$$D_{ii} = \frac{k_B\Theta}{m^*}\tau = \frac{k_B\Theta}{e}\mu \quad (5.10)$$

whereas the Einstein relation for a Fermi gas (Eq. (2.65)) is obtained by replacing v_F with the Fermi energy.

A typical value for electron gases in Ga[Al]As HEMTs at low temperature is $D = 0.1 \text{ m}^2/\text{s}$.

Diffusion constant and mobility are thus intimately related. For a simple interpretation, look at Fig. 5.1. A gradient in the electrostatic potential, as well as a gradient in the chemical potential, leads to a spatially varying electrochemical potential and causes drift or diffusion, respectively.

It is worth emphasizing that, due to the Einstein relation, we can calculate the components of the conductivity tensor σ_{ij} from the velocity autocorrelation function. We recall that $\sigma_{ij} = ne\mu_{ij}$ and write the classical conductivity as

$$\sigma_{ij} = \frac{ne^2}{k_B\Theta} \int_{t=0}^{\infty} C_{v_i v_j}(t) dt \quad (5.11)$$

This is the simplest version of the *Kubo formula* [181]. It is frequently used in numerical simulations of the conductivity, where electrons are injected in random directions in the potential landscape and $C_v(t)$ is calculated. For a review of this technique, see [166].

For a degenerate electron gas (see Section 2.6.4) in two dimensions (a situation that we will encounter frequently below), we obtain instead

$$\sigma_{ij} = \frac{m^* e^2}{\pi \hbar^2} \int_{t=0}^{\infty} C_{v_i v_j}(t) dt \quad (5.12)$$

The diffusion constant enters in some of the scales to be introduced below. The Einstein relation makes clear that D can be obtained experimentally from the mobility.

5.4

Dephasing time and phase coherence length

Elastic scattering events, which determine the electron mobility at low temperatures, do not cause dephasing, since the phase shift experienced by the scattered electrons is reproducible. If the scattering event can, in any way, be regarded as a measurement of the electron's location, dephasing takes place. This is the case in spin-flip scattering events at magnetic impurities, or for electron–phonon scattering. Electron–electron scattering does cause dephasing, since energy is transferred between the scattering partners. However, the latter kind of scattering does not, in general, cause resistance, since the total momentum of the electron system remains unchanged. Therefore, electron–electron scattering gives to a first approximation no contribution to the resistivity. Transport effects that rely on interference of electronic wave functions, however, can be used to determine the *dephasing time* τ_ϕ . The theory for the magnitudes and the parametric dependence of τ_ϕ is developed in [3]. For two dimensions and in the diffusive regime, one finds

$$\frac{1}{\tau_\phi} = \begin{cases} \frac{\pi}{2} \frac{(k_B T)^2}{\hbar E_F} \ln \left(\frac{E_F}{k_B T} \right), & k_B T > \hbar / \tau_D \\ \frac{k_B T}{2m^* D} \ln \left(\frac{m^* D}{\hbar} \right), & k_B T < \hbar / \tau_D \end{cases} \quad (5.13)$$

This linear relation between $1/\tau_\phi$ and T at low temperatures, which changes to a quadratic dependence at higher temperatures, is found experimentally in reasonable agreement with the theoretical expressions; see e.g. [55].

The *phase coherence length* ℓ_ϕ is the distance the electrons travel before their phase is randomized. For $\tau_\phi < \tau$, one has $\ell_\phi = v_F \tau_\phi$. For $\tau_\phi > \tau$, however, the electrons get scattered elastically within the phase coherence time, and the distance they travel within τ_ϕ gets reduced. For $\tau_\phi \gg \tau_D$, which is often the case at low temperatures, $\ell_\phi = \sqrt{D \tau_\phi}$. Typical dephasing times in mesoscopic samples are of the order of 1 ps.

We will meet the dephasing time in Chapter 8, where we will also see how it can be determined experimentally.

5.5

Electron–electron scattering time

The *electron–electron scattering time* τ_{ee} is the average time of flight for the electrons between successive electron–electron scattering events. In a simple picture [12], we expect that $\tau_{\text{ee}} \propto 1/T^2$. Let us assume that a single electron (labeled 1) has an energy Δ above the Fermi energy of an electron gas at zero temperature: $E_1 = E_F + \Delta$. Consider how this electron can scatter with an electron (labeled 2) in the Fermi sea. For the energy of the second electron, $E_2 \leq E_F$ holds. In addition, both the final states (with energies E_3 and E_4) that the electrons occupy after the scattering must be empty: $E_3, E_4 > E_F$. Since $E_1 + E_2 = E_3 + E_4 > 2E_F$, it follows that

$$(E_1 - E_F) + (E_2 - E_F) = \Delta + (E_2 - E_F) > 0$$

Therefore, only electrons within a shell of thickness Δ below the Fermi edge can scatter at 1. This is a fraction Δ/E_F of all the electrons. Furthermore, E_3 and E_4 must be inside $[E_F, E_F + \Delta]$. This means that the electron–electron scattering probability is $\propto (\Delta/E_F)^2$. For a thermally smeared Fermi gas, we can identify $\Delta \approx k_B\Theta$, and so $\tau_{\text{ee}} \propto 1/T^2$ results. This argument has been verified to be approximately true by more sophisticated calculations in [119], which derived for the ballistic regime

$$\frac{1}{\tau_{\text{ee}}} = \frac{E_F}{4\pi\hbar} \left(\frac{\Delta}{E_F} \right)^2 \left[\ln \left(\frac{E_F}{\Delta} \right) + \ln \left(\frac{2k_{\text{TF}}}{k_F} \right) + \frac{1}{2} \right]$$

where k_{TF} is the Thomas–Fermi screening vector; see Chapters 2 and 3.

Although interactions are not treated explicitly, τ_{ee} will occasionally pop up, in particular in Chapter 8. It can be determined experimentally by magneto-resistivity measurements [54].

5.6

Thermal length

The *thermal length* ℓ_Θ specifies the length scale over which thermal smearing takes place. From the uncertainty relation $\hbar \leq k_B T \cdot \tau_T$ one obtains a “thermal time”, below which it is not possible to determine the energy of the electron better than $k_B\Theta$. This time scale corresponds to a length scale $\ell_\Theta = \sqrt{D\tau_T} = \sqrt{\hbar D/k_B T}$.

We will meet the thermal length in particular when phase coherence effects are discussed, namely in Chapter 8.

5.7

Localization length

The *localization length* ℓ_ξ is the average length over which an electronic state extends in a sample. The disorder potential localizes the states on this length scale. As we shall see in the following chapter, magnetic fields can tune the localization length over wide ranges.

5.8

Interaction parameter (or gas parameter)

The *interaction parameter* r_s is the ratio of the (unscreened) Coulomb energy between two electrons at their average distance, and their kinetic energy at the Fermi edge. In two dimensions,

$$r_s = \frac{e^2 / 4\pi\epsilon\epsilon_0 r}{m^* v_F^2 / 2} = \frac{e^2 m^*}{\epsilon\epsilon_0 h^2} \frac{1}{\sqrt{n_e}}$$

For 2DEGs in GaAs, $r_s \approx 1$. In Si MOSFETs as well as in hole gases residing in GaAs, r_s can be much larger, and values up to $r_s \approx 20$ have been reported.

Although we essentially treat the electron gases as non-interacting or weakly interacting, it is worth keeping in mind that this is not really true in low-dimensional electron gases. Strictly speaking, the validity of Fermi liquid theory in systems with $r_s > 1$ is questionable.

5.9

Magnetic length and magnetic time

A magnetic field sets a length scale as well, namely the spatial extension of wave functions in the magnetic field. It is given by the *magnetic length* $\ell_B = \sqrt{\hbar/eB}$, which corresponds to the width of the ground state of a quantizing magnetic field, as will be discussed in more detail in the next chapter. Also of importance is the *cyclotron radius* r_c , i.e. the radius of the circle the electrons follow in a magnetic field: $r_c = k_F \ell_B^2$, as can be easily checked. The *magnetic time* τ_B is the time an electron needs to diffuse across the area $\frac{1}{2}\ell_B^2$. It is given by $\tau_B \equiv \ell_B^2/(2D)$.

The magnetic length will be of particular importance in Chapters 6 and 7.

Exercises

E5.1 The quantities listed in Table 5.1 have been determined experimentally in Si MOSFETs and in GaAs HEMTs. Calculate the scattering time τ , the diffusion constant D , the Fermi wavelength λ_F , the phase coherence length ℓ_ϕ , the inelastic scattering length ℓ_{in} , the thermal length ℓ_T , and the interaction parameter r_s .

Tab. 5.1 Data for Exercise E5.1.

Physical quantity/material	GaAs ($T = 4.2 \text{ K}$)	Si ($T = 4.2 \text{ K}$)
electron density (10^{15} m^{-2})	4	0.7
electron mobility ($\text{m}^2/\text{V s}$)	100	4
effective mass, m_e	0.067	0.19
dephasing time (10^{-12} s)	30	10

Which material would you prefer for ballistic electron experiments, which for investigations of phase coherent electrons, and which for studying electron-electron interactions?

E5.2 Derive the magneto-resistivity tensor, Eq. (2.59), from Eq. (5.4) [Hint: Note that, in a magnetic field, the velocity correlation function becomes $C_{v_i v_j}(t, B) = \langle v_i(0)v_j(t, B) \rangle$.]

Further Reading

More about the relevant quantities in mesoscopic transport can be found in [65], as well as in the review article [27]. For an introduction to the Kubo formalism, see [347].

6**Magneto-transport Properties of Quantum Films**

Transport experiments in external magnetic fields are very common in mesoscopic physics. With the magnetic field, we can reversibly and – if we perform our experiment carefully enough – non-destructively tune various scales, such as the magnetic length, the effective mass, or the cyclotron energy. We have seen already that measuring the Hall resistivity in small magnetic fields allows us to determine the carrier density. In this chapter, we are mainly interested in strong magnetic fields, such that $\omega_c\tau \geq 1$. This condition simply means that the electrons can complete at least one cyclotron orbit before they get scattered. It is immediately clear that new effects can be expected in this regime. Recall Bohr's atomic model: discrete states are obtained from interference of electronic waves circulating around the nucleus. For an interference to be constructive, the circumference of the trajectory must be an integer multiple of the electronic wavelength. A similar thing happens in strong magnetic fields. Here, the electrons are forced to circulate in cyclotron orbits. The result, known as *Landau quantization*, is discussed in Section 6.1. In particular, it is a very important ingredient to the quantum Hall effect, which is the topic of Section 6.2. In Section 6.3, we return to intermediate magnetic fields and show how the magneto-oscillations observed in the longitudinal direction (Shubnikov–de Haas oscillations) can be analyzed to obtain the quantum scattering time and the effective mass. In the subsequent section (Section 6.4), we give a small selection of further magneto-transport experiments.

Up to that point, the magnetic field has been perpendicular to the plane of the electron gas. The basic effects in parallel magnetic fields are presented in Section 6.5, which concludes this chapter.

If not stated otherwise, the magnetic field is homogeneous and points in the z -direction, perpendicular to the plane of the 2DEG. In this case, we call it B . Magnetic fields in the plane of the electron gas are referred to as parallel magnetic fields, and are denoted by $B_{||}$.

6.1

Landau quantization

6.1.1

Two-dimensional electron gases in perpendicular magnetic fields

The Schrödinger equation of a free 2DEG in a magnetic field B reads¹

$$\left[\frac{(\vec{p} + e\vec{A})^2}{2m^*} + V(z) \right] \Phi(\vec{r}) = E\Phi(\vec{r}) \quad (6.1)$$

where \vec{A} denotes the vector potential. The z -direction is of no further interest to us, since B does not influence the electronic motion in that direction. We therefore assume that the z -direction can be treated separately, leading to a quantized energy E_z , which is the conduction band bottom of the 2DEG. We choose the Landau gauge, $\vec{A} = (-By, 0, 0)$, for mathematical simplicity.² The x - y Hamiltonian emerging from Eq. (6.1) reads

$$H_{xy} = \frac{1}{2m^*} [(p_x - eBy)^2 + p_y^2] \quad (6.2)$$

With the ansatz

$$\Phi(x, y) = \Psi(y)e^{ik_xx} \quad (6.3)$$

a one-dimensional Schrödinger equation in the y -direction is obtained,

$$\left[-\frac{\hbar^2}{2m^*} \frac{\partial^2}{\partial y^2} + \frac{\hbar^2 k_x^2}{2m^*} - \frac{\hbar k_x eBy}{m^*} + \frac{e^2 B^2 y^2}{2m^*} \right] \Psi(y) = (E - E_z)\Psi(y) \quad (6.4)$$

while plane waves are the eigenfunctions in the x -direction of the separated Hamiltonian. The Schrödinger equation for the y -direction can be mapped onto the harmonic oscillator equation

$$\left[-\frac{\hbar^2}{2m^*} \frac{\partial^2}{\partial v^2} + \frac{1}{2} m^* \omega^2 v^2 \right] u(v) = Eu(v) \quad (6.5)$$

by

$$\omega \rightarrow \omega_c = \frac{eB}{m^*} \quad \text{and} \quad v \rightarrow y - \frac{\hbar k_x}{m^* \omega_c} \quad (6.6)$$

The cyclotron frequency ω_c is the angular frequency of the electron in the magnetic field.

1) We define e as positive; the electronic charge is thus $-e$.

2) It is instructive to solve this problem in the symmetric gauge, $\vec{A} = 0.5(-By, Bx, 0)$, and use polar coordinates; see e.g. [87] for details.

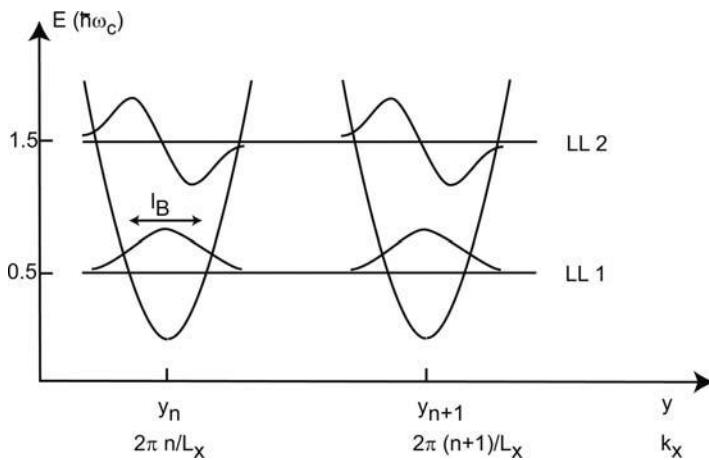


Fig. 6.1 Electronic states in a Landau level. The positions of the harmonic oscillator potentials y_n in the y -direction are given by the wave numbers k_x that satisfy the boundary condition.

For simplicity, let us consider a rectangular sample of area $L_x \times L_y$. Since k_x quantizes as a consequence of the periodic boundary conditions according to

$$k_{x,n} = \frac{2n\pi}{L_x} \quad (6.7)$$

the harmonic oscillators in the y -direction are centered at the positions

$$y_n = \frac{n\pi\hbar}{m^*\omega_c L_x} \quad (6.8)$$

The eigenfunctions of the $x-y$ Hamiltonian are thus plane waves in the x -direction, multiplied with Hermite polynomials in the y -direction, as shown schematically in Fig. 6.1.

Question 6.1: Check that the full width at half-maximum (FWHM) of the ground state in the y -direction equals the magnetic length.

The corresponding energy eigenvalues are

$$E_j = \hbar\omega_c(j - \frac{1}{2}) \quad (6.9)$$

with j being a positive integer. Besides spin and valley degeneracies, the degeneracy of each energy level is given by the number of allowed wave numbers in the x -direction. The states of energy E_j form the j th *Landau level*. In order to determine the degeneracy of a Landau level, we can use the fact that

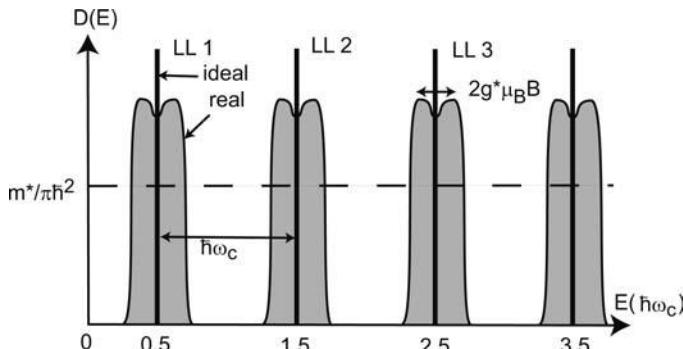


Fig. 6.2 Ideal and real densities of states of a Landau-quantized 2DEG that is spin degenerate at $B = 0$. The δ functions broaden due to fluctuations of the conduction band bottom, while the spin degeneracy is lifted, and a Zeeman doublet results for non-zero effective g-factors g^* .

the integrated density of states is independent of the magnetic field. Hence, the number of states per unit area in a Landau level must be $g_s m^* / (2\pi\hbar^2)$, multiplied by $\hbar\omega_c$; see Fig. 6.2. A degeneracy of $g_s / (2\pi\ell_B^2)$ per unit area in each energy level is obtained (g_s counts both the spin and valley degeneracies). Hence, the density of states of an ideal 2DEG in a perpendicular magnetic field reads

$$D(E) = \frac{g_s}{2\pi\ell_B^2} \delta(E - E_j) \quad (6.10)$$

with E_j given by Eq. (6.9).

Question 6.2: Calculate the degeneracy of a Landau level by counting the states with Eq. (6.8). Use the condition that all the harmonic oscillators must have their center y_n inside the sample. Assume further that the magnetic length is small compared to the sample size.

In real samples, the Landau levels are broadened with an approximately Gaussian shape by potential fluctuations, and split via the two alignments of the electron spin in the magnetic field (Fig. 6.2). The spin splitting is described by the effective g -factor g^* . For bulk GaAs, $g^* = -0.44$.³

In general, the Landau level at the Fermi energy is only partly occupied, and it is thus useful to introduce a quantity that measures the degree of filling of Landau levels. This is the task of the *filling factor* ν , defined as

$$\nu = \frac{g_s E_F}{\hbar \omega_c} \quad (6.11)$$

3) A magnetic field can strongly modify g^* .

For the frequent case of $g_s = 2$ (spin degeneracy), $\nu = 2j$ means that j Landau levels are completely filled. Furthermore, in sufficiently strong magnetic fields, the spin degeneracy may be lifted due to the Zeeman effect. In that case, an odd integer value of ν means that one spin direction of Landau level $j = \nu/2$ is full, while the other is empty.

6.1.2

The chemical potential in strong magnetic fields

In experiments, it is common to vary the Landau level occupation by tuning either the electron density or the magnetic field. Suppose B is fixed, the temperature is zero, and we tune the electron density. Let us for simplicity assume that there is no valley degeneracy, i.e. $g_s = 2$, and there is no spin splitting. The integrated density of states in each Landau level D_{LL} is then given by

$$D_{\text{LL}} = \frac{1}{\pi \ell_B^2} \quad (6.12)$$

The Fermi level E_F is independent of the electron density n as long as, in the highest occupied Landau level, there are empty states left. In this case, $E_F = (j - \frac{1}{2})\hbar\omega_c$. At electron densities $n_j = jD_{\text{LL}}$, all Landau levels are either full or empty, and the Fermi energy equals $E_F = j\hbar\omega_c$; see Fig. 6.3(a). In the latter case, we would classify the system as an insulator, since the density of states at the chemical potential is zero. Remarkably, a 2DEG in a magnetic field experiences a sequence of metal-insulator transitions as a function of n . In our ideal system with the density of states composed of δ functions (Eq. (6.10)), the insulating phases are just points along the n -axis. As we will see shortly, the insulating behavior extends over non-zero intervals in real samples, since the electronic states in the wings of a peak in $D(E)$ tend to be localized for a real sample.

Similar parametric metal-insulator transitions are, of course, also found as a function of B with the electron density fixed. Here, the insulating points correspond to magnetic fields

$$B_j = \frac{\pi \hbar n}{e} \frac{1}{j}$$

It is easily verified that at these magnetic fields, E_F changes from

$$E_F(B_j - \delta B) = \frac{\hbar e B_j}{m^*} \left(j - \frac{1}{2}\right)$$

via

$$E_F(B_j) = \frac{\hbar e B_j}{m^*} j = E_F(B=0)$$

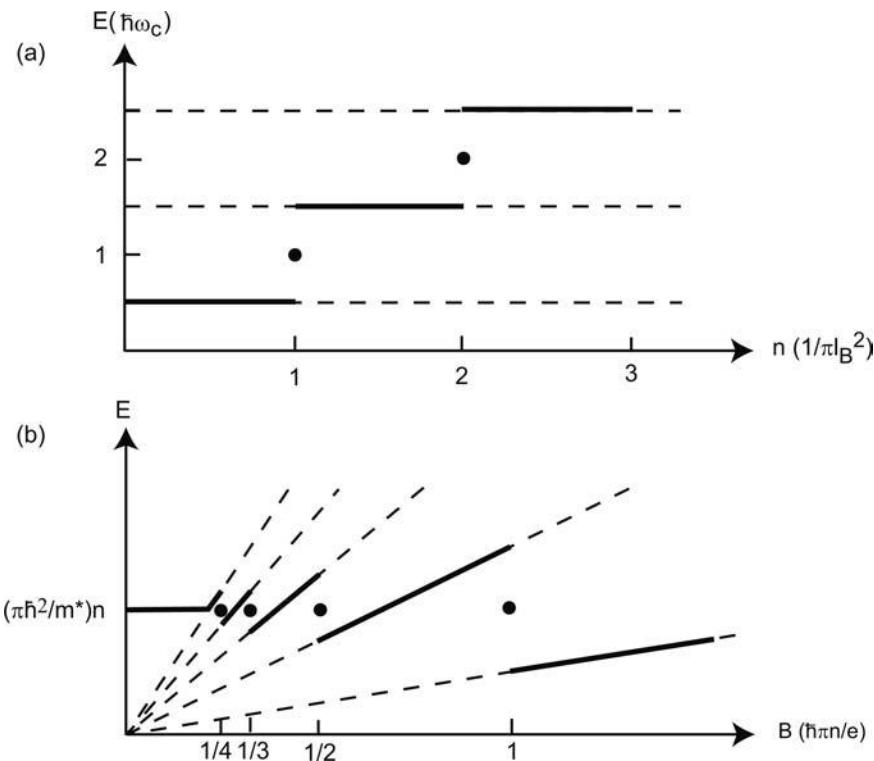


Fig. 6.3 Evolution of the Fermi level (a) as a function of electron density in a fixed magnetic field, and (b) as a function of the magnetic field with n fixed. An ideal density of states is assumed. In both scenarios, metal-insulator transitions exist, with insulating phases of zero width in the parameter coordinate.

to

$$E_F(B_j + \delta B) = \frac{\hbar e B_j}{m^*} \left(j + \frac{1}{2}\right)$$

where δB denotes an arbitrarily small magnetic field. This behavior is depicted in Fig. 6.3(b).

The density of states and its evolution in magnetic fields can be nicely detected by a powerful tool known as *capacitance spectroscopy*. The experimental setup resembles somewhat that used to measure differential conductances sketched in Fig. 4.32 (see Fig. 6.4(a)). The sample used for the experiment in Fig. 6.4 was a GaAs–Al_xGa_{1-x}As interface without modulation doping. Instead, a highly doped layer was defined 100 nm away from the heterointerface. In this structure, a 2DEG can be generated at the heterointerface by applying positive voltages between the top gate and the doping layer. The electrons reach the interface by tunneling across the un-

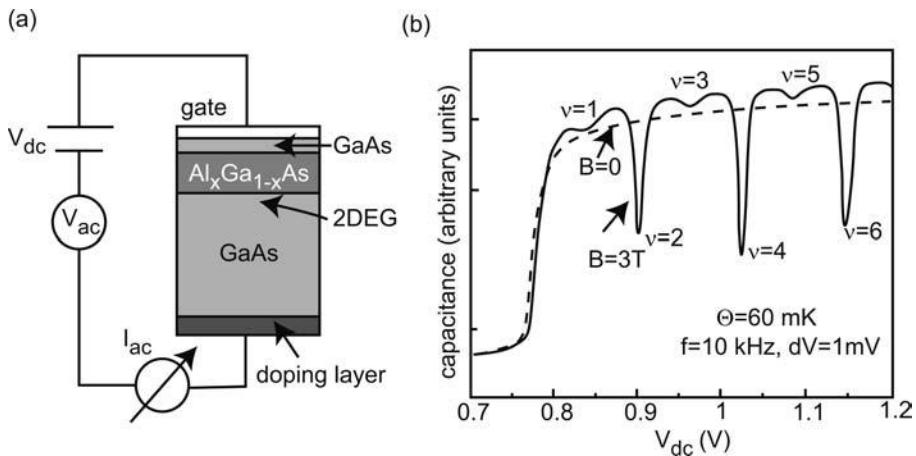


Fig. 6.4 (a) Schematic setup for measuring the capacitance of an electron gas. The quantum well is empty at $V_g = 0$ and can be filled with electrons by applying a positive voltage between the gate and the doping layer. (b) The measured capacitance shows the filling of the 2DEG at $V_g = 0.77\text{ V}$, as well as the modulated density of states in perpendicular magnetic fields. Adapted from [78].

doped GaAs spacer layer. In addition, a small AC signal is superimposed, and the current at a phase difference of $\pi/2$ is measured with a lock-in amplifier.

In order to deposit charge on the capacitor formed by the 2DEG and the gate, the voltage source has to do both electrostatic and chemical work on the system. The density of states in the metal electrode is very large. Hence, its chemical potential will remain constant. The density of states of the 2DEG, however, is much lower. As charge is added, the Fermi level in the electron gas will therefore change significantly, namely by an amount that depends on the density of states, as well as on the total charge added. Hence, the voltage can be split into two parts, V_{chem} and V_{elstat} , that perform the chemical and the electrostatic work, respectively. Changing the charge on the capacitor by dq thus requires

$$dV_{\text{elstat}} = \frac{1}{C} dq$$

plus

$$dV_{\text{chem}} = \frac{1}{e} d\mu = \frac{1}{e} \frac{d\mu}{dn} dn = \frac{1}{e^2} \frac{d\mu}{dn} dq = \frac{1}{e^2 D(E)} dq$$

Here, dn is the electron density change in the 2DEG, and μ denotes its chemical potential. Thus, the total voltage change equals

$$dV = dV_{\text{elstat}} + dV_{\text{chem}} = \left[\frac{1}{C} + \frac{1}{e^2 D(E)} \right] dq$$

It becomes apparent that the effective capacitance is the geometric capacitance in series with a “chemical capacitance”, i.e.

$$\frac{1}{C_{\text{eff}}} = \frac{1}{C} + \frac{1}{e^2 D(E)} \quad (6.13)$$

and we can determine the density of states by capacitance measurements. This explains the observations in Fig. 6.4(b). In the absence of magnetic fields, the capacitance essentially experiences a jump as the gate voltage fills the potential well at the heterointerface. Since $D(E)$ of a 2DEG is constant, no further structure is observed. This changes as a magnetic field is applied, due to the formation of Landau levels. Once the geometrical capacitance is known, it is straightforward to extract the density of states.

Capacitance spectroscopy is also possible by applying the voltage directly between the top gate and a 2DEG. This method is inferior in high magnetic fields, though. For a discussion of this issue, see Paper P6.1.

6.2

The quantum Hall effect

6.2.1

Phenomenology

Back in 1980, K. von Klitzing and coworkers [176] discovered a quantization of the Hall resistance in 2DEGs residing in Si MOSFETs. Examples are shown in Fig. 1.3, as well as in Fig. 6.5. Plateaux were observed at integer filling factors, i.e. for

$$\rho_{xy}(B) = \frac{1}{\nu} \frac{h}{e^2}$$

Subsequent experiments showed that this quantization is *universal*, in the sense that it is observed in all kinds of materials, provided the electron gas is two-dimensional. The accuracy $\delta\rho_{xy}/\rho_{xy}$ can be of the order of 3×10^{-10} (Fig. 6.6), such that the $\nu = 1$ quantum Hall plateau has been chosen as the resistance standard, with a resistance of $R_Q = 25\,812.807\,\Omega$ by definition [247].

Furthermore, another quantization of ρ_{xy} has been discovered [305]. In extremely high-quality samples, additional resistance plateaus are observed at

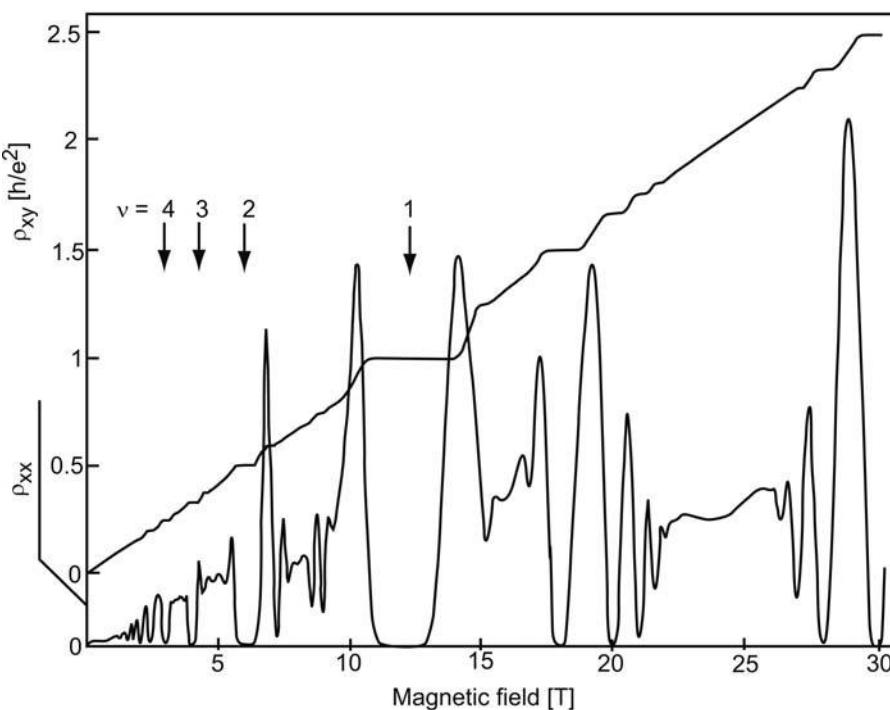


Fig. 6.5 The quantum Hall effect and Shubnikov–de Haas oscillations in a Ga[Al]As HEMT, measured in a dilution refrigerator at a temperature of 100 mK. A filling factor of $v = 1$ is reached at $B \approx 12$ T. Besides the integer quantum Hall effect, pronounced structures at fractional filling factors are observed. After [330].

$\rho_{xy} = h/ke^2$, with k being a rational number. This type of quantization is very pronounced in Fig. 6.5, but its discussion is beyond our scope. Here, we restrict ourselves to the following remark. Consider the data of Fig. 6.5 around $v = 1/2$, i.e. $B \approx 24$ T. The data resemble the integer quantum Hall effect (QHE) around $B = 0$. This observation can be substantiated by theory. One picture of the fractional QHE is that so-called composite fermions, which are quasi-particles composed of one electron and two magnetic flux quanta, form in strong magnetic fields. These quasi-particles (with an effective mass that differs from m^*) then undergo Landau quantization in an effective magnetic field which remains after the flux quanta used to form the composite fermions are subtracted. For further information on this *fractional quantum Hall effect*, the reader is referred to the specialized literature cited at the end of this chapter.

Interestingly, the accurate determination of e^2/h is highly relevant for quantum electrodynamics, since this ratio is contained in the fine structure constant, $\alpha = \frac{1}{2}\mu_0 c(e^2/h)$, which describes the coupling of elementary particles

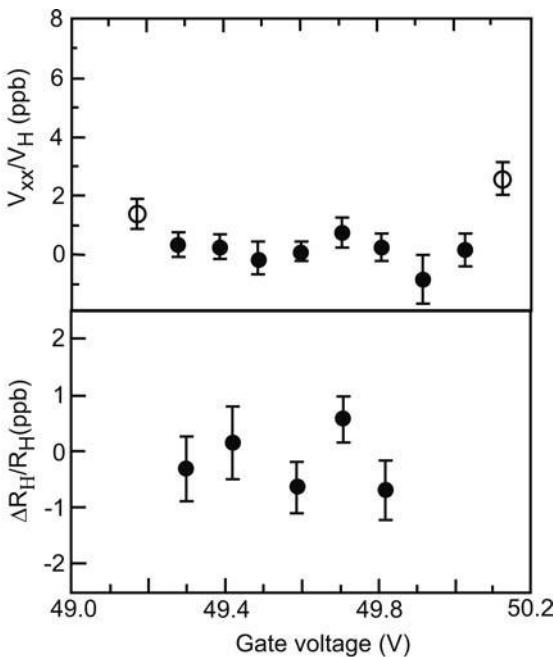


Fig. 6.6 The relative accuracy of ρ_{xy} in the $\nu = 1$ quantum Hall plateau, and the corresponding longitudinal voltage drop, measured as a function of the electron density, which is changed via a gate voltage. After [167]. For the accuracy $\Delta R_H/R_H$, a value of 3×10^{-10} is obtained.

to electromagnetic fields, and as such represents the expansion parameter in this theory.

The QHE is closely connected to another highly remarkable effect, namely the magneto-oscillation of the longitudinal resistivity ρ_{xx} (see Fig. 6.5). These oscillations are known as *Shubnikov-de Haas oscillations*. In fact, in the regions of quantized Hall resistance, ρ_{xx} becomes zero, with an accuracy comparable to that of ρ_{xy} (Fig. 6.6). The striking correlation between ρ_{xx} and ρ_{xy} suggests a common explanation. The relation between these two quantities will become clear in Chapter 7.

6.2.2

Toward an explanation of the integer quantum Hall effect

Besides the Landau quantization, disorder is an essential ingredient in understanding the origin of the quantum Hall effect. Suppose we have adjusted the 2DEG to one of the insulating points, characterized by

$$B_j = \frac{hn}{2e} \frac{1}{j}$$

The classical Hall resistivity at such a point is given by

$$\rho_{xy,j} = -\frac{B_j}{en} = -\frac{\hbar}{2e^2} \frac{1}{j} \quad (6.14)$$

which are the resistances of the observed plateaus.⁴ Clearly, in an insulating regime, charge cannot flow inside the sample, and the differential Hall conductance should be zero. Therefore, we expect the Hall conductance to remain constant and $\sigma_{xx} = 0$ (and thus $\rho_{xx} = 0$).

But in our model system, the insulating interval is only a point. It is here that the “real” properties of the sample enter the explanation. There is disorder in the system, which broadens the δ functions in the density of states. With the help of Fig. 6.7, we argue that the states in the wings of the peaks of $D(E)$ tend to be localized. Consider a cross section of the potential landscape in the y -direction. As long as the magnetic length is small compared to the length scale of the potential fluctuations, the energies of the Landau levels will just follow the potential fluctuations. We study the first Landau level as an example. The same line of arguing holds for higher Landau levels. Clearly, the states inside one Landau level now have different energies, and the shape and width of the peak in the density of states reflect the energy distribution of the disorder potential. Consider states in the low-energy wing of the peak in $D(E)$, i.e. at the energy E_1 in Fig. 6.7. Apparently, the electrons reside in minima of the potential landscape. Close to the center of such a puddle, where the local electric field is weak or zero, the electrons move in cyclotron orbits. At the puddle edge, though, they skip along the potential wall during their cyclotron motion.

In order to see this, it pays to revisit the $\vec{E} \times \vec{B}$ drift, i.e. the motion of charged particles under crossed electric and magnetic fields. In the equation of motion for electrons in the diffusive regime (Eq. (2.46)), it was assumed that $\omega_c \tau \ll 1$, such that the magnetic field deflects the electrons that move through the sample with the drift velocity, which is determined by the electric field, on the one hand, and by the friction term, on the other. Now, however, the magnetic field is much stronger, and the electrons are able to complete cyclotron circles without being scattered. Consider the motion on a time scale between $1/\omega_c$ and τ . There is no diffusive scattering, and the classical equation of motion now reads

$$m^* \frac{d^2 \vec{r}}{dt^2} = -e \left(\vec{E} + \frac{d\vec{r}}{dt} \times \vec{B} \right) \quad (6.15)$$

4) Recall that, in our discussion of the Landau quantization, we have assumed spin degeneracy. For spin-split Landau levels, we obtain

$$\rho_{H,v} = -\frac{\hbar}{e^2} \frac{1}{v}.$$

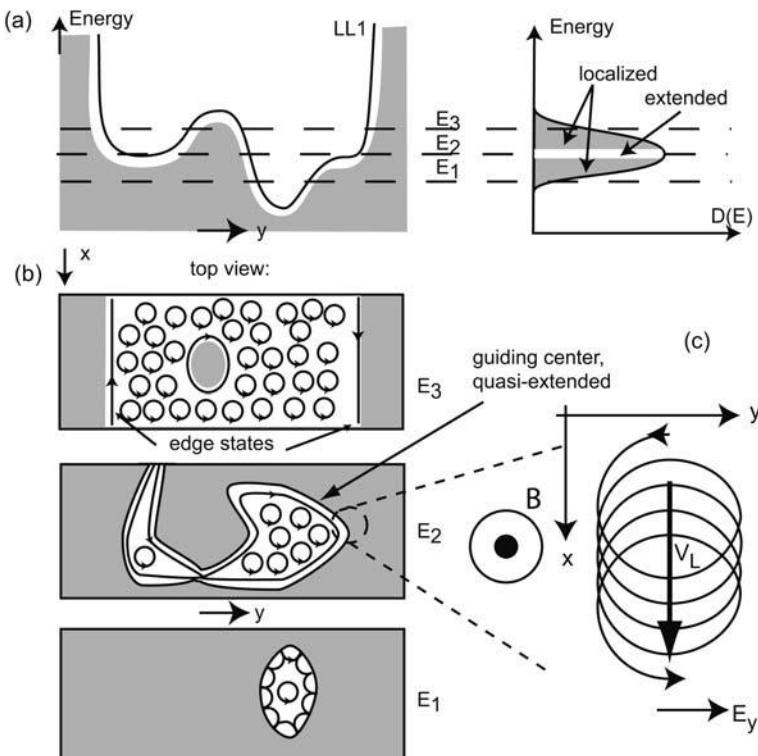


Fig. 6.7 (a) Cross section of the potential landscape in the y -direction across a Hall bar. The first Landau level (LL 1) follows the energy of the conduction band bottom. The resulting peak in the density of states is shown to the right. (b) Top views of the samples sketched at energies E_i , as indicated in (a). Gray areas denote regions where the energy of the first Landau level is larger than E_i . At energy E_3 , the states are localized at potential maxima. At low energy E_1 , the electrons are caught in potential minima. At the intermediate energy E_2 , however, the elec-

tron puddles merge, and electrons may travel across the whole sample via skipping orbits at the edges of the Fermi sea. This leads to extended states around the center of the peak in $D(E)$. (c) A close-up of the electronic motion in the presence of an electric field in the y -direction. The cyclotron motion is superimposed on the motion of the guiding center, which moves perpendicular to both fields at constant velocity. Note for the orientation of the electronic motion that we are looking at the sample in the $-z$ -direction, opposite to the direction of the magnetic field.

where \vec{r} is the position of the electron in the (x, y) plane. Suppose the electric field points in the y -direction, $\vec{E} = (0, E_y, 0)$, and the magnetic field, as usual, points in the z -direction. Then Eq. (6.15) is solved by

$$\begin{aligned} x(t) &= \frac{E_y}{\omega_c B} \sin(\omega_c t) + \frac{E_y}{B} t + X_0 \\ y(t) &= -\frac{E_y}{\omega_c B} \cos(\omega_c t) + Y_0 \end{aligned} \quad (6.16)$$

It is common to separate the motion into the motion of the *guiding center*

$$(X(t), Y(t)) = (X_0 - v_L t, Y_0) \quad (6.17)$$

and the motion relative to the guiding center

$$(x_r(t), y_r(t)) = (r_c \sin(\omega_c t), -r_c \cos(\omega_c t)) \quad (6.18)$$

Here, v_L denotes the drift velocity due to the Lorentz force, and $r_c = E_y/\omega_c B = v_L/\omega_c$ is the cyclotron radius. We thus see that the electrons perform a cyclotron motion around the guiding center, which moves at constant velocity in the direction perpendicular to both fields, as sketched in Fig. 6.7(c). It is straightforward to write down the components of the conductivity tensor for such a situation. Here, we average over the (fast) cyclotron motion, such that only the motion of the guiding center gives a contribution. Since

$$j_x = -ne(dx/dt) = \sigma_{xx}E_x + \sigma_{xy}E_y$$

and

$$j_y = \sigma_{yx}E_x + \sigma_{yy}E_y$$

it follows that

$$\sigma_{xy} = \frac{-ne}{B} \quad \text{and} \quad \sigma_{yy} = 0 \quad (6.19)$$

Likewise, we obtain the remaining two components by considering an electric field in the x -direction. The corresponding components of the resistivity tensor are $\rho_{xx} = \rho_{yy} = 0$ and $\rho_{xy} = -\rho_{yx} = -B/(ne)$. It may seem strange that both the conductivity and the resistivity can be zero at the same time. Remember, however, that we are dealing with a resistivity and conductivity *tensor*. The observation $\sigma_{xx} = 0$ implies that no current flows in the x -direction when a voltage is applied in the x -direction. On the other hand, $\rho_{xx} = 0$ means that applying a current in the x -direction causes no voltage drop in the x -direction. This is no contradiction, provided the equipotential lines are parallel to the x -direction if a voltage or current is applied along this axis (Fig. 6.8). It has been verified experimentally that this is in fact the case inside a quantum Hall plateau [177].

Suppose now that we change the electron density of a perfectly clean system and apply a negligibly small voltage in the x -direction. The only electric field of relevance is the Hall field in the y -direction. The electrons move parallel to the x -direction, as long as we are in a metallic state. As we fill one initially empty Landau level, i.e. $\delta n = D_{LL}$, the Hall conductance changes by $\delta\sigma_{xy} = -2e^2/h$. Nevertheless, $\sigma_{xy}(n)$ is a straight line, since the insulating regions are just points on the n -axis. Also, σ_{xx} is zero for all electron densities.

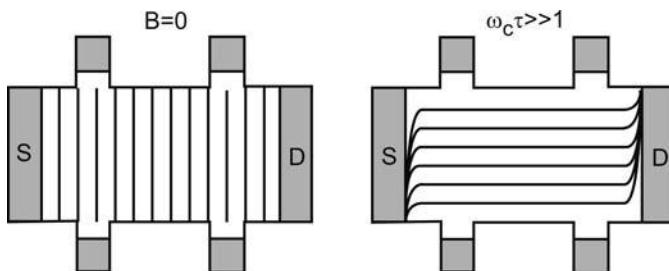


Fig. 6.8 Equipotential lines in a 2DEG at $B = 0$ (left) and in a strong magnetic field (right).

The electrons in the low-energy tail of a peak in $D(E)$, however, are localized. They either perform cyclotron orbits, or they circle along the edge of the puddle, in the direction of rotation opposite to that of the cyclotron motion. Such orbits are called *skipping orbits*, which we will revisit in Chapter 7. In any case, none of the electrons in such a puddle can carry a current across the sample. Thus, we should modify our definition of a metal accordingly and speak of a metal only if the density of *extended* states at the Fermi energy is non-vanishing. In such a situation, we can fill the Landau level while the system is in an insulating state, which means that $\delta\sigma_{xy}/\delta n = 0$.

A somewhat different kind of localization occurs in the high-energy wing of the peak in $D(E)$. Consider the situation at energy E_3 . The states at this energy correspond to skipping orbits that circle around maxima in the potential landscape. Hence, we conclude that, also for Fermi energies in the high-energy tail of the density of states peak, the system behaves in an insulating manner.

At an intermediate energy E_2 , the electron puddles and the potential hills have about equal sizes on average, and the localized skipping orbits of adjacent structures will merge. At this energy, the electron may *percolate* across the whole sample, and consequently the sample behaves in a metallic way. Under these conditions, σ_{xy} can change and σ_{xx} becomes non-zero, since the electrons may traverse the sample from source to drain. That such a state does in fact exist is shown more rigorously in the specialized literature (see the further reading at the end of this chapter). Here, we just state that energy E_2 represents a percolation threshold [288], and that the electron gas undergoes a percolation transition as the Fermi energy is swept across energy E_2 .

Thus, the disorder does something truly remarkable: it increases the insulating regions of the parameter range (the parameter is the electron density or the magnetic field) from points (Fig. 6.3) to extended intervals in real samples, while the extended metallic regions of the ideal sample are reduced to very small intervals. This allows the observation of the conductance steps.

Question 6.3: Sketch $\sigma_{xy}(n)$ and $\sigma_{xx}(n)$ for an ideal sample and for a real sample with disorder. Here, n is the two-dimensional electron density.

We will revisit the QHE in Section 7.3, and provide an alternative view, based on the observation that the current in the QHE regime is carried via effectively one-dimensional states. A more fundamental explanation of the quantum Hall effect has been provided in [187]. It derives the exact quantization from (i) the gauge invariance and (ii) the existence of a mobility gap around the Fermi energy. A different approach by Thouless [301] derives the exact quantization of the Hall resistance by treating the random potential within perturbation theory. These models are beyond our scope, however. We will instead discuss another point of view in the following chapter, which is based on transport in quantum wires that effectively form at the sample edge [137] as indicated in Fig. 6.7(b).

6.2.3

The quantum Hall effect and three dimensions

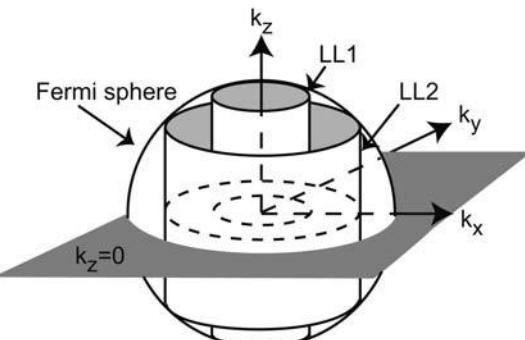
By now, you may be wondering whether the quantum Hall effect occurs only in quantum films. In fact, it vanishes in three-dimensional free electron gases. The reason is illustrated in Fig. 6.9. A magnetic field pointing in the z -direction quantizes the motion in the (x, y) plane. The motion in the z -direction, however, remains unaffected. The Fermi sphere “condenses” into cylinders of radii

$$k_{xy} = \sqrt{k_x^2 + k_y^2} = \sqrt{\frac{2m^*\omega_c(j - \frac{1}{2})}{\hbar}}$$

which extend along the z -direction. The Landau levels have evolved into Landau bands with a one-dimensional density of states. Thus, no matter what magnetic field we apply or how large the Fermi energy is, there are always states at the Fermi energy, there are no insulating regions in the parameter space, and consequently there are no metal–insulator transitions. Confining the electron gas in the z -direction corresponds to a Fermi circle parallel to the (x, y) plane, the cylinders are projected on circles, and parametric metal–insulator transitions are again possible. Note that this line of arguing is based on a *free* electron gas in the z -direction. The density of states in this direction has been modified by Stormer and coworkers [292], who grew a periodic sequence of GaAs quantum wells in the z -direction, separated by $\text{Al}_{0.18}\text{Ga}_{0.82}\text{As}$ barriers. This periodic superlattice generated bands of width b , with bandgaps in the meV regime, i.e. comparable to $\hbar\omega_c$ for moderate magnetic fields. The corresponding density of states, shown in Fig. 6.9(b), thus develops gaps in sufficiently large magnetic fields, such that the quantum Hall effect should be visible as soon as b gets smaller than $\hbar\omega_c$. This has been experimentally ver-

ified in [292]. Hence, although the electron gas does not need to be strictly two-dimensional, a sufficiently large anisotropy is a prerequisite for the quantum Hall effect to occur.

(a)



(b)

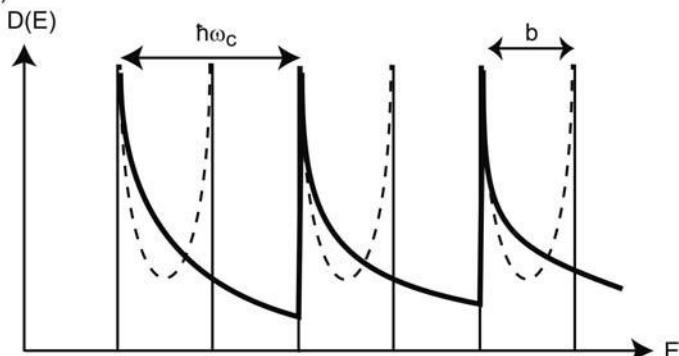


Fig. 6.9 Landau quantization in three-dimensional systems, drawn for two occupied Landau bands. (a) Under sufficiently large magnetic fields, the Fermi sphere of a free electron gas at $B = 0$ condenses into Landau levels in the (x, y) plane, while k_z remains continuous. (b) The density of states for a free electron gas (bold line), and of a periodic superlattice in the z -direction (dashed lines).

6.3

Elementary analysis of Shubnikov–de Haas oscillations

We now turn our attention to the Shubnikov–de Haas (SdH) oscillations at small and intermediate magnetic fields, i.e. for $\omega_c\tau < 1$. Here, the quantum Hall effect is weak, and $\rho_{xx}(B)$ oscillates, but does not vanish (see Fig. 1.3). In this regime, the magnetic field causes just a weak modulation of the density of states. As a consequence, the density of states at the Fermi level, as well as the screening properties of the electron gas, oscillate as the electron density or

the magnetic field is tuned. This is reflected in the longitudinal resistivity. The scattering theory of such a system is hampered by several difficulties, which are discussed in the papers of Ando [8]. For short-range scattering potentials that are weak compared to the Fermi energy, however, Ando has derived an analytic expression for $\rho_{xx}(B)$, which in the limit $\omega_c\tau \ll 1$ reads

$$\rho_{xx}(B) = \rho_{xx}(0) \left[1 - 4 \cos \left(\frac{\pi \hbar n}{eB} \right) D(m^*, \Theta) E(m^*, \tau_q) \right] \quad (6.20)$$

which we call the Ando formula. The resistivity according to this formula is illustrated in Fig. 6.10. The expression $D(m^*, \Theta)$ is known as the Dingle term. It contains the temperature dependence and depends on the effective mass:

$$D(m^*, \Theta) = \frac{x}{\sinh x}, \quad x = \frac{2\pi^2 k_B}{\hbar e B} m^* \Theta \quad (6.21)$$

The exponential term $E(m^*, \tau_q)$ in Eq. (6.20) equals

$$E(m^*, \tau_q) = e^{-\pi/(\omega_c \tau_q)} \quad (6.22)$$

and depends on the quantum scattering time.

Therefore, both the effective mass as well as τ_q can be extracted from measuring the temperature dependence of the SdH oscillations. We pick a single, suitable SdH resonance. Its amplitude is given by

$$A = 8\rho_{xx}(0)D(m^*, \Theta)E(m^*, \tau_q) \quad (6.23)$$

This can be rewritten as

$$\ln \left(\frac{A}{\Theta} \right) = C_1 - \ln \left[\sinh \left(\frac{2\pi^2 k_B}{\hbar e B} m^* \Theta \right) \right] \quad (6.24)$$

Here, C_1 denotes a constant, which has no further interest for our purposes. For sufficiently small magnetic fields and sufficiently high temperatures (this is what we mean by "a suitable SdH resonance"), $\ln(\sinh x) \approx x$. Thus, by plotting $\ln(A/\Theta)$ vs. Θ , a straight line with a slope of

$$\frac{2\pi^2 k_B}{e B \hbar} m^* \quad (6.25)$$

is obtained. Besides analyzing cyclotron resonances, this is a common way to determine effective masses. Once we know m^* , we can exploit the exponential term and determine τ_q along similar lines. This time, a measurement of $\rho_{xx}(B)$ at fixed temperature, which extends over many SdH oscillations, is analyzed. We rewrite the expression for the oscillation amplitude as

$$AB \sinh \left(\frac{2\pi^2 k_B \Theta m^*}{e B \hbar} \right) = 8\rho_{xx}(0) \frac{2\pi^2 k_B \Theta m^*}{e \hbar} e^{-\pi/(\omega_c \tau_q)} \implies Y = \ln \left[AB \sinh \left(\frac{2\pi^2 k_B \Theta m^*}{e B \hbar} \right) \right] = C_2 - \frac{\pi m^*}{e} \frac{1}{\tau_q} \frac{1}{B} \quad (6.26)$$

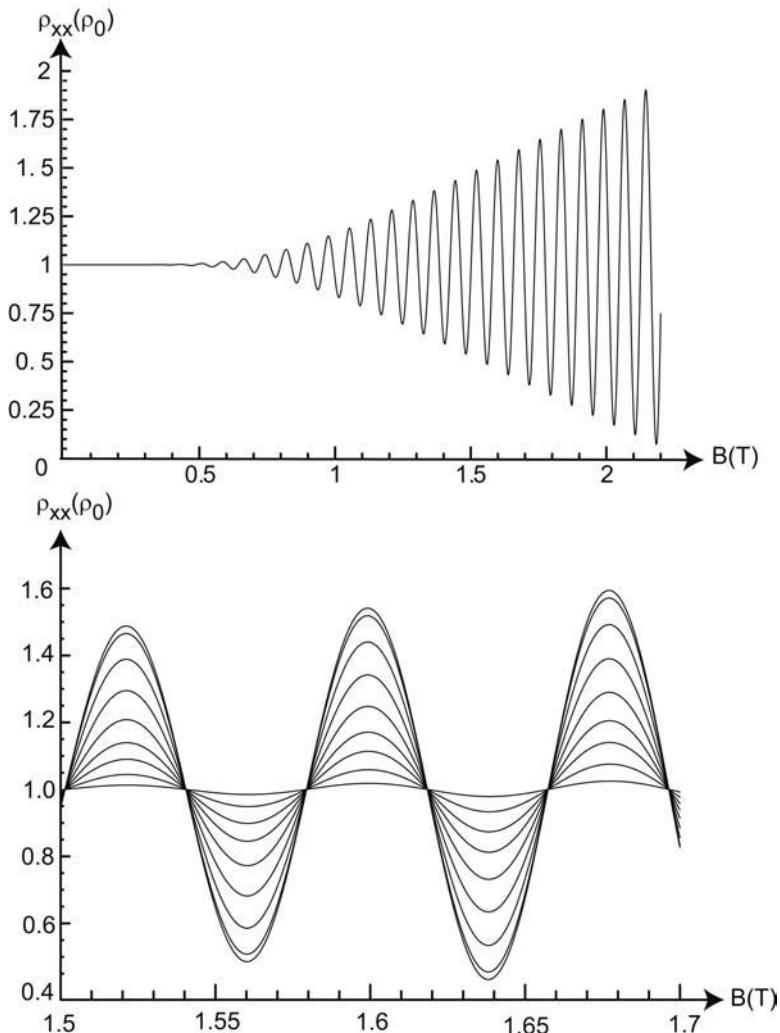


Fig. 6.10 Top: SdH oscillations of a 2DEG as a function of B , as described by the Ando formula (Eq. (6.20)). Bottom: Temperature dependence of some oscillations. The temperatures are $\Theta = 1, 2, 4, 6, 8, 10, 12, 15$, and 20 K . See also Exercise E6.2.

and plot Y as a function of $1/B$ (known as a *Dingle plot*). The slope of this straight line equals

$$\frac{\pi m^*}{e\tau_q} \quad (6.27)$$

Fig. 6.11 shows the results of such an analysis performed on a GaN–Al_xGa_{1-x}N HEMT structure.

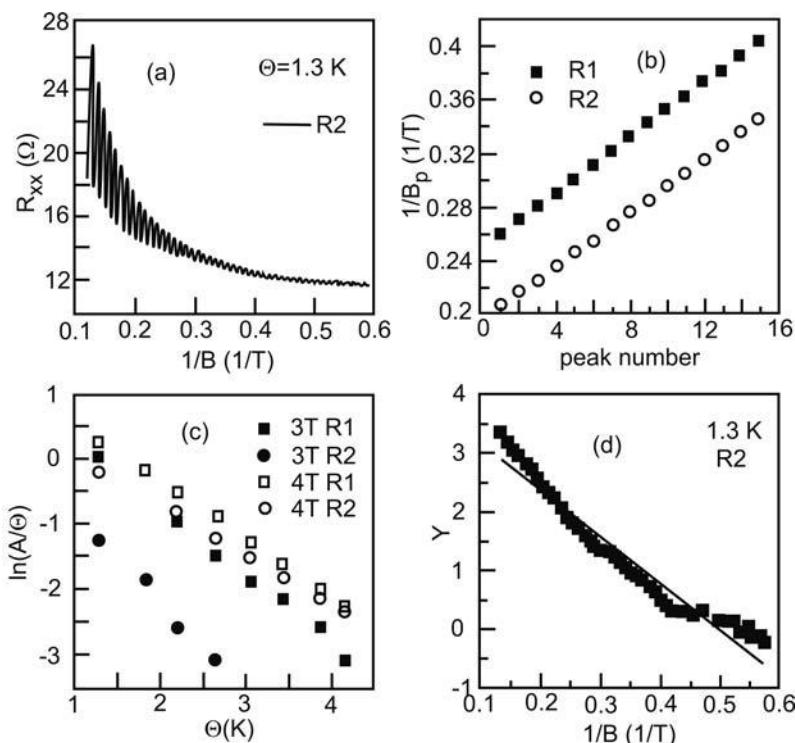


Fig. 6.11 Standard analysis of SdH oscillations (a), performed on two different samples (R1 and R2) of a 2DEG at a modulation-doped $\text{GaN}-\text{Al}_x\text{Ga}_{1-x}\text{N}$ interface. The electron density was $n = 4.8 \times 10^{16} \text{ m}^{-2}$ (b). A quantum scattering time of $\tau_q = 0.5 \text{ ps}$ (c) and an effective mass of $m^* = (0.215 \pm 0.006)m_e$ (Dingle plot, (d)) have been found. After [262].

6.4

Some examples of magneto-transport experiments

There are many further interesting magneto-transport experiments on quantum films, and we provide a few examples.⁵

6.4.1

Quasi-two-dimensional electron gases

For Fermi energies above the second quantized energy level of the confining potential in the z -direction, the electron gas is no longer strictly two-dimensional, and we speak of a *quasi-2DEG*. While the Hall slope measures

5) We will see another very important example in Chapter 8, namely how electronic phase coherence manifests itself in the magneto-resistivity.

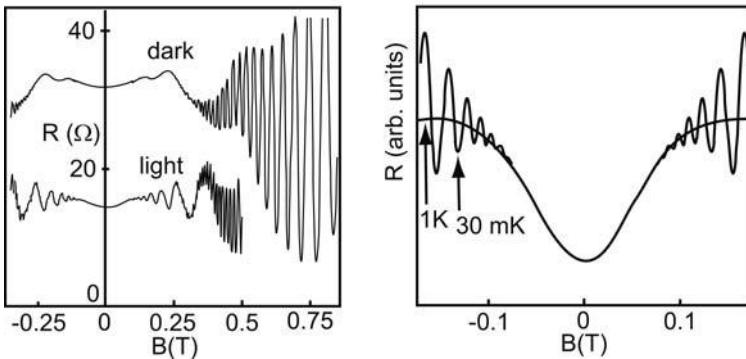


Fig. 6.12 Longitudinal magneto-resistivity of a 2DEG in a GaAs–Al_xGa_{1–x}As HEMT with two occupied subbands (left). Here, “light” indicates that the sample has been illuminated by light of a frequency below the GaAs bandgap. This ionizes residual neutral donors and thus increases the electron density in the

2DEG. The magneto-oscillations are modulated in both cases. The two SdH frequencies correspond to the partial electron densities in the two subbands. In addition, a positive magneto-resistivity around $B = 0$ is observed (right). After [156].

the total electron density, each two-dimensional subband causes a Shubnikov-de Haas oscillation, provided the scattering times are sufficiently large. This results in a modulation of SdH oscillations (Fig. 6.12). The scattering times in the upper subband, however, can be small, such that the corresponding SdH oscillation may not be observable. In this case, their occupation can be detected as the difference between the total electron density (as obtained from the Hall slope) and the density of the lowest subband (determined from the SdH oscillation period of the lower subband).

A further signature of multi-subband occupation is a parabolic and positive magneto-resistivity around $B = 0$. The subbands can be regarded as resistors in parallel, such that the total conductivity tensor is obtained by simple addition of the individual subband conductivity tensors. The two subbands are characterized by different scattering times τ_1 and τ_2 . We further assume that the effective masses in both subbands are identical, and that inter-subband scattering can be neglected, i.e. the scattered electrons remain in their original subband during scattering events.⁶ The total magneto-conductivity tensor is then given by

$$(\underline{\sigma}) = \frac{n_1 e^2 \tau_1}{1 + (\omega_c \tau_1)^2} \begin{pmatrix} 1 & -\omega_c \tau_1 \\ \omega_c \tau_1 & 1 \end{pmatrix} + \frac{n_2 e^2 \tau_2}{1 + (\omega_c \tau_2)^2} \begin{pmatrix} 1 & -\omega_c \tau_2 \\ \omega_c \tau_2 & 1 \end{pmatrix}$$

6) Inter-subband scattering can be included in the analysis, see [344].

The longitudinal resistivity ρ_{xx} is obtained by matrix inversion as discussed in Chapter 2. For small magnetic fields, a Taylor expansion gives the expression

$$\rho_{xx}(B) = \rho_{xx}(0) \left[1 + \frac{n_1 \tau_1 n_2 \tau_2 (\tau_1 - \tau_2)^2 e^2}{m^*{}^2 (n_1 \tau_1 + n_2 \tau_2)^2} B^2 \right] \quad (6.28)$$

Typical experimental results show such a behavior, as illustrated in Fig. 6.12.

Question 6.4: Derive Eq. (6.28), and determine the conditions for which the Hall slope measures the total electron density $n = n_1 + n_2$ of both subbands.

6.4.2

Mapping of the probability density

Perturbation theory tells us that inserting a δ function $U_0 \delta(z - z_0)$ in a potential generates, to first order, energy shifts ΔE_i of the energy eigenvalues E_i , where ΔE_i is proportional to the probability density of the corresponding eigenstate at z_0 :

$$\Delta E_i = U_0 |\psi(z)|^2 \quad (6.29)$$

With U_0 known, $|\psi(z)|^2$ can therefore be determined by measuring ΔE_i (see [204]). If we could scan the δ function across the potential, the probability density could be mapped this way. This idea can in fact be realized experimentally in parabolic quantum wells, where the conduction band bottom varies parabolically in the growth direction. It is a unique property of a parabolic potential that superposition of a constant electric field displaces the potential without changing its shape – see Exercise E6.3. In the sample depicted in Fig. 6.13, the quantum well contains a $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ spike at its as-grown center, while the parabola can be shifted by applying a voltage applied to the two electrodes haloing the parabola [259]. In the experiment shown in Fig. 6.13, two subbands were occupied, and, by measuring the two subband densities via SdH oscillations, the energy shifts have been detected. Hence, $|\psi_2(z)|^2 - |\psi_1(z)|^2$ has been measured (see Figs. 6.13(c) and (d)).

6.4.3

Displacement of the quantum Hall plateaux

According to our model in Section 6.2, the quantum Hall plateaus are centered around integer filling factors. In other words, if we extrapolate the classical Hall slope into the quantum Hall regime, it should intersect the plateaus at their center. Here, we have implicitly assumed that the peaks of the density of states are symmetrical, which is not necessarily the case. Their shape depends on the character of the scatters. For example, predominantly repulsive

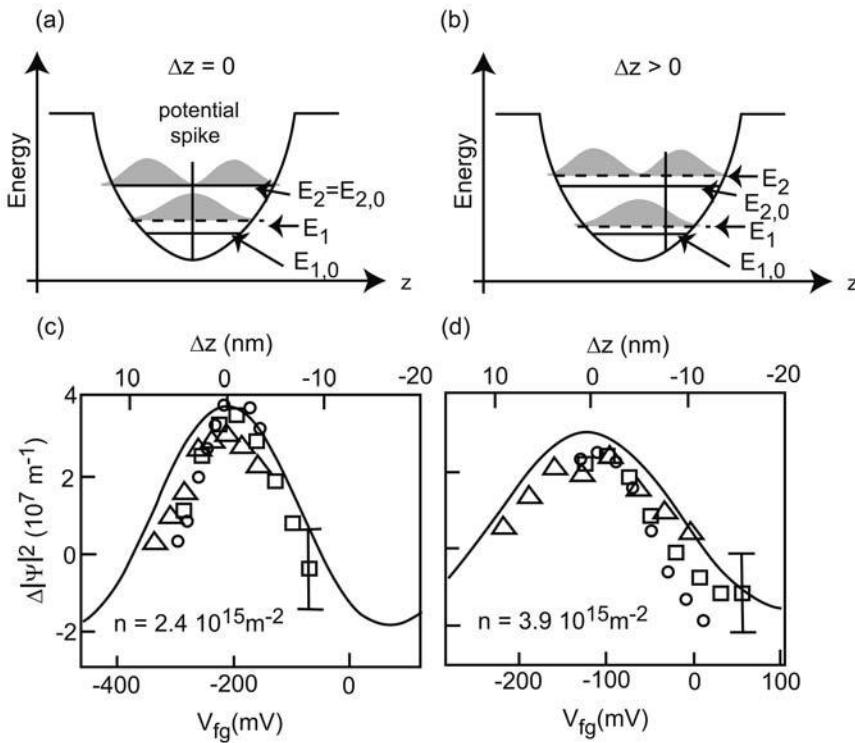


Fig. 6.13 (a) Conduction band of a parabolic Ga[Al]As quantum well. A potential spike has been grown at the center of the parabola, which shifts the energy levels as compared to the same potential without the spike (dashed lines). The shift is proportional to $|\Psi^2|$ at the position of the spike. (b) By applying a constant electric field in the z -direction, the confining parabola is displaced with respect

to the spike, and the energy levels shift accordingly. (c, d) Measured differences of the probability density between subbands 1 and 2 as a function of z for two different electron densities. The different symbols denote different spike heights, i.e. an Al concentration of $x = 0.05, 0.1$, and 0.15 , respectively. Adapted from [259].

scattereders shift the center of gravity of a Landau level toward higher energies, as shown schematically in Fig. 6.14(a). This asymmetry affects the position of the quantum Hall plateaus. Suppose the scatterers are predominantly repulsive, we start out from LL j completely filled, and we increase the magnetic field. As long as we deplete localized states, no changes in the resistivities are observed. As can be seen from Fig. 6.14(c), this means that the delocalized states reach the Fermi level at larger magnetic fields as compared to the symmetric situation, and the jump in ρ_{xy} , as well as the peak in ρ_{xx} , are shifted correspondingly with respect to the classical Hall slope. This effect has been studied systematically in [144] (Fig. 6.15) and is occasionally used to obtain further information on the scatterers.

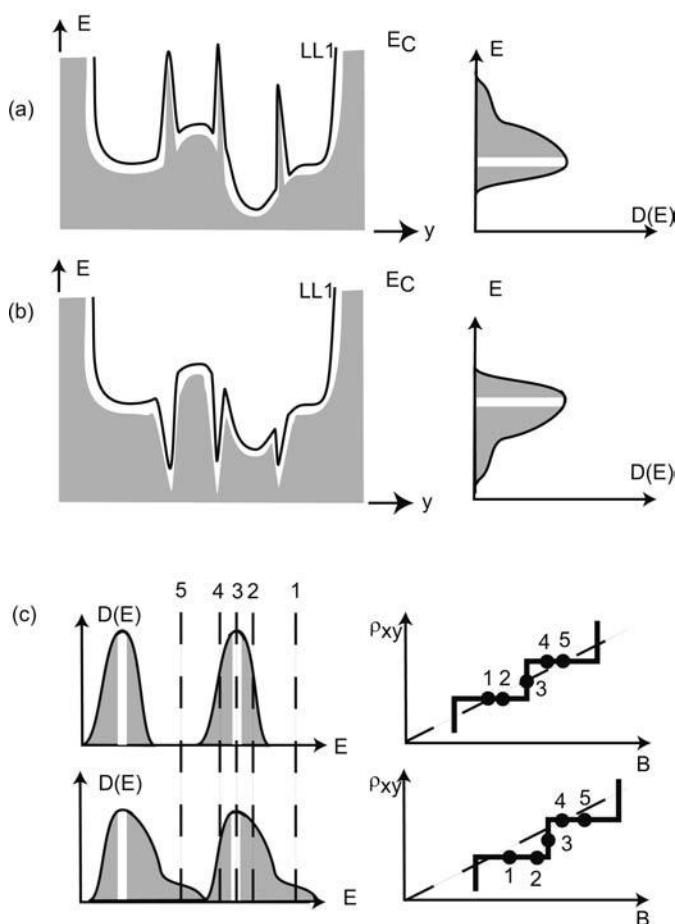


Fig. 6.14 Repulsive scatterers shift a fraction of the states within a peak of the density of states to higher energies, which results in a shift of the quantum Hall plateaus to larger magnetic fields. Likewise, predominantly attractive scatterers shift the quantum Hall plateaus to smaller magnetic fields.

6.5 Parallel magnetic fields

In comparison to the quantum Hall effect, a magnetic field in the plane of the 2DEG produces much less spectacular results. Nevertheless, investigating the transport properties as a function of a parallel magnetic field B_{\parallel} is a useful tool. The density of states can be tuned, and spin effects can be investigated without being buried in the dominating orbital effects. Here, we discuss how B_{\parallel} affects the density of states and discuss the consequences.

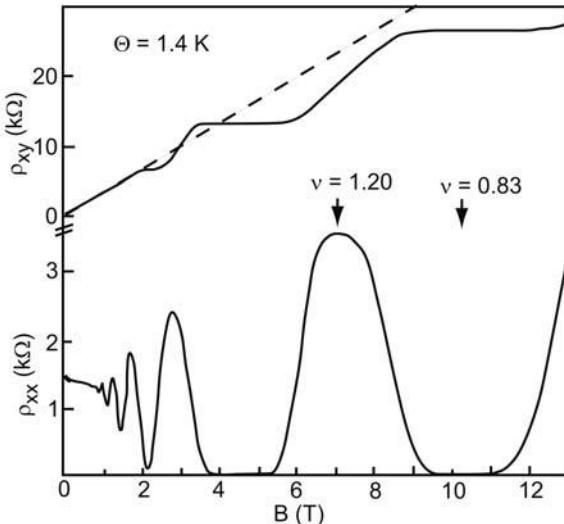


Fig. 6.15 Quantum Hall plateaus can be shifted with respect to the extrapolated classical Hall trace (dashed line). Adapted from [144].

A 2DEG in a homogeneous magnetic field of arbitrary orientation is a very complicated problem in its most general form. For example, the classical dynamics of a square well in tilted magnetic fields is chaotic [105], which leads to the corresponding quantum mechanical signatures. The evolution of the energy levels as a function of a tilted magnetic field can be studied using perturbation theory [34, 260]. An analytical solution is possible for a parabolic confinement [198]. Besides tuning the spin splitting, B_{\parallel} has two effects. First of all, the parabolic confinement generated by B_{\parallel} adds to the electrostatic confinement, such that the subbands shift to higher energies. This is sometimes referred to as *diamagnetic shift*. Second, the effective mass in the direction perpendicular to B_{\parallel} , but in the plane of the electron gas, increases. We have seen this behavior already in its extreme version in the QHE, where the effective mass goes to infinity.

Here, we restrict ourselves to a simple case, which reveals these properties analytically. Other potential shapes show a similar qualitative behavior. We consider a parabolic quantum well with an electrostatic potential in the growth direction, given by

$$V(z) = \frac{1}{2}m^*\omega_0^2 z^2 \quad (6.30)$$

The magnetic field is applied in the x -direction, and we choose the gauge $\vec{A} = (0, -zB_{\parallel}, 0)$, which gives $\vec{B} = \vec{\nabla} \times \vec{A} = (B_{\parallel}, 0, 0)$. The Schrödinger equation now reads

$$\frac{1}{2m^*}[p_x^2 + (p_y - eB_{\parallel}z)^2 + p_z^2]\Psi(\vec{r}) + \frac{1}{2}m^*\omega_0^2 z^2\Psi(\vec{r}) = E\Psi(\vec{r}) \quad (6.31)$$

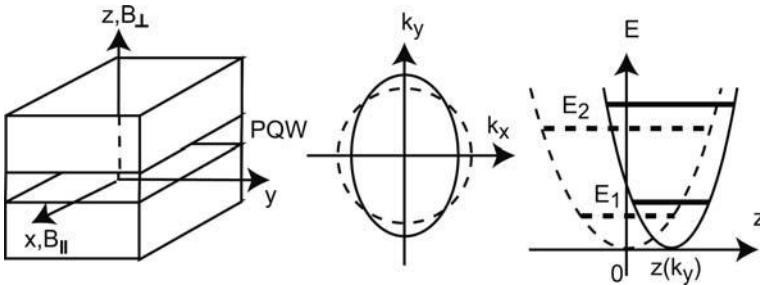


Fig. 6.16 Left: Definition of the coordinate system and the magnetic field directions for a parabolic quantum well in perpendicular and parallel magnetic fields. Center: Corresponding shape of the two-dimensional Fermi sphere in a magnetic field applied in the x -direction (full line), in comparison to the Fermi sphere for $B = 0$ (dashed line). Right: The

confining potential in the z -direction for $B = 0$ (dashed line) shifts and narrows for a state with wave number k_y when a parallel magnetic field is applied (full line). This leads to a diamagnetic shift of the energy levels, and an increase in the subband separation, as indicated for the first two subbands.

By applying the momentum operators to the wave function ansatz

$$\Psi(\vec{r}) = \exp(ik_x x) \exp(ik_y y) \Phi(z)$$

this can be written as

$$\left[\frac{\hbar^2 k_x^2}{2m^*} + \frac{\hbar^2 k_y^2}{2m^*} + \frac{1}{2} m^* \omega^2(B_{\parallel}) \left(z^2 - \frac{2\hbar k_y \omega_c}{m^* \omega^2(B_{\parallel})} z \right) + \frac{p_z^2}{2m^*} \right] \Psi(\vec{r}) = E \Psi(\vec{r}) \quad (6.32)$$

with

$$\omega(B_{\parallel}) = \sqrt{\omega_0^2 + \omega_c^2} \quad (6.33)$$

To complete the square, we add and subtract

$$z^2(k_y) = \left[\frac{\hbar k_y \omega_c}{m^* \omega^2(B_{\parallel})} \right]^2 \quad (6.34)$$

and obtain

$$\left[\frac{\hbar^2 k_x^2}{2m^*} + \frac{\hbar^2 k_y^2}{2m^* \omega_0^2(B_{\parallel})} + \frac{1}{2} m^* \omega^2(B_{\parallel}) [z - z(k_y)]^2 + \frac{p_z^2}{2m^*} \right] \Psi(\vec{r}) = E \Psi(\vec{r}) \quad (6.35)$$

where

$$m_y^*(B_{\parallel}) = m^* \frac{\omega^2(B_{\parallel})}{\omega_0^2} \quad (6.36)$$

The solution of Eq. (6.23) consists of free electrons in the x - and y -directions, with an effective mass in the y -direction which increases as B_{\parallel} is increased.

Intuitively, we can think of the electron trajectories being bent by B_{\parallel} in the z -direction, such that the electron has more difficulties moving in the y -direction. Hence, the Fermi sphere is deformed into an ellipse (Fig. 6.16). As we have seen already during the discussion of the Fermi surface of Si in Chapter 3, the average effective mass is given by

$$m^*(B_{\parallel}) = \sqrt{m_x^* m_y^*} = m^* \frac{\omega(B_{\parallel})}{\omega_0} \quad (6.37)$$

Therefore, the density of states increases for each subband as B_{\parallel} increases. The total density of states is given by

$$D_{\text{PQW}}(E, B_{\parallel}) = \frac{m^*(B_{\parallel})}{\pi \hbar^2} \sum_{j=0}^{\infty} \theta(E - E_j(B_{\parallel})) \quad (6.38)$$

with the subband energies

$$E_j(B_{\parallel}) = (j + \frac{1}{2}) \hbar \omega(B_{\parallel}) \quad (6.39)$$

Note further that the states in each subband are centered at positions given by $z(k_y)$. Electrons that move in the $+k_y$ -direction ($-k_y$ -direction) are predominantly located at more negative (positive) z (see Fig. 6.16). This effect just describes the deflection of the moving electrons in a magnetic field.

In order to check this model experimentally, we can apply a parallel magnetic field to a parabolic quantum well and probe the effective mass by temperature-dependent Shubnikov-de Haas measurements in an additional perpendicular magnetic field. True, the Hamiltonian has to be modified, but as long as the perpendicular magnetic field B_{\perp} is sufficiently small, we can treat it as a perturbation, which leaves the modifications imposed by B_{\parallel} unchanged. If more than one subband is occupied at $B = 0$, we should be able to see the magnetic depopulation of the upper subbands as B_{\parallel} increases. In a standard experimental setup, however, there is only one magnetic field direction available. This problem can be solved by measuring the magneto-resistivity with the sample tilted with respect to the magnetic field direction. After performing this experiment for a sequence of different tilt angles, the magnetic field can be disentangled into its parallel and perpendicular components and we are able to analyze $\rho_{xx}(B_{\perp}, B_{\parallel})$.

In Fig. 6.17, some raw data of such an experiment are shown. Here, three subbands were occupied in the parabolic quantum well at $B_{\parallel} = 0$. At zero tilt angle, the effect of a purely parallel magnetic field on the resistivity can be studied. One observes a minimum in $\rho_{xx}(B_{\parallel})$ at $B_{\parallel} \approx 0.7$ T, and a sharp decrease at $B_{\parallel} = 2.2$ T, which are attributed to the depopulation of the third and second subbands, respectively. While the origin of the minimum is not well understood, the resistivity drop at the depletion of the second subband

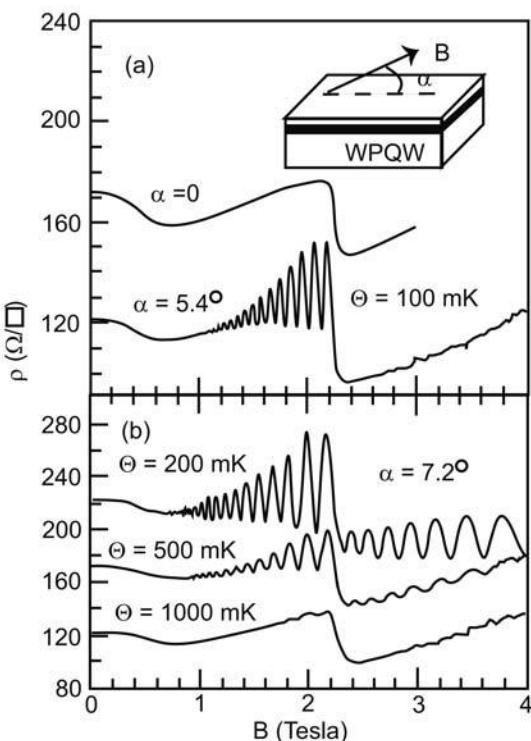


Fig. 6.17 Magneto-resistivity measurements in a parabolic quantum well in tilted magnetic fields. (a) Plot of $\rho_{xx}(B)$ for a tilt angle of zero ($\theta = 0$). The minima can be attributed to the depletion of subbands 3 and 2, respectively. As the sample is tilted, a perpendicular magnetic field component generates SdH oscillations. (b) Temperature-dependent measurements allow one to determine the effective electron mass. After [88].

has two reasons. First of all, the electrons in the second subband suffer a lot of scattering at low subband densities N_2 , which increases the resistivity. Second, inter-subband scattering is no longer possible for $B_{\parallel} > 2.2$ T. As the sample is tilted, the perpendicular magnetic field induces Shubnikov–de Haas oscillations, which can be used to determine the subband densities, once the oscillations have been attributed to a particular subband. In the example shown in Fig. 6.17, the oscillation in 1 T $< B < 2$ T is attributed to the second subband, while oscillations at higher magnetic fields stem from the first subband.

In addition, Hall measurements can be performed in order to determine the total electron density. Hence, the electron densities N_1 and N_2 can be determined, as shown in Fig. 6.18(a). As expected, the upper subband is depleted by B_{\parallel} , while N_1 approaches the total electron density at strong parallel magnetic fields. Furthermore, temperature-dependent measurements can be used

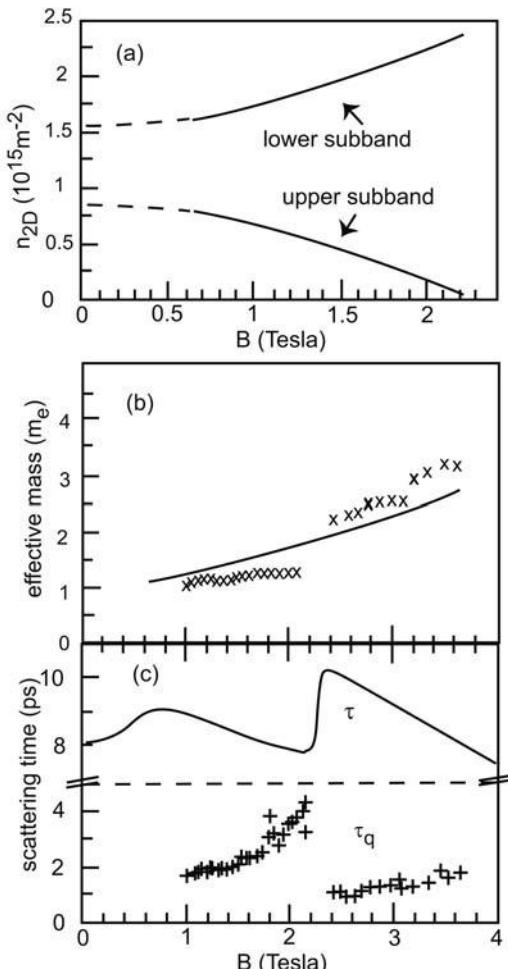


Fig. 6.18 Analyzing data of the type as shown in Fig. 6.17 gives not only electron densities of the lowest two subbands (a), but also the effective mass (b) and the scattering times (c) as functions of B_{\parallel} . While $m^*(B_{\parallel})$ is in reasonable agreement with the model

described in the text (full line in (b)) and the behavior of τ is as expected, it is somewhat surprising that τ_q increases strongly in the second subband as this band gets depleted. After [88].

to determine the effective mass and the quantum scattering time, as shown in Figs. 6.18(b) and (c), respectively. While the average effective mass agrees reasonably well with the model described above, it is found that the quantum scattering time τ_q increases strongly as the second subband gets close to depletion, an effect that is poorly understood. Apparently, the electrons screen the small-angle scattering potential much better when they are in the first subband.

Papers and Exercises

P6.1 Go through [282] and discuss the difficulties encountered when capacitances in high magnetic fields are measured by applying a voltage directly between a top gate and the electron gas.

P6.2 The helium fountain effect (see Section 4.2) has been used in [173] to detect the spatial distribution of dissipation in the quantum Hall effect. Discuss this entertaining experiment in relation to the equipotential lines sketched in Fig. 6.8.

P6.3 An interesting relation between Shubnikov–de Haas oscillations and the quantum Hall effect has been reported in [290]. What is the explanation suggested by Simon and Halperin [278]?

P6.4 Discuss the quantum Hall effect observed in graphene, as reported in [223].

E6.1 Analyze the data of Fig. 1.3, measured on a 2DEG in Ga[Al]As. Enumerate the quantum Hall plateaus. Determine the electron density both from the Hall slope and from the Shubnikov–de Haas oscillations. Extract the mobility and the scattering time. Estimate the effective g -factor.

E6.2 Figure 6.10 shows measurements of the resistivity of a two-dimensional electron gas as a function of magnetic field and temperature. The electron density has been determined from the Hall slope to be $n = 6.2 \times 10^{15} \text{ m}^{-2}$, while from measuring $\rho_{xx}(B = 0)$ the mobility is found to be $\mu = 77 \text{ m}^2/\text{V s}$.

(a) Determine τ , m^* , and τ_q .

(b) What material could it be? What does the ratio τ/τ_q tell you?

E6.3 A one-dimensional harmonic oscillator (characterized by ω_0) is placed in a constant electric field F .

Solve the Schrödinger equation. Show that the parabola gets displaced in both space and energy, but maintains its shape. Determine the location and the energy of the potential minimum as a function of F .

Further Reading

The full story of the quantum Hall effect is much more complicated than the simple picture developed here would suggest. The fractional quantum Hall

effect has been left aside. For a full theoretical discussion, the reader is referred to [87]. More elementary introductions are given in [136, 246, 341]. More on quantum films in general can be found in [90].

7

Quantum Wires and Quantum Point Contacts

Quantum wires are quasi-one-dimensional systems, i.e. their width w must be comparable to the Fermi wavelength. In analogy to our previous notation in two dimensions, the wire is strictly one-dimensional if only the mode with the lowest energy is occupied. The wire is called *diffusive* if its length L is much larger than the elastic mean free path ℓ_e (Fig. 7.1(a)). In this case, the electrons will suffer many elastic scattering events during their trip along the wire. Note that the trajectories indicated by an arrow are only meaningful for $\lambda_F \ll w$, which means that the number of occupied modes is sufficiently large, and a localized wave packet can be constructed. This is not necessarily true in quantum wires. If only a few modes are occupied, the semiclassical picture breaks down, and we should think of the electrons as plane waves inside the quantum wire. We will study the basic magneto-resistance properties of diffusive quantum wires in Section 7.1. In the opposite limit, $L \ll \ell_e$, there is no elastic scattering in the wire, except for boundary scattering at the walls (Fig. 7.1(b)). Such wires are often very short and form a point-like contact between the left and the right reservoir. Such short ballistic quantum wires are usually called *quantum point contacts* (Fig. 7.1(c)).

One of the central questions in this chapter is the resistance of ballistic quantum wires. “Well,” you might say, “there should be no resistance in a ballistic wire.” Whether this is true or not depends on what exactly we mean by “the resistance of the wire”. It turns out that a two-terminal measurement gives quantized resistances, which very closely resemble the quantum Hall effect, and there is in fact a surprising relation. If, however, the resistance is measured in a four-terminal geometry using suitable voltage probes, the resistance is in fact zero. The formalism used to describe transport in ballistic wires has been developed by R. Landauer and M. Büttiker. An introduction to ballistic quantum wires is given in Section 7.2. In Section 7.3, we will discuss the quantum Hall effect and the Shubnikov–de Haas (SdH) oscillations in terms of transport through ballistic quantum wires. This includes introducing the Landauer–Büttiker formalism. It will be established that quasi-one-dimensional *edge states* carry the current in the quantum Hall regime.

All the quantum wires we will have discussed up to this point reside in semiconductor hosts, but this is not the only way of realizing a quantum wire experimentally. Some alternative approaches will be presented in Section 7.4, which rounds off our discussion of quantum wires.

Throughout this chapter, the wires extend in the x -direction, while the motion in the y - and z -directions is quantized. Furthermore, we use the effective mass approximation, except when carbon nanotubes are discussed in Section 7.5. Hence, the density of states of a quantum wire (QWR)¹ is given by

$$D_{\text{QWR}}(E) = \sum_{j=1}^{\infty} D_1(E - E_j) \quad (7.1)$$

where $D_1(E - E_j)$ denotes the one-dimensional density of states with a band bottom at energy E_j (see Fig. 7.2). The singularities at the mode bottoms are in reality smeared out by both disorder and temperature, and thus do not cause

¹⁾ The acronym QWR is used for “quantum wire” to make a distinction from QW for “quantum well”.

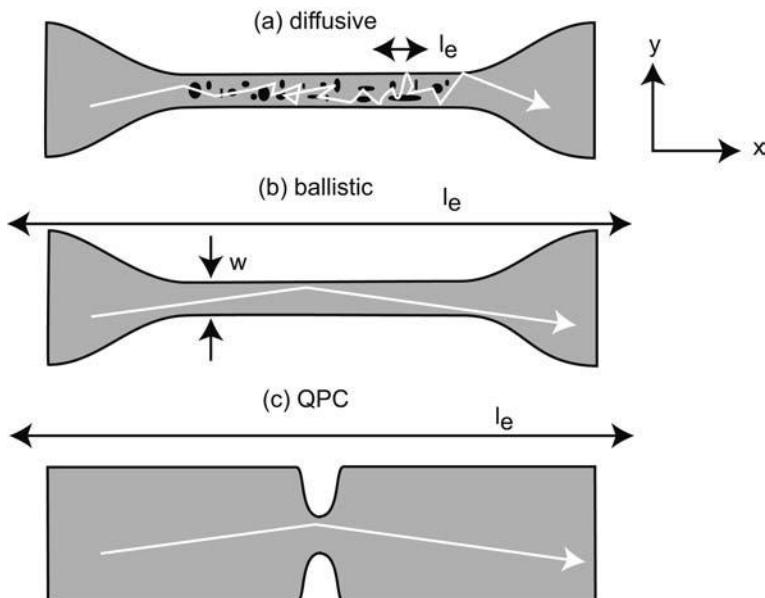


Fig. 7.1 (a) A diffusive wire contains many scatterers. Its transport properties can be described by the Boltzmann equation. This clearly becomes questionable for wires of length $L \approx l_e$. (b) A wire with $L > l_e$ is called ballistic. The electrons are scattered at the confining walls only. (c) A ballistic wire with $w \approx L \gg l_e$ is called a “quantum point contact”. Adapted from [27].

any difficulties. The electron density in the ideal QWR is given by

$$n_{\text{QWR}} = \frac{2}{\pi\hbar} \sum_{j=1}^{\infty} \sqrt{2m^*(E - E_j)} \theta(E - E_j) \quad (7.2)$$

Here, we have assumed a spin degeneracy of 2, and $\theta(E - E_j)$ denotes the Heaviside step function.

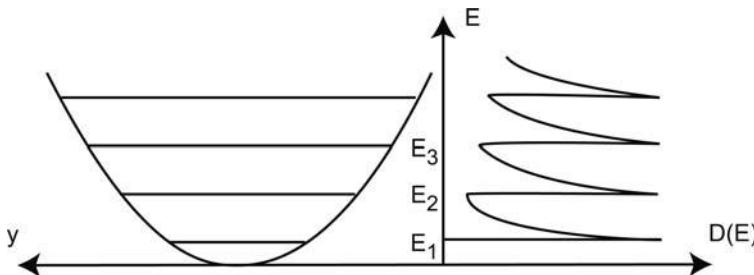


Fig. 7.2 Schematic energy spectrum of a parabolic quantum wire, and the corresponding density of states.

7.1 Diffusive quantum wires

7.1.1 Basic properties

An experimental realization of a diffusive quantum wire in a Ga[Al]As HEMT is shown in Fig. 7.3. For such wires, a parabolic confinement in the y -direction is an excellent approximation (z is, as usual, the growth direction):

$$V(y) = \frac{1}{2}m^*\omega_0^2y^2 \quad (7.3)$$

In fact, it is not easy to detect experimentally a non-parabolic confinement. The characteristic quantities of the wire can be determined by – you guessed right – magneto-transport experiments. How does a magnetic field influence the energy levels in a quantum wire? The problem is similar to that of a parabolic quantum well in a parallel magnetic field, as studied in Section 7.5. We have to add the potential (7.3) to the Hamiltonian

$$H = \frac{1}{2m^*}p_y^2 + \frac{1}{2}m^*\omega_c^2(y - y_n)^2 \quad (7.4)$$

known from the Landau quantization. We use the ansatz for the wave function

$$\Phi(x, y) = e^{ik_xx}\psi(y) \quad (7.5)$$

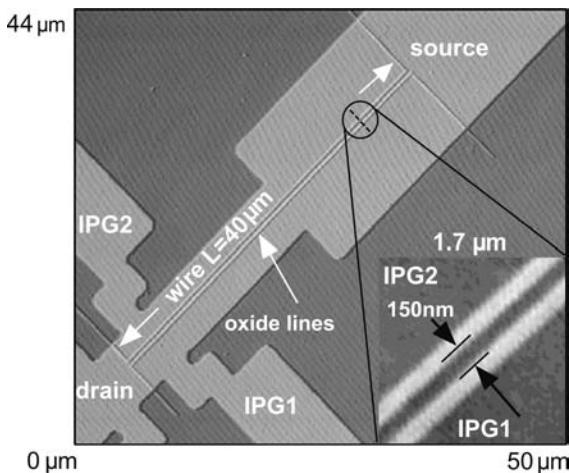


Fig. 7.3 Top view of a diffusive quantum wire ($L = 40 \mu\text{m}$, geometric width $w_g = 150 \text{ nm}$), patterned into a Ga[Al]As HEMT by local oxidation. The bright lines define the walls (a close-up is shown in the inset). The electrostatic wire width w can be tuned by applying voltages to the two in-plane gates IPG1 and IPG2.

and obtain (see Exercise E7.1)

$$H_{xy} = \frac{1}{2m^*} p_y^2 + \frac{1}{2} m^* \omega(B)^2 (y - \bar{y}_n)^2 + \frac{\hbar^2 k_x^2 \omega_0^2}{2m^* \omega(B)^2} \quad (7.6)$$

with

$$\omega(B) = \sqrt{\omega_0^2 + \omega_c^2} \quad \text{and} \quad \bar{y}_n = y_n \frac{\omega_c^2}{\omega^2(B)} \quad (7.7)$$

The last term in Eq. (7.6) describes plane waves in the x -direction. They have the usual free electron dispersion, with the magnetic mass

$$m^*(B) = m^* [\omega(B)/\omega_0]^2 \quad (7.8)$$

which is *larger* than the effective mass at $B = 0$. The remaining terms in Eq. (7.6) give an effective confinement in the y -direction, characterized by $\omega(B)$. The energy eigenvalues therefore depend on B and are given by

$$E_j(B) = \hbar \omega(B) (j - \frac{1}{2}) \quad (7.9)$$

with $j = 1, 2, \dots$ enumerating the one-dimensional modes. The resulting density of states is sketched in Fig. 7.2.

Question 7.1: How do the wire modes evolve into Landau levels in the limit $\omega_c \gg \omega_0$? What happens to the magnetic mass and to the electron velocity in the x -direction?

The magnetic field thus increases the confinement strength and may depopulate the magnetoelectric modes by squeezing them above the Fermi level, very similar to the diamagnetic shift discussed in Chapter 6. The density of states at the Fermi level therefore oscillates as a function of B , which is reflected in the resistivity. In contrast to the SdH oscillations, the corresponding magneto-oscillations are not periodic in $1/B$. At large magnetic field ($\omega_0 \ll \omega_c$), the electrons feel an effectively two-dimensional potential governed by the magnetic confinement. In this case, the magneto-transport properties should approach those of a 2DEG in the quantum Hall regime. At small magnetic fields, on the other hand, the mode spacing approaches $\hbar\omega_0$ as B is decreased, and the level spacing becomes approximately independent of B . The number j of occupied modes as a function of B , ω_0 and n_{QWR} has been calculated by Berggren [30]. The authors obtain (see Exercise E7.1)

$$j(n_{QWR}, \omega_0, 1/B) = \left(\frac{3\pi}{4} n_{QWR} \omega_0 \right)^{2/3} \left(\frac{\hbar}{2m^*} \right)^{1/3} \frac{1}{\omega(B)} \quad (7.10)$$

This equation can be used to determine the characteristic wire parameters by fitting $j(n_{QWR}, \omega_0, 1/B)$, using ω_0 and n_{QWR} as fit parameters. A measurement of $\rho_{xx}(B)$, performed on the quantum wire of Fig. 7.3, including the result of such a fitting procedure, is shown in Fig. 7.4.

7.1.2

Boundary scattering

In quantum wires, the electrons hit the confining wall much more frequently than in a 2DEG. Whether this boundary scattering contributes to the resistivity is a legitimate question. This would be the case if boundary scattering changed the electron momentum in the x -direction. Smooth walls, i.e. walls that show spatial variations only on length scales much larger than the Fermi wavelength, scatter the electrons specularly and thus do not cause additional resistivity. The smoothness of the wall is hence a critical quantity. It has been experimentally demonstrated that, usually, boundary scattering is almost completely specular, unless the walls are made intentionally rough. Such wires with highly diffusive walls have been fabricated by ion implantation [300]. Here, a maximum in $\rho_{xx}(B)$ at $w \approx \frac{1}{2}r_c$, known as a “wire peak”, can be observed (see Fig. 7.5). It can be shown that the magnetic field determines how sensitive the electrons are to the specularity of the confinement. The highest sensitivity is reached for cyclotron radii of roughly twice the electronic wire width w , in detail $w = 0.55r_c$. In a simple picture, we can imagine that, at this magnetic field, the fraction of the electron trajectories close to the wire edge is maximized. For the details of this effect, which is beyond our scope here, see [27] and [300], as well as references therein.

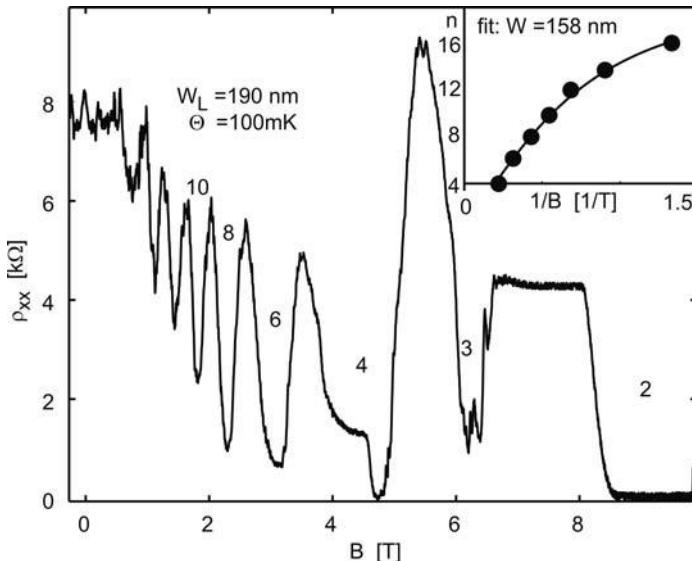


Fig. 7.4 The magneto-resistivity of the wire shown in Fig. 7.3. The most prominent feature is an oscillation as a function of B , which detects the magnetic depopulation of the wire modes. Further structures are a pronounced maximum at $B = 0$, and fluctuations at small magnetic fields. These are due to phase

coherence effects and will be discussed in Chapter 8. The positions of the oscillation minima are plotted vs. $1/B$ as full circles in the inset. Here, the line is a least squares fit to Eq. (7.10), which gives an electronic wire width of 158 nm.

The measurements in Fig. 7.5 show a specularity of only 70% (which means that 70% of the boundary scattering events take place in a specular fashion), while in top gate defined wires the specularity is so high that observing a wire peak is quite hard (inset in Fig. 7.5). In fact, a wire peak is not visible in Fig. 7.4, although this particular wire is 40 μm long. Therefore, for QWRs defined by top gates, by cleaved edge overgrowth, or by local oxidation, we can safely neglect boundary scattering.

7.2

Ballistic quantum wires

7.2.1

Phenomenology

In 1988, van Wees et al. [317] and Wharam et al. [326] investigated the transport properties of quantum point contacts (QPCs), of the shape sketched in Fig. 7.1(c). The QPCs were created by applying suitable voltages to a split

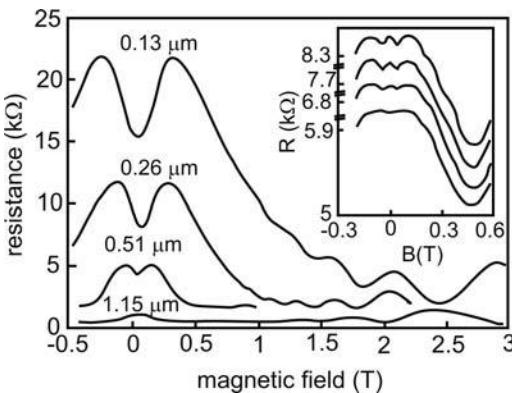


Fig. 7.5 Diffusive scattering at wire boundaries causes a magneto-resistivity peak at $w = 0.55r_c$. The main figure shows these peaks for wires of different widths (stated in the figure), studied at QWRs made by ion beam implantation, which generates particularly rough, diffusive walls. Wire peaks in QWRs defined by top gates of similar widths (shown in the inset) are much weaker. The length of the wires was $L = 12 \mu\text{m}$. After [300].

gate (see the inset in Fig. 7.6) on top of a Ga[Al]As HEMT structure. With such geometries, the QPC is imposed on the 2DEG by tuning the gate voltage to negative values, such that the electron gas underneath gets depleted. By further decreasing the gate voltage, the lateral electric stray field, and with it the lateral depletion zone around the gates, increases. This can be used to tune the electronic width, and with it the number j of occupied modes of the QPC, ideally all the way down to zero.

Once a small background resistance, which stems from the 2DEG between the QPC and the voltage probes, has been subtracted, the conductance of such QPCs turned out to be quantized in units of $j \times 2e^2/h$, *in the absence of magnetic fields* – see Fig. 7.6 (and Fig. 1.2, by the way). This quantization, of course, very closely resembles the quantum Hall effect as a function of the electron density. In QPCs, however, we must remember the following:

- There is no Landau quantization, and there are no scatterers in the region of interest.
- The accuracy is typically of the order of $\delta R/(h/2e^2) \approx 10^{-2}$, much smaller than the accuracy of ρ_{xy} inside a quantum Hall plateau.
- Subsequent studies further revealed that the conductance quantization vanishes in longer quantum wires, typically for $L > 2 \mu\text{m}$, although signatures of quantization have been observed in wires up to 20 μm long [333]. This is also in contrast to the quantum Hall effect (QHE), which can be observed in samples of millimeter sizes.

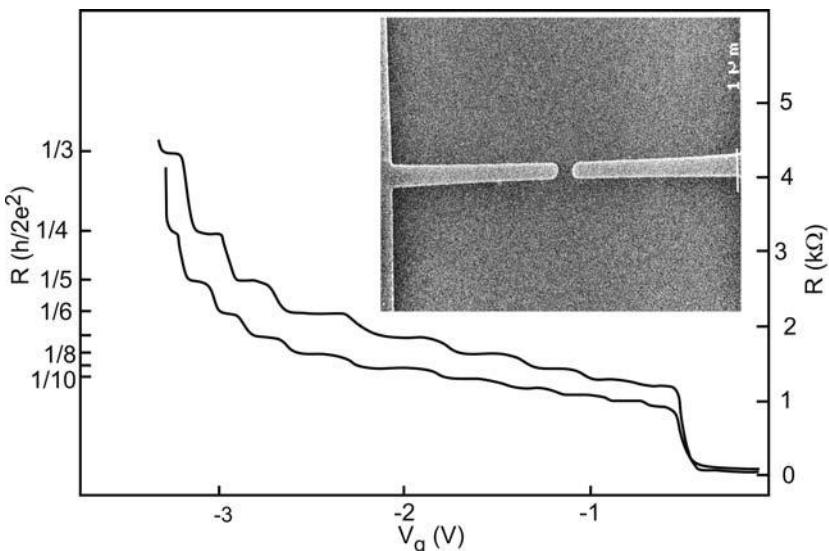


Fig. 7.6 Inset: Scanning electron micrograph of a split gate structure used to define a quantum point contact. The brighter areas are gold electrodes on top of a Ga[Al]As HEMT. Main figure: Resistance as a function of voltage V_g applied to the split gates with respect to the 2DEG. At a threshold voltage $V_{g,\text{th}} = -500 \text{ mV}$, the resistance sharply increases by about $1 \text{ k}\Omega$. Here, the

2DEG underneath the gates is depleted and the QPC is defined. As V_g is further reduced, quantized steps in the resistance at $R_j = (1/j)(2e^2/h)$ are observed. The two traces are taken for two different carrier concentrations in the 2DEG, which has been changed by illumination with an infrared light-emitting diode. The temperature was 60 mK . The measurements are adapted from [326].

As we shall see, there is in fact a close relation between the QHE and the conductance quantization in a QPC. Before we study this connection, some more obvious questions need to be discussed: Why is there a resistance at all in QPCs, although there are no scatterers around? Should we not just measure the resistance in series with the QPC? And why is the conductance quantized in units of $2e^2/h$?

7.2.2

Conductance quantization in QPCs

Essentially, the previous questions can be answered in two steps:

1. There is no backscattering either inside the QPC or at its exit.
2. The occupation of the states in close proximity to the QPC is not described by a Fermi–Dirac distribution.²

2) This is actually true for many mesoscopic transport scenarios.

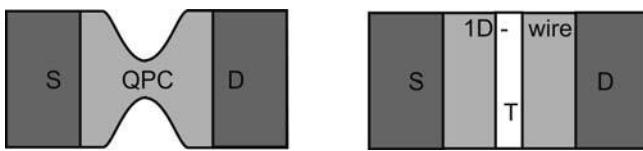


Fig. 7.7 A QPC attached to source and drain (left) and its idealized model (right), where the transition regions are one-dimensional leads, and the constriction is a barrier with transmission probability T .

To begin with, we look in more detail at the geometric shape of a QPC (Fig. 7.7). Clearly, the QPC is connected to source S and drain D via two transition regions, which are quasi-one-dimensional. In a rather crude, but nevertheless very insightful, approximation, we replace the transition region by ballistic, strictly one-dimensional QWRs and the QPC itself by a barrier with transmission probability T . We do so since we plan to calculate the conductance of the QPC from a scattering approach, where incoming electronic plane waves are scattered into outgoing plane waves, which are eigenfunctions of the one-dimensional wires. For simplicity, we assume an energy-independent transmission probability. The QPC is open for $T = 1$.

We now calculate the conductance for our model QPC. For this purpose, recall that the current density in its simplest form is given by $j = -nev$, where n is the three-dimensional carrier density and v is the velocity of the electrons. The corresponding one-dimensional expression is obtained by integrating over the cross section of the current; it reads $I = -n_1ev$. Here, I is the current and n_1 is the one-dimensional electron density. This simple relation is generalized to our model system as follows. Suppose a voltage $V = (\mu_S - \mu_D)/(-e)$ drops between source and drain. The reservoirs fill the connected states of the wire with k -vectors pointing away from the reservoir (outgoing states), up to their respective electrochemical potentials. Now, I , n_1 and \vec{v} depend on energy. Furthermore, the density of right-moving electrons at energy E is given by the density of states on the side $j = S, D$ of the barrier, multiplied by the corresponding Fermi function:

$$\vec{n}_j(E) = \vec{D}_j(E)f(E - \mu_j)$$

Similarly, we have for the density of left-moving electrons:

$$\overleftarrow{n}_j(E) = \overleftarrow{D}_j(E)f(E - \mu_j)$$

Here, $\vec{D}_j(E)$ ($\overleftarrow{D}_j(E)$) is the density of states for right-moving (left-moving) electrons. An electron contributes to the current if it traverses the barrier.³

3) We consider a coherent scenario. In the case where the transmission is incoherent, the probability of the states on the other side being empty has to be included, i.e. the transmission probability has to be multiplied by $[1 - f(E - E_j)]$.

The spectral current $I(E)$ is therefore given by

$$\begin{aligned} I(E) &= eT[\vec{n}_S(E)v_S(E) - \vec{n}_D(E)v_D(E)] \\ &= eT[\vec{D}_S(E)f(E - \mu_S)v_S(E) - \vec{D}_D(E)f(E - \mu_D)v_D(E)] \end{aligned} \quad (7.11)$$

where T is taken to be independent of energy. Assuming just a spin degeneracy of 2, we have

$$\vec{D}_j(E) = \vec{D}_j(E) = \frac{1}{2}D_1(E - E_{1,j}) = \frac{1}{2\pi\hbar}\sqrt{\frac{2m^*}{E - E_{1,j}}} \quad (7.12)$$

where $E_{1,j}$ denotes the bottom of the mode on side j . Since the electron velocity in the mode is given by

$$v_j(E) = \sqrt{\frac{2(E - E_{1,j})}{m^*}} \quad (7.13)$$

Eq. (7.11) simplifies to

$$I(E) = \frac{eT}{\pi\hbar}[f(E - \mu_S) - f(E - \mu_D)] \quad (7.14)$$

For zero temperature,⁴ the Fermi functions become Heaviside step functions $\theta(\mu_j - E)$, and we obtain a total current of

$$I = \frac{eT}{\pi\hbar} \int_{E=0}^{\infty} [\theta(\mu_S - E) - \theta(\mu_D - E)] dE = \frac{2e^2}{h}TV \quad (7.15)$$

Therefore, the conductance for a mode with $T = 1$ equals

$$G = \frac{2e^2}{h} \quad (7.16)$$

It is quantized because the energy dependence of the one-dimensional density of states and that of the electron velocity cancel each other.

Apparently, the quantized conductance follows quite naturally from this simple consideration. The result is nevertheless quite surprising. For $T = 1$, the electrons suffer no scattering at the QPC. A finite conductance for such a scenario is counter-intuitive. The answer can be found in the subtlety that the voltage drop across the QPC is *not* the difference between the source and the drain potentials, divided by e . Within the mean free path around the barrier, transport is ballistic, and therefore states moving away from the barrier are only occupied if an electron has been scattered into them by reflection at the

4) The effect of $\Theta > 0$ is considered in Exercise E7.2.

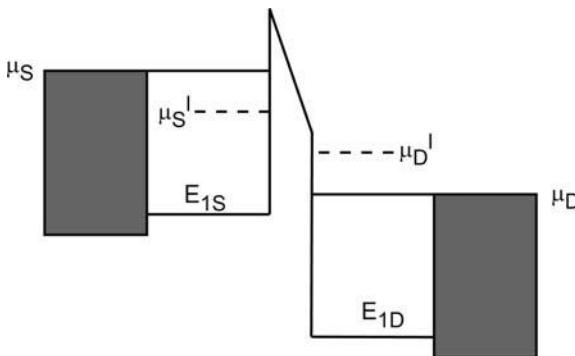


Fig. 7.8 Electrochemical potentials of the system of Fig. 7.7, both at the reservoirs and close to the barrier (dashed lines).

barrier. Clearly, the occupation probability of the states does not have the form of a Fermi–Dirac distribution. This point has been quantified in [49].

In order to explain how this fact can shed light on the issue, we again allow $0 \leq T \leq 1$ and consider the open channel as a special case later on. For the sake of simplicity, let us assume that the source–drain bias voltage is sufficiently small, which means that the density of states can be set constant in this small energy range of interest. Therefore, we obtain

$$\overleftrightarrow{D}_S(E) = \overleftarrow{D}_S(E) = \overrightarrow{D}_D(E) = \overleftarrow{D}_D(E)$$

Consider the scenario depicted in Fig. 7.8. To the left side of the barrier, all right-moving states are occupied up to μ_S , since they are connected to the source potential. On this side, a left-moving state close to the barrier is occupied only if an electron has scattered into it, which can happen by backscattering at the barrier, or by forward scattering of drain electrons across the barrier. Hence, the electrochemical potential at the left side must be smaller than μ_S . We denote this local chemical potential by μ_S^l . Likewise, some right-moving states on the right side with energies within $[\mu_D, \mu_S]$ are occupied by electrons that have been transmitted through the barrier. This results in a local potential at the right side $\mu_D^l > \mu_D$.

Recall that the chemical potential μ of a metal can be defined as that energy for which the number of empty states below μ equals the number of occupied states above μ . This definition is very convenient here, and we use it to calculate μ_S^l and μ_D^l . On the left side, the only empty states with $E < \mu_S^l$ are those left-moving states that have not been occupied by reflection of electrons at the barrier. The density of these can now be obtained from

$$n_{S,\text{empty}}(E < \mu_S^l) = \int_{\mu_D}^{\mu_S^l} \overleftarrow{D}_S(E) T dE = \overleftarrow{D}_S T (\mu_S^l - \mu_D) \quad (7.17)$$

Here, we do not count states below μ_D , since these contributions cancel each other. The density of occupied states $n_{S,\text{occ}}(E > \mu_S^\ell)$ is the sum of two components. First, all right-moving states on the left side in $[\mu_S^\ell, \mu_S]$ are occupied, since they get filled by the source reservoir. Second, there are left-moving states that got populated by backscattering at the barrier. Hence,

$$n_{S,\text{occ}}(E > \mu_S^\ell) = \overrightarrow{D}_S(\mu_S - \mu_S^\ell) + (1 - T)\overleftarrow{D}_S(\mu_S - \mu_S^\ell) \quad (7.18)$$

Since

$$\overleftarrow{D}_S = \overrightarrow{D}_S$$

we obtain the local chemical potential at the left side via

$$\begin{aligned} n_{S,\text{empty}}(E < \mu_S^\ell) &= n_{S,\text{occ}}(E > \mu_S^\ell) \implies \\ \mu_S^\ell &= \mu_S - \frac{1}{2}T(\mu_S - \mu_D) \end{aligned} \quad (7.19)$$

A corresponding consideration for the part of the wire attached to drain gives

$$\mu_D^\ell = \mu_D + \frac{1}{2}T(\mu_S - \mu_D) \quad (7.20)$$

Question 7.2: Derive Eq. (7.20).

The local voltage drop at the barrier is thus given by

$$V^\ell = \frac{1}{e}(\mu_S^\ell - \mu_D^\ell) = \frac{1}{e}(\mu_S - \mu_D)(1 - T) \quad (7.21)$$

and we find the conductance of the barrier to be

$$G_{\text{barrier}} = \frac{eI}{V^\ell} = \frac{2e^2}{h} \frac{T}{1 - T} \quad (7.22)$$

which certainly makes a lot of sense: as T approaches unity, the barrier conductance goes to infinity. Since the overall conductance between source and drain equals

$$G_{SD} = \frac{2e^2}{h}T \quad (7.23)$$

there must be a resistance in series with the barrier of

$$R_{\text{contact}} = \frac{1}{G_{SD}} - \frac{1}{G_{\text{barrier}}} = \frac{h}{2e^2} \quad (7.24)$$

This is called the *contact resistance*, since it occurs at the interface between the reservoirs and the one-dimensional wire. It arises from the electrons that have

passed the quantum wire and enter the contacts in a distribution that differs from the Fermi function. The electrons coming from S are injected into the drain contact at energies above μ_D , dissipate their excess energy via inelastic scattering events, and finally end up in a Fermi–Dirac distribution. This relaxation causes the contact resistance at the drain side. Similarly, the electrons that pass the wire from D to S find all states at their energy in S occupied, so scattering is required to let them enter. For $T = 1$, the right-moving states in the wire are occupied up to μ_S , while the left-moving states are populated up to μ_D . The local chemical potential inside the wire is $(\mu_S + \mu_D)/2$. The voltage thus drops by $V/2$ at both the entrance and the exit.

Our purpose here was to shed some light on the conductance quantization in quasi-one-dimensional systems. Along the way, many questions remain, and a lot of assumptions have certainly been made. To work out all these details is a formidable task, and we briefly outline some of them below.

First of all, we have considered only a single mode. The Landauer formula [186] (see also [45])

$$G = \frac{2e^2}{h} \sum_{\alpha,\beta} |t_{\alpha\beta}|^2 \quad (7.25)$$

gives the conductance of a multimode QPC. The transmission amplitude from mode β to mode α is $t_{\alpha\beta}$. The partial conductances of the individual modes are thus additive for negligible coupling between different modes. This assumption implies that the electrons remain in their initial mode throughout their trip across the QPC. This kind of transport is called *adiabatic*. It requires that the width of the channel changes smoothly on the scale of the Fermi wavelength, which is usually the case in the experiments under discussion. The properties of adiabatic constrictions and the consequences of non-adiabaticity in experimental realizations are discussed in [335].

Another effect not considered are reflections of the electronic wave functions, which may take place at the entrance and the exit of the wire. Multiple reflections may lead to transmission resonances, as we will discuss in more detail in Chapter 8. It is intuitively clear that a smooth potential shape in the above sense suppresses backscattering of electrons at the exit, and interference effects are absent. Extensive numerical simulations have quantified how the QPC conductance depends on the potential shape. An example is shown in Fig. 7.9. In some experiments, weakly pronounced oscillations, superimposed on conductance steps, have been observed, but the agreement with theoretical considerations is poor. It has turned out to be very difficult to fabricate samples with sufficiently sharp contact regions. Thermal smearing further hampers a clear observation.

The conductance steps vanish at a characteristic temperature given by the energy spacing of the modes. As can be seen in Exercise E7.2, a sharp trans-

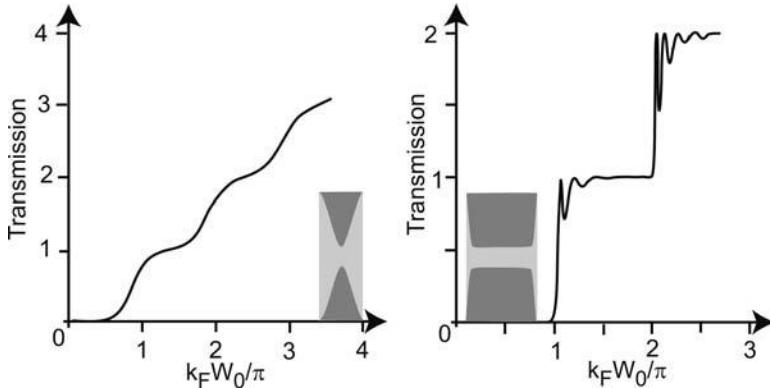


Fig. 7.9 Transmission as a function of the Fermi wavenumber, k_F , and the electronic width at its narrowest point, W_0 , calculated with the recursive Green's function technique [94], for zero temperature. A QPC with a smooth transition region (left; the geometry is shown in the inset) shows smooth transmission steps. In longer wires with sharp contacts to the reservoirs, resonances due to interference effects are found (right). After [199].

mission step at $\Theta = 0$ is thermally smeared at higher temperatures according to

$$G = \frac{2e^2}{h} f(E_1 - \mu) \quad (7.26)$$

which gives a characteristic temperature of

$$\Theta_{\text{char}} = \frac{\Delta}{2k_B \ln(3 + 2\sqrt{2})} \approx \frac{\Delta}{3.52k_B} \quad (7.27)$$

In the sample shown in Fig. 1.2, for example, the temperature for which the steps vanish is roughly 20 K, and the mode spacing is therefore of the order of 6 meV. One might thus expect that the steps become infinitely sharp as the temperature approaches zero. This is not observed experimentally. Rather, the steepness of the conductance steps tends to saturate as Θ is reduced below a few hundred mK. In [48], the potential step of our model is replaced by a saddle-shaped potential,

$$V(x, y) = \frac{1}{2}m^*(\omega_y^2 y^2 - \omega_x^2 x^2) \quad (7.28)$$

which should represent an excellent approximation for typical QPC geometries. The transmission probability of this potential is given by

$$T(\epsilon) = \frac{1}{1 + \pi\epsilon} \quad (7.29)$$

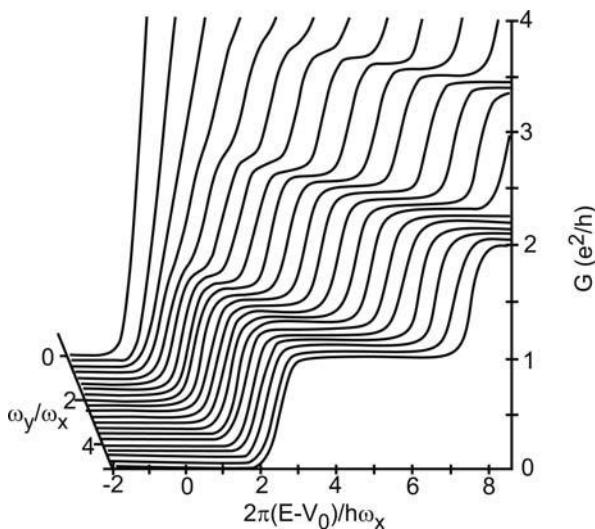


Fig. 7.10 Conductance of a QPC with a saddle point potential, characterized by ω_x and ω_y , calculated for zero temperature. The steps vanish as soon as ω_y becomes smaller than ω_x . After [48].

In this model, the energy separation between the modes equals $\hbar\omega_y$, while $\hbar\omega_x$ determines the width of the transition range between adjacent conductance steps (Fig. 7.10).

7.2.3

Magnetic field effects

As in diffusive quantum wires, the modes of a QPC get depleted by perpendicular magnetic fields (Fig. 7.11), since the magnetic field increases the confinement and the energy separation between the modes becomes larger. Simultaneously, quantum Hall states are formed in the 2DEG, which influences the resistance measurements. We can elegantly explain the data of Fig. 7.11, once the Landauer-Büttiker formalism has been introduced, and we will come back to this issue later. In weak perpendicular magnetic fields, QPCs show a negative magneto-resistance in addition, an effect discussed in Paper P7.2. Diamagnetic shifts of the QPC modes can, however, be conveniently studied in parallel magnetic fields, since here the effect that B exerts on the 2DEG is negligible.

This kind of spectroscopy has been performed in [261] on a QPC residing in a parabolic quantum well with two occupied subbands in the growth direction (Fig. 7.12). Transverse modes in both the y - and z -directions contribute to the current. We label the channels in the y -direction by l and in the z -direction by m , respectively. Since the confinement in the z -direction is significantly larger

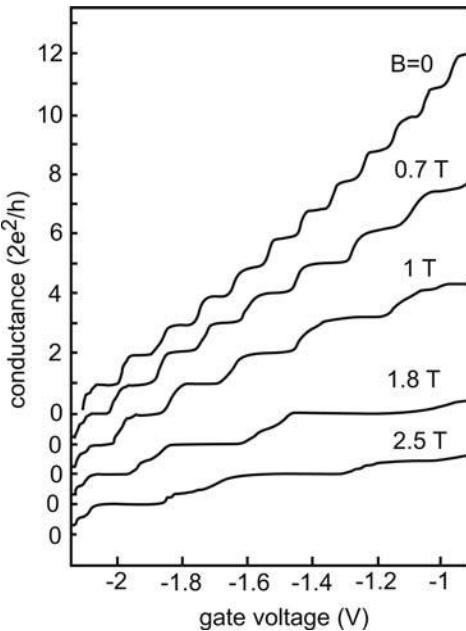


Fig. 7.11 Depopulation of QPC modes by a magnetic field. As B is increased, the width of the conductance plateaus increases. This reflects the increasing magnetoelectric confinement in the y -direction. At $B = 1.8$ T and above, the spin degeneracy gets lifted and additional plateaus evolve at odd integers of e^2/h . After [318].

than in the y -direction, we can think of the mode structure as being composed of ladders denoted by m , each with rungs labeled by l .⁵ The total number of modes carrying current is given by

$$j = \sum_{m,l=1}^{\infty} \theta(\mu - E_{m,l}) \quad (7.30)$$

Here, μ is the chemical potential (we assume that the source-drain voltage is negligibly small), and $E_{m,l}$ is the energy of the mode bottom.

To keep things simple, we restrict ourselves to the case of two occupied subbands in the growth direction. We again expect conductance quantization in units of $2e^2/h$. However, if two degenerate modes (belonging to different ladders) cross the chemical potential, the conductance will change by $4e^2/h$. Such a degeneracy shows up in the experiment as a suppression of conductance steps (Fig. 7.12). Via diamagnetic shifts, the energies of the modes, and thus the degeneracies, can be tuned by in-plane magnetic fields. The transconductance dG/dU_{sg} (U_{sg} is the voltage applied to the split gates) emphasizes the diamagnetic shift of the modes, and is a good representation of the QPC's

5) This is possible only if the Hamiltonian is separable.

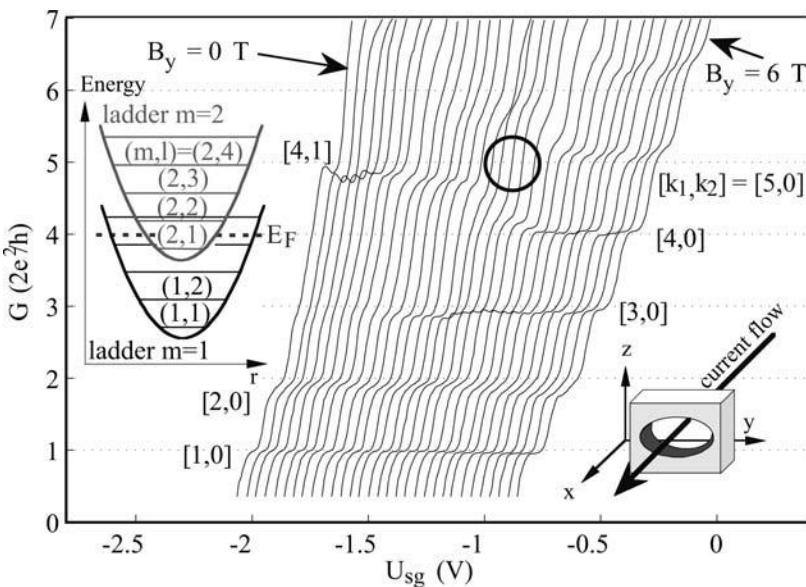


Fig. 7.12 The left inset shows a schematic mode spectrum of a QPC, located in a parabolic quantum well with two occupied subbands in the growth direction, plotted as a function of the spatial coordinate r . Each subband contains a ladder of equidistant modes; m is the subband index, while l labels the mode within a ladder. The main figure shows

the conductance as a function of the QPC gate voltage U_{sg} , measured for magnetic fields $B_\perp = 0$ to 6 T applied in the y -direction (see right inset). Conductance steps can be suppressed and recovered by the magnetic field, as in the encircled region. The labels $[k_1, k_2]$ denote the number of occupied modes in subbands 1 and 2. After [261].

energy spectrum at the same time. In Figs. 7.13(a) and (b), such measurements are shown for B applied in the y - and x -directions, respectively. Dark regions correspond to low transconductance: here, the chemical potential lies in between two adjacent modes. The bright lines (high transconductance) reflect the mode spectrum.

For both magnetic field directions, the two subband ladders are visible, with each ladder having its own characteristic dispersion. As a function of B_y , the levels of both ladders show a positive dispersion, which is much stronger in the $m = 2$ ladder. For magnetic fields in the x -direction, the dispersion of the $m = 1$ states is reversed: their energy decreases as B_x is increased. In both cases, mode crossings are clearly visible. The measurements in Figs. 7.13(a) and (b) are compared to model calculations in Figs. 7.13(c) and (d). It is reasonable to assume that the energy in the QPC is proportional to U_{sg} . The lever arm $\alpha = dE/dU_{sg}$ can be estimated by temperature-dependent measurements, which give the energy spacing between adjacent modes according to Eq. (7.27). A value of $\alpha \approx 0.02$ eV/V is found. Furthermore, we assume parabolic confinement in the y -direction as well. The confining potential at

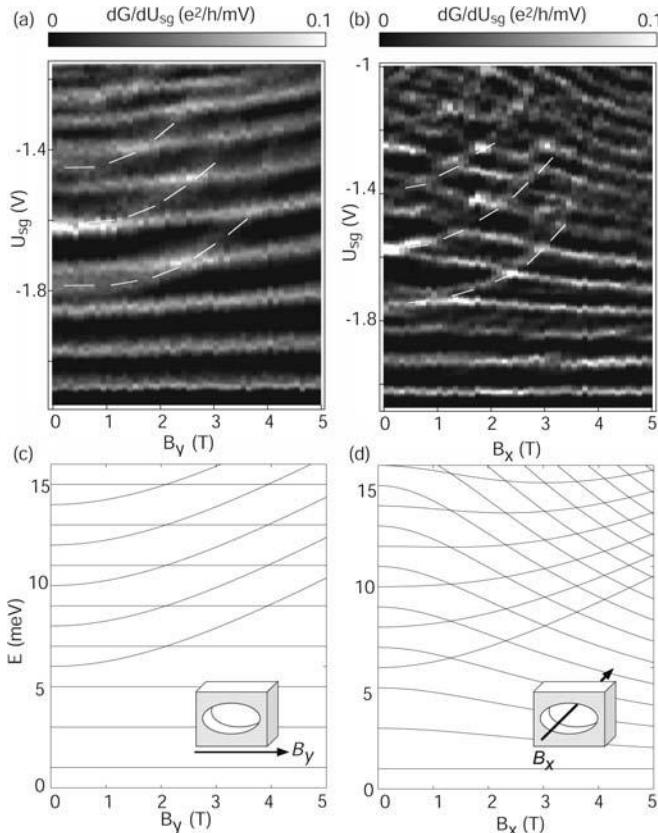


Fig. 7.13 Gray scale plot of the transconductance dG/dU_{sg} as a function of (a) B_y and (b) B_x . The bright lines represent the QPC modes. (c, d) The measurements are compared to the energy spectrum of an elliptical parabolic confinement. After [261].

the narrowest point of the QPC, which determines the number of transmitted modes, is then given by

$$V(y, z) = \frac{1}{2}m^*(\omega_y^2y^2 + \omega_z^2z^2) \quad (7.31)$$

After our previous discussion of parallel magnetic fields in Section 6.5, it is intuitively clear that B_y causes magnetic shifts of the modes' z -component and enhances the effective mass in the x -direction, while B_x does not modify the effective mass, but causes magnetic shifts of the y - and z -components. In [264], the Schrödinger equation has been solved for a magnetic field B_y applied in the y -direction. The energy eigenvalues are found to be

$$E_{ml}^y = \hbar\omega_y(l - \frac{1}{2}) + \hbar\sqrt{\omega_z^2 + \omega_c^2}(m - \frac{1}{2}) \quad (7.32)$$

For a magnetic field B_x in the x -direction, the energy levels are given by [267]

$$E_{ml}^x = \hbar\omega_1(l - \frac{1}{2}) + \hbar\omega_2(m - \frac{1}{2})$$

where

$$\omega_{1,2}^2 = \frac{1}{2}(\omega_c^2 + \omega_y^2 + \omega_z^2) \pm \sqrt{(\omega_c^2 + \omega_y^2 + \omega_z^2)^2 - \omega_y^2\omega_z^2} \quad (7.33)$$

For $\omega_y = \omega_z$, Eq. (7.33) becomes the Fock–Darwin spectrum [64, 99]. The calculated spectra agree very well with the experimental result for $\omega_y = 2$ meV and $\omega_z = 5$ meV.

7.2.4

The “0.7 structure”

In [299] a very clearly pronounced additional conductance plateau at $G \approx 0.7 \times (2e^2/h)$ has been reported (the “0.7 structure” or “0.7 feature”). Subsequent experiments have confirmed this observation and studied its parametric behavior. Not only does this plateau remain unexplained within our model, it has some additional puzzling features (see Fig. 7.14), namely:

- as the temperature is increased, it does not suffer from thermal smearing, but becomes more prominent instead;
- it emerges from the spin-split plateau at $G = e^2/h$ as a strong parallel magnetic field is reduced;
- its presence or absence seems to depend randomly on the sample.

In fact, the origin of this plateau has not yet been clarified unambiguously, although there is strong evidence that it is caused by electronic correlation effects known as Kondo correlations [62].

7.2.5

Four-probe measurements on ballistic quantum wires

Measuring the resistance of a ballistic QWR without the contact resistance clearly requires voltage probes to be attached in between the two contacts. This has to be done without disturbing the current flow. For example, if the electrons get backscattered at the probes along their trip from source to drain, the resistance will be increased. In other words, the resistance R_{wp} between the wire and the probe has to be large compared to R_{SD} . Such an experiment has been performed on quantum wires defined by cleaved edge overgrowth (see Chapter 5). The sample layout is reproduced in Fig. 7.15. With all gates grounded, a wire extends along the cleaved edge, which is coupled to the 2DEG that resides in the quantum well. The length scale for scattering between the wire modes and the states in the 2DEG is $\ell_{2D-1D} \approx 6 \mu\text{m}$.

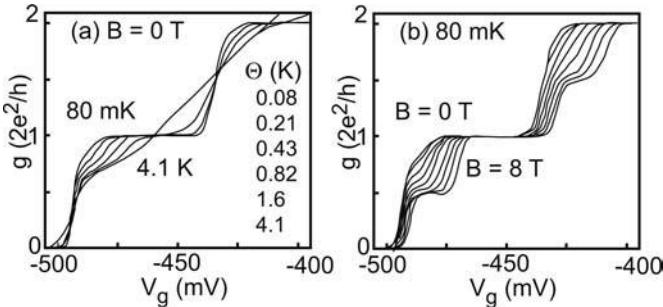


Fig. 7.14 Conductance of a QPC as a function of the split-gate voltage. (a) The “0.7 structure” shows a different temperature dependence than the conventional conductance steps. (b) In strong parallel magnetic fields, the spin degeneracy is removed and additional plateaus are visible at odd integers of e^2/h . As B is reduced, the spin-split plateau at $G = 0.5e^2/h$ evolves into the “0.7 structure”. After [62].

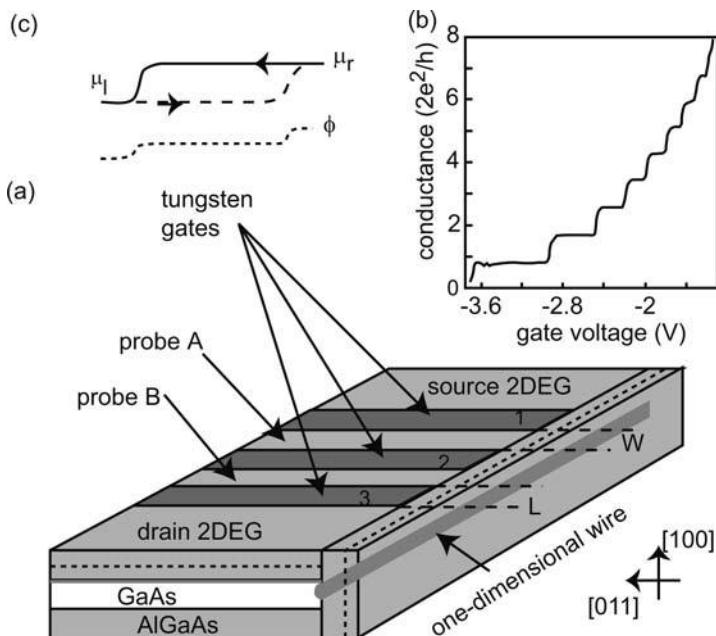


Fig. 7.15 (a) Sample geometry for four-terminal resistance measurements on a ballistic quantum wire. A quantum well is grown in the [100] direction, and three tungsten gate stripes of width $W = 2 \mu\text{m}$, separated by $L = 2 \mu\text{m}$, are evaporated on top. The wafer is then cleaved, and a modulation-doped layer of $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ is grown along the [110]

direction. The wire extends along the cleave. (b) Conductance quantization as a function of the voltage at gate 2, with the other gates grounded. (c) Spatial variation of the electrochemical potentials of left- and right-moving electrons along the wire, as discussed in the text. In addition, the corresponding potential drop ϕ is shown. After [241].

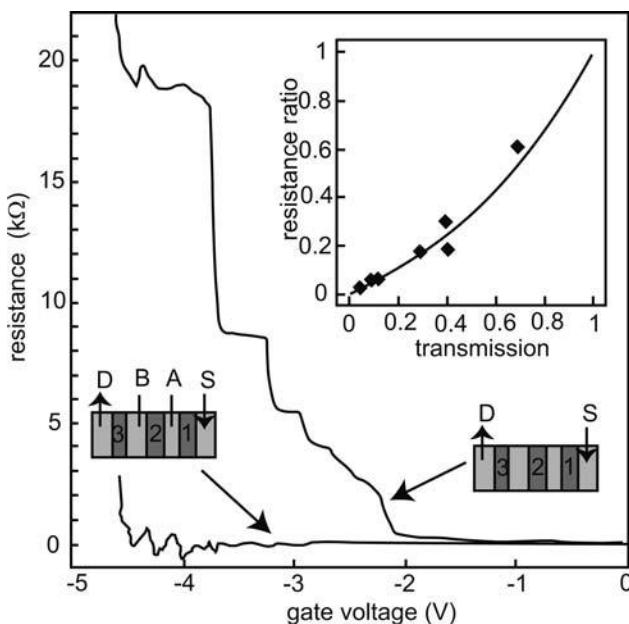


Fig. 7.16 Main figure: The two-terminal resistance of the wire (upper trace) along gate 2 and its four-terminal resistance (lower trace) are compared. The inset shows the ratio between the four-terminal resistance and the two-terminal resistance as a function of the probability T for electrons to be transmitted between the wire and the probe. The lower trace in the main figure has been performed at $R_{wp} = 250 \text{ k}\Omega$, i.e. $T \approx 0.05$ in the inset. After [241].

Negative voltages applied to the tungsten gate stripes deplete the 2DEG underneath. The ballistic wire of length $L = 2 \mu\text{m}$ can be tuned by activating gate 2, while gates 1 and 3 remain grounded. In this operation mode, the 2DEG areas to the left and right of gate 2 serve as source and drain. Clear resistance quantization is observed (upper trace in Fig. 7.16; this trace is equivalent to the conductance trace in Fig. 7.15). The plateaus deviate from the expected $G = je^2/h$ by up to 25%, an effect that has its origin in non-ideal contacts to source and drain in this particular sample geometry. By additionally activating gates 1 and 3 to appropriate voltages, strictly one-dimensional leads are generated along the corresponding regions of the wire. Also, the 2DEG regions between the gates now form two voltage probes that couple weakly to the wire, since their width is smaller than ℓ_{2D-1D} . Clearly, the voltage probes are located in between the contact regions of the wire to source and drain, and the measured voltage difference between A and B in Fig. 7.16 should be zero. This is in fact the case for all plateaus, except close to pinch-off around a gate voltage of -4.6 V .

These measurements have confirmed in a beautiful way the model of contact resistances and ballistic transport in one-dimensional systems. There is actually a quite different kind of experiment which proves this as well, namely four-terminal resistance measurements on 2DEGs in the quantum Hall regime. This is the topic of the following sections.

7.3

The Landauer–Büttiker formalism

In the previous section, we have argued that the quantized conductance of ballistic quantum wires stems from contact resistances. We have also seen that four-probe measurements give a resistance of zero, as expected from their interpretation in terms of contact resistances. In fact, a conceptually very similar system is a 2DEG in the quantum Hall regime. As already indicated in Section 6.2, the electrons skip along the edge of the Hall bar in strong magnetic fields. The origin of this dynamics is illustrated in Fig. 7.17.

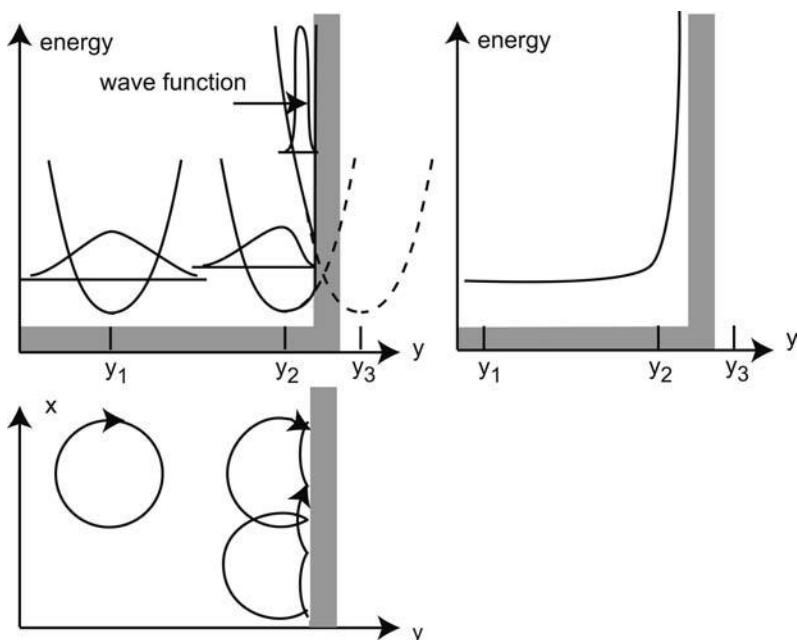


Fig. 7.17 Modification of the magnetoelectric confinement of the electrons as they approach the edge of the 2DEG (top left). The undisturbed cyclotron motion at y_1 is increasingly squeezed as the guiding center approaches the edge (positions y_2 and y_3). As a consequence, the energy of the Landau level increases (right), while the electrons delocalize along the x -direction (bottom).

7.3.1

Edge states

At the edge of the 2DEG, the conduction band bottom increases sharply and modifies the combined potential of the Landau harmonic oscillators and the electrostatic confinement. The increased confinement shifts the Landau levels to higher energies. Each LL crosses the Fermi level at some point, and consequently the density of states at the Fermi level is always larger than zero. As sketched in Fig. 7.17 as well, the electrons skip along the edge. Therefore, we speak of *skipping orbits* and *edge states*. Edge states have several peculiar features, which become self-evident immediately. First of all, they are one-dimensional: the electron motion is confined perpendicular to the sample edge, but is free in the direction parallel to it. Second, all the electrons at one sample edge move in the same direction, while the electrons at the opposite edge move in the opposite direction. In the bulk, all electrons are localized at potential modulations, except for special filling factors, as already shown in Section 6.2. The resulting edge state configuration with the directions of current flow is shown in Fig. 7.18.⁶

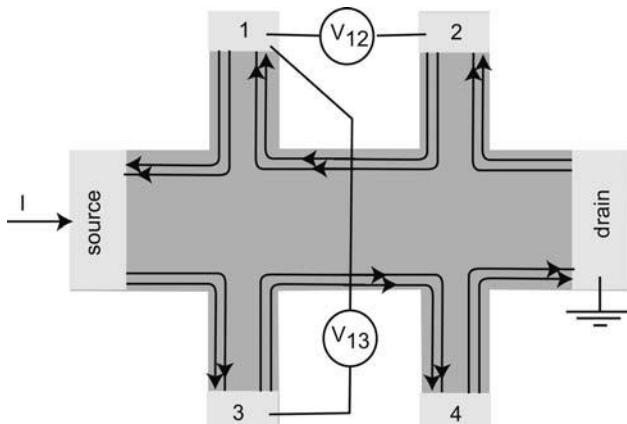


Fig. 7.18 Top view of a Hall bar in a strong magnetic field. Current flows in one-dimensional edge states only, in the directions indicated by the arrows. Here, two Landau levels are occupied.

There is no backscattering in edge states, i.e. the elastic mean free path approaches infinity. Suppose an electron in an edge state hits a scatterer close to the edge. Its momentum right after the scattering event may be reversed, but the strong magnetic field bends the momentum back into the forward direction. In order to be backscattered, the electron has to traverse the whole Hall bar and reach the opposite edge! Hence, backscattering is greatly reduced. It

6) Since the electrons circulate around the edge, one speaks of a *chiral* Fermi liquid.

follows that a 2DEG in the quantum Hall regime comes very close to an ideal ballistic quantum wire: it is one-dimensional and backscattering is absent. We can even attach voltage probes inside the quantum wires without inducing backscattering. Therefore, the voltage drop between, for example, contacts 1 and 2 in Fig. 7.18 should be zero. You will have realized, of course, that this is exactly what we measure in a Shubnikov–de Haas experiment. In [47], the Landauer formula has been generalized to an arbitrary number of contacts, such that circuits of ballistic quantum wires can be treated. The concept is known as the Landauer–Büttiker formalism.

Consider a circuit of ballistic quantum wires, like, the system of Fig. 7.18. We define the *direct* transmission probability of contact p into contact q as $T_{q \leftarrow p} = T_{qp}$. It is possible to have $T_{qp} > 1$, since more than one mode may connect the two contacts. Note that T_{qp} does not have to be an integer. Note further that, within this notation, T_{pp} is a backscattering probability. The *total current emitted by contact p* is denoted by I_p , while μ_p is the electrochemical potential of contact p . Again, an “ideal” contact absorbs all incoming electrons and distributes the emitted electrons equally among all outgoing modes, such that they are filled up to μ_p , assuming zero temperature.

In this notation, the Landauer formula generalizes to the Büttiker formula

$$I_p = \frac{2e}{h} \sum_q (T_{qp}\mu_p - T_{pq}\mu_q) \quad (7.34)$$

which is a direct consequence of current conservation. We proceed by applying the Büttiker formula to the sample shown in Fig. 7.18. It gives a system of six linearly dependent equations, one for each contact:

$$\begin{pmatrix} I_s \\ I_d \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{pmatrix} = \begin{pmatrix} G & 0 & -G & 0 & 0 & 0 \\ 0 & G & 0 & 0 & 0 & -G \\ 0 & 0 & G & -G & 0 & 0 \\ 0 & -G & 0 & G & 0 & 0 \\ -G & 0 & 0 & 0 & G & 0 \\ 0 & 0 & 0 & 0 & -G & G \end{pmatrix} \begin{pmatrix} V_s \\ V_d \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix}$$

with $G = j(2e^2/h)$. By choosing $\mu_d = 0$ as a reference potential, and after eliminating the drain current as a consequence of current conservation (remember that the voltage probes measure the potentials without drawing current), we can eliminate the drain row and column, and the following matrix equation results:

$$\begin{pmatrix} I_s \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{pmatrix} = \begin{pmatrix} G & -G & 0 & 0 & 0 \\ 0 & G & -G & 0 & 0 \\ 0 & 0 & G & 0 & 0 \\ -G & 0 & 0 & G & 0 \\ 0 & 0 & 0 & -G & G \end{pmatrix} \begin{pmatrix} V_s \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix}$$

Its solution gives

$$V_s = V_3 = V_4, \quad V_1 = V_2 = 0, \quad I_s = GV_s \quad (7.35)$$

a result that you may have guessed, considering, for example, that probes 1 and 2 are resistanceless connected to drain. Therefore, we find the longitudinal resistance

$$R_{xx} = \frac{V_1 - V_2}{I_s} = \frac{V_3 - V_4}{I_s} = 0 \quad (7.36)$$

and the Hall resistance

$$R_{xy} = \frac{V_3 - V_1}{I_s} = \frac{V_4 - V_2}{I_s} = \frac{h}{2je^2} \quad (7.37)$$

Within the edge state picture, the quantized Hall resistance is obtained, and the longitudinal resistance vanishes. The accuracy of the quantization is so much more accurate than in a QPC because backscattering is greatly suppressed. Let us now consider what happens as we increase the magnetic field, such that the uppermost occupied LL gets depleted. The corresponding edge state, which is the innermost occupied one, is depopulated as well. Since the velocity in the x -direction of the electrons in edge state j is given by

$$v_j(k_x) = \frac{1}{\hbar} \frac{\partial E_j(k_x)}{\partial k_x} = \frac{1}{\hbar} \frac{\partial V(y(k_x))}{\partial y} \frac{\partial y}{\partial k_x} = \frac{1}{eB} \frac{\partial V(y)}{\partial y} \quad (7.38)$$

where we have used $y(k_x) = \hbar k_x / eB$, the velocity of the electrons approaches zero as the edge state gets depleted. As a consequence, the edge state begins to soften and the electron trajectories penetrate into the bulk. Finally, the electrons can percolate all the way to the opposite edge, backscattering sets in, and the conductance quantization vanishes.

Haug et al. [145] have performed an instructive experiment related to this picture (see Fig. 7.19). A gate stripe extends across a Hall bar inside an area that can be measured by four voltage probes. Biasing the gate tunes the electron density, and thus the number of occupied Landau levels, underneath. If the filling factor under the gate is smaller than outside the gated area, edge states get redirected at the gate. This changes the transmission probabilities in Eq. (7.34).

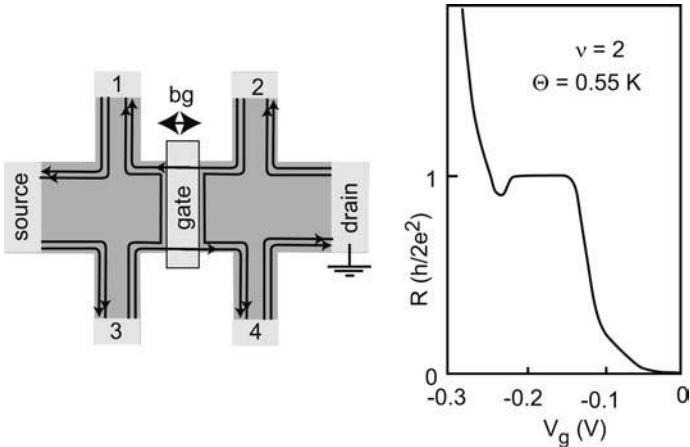


Fig. 7.19 Left: Sample geometry to control backscattering between edge states. A top gate covers the Hall bar in between four voltage probes. At suitable gate voltages, the inner one of the two edge states gets reflected. Right: For a 2DEG in the regime of filling vector 2, with spin-split edge states, a plateau

at $R_{xx} = h/2e^2$ is observed as a function of the gate voltage, once the reflection of the inner edge state at the gate is complete. After [145]. The dip around a gate voltage of -0.2 V can be explained within a trajectory network formed below the gate.

In Exercise E7.3 the resistances of this system will be calculated. The result for filling factor N in the ungated region and M in the gated region is

$$\begin{aligned} R_{12} &= R_{34} = \frac{h}{e^2} \left(\frac{1}{M} - \frac{1}{N} \right) \\ R_{13} &= R_{24} = \frac{h}{e^2} \frac{1}{N} \\ R_{14} &= \frac{h}{e^2} \left(\frac{1}{M} - \frac{2}{N} \right) \\ R_{23} &= \frac{h}{e^2} \frac{1}{M} \end{aligned} \quad (7.39)$$

Note that the results of some measurements now depend on the direction of the magnetic field.

The Landauer–Büttiker formalism is a powerful tool, which allows to treat a variety of problems very elegantly. Further examples are treated in the exercises.

7.3.2

Edge channels

So far, we have interpreted edge states as guiding centers of electron trajectories in strong magnetic fields. Within this picture, the trajectories of electrons moving in different edge states intersect, and we may expect a strong inter-

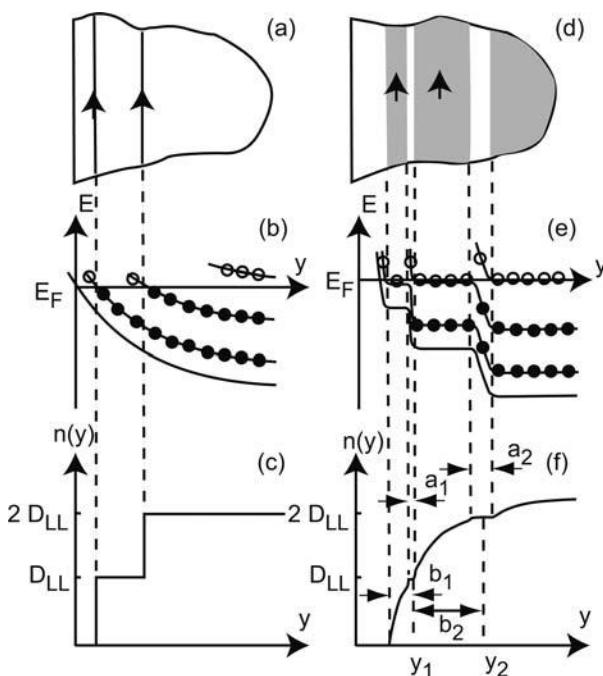


Fig. 7.20 (a) Guiding center trajectories within the edge state picture, for the case of filling factor 2. Along these lines, the system is metallic, while it is insulating everywhere else. (b) The corresponding energies of the Landau levels. Full circles denote occupied states. (c) The resulting spatial variation of the electron density. (d) In the edge channel picture, the potential gets screened in the

metallic regions, and (e) the potential drop concentrates within the insulating regions. The edge states evolve into metallic stripes of non-zero width, separated by insulating stripes. The stripe width is determined by the electrostatics of the configuration. (f) The resulting electron density close to the edge. After [53].

edge state scattering rate. In fact, the edge states are spatially separated in sufficiently strong magnetic fields, and inter-edge state scattering is suppressed.⁷ This can be understood by studying the effects of screening at the edge (see Fig. 7.20). At points where the edge states intersect the Fermi level, the system has a metallic character. Here, the electrons in the edge state are able to screen the confining potential, and edge channels are formed. The potential drop is concentrated in the insulating regions. The electrostatics of edge states, which is the topic of Paper P7.3, was considered first in [53]. Note that just before the innermost occupied edge channel gets emptied, its width approaches infinity and extends all the way to the opposite edge. As within our picture of the previous section, backscattering becomes possible under these circumstances.

7) Note that this kind of scattering does not show up in the resistance, unless special geometries are considered, like that of Exercise E7.4.

7.4

Further examples of quantum wires

Conductance quantization was first detected in QPCs defined in 2DEGs by gate voltages, and the vast majority of experiments have been performed on such systems. But these results have also triggered the search for similar effects in other materials, in particular conventional metals and carbon nanotubes. Even quantized transmission of light through a pinhole has since been discovered [214].

7.4.1

Conductance quantization in conventional metals

Once the behavior of QPCs in semiconductors was known, observing conductance quantization turned out to be possible in a very simple experiment: Just pull a metallic wire and measure its resistance simultaneously. Right before it breaks, you will observe quantized conductance. This experiment is performed in several undergraduate lab courses, and you can even do it on your kitchen table using a household wire [60].

Most of these experiments, however, are done in a more controlled setup. For example, a thin metal wire is mounted as the tip of a scanning tunneling

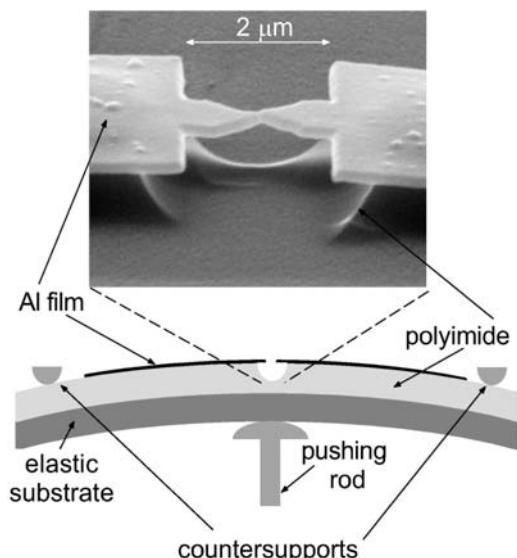


Fig. 7.21 A mechanically controlled break junction for observing conductance quantization in an Al QPC. The elastic substrate is bent by a pushing rod with a piezoelectric element. The thin Al bridge, fabricated by electron beam lithography, can be broken and reconnected for many cycles. Taken from [263].

microscope (STM) and pushed toward a metal surface. Fig. 7.21 shows a mechanically controlled break junction, another widely used setup. Fig. 7.22 reproduces a typical experimental current trace as a function of time over which the junction is deformed. Clear conductance steps can be observed, although they do not necessarily have the “right” values. This is attributed to the details of the breaking process: possibly, several QPCs are generated in parallel. Disorder in the junction may modify the plateau values as well. Conductance histograms taken over many cycles of breaking and reconnecting the junction, however, show that the conductance is predominantly quantized in units of $\approx 2e^2/h$. The details of these experimental results contain a lot of information. To a crude approximation, it may be assumed that, right before the wires break, the current is carried via a single atom. The degeneracies of the conducting modes of the atomic junction can be material-specific, as has been demonstrated, e.g. in [196].

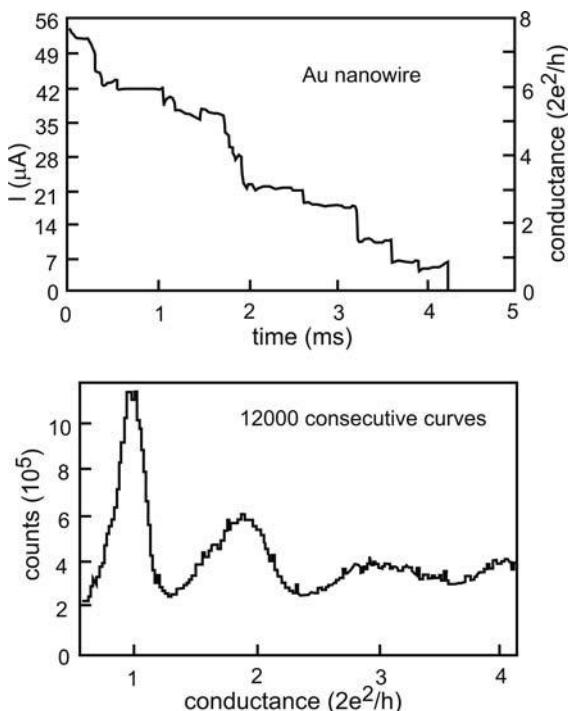


Fig. 7.22 Top: Conductance as a function of the deformation time of a gold junction (an STM setup was used). The observed steps are usually not quantized in units of $2e^2/h$. Bottom: A histogram of many consecutive sweeps of the upper type, however, reveals that steps of the expected height dominate. Adapted from [61].

7.4.2

Molecular wires

The possibilities of molecular design are absolutely amazing. The synthesis techniques are of utmost importance for a tremendously wide range of applications, ranging from, say, foams for thermal isolation through novel medications to applications in semiconductor processing, such as photo- or electron beam resists. Should it not also be possible to design molecules with an electronic functionality relevant for applications? It is in fact possible, and a research branch named *molecular electronics* has emerged from this question. In its broadest sense, it spans all electronic applications of materials composed of molecules, including, for example, organic semiconductors like those presented in Chapter 3. It has already been demonstrated that elementary operations like switches [75] or memory cells [59, 251] can be made from single molecules attached to wires. In a different kind of application, molecules are also used as templates for the preparation of metallic wires [339] and even more complicated structures like quantum interference devices [152].

Here, we consider individual objects that can be regarded as single-molecule quantum wires. The most prominent and best studied example is carbon nanotubes, provided we classify them as “molecules”. Carbon nanotubes are comparatively large and can therefore be processed in relatively straightforward ways. As a downside, they do not represent an absolute size limit, since other suitable, typically aromatic, molecules are still smaller. However, these systems typically do not show the behavior of quantum wires, but instead that of quantum dots; they are therefore discussed in Chapter 10.

7.4.2.1 Carbon nanotubes

Carbon nanotubes (CNs) have enjoyed wide popularity since 1992. They have unique structural, mechanical, and electronic properties, which are treated in several excellent books, as well as in review articles (for references, see the further reading at the end of this chapter). Before we look at the quantum wire aspects of CNs, we should define what we actually mean by a *carbon nanotube*.

We have seen in Chapter 2 that graphite consists of two-dimensional, weakly coupled sheets of honeycomb lattices. By laser ablation, for example, it is quite easy to produce individual graphite sheets. In 1991, it was discovered that, under certain experimental conditions, these sheets roll up and form hollow carbon cylinders (carbon nanotubes, CNs) with diameters of a few nanometers only [160]. Their length, however, can be many micrometers. One distinguishes between single-walled and multi-walled CNs, depending on the number of concentric carbon cylinders. Of particular interest are single-walled CNs, since these well defined systems can be treated theoretically to a

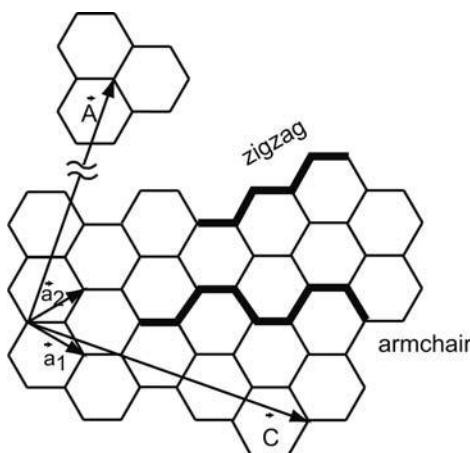


Fig. 7.23 Scheme of a sheet of graphite used to roll up a CN.

considerable depth. In Fig. 7.23, a graphite sheet used to form a CN is shown. The *chirality vector* \vec{C} defines a stripe that is cut out of the 2D lattice. The CN is formed by connecting the edges of this stripe, located at the bottom and the top of \vec{C} . Hence, $|\vec{C}|$ is the diameter of the CN. It has become common practice to characterize the structure of a CN by the coordinates (n_1, n_2) of \vec{C} with respect to the lattice vectors \vec{a}_1 and \vec{a}_2 of the graphite sheet. For example, the CN shown in Fig. 1.6 is a $(10, 5)$ tube, which means that $\vec{C} = 10\vec{a}_1 + 5\vec{a}_2$. Note that it suffices to consider tubes with $0 < n_2 < n_1$.

For apparent reasons, CNs with cross sections along the bold lines are called “zigzag” and “armchair” tubes, respectively. The elementary lattice vector of the resulting CN, which is perpendicular to \vec{C} in the graphite sheet, is denoted by \vec{A} . Its length is the minimal distance for which top and bottom see identical environments. The additional periodic boundary condition modifies the electronic band structure. Clearly, a CN is a quasi-one-dimensional system. It is furthermore obvious that the wave vector k_y perpendicular to the tube axis gets quantized, such that one-dimensional modes emerge from the two-dimensional energy dispersion of the graphite sheet (Fig. 2.3). Depending on the direction of the CN axis (parallel to \vec{A}) and the circumference of the CN, the k_x of one mode may or may not hit a K-point of the graphite sheet’s Brillouin zone (see Fig. 7.24), and hence a metallic or a semiconducting CN results. It can be shown that CNs are metallic if $(2n_1 + n_2)$ is an integer multiple of 3 (see e.g. the further reading at the end of this chapter). The energy dispersion of these two classes of CNs are reproduced in Fig. 7.25.

The calculated densities of states (Fig. 7.26 gives an example) agree very well with experimental results [224,328] obtained by scanning tunneling spectroscopy. The quasi-one-dimensional character is apparent. The mode separa-

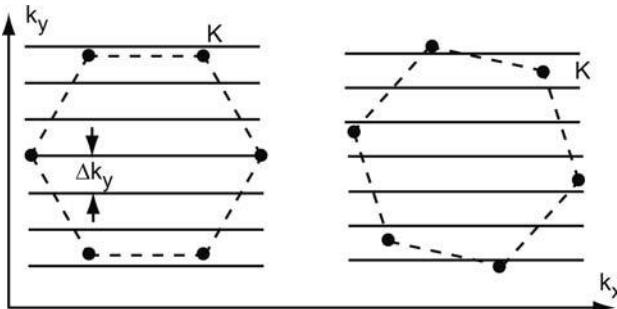


Fig. 7.24 By periodic boundary conditions in the y -direction, the Brillouin zone of the graphite sheet (dashed hexagon) gets partly quantized, and one-dimensional modes (lines) in the x -direction, i.e. along the CN axis, result. Left: If some K-points fall on the modes, the CN is metallic. Right: The 1D modes miss the K-points. The CN is semiconducting.

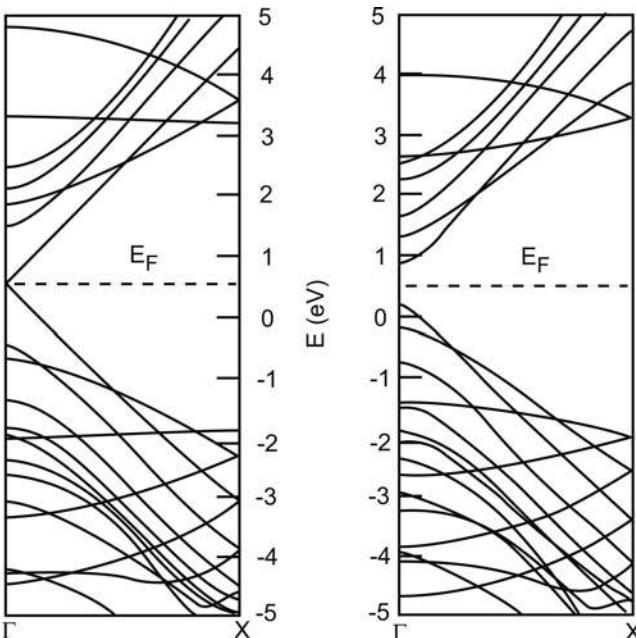


Fig. 7.25 Band structures of (a) metallic and (b) semiconducting CNs. After [138].

tion is of the order of 100 meV, much larger than in lithographically patterned QWRs.

Making low-resistance contacts to CNs has turned out to be very difficult. Such samples are typically prepared by depositing CNs on an insulating substrate that contains some metallic electrodes. By chance, a CN will make con-

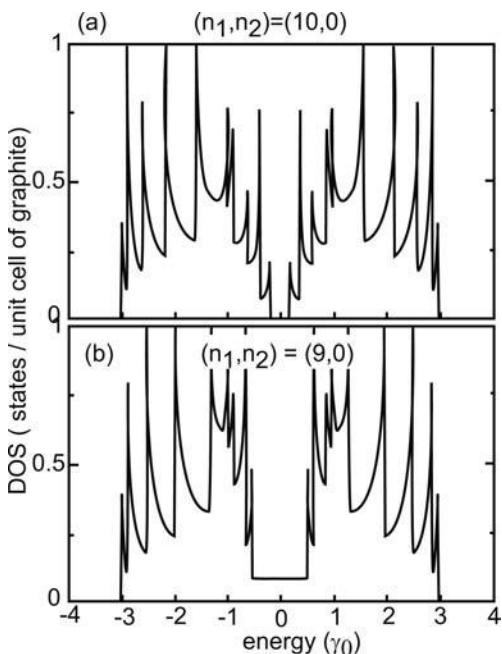


Fig. 7.26 The calculated density of states of (a) semiconducting and (b) metallic carbon nanotubes. The Fermi energy is at $E = 0$. The bandgap equals 0.7 eV in the semiconducting CN. Adapted from [257].

tacts with two or more electrodes, and multi-terminal transport measurements become possible. Usually, the contact between, say, a gold electrode and a CN in such a sample is a tunnel barrier. It has been shown that the contact resistance can be reduced by electron beam irradiation of the contact region [18]; also, the contact becomes much better when the electrodes are patterned *on top of* the deposited CNs. In neither of these setups, however, could conductance steps be observed. An experiment that demonstrates conductance quantization in multi-walled CNs has been performed in [103]: the authors immersed a CN attached to the tip of an STM into liquid mercury, and observed conductance steps as a function of the tip position at room temperature.

In metallic CNs, there are two spin-degenerate modes at the Fermi level, and we expect a quantized conductance of $G = 4e^2/h$, in the case when a single cylinder of a multi-walled CN couples to the reservoir. Instead, steps of height $2e^2/h$ are observed. The origin of this discrepancy has remained unexplained. Possibly, the spin degeneracy is lifted by electron-electron interactions.

7.5

Quantum point contact circuits

7.5.1

Non-Ohmic behavior of QPCs in series

Combinations of QPCs offer a variety of experimental options. Even the most elementary one, namely just two QPCs, already raises interesting questions. In Fig. 7.27, the resistance of two QPCs in series, with a separation much smaller than the elastic mean free path, is shown as a function of the voltages applied to the two split gates. It is immediately apparent that Ohm's law is violated: the series resistance is much smaller than the sum of the individual QPC resistances. We denote the individual conductances of QPC_1 and QPC_2 by G_1 and G_2 , respectively. The series resistance of both QPCs between the injector and the collector is denoted by G_{ic} . Experimentally, $G_{\text{ic}} \approx \min\{G_1, G_2\}$ was found in this experiment. Apparently, the QPC that is narrower determines the total resistance, while the electrons pass the second one with no or little further resistance. One is therefore tempted to guess that the contact resistance of the second QPC is strongly reduced.

This behavior can be explained by the ballistic character of the electron motion in between the QPCs, in combination with the adiabatic coupling mentioned in Section 7.2.2. Suppose the electrons exit the QPC in an adiabatic fashion, i.e. they remain in the mode they used to pass through the QPC. As the width of the constriction gets wider, the energy of the mode, and with it

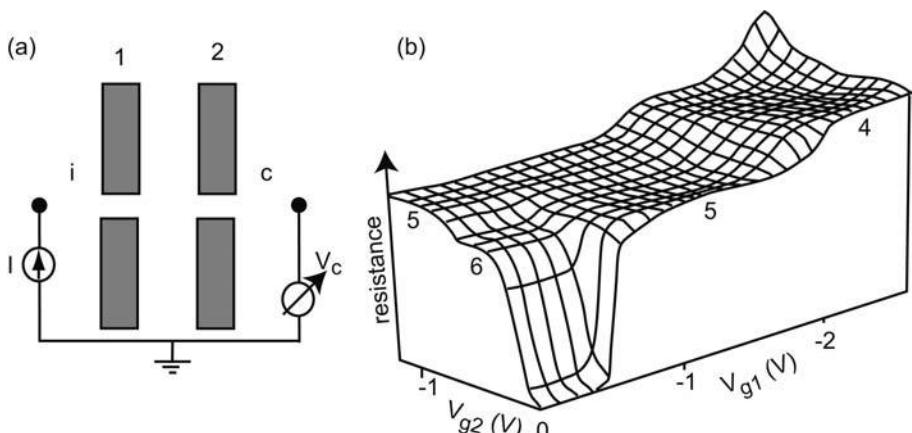


Fig. 7.27 (a) Sketch of two QPCs in series. (b) The resistance as a function of the voltages V_{g1} and V_{g2} applied to the split gates 1 and 2. The behavior is non-ohmic, and indicates that the total resistance roughly equals the individual resistance of the QPC with fewer occupied modes. The temperature was about 300 mK. Adapted from [327].

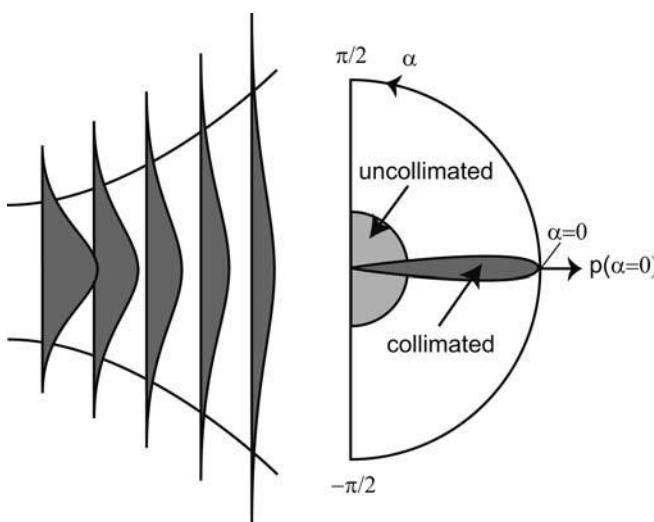


Fig. 7.28 Adiabatic spreading of the wave function belonging to the lowest mode of a QPC at its exit. As the channel gets wider, the transverse electron momentum k_y is reduced. The result is sketched to the right: compared to the uncollimated beam, the probability density $p(\alpha)$ for the electrons to be emitted in the x -direction peaks around $\alpha = 0$.

the transverse momentum p_y , gets reduced. Consequently, the electrons in that mode get *collimated* as they exit the first QPC: the ejection angle is smaller than $\pm\pi/2$, while the angular probability distribution of the electrons in the 2DEG reflects the sum of the probability densities of the QPC modes that carry the current (see Fig. 7.28). This spatial distribution has been verified experimentally by an ingenious experiment described in [304]. Thus, the electrons more or less already have the correct transverse momentum needed for entering the second QPC.

To be more quantitative, we once again use the Landauer–Büttiker formula. There is a direct transmission of electrons from the injector to the collector T_{ic} . However, since, in the setup of Fig. 7.27, the central region in between the QPCs floats and the collector is grounded, current conservation tells us that the injected flow of electrons will either escape back into the injector or finally make it into the collector, after some scattering events that will take place far away from the QPC region. If we assume that the middle area is large enough for equilibrating these electrons, we can speak of the chemical potential of the middle μ_m . Thus, either electrons are transmitted directly from the injector into the collector, or they get absorbed and re-emitted by an effective reservoir at potential μ_m .

For the sake of simplicity, let us study the case of equal QPC conductances, i.e. $G_1 = G_2 = N \times 2e^2/h$. The Landauer–Büttiker equations then read

$$\begin{aligned}\frac{h}{2e^2}I_i &= NV_i - T_{mi}V_m - T_{ci}V_c \\ \frac{h}{2e^2}I_c &= NV_c - T_{mc}V_m - T_{ci}V_i \\ 0 &= [N_m - (1 - T_{mm})]V_m - T_{mi}V_i - T_{mc}V_c\end{aligned}\tag{7.40}$$

If we assume $T_{mi} = T_{mc}$ and use $I_i = -I_c$, addition of the first two equations of (7.40) gives

$$R_{ic} = \frac{V_i - V_s}{I_i} = 2 \frac{h}{2e^2} \frac{1}{(N + T_{ic})}\tag{7.41}$$

This tells us that the series resistance is just the resistance of one QPC if $T_{ic} = N$, when all the electrons are directly transmitted. If $T_{ic} = 0$, the resistances follow Ohm's law.

If the electrons were ejected from the first QPC with equal probability into all directions, only a fraction $\approx w/(2\pi L)$ would be directly transmitted for $w \ll L$. Here, L is the separation of the QPCs, and w is the effective QPC width.

That the collimation effect actually determines the non-ohmic addition of QPC resistances in series can be seen by studying T_{ic} as a function of magnetic field. We expect that the collimated electron beam pattern gets deflected due to the Lorentz force, and T_{ic} should show a strong peak at $B = 0$, haloed by smaller side peaks at non-zero magnetic fields for QPCs with more than just one conducting channel. Such experiments have been performed by Molenkamp et al. [212] and by Shepard et al. [273].

7.5.2

QPCs in parallel

The resistance of QPCs in parallel in a ballistic circuit (see Fig. 7.29(a)) is a periodic function of the magnetic field. The electrons ejected from the first QPC are forced on cyclotron orbits. For the correct polarity of the magnetic field, the deflected electron beam is directly injected into the second QPC, provided the separation between the QPCs is an integer multiple j of the cyclotron diameter. A fraction of the electrons thus gets caught at the QPCs and circulates around the separating barrier. Consequently, the resistance shows maxima for $jr_c = s$, where s is the separation. This can be nicely seen in Fig. 7.29(b), where these resistance maxima have been measured in a one-dimensional array of 43 QPCs in parallel, with a spacing of $s = 4.6 \mu\text{m}$. Since the electron density was

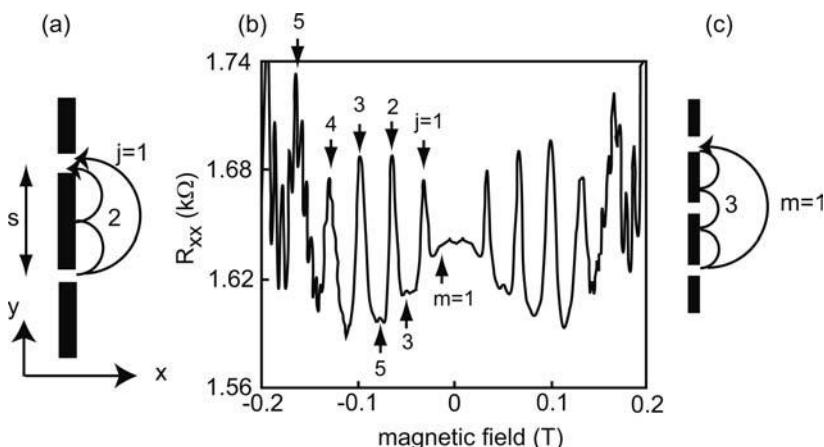


Fig. 7.29 The magneto-resistance of QPCs in parallel (a) shows periodic peaks (b), due to commensurability between the cyclotron orbits and the QPC spacing s (a, c). Adapted from [217].

$n = 2.2 \times 10^{15} \text{ m}^{-2}$, a peak spacing of

$$\Delta B = \sqrt{8\pi n} \frac{\hbar}{es} \approx 34 \text{ mT}$$

is observed. Note that the peak resistance is significantly larger than the resistance at $B = 0$, while it is known that the resistance of an individual QPC drops as B is increased – see Paper P7.2. Owing to the arrangement of many QPCs in parallel, additional resonances are found in Fig. 7.29, which are labeled by m and correspond to trajectories that obey the resonance condition for next-nearest-neighbor QPCs.

In this apparently simple explanation, two crucial assumptions have been implicitly made. First of all, the scattering at the separating barrier must have a large specularity. Here, the barriers have been defined by a shallow etch of the Ga[Al]As surface, which gives a specularity of 1, to a good approximation [217]. Second, despite the collimation effect, the electrons are not strictly ejected in the x -direction from the first QPC. However, a simple geometric consideration shows that electrons that exit the QPC at an arbitrary angle in the x -direction all get focused at separations $\delta y = j\sqrt{8\pi n}/(eB)$, where δy denotes the distance of the focus from the QPC in the y -direction. Further details of this *magnetic electron focusing*, which is well known from normal metals as well as from charged particles in vacuum tubes, can be found in Paper P7.4.

7.6

Semiclassical limit: Conductance of ballistic two-dimensional systems

Armed with the Landauer–Büttiker formalism, we are now in a position to discuss how the conductance of a ballistic, quasi-two-dimensional sample can be treated. Recall that, at liquid helium temperatures, the elastic mean free path in 2DEGs can easily become as large as $10\text{ }\mu\text{m}$, which allows us to define rather complex ballistic electronic circuits. But how shall we model this? Clearly, the Boltzmann model is of no help: it is based on momentum relaxation by random scattering. On the other hand, we have just seen how, in quasi-1D systems, the conductance is obtained from transmission probabilities. As the QWR width becomes wider, it approaches a 2D sample, and the number of occupied modes becomes very large. They will no longer be resolvable at some point, since their natural width in energy is larger than the mode spacing. Nevertheless, there is no fundamental reason why the Landauer–Büttiker formalism should not be applied to such a quasi-2D system. This consideration gives us a powerful tool to analyze ballistic, two-dimensional structures: (i) determine the number of modes from the sample width and the electron density, (ii) calculate the transmission from lead p into lead q by an appropriate method (e.g. by classical simulations), and (iii) insert both into Eq. (7.34). This approach goes back to [26], where it was applied to calculate the components of the resistance tensor of ballistic crossovers.

Here, we demonstrate this technique by studying a simple yet interesting example known as a *magnetic barrier* [188], which is a strongly localized magnetic field peak in the transport direction (x -direction), oriented perpendicular to the plane of the 2DEG (see Fig. 7.30). A magnetic barrier can be prepared by a ferromagnetic disk on top of a Hall bar, which is magnetized in the x -direction by an in-plane magnetic field $B_{||}$. The fringe field can be approximated by a magnetic dipole field. At the location of the 2DEG, its z -component is strongly localized, with extremal points below the edge of the disk (Fig. 7.30(a)). We call one of the two peaks in $B_z(x)$ the magnetic barrier. That one centered at $x = 0$ is given by

$$B_z(x) = \frac{\mu_0 M(B_{||})}{4\pi} \ln \left(\frac{x^2 + z_0^2}{x^2 + (z_0 + h)^2} \right) \quad (7.42)$$

where $\mu_0 M$ is the magnetization and h is the thickness of the ferromagnetic disk, while z_0 denotes the distance of the 2DEG from the surface (Fig. 7.30(b)). We can measure the barrier resistance by supplying a current in the x -direction and probing the voltage drop at contacts 1 and 2 [182].

As $|B_{||}|$ is increased, an increase in R_{xx} is observed in Fig. 7.31(a), which is very steep around $B_{||} = 0$ and saturates at larger magnetic fields. The minimum is not exactly at $x = 0$, which is a consequence of the hysteretic behavior of the magnetization. This measurement is readily understood qual-

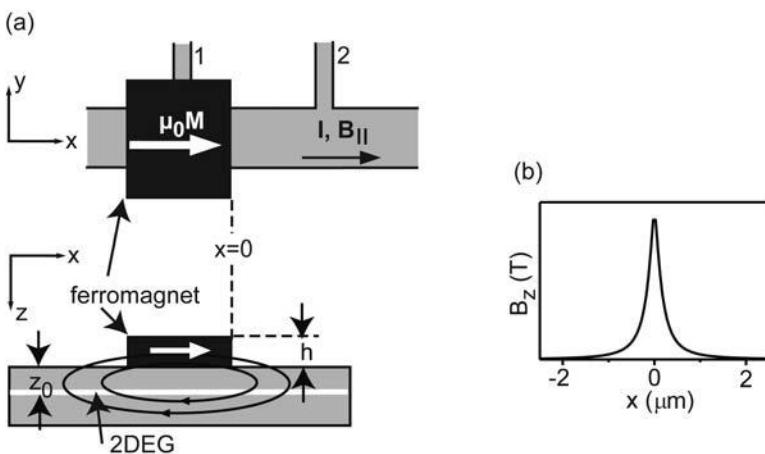


Fig. 7.30 Experimental realization of a magnetic barrier. (a) Top view and cross section. (b) Shape of the barrier $B_z(x)$ that emerges around $x = 0$ according to Eq. (7.42).

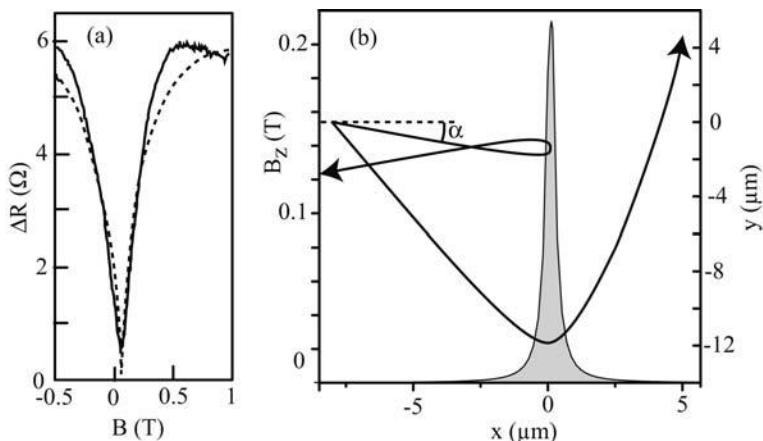


Fig. 7.31 (a) Measurement of a magnetic barrier resistance as a function of B_{\parallel} (full line), in comparison with the calculation obtained from the Landauer–Büttiker formalism (dashed line). Adapted from [310]. (b) Trajectories of electrons injected toward the magnetic barrier (gray area) at different angles $\alpha_0 = 15^\circ$ and 60° , respectively.

itatively: As the electrons enter the magnetic barrier region, they experience a Lorentz force with an x -dependent cyclotron orbit. Depending on the angle of incidence and on the magnetic barrier height, i.e. on B_{\parallel} , they get transmitted or reflected (see Fig. 7.31(b)). Once the magnetization of the ferromagnet is saturated, B_{\parallel} has no further effect on the magnetic barrier, and R_{xx} remains constant.

We proceed by describing the resistance of the barrier with the adapted Landauer–Büttiker formalism. For simplicity, we consider a two-terminal configuration and assume that the resistance in series with the magnetic barrier is negligible. For this scenario, Eq. (7.34) reads

$$\frac{h}{2e}I = T(\mu_S - \mu_D) \implies G = \frac{2e^2}{h}T \quad (7.43)$$

We have a large number of modes N , and express T as

$$T = N\langle T \rangle \quad (7.44)$$

where $\langle T \rangle$ denotes the transmission probability of the electrons at the Fermi level, averaged over all modes. To establish a connection to a semiclassical scenario, we can construct wave packets by superposition of the electronic wave functions of the occupied modes, which will result in classical particles moving with the Fermi velocity and which are injected at, say, $x = x_0$, toward the barrier with a constant angular distribution. Thus, $\langle T \rangle$ is obtained from averaging over the transmission probabilities $T(x_0, y, \alpha_0)$, where α_0 denotes the angle between the injection wave vector and the x -direction. In addition, each angle of incidence has to be weighted by the partial flux it carries in the x -direction, which is proportional to v_x and hence to $\cos \alpha_0$. According to these considerations, the conductance is given by [170]

$$G = \frac{2e^2}{h}N\langle T \rangle$$

with

$$\langle T \rangle = \int_{y=0}^W \int_{\alpha_0=-\pi/2}^{\pi/2} \frac{1}{W} \frac{\cos \alpha'_0}{2} T(x_0, y, \alpha'_0) d\alpha'_0 dy \quad (7.45)$$

where the factor $1/2$ originates from the normalization condition for the incident flux, i.e.

$$\langle v_x \rangle = v_F \int_{\alpha_0=-\pi/2}^{\pi/2} \cos \alpha'_0 d\alpha'_0 = \frac{1}{2}v_F$$

Moreover, N can be expressed as $N = k_F W / \pi$.

The transmission probability $\langle T \rangle$ may be obtained from a numerical simulation of an ensemble of trajectories, for example. The case of the magnetic barrier, however, can be solved analytically. Consider the effect of the barrier on an electron injected toward the barrier at x_0 in a direction given by α_0 . As the barrier is approached, the electron experiences an increasing Lorentz

force, which results in an x -dependent cyclotron motion, with decreasing cyclotron radius. The electron will be reflected if, at some position x , the electron moves in the $+y$ -direction. This depends not only on the shape and strength of the barrier, but also on the angle of incidence. The magnetic barrier can thus be thought of as a ballistic angular filter. Above a critical angle of incidence $\alpha_0 = \alpha_c$, all electrons will be reflected. We can obtain α_c from a geometric consideration. Consider the cyclotron orbit of infinitesimal length $ds = r_c(x) d\alpha$ that the electron performs at x . Along ds , it moves $dx = ds \cos \alpha d\alpha$ in the transport direction. Since $r_c(x) = v_F m^* / (e B_z(x))$, we can determine the angle $\alpha(x)$ of the electron velocity with respect to the x -direction from

$$\frac{e}{m^* v_F} \int_{x_0}^x B_z(x') dx' = \int_{\alpha_0}^{\alpha} \cos \alpha' d\alpha'$$

Reflection occurs if $\alpha = \pi/2$ is reached at any position x . The critical angle of incidence is therefore given by

$$\sin \alpha_c = 1 - \frac{e}{m^* v_F} \int_{x_0}^{\infty} B_z(x') dx' \quad (7.46)$$

which also implies that the barrier is closed for all electrons if

$$\int_{x_0}^{\infty} B_z(x') dx' \geq \frac{2m^* v_F}{e}$$

With this result, it is straightforward to calculate $\langle T \rangle$ for a magnetic barrier, provided we neglect corrections due to effects at the edge of the Hall bar. Clearly, $T(x_0, y, \alpha_0) = 1$ for $-\pi/2 \leq \alpha_0 < \alpha_c$, and = 0 otherwise. Thus,

$$G = \frac{2e^2 k_F W}{h \pi} \left(1 - \frac{e \Phi_b}{2m^* v_F} \right), \quad \alpha_c > -\pi/2$$

$$G = 0, \quad \alpha_c = -\pi/2 \quad (7.47)$$

with

$$\Phi_b = \int_{x_0}^{\infty} B_z(x') dx'$$

If the electrons are injected well before the magnetic barrier, one obtains

$$\Phi_b = \int_{-\infty}^{\infty} B_z(x') dx' = \frac{1}{2} \mu_0 M h$$

In this approximation, the conductance is independent of z_0 . Then Eq. (7.47) describes the experimental data reasonably well (see Fig. 7.31(a)). Note that, in order to compare experiment and model, the magnetization characteristic of the ferromagnet $\mu_0 M(B_{\parallel})$ has to be determined independently, which can be done by the technique of *Hall magnetometry* [215].

What happened to the contact resistance in this picture? In Eq. (7.47), it is included since we started from the two-terminal configuration. Alternatively, we could perform a more complicated calculation modeling the four-probe geometry of Fig. 7.30, and determine the chemical potential of the two voltage probes under the condition that they draw zero current [26]. In such a case, the result, $G = (\mu_1 - \mu_2)/eI$, would not contain the contact resistance. In our classical model in a two-terminal configuration, it enters implicitly via the assumption of point-like electrons. Such a wave packet can only be constructed from an infinite number of wave functions, which means that the electron gas must have an infinite width at the point of injection.

The method presented here is one way to determine the conductance of non-diffusive samples, i.e. of samples with scattering properties that do not lead to an exponential and homogeneous momentum relaxation. It is a powerful tool that can be applied to a wide variety of systems. We will get to know a different, complementary, technique applicable to such cases in Chapter 11, where we investigate transport through artificial crystals.

7.7

Concluding Remarks

As the feature size of our electronic devices keeps decreasing, the quantum wire aspect of an electrical connection will become more and more important. Conductance quantization in quantum wires is one of the milestones in mesoscopic physics, and its detailed explanation is not trivial. Throughout this chapter, we have treated the electrons as non-interacting. It is known, however, that the Landauer formula remains valid for an interacting region connected to non-interacting leads, which host the eigenstates of the incoming and outgoing waves.

Both non-interacting and interacting electron gases in one dimension can be mapped onto non-interacting bosons, as long as the energy of the excitations is small, i.e. the energy dispersion can be assumed as linear. The resulting system is a *Luttinger liquid*. Luttinger liquids are distinctly different from Fermi liquids. For example, they show spin-charge separation, which means that the spin and the charge are decoupled and have different group velocities. The Luttinger liquid aspects of one-dimensional electron systems are currently an

active field of research. The interested reader is referred to [265] as a starting point for further studies.

Papers and Exercises

P7.1 In [120], the authors derive the QPC conductance quantization by assuming that the narrowest point of the constriction determines the resistance. Discuss what determines the conductance in this model, and why it is quantized.

P7.2 At weak magnetic fields, QPCs show a negative magneto-resistance, as reported first in [157]. Develop a picture of this effect.

P7.3 In [53], the electrostatics of edge channels is developed. Quantify the edge channel geometry illustrated in Fig. 7.20.

P7.4 The paper by van Houten et al. [158] discusses in detail the effect of magnetic electron focusing in 2DEGs. Start by studying the formation of caustics (appendix C therein). Next, follow the instructive derivation of the expressions for the four-terminal experiments of section IV, as elaborated in appendix D and section V of that paper [158].

E7.1 An infinitely long quantum wire extends along the x -direction. In the z -direction, only the lowest subband is occupied. The confinement in the y -direction is parabolic: $V(y) = \frac{1}{2}m^*\omega_0^2y^2$. A magnetic field is applied in the z -direction.

- (a) Write down the two-dimensional Schrödinger equation (use the Landau gauge $\vec{A} = (By, 0, 0)$, and ignore the z -direction). Solve the equation using the ansatz $\Phi(x, y) = e^{ik_xx}\psi(y)$. Show that, by suitable substitution, the problem is equivalent to a harmonic oscillator.
- (b) Interpret the above results. Focus in particular on the evolution of the energy levels as a function of B , as well as on the energy dispersion in the x -direction. Discuss further the limits $\omega_0 \rightarrow 0$ and $\omega_c \rightarrow 0$.
- (c) Derive Eq. (7.10). [*Hint:* Use the density of states of a parabolic quantum wire and approximate the sum by an integral.]

E7.2 Consider the transmission of the lowest mode of a QPC, modeled by the transmission function $T(E) = \theta(E - E_1)$. Show that the conductance G

as a function of the chemical potential in the reservoirs and of E_1 has the form of Eq. (7.17). Calculate also the characteristic temperature for which the conductance quantization is thermally smeared. Assume that adjacent modes are separated by an energy difference Δ . Assume that the source-drain bias voltage is infinitely small.

E7.3 In this exercise, the resistances measured on the sample shown in Fig. 7.19 will be calculated. We assume that the filling factor in the ungated region is N ; below the gate, the filling factor is given by $M \leq N$. Both filling factors are integers. The spin degeneracy is lifted in all edge states.

- Set up the matrix equation obtained within the Landauer–Büttiker formalism.
- Calculate the resistances R_{ij} , for $i, j = 1, \dots, 4$. Explain the plateau at $R_{xx} = h/2e^2$ in Fig. 7.19.

E7.4 Conventional Hall geometries are insensitive with respect to electron scattering between edge states at the same edge. Nevertheless, such scattering exists, and can be characterized by a coupling p between adjacent edge states along a distance L . This coupling p is defined as $p = (\Delta\mu - \Delta\mu^*)/\Delta\mu$, where $\Delta\mu$ and $\Delta\mu^*$ denote the potential differences between the edge states at the beginning and at the end, respectively, of the distance L . (Is this a meaningful definition?) We study the coupling between the two spin-split edge states of the first Landau level. Experimentally, a Hall bar is adjusted in the regime of filling factor 2. The sample contains two gates across the Hall bar (see Fig. 7.32). These gates are biased to a regime where edge state 1 is transmitted, while edge state 2 is reflected. The distance between the gates is our length L .

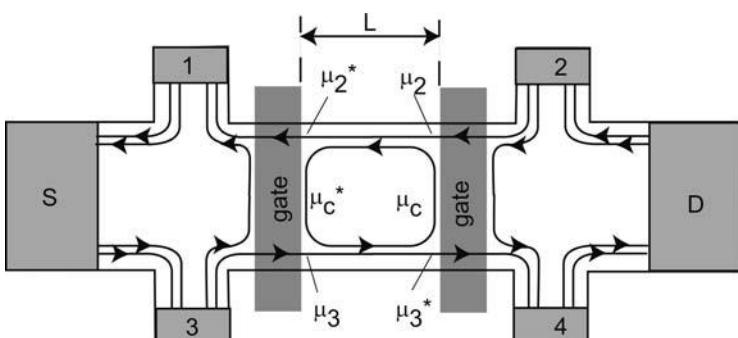


Fig. 7.32 Sample geometry for Exercise E7.4.

- (a) Assume that p is equal for both edges. Write down the matrix equation and determine μ_i ($i = 1, \dots, 4, 2^*, 3^*, c$, and c^*), as well as the current $I_S = I_D$. Note that, due to charge conservation, $\mu_3 + \mu_c^* = \mu_3^* + \mu_c$, and, similarly, $\mu_2 + \mu_c = \mu_2^* + \mu_c^*$.
- (b) The equilibration length L_{eq} is defined as the distance the electrons have to travel before the potential difference $\Delta\mu$ between two adjacent edge states decreases to $1/e = 0.368$ of its initial value. Determine L_{eq} from p . How large is L_{eq} for typical experimental numbers of $R_{12} = 0.53h/e^2$ and $L = 50 \mu\text{m}$?

E7.5 This exercise is about forming carbon nanotubes.

- (a) Express the \vec{C} shown in Fig. 7.23 in terms of the lattice vectors \vec{a}_1 and \vec{a}_2 of the graphite sheet. Determine n_1 and n_2 . Give the general condition for (n_1, n_2) for (i) armchair and (ii) zigzag tubes.
- (b) Calculate $\vec{A}(n_1, n_2)$ for the \vec{C} in Fig. 7.23.
- (c) Calculate the fundamental reciprocal lattice vector \vec{B} of the CN. Draw the first Brillouin zone of the CN in the Brillouin zone of a graphite sheet. What is the mode spacing Δk_y due to quantization along the CN circumference? Illustrate these modes in the Brillouin zone.
- (d) Is this particular CN metallic or insulating?
- (e) Consider the energy dispersions of Fig. 7.25. The metallic and semiconducting tubes are $(12, 0)$ and $(13, 0)$, respectively. Estimate the effective mass in the conduction/valence band of the semiconducting tube close to the band extremal points. What is the effective mass at the Fermi level of the metallic tube? Calculate the density of states around the Fermi level for the metallic CN. Why is it constant, although the system is one-dimensional? Does the chemical potential depend on temperature?

E7.6 Use the Landauer–Büttiker formalism to calculate V_c/I in Fig. 7.33(a), and I_c/V_i in Fig. 7.33(b). Assume identical QPCs in both cases. Discuss the advantages and disadvantages of both these setups.

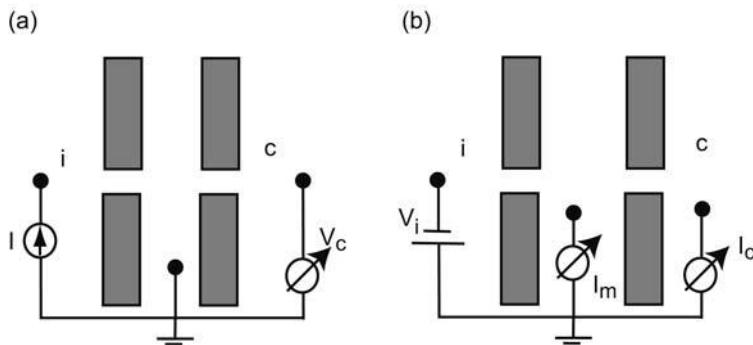


Fig. 7.33 Measurement configurations considered in Exercise E7.6.

Further Reading

More on quantum wires in general can be found in [27]. The reader is also referred to [65], which in particular provides a detailed discussion of many aspects related to conductance quantization, as well as to [90]. An excellent review on QPCs is the article [159]. Finally, carbon nanotubes are the topic of two recent books, namely [258] and [141].

8**Electronic Phase Coherence**

In the previous chapters, we have implicitly assumed that the electrons are phase coherent in the confined directions, but have not considered the possible consequences of phase coherence in the extended directions. Do such effects exist at all? The answer is “Yes”, as you may have guessed. Coherence manifests itself in electronic interferences, which can take place within the time scale τ_ϕ and the corresponding length scale ℓ_ϕ . Interference effects remind us of wave optics, and there are in fact many analogies. The most important signatures of electronic phase coherence in diffusive systems are Aharonov–Bohm type effects, weak localization, and universal conductance fluctuations, which are the topic of Sections 8.1, 8.2, and 8.3, respectively. In Section 8.4, we have a look at phase coherence in ballistic systems. The final section (Section 8.5) introduces the resonant tunneling effect, i.e. the transmission barriers of tunnel barriers in series with a spacing smaller than the phase coherence length.

8.1**The Aharonov–Bohm effect in mesoscopic conductors**

In 1959, Aharonov and Bohm published a seminal gedanken experiment [1]. The authors predicted that the partial waves of a charged particle enclosing an electrostatic or magnetic potential experience a magnetic phase shift, even if the electric and magnetic fields vanish in the regions of non-zero probability density. This phase should be distinguished from the dynamic phase, which is the frequency of the plane wave electronic states at the Fermi energy, integrated over time. Interferences as a function of the relative phase shift occur, which are known as the electrostatic and the magnetic Aharonov–Bohm (AB) effect, respectively.

In Fig. 8.1, an experimental setup suited to test this prediction is shown. A ring is patterned out of a metal or a 2DEG, with a circumference smaller than the phase coherence length. Suppose a conducting ring encloses a magnetic vector potential \vec{A} that generates a constant magnetic field perpendicular to

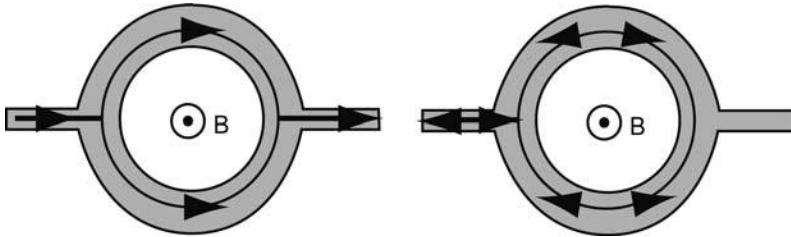


Fig. 8.1 Sketch of a sample used to study the Aharonov–Bohm effect. Interfering trajectories that cause h/e and $h/2e$ oscillations are drawn to the left and to the right, respectively.

the plane of the ring.¹ The phase collected by the electrons during their passage through branch j of the ring (j denotes the upper or the lower semicircle) is given by

$$\begin{aligned}\phi_j &= \frac{e}{\hbar} \int_j \vec{A} d\Gamma \\ \phi_{\text{upper}} &= \frac{eBS}{2\hbar}, \quad \phi_{\text{lower}} = -\frac{eBS}{2\hbar}\end{aligned}\quad (8.1)$$

Here, S denotes the area enclosed by the ring, and Γ is the parameterized trajectory. The total transmission probability T is obtained by summing up all the probability amplitudes and calculating the absolute value of the square. For now, we neglect multiple reflections at the entrance and exit of the ring. Let us further assume that, for $\vec{A} = 0$, both branches have identical transmission amplitudes t_0 . The total transmission probability T is obtained from

$$\begin{aligned}t_j &= t_0 e^{i\phi_j} \quad \Rightarrow \\ T &= (t_{\text{upper}} + t_{\text{lower}})^* (t_{\text{upper}} + t_{\text{lower}}) = 2T_0 \left[1 + \cos \left(\frac{eBS}{\hbar} + \phi_0 \right) \right]\end{aligned}\quad (8.2)$$

Apparently, T oscillates as a function of B , with a period of one magnetic flux quantum $\Phi_0 = h/e$ that penetrates the ring. Experimentally, the AB effect on metallic loops was first observed in gold rings [316] (see Fig. 8.2). It was later reproduced in rings defined in a Ga[Al]As heterostructure [303] (see Fig. 1.2 for such data). In these experiments, the amplitudes were much smaller than predicted by Eq. (8.1), since ℓ_ϕ was smaller than the ring circumference, and thus only a fraction of the electrons could pass the ring coherently.

So far, interferences have been taken into account to first order only. For sufficiently large ℓ_ϕ , there are of course also higher-order interferences, for example, between two partial waves that have traversed both arms clockwise

1) This is not exactly the original proposal in [1], since the branches of the ring are penetrated by the magnetic field. For very narrow arms compared to the ring diameter, however, this modification is irrelevant.

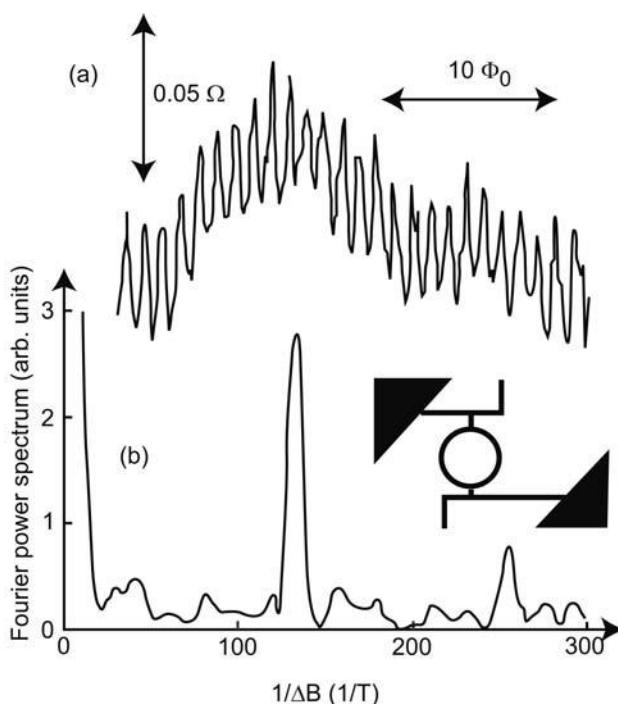


Fig. 8.2 AB effect as observed in a gold ring (the sample is shown in the inset). In (b) the Fourier spectrum of the raw data in (a) is shown. A strong eS/h frequency is observed, while the second order $2eS/h$ is much weaker. After [316].

and counterclockwise, respectively (see Fig. 8.1). Owing to their interference at the ring's entrance, the reflection probability is a periodic function of B . The period of these oscillations is half the AB period, $h/2eS$. They are known as Altshuler–Aronov–Spivak (AAS) oscillations [2] and are of particular relevance in mesoscopic physics. Since both trajectories traverse exactly the same path, their phase difference at the ring's entrance is always zero, independent of the size and shape of the loop. The interference is constructive, which leads to a backscattering probability that is enhanced compared to the classical value. One speaks of *coherent backscattering*. In an ensemble of AB rings, all the initial phases ϕ_0 are randomly distributed, while they are always zero for the AAS oscillations. We therefore expect that, in ensembles of rings, the AB oscillations average to zero, while the AAS oscillations survive ensemble averaging. This was demonstrated first in [272] (Fig. 8.3), long before the AB oscillations could be detected in individual rings. This experiment was carried out with an insulating cylinder (diameter about $1.5\mu\text{m}$), coated with a thin Mg film. Measuring the resistance between the top and the bottom of the cylinder in a magnetic field along the cylinder axis can be thought of as ensemble averag-

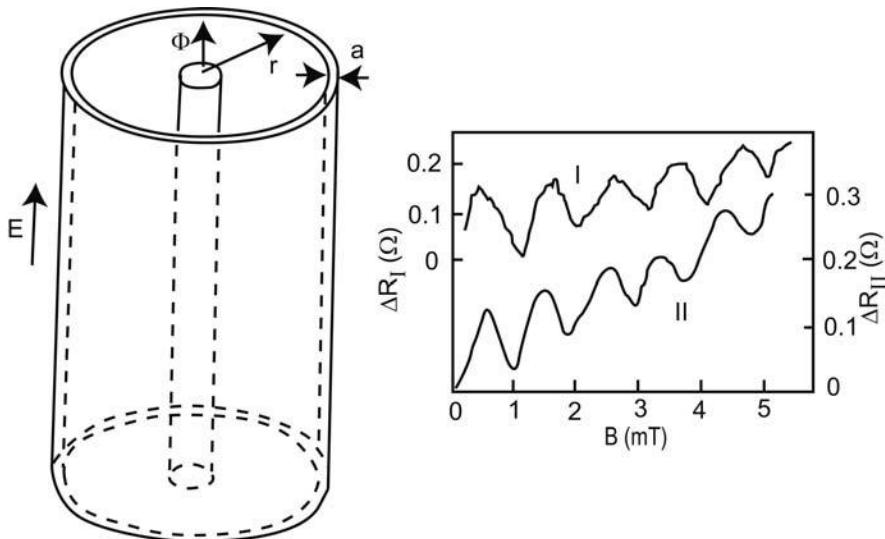


Fig. 8.3 The experiment of Sharvin and Sharvin [272]. The resistance as a function of the magnetic field along the cylinder axis oscillates with a period of $h/2eS$, i.e. it shows AAS oscillations. AB oscillations are absent. After [272].

ing over many rings in parallel. Owing to the small variations in the diameter, several oscillation periods can be observed.

Ensemble averaging was systematically investigated in [308]. Different numbers ($j = 1, 3, 10$, and 30) of silver rings were patterned in series, and the amplitudes of the various oscillation periods were investigated. It was found that the amplitude of the AB oscillations drops with $1/\sqrt{j}$ while the amplitude of the AAS oscillations remains constant (Fig. 8.4).

8.2

Weak localization

Imagine an experiment where a lot of rings with a broad size distribution are measured simultaneously. What will remain of the magneto-oscillations? We surely can no longer expect to observe them, since each loop area has its own period, which will ensemble average.² Note, however, that all the resonant backscattering waves are in phase at $B = 0$, and the resistance is enhanced as compared to the incoherent case.

To be a bit more specific, we follow the line of arguing presented in [27] and consider the probability $P(\vec{r}_1, \vec{r}_2, t)$ for the electron to move from \vec{r}_1 to \vec{r}_2

2) Small magnetic fields are assumed, such that Shubnikov–de Haas oscillations are absent.

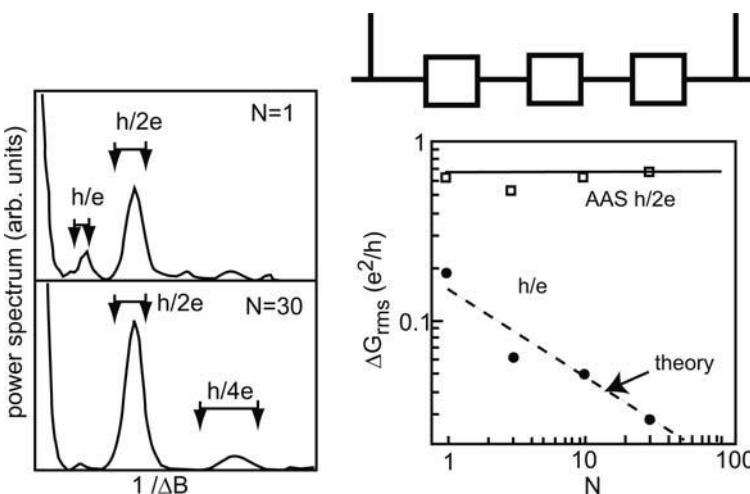


Fig. 8.4 Top right: Three Ag loops in series. The size of the loops is $940\text{ nm} \times 940\text{ nm}$. The width of the Ag wires is about 80 nm. Left: The Fourier spectra of the magneto-resistance oscillations observed in one ring (top) shows both h/e and $h/2e$ components. For 30 loops in series (bottom), however, the h/e component is absent, while a weak $h/4e$ component has emerged. Right: The relative strengths of the Fourier peaks are plotted vs. the number N of rings in series. After [308].

within the time t . Thus $P(\vec{r}_1, \vec{r}_2, t)$ is the squared sum of all the probability amplitudes A_i for this propagation within t :

$$P(\vec{r}_1, \vec{r}_2, t) = \left| \sum_i A_i \right|^2 = \sum_i |A_i|^2 + \sum_{i \neq j} A_i A_j^2 \quad (8.3)$$

The first term on the right-hand side is the classical probability for the electron to propagate between the two points along any path within t . The second term results from interferences. Since the phases of A_i are uncorrelated, the interference term averages to zero, with one exception: for $\vec{r}_1 = \vec{r}_2$, we can form pairs of trajectories that correspond to identical paths, traversed in opposite directions. In other words, the two paired propagators can be mapped on each other by time inversion. At $B = 0$, their phases are identical, and interference is constructive. Such pairs thus give a non-vanishing contribution to the interference term. Since $|A_i + A_i^{\text{time reversed}}|^2 = 4|A_i|^2$, we find an enhancement of the backscattering probability by a factor of 2, compared to the classical value, for such propagators. As the magnetic field is increased, the contribution of the largest rings in the ensemble to the resistivity will oscillate rapidly, while the phase difference in the smallest rings will remain essentially unchanged. Hence, the larger the magnetic field, the fewer rings will contribute to the constructive interference, and the resistance should drop to its classical value,

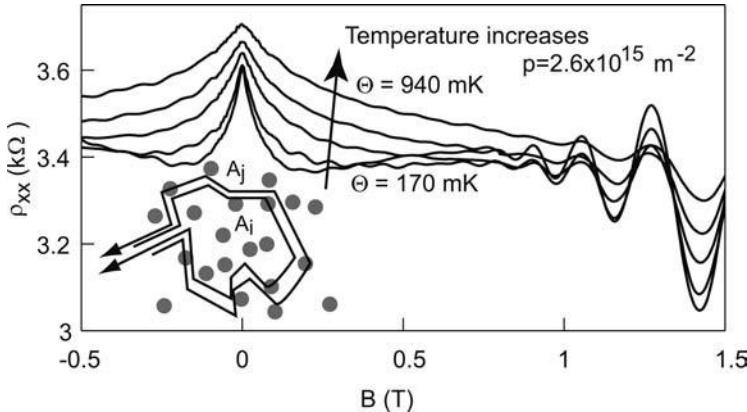


Fig. 8.5 Inset: A fraction of the electronic trajectories in a diffusive 2DEG form closed loops and lead to coherent backscattering. Main figure: WL peak as a function of B , and for various temperatures between 170 and 940 mK. The sample was a Si–SiGe quantum well containing a hole gas. Adapted from [271].

once the phase shift in the smallest rings is of the order of π . This is exactly the situation encountered in a diffusive electron gas. The ensemble of loops is formed by the elastic scatterers (see Fig. 8.5). This coherent backscattering at randomly distributed scatterers is called *weak localization* (WL). The localization of the electrons due to coherent backscattering is thereby distinguished from strong localization, which takes place in highly disordered samples.

Experimentally, we can thus expect an increased resistivity in diffusive samples due to WL at zero magnetic field, which is reduced to its classical value as the magnetic field increases. Of course, there will be no AAS oscillation because of the averaging. This is observed in experiments (see Fig. 8.5). The functional form of this WL peak depends on several parameters. For a two-dimensional system, i.e. for a sample width larger than ℓ_ϕ , Altshuler et al. [3] have derived the magnetic field dependence of the WL correction to the classical conductivity:

$$\Delta\sigma_{xx}^{\text{WL}}(B) = \frac{e^2}{2\pi^2\hbar} \left[\Psi\left(\frac{1}{2} + \frac{\tau_B}{2\tau_\phi}\right) - \Psi\left(\frac{1}{2} + \frac{\tau_B}{2\tau}\right) + \ln\left(\frac{\tau_\phi}{\tau}\right) \right] \quad (8.4)$$

which at $B = 0$ reduces to

$$\delta\sigma_{xx}^{\text{WL}} = -g_s \frac{e^2}{4\pi^2\hbar} \ln\left(1 + \frac{\tau_\phi}{\tau}\right) \quad (8.5)$$

Here, $\Psi(x)$ is the digamma function. For large arguments, it can be approximated by

$$\Psi(x) \approx \ln(x) - \frac{1}{x}$$

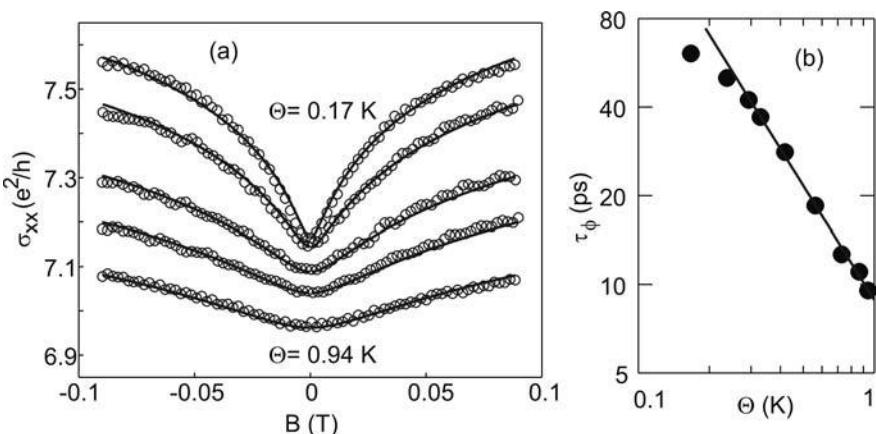


Fig. 8.6 (a) The data of Fig. 8.5 (circles), translated into longitudinal conductivity $\sigma_{xx}(B)$. The lines are least squares fits according to Eq. (8.4). (b) Temperature dependence of τ_ϕ , as determined from these fits. The line represents the theoretically expected $1/T$ dependence. Adapted from [271].

Fig. 8.6 shows that typical data can be fitted very well to the Altshuler formula.³ The fit parameter is the phase coherence time.

It is widely accepted that, at low temperatures, the dephasing occurs via quasi-elastic electron-electron collisions [3]. For such a type of dephasing, one expects a characteristic $\tau_\phi \propto 1/T$ dependence, which is usually found in experimental data (see Fig. 8.6), except at very low temperatures. Here, τ_ϕ saturates in most experiments. A plausible explanation is that the electron temperature deviates from the bath temperature of the helium mixture in dilution refrigerators at very low temperatures.

8.3

Universal conductance fluctuations

Let us take another look at the quantum wire already presented in Figs. 7.3 and 7.4. In Fig. 8.7, its two-probe conductance as a function of the electron density and of magnetic fields in the regime $\omega_c < 0.45\omega_0$, where Landau quantization can be neglected, is shown. Around $B = 0$, a pronounced dip in the conductance can be seen, which is due to weak localization. In addition, conductance fluctuations are observed as a function of both parameters outside the weak localization peak. Cross sections of the conductance as a function of only one parameter with the second parameter fixed show this more clearly.

Note that these fluctuations are *not* noise: the features are fairly symmetric with respect to magnetic field inversion, and the features shift toward larger

3) Note that, in order to transform ρ_{xx} into σ_{xx} , the Hall resistivity has to be measured as well.

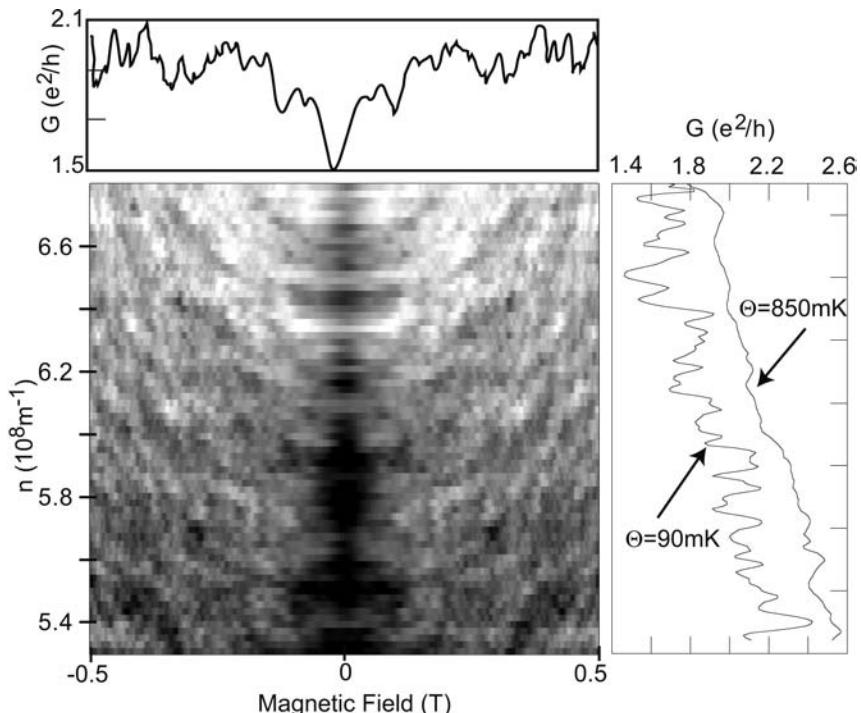


Fig. 8.7 Main figure: Conductance of the quantum wire shown in Figs. 7.3 and 7.4, as a function of the one-dimensional electron density n and the magnetic field B . The gray scale ranges from $G = 1.4e^2/h$ (dark) to $2.5e^2/h$ (light). Cross sections as a function of B for constant $n = 6 \times 10^{-8} \text{ m}^{-1}$, and as

a function of n at $B = 0$, are shown to the top and to the right, respectively. The temperature is 90 mK. The figure on the right shows the thermal smearing of the fluctuations as the temperature is moderately increased to 850 mK.

electron densities as $|B|$ is increased. In addition, the fluctuations smear out rapidly as the temperature is increased to about 1 K. The typical fluctuation amplitude is of the order of $0.2e^2/h$. Characteristic fluctuation periods are $\approx 20 \text{ mT}$ and $5 \times 10^6 \text{ m}^{-1}$, respectively. These fluctuations are *parametric*, i.e. they are perfectly reproducible as a function of the parameters, but they are nevertheless fluctuations, since there is no way to control their individual appearance. Also, the fluctuations look different in other samples with identical macroscopic properties, and they change in an individual sample when we warm it to room temperature and cool it down again.

Apparently, the fluctuations depend on the mesoscopic structure of the sample. It is known that thermal cycling changes the impurity configuration of the sample, but not the macroscopic features, like the impurity density or the scattering times. It can thus be assumed that the fluctuations somehow characterize the specific impurity configuration in the sample during a particular

cooling down. Note that the amplitude and the width of the weak localization peak fluctuate as well. This is an indication that the measurement does not average over a huge number of weak localization loops, but just a few of them. Furthermore, the fluctuations show a very strong temperature dependence at ultra-low temperatures below 1 K, while the elastic scattering times are essentially independent of temperature in this regime. This suggests that these fluctuations are again a manifestation of electronic phase coherence: we have just seen in Section 8.2 on weak localization that $\tau_\phi \propto 1/\Theta$ in this temperature range. This very qualitative line of arguing essentially sketches the generally accepted interpretation of these conductance fluctuations.

Similar fluctuations can be observed in many mesoscopic samples where the sample size is comparable to the phase coherence length, but larger than the elastic mean free path. The quantum wire under consideration here, for example, has a length of $L = 40 \mu\text{m}$, while $\ell_d = 5.7 \mu\text{m}$ and $\ell_q = 460 \text{ nm}$ for the 2DEG of which the wire was made. The phase coherence length inside the wire is $\ell_\phi \approx 7 \mu\text{m}$. In most samples with $\ell_e < L < \ell_\phi$ and for negligible temperature, it is found that the average fluctuation amplitude depends on neither the sample size nor the strength and configuration of the elastic scatterers. This is quite remarkable; apparently, more scatterers does not imply more pronounced smearing of the fluctuations. Therefore, they are often referred to as *universal conductance fluctuations* (UCF).

Unfortunately, there is no simple, intuitive picture for UCF, in contrast to weak localization. Furthermore, the quantitative description of the fluctuations strongly depends on many length scales, in particular the width and length of the sample, as well as on ℓ_e , ℓ_ϕ and ℓ_T . We therefore refrain from a discussion of UCF in all these regimes, and exemplify it by some qualitative arguments for our wire. All theoretical models for parametric UCF are based on the *ergodicity theorem*. Suppose that we measure the resistivity of various samples with identical macroscopic parameters and sample sizes of a few elastic mean free paths only. Furthermore, the samples are cooled to a regime where $\ell_\phi > \ell_e$. Even with identical resistivities over macroscopic length scales, we can no longer expect to measure identical resistances in different samples, since the exact number of scatterers will be sample-dependent. In addition, the microscopic configuration of elastic scatterers is sample-specific, even if their numbers are identical, and consequently the interference pattern of the electron waves, and with it the transmission and reflection probabilities, will be unique for each sample. As in weak localization, the interferences of the electronic wave functions generate localization and reduce the conductance.

Experimentally, it is a rather tedious task to fabricate and measure sufficiently many samples for good statistics. It is, however, generally accepted that the systems under consideration here behave *ergodically*. Suppose that we

measure a quantity q in an ensemble of samples with identical macroscopic parameters, and determine the variance of q . In a second experiment, we measure the same quantity in only one member of the ensemble as a function of a parameter p (which can be, for example, the electron density or the magnetic field), and average the quantity over p . By definition, a system is called *ergodic* (with respect to q) if the two averaging procedures give the same mean value and the same variance. This definition assumes that both the number of samples in the ensemble and the scan range of the parameter are infinite. In a simple picture, we can imagine that, in both experiments, we average over many micro-states of the system. Non-ergodic samples are thus samples where tuning the external parameter does not induce sufficient transitions between micro-states. This can be due to metastable states, for example. For a more detailed discussion of ergodicity, see for example [229].

By what amount do we have to change p before the micro-state of the system can be regarded as different, or before the values $q(p)$ of the measured quantity can be regarded as statistically independent, respectively? The answer is given by the autocorrelation function⁴ $C_q(\Delta p)$. At $\Delta p = 0$, we have $C_q(0) = C_0$. For a fluctuating q as a function of p , typically $C_q(\Delta p)$ drops to zero within a certain range of Δp . The autocorrelation value p_C is the parameter value for which the autocorrelation function has dropped to $0.5C_0$, and the data can be considered as statistically independent as soon as p differs by at least p_C . As long as $\ell_\phi > L$, the average UCF amplitude has been found to be of the order of e^2/h , independent of the number of elastic scatterers and of L . This is quite surprising, since it means that averaging over disorder does not weaken the fluctuations, as long as the sample is phase coherent. This result can be derived if one assumes strong correlations between the transmission amplitudes of different paths across the whole sample, while the reflection amplitudes are uncorrelated. An intuitive argument for such a scenario would be that electrons that manage to cross the sample do so via identical sections of trajectories which contain many scatterers, while a single backscattering event suffices for reflection [189].

In many circumstances, however, the non-zero temperature cannot be neglected. Its effect on the length scales is twofold. First of all, it may reduce ℓ_ϕ to the regime of $\ell_e < \ell_\phi < L$. Second, the thermal length comes into play. In [25], it has been argued that, for a quantum wire with $w \ll \ell_\phi$, the fluctuation amplitude can be approximated by

$$\delta G = \frac{1}{\beta} \frac{e^2}{h} \sqrt{\frac{12(\ell_\phi/L)^3}{1 + [9/(2\pi)](\ell_\phi/\ell_T)}} \quad (8.6)$$

⁴) For a brief introduction to correlation functions, see Appendix B.

which, for the QWR under study here, agrees fairly well with experiment if a phase coherence length inside the wire of $\ell_{phi} \approx 7 \mu\text{m}$ and a thermal length of $\ell_T \approx 11 \mu\text{m}$ are assumed.

We conclude this section by discussing an experiment in which the impurity configuration has been changed parametrically [148]. The QWR under discussion was tuned by varying the two in-plane gates (see Fig. 7.3) such that $\delta V_{as} = \delta V_1 = -\delta V_2$. In Exercise E6.3, it was shown that a constant electric field displaces a parabolic potential without changing its shape. By analyzing the magneto-conductance oscillations in large magnetic fields, one can ensure that ω_0 remains constant in such an experiment. Within the approximation of a parabolic confinement, the QWR is thus spatially displaced in the y -direction as δV_{as} is scanned (Fig. 8.8(a)). The displacement δy equals the change of the wire width as a function of one in-plane gate voltage with the second gate voltage fixed. Shifts up to $\delta y = 12 \text{ nm}$ are possible before a leakage current between wire and in-plane gates sets in. Furthermore, δy is linear in δV_{pgi} ($i = 1, 2$) within experimental accuracy, and a lever arm $\delta y/\delta V_{as} = 80 \text{ nm/V}$ is measured.

In Fig. 8.8(b), the conductance of the wire is shown as a function of its displacement. Again, reproducible conductance fluctuations are observed, with a temperature dependence similar to the fluctuations as a function of B and n . The average period and amplitude are $\delta y \approx 2 \text{ nm}$ and $\delta G \approx 0.15e^2/h$, respectively. Apparently, the interference pattern can also be changed by shifting the wire through the crystal. Intuitively, this is quite clear, since the interference pattern depends on the potential landscape at which the electronic waves scatter. As the wire is shifted through the crystal, the scattering potential is scanned, and the interference pattern changes accordingly. But what determines the fluctuation period of $\delta y \approx 2 \text{ nm}$?

We denote the relevant density of bumps in the scattering potential by $1/d^2$ and try to relate d to a characteristic length scale of the sample. On average, the number of bumps inside the wire should change by one as the wire is displaced by $\delta y = d^2/2L$. The factor of $1/2$ enters because the bumps may enter the wire on one side as well as exit it on the other. One finds $\delta y = 2.7 \text{ nm}$ for $d = \ell_q$, the quantum scattering length. This indicates that the conductance fluctuations are caused by all scatterers, and not just by the large-angle scatterers that determine ℓ_e , which enter or leave the wire region as it is displaced. It should be mentioned that shifting the impurity configuration within the wire also generates conductance fluctuations, which, however, have a characteristic fluctuation period of the order of the Fermi wavelength $\lambda_{F,1}$ of the electrons in the lowest one-dimensional subband, as has been numerically demonstrated in [50] for the case of displacing a single scatterer. In the experiment under discussion here, however, this length scale cannot be seen, since $\lambda_{F,1} \approx 30 \text{ nm}$, which is larger than the displacement range.

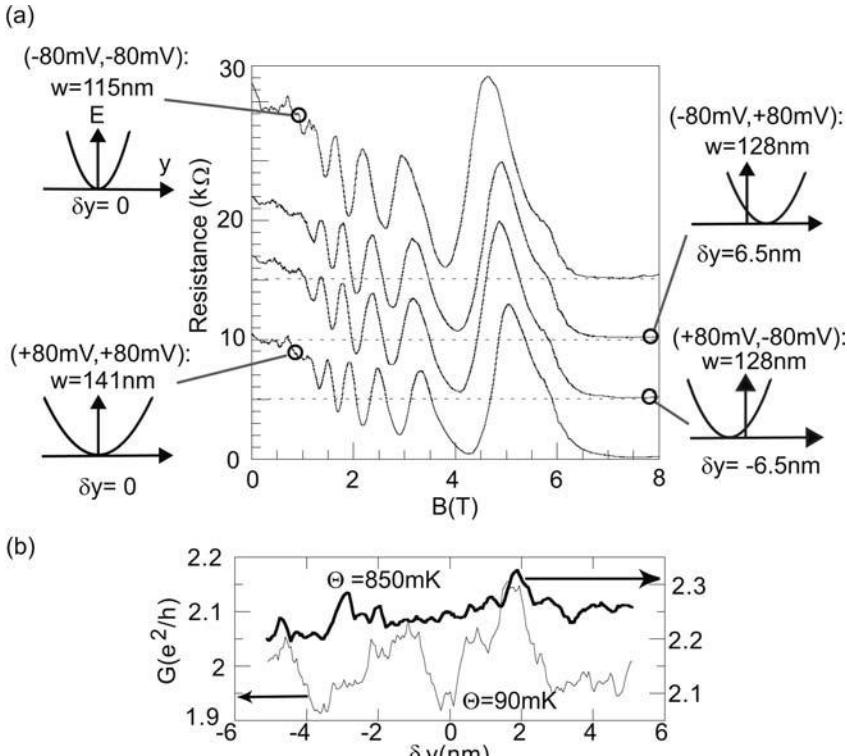


Fig. 8.8 (a) Magneto-resistance of the QWR (the sample is shown in Fig. 7.3) for different combinations of in-plane gate voltages (V_1, V_2). Adjacent traces are offset by 5 k Ω , and the dashed lines denote the corresponding zeroes. Fitting the oscillations to Eq. (7.10) shows that the wire width w and

ω_0 do not alter for antisymmetric gate voltage changes, i.e. for $V_1 + V_2 = \text{constant}$. The potential shape and position of the QWR are indicated by the sketches outside the main figure. (b) Measured conductance fluctuations as a function of the wire displacement δy , shown for two temperatures.

Question 8.1: How do you explain that in Fig. 8.7, the features in the conductance shift to larger electron densities as $|B|$ is increased?

8.4

Phase coherence in ballistic 2DEGs

Electronic phase coherence in ballistic 2DEGs is a relatively unexplored territory. All the experiments discussed so far in this chapter have relied on the diffusive character of the sample; scattering at impurities was essential for obtaining information on τ_ϕ . An important experiment addressing the issue

of phase coherence in the ballistic regime has been performed in [336] (see Fig. 8.9). Electrons were injected into a ballistic 2DEG via a quantum point contact acting as an emitter (E), and collected using a second QPC (C). Since the sample in between the emitter and collector is ballistic, only those electrons that move very close to the straight line that connects E and C contribute to the voltage buildup. The upper half-plane of the 2DEG was partly covered by a gate of length L , which could be used to tune the electron density, and hence the phase shift, of the electrons underneath. Hence, approximately 50% of the electrons entering the collector have passed below the gate. Assuming a linear relation between the gate voltage V_G and the Fermi energy E_F in the 2DEG, the phase shift that the electrons acquire underneath the gate is given by

$$\delta\phi = Lk_{F,0}\delta(\sqrt{1 - V_G/V_T}) \quad (8.7)$$

where $k_{F,0}$ denotes the Fermi wave vector below the gate for $V_G = 0$, and $V_D = -290$ mV is the threshold gate voltage, at which the 2DEG below the gate gets depleted.

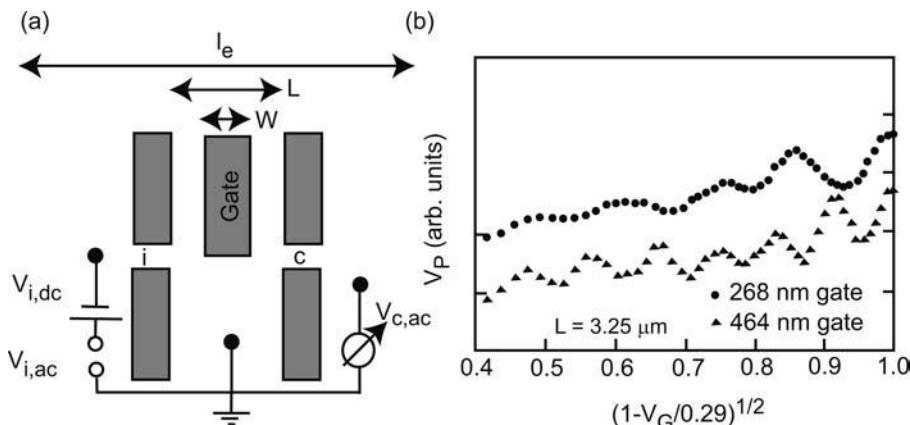


Fig. 8.9 (a) Sketch of the experimental setup used to investigate the dephasing in ballistic 2DEGs. The 2DEG is enclosed by two split gates acting as emitter (E) and collector (C), separated by a distance $L < \ell_e$. Note that the ballistic 2DEG is grounded, and the AC voltage buildup is measured behind the collector QPC. (b) Voltage buildup at C, as a function of $\sqrt{1 - V_G/V_T}$. Periodic oscillations are observed, with a period that drops as the gate width W is increased. The measurements were performed at $\Theta = 1.4$ K. Adapted from [336].

llector QPC. (b) Voltage buildup at C, as a function of $\sqrt{1 - V_G/V_T}$. Periodic oscillations are observed, with a period that drops as the gate width W is increased. The measurements were performed at $\Theta = 1.4$ K. Adapted from [336].

Question 8.2: Check Eq. (8.7) with the assumptions stated in the text.

The phase shifted electrons that traversed underneath the gate interfere with those that bypass the gated region, and the resulting current should therefore be periodic with a period of

$$\delta\phi = 2\pi \quad \Rightarrow \quad \delta(\sqrt{1 - V_G/V_T}) = 2\pi/k_{F,0} \quad (8.8)$$

In fact, the voltage buildup at the collector as a function of the gate voltage showed quasi-periodic oscillations with the expected period, as shown in Fig. 8.9(b). The data shown here have been obtained for a pure AC excitation at E, i.e. for the special case of $V_{E,DC} = 0$. The experiment, however, has been carried out for a variety of DC excitation voltages, in order to compare the experiment with theoretical considerations. Note that only the AC component was measured at the collector.

From the relative oscillation amplitude $a(L, V_{E,DC})$, which is the peak-to-peak amplitude divided by the average voltage buildup, the phase coherence length was calculated via the relation

$$\ell_\phi = -\frac{L}{\ln[a(L, V_{E,DC})]} \quad (8.9)$$

This equation has its origin in the assumption that complete dephasing takes place via single electron-electron scattering events, which occur randomly.⁵ This conclusion is drawn because the theoretical expression for the electron-electron scattering length in ballistic 2DEGs, derived in [119], shown as full lines in Fig. 8.10, agrees very well with the experimental data.

The data imply that $\ell_\phi \propto 1/V_{E,DC}^2$ in this case. If we, somewhat sloppily, identify $V_{E,DC}$ with an effective temperature via $eV_{E,DC} \approx k_B\Theta$, this result indicates that, in the ballistic regime, $\ell_\phi \propto 1/\Theta^2$, which is different from the result in the diffusive regime. Furthermore, $\ell_\phi \approx 100 \mu\text{m}$ was found for $V_{E,DC} = 0$, which corresponds in the ballistic regime to a dephasing time of $\tau_\phi = \ell_\phi/v_F \approx 37 \text{ ps}$, which is the same order of magnitude as found in diffusive systems.

8.5

Resonant tunneling and s-matrices

Tunnel barriers are an essential part of many nanostructures. They represent small, often tunable, resistors with little or no dissipation.⁶ The transmission

- 5) The concept behind this relation is the Poisson distribution of uncorrelated events, an issue discussed in further detail in Exercise E8.2.
- 6) The electrons that tunnel elastically through a barrier are injected into the collector at an energy eV above the Fermi level. Dissipation occurs inside the collector, within the inelastic scattering length.

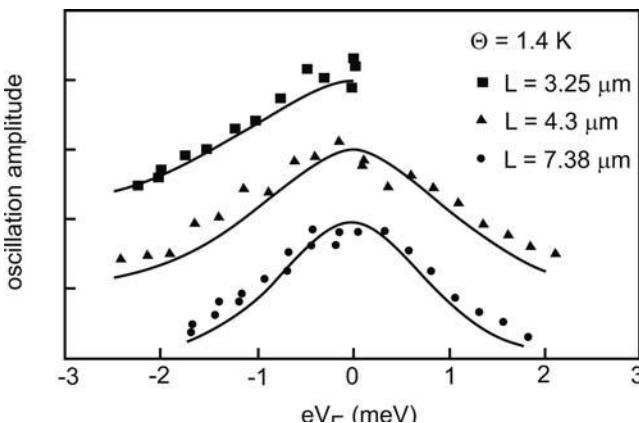


Fig. 8.10 The symbols denote the oscillation amplitudes as measured for various gate lengths L and under different DC emitter voltages $V_{E,DC}$. The solid lines are theoretical curves as expected for complete dephasing in single electron-electron scattering events at random locations. The good agreement suggests that this is in fact the actual dephasing mechanism. (Taken from [336]).

probability T of a rectangular barrier of height V_0 and width a as a function of the energy E of the incident particle is a standard example in elementary quantum mechanics. It is given by

$$T(E) = \begin{cases} \frac{4E(V_0 - E)}{4E(V_0 - E) + V_0^2 \sinh^2[\sqrt{2m(V_0 - E)}a/\hbar]}, & E \leq V_0 \\ \frac{4E(E - V_0)}{4E(E - V_0) + V_0^2 \sin^2[\sqrt{2m(E - V_0)}a/\hbar]}, & E \geq V_0 \end{cases} \quad (8.10)$$

(see Fig. 8.11). We speak of tunneling if $E \leq V_0$. In semiconductors, such barriers can be designed by incorporating an AlAs layer in GaAs during growth. Another widespread experimental realization is by quantum point contacts in the pinch-off regime, as discussed in Chapter 7. In metallic nanostructures, a tunnel barrier can be easily formed by depositing a metal layer on top of an oxidized metal.

The s -matrix \underline{s} of a tunnel barrier relates the outgoing wave functions to the incoming wave functions via

$$\vec{b} = \underline{s}\vec{a}, \quad \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} r_{11} & t_{12} \\ t_{21} & r_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad (8.11)$$

The coefficients of the s -matrix can be calculated from elementary quantum mechanics. How exactly they depend on the barrier parameters is of marginal interest only for our purposes. Here, we simply note that the off-diagonal elements t_{ii} correspond to transmission amplitudes, while the diagonal elements

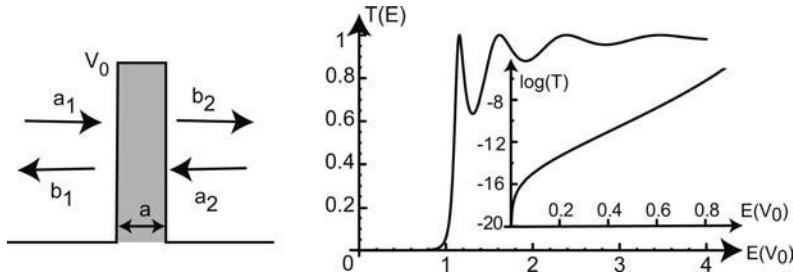


Fig. 8.11 Transmission of a single rectangular barrier (sketched to the left), plotted for $a = 8\hbar/\sqrt{2m}$ (m is the mass of the particle). The inset shows a logarithmic plot of the tunneling regime, showing that, in the tunneling regime, T increases approximately exponentially with energy.

r_{ii} represent reflection amplitudes. Owing to conservation of probability current density, \underline{s} has to be unitary:

$$\underline{s}^T \underline{s} = \underline{1} \quad \implies$$

$$|r_{11}| = |r_{22}|, \quad |t_{12}| = |t_{21}|, \quad r_{11}^* r_{11} + r_{22}^* r_{22} = 1, \quad r_{11} t_{12}^* + t_{21} r_{22}^* = 0$$

Since the tunnel barrier in Fig. 8.11 is invariant under reflection, the relations

$$r_{11} = r_{22} = r, \quad t_{12} = t_{21} = t$$

hold in this case. For tunnel barriers of a different shape, the transmission probabilities of course differ from the simple expressions in Eq. (8.10).

Methods for calculating such transmission coefficients are well established in elementary quantum mechanics. For our purposes, however, the relation between barrier shape and the s -matrix is of no further interest. Any barrier can be characterized by an s -matrix of the type (8.11). We will combine the s -matrices of individual barriers to calculate the transmission probability of more complicated structures, in particular systems with two tunnel barriers in series. The *double barrier* is sketched in Fig. 8.12. Each barrier is characterized by its s -matrix. We are interested in the transmission of the double barrier structure.

Suppose the transport between the barriers is completely coherent, i.e. the distance L between the barriers is much smaller than the phase coherence length. As the electrons travel between the barriers, they collect a phase θ . Each time the wave hits barrier j , a fraction $(1 - t_j)$ gets reflected, which leads to interference. The total transmission amplitude from source to drain t_{sd} is obtained by summing up the partial transmission amplitudes along all the trajectories the electron wave can take. There are infinitely many such trajectories, since the electron can experience an arbitrary number of round trips

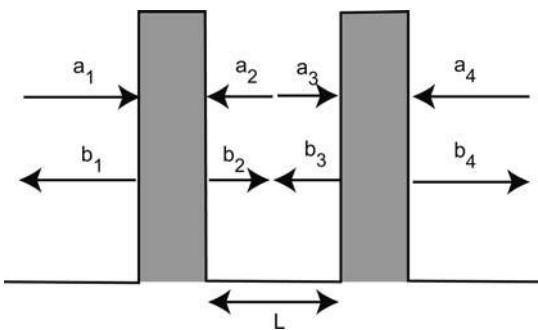


Fig. 8.12 Sketch of a resonant tunneling structure, formed by two tunnel barriers in series separated by a distance L , and the corresponding wave function amplitudes.

between the barriers before leaving the structure. Therefore,⁷

$$\begin{aligned} t_{\text{sd}} &= t_1 e^{i\theta} t_2 + t_1 e^{i\theta} r_2 e^{i\theta} r_1 e^{i\theta} t_2 + \dots \\ &= t_1 t_2 e^{i\theta} \left[1 + \sum_{j=1}^{\infty} (r_1 r_2 e^{2i\theta})^n \right] = \frac{t_1 t_2 e^{i\theta}}{1 - r_1 r_2 e^{2i\theta}} \end{aligned}$$

which gives the transmission probability

$$T = t_{\text{sd}}^* t_{\text{sd}} = \frac{T_1 T_2}{1 + R_1 R_2 - 2\sqrt{R_1 R_2} \cos \theta} \quad (8.12)$$

This transmission is plotted as a function of θ in Fig. 8.13.

Eqation (8.12) is nothing other than the well known Airy formula describing the transmission of coplanar optical resonators, so-called *etalons*. The properties of T are therefore extensively discussed in textbooks on wave optics. In optical resonators of this type, the finesse $F^* = (\pi\sqrt{1-T})/T$, which essentially measures the average number of partial waves interfering with each other, is an important quantity. For sufficiently small T , the full width at half-maximum (FWHM) of a resonance is given by $\text{FWHM} = 4 \arcsin[T/(2\sqrt{1-T})]$. For $T_i \ll 1$, the FWHM can be approximated by $\text{FWHM} \approx 2T/\sqrt{1-T}$. Furthermore, in this regime, the system can be approximated by a damped harmonic oscillator, where the oscillation is the wave bouncing back and forth between the two barriers, and the damping is provided by tunneling out of the resonator. Such systems have resonances at an energy E_0 and a homogeneous linewidth of Lorentzian shape, which can be written as

$$T(E) = \frac{\Gamma_1 \Gamma_2}{\frac{1}{4}(\Gamma_1 + \Gamma_2)^2 + (E - E_0)^2} \quad (8.13)$$

7) The electron waves experience phase shifts during the reflection at the barriers, which are neglected here.

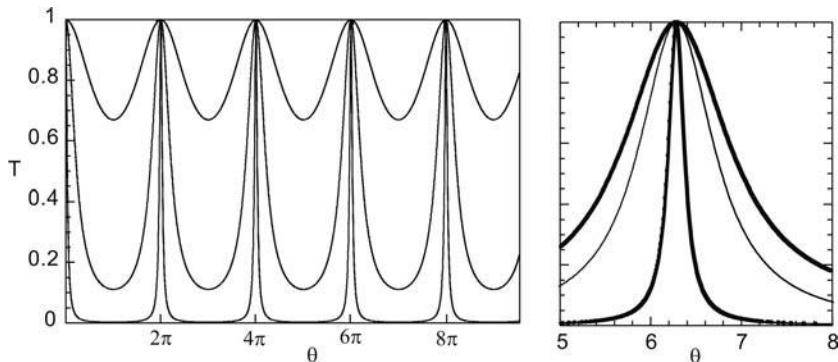


Fig. 8.13 Left: Coherent transmission of a double barrier as a function of the phase collected during one round trip between the barriers, shown for equal individual barrier transmissions $T_b = 0.9, 0.5$, and 0.1 , respectively. Right: Lorentzian fit (bold lines) for $T_i = 0.5$ and 0.1 (thin lines). For the latter barrier transmission, the fit according to Eq. (8.13) is essentially indistinguishable from Eq. (8.13).

Here, Γ_i denotes the coupling constant of barrier i . It is given by $\Gamma_i = \hbar\nu T_i$, and ν is known as the attempt frequency, i.e. the frequency at which the electron hits barrier i and tries to tunnel. It is given by $\nu = v/2L = \hbar k/2Lm^*$, with v being the velocity of the electron. Hence, Γ_i/\hbar represents the tunnel rate, or in other words the number of tunnel events across barrier i per unit time. In addition, the electron phase θ has been mapped onto the electron energy via energy

$$E(\theta) = \frac{\hbar^2 \theta^2}{2m^* L^2}$$

in Eq. (8.13). The quality of this approximation even for not so small T_i is demonstrated in Fig. 8.13.

Thus, the double barrier can be thought of as an electron interferometer. A resonance occurs when the Fermi wavelength is commensurable with L , i.e. $n \times \lambda_F/2 = L$.⁸ Within the s -matrix formalism, this equation is easily obtained by multiplication of the s -matrices for the two barriers with that describing the electron transfer from one barrier to the other.

With s -matrices, we can treat the double barrier transmission in a more general way. The s -matrices of the two individual barriers read

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} r_1 & t_1 \\ t_1 & r_1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad \begin{pmatrix} b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} r_2 & t_2 \\ t_2 & r_2 \end{pmatrix} \begin{pmatrix} a_3 \\ a_4 \end{pmatrix}$$

8) Note that this is not exactly true, since the wave function penetrates into the barrier. However, for small transmission amplitudes, this is an excellent approximation.

while the incoming wave functions are related to the outgoing wave functions via $a_3 = b_2 e^{i\theta}$ and $a_2 = b_3 e^{i\theta}$. Let us further assume that a wave is incoming only from the left with amplitude 1, $a_1 = 1$, and no left-moving wave exists to the right-hand side of the double barrier, $a_4 = 0$. This results in a vector \vec{b} of outgoing amplitudes as a function of incoming amplitudes $\vec{a} = (1, b_3 e^{i\theta}, b_2 e^{i\theta}, 0)$, related by

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} r_1 & t_1 & 0 & 0 \\ t_1 & r_1 & 0 & 0 \\ 0 & 0 & r_2 & t_2 \\ 0 & 0 & t_2 & r_2 \end{pmatrix} \begin{pmatrix} 1 \\ b_3 e^{i\theta} \\ b_2 e^{i\theta} \\ 0 \end{pmatrix}$$

Solving for the transmission amplitude b_4 gives

$$b_4 = \frac{t_1 t_2 e^{i\theta}}{1 - r_1 r_2 e^{2i\theta}}$$

leading to the transmission amplitude $T = b_4^* b_4$ of Eq. (8.12). In this particular example, we could easily guess the result by summing up the interference paths. In more complex structures, however, it may not be so easy to do this, and the s-matrices prove to be very useful. We will see an example of this below.

Note that thermal smearing has been neglected. It will be discussed in Exercise E8.4.

Owing to inelastic scattering events, electrons may lose their phase coherence as they traverse the double barrier. In the case of complete incoherence, we do not have to sum up the transmission amplitudes, but rather the transmission probabilities of all trajectories. In that case, the result is

$$T_{\text{sd}}^{\text{inc}} = T_1 T_2 + T_1 R_2 R_1 T_2 + \dots = \frac{T_1 T_2}{1 - R_1 R_2} \quad (8.14)$$

It should be noted that, in real samples, transport is quite often partly coherent. M. Büttiker found an elegant model for this general situation [46]. The incoherent part of the transmission is modeled by a reservoir in between the barriers, which absorbs and re-ejects those electrons whose phase coherence gets lost.

We conclude this section by discussing the transmission of a quantum ring in terms of the s-matrix formalism. Earlier on, we already studied the transmission of an open ring as a function of the magnetic field, which revealed the Aharonov–Bohm effect. The spectrum of an isolated ring is also well known: in the simplest model, a one-dimensional wire (length $2\pi R$) is bent into a ring, imposing periodic boundary conditions

$$\ell\lambda = 2\pi R, \quad \ell = 0, \pm 1, \pm 2, \dots$$

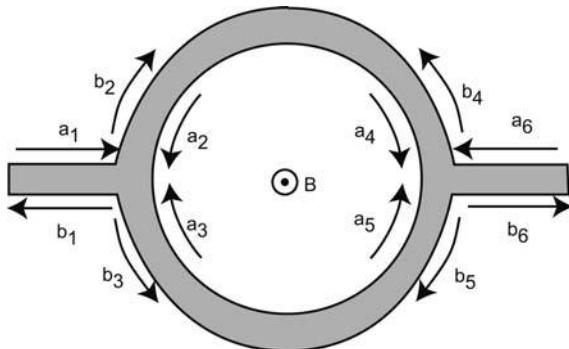


Fig. 8.14 Schematic sketch of the quantum ring under study and the nomenclature of the partial wave functions.

where λ is the electronic wavelength. As a consequence, the wave number is quantized in units of $1/R$. A magnetic field perpendicular to the plane of the ring induces a phase shift of $\Delta\phi = 2\pi\Phi/\Phi_0$, where $\Phi = BA$ is the magnetic flux through the ring (A denotes the ring area), and $\Phi_0 = h/e$ is the magnetic flux quantum. This corresponds to a magnetic wave vector of $k_m = \Delta\phi/2\pi R = (1/R)\Phi/\Phi_0$, and the energy spectrum is given by

$$E_\ell = \frac{\hbar^2}{2m^*R^2}(k_\ell + k_m)^2 = \frac{\hbar^2}{2m^*R^2}(\ell + \Phi/\Phi_0)^2 \quad (8.15)$$

The states are characterized by their angular momentum $\hbar\ell$. This energy spectrum is treated in Exercise E8.3.

Suppose we now couple the ring to two reservoirs to the left and right via tunable tunnel barriers. How will the spectrum of the isolated ring evolve into the Aharonov–Bohm effect observed in open rings? The s-matrices offer a very elegant way to study this evolution. For simplicity, we assume that both tunnel barriers are equal and that the two branches of the ring have the same length (Fig. 8.14).

The junction can be described by the so-called Shapiro matrix

$$(s_{Sh}) = \begin{pmatrix} c & \sqrt{\epsilon} & \sqrt{\epsilon} \\ \sqrt{\epsilon} & a & b \\ \sqrt{\epsilon} & b & a \end{pmatrix}$$

Here c (a) represent the reflection amplitudes for electrons hitting the junction from lead 1 (2 or 3, respectively), while $\sqrt{\epsilon}$ and b are transmission amplitudes. Unitarity of the s-matrix is given for

$$\epsilon = \frac{1}{2}(1 - c^2), \quad a = -\frac{1}{2}(1 + c), \quad b = \frac{1}{2}(1 - c)$$

or

$$\epsilon = \frac{1}{2}(1 - c^2), \quad a = \frac{1}{2}(1 - c), \quad b = -\frac{1}{2}(1 + c)$$

The second set of relations corresponds to two ring branches which become decoupled from each other as c approaches zero. Therefore, the first solution describes the situation of interest. Since c is a measure of the coupling of the ring to the leads, we will express the transmission T_{ring} of the ring as a function of c . The matrix (s_{Sh}) is the s -matrix for the left and right junction. The incoming amplitudes are coupled to the outgoing ones via $\vec{b}_{l,r} = (s_{\text{Sh}})\vec{a}_{l,r}$ with $\vec{b}_l = (b_1, b_2, b_3)$, $\vec{b}_r = (b_4, b_5, b_6)$, $\vec{a}_l = (a_1, a_2, a_3)$, and $\vec{a}_r = (a_4, a_5, a_6)$. As above, we assume a wave incoming from the left only, with amplitude 1, and denote the phase collected from the vector potential by ϕ , such that the incoming and outgoing waves inside the ring are related via

$$\vec{a} = (\vec{a}_l, \vec{a}_r) = (1, b_4 e^{i\theta} e^{i\phi}, b_5 e^{i\theta} e^{-i\phi}, b_2 e^{i\theta} e^{-i\phi}, b_3 e^{i\theta} e^{i\phi}, 0)$$

This leads to the system of equations

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix} = \begin{pmatrix} c & \sqrt{\epsilon} & \sqrt{\epsilon} & 0 & 0 & 0 \\ \sqrt{\epsilon} & a & b & 0 & 0 & 0 \\ \sqrt{\epsilon} & b & a & 0 & 0 & 0 \\ 0 & 0 & 0 & a & b & \sqrt{\epsilon} \\ 0 & 0 & 0 & b & a & \sqrt{\epsilon} \\ 0 & 0 & 0 & \sqrt{\epsilon} & \sqrt{\epsilon} & c \end{pmatrix} \begin{pmatrix} 1 \\ b_4 e^{i\theta} e^{i\phi} \\ b_5 e^{i\theta} e^{-i\phi} \\ b_2 e^{i\theta} e^{-i\phi} \\ b_3 e^{i\theta} e^{i\phi} \\ 0 \end{pmatrix} \quad (8.16)$$

The transmission probability is given by $T_{\text{ring}}(c, \theta, \phi) = b_6^* b_6$. After solving Eq. (8.16) for b_6 and after some algebra, one finds the somewhat lengthy expression

$$T_{\text{ring}}(c, \theta, \phi) = b_6^* b_6 = \frac{16(1 - c^2)^2 \cos^2 \phi \sin^2 \theta}{A + B + C + D + E} \quad (8.17)$$

with

$$\begin{aligned} A &= 5 - 4c + 6c^2 - 4c^3 + 5c^4 \\ B &= (1 + c)^4 \cos^2(2\phi) \\ C &= -4(1 - c)^2(1 + c^2) \cos(2\theta) \\ D &= -2(1 + c)^2 \cos^2 \phi [2(1 + c^2) \cos(2\theta) - (1 - c)^2] \\ E &= 8c^2 \cos(4\theta) \end{aligned}$$

Fig. 8.15 shows how the transmission as a function of the dynamic phase θ and the magnetic phase ϕ evolves as the reflection amplitude is reduced. Fig. 8.15(a) corresponds to an open ring, showing essentially Aharonov–Bohm oscillations. Note that here the phase coherence length is infinite. In order to recover the sinusoidal magneto-oscillations typical for the Aharonov–Bohm effect, we would have to expand Eq. (8.17) in a Fourier series and plot

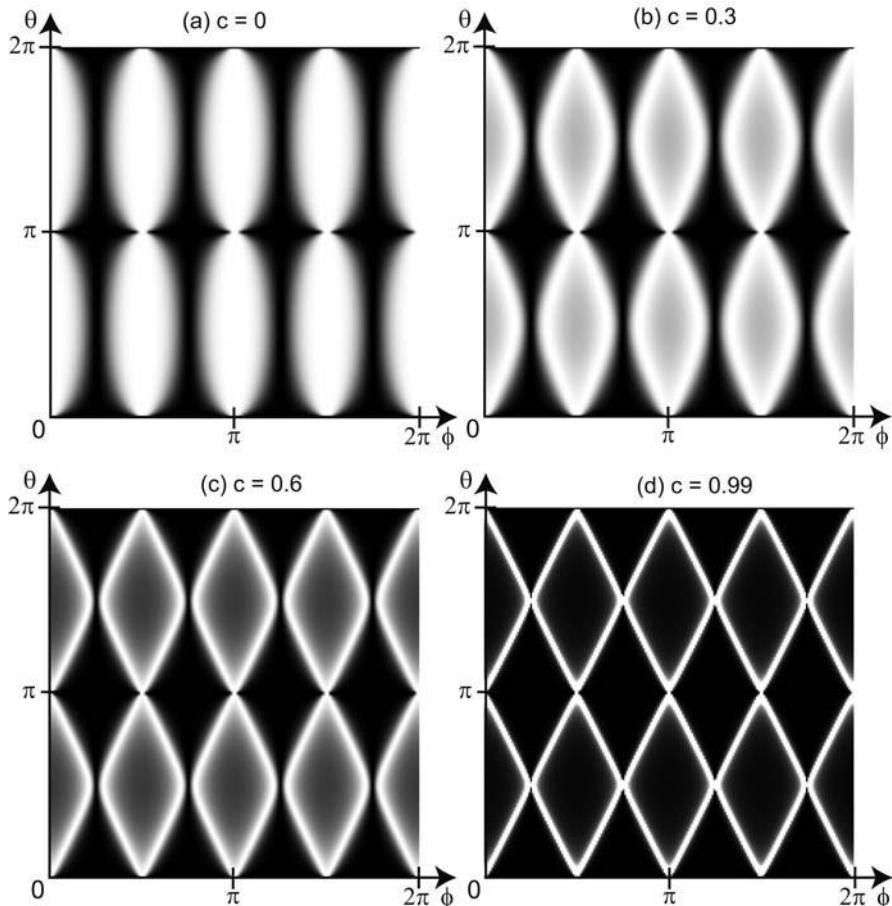


Fig. 8.15 Transmission of an ideal quantum ring as a function of θ and ϕ for different reflection amplitudes at the ring entrances. Black corresponds to $T_{\text{ring}} = 0$, white to $T_{\text{ring}} = 1$.

the first order only. The second order gives the Altshuler–Aronov–Spivak oscillations. Fig. 8.15(d) shows the transmission for a reflection amplitude close to 1 (namely $c = 0.99$). Here, the parabolas of Eq. (8.15) are found (remember that $E \propto \theta^2$). In Figs. 8.15(b) and (c), the transmission is plotted for $c = 0.2$ and 0.4, respectively. Hence, as c increases, the transmission gets more and more concentrated at the edges of the ellipsoidal regions of high transmission in Fig. 8.15(a). Simultaneously, the shape of these ellipsoid-like regions evolves into diamond-like structures.

Papers and Exercises

P8.1 Go through [25]; explain the *flux cancellation effect* and the theoretical expression for the autocorrelation function of UCF in magnetic fields.

P8.2 What is *weak antilocalization*? To answer this question, consult [77].

E8.1 Consider the geometry depicted to the left in Fig. 8.1. Assume a constant magnetic field is present in the z -direction, and calculate the phase difference between partial waves that traverse the upper and the lower branch. Show that $\Phi_{\text{upper}} - \Phi_{\text{lower}} = 2\pi\Phi/\Phi_0$, where $\Phi_0 = h/e$ denotes the magnetic flux quantum.

E8.2 The relation between random events and the Poisson distribution is applicable to many situations. Here, it will be discussed using Eq. (8.9) as an example: within the assumptions described in the text, random electron-electron (e-e) scattering events determine the amplitude of the Aharonov-Bohm type oscillations of Fig. 8.9.

- (a) We denote the average number of e-e scattering events per unit time, i.e. the e-e scattering rate, by γ . Clearly, the exact number of scattering events j within a time t will fluctuate around its average, which equals simply $t\gamma$. What is the probability $P(j)$ that exactly j events take place within t ($P(j)$ is the *Poisson distribution*)? [Hint: Divide the time interval into a large number of sections of equal size, such that more than one event per interval does occur. Count all the possible arrangements of the sections under the constraint that j of them are occupied.]
- (b) Use $P(j)$ to define a meaningful e-e scattering length. How does Eq. (8.9) emerge from this?

E8.3 Consider a ring (radius r) with only one radial mode occupied (i.e. the ring has been formed out of a strictly one-dimensional wire).

- (a) Calculate the energy spectrum of the ring as a function of a homogeneous magnetic field perpendicular to the ring area. It makes life easier to use cylindrical coordinates and gauge the vector potential as $\vec{A} = (0, rB/2, 0)$. Use the wave function ansatz $\Psi(\phi) = (2\pi r)^{-1/2} e^{i\ell\phi}$ with ℓ being an integer. [What is the physical meaning of ℓ ?]
- (b) Calculate the current flowing in the ring as a function of ℓ for zero temperature. How do you interpret this result? Compare the result to a current generated by a single electron circulating in the loop.

- (c) Estimate the current for an odd number of electrons in the ring at $B = 0$. Assume realistic ring diameters and electron densities. How could one measure this current?

E8.4 *Thermal smearing of resonant tunneling peaks.* Calculate, in analogy to our treatment of the thermal smearing of quantized conductance steps in Chapter 7, how a transmission resonance of a double barrier is modified by a non-zero temperature.

- (a) Consider the limiting case of a purely thermally broadened resonance, i.e. $T(\Theta=0, E) = \delta(E - E_r)$. Show that the line shape is the derivative of the Fermi function. Calculate its linewidth (FWHM).
- (b) How does the line shape look like for a Lorentzian-shaped resonance $T(\Theta = 0, E)$? How would you determine experimentally the thermal and the Lorentzian contribution to the line width?

Further Reading

The reader is encouraged to study the excellent treatment of phase coherent electrons in mesoscopic samples in Section 6 of the book by Beenakker and van Houten [27].

9

Single-Electron Tunneling

The charge stored on a capacitor is not quantized: it consists of polarization charges generated by displacing the electron gas with respect to the positive lattice ions and can take arbitrary magnitudes. The charge transfer across a tunnel junction, however, is quantized in units of the electron charge (*single-electron tunneling*), and may be suppressed due to the Coulomb interaction (*Coulomb blockade*). These simple facts lay the foundation for a new type of electronic device called single-electron tunneling (SET) devices. Coulomb blockade was first suggested back in 1951 by Gorter [123], who explained earlier experiments [164]. It remained largely unnoticed until, almost 40 years later, Fulton and Dolan built a transistor based on single-electron tunneling [109]. After introducing the concept of Coulomb blockade in Section 9.1, we will discuss basic single-electron circuits, in particular the double barrier and the single-electron transistor, in Section 9.2. Some examples and applications are given in Section 9.3.

9.1

The principle of Coulomb blockade

Consider a tunnel junction biased by a voltage V . The equivalent circuit of a tunnel junction consists of a “leaky” capacitor, i.e. a resistor R in parallel with a capacitor C (Fig. 9.1). For charges $|q| < e/2$, an electron tunneling across the barrier would increase the energy stored in the capacitor. This effect is known as *Coulomb blockade* [191]. For $|q| > e/2$, the tunneling event reduces the electrostatic energy, and the differential conductance is given by $dI/dV = 1/R$. Experimentally, it is far from easy to observe Coulomb blockade at a single tunnel barrier, for two reasons.

First of all, in order to avoid thermally activated electron transfers, $e^2/(8C) \geq k_B\Theta$ is required.

Question 9.1: A typical tunnel junction patterned by angle evaporation is formed by a thin oxide layer (thickness 5 nm, dielectric constant $\epsilon \approx 5$). Estimate the maxi-

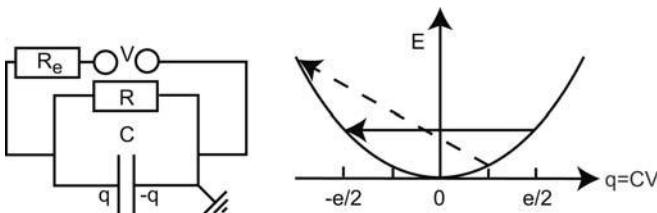


Fig. 9.1 Equivalent circuit and energy diagram of a single tunnel junction. The resistor R_e represents the low-frequency impedance of the environment.

mum area of the capacitor plates for Coulomb blockade to be observed at (a) 4.2 K and (b) 300 K.

Second, the resistance of the tunnel junction has to be “sufficiently large”. We can speak of individual electrons tunneling through the barrier only if the tunnel events do not overlap, which means that the time between two successive events $\delta t \approx eR/V$ must be large compared to the duration τ of a tunnel event, which can be estimated as $\tau \approx \hbar/eV$ [178]. This leads to the condition $R \gg \hbar/e^2$. Furthermore, quantum fluctuations can destroy the Coulomb blockade as well. So far, we have neglected the fact that the tunnel junction is coupled to its environment, which is modeled by the resistance R_e in Fig. 9.1. More generally, the environment represents a frequency-dependent impedance, although here we restrict ourselves to very small frequencies, such that the impedance can be replaced by R_e .

In fact, our above line of arguing implicitly assumes the so-called *local rule*, which states that the tunneling rate across the junction is governed by the difference in electrostatic energy right before and right after the tunnel event. According to the *global rule*, on the other hand, the tunnel rate is determined by the electrostatic energy difference of the whole circuit. Since the environment inevitably includes some capacitances much larger than the capacitance of the tunnel junction, we may expect that, in this case, the Coulomb blockade vanishes.

The influence of the electromagnetic environment on the performance of tunnel junctions is discussed in detail in [125]. Here, we just give a simple argument. The local rule holds provided the tunnel junction is sufficiently decoupled from the environment. In the leads, quantum fluctuations of the charge take place. An estimate based on the Heisenberg uncertainty relation tells us what “sufficiently decoupled” actually means: for quantum fluctuations with a characteristic energy amplitude δE , the uncertainty relation $\delta E \delta t \geq \hbar/2$ holds. Coulomb blockade is only visible for energy fluctuations at the junction much smaller than $e^2/8C$, while the time scale is given by the time constant of the circuit: $\delta t \approx \tau = R_e C$.

Hence, Coulomb blockade can be observed on a single tunnel junction only if the resistance of the environment is of the order of the resistance quantum h/e^2 or higher. The influence of the environmental resistance on the Coulomb blockade has been calculated in [70] and is shown in Fig. 9.2. These considerations imply that it is not so easy to observe Coulomb blockade at a single tunnel junction. Since the environment has to be sufficiently decoupled, the resistance of the leads has to be larger than h/e^2 . This generates Joule heating, which in turn makes it difficult to keep the electron temperature below $e^2/2Ck_B$. Nevertheless, Coulomb blockade has been observed in single tunnel junctions biased via wires of sufficiently high resistance (Fig. 9.3).

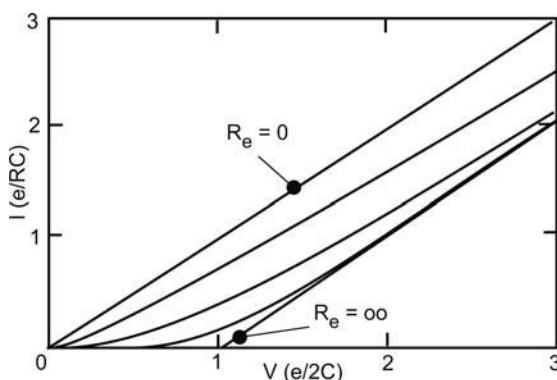


Fig. 9.2 Evolution of the I - V characteristic of a single tunnel junction as the resistance of the environment R_e is increased. For $R_e > h/2e^2$, the Coulomb gap becomes clearly visible. The traces are shown for $R_e/R = 0, 0.1, 1, 10$, and ∞ . After [70].

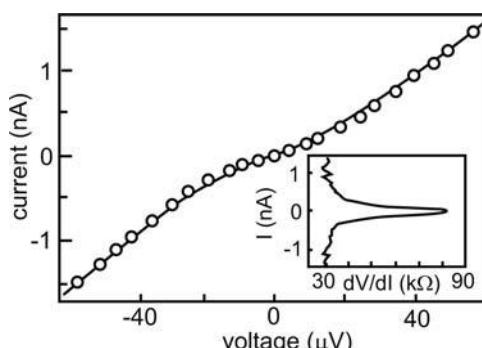


Fig. 9.3 The I - V characteristic of Al-Al₂O₃-Al tunnel barriers, fabricated by angle evaporation. In order to suppress quantum fluctuations, the cross section of the Al wires is only $10\text{ nm} \times 10\text{ nm}$. The superconductive state has been destroyed by applying a magnetic field. After [57].

The limitations imposed by the need to decouple the environment from the tunnel junction can be relaxed by using two tunnel junctions in series (Fig. 9.4), since here quantum fluctuations at the island in between the junctions are strongly suppressed [125]. The number of electrons at the enclosed island can change only by tunneling across one of the barriers, an event essentially free of dissipation. The energy relaxation will take place somewhere in the leads, far away from the island. The resistance of relevance for the suppression of the quantum fluctuations is now that of a tunnel barrier, while the capacitance corresponds to the total capacitance of the island to its environment. Therefore, quantum fluctuations at the island can be suppressed easily without running into heating problems.

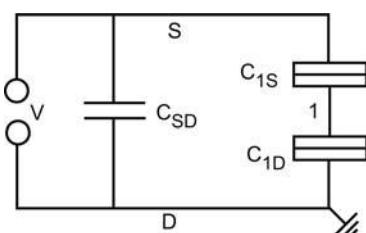


Fig. 9.4 A double barrier structure attached to source (S) and drain (D). C_{SD} denotes a residual capacitance between the two leads.

Question 9.2: The self-capacitance of a metallic grain is sometimes estimated by $C_{self} = V/q$, where V denotes the potential of the grain and q the charge transferred onto it from infinity (at zero potential). For a sphere, C_{self} equals $4\pi\epsilon_0 r$, whereas, for a circular disk, $C_{self} = 8\epsilon\epsilon_0 r$ (r denotes the radius of the island). Estimate C_{self} and the charging energy for some reasonable grain radii.

9.2

Basic single-electron tunneling circuits

Before we discuss single-electron tunneling in the double barrier system, it is useful to have a look at the problem from a more general point of view, which is then used to analyze specific examples including the circuit of Fig. 9.4.

Consider an arrangement of $(n + m)$ conductors embedded in some insulating environment. Each conductor i is at an electrostatic potential V_i , has a charge q_i stored on it, and has a capacitance C_{iD} to drain (ground).¹ Between

¹) In publications, one frequently encounters an “antisymmetric bias condition”, where a voltage of $V_S = +V/2$ is applied to the source, and the drain voltage is $V_D = -V/2$. The electrostatics is different in that case.

each pair of conductors i and j , there is a mutual capacitance C_{ij} . Some of these capacitances may belong to tunnel junctions, which allow electron transfers between the corresponding conductors. Furthermore, we assume that m conductors are connected to voltage sources, which we call *electrodes*, while the n remaining ones are *islands*.² For convenience, we enumerate the n islands from 1 to n , and the m electrodes from $n+1$ to $n+m$.

The charges and potentials of the islands can be written in terms of an island charge vector \vec{q}_I and potential vector \vec{V}_I , respectively. Similarly, charge and potential vectors can be written down for the electrodes, \vec{q}_E and \vec{V}_E . The state of the system can be specified by the total charge vector $\vec{q} = (\vec{q}_I, \vec{q}_E)$. Equivalently, it can be characterized by the total potential vector defined as $\vec{V} = (\vec{V}_I, \vec{V}_E)$. Charge and potential vectors are related via the capacitance matrix $\underline{\mathcal{C}}$:

$$\vec{q} = \underline{\mathcal{C}} \vec{V} \quad (9.1)$$

We write $\underline{\mathcal{C}}$ as

$$\underline{\mathcal{C}} = \begin{pmatrix} \mathcal{C}_{II} & \mathcal{C}_{IE} \\ \mathcal{C}_{EI} & \mathcal{C}_{EE} \end{pmatrix} \quad (9.2)$$

The capacitance submatrices between type A and type B conductors (A, B can be electrodes or islands) are denoted by \mathcal{C}_{AB} . Note that the ground is *not* a conductor in terms of our definition, and that $\underline{\mathcal{C}}$ is symmetric. The matrix elements of $\underline{\mathcal{C}}$ are given by (see Appendix B)

$$(\underline{\mathcal{C}})_{ij} = \begin{cases} -C_{ij} & j = 1, \dots, n+m; j \neq i \\ C_{iD} + \sum_{k=1; k \neq i}^{n+m} C_{ik} & j = i \end{cases}$$

The electrostatic energy³ E is given by the energy stored at the islands, minus the work done by the voltage sources. Minimizing this energy gives us the ground state.

As we shall see, in single-electron circuits, usually the voltages applied to the electrodes are parametrically changed, and the initial island charge vector \vec{q}_I given. As \vec{V}_E is changed, the potential difference between two conductors connected by a tunnel junction may become sufficiently large for electrons to tunnel, resulting in a new charge configuration. Such charge rearrangements will take place as soon as the electrostatic energy of the new configuration is

- 2) The electrostatics of such systems in terms of the capacitance matrix is discussed in Appendix C.
- 3) The electrostatic energy is the free energy $E = U - \mu N$, where U is the total energy, μ is the electrochemical potential, and N is the number of electrons.

equal to, or smaller than, the energy of the original configuration. The charge transfer can be specified by the change of the charge vector $\Delta\vec{q} = \vec{q}_{\text{new}} - \vec{q}$. For a system initially in its ground state, we can find the parametric transition to a new ground state from the condition

$$\Delta E = E_{\text{new}} - E \leq 0 \quad (9.3)$$

It may look very cumbersome to calculate the energy differences of all the possible charge transfers and find its minimum. Usually, however, only very few electron transfers have to be considered.

In Eq. (9.3) ΔE is given by⁴

$$\Delta E[\vec{V}_E, \vec{q}_I, \Delta\vec{q}] = \Delta\vec{q}_I \underline{C}_{II}^{-1} [\vec{q}_I + \frac{1}{2}\Delta\vec{q}_I - \underline{C}_{IE} \vec{V}_E] + \Delta\vec{q}_E \vec{V}_E \quad (9.4)$$

This equation is an important relation, which can be used to analyze Coulomb blockade in all systems that can be characterized by a capacitance matrix. Note that it cannot be used to study Coulomb blockade at the single junction, since the crucial time scale involved there does not enter the formalism leading to Eq. (9.4). We are now ready to study the double barrier shown in Fig. 9.4.

9.2.1

Coulomb blockade at the double barrier

The system consists of one electrode (source S) and one island (1). In the following, islands will be labeled by arabic numbers and electrodes by capital letters. The capacitance matrix reads

$$\underline{C} = \begin{pmatrix} C_{11} & -C_{1S} \\ -C_{1S} & C_{SS} \end{pmatrix}$$

with $C_{11} = C_{1S} + C_{1D}$ and $C_{SS} = C_{1S} + C_{SD}$. The charge on the island is given by the number n of electrons tunneled onto it, plus an arbitrary background charge q_0 , induced by the environment: $q = q_0 - ne$. Four different charge transfers are relevant. An electron can hop in both directions across C_{1S} or C_{1D} . For electron transfers across C_{1S} , we have $\vec{V} = (V_1, V)$, $\vec{q} = (q_0 - ne, q_S)$, and $\Delta\vec{q} = \pm e(-1, 1)$. Here “+ (−)” corresponds to a transfer of one electron from S to 1 (1 to S). Consequently, the energy difference reads, according to Eq. (9.4),

$$\Delta E[V, q_0 - ne, \pm e(-1, 1)] = \frac{e}{C_{11}} \left[\frac{e}{2} \pm (ne - q_0 + C_{1D}V) \right] \quad (9.5)$$

For tunnel events across C_{1D} , $\Delta\vec{q} = \pm e(-1, 0)$. Here, “+ (−)” corresponds to a transfer of one electron from D to 1 (1 to D). This gives

$$\Delta E[V, q_0 - ne, \pm e(-1, 0)] = \frac{e}{C_{11}} \left[\frac{e}{2} \pm (ne - q_0 - C_{1S}V) \right] \quad (9.6)$$

4) For a derivation of Eq. (9.4), see Appendix C.

Coulomb blockade is established only if all four energy differences are positive. This defines a voltage interval of vanishing current:

$$\begin{aligned} \text{Max}\left\{\frac{1}{C_{1S}}[-q_0 + e(n - \frac{1}{2})], \frac{1}{C_{1D}}[q_0 - e(n + \frac{1}{2})]\right\} \\ < V < \text{Min}\left\{\frac{1}{C_{1S}}[-q_0 + e(n + \frac{1}{2})], \frac{1}{C_{1D}}[q_0 - e(n - \frac{1}{2})]\right\} \quad (9.7) \end{aligned}$$

Let us study some special scenarios.

1. *No background charges.* The simplest situation is $n = 0$, no background charges ($q_0 = 0$), and identical junction capacitances $C_{1S} = C_{1D} = C_{11}/2$. Now Eq. (9.7) reads $-e/C_{11} \leq V \leq e/C_{11}$. For $V = 0$, we get

$$\Delta E[0, 0, e(\mp 1, \pm 1)] = \Delta E[0, 0, e(\pm 1, 0)] = e^2/(2C_{11})$$

All four charge transfer processes are suppressed (Fig. 9.5(a)). Applying a positive voltage $V = e/C_{11}$ to the source means that

$$\begin{aligned} \Delta E[V, 0, e(-1, 1)] &= e^2/C_{11} > 0 \\ \Delta E[V, 0, e(1, -1)] &= 0 = \Delta E[V, 0, e(-1, 0)] \end{aligned}$$

and

$$\Delta E[V, 0, e(1, 0)] = e^2/C_{11} > 0$$

At this voltage, an electron can either tunnel from drain to the island or from the island to source (Fig. 9.5(b)). Both processes have the same probability.

Question 9.3: Suppose that an electron has just tunneled from drain onto the island under these conditions. The system is in the state depicted in Fig. 9.5(b). Show that, now, an electron will tunnel from the island to source, and a current is established. Calculate the energy differences indicated in Fig. 9.5(c).

The system thus oscillates between the situations depicted in Figs. 9.5(b) and (c). In each oscillation cycle, a single electron is transferred from drain to source. In addition, the tunnel events show a pair correlation. Shortly after an electron has tunneled from drain to the island, a tunneling process from the island to drain will take place, and vice versa.

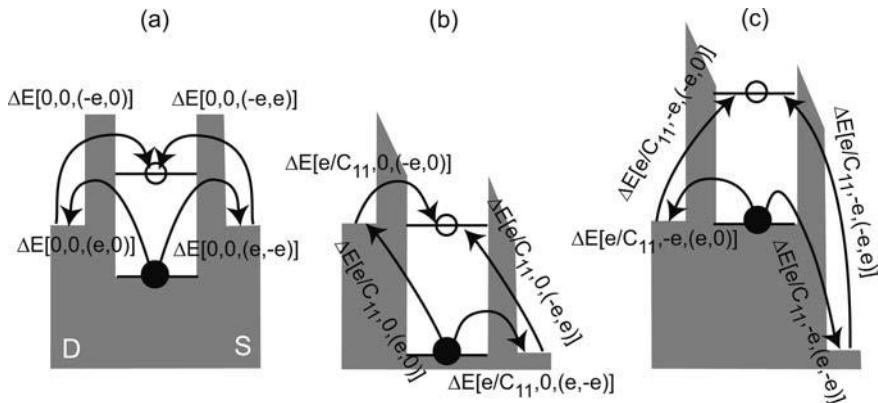


Fig. 9.5 Energy differences of the four electron transfers at the double barrier. Open circles denote empty states, while full circles correspond to occupied states. (a) No voltage is applied ($V = 0$), and Coulomb blockade is established. (b) $V = e/C_{11}$.

Electrons can hop from drain onto the island, as well as from the island to source. (c) Differences in the electrostatic energy after an electron has, starting from the situation in (b), tunneled from drain onto the island.

2. *Effect of a background charge q_0 .* Let us assume that $n = 0$, and $C_{1S} = C_{1D}$, which leads to the condition for Coulomb blockade

$$\begin{aligned} \text{Max}\left\{\frac{2}{C_{11}}\left(-q_0 - \frac{e}{2}\right), \frac{2}{C_{11}}\left(q_0 - \frac{e}{2}\right)\right\} \\ < V < \text{Min}\left\{\frac{2}{C_{11}}\left(-q_0 + \frac{e}{2}\right), \frac{2}{C_{11}}\left(q_0 + \frac{e}{2}\right)\right\} \end{aligned}$$

This means that, by a non-zero q_0 , the Coulomb gap can be reduced, but never be increased. In fact, for $q_0 = (j + \frac{1}{2})e$ with j being an integer, the Coulomb gap vanishes completely. Background charges can seriously hamper the observation of the Coulomb blockade, especially when they are time-dependent.

Question 9.4: Draw the energy diagram corresponding to Figs. 9.5(a)–(c) for $q_0 = e/4$. Assume equal capacitances.

Question 9.5: Show that for $C_{1S} \neq C_{1D}$, the larger capacitance determines the Coulomb gap, which gets reduced compared to the Coulomb gap for identical junctions.

Coulomb blockade in metallic islands has been known for a long time. As an example of the early indications, we take a look at an experiment of Giaever and Zeller [117]. The authors measured the current–voltage characteristic of a granular Sn film sandwiched between an oxide layer and metallic electrodes (Fig. 9.6). The average diameter of the Sn granules was 11 nm, such that single-electron tunneling is expected to play a role at low temperatures. The system contains an ensemble of double barriers in parallel. Therefore, we expect to observe a gap in the I – V characteristic around $V = 0$ that corresponds to the average single-electron charging energy. Leakage currents through the oxide in between the islands are quite small, since the conductance of tunnel barriers decreases exponentially with increasing barrier thickness. At zero magnetic field, both the Al electrodes as well as the Sn granules are in the superconductive state, and the superconductive energy gap strongly influences the transport measurements.⁵ However, by applying a magnetic field, the superconductive state is destroyed and our previous model becomes applicable. The Coulomb gap manifests itself in an increased differential resistance around $V = 0$, compared to that observed at larger voltages.

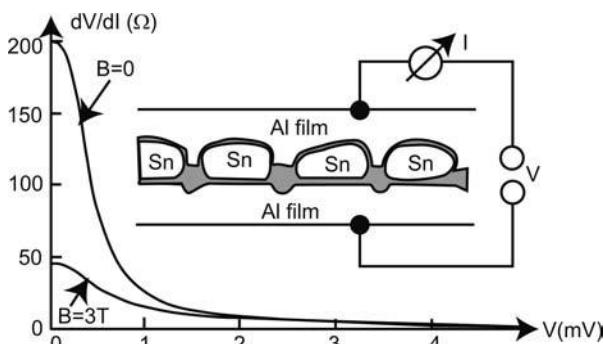


Fig. 9.6 The experiment of Giaever and Zeller. After [117]. A granular Sn film was embedded in an oxide layer and covered on both sides by Al, which acted as source and drain.

9.2.2

Current–voltage characteristics: The Coulomb staircase

Besides the Coulomb gap around $V = 0$, the Coulomb blockade generates under certain conditions a staircase-like structure in the current–voltage characteristic, known as a *Coulomb staircase*. In contrast to our earlier considerations concerning transport through mesoscopic structures, we study here a system of interacting electrons, and a charge transfer changes the electrostatic energy

5) Some information about the interplay of superconductivity and single-electron tunneling can be gained from Paper P10.4.

as well. To include the interaction, we use the so-called transfer Hamiltonian model, which allows us to relate the change in energy ΔE due to a tunnel event with a tunnel rate $\Gamma(\Delta E)$. For the transmission coefficients calculated in earlier chapters, we always assumed that the energy is conserved. Here, however, the electrostatic energy changes as an electron tunnels, and the voltage sources do some work on the system.

Such situations can be conveniently dealt with by using Fermi's golden rule, which originates in time-dependent perturbation theory. The transfer Hamiltonian model starts from an impenetrable barrier, separating two electron gases. Tunneling is treated as a perturbation and is described by a perturbation Hamiltonian H_t , which is of no further interest to us here. The interested reader is referred to [90] for details. Applied to a tunnel barrier, Fermi's golden rule states that the transition rate for an electron in the initial state $|i\rangle$ to a final state $|f\rangle$ on the other side of the tunnel barrier is given by

$$\Gamma_{i \rightarrow f} = \frac{2\pi}{\hbar} |\langle i | H_t | f \rangle|^2 \delta(E_f - E_i - \Delta E) \quad (9.8)$$

Here, E_i and E_f denote the energies of the initial and final states with respect to the bottom of the conduction band, and the matrix element $\langle i | H_t | f \rangle$ describes the coupling of the left-hand side to the right-hand side of the tunnel barrier. This transition rate is just the transmission probability per unit time. In order to determine the total transition rate $\Gamma(\Delta E)$, we have to make the following considerations.

1. The tunneling rate at energy E will be proportional to the spectral electron density $n(e) = D_i(E)f(E)$. Here the index i denotes the side of the barrier that hosts state i , D_i is the relevant density of states, and $f(E)$ denotes the Fermi–Dirac distribution function.
2. Since we are dealing with fermions, the electrons can tunnel only into an empty state $|f\rangle$. The transfer rate for an electron in $|i\rangle$ will thus be proportional to $D_f(E + \Delta E)[1 - f(E + \Delta E)]$.
3. We have to integrate over all energies at which states with non-zero tunneling probability exist. These are all the states above the maximum of the conduction band bottoms on both sides $E_{cb,max}$.

Therefore, the total transition rate is given by

$$\begin{aligned} \Gamma_{1 \rightarrow 2}(\Delta E) &= \frac{2\pi}{\hbar} \int_{E_{cb,max}}^{\infty} |\langle i | H_t | f \rangle|^2 D_i(E) D_f(E - \Delta E) \\ &\times f(E)[1 - f(E - \Delta E)] dE \end{aligned} \quad (9.9)$$

Now, 1 and 2 denote the conductors that contain the initial and final states, respectively. For large energy barriers, we can safely assume that the matrix elements of H_t will be approximately independent of energy. Second, we assume that the density of states does not depend on energy, either since the electron gas is two-dimensional, or since the voltage drop is sufficiently small. Furthermore,

$$f(E)[1 - f(E - \Delta E)] = \frac{f(E) - f(E - \Delta E)}{1 - \exp(\Delta E/k_B\Theta)}$$

If we further consider only cases where the temperature is sufficiently low, we can approximate the Fermi functions by step functions, and obtain

$$\Gamma_{1 \rightarrow 2}(\Delta E) = \frac{1}{Re^2} \frac{\Delta E}{1 - \exp(\Delta E/k_B\Theta)} \quad (9.10)$$

Here, the resistance R of the tunnel barrier has been defined as

$$R = \frac{\hbar}{2\pi e^2 |\langle i|H_t|f\rangle|^2 D^2} \quad (9.11)$$

(see Exercise E9.2). The current is then obtained from the difference of tunnel rates in both directions,

$$I = e[\Gamma_{1 \rightarrow 2}(\Delta E_{1 \rightarrow 2}) - \Gamma_{2 \rightarrow 1}(\Delta E_{2 \rightarrow 1})]$$

Let us apply this result to the island of Fig. 9.4. For a steady state, the average charge at the island is constant, and the current from source to the island is given by

$$I(V) = e \sum_{n=-\infty}^{\infty} p(n)[\Gamma_{1 \rightarrow S}(\Delta E_{1 \rightarrow S}(n)) - \Gamma_{S \rightarrow 1}(\Delta E_{S \rightarrow 1}(n))] \quad (9.12)$$

Equivalently, $I(V)$ can be expressed in terms of the drain tunneling rates. Here, we denote the tunneling rate from 1 to source by $\Gamma_{1 \rightarrow S}(\Delta E_{1 \rightarrow S}(n))$, while the reverse process is denoted accordingly.

Of course, the energy differences now depend on the number of excess electrons n stored on the island. The probability of finding n electrons on the island is denoted by $p(n)$. We expect this function to be peaked around one number, which is given by the sample parameters and by V . The steady state condition furthermore requires that the probability for making a transition between two charge states (characterized by n) is zero. This means that the rate of electrons entering the island occupied by n electrons equals the rate of electrons leaving the island when occupied by $(n + 1)$ electrons:

$$\begin{aligned} & p(n)[\Gamma_{1 \rightarrow S}(\Delta E_{1 \rightarrow S}(n)) + \Gamma_{1 \rightarrow D}(\Delta E_{1 \rightarrow D}(n))] \\ &= p(n+1)[\Gamma_{S \rightarrow 1}(\Delta E_{S \rightarrow 1}(n+1)) + \Gamma_{D \rightarrow 1}(\Delta E_{D \rightarrow 1}(n+1))] \end{aligned} \quad (9.13)$$

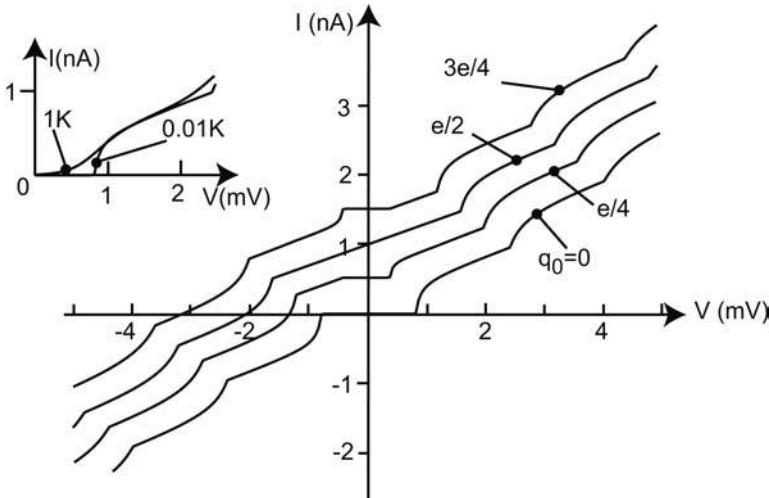


Fig. 9.7 Coulomb staircase as calculated from Eq. (9.12), for different background charges q_0 . The structure is periodic in q_0 , with a period of one elementary charge. Typical sample parameters have been assumed, namely $C_{1S} = C_{1D} = 0.1 \text{ fF}$, $R_{1S} = 20 \text{ M}\Omega$, $R_{1D} = 1 \text{ M}\Omega$, at a temperature of $T = 10 \text{ mK}$. The inset shows the thermal smearing of the Coulomb gap (for $q_0 = 0$) as the temperature is increased to 1 K.

We are now ready to calculate the $I(V)$ characteristic. Equation (9.13), together with the normalization condition

$$\sum_{n=-\infty}^{\infty} p(n) = 1$$

allows us to obtain $p(n)$, which we insert in Eq. (9.12). This requires some numerics, which is considerably simplified by the fact that only a few occupation numbers have non-vanishing probabilities.

Fig. 9.7 shows staircases calculated from Eq. (9.12) for different background charges. The staircases are periodic in q_0 with a period of one elementary charge. Qualitatively, the staircase can be understood as follows: Suppose the tunnel rate across junction S is much larger than that across junction D, and the voltage applied is positive. The voltage now drops completely across junction D, i.e. $V_{1D} \approx V$. From Eq. (9.4), we calculate from $\Delta E[V, -ne, e(-1, 0)] = 0$ the threshold voltages $V(n_0)$ and $V(n_0 + 1)$, which differ by $\Delta V = e/C_{1S} \approx e/C_{11}$. If the voltage is increased by this amount, an additional electron can jump on the island via the drain junction. This increases the current (which is governed by $\Gamma_{1 \rightarrow D}$ and by $\Gamma_{D \rightarrow 1}$) by $\Delta I = e/R_{1D}C_{11}$ for sufficiently low temperatures, as can be seen by inserting

$$e\Delta V = \Delta E[V, -(n+1)e, e(-1, 0)] - \Delta E[V, -ne, e(-1, 0)]$$

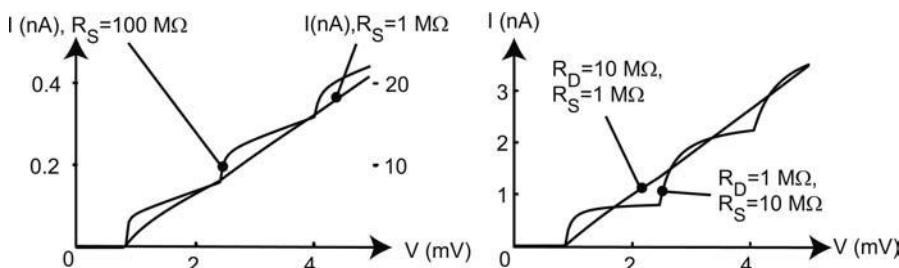


Fig. 9.8 Steps of the Coulomb staircase for various sample parameters, as calculated for $T = 10 \text{ mK}$. Left: $C_{1S} = C_{1D} = 0.1 \text{ fF}$, $R_{1D} = 1 \text{ M}\Omega$. For $R_{1S} = R_{1D}$, the steps are absent, while for $R_{1S} = 100R_{1D}$, they are quite pronounced. Right: Coulomb staircase of an island with two junctions of both different capacitances and different resistances, i.e. $C_{1S} = 0.1 \text{ fF}$, $C_{1D} = 1 \text{ aF}$.

in Eq. (9.12). The markedness of the staircase steps strongly depends on the sample parameters (Fig. 9.8). The steps become most pronounced if both the resistance and the capacitance of one junction are large compared to those of the second junction. Experimentally, however, this is hard to achieve, since small tunnel resistances tend to correspond to small capacitances as well. An analytical model for the Coulomb staircase in this limit is discussed in Paper P9.2.

Particularly beautiful Coulomb staircases have been observed in scanning tunneling experiments on clusters, where the experimental setup consists of a conducting granule or cluster, deposited on an insulating layer on top of a conducting substrate. The tip of a scanning tunneling microscope (STM) is positioned on top of the cluster (Fig. 9.9(a)) and the current is measured as a function of the voltage applied to the STM tip with respect to the substrate [6]. In such experiments, the resistance of one barrier is given by the distance between tip and granule, which can be changed over a wide range. Fig. 9.9(b) shows typical experimental data.

9.2.3

The SET transistor

In 1987, Fulton and Dolan [109] published a seminal experiment: By angle evaporation, a small metallic island was patterned, coupled to source and drain via tunnel barriers with cross sections in the range of $50 \text{ nm} \times 50 \text{ nm}$. In addition, a third electrode (the *gate electrode*) was defined such that the gate-island resistance approaches infinity, and thus couples to the island only capacitively. In this way, the effective background charge and thus the width of the Coulomb gap can be tuned continuously with the gate voltage, and, for sufficiently small source-drain voltages, the current flowing from source

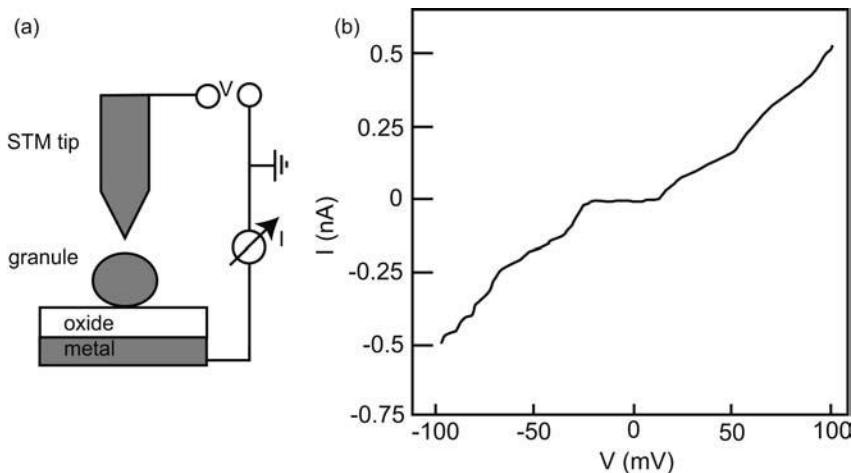


Fig. 9.9 (a) One experimental setup for measuring the Coulomb staircase. (b) Experimental data, a least squares fit of which gives the parameters $C_S = 2 \text{ aF}$, $C_D = 4.14 \text{ aF}$, $R_S = 34.9 \text{ M}\Omega$, $R_D = 132 \text{ M}\Omega$, and an offset charge of $0.12e$. Here, the granule was a small indium droplet on top of an oxidized conducting substrate. The temperature was 4.2 K . The measurement is adapted from [6].

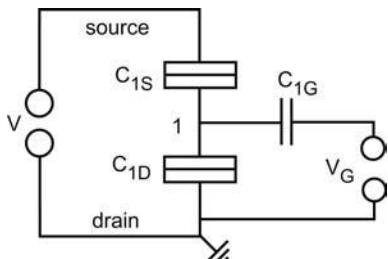


Fig. 9.10 Schematic diagram of a SET transistor.

to drain can be controlled. The system constitutes a transistor based on the Coulomb blockade and is known as a single-electron tunneling (SET) transistor. Its equivalent circuit is shown in Fig. 9.10.

For simplicity, let us assume that the background charge vanishes for zero gate voltage. This is no restriction of generality, since additional background charges can always be compensated for by a gate voltage offset. The inverse capacitance matrix now reads $(\underline{\mathcal{C}}^{-1})_{11} = 1/C_{11}$, and $(\underline{\mathcal{C}}^{-1})_{ij} = 0$ otherwise. Furthermore, $C_{IE} = (-C_{1S}, -C_{1G})$. The electrode voltage vector is given by $\vec{V}_E = (V, V_G)$, while the island charge vector reads $\vec{q}_I = -ne$. The Coulomb gap is given by the onset of the same tunneling events as for the single island studied above. Now, however, the Coulomb gap depends upon the gate

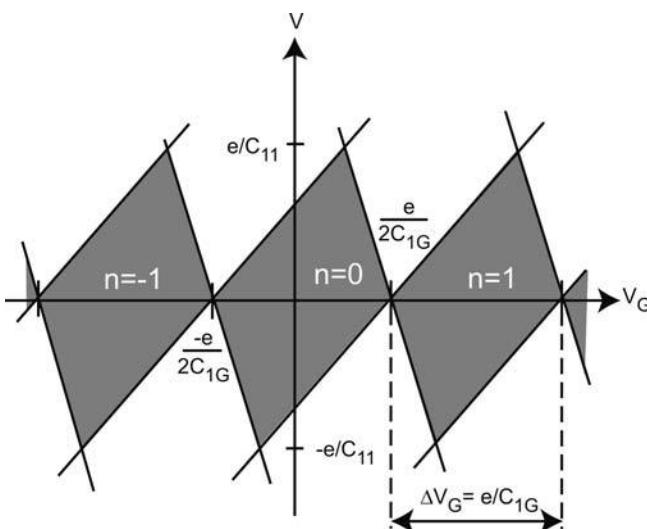


Fig. 9.11 Stability diagram of a single-electron transistor. Within the diamonds, Coulomb blockade is established, while outside, a current flows between source and drain. The slopes of the boundaries are given by $C_{1G}/(C_{11} - C_{1S})$, and by $-C_{1G}/C_{1S}$, respectively.

voltage. The corresponding energy differences are

$$\begin{aligned}\Delta E[(V, V_G), -ne, \pm e(-1, 1)] &= \frac{e}{C_{11}} [\frac{1}{2}e \pm (C_{11} - C_{1S})V \pm ne \mp C_{1G}V_G] \\ \Delta E[(V, V_G), -ne, \pm e(-1, 0)] &= \frac{e}{C_{11}} [\frac{1}{2}e \mp C_{1S}V \pm ne \mp C_{1G}V_G]\end{aligned}$$

Coulomb blockade is established if all four energy differences are positive. For each n , this condition defines a stable, diamond-shaped region in the (V_G, V) plane, with the four boundaries given by the onset conditions:

$$\begin{aligned}\Delta E[(V, V_G), -ne, \pm e(-1, 1)] = 0 \quad &\Rightarrow \\ V(V_G, n) &= \frac{C_{1G}}{C_{11} - C_{1S}}V_G - \frac{e(n \pm \frac{1}{2})}{C_{11} - C_{1S}} \\ \Delta E[(V, V_G), -ne, \pm e(-1, 0)] = 0 \quad &\Rightarrow \\ V(V_G, n) &= -\frac{C_{1G}}{C_S}V_G + \frac{e(n \pm \frac{1}{2})}{C_{1S}}\end{aligned}\tag{9.14}$$

These stable regions are known as Coulomb diamonds, and line up along the V_G -axis (Fig. 9.11).

Fig. 9.12 shows a measurement of the stability diagram of a Al-Al₂O₃ single-electron transistor. The experimentally obtained shape of the Coulomb diamonds, as well as the current–voltage characteristic, agree very well with

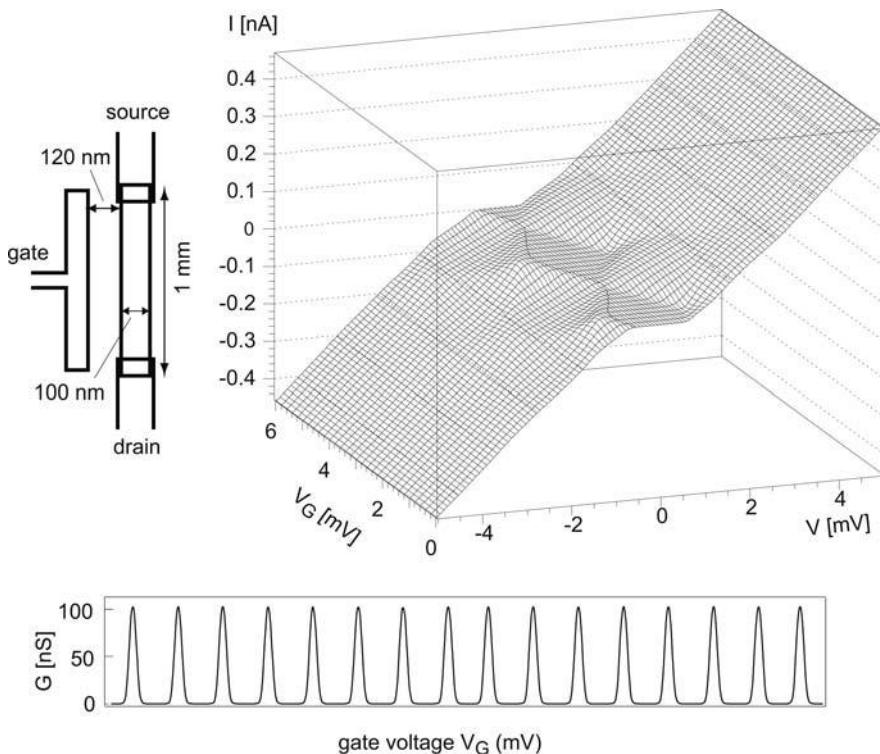


Fig. 9.12 Stability diagram of an Al-Al₂O₃ SET transistor (its dimensions are shown to the left), measured at a temperature of 30 mK. At the bottom, Coulomb blockade oscillations are shown for $V = 10 \mu\text{V}$. Adapted from [111].

the model just developed. For $|V| < e/C_{11}$, the current oscillates strongly as a function of the gate voltage, an effect known as *Coulomb blockade oscillations*. Current peaks occur at $V_G = (e/C_{1G})(n + \frac{1}{2})$. In each gate voltage period $\Delta V_G = e/C_{1G}$, n changes by one. It is important to point out that these oscillations have nothing to do with resonant tunneling. Neither did we assume phase coherence, nor does the nearest-neighbor spacing of the energy levels have to be larger than $k_B\Theta$! In fact, for the system shown in Fig. 9.12, the level spacing is well below 1 μeV. We shall see in the following chapter on quantum dots how single-electron tunneling coexists with size quantization. The weak structures outside the diamonds correspond to Coulomb staircases for each gate voltage, telling us that the two tunnel barriers are not identical.

The line shape of the Coulomb blockade resonances in the limit of negligible source-drain voltage has been derived in [184] and in [23]. The typical experimental situation is characterized by $h\Gamma \ll \Delta \ll k_B\Theta \ll E_C$. This is known as the *metallic regime*. Here, Γ denotes the coupling of the island to the

leads, while Δ is the spacing of the discrete (kinetic) energy levels of the island. Coulomb blockade is well pronounced in this regime, but many energy levels carry current. The line shape of the conductance resonances is given to a good approximation by

$$G(E) = \frac{e^2 D_{\text{island}}}{2} \frac{\Gamma^S \Gamma^D}{\Gamma^S + \Gamma^D} \cosh^{-2} \left(\frac{E - E_{\max}}{2.5 k_B \Theta} \right) \quad (9.15)$$

Here, D_{island} is the density of states in the dot, $\Gamma^{\text{S}, \text{D}}$ denote the couplings of the dot to source and drain, while E_{\max} is the energy at the peak amplitude. Note that the gate voltage can be transformed into an energy via $\delta E = eC_{1G}/C_{11}\delta V_G$. Increasing the temperature thus broadens the resonances, but does not change the peak conductance. Since the conductance of an individual energy level of the island scales as $1/\Theta$ (see Exercise E8.4), and the number of contributing states is proportional to Θ , the total temperature dependence of the peak conductance just cancels [23].

It is important to realize that Coulomb oscillations do not measure the density of states of the island, but the addition spectrum. The density of states tells us how many electrons can be in the system at a particular energy, for a *fixed* number of electrons. The addition spectrum, on the other hand, tells us at which energies electrons can be *added to* the system. If the system is interacting, these two quantities are different, a fact that is clearly demonstrated here. Besides being a somewhat unconventional transistor with an oscillatory transconductance dI/dV_G , this device is extremely sensitive to charges in the vicinity of the island and can thus be used as an electrometer, as used, for example, to study the electrochemical potential in semiconductor heterostructures [161, 321]. Particularly appealing is the integration of a SET transistor in the tip of a scanning probe microscope, which results in an electrometer of both high spatial and charge resolution [131, 340]. The charge resolution is ultimately limited by shot noise; a sensitivity of 10^{-4} electrons has been demonstrated experimentally in [348].

Question 9.6: Estimate the charge resolution δq achievable with the single-electron transistor of Fig. 9.12. Assume the operation point is in the wing of a Coulomb blockade resonance, and assume a current resolution of 10fA .

In transistor operation, its advantage is low power consumption, since, for switching, the charge needs to be changed by only a small fraction of e . Schemes for a digital logic based on single-electron tunneling have been developed, and experimental implementations are being investigated [7, 178]. One problem is to reduce the island size sufficiently in order to operate the devices at room temperature. To date, there are several reports on SET tran-

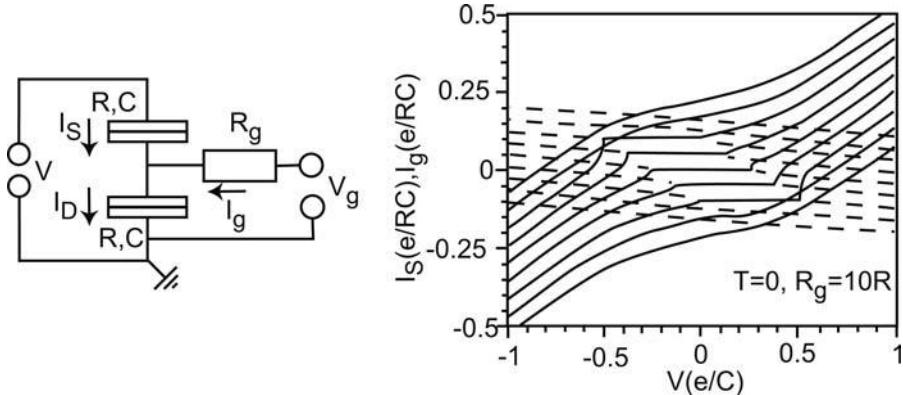


Fig. 9.13 Current–voltage characteristics of a resistively coupled single-electron transistor. Shown is both the source current (solid lines) and the gate current (dashed lines) for V_G varying from $-e/2C$ to $e/2C$ in steps of $e/8C$. The traces are offset vertically for clarity.
(Adapted from [179].)

sistors operating at room temperature (see e.g. [275]), but production of such devices is by no means standard. In addition, the switching is strongly disturbed by fluctuating background charges, although a charge stability of 0.01 elementary charges over weeks has been demonstrated in silicon-based SET transistors [349]. Furthermore, the voltage gain in such transistors is limited.

These limitations can be overcome, in principle, by using resistively coupled single-electron transistors. The circuit is shown in Fig. 9.13: the gate couples to the island via a gate resistance $R_G \gg h/(2e^2)$. In describing this device, Eq. (9.10) has to be modified, since charge can also flow from the gate onto the island:

$$\begin{aligned} p(q) & \left[\Gamma_{S \rightarrow 1}(\Delta E(q)) + \Gamma_{D \rightarrow 1}(\Delta E(q)) + \frac{1}{R_G C_{11}} \frac{\partial}{\partial q} (q - V_G C_{11} + V C_{1D}) \right] \\ & = p(q + e) [\Gamma_{1 \rightarrow S}(\Delta E(q + e)) + \Gamma_{1 \rightarrow D}(\Delta E(q + e))] \end{aligned} \quad (9.16)$$

Now $p(q)$ is the probability density of finding the total charge q on the island. The corresponding current–voltage characteristics are shown in Fig. 9.13.

In this device, the gate voltage keeps the island potential fixed at long time scales ($t \gg 1/R_G C_{11}$). If, however, V is sufficiently large and an electron can tunnel from S into the island, the gate response is too slow to prevent an additional voltage buildup at the drain junction, and the electron is able to tunnel to drain. If $|V_G| > e/C_{1D}$, a gate current starts to flow, and the island is open. Therefore, there is only one Coulomb diamond, centered around $(V, V_G) = (0, 0)$. The transconductance is no longer oscillatory in V_G , and the device is much less sensitive to fluctuating background charges. Fabricating such a transistor, however, hits some experimental difficulties that have yet to

be overcome: the heating problem is similar to that in a single tunnel junction, and the stray capacitance between gate and island should be negligible. In addition, this Coulomb diamond is much more sensitive to thermal smearing and noise than those in “conventional” SET transistors [179].

9.3

SET circuits with many islands: The single-electron pump

As an example of a more complex SET circuit, we study the system of two islands in series, also known as a *single-electron pump* (Fig. 9.14).

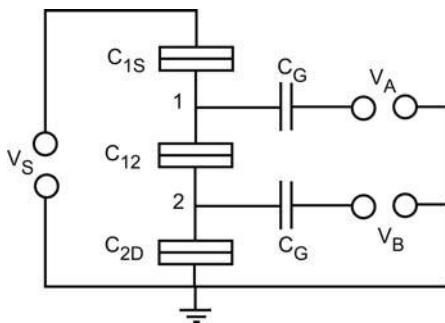


Fig. 9.14 Circuit of two islands in series. Each island can be tuned by a nearby gate electrode.

Via a tunnel junction, island 1 is coupled to source and island 2 to drain. The total capacitances C_{11} of both islands are assumed to be equal. Furthermore, we neglect several capacitance matrix elements (except those shown in Fig. 9.14) and assume that electrode A (B) couples only to island 1 (2), with equal capacitances. Nevertheless, V_B influences V_1 via the inter-island capacitance C_{12} and vice versa. We will not study the effect of a source–drain bias voltage. Rather, we are interested in the ground state of the system as a function of V_A and V_B . We assume that we can probe this state by applying a negligibly small source–drain voltage. Hence, we set $V_S = 0$. The island charge vector is given by $-e(n_1, n_2)$, and the electrode voltage vector by $(V_A, V_B, 0)$. The capacitance matrices of interest are

$$\underline{\underline{C}}_{II} = \begin{pmatrix} C_{11} & -C_{12} \\ -C_{12} & C_{22} \end{pmatrix}$$

$$\underline{\underline{C}}_{IE} = \begin{pmatrix} -C_G & 0 & -C_{1S} \\ 0 & -C_G & 0 \end{pmatrix}$$

with $C_{11} = C_{22} = C_{1S} + C_{12} + C_G = C_{2D} + C_{12} + C_G$. Six electron transfers are of importance.

1. An electron tunnels between source and 1, $\Delta\vec{q}_I = e(\pm 1, 0)$, $\Delta\vec{q}_E = e(0, 0, \mp 1)$. The onset of this transfer is determined by

$$\begin{aligned}\Delta E[\vec{V}_E, -e(n_1, n_2), \Delta\vec{q}] &= 0 \implies \\ C_{11}[C_G V_A - (n_1 \mp \frac{1}{2})e] &= -C_{12}[C_G V_B - n_2 e]\end{aligned}\quad (9.17)$$

2. An electron is transferred between drain and 2, $\Delta\vec{q}_I = e(0, \pm 1)$, $\Delta\vec{q}_E = (0, 0, 0)$, which gives

$$C_{11}[C_G V_B - (n_2 \mp \frac{1}{2})e] = -C_{12}[C_G V_A - n_1 e] \quad (9.18)$$

3. Finally, electrons can be exchanged between 1 and 2, $\Delta\vec{q}_I = e(\pm 1, \mp 1)$, $\Delta\vec{q}_E = (0, 0, 0)$, leading to

$$V_A - \frac{e}{C_G}(n_1 \mp \frac{1}{2}) = V_B - \frac{e}{C_G}(n_2 \pm \frac{1}{2}) \quad (9.19)$$

These boundaries define regions of stable electron configurations in the (V_A, V_B) plane, each of which is characterized by the island charge vector that corresponds to the lowest energy. For $C_{12} \rightarrow 0$, islands 1 and 2 are no longer coupled. It becomes impossible to influence island 1 by V_B and vice versa. In this limit, the stability diagram consists of squares given by conditions 1 and 2. Condition 3 plays no role, since the corresponding lines just touch two corners of the square (Fig. 9.15(a)).

Question 9.7: Investigate the stability diagram of the double island in the limit of connected islands.

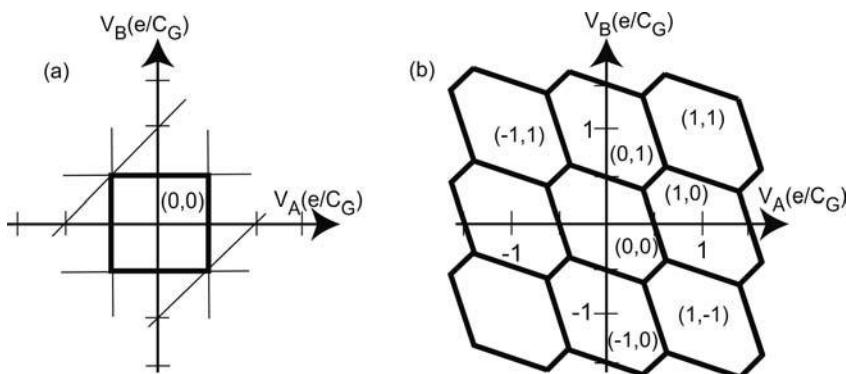


Fig. 9.15 Stability diagram of the two-island system of Fig. 9.14, (a) for completely decoupled islands and (b) for an inter-dot capacitance $C_{12} = C_G$.

The general situation is shown in Fig. 9.15(b): the boundaries (1) and (2) tilt for $C_{12} > 0$, and the stable regions develop a hexagonal shape. A current can pass from source to drain only if electrons can tunnel between the two islands as well as between island 1 (2) and source (drain). This degeneracy exists only at the corners of the elongated hexagons.

Question 9.8: Study the effect of cross capacitances on the stability diagram. Consider equal capacitances between gate A (B) and island 2 (1), which are much smaller than C_G .

The charge configuration of the double island system can be directly monitored by coupling a SET transistor to each island (Fig. 9.16). In this setup, the SET transistor labeled by 3 (4) serves as an electrometer to measure the charge on island 1 (2) [5]. In Fig. 9.16(a), the current through the double island is shown as a contour plot. As expected, current flows predominantly at the corners of the stable regions. Figs. 9.16(b) and (c) show the conductance of the electrometers 3 and 4, respectively, which is a measure of the charge on island 1 (2). The transition of the island charges is clearly visible as a sharp increase of the electrometer conductance along the direction that corresponds to changing the charge at the measured island. In Fig. 9.16(d), the difference signal of the two electrometers is shown, which emphasizes that, in each stable region, the charge configuration is really a different one.

In [245], it has been demonstrated for the first time that, with this device, electrons can be “pumped” by the gate voltages. The current can even be made to flow in the opposite direction of the source–drain bias voltage drop. In order to understand this experiment, we first consider the effect of a non-zero bias voltage: it shifts the boundaries of the stability diagram and generates triangular regions at the corners of the hexagons. Inside the triangles, Coulomb blockade becomes impossible. In order to operate the pump, the DC component of the gate voltages V_A and V_B are adjusted such that the device is located within one of these triangles (Fig. 9.17(a)).

Question 9.9: Calculate the shifts of the boundaries given in Fig. 9.17(a).

In addition, an AC voltage is applied to gates A and B, with a phase shift of (not necessarily exactly) $\pm\pi/2$. For sufficiently large AC amplitudes, the trajectory of the device state is a circle enclosing the triangle. Circling around the triangle labeled “P” in the positive direction corresponds to a sequence of states $(n_1, n_2) \rightarrow (n_1 + 1, n_2) \rightarrow (n_1, n_2 + 1) \rightarrow (n_1, n_2)$. This means that, for each round trip, one electron is transferred from source to drain, independent

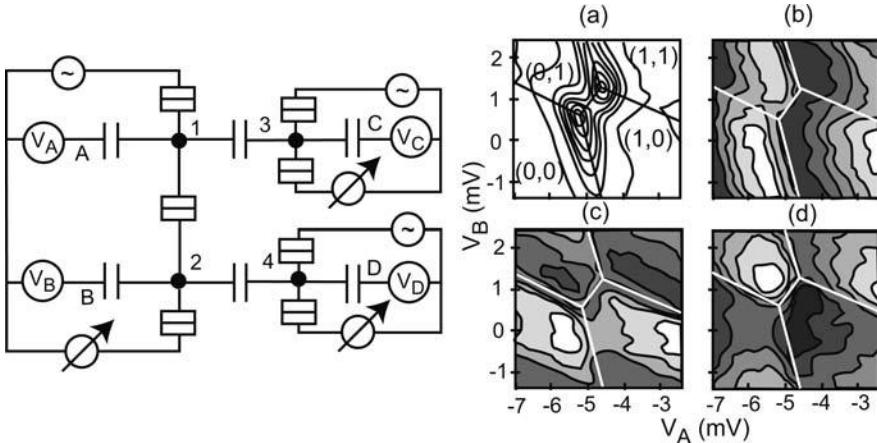


Fig. 9.16 Measurement of the stability diagram of the double island system. Left: Equivalent circuit of the double island system 1 and 2, with each island coupled to a SET transistor acting as an electrometer. Right: (a) conductance of the double island as a function of the gate voltages V_A and V_B in a contour diagram; (b, c) conductance of electrometer 3 (4), respectively; (d) difference signal of the two electrometers. (Adapted from [5].)

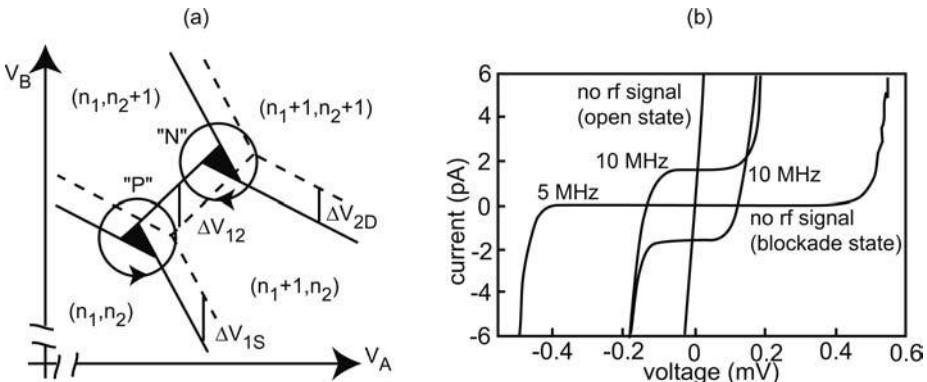


Fig. 9.17 (a) A non-zero bias voltage shifts the boundaries of the stability diagram in the V_B direction by

$$\Delta V_{1S} = -\frac{V_s}{C_G} \left(C + \frac{C^2}{C_{12}} - C_{12} \right)$$

$$\Delta V_{2D} = -\frac{V_s}{C_G} \frac{C_{12}^2}{C}$$

$$\Delta V_{12} = \frac{V_s}{C_G} C_{12}$$

respectively. As a result, triangular shaped regions are formed in which Coulomb blockade no longer exists. The circles denote the trajectories of the device as small AC voltages are applied to gates A and B. (b) Operation of the electron pump at different frequencies. The actual phase shift of the AC signal was $\pm 130^\circ$. Also shown are the I - V characteristics in the center and at a corner of a stable region, without an AC voltage applied. After [193].

of the direction and magnitude of the bias voltage. The current plateaus of the single-electron pump are shown in Fig. 9.17(b) as a “P” point is encircled with two different frequencies in positive (phase shift $\pi/2$) and negative (phase shift $-\pi/2$) directions. Note that the current is independent of the sign of V_S within a window around $V_S = 0$.

Also shown is the current–voltage characteristic when no AC signal is applied. Here, the current plateaus are absent. Provided the trajectory encloses the triangle completely and the AC amplitude is sufficiently small, such that other electron transfers are impossible, the current is coupled to the frequency via

$$I = ef$$

Furthermore, for the system to follow the frequency, f has to be smaller than the inverse time constant $1/\tau$ of the device, given by roughly $\tau = R_{12}C_{12}$. Encircling type “N” points in the same direction, or switching the direction in type “P” points, respectively, reverses the sign of the current.

Frequencies are the most accurate quantities we have in physics (the “NIST-F1 standard” is currently the frequency standard in the US and has an accuracy of 10^{-15}). This raises the question whether the single-electron pump can be used as a current standard, with the current coupled to a frequency (at present, currents can be defined with a relative accuracy of 10^{-6} [203]). Here, the low current that can be pumped through a single-electron pump constitutes a problem. We may, however, rephrase this question and ask: How accurate is the number of electrons pumped? It turns out that the accuracy is dominated by multi-junction tunneling events, so-called cotunneling. Even with Coulomb blockade established, an electron may tunnel onto the island virtually. If this electron, or a different one, tunnels off the island across the second barrier, a real current results. Cotunneling can be suppressed by increasing the number of tunnel junctions [14, 15, 203]. Fig. 9.18 shows an example where the cotunneling has been suppressed by placing high on-chip resistors in series with the SET device [193].⁶

Keller and coworkers [172] used an electron pump (see Fig. 9.19) that consists of six islands in series to charge a capacitor with an accuracy of 10^{-8} , i.e. the uncertainty is one electron for 10^8 pumped electrons. By measuring the voltage drop V across the capacitor after pumping N electrons, the capacitance $C = Ne/V$ could be determined with a standard deviation of 3×10^{-7} .

6) The results shown in Figs. 9.16 and 9.17 have actually been obtained with a thin-film Cr resistor located at the entrance and exit of the electron pump (see [193]).

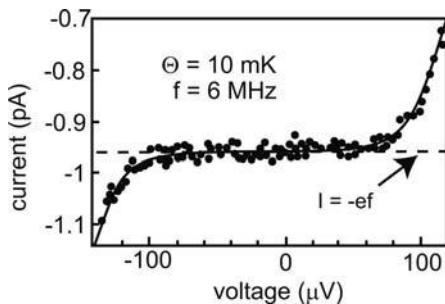


Fig. 9.18 Comparison between the observed current plateau of the single-electron pump (circles) and the current I expected from $I = ef$. Close to the center of the plateau, a relative error of 10^{-6} is found. Here, cotunneling has been suppressed by resistors in series with the single-electron pump. Adapted from [193].

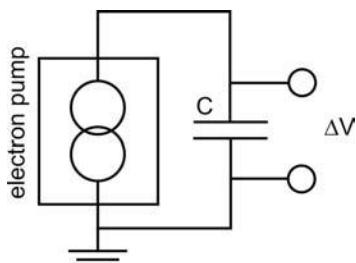


Fig. 9.19 Principle of the capacitance standard: the single-electron pump, consisting of several SET transistors in series, transfers a well-defined number of electrons onto the plate of a capacitor, and the voltage drop is measured.

Papers and Exercises

- P9.1** In [110], a single-electron transistor is used for detecting charge rearrangements in the substrate. How does this work?
- P9.2** Hanna and Tinkham [140] developed an analytical model for the Coulomb staircase in the limit of strongly differing junction couplings. Work out their model and reconstruct the authors' " $I(V)$ phase diagram" in Fig. 1b of that paper.
- P9.3** Geerligs et al. [113] demonstrated the operation of a *single-electron turnstile*, a slightly different concept for counting electrons than the single-electron pump. Explain the pumping mechanism of the single-electron turnstile.

P9.4 Superconductivity adds a new and exciting twist to single-electron tunneling. Work out the basic modifications due to superconductivity. A good starting point is Fitzgerald et al. [96].

E9.1 The “single-electron tunneling box” consists of an island in between a tunnel barrier and a capacitor with infinite resistance (see Fig. 9.20). The tunnel resistance is sufficiently high to suppress quantum fluctuations. Calculate the number of excess electrons on the island as a function of the voltage.

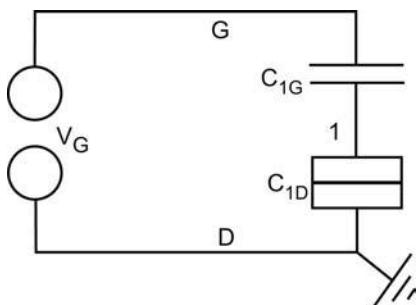


Fig. 9.20 Equivalent circuit of the SET box for Exercise E9.1.

E9.2 Calculate the current through a tunnel barrier in the absence of single-electron charging effects. Show that our definition of the resistance in Eq. (9.11) is reasonable for small voltages applied, since Ohm’s law is obtained.

E9.3 Modify the double island system of Fig. 9.14 such that both source and drain couple to island 1 only. Island 2 “dangles” (see Fig. 9.21). In the limit of zero source–drain bias voltage, what does the phase diagram in the (V_A, V_B) plane look like? Discuss the relevance of direct electron transfers between island 2 and the source/drain contacts. Assume identical capacitances.

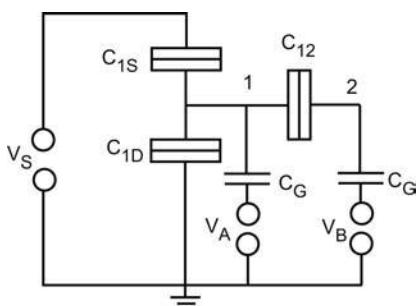


Fig. 9.21 Sketch of the double island system of Exercise E9.3.

Further Reading

A classic review article was written at the beginning of the “single-electron tunneling age” by Averin and Likharev [13]. A stimulating book containing collections of articles on various aspects of single-electron tunneling phenomena is [124]. Furthermore, [178] is an article entitled “Coulomb blockade and digital single electron devices”, which focuses on the relevant aspects of a future single-electron logic.

10 Quantum Dots

A conducting island of a size comparable to the Fermi wavelength in all spatial directions is called a *quantum dot*. The properties of quantum dots are very similar to those of atoms, and sometimes you hear that quantum dots are artificial atoms. The differences are essentially the size (0.1 nm for atoms vs. \approx 100 nm for quantum dots), and the shape and strength of the confining potential. In atoms, the electrons are bound by the attractive forces exerted by the nucleus; whereas in quantum dots, a mean electric field generated by background charges and gate voltages holds the electrons together. The number of electrons in an atom can be changed by ionizing it, which can be done by irradiating it with electromagnetic waves, or by applying a strong electric field. In quantum dots, the electron number is typically altered by tuning the confinement potential. An equivalent process in atoms would be to replace the nucleus by a neighbor in the periodic table.

The length scales imply characteristic energy scales in quantum dots that differ from those in atoms by roughly four orders of magnitude. The energy level spacing in atoms is of the order of 1 eV, while in quantum dots it is typically 0.1 meV. The ionization energy is in the range of 10 eV for atoms and about 1 meV in quantum dots. Therefore, quantum dots open up novel experimental possibilities not available in atoms. For example, it is easy to break Hund's rules in a quantum dot by applying a magnetic field. Doing this in an atom requires a magnetic field of the order of 10^4 T, two orders of magnitude larger than the strongest magnetic fields available in the laboratory. Such possibilities, combined with the high tunability of quantum dots, have boosted quantum dot research in the past 10 years. The option of tailoring their optical and electronic properties promises a variety of applications as well. Quantum dot lasers with particularly low threshold currents have been built, and it is envisaged to transfer many concepts of quantum optics related to the interaction of photons with atoms into a solid state environment, a goal that would pave the road for novel applications, such as quantum computation.

Here, we restrict ourselves to the basic transport properties of quantum dots. From a fundamental point of view, the possibility to probe a small entity of confined, interacting electrons by transport experiments is exciting. Clearly,

this aspect is essential for any kind of optoelectronic quantum dot devices as well.

We begin this chapter with a brief survey of the transport phenomenology of quantum dots. The elementary constant interaction model will be introduced in Section 10.2. It offers a crude and simple way to separate the interaction effects from size quantization, and allows the interpretation of a variety of observations in a straightforward manner. However, many experiments contradict this simple model, as demonstrated in Section 10.3. In Section 10.4, we will have a look at the line shapes of the conductance resonances, which offer additional information and complement the information on the discrete energies that is extracted from the resonance positions. Finally, the chapter is concluded with a look at further experimental realizations of quantum dots that do not rely on semiconductor heterostructures.

10.1

Phenomenology of quantum dots

The majority of transport experiments on quantum dots have been performed on samples made by the top-down approach, namely by lateral patterning of a semiconductor heterostructure. We pick this realization as an example to introduce the transport phenomenology of quantum dots. Other systems are mentioned in Section 10.5.

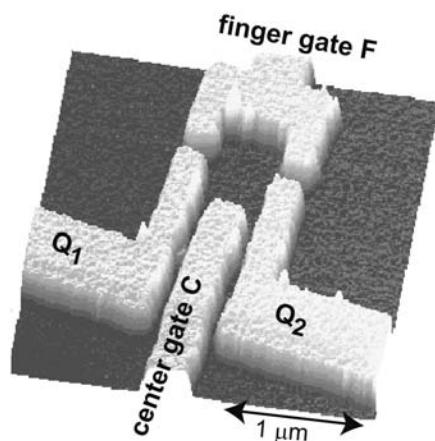


Fig. 10.1 Atomic force micrograph of a gate geometry used to generate a quantum dot in a Ga[Al]As heterostructure. The gold electrodes (bright) have a height of 100 nm. The two QPCs formed by the gate pairs F–Q₁ and F–Q₂ can be tuned into the tunneling regime, such that a quantum dot forms in between the two barriers. Its electrostatic potential can be varied by changing the voltage applied to the center gate.

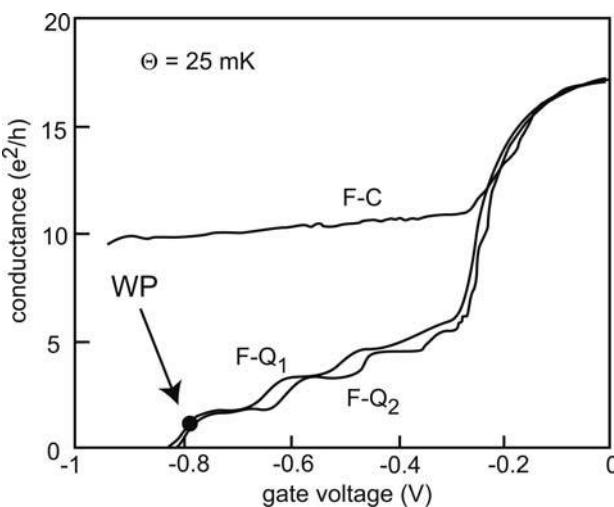


Fig. 10.2 Parametric conductances of different pairs of gates of the sample shown in Fig. 10.1. Sweeping the gate pairs F–Q₁ and F–Q₂ show typical QPC characteristics. Here, all remaining gates have been grounded. Gate pair F–C forms a somewhat poorly de-

fined QPC, which cannot be pinched off. Nevertheless, the formation of the channel due to depletion is clearly seen. “WP” denotes the working point of the gates F, Q₁, and Q₂ used to operate the dot as a single-electron transistor.

The gate structure of Fig. 10.1, defined on top of a Ga[Al]As heterostructure, is designed to impose and tune a quantum dot in the 2DEG underneath. In combination with the finger gate F, the gates Q₁ and Q₂ form two quantum point contacts (QPCs), which can be tuned independently (see Fig. 10.2). Suppose we now adjust the finger gate and the center gate such that the electron gas underneath is depleted. As the conductance of both QPCs is reduced below $2e^2/h$ by tuning the Q gates, the electron puddle in between the gates gets disconnected from the environment, and a closed quantum dot is formed, weakly coupled to source and drain via tunnel barriers. In this regime, conductance oscillations as a function of any of the gate voltages are observed (Fig. 10.3). The voltages applied to F, Q₁ and Q₂ strongly change the QPC conductances as well, which limits the tuning range of the dot. Therefore, the dot is usually tuned with a center gate voltage. This gate is designed to couple well to the dot, with little influence on the QPC transmission. A typical conductance trace as a function of center gate voltage was shown earlier in Fig. 1.5 in the Introduction.

Qualitatively, such oscillatory behavior is expected from both single-electron tunneling, as well as from the resonant tunneling discussed in Chapter 8. In fact, when this kind of oscillation was first observed, its explanation was not immediately clear. Only additional experimental studies revealed that, in fact, both of these effects play an important role. For typical experimental

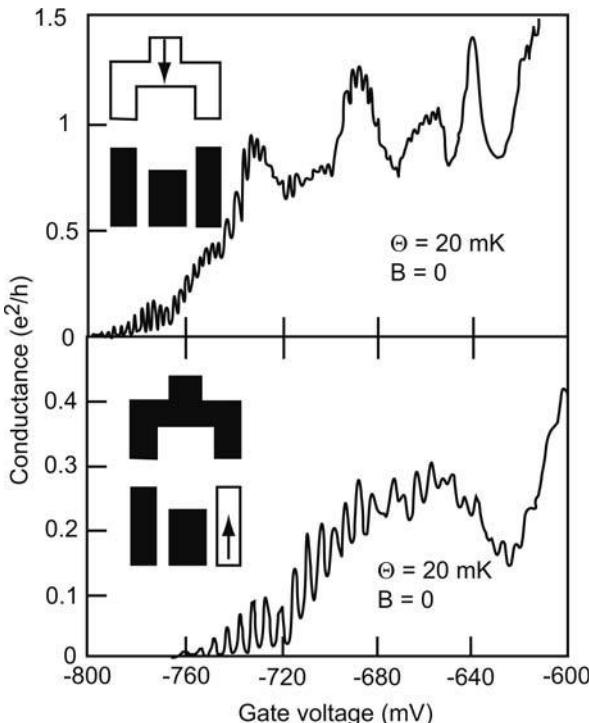


Fig. 10.3 Gate voltage characteristics with the dot defined (all gates are activated). Short-period transmission resonances are seen, which are modulated with a much larger period. Note that the resonances set in at a threshold gate voltage.

parameters, the Coulomb blockade determines the coarse features, while size quantization, i.e. the quantization of the kinetic energy inside the dot, is responsible for the fine structure [155, 210, 269]. A large amount of information on quantum dots has been collected by investigating their conductance as a function of gate voltage and a second, independent, parameter. In Fig. 10.4, the peak positions of the conductance resonances are plotted as a function of the gate voltage and a (perpendicular) magnetic field. While the raw data look rather structureless at first sight, a further investigation reveals a rich fine structure. First of all, the peak spacing in gate voltage is not constant – see the inset in Fig. 10.4(a). There is a general trend toward smaller peak spacings as the gate voltage is increased. This is partly due to a geometrical effect, since the edge of the dot approaches the gate electrode as we fill in electrons. As a consequence, the capacitance between the dot and the gate increases. On top of this effect, fluctuations in the peak spacings are apparent, as we saw already in Fig. 1.5 in the Introduction. The fine structure becomes more visible once a constant amount of the peak spacings has been subtracted (Fig. 10.4(b)). We

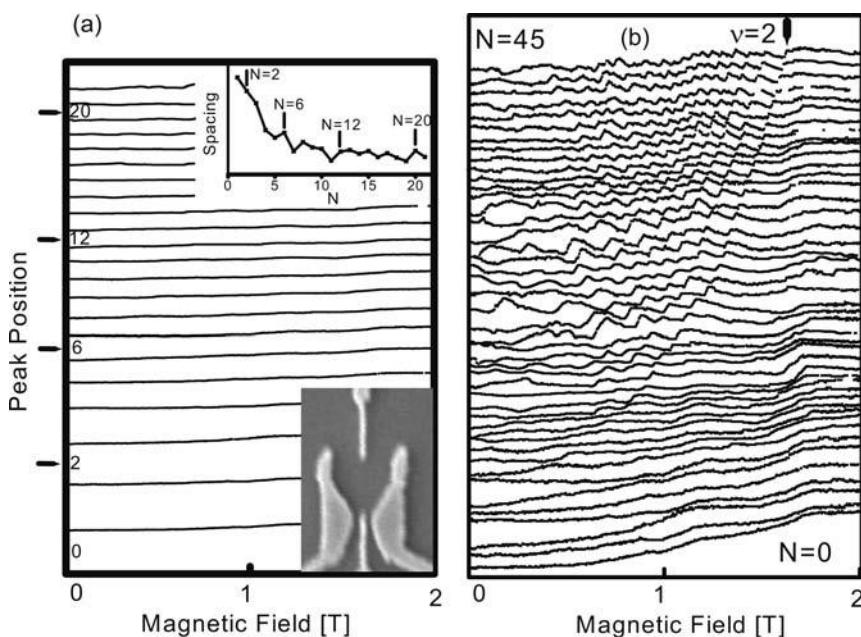


Fig. 10.4 (a) Positions of 22 consecutive conductance resonances as a function of the gate voltage and the magnetic field. The gate geometry of the sample is shown in the lower inset. The quantum dot has an approximately triangular shape with a width and height of about 450 nm. The center gate is in the center at the bottom. The upper inset shows the peak spacings at $B = 0$. The numbers in-

dicate dot occupations for which particularly large spacings are expected within the Fock–Darwin model (see text). (b) The data of (a) up to 45 electrons in the dot, with a constant peak spacing removed. $N = 0$ indicates the region where the dot is empty, while $v = 2$ denotes the Landau level filling factor inside the dot; see text. After [56].

divide this pattern into three regimes. At very low magnetic fields, the spacings fluctuate, with a certain tendency to bunch together for small occupation numbers. At intermediate magnetic fields ($B \approx 1$ T; and occupation numbers $N > 20$ in this example), the peak positions show quasi-periodic cusps. This pattern changes abruptly as B is increased, with a transition magnetic field that increases with the occupation number.

While the details of this overall pattern depend on the sample, and many additional effects have been found in particularly designed quantum dots and under appropriate experimental conditions, such a phenomenology is a common feature of most samples. In the following, we focus on the interpretation of this overall pattern. Before we begin, however, further experimental results need to be mentioned that contain additional information. Let us first look at the current through a quantum dot as a function of a gate voltage and of the source–drain bias voltage V (Fig. 10.5).

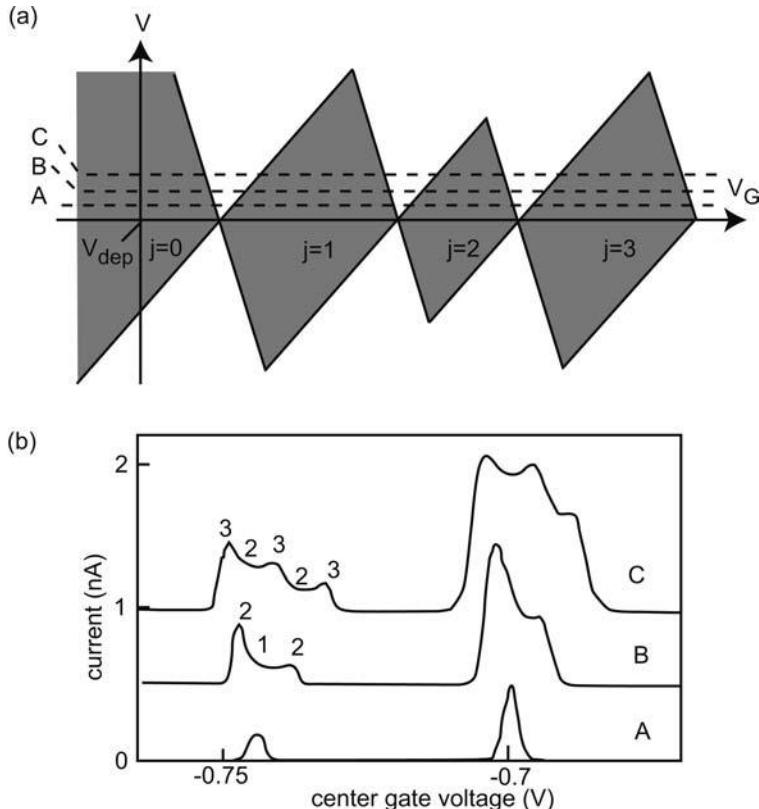


Fig. 10.5 (a) Sketch of the current through a quantum dot as a function of the gate voltage and the source–drain bias voltage V . Diamond-shaped regions of suppressed current exist. (b) As V is increased from A to C, the gate voltage sweeps reveal a fine structure of the conductance resonances. After [169].

Regions of suppressed current are observed, as sketched in Fig. 10.5(a). They resemble the Coulomb diamonds encountered already in Chapter 9. Here, however, their size fluctuates, while their shapes are essentially identical. At small source–drain voltages, the conductance resonances as a function of the gate voltage look similar to Coulomb blockade oscillations, although their amplitude fluctuates from resonance to resonance. As the bias voltage is increased, however, a fine structure emerges (Fig. 10.5(b)), which is absent in single-electron transistors. Finally, the amplitude of the resonances can be tuned by a magnetic field (see Fig. 10.6). In fact, it may change by orders of magnitude and can be suppressed below the noise level.

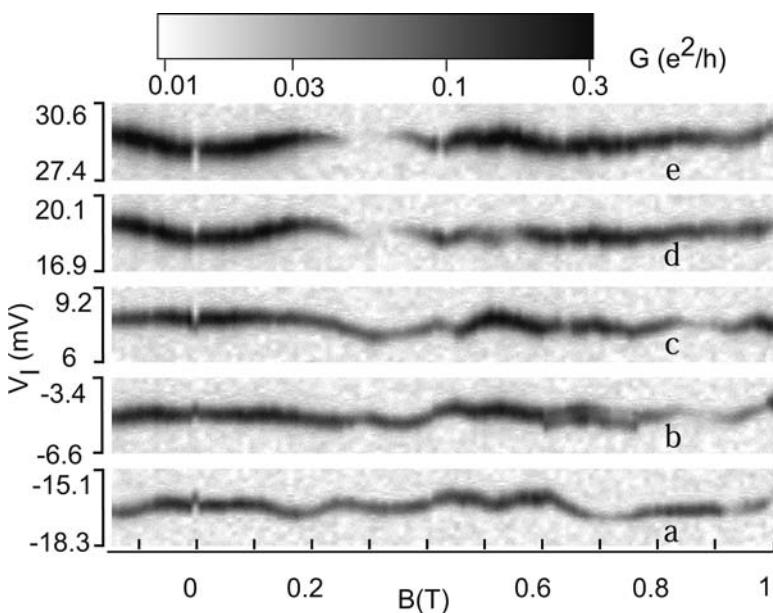


Fig. 10.6 Evolution of five consecutive conductance resonances in a magnetic field. The peak positions fluctuate by about 20% of their spacing, while the amplitude varies by up to 100%. After [197].

10.2 The constant interaction model

How shall we interpret these observations? Clearly, a quantum dot is a quasi-zero-dimensional system. Within a single-particle picture, its density of states consists of a sequence of peaks, with positions determined by the size and shape of the confining potential, as well as by the effective mass of the host material. A very simple estimation for the average nearest-neighbor spacing Δ of these energy levels is obtained by starting from the two-dimensional density of states, $D_2(E) = m^*/\pi\hbar^2$. For a spin degeneracy of 2, there are $m^*A/\pi\hbar^2$ states per energy interval of unit length in an area A , and thus a spin-resolved, average energy level spacing of

$$\Delta \approx \pi\hbar^2/m^*A \quad (10.1)$$

is expected. The second energy scale of relevance is set by the single-electron charging energy. For a sufficiently weak coupling to the leads, i.e. when the conductances of the barriers that connect the dot to source and drain are below $2e^2/h$, the electrons of the dot are strongly localized, and Coulomb blockade comes into play.

Question 10.1: Show that Eq. (10.1) is obtained from the energy spectrum of a two-dimensional square well in the limit of large quantum numbers; estimate Δ and the single-electron charging energy for a GaAs quantum dot with a diameter of 300 nm.

Apparently, in order to add an electron to the dot, the *addition energy* is required, which is the sum of the electrostatic and the kinetic parts of the energy, as discussed already during our discussion of capacitance spectroscopy in Section 6.1. One estimate for the addition energy of the j th electron would be simply to add the single-electron charging energy and the single-particle separation of the j th energy level from its occupied neighbor. This approach seems perfectly satisfactory at first sight, but in fact makes the crucial assumption that the kinetic energy of the dot states is independent of the number of electrons in the dot. Owing to the electron–electron interactions, screening, as well as exchange and correlation effects, this is not strictly the case. It is well known that interactions strongly modify the energy spectra of atoms. Even in the case of the helium atom with just two electrons, the addition spectrum is tremendously complicated. The approach outlined above, known as the *constant interaction* (CI) model, disregards such difficulties. The CI model is a valuable tool for analyzing quantum dot addition spectra, and it provides a good explanation of the data in several cases.

It is straightforward to include the additional discrete energy levels in our single-electron tunneling model of Chapter 9. Suppose state j with single-particle energy ϵ_j is the highest occupied state in the quantum dot. An additional electron will occupy the empty state with the lowest energy, ϵ_{j+1} . This energy can simply be added to the energy difference ΔE in Eq. (9.4). Likewise, for processes that reduce the electron number in the dot, we subtract ϵ_j , the energy of the highest occupied level, in that equation. In the previous chapter, we took it for granted that the number of electrons in the island is large compared to the number of additional charges we forced on the island with the gate voltage. This is no longer necessarily the case in quantum dots, since the electron densities are smaller by a factor of ≈ 1000 . Typical occupation numbers range between 0 and 100. It is thus natural to define the number j of electrons as zero for the empty dot. In principle, the dot can be filled with holes, but this would require the dot potential to be tuned across the bandgap of the semiconductor host. We do not consider this possibility.¹

For simplicity, let us assume that there is no background charge. The boundaries of the stable regions that form the diamonds (Eqs. (9.13)) are modified and now read as follows.

For $\delta\vec{q} = e(-1, 1)$:

1) Note that, although we speak of quantum dots in electron gases, everything is analogous for hole gases.

$$V(V_G, j) > \frac{C_{1G}}{C_{11} - C_{1S}}(V_G - V_{\text{dep}}) - \frac{e(j + \frac{1}{2})}{C_{11} - C_{1S}} - \epsilon_{j+1} \frac{C_{11}}{e(C_{11} - C_{1S})}$$

For $\delta\vec{q} = e(1, -1)$:

$$V(V_G, j) < \frac{C_{1G}}{C_{11} - C_{1S}}(V_G - V_{\text{dep}}) - \frac{e(j - \frac{1}{2})}{C_{11} - C_{1S}} - \epsilon_j \frac{C_{11}}{e(C_{11} - C_{1S})}$$

For $\delta\vec{q} = e(-1, 0)$:

$$V(V_G, j) < -\frac{C_{1G}}{C_{1S}}(V_G - V_{\text{dep}}) + \frac{e(j + \frac{1}{2})}{C_{1S}} + \epsilon_{j+1} \frac{C_{11}}{eC_{1S}} \quad (10.2)$$

For $\delta\vec{q} = e(1, 0)$:

$$V(V_G, j) > -\frac{C_{1G}}{C_{1S}}(V_G - V_{\text{dep}}) + \frac{e(j - \frac{1}{2})}{C_{1S}} + \epsilon_j \frac{C_{11}}{eC_{1S}}$$

Here, V_{dep} denotes the depletion voltage as indicated in Fig. 10.5, while V_G is the gate voltage used to tune the dot.

For the special case of $j = 0$, of course, we cannot remove a further electron from the island, and thus the second and fourth inequalities in (10.2) do not apply. The stability diagram thus consists of a semi-infinite set of diamonds of equal shape. Their sizes vary due to the varying single-particle level spacings. In addition, a “semi-diamond” is obtained for $j = 0$ (see Fig. 10.5). In analogy to the diamonds in the pure electrostatic case, their maximum extension in the V -direction equals

$$\Delta V = \frac{1}{e} \left(\frac{e^2}{C_{11}} + \epsilon_{j+1} - \epsilon_j \right)$$

The peak spacing in gate voltage at $V \approx 0$ is given by

$$\Delta V_G = \alpha \left(\frac{e^2}{C_{11}} + \epsilon_{j+1} - \epsilon_j \right) \quad (10.3)$$

The ratio $\alpha = C_{11}/eC_G$ is a lever arm that translates the addition energies into gate voltages, while the dot’s total energy changes by $(e^2/C_{11}) + \epsilon_{j+1} - \epsilon_j$ as one electron is added.

Question 10.2: The quantum dots of Figs. 10.1 and 10.4 have more than one gate. How does this enter in Eq. (10.3)?

Within the CI model, we can subtract E_C from the measured peak spacings according to Eq. (10.3), as has been done to get Fig. 10.4(b) from Fig. 10.4(a). The remainder should correspond to the single-particle energy spectrum of

the dot. In order to appreciate the effect of a magnetic field, we consider a model that has the advantage of being analytically solvable, namely the Fock–Darwin model [64, 99] encountered already in Chapter 7. Now, the electron motion is no longer free in the third direction, and the parabolic potential is orientated in the (x, y) plane. The energy spectrum in this plane is not changed by these modifications.

Hence, we assume that the quantum dot is circular in shape and has a parabolic confinement potential,

$$V(x, y) = \frac{1}{2}m^*\omega_0^2(x^2 + y^2) = \frac{1}{2}m^*\omega_0^2r^2$$

with r being the radius of the dot. The corresponding Schrödinger equation can be solved analytically, even with a magnetic field applied in the z -direction [165]. The energy spectrum is given by

$$E_{n,l}(B) = (2n + |l| + 1)\hbar\sqrt{\omega_0^2 + \frac{1}{4}\omega_c^2 - \frac{1}{2}l\hbar\omega_c} \pm g^*\mu_B B \quad (10.4)$$

The radial quantum number is $n = 0, 1, 2, \dots$, while l is the angular momentum quantum number, i.e. $l = 0, \pm 1, \pm 2, \dots$. At $B = 0$, the energy levels are located at $(j+1)\hbar\omega_0$, with $j = 2n + |l|$, and with an orbital degeneracy of j . In addition, there is a twofold spin degeneracy. In analogy to atomic energy spectra, we can speak of the j th Fock–Darwin shell. The orbital degeneracies in each shell get removed by a perpendicular magnetic field, since all states within one shell have different angular momenta and thus respond differently to the magnetic field. For now, let us assume that the effective g -factor of the dot g^* is negligible, such that the levels remain spin-degenerate. This spin-degenerate Fock–Darwin spectrum is shown for $\omega_c \leq \omega_0$ to the left in Fig. 10.7. We see that a sufficiently strong magnetic field induces level crossings. For the confining strength shown in this figure ($\hbar\omega_0 = 1$ meV), for example, a crossing of level $(n, l) = (0, 2)$ with $(n, l) = (0, -1)$ occurs at $B \approx 0.4$ T, and the ground state configuration of the dot changes. Similar level crossings occur more frequently at higher energies.

Sometimes, the behavior expected within the Fock–Darwin model agrees even quantitatively with the experimental observations; see Paper P10.1. In the typical experiment shown in Fig. 10.4, of course, one cannot expect perfect agreement, since the dot’s confining potential is neither circular nor strictly parabolic.² The Fock–Darwin model predicts filled shells for $N = 2, 6, 12, 20, \dots$. Although these spacings are slightly enhanced in Fig. 10.4(b), there is certainly no 1 : 1 correspondence. Note further that there is also no spin pairing visible. The CI model is thus a reasonable first ap-

2) Numerical simulations have revealed that, in many samples, a circular dot shape and a parabolic confinement are actually better approximations than might be expected from the gate geometry [185].

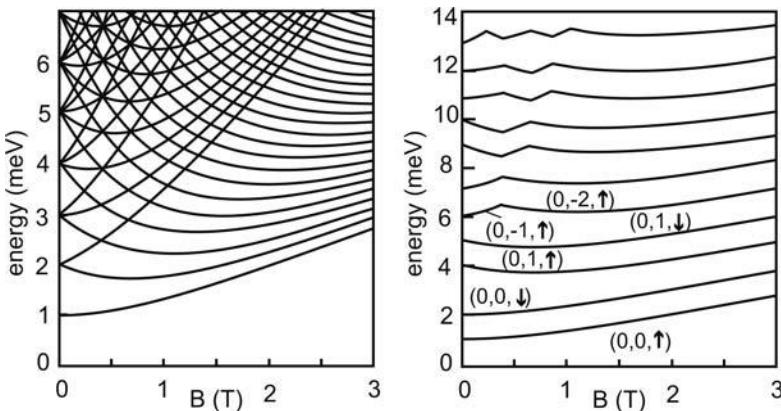


Fig. 10.7 A section of the Fock–Darwin spectrum (left), calculated for $\hbar\omega_0 = 1$ meV, and the predicted evolution of the conductance resonances as a function of the gate voltage and the magnetic field (right), when the single-electron charging energy equals 1 meV as well. The labeling is explained in the text.

proximation for this regime, although it cannot explain all the details of the experimental observations.

10.2.1

Quantum dots in intermediate magnetic fields

We proceed by looking at the quasi-periodic cusps of the peak positions that can be seen in Fig. 10.4(b). To begin with, it is useful to transform the Fock–Darwin model to a different set of quantum numbers, which emphasizes the behavior of the energy levels in magnetic fields. Intuitively, we expect that the stronger B is, the less important the electrostatic confinement should be, and the Fock–Darwin levels should bunch together and form Landau levels. It is therefore appropriate to relabel the energy levels by the Landau level quantum number $m = 1, 2, 3, \dots$, and a quantum number p that enumerates the energy levels within a Landau level. With the transformation [165]

$$p = n + \frac{1}{2}(|l| + l)$$

and

$$m = n + \frac{1}{2}(|l| - l) + 1$$

the energy levels of the Fock–Darwin spectrum read

$$E_{m,p} = \hbar(m + p)\sqrt{\omega_0^2 + \frac{1}{4}\omega_c^2} + \frac{1}{2}\hbar(m - p - 1)\omega_c \quad (10.5)$$

Here, we again neglect spin splitting. In the regime of filling factors $2 < \nu < 4$, this spectrum develops a very simple structure (see Fig. 10.8).

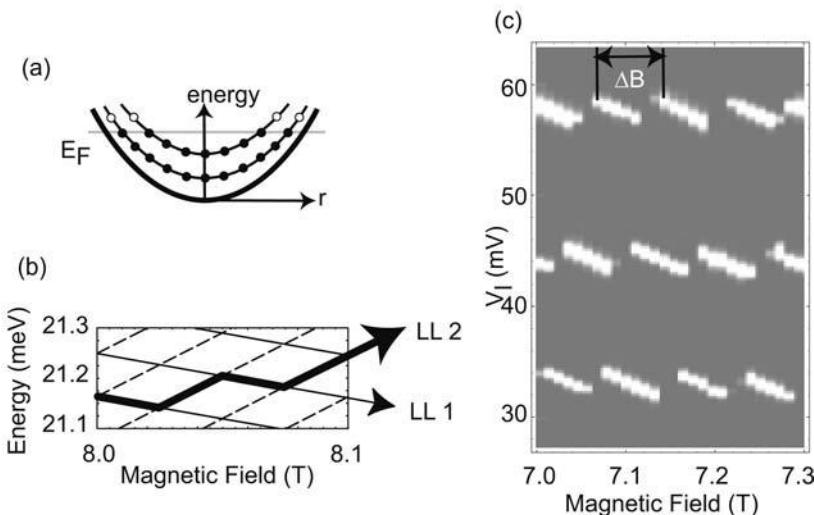


Fig. 10.8 (a) Two Landau levels in a parabolic quantum dot. Below the Fermi energy, the discrete states of each Landau level are occupied, as indicated by the full circles. Higher Landau levels are not shown, since all their states are empty. The filling factor is thus in the regime $2 < \nu < 4$. (b) A section of the corresponding energy spectrum, calculated for a circular dot with a parabolic confinement of $\hbar\omega_0 = 1$ meV. States be-

longing to LL 1 (thin full lines) reduce their energy as B is increased, while those states belonging to LL 2 (dashed lines) are running upwards in energy. The bold lines represent the Fermi level as a function of the magnetic field B , for a constant number of electrons in the dot. (c) The conductance of a quantum dot as a function of the gate voltage and the magnetic field in this regime. Bright areas correspond to a high conductance.

States with $m = 1$ decrease in energy as B is increased, while Landau level 2 (LL 2) states show a positive magneto-dispersion. This fact simply reflects the increasing degeneracy of LL 1 and the depopulation of LL 2 as B is increased. The energies as a function of B develop a quasi-periodic pattern of diamonds (see Fig. 10.8(b)). At constant electron number, the electrochemical potential of the dot, and with it the energy of the transmission resonance, moves in zigzag lines as B is tuned. The period can be approximated by $\Delta B \approx (\omega_0/\omega_c)^2 B$. Furthermore, the energy levels are approximately equally spaced in energy, with a spacing of

$$\Delta E = E_{m,p+1} - E_{m,p} \approx \hbar \frac{\omega_0^2}{\omega_c}$$

which is independent of the energy level quantum numbers. These approximate expressions are part of Exercise E10.1. The spatial location of these states is shown schematically in Fig. 10.8(a). At the Fermi level, the $m = 1$ states are close to the edge of the dot, while states with $m = 2$ are located toward the dot center. This means that states belonging to LL 1 couple much better to the leads than do LL 2 states.

Although we have developed this scenario within the Fock–Darwin model, most conclusions remain valid for other dot shapes and confining potentials. Independently of these details, the LL2 states couple more weakly to the leads, since they are further inside the dot, and thus the tunnel barrier they form with the reservoirs is larger. These states will get depopulated as B is increased, independently of the confinement shape. In fact, the measurement shown in Fig. 10.8(c) agrees reasonably well with the Fock–Darwin model. Bright lines of high conductance with a negative slope are observed. They measure the magnetic field dispersion of LL1 states. The LL2 states couple very weakly to the leads, owing to the exponentially suppressed tunneling. The conductance via those states is not detectable at the low source–drain bias voltages used in the experiment. For a further interpretation of these data, see Exercise E10.1.

Experimentally, different confinement potentials can be established by, for example, varying the parameters of the sample and of the fabrication process accordingly. The corresponding data in a quantum dot with an approximate hard-wall confinement are discussed in Fig. 10.9. The energy spectrum of a circular disk with hard walls cannot be solved analytically. Rather, the spectrum is obtained by numerical calculation of the zeros of the hypergeometric function usually denoted as ${}_1F_1$ in the literature of special functions [112]. This energy spectrum is shown schematically in Figs. 10.9(a) and (b): most strikingly, the density of states at the Fermi level in LL2 is higher than in LL1, provided the Fermi level is not far above the energy of LL2 at the center of the dot. Fig. 10.9(c) shows a corresponding measurement [106].

The reconstruction of the energy spectrum of this dot reveals that the spacing between LL1 states is significantly larger as compared to that for LL2 states, which indicates a steep-wall confinement.

10.2.2

Quantum rings

As a third example, we have a look at the reconstructed energy spectrum of a quantum ring in moderate magnetic fields (Fig. 10.10) [107]. From the single-particle spectrum of a one-dimensional quantum ring (the topic of Exercise E8.3), we expect a pattern formed by a set of parabolas

$$E_{l,n} = \frac{\hbar^2}{2m^*r^2}(l+n)^2 \quad (10.6)$$

Here, $n = eBr^2/2\hbar$ is the number of magnetic flux quanta that penetrate the ring, and $l = 0, \pm 1, \pm 2, \dots$ is the angular momentum quantum number. This is partly observed in the data of Fig. 10.10. The periodicity and the amplitude of the well pronounced zigzag lines are in agreement with the expectations from the single-particle spectrum. In addition, quasi-dispersionless states are

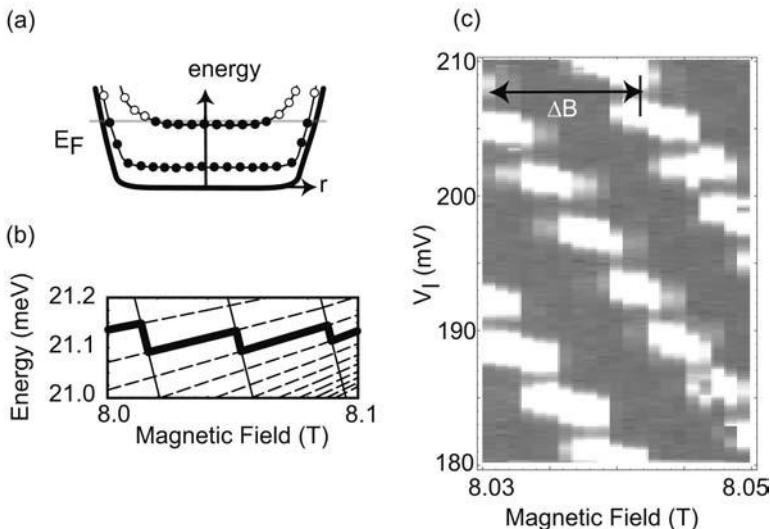


Fig. 10.9 (a) Landau levels in a quantum dot with an approximate hard-wall confinement. (b) A section of the calculated energy level diagram for $2 \leq \nu \leq 4$. (c) A corresponding set of experimental data.

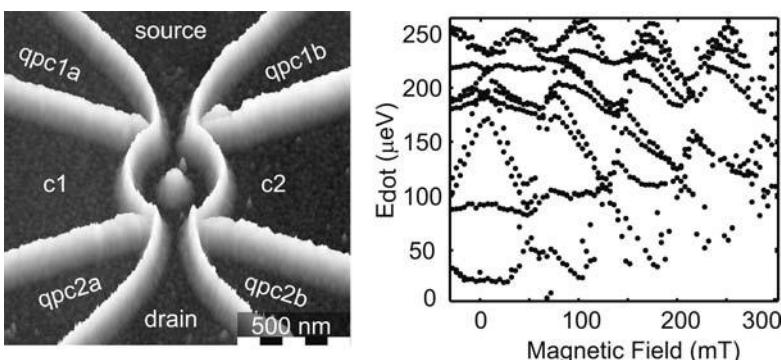


Fig. 10.10 Left: Sample geometry of a quantum ring, defined in the 2DEG of a Ga[Al]As HEMT by local oxidation of the surface. As usual, the QPC gates tune the dots coupling to source and drain, while the ring can be tuned with the center gates. The ring contains about 100 electrons. Right: The reconstruction of the energy spectrum.

observed, which most likely reflect the imperfections of the ring. Owing to azimuthal thickness variations, the actual states may be an admixture of various eigenstates with different l , which damp the amplitude of the zigzag lines. Although strong deviations from the single-particle spectrum of a perfect one-dimensional ring are observed, the CI model thus seems to be a good approximation.

10.3

Beyond the constant interaction model

The CI model is bound to fail if residual interactions come into play, like exchange and correlation effects, or like screening that depends on the number of electrons in the dot. Actually, the absence of spin pairing in Fig. 1.5 is a good example for exchange and correlation energies which can be dealt with only numerically. We now look at some effects in quantum dots that remain unexplained within the CI model, but can be interpreted within simple models for residual interactions.

10.3.1

Hund's rules in quantum dots

Hund's rules tell us in what sequence the states within an atomic shell are filled with electrons. Hund's first rule states that the total spin gets maximized, without violating the Pauli principle. This rule originates in the exchange interaction, due to which electrons with parallel spin are kept spatially separated, which reduces their mutual Coulomb energy. Hund's second rule forces the electrons to maximize the total orbital angular momentum, under the constraint of Hund's first rule. How the first six electrons are filled in a Fock–Darwin potential at $B = 0$ according to Hund's rules is shown in Fig. 10.11. The third electron filled in this potential occupies the level $(n, l) = (0, 1)$.³ The fourth electron occupies the $(0, -1)$ level, with the same spin direction as the third electron. In analogy to the nomenclature used in atomic physics, we denote the electronic configurations by $^{2S+1}L_J$, where S is the total spin, J the total angular momentum, and L the total orbital momentum, which is usually denoted by S for $L = 0$, P for $L = 1$, and so on. Forcing the fourth electron in the $(0, 1)$ level as well requires the exchange energy Δ_{xc} to be paid, and the configuration 3D_3 results. The fifth electron goes again in level $(0, 1)$, with its spin in the opposite direction, while the sixth one fills the last empty state in this shell. How Hund's rules can be broken in quantum dots by applying magnetic fields is the topic of Paper P10.1.

10.3.2

Quantum dots in strong magnetic fields

An obvious failure of the CI model occurs in quantum dots under strong magnetic fields, i.e. for filling factors $\nu < 2$. In Fig. 10.4, this regime is located above the upper magnetic field threshold for the cusps. The experimental findings are summarized in Fig. 10.12. In the previous section, we have seen

3) The spin orientation is determined by Hund's third rule, which states how the total spin and the total angular momentum couple. This depends on the host material.

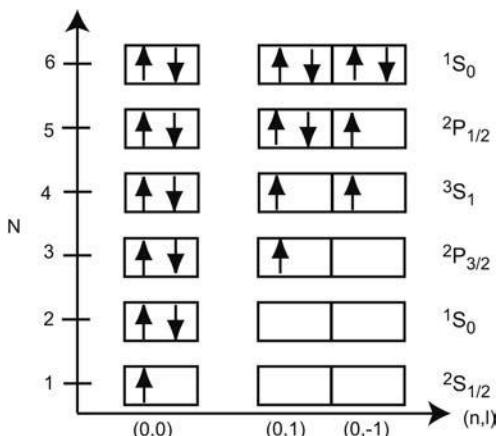


Fig. 10.11 Schematic occupation of the Fock–Darwin energy levels (n, l) as the dot is filled with N electrons, according to Hund's rules.

that frequent and quasi-periodic level crossings are observed in quantum dots for filling factors just above 2, which is in agreement with a single-particle spectrum. For filling factors below 2, there are no orbital level crossings. Only the spin splitting causes slightly different magneto-dispersions for spin-down states as compared to spin-up states. Very infrequent level crossings are therefore expected within a single-particle picture and for reasonable effective g -factors. However, in [207], frequent level crossings have been observed in this regime, which remain unexplained within the CI model.

To understand this effect, we revisit the Chklovskii picture of edge channels mentioned in Section 7.3, and adapt it to a quantum dot. Imagine that the edge channel configuration of Fig. 7.20 is bent to form a circle. Qualitatively, we see right away that, owing to the spin splitting and the modulated screening properties at the edge and the resulting electronic structure, a dot with only the spin-up and spin-down sublevels of one Landau level occupied segregates into a metallic ring around its edge and a metallic disk at its center, separated by an insulating stripe.⁴ The resulting structure is sketched in Fig. 10.13. The location and width of the insulating stripe are determined by the effective g -factor inside the dot. From an electrostatic point of view, an additional capacitance is formed between the ring and the disk. The system can be thought of as a variation of a double island system discussed in Exercise E9.3: the ring corresponds to the first island coupled to the leads. The second island, i.e. the central disk, couples only indirectly to the leads, via the ring.

4) This picture is, as in the case of straight edge channels, only physically meaningful if the insulating stripe is wider than the magnetic length.

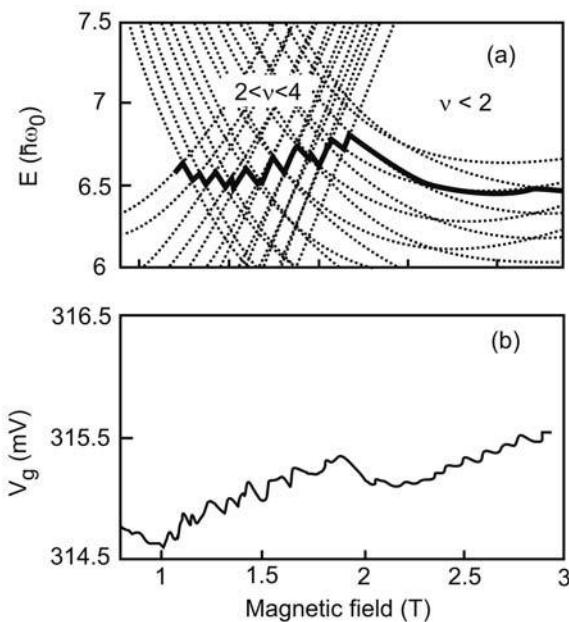


Fig. 10.12 (a) The dotted lines represent the magnetic field dispersion of the levels 30–50 of the Fock–Darwin spectrum, including spin splitting. The bold line follows the energy of the 39th state. Below filling factor 2, levels cross only because the two spin directions have different magneto-dispersions. Hence,

very rare level crossings are expected in this regime. (b) Experimentally, however, rapid oscillations of the conductance peak positions are found, as exemplified by a resonance observed for about 39 electrons in the dot. After [207].

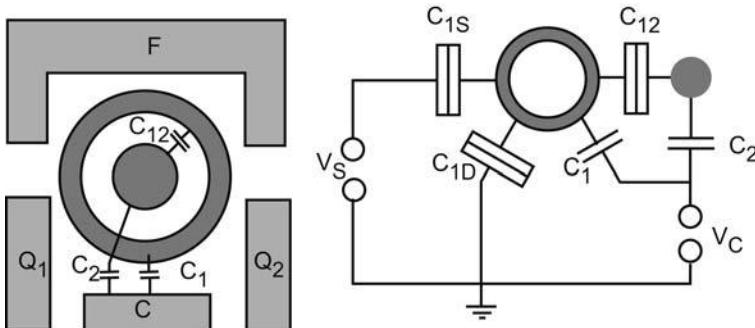


Fig. 10.13 Electrostatic structure of a quantum dot in strong magnetic fields (left). The dark regions denote metallic areas of the dot, while the white stripe in between is insulating. The metallic ring close to the dot's edge is formed by the edge channel of the spin-up sublevel of Landau level 1, while the disk at

the dot's center is formed by the spin-down sublevel. An intra-dot capacitance emerges, and the capacitance of the dot with respect to the gates is split up into two components stemming from the disk and the ring. The system is equivalent to the double island circuit shown to the right.

The analogy is not complete, though. First of all, the intra-dot capacitance C_{12} is a function of the magnetic field and of the electron density. Second, tuning both islands independently with two gate voltages is impossible. Rather, the gate voltage couples differently to the two islands, with a ratio that changes with the gate voltage. As the magnetic field is increased, the spin-down sublevel gets depopulated, and the electrons get transferred to the spin-up sublevel. Such a transfer, however, requires the intra-dot charging energy to be overcome, since the electron cannot be transferred to the spin-up sublevel within the disk, as all the states of this sublevel lie well below the Fermi level and are thus occupied.⁵ The period of the zigzag lines of the transmission peaks as a function of B are now a measure for the intra-dot capacitance. Quantitative calculations within such a model agree well with the experimental data (see Fig. 10.14).

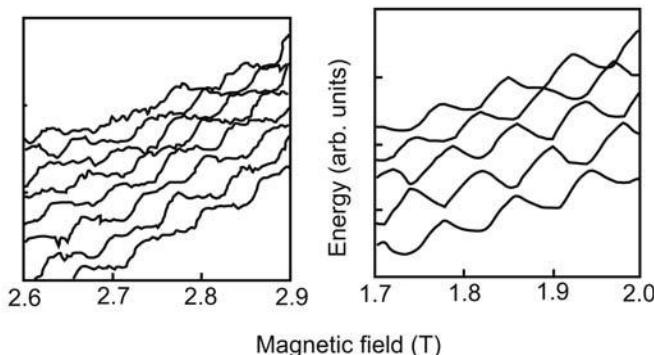


Fig. 10.14 The measured (left) and calculated (right) evolution of consecutive conductance peak positions for quantum dot filling factors below 2 agree reasonably well, in particular with respect to the periodicity, if this electrostatic dot structure is taken into account. After [207].

10.3.3

The distribution of nearest-neighbor spacings

Above an occupation number of about 20 and for small magnetic fields, analytical single-particle energy spectra bear no resemblance to the observed energy level spacings of quantum dots. This experimental fact is usually explained in terms of *quantized chaos*. Owing to the tremendous relevance of the underlying theory for all kinds of mesoscopic phenomena, we introduce the concept using some properties of quantum dots as an example.

A classical system is chaotic if its evolution in time depends exponentially on changes of the initial conditions. To make the connection, consider a point-like mass confined in a two-dimensional box. The point mass moves in a con-

5) The electrostatics of this system is the topic of Paper P10.2.

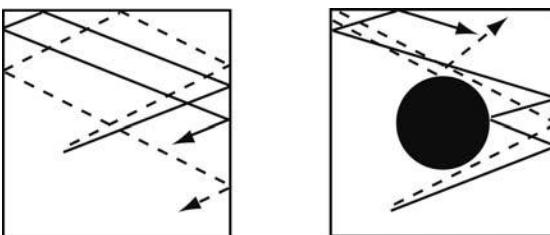


Fig. 10.15 In a classically chaotic geometry like the Sinai billiard (right), trajectories diverge exponentially as the initial conditions change by arbitrary small amounts. This is not the case for regular structures, like the square to the left.

stant potential, i.e. along straight lines, and experiences specular reflection at the walls. The trajectory of the point mass can be parameterized in a suitable way, and its position at time t can be written as $p_1(t)$. Now, suppose we start the motion of the point mass with a slightly changed initial condition $p_2(0) = +p_1(0)\delta p$, and we ask how $\Delta p(t) = p_2(t) - p_1(t)$ evolves over time. It turns out that there are two fundamentally different kinds of evolutions, which depend on the shape of the box. If, for long time scales ("long" meaning that the point mass has hit the wall many times), $\Delta p(t)$ diverges exponentially, the box is called *chaotic*. If the divergence is non-exponential, the box is *regular*. It turns out that only very few structures are regular, e.g. a square box, or a circle. Most shapes show a chaotic classical dynamics. A famous example widely discussed in the literature is the so-called Sinai billiard, which consists of a square box with a circular pillar at its center (see Fig. 10.15). Quantizing a classically chaotic system can be done by using the Gutzwiller trace formula [132]. The properties typical for classically chaotic systems, like the exponential sensitivity on initial conditions, get lost during the quantization process, but there are nevertheless remnants of classical chaos in the quantum regime.⁶

The two most widely investigated remnants of chaos in quantum dots are the probability distribution of nearest-neighbor peak separations (the NNS distribution), which we will discuss below, and the distribution of the transmission resonance amplitudes, which is the topic of Paper P10.3.

Experimentally, the NNS distribution $p(s)$ is obtained in a straightforward way: determine the spacings s_j of adjacent energy levels E_j and E_{j+1} , plot them in a histogram, and normalize the distribution properly. For randomly placed energy levels, a Poisson distribution is obtained, i.e. $p(s) = \exp(-s)$. For some quantum mechanical systems, we can say right away what the NNS distribution looks like. For the Fock–Darwin potential at $B = 0$, for example,

6) The subfield of chaos theory that investigates these remnants is referred to as *quantized chaos*. For further information, see [133] and [134].

it consists of two δ functions, one at $s = 0$, which contains the degenerate levels including spin degeneracies, and one at $s = \hbar\omega_0$. Each regular system has its characteristic $p(s)$, i.e. $p(s)$ is non-universal. This is not the case for classically chaotic systems. Naively, one is probably tempted to assume that, in a chaotic system, the positions of the energy levels are completely random, and thus $p(s)$ would be a Poisson distribution. This, however, is not the case. Rather, $p(s)$ takes one of three universal forms, which depends solely on the symmetry properties of the Hamiltonian. These distributions can be calculated within the so-called *random matrix theory* (RMT), which has turned out to be highly successful in many branches of physics, including several aspects of mesoscopic transport [24].

We now take a look at the concept of RMT by sketching the derivation of $p(s)$. Suppose we represent the Hamiltonian of our quantum dot in some basis, such that it can be written as a Hamiltonian matrix H . For a Hamiltonian that is invariant under time inversion (i.e. no magnetic field should be present), the matrix is Hermitian, and $p(H)$ should be invariant under orthogonal basis transformations. If time reversal symmetry is broken, the Hamiltonian matrix is unitary, and $p(H)$ should not change under unitary transformations. These conditions clearly set some constraints on the matrix elements. It is assumed that, within these constraints, the matrix elements are completely random for classically chaotic systems. The above conditions define the orthogonal and the unitary ensemble of random matrices, called *Gaussian orthogonal ensemble* (GOE) and *Gaussian unitary ensemble* (GUE), respectively. For an arbitrary large number of levels, the Wigner–Dyson distributions for $p(s)$ result from the ensemble properties. The calculation is carried out in [211]. These complicated distributions can be very well approximated by the Wigner surmises, which are the corresponding distributions for a two-level system with the same symmetry properties [42]. One distinguishes between pure distributions, where a spin degeneracy is absent, and bimodal distributions, where a δ function at $s = 0$ is introduced ad hoc, which takes a twofold spin degeneracy into account. Their most important features are summarized in Table 10.1.

The Wigner surmises for spin-degenerate systems are plotted in Fig. 10.16. Furthermore, calculating the Wigner surmise for the orthogonal ensemble is the topic of Exercise E10.3.

Up to now there has been no strict proof that quantized chaos should obey the predictions of RMT. The crucial point is whether the Hamiltonian matrix elements of a chaotic system are really random. Empirically, however, the agreement between RMT and experimental results is overwhelming. For example, the measured NNS distributions of microwave cavities and the excitation spectra of nuclei or of a hydrogen atom in a strong magnetic field are indistinguishable to the Wigner surmise. Numerical simulations show very

Tab. 10.1 Properties of the Wigner surmises of relevance. The average spacing is denoted by \bar{s} , while σ is the standard deviation.

Ensemble	GOE	GUE	Bimodal GOE	Bimodal GUE
$p(s)$	$\frac{\pi}{2}se^{-\pi s^2/4}$	$\frac{32}{\pi^2}s^2e^{-4s^2/\pi}$	$\frac{1}{2}[\delta(s) + p(s)]$	$\frac{1}{2}[\delta(s) + p(s)]$
\bar{s}	1	1	0.5	0.5
σ	$\sqrt{\frac{4}{\pi} - 1}$	$\sqrt{\frac{3\pi}{8} - 1}$	$\sqrt{\frac{2}{\pi} - \frac{1}{4}}$	$\sqrt{\frac{3\pi}{16} - \frac{1}{4}}$

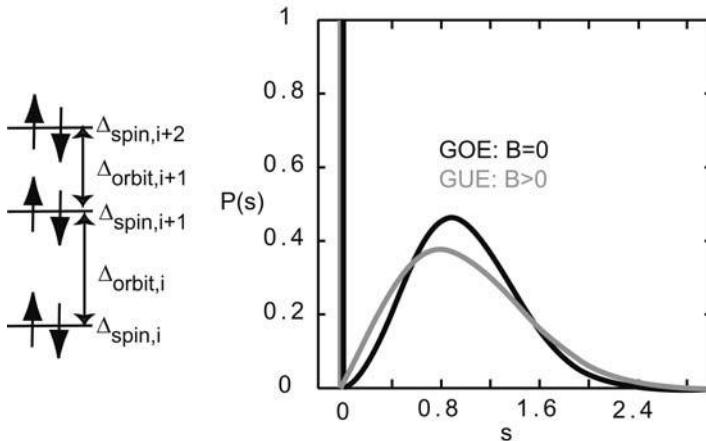


Fig. 10.16 Bimodal Wigner surmises for the Gaussian orthogonal (GOE) and the Gaussian unitary (GUE) ensembles.

good agreement as well (see Fig. 10.17). Most strikingly, small level separations are suppressed in chaotic systems, an effect known as level repulsion. It can be traced back to more anticrossings in systems with reduced symmetry.

In quantum dots, however, the experimentally obtained NNS distributions deviate from the bimodal Wigner–Dyson distribution expected within the constant interaction model. Experimentally, one subtracts the single-electron charging energy from the measured addition spectrum and plots a histogram of the remaining peak spacings. A typical example of such a measurement is shown in Fig. 10.18. There $P(s)$ does not resemble the expected traces at all. There is no signature of a bimodal distribution, nor does the FWHM agree well with the RMT prediction (see e.g. [233], [277] or [280]). It is natural to suspect that residual electron–electron interactions beyond the CI model cause the discrepancy. Clearly, the spin degeneracy is removed by such interactions, which broadens the spin peak and displaces it to a non-zero value. If these

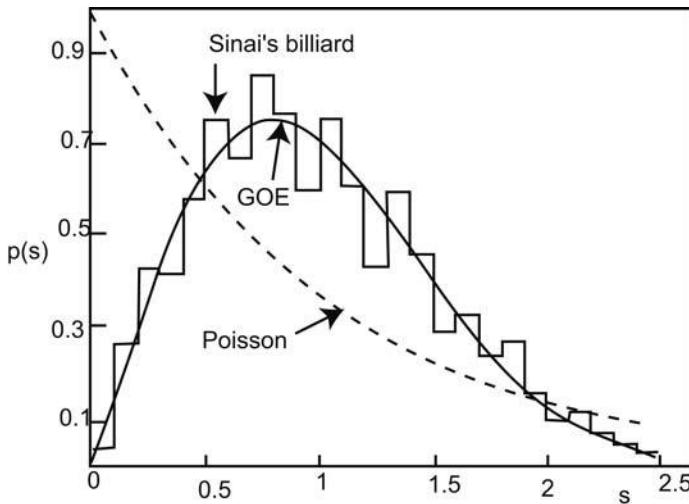


Fig. 10.17 Numerically calculated NNS distribution of a Sinai billiard (histogram, about 1000 eigenvalues have been used) vs. the GOE Wigner surmise (full line). For comparison, the Poisson distribution (dashed line) is shown as well. After [41].

residual interactions are strong enough, they can deform the bimodal Wigner surmise distributions into singly peaked distributions of a different shape. In that respect, the experimental NNS distributions serve as a reference measurement, to optimize various models for residual interactions in small electronic systems [306].

10.4

Shape of conductance resonances and current–voltage characteristics

As we have just seen, tuning a quantum dot with a gate voltage and looking at its conductance as a function of small source–drain bias voltages is an important experimental technique. So far, we have extracted information from the peak positions. It is self-evident to ask what kind of information is contained in the line shapes and the amplitudes of the conductance resonances. Clearly, the fluctuating amplitude is a measure for the coupling of the current-carrying state in the dot to the leads, and hence of the corresponding wave function amplitude close to the tunnel barriers. It varies from state to state and as a function of magnetic field. This fact can be used to compare the statistical properties of the wave functions with RMT – see Paper P10.3. Note that this is in marked contrast to the single-electron transistors discussed in Chapter 9, where many states couple very weakly to the leads, which results in an approximately constant peak amplitude.

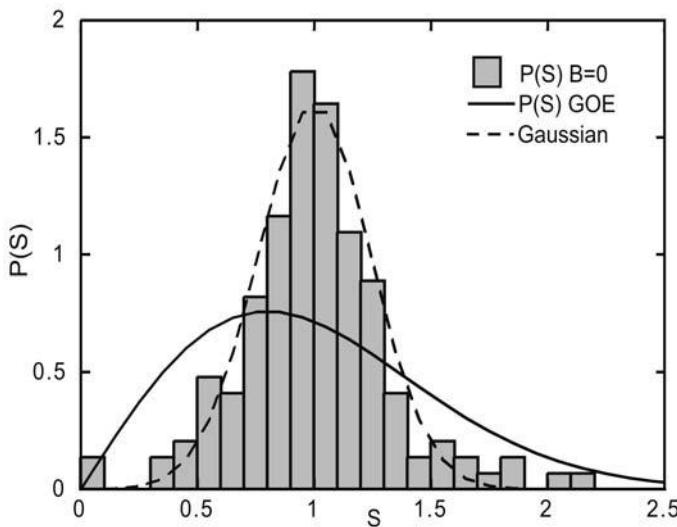


Fig. 10.18 Measured NNS distribution of a quantum dot in GaAs (histogram), in comparison to the Wigner surmise. The distribution look more like a Gaussian, and the bimodal structure is absent. After [277].

First of all, these considerations allow us to interpret the results of Fig. 10.5(b) qualitatively, as shown in Fig. 10.19. In (a), no current flows, since the source and drain electrochemical potentials μ_S and μ_D lie inside the Coulomb gap. As V_G is increased (b), level 2 gets aligned with μ_D , and an electron may tunnel into this level, a process that establishes the scenario depicted in (c). Now, one of the electrons in state 1 or 2 may tunnel into the source, such that both levels contribute to the coupling between the dot and source. As V_G is increased further, scenario (d) gets established at some point. Here, only level 2 contributes to the current, and the system oscillates between (d) and (e). Hence, the overall conductance will be smaller than in (b) and (c). Finally, consider the situation depicted in (f) and (g) at higher gate voltages. Here, the two empty states 2 and 3 lie in between μ_S and μ_D , which increases the coupling between the dot and drain. One of them will get occupied and re-emptied by the electron tunneling into the source. Hence, we expect a conductance resonance with a doublet shape, as observed in Fig. 10.5(b), trace B.

From these considerations, it becomes clear that the single-particle level spacings can be determined by high-bias transport experiments. In fact, a gate is not even necessary. As we increase the source–drain voltage, additional quantum dot states subsequently become accessible for transport, which is reflected in current steps, or in peaks in the differential conductance dI/dV , as a function of V (see Fig. 10.20). This is an important experimental tool for investigating quantum dots where a gate electrode is impossible to define; some

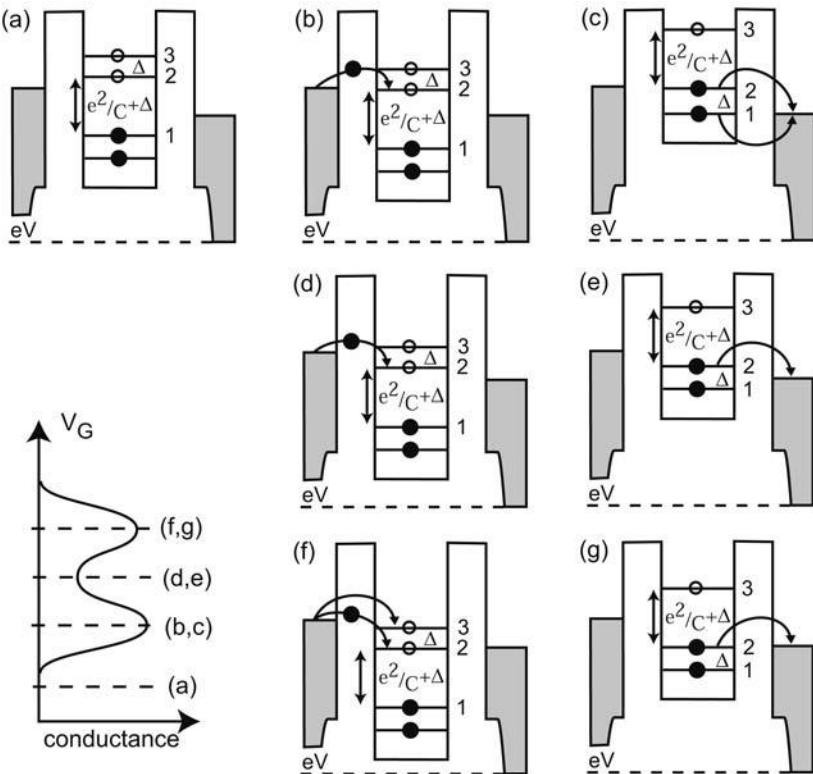


Fig. 10.19 Free energy diagrams of a quantum dot with a bias voltage comparable to the level spacing Δ applied. Full circles denote occupied dot states, while open circles indicate empty states. To the lower left, the resulting conductance resonance and the corresponding energies for each scenario (a)–(g) are sketched. This should be compared to the observed trace B in Fig. 10.5(b).

examples will be mentioned in the next section. The underlying theory has been developed in [16].

We have already discussed the resonance line shape at negligible source-drain voltages in the metallic regime $h\Gamma \ll \Delta \ll k_B\Theta \ll E_C$. This situation is usually not encountered in quantum dots. In many experiments, $k_B\Theta$, Δ as well as $h\Gamma$ are actually of the same order of magnitude. There is no general expression for the line shape for arbitrary values of these quantities. Complications arise due to the Coulomb interactions, which correlate the tunneling events across the two barriers, and due to the fact that the distribution function inside the quantum dot is usually not a Fermi-Dirac function. For small couplings $h\Gamma \ll k_B\Theta, \Delta$, the corresponding theory has been worked out in [23]. Even in this regime, the line shapes have to be calculated numerically. An analytical result is available, however, for one important limiting case, namely

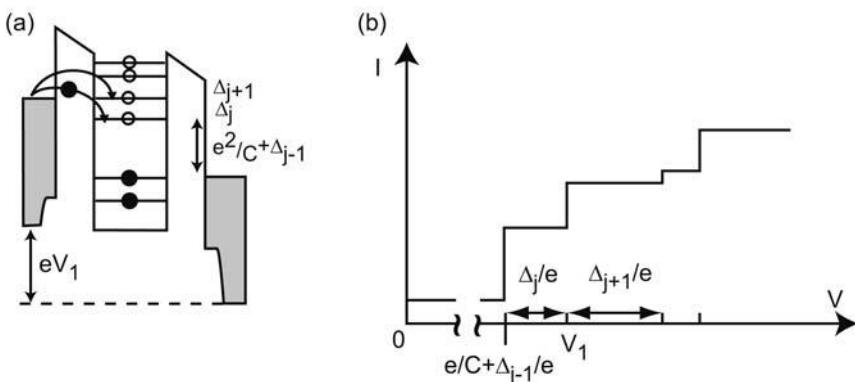


Fig. 10.20 (a) Spectroscopy on quantum dots by I - V measurements. At the voltage V_1 , two energy levels can carry current. (b) Tuning the voltage changes the number of current-carrying states, which is observed as steps in the I - V characteristic.

for $h\Gamma \ll k_B\Theta \ll \Delta \ll E_C$. In this regime, only a single level of negligible homogeneous broadening carries current. Now, the line shape equals

$$G(E) = \frac{e^2}{4k_B\Theta} \frac{\Gamma^S \Gamma^D}{\Gamma^S + \Gamma^D} \cosh^{-2} \left(\frac{E - E_{\max}}{2k_B\Theta} \right) \quad (10.7)$$

which is a generalization of the line shape discussed in Exercise E8.4. Note that the peak conductance now increases as $1/\Theta$, as long as $h\Gamma \ll k_B\Theta$, in contrast to the peaks in the metallic regime.

10.5

Other types of quantum dots

A laterally confined region in a semiconductor heterostructure is just one variation of a quantum dot. In this section, some other types are presented. It should be noted that large ensembles of quantum dots have been investigated for a long time – see, for example, the experiment by Giaever and Zeller [117] in Fig. 9.6. Later on, capacitance measurements on self-assembled quantum InAs dots embedded in GaAs have demonstrated for the first time experimentally that the single-particle level spacing Δ can be larger than the single-electron charging energy [79]. Also, single-electron charging of individual atoms which are part of metal-organic molecules has been demonstrated by measurements on arrays of such molecules [93]. Below, we will have a glance at experiments in which individual quantum dots are probed.

10.5.1

Metal grains

As pointed out in the previous chapter, the very first experiments related to single-electron tunneling have been performed on granular films of extremely small metal grains, which were embedded in an insulating matrix. In principle, the size of the Sn grains fabricated by Giaever and Zeller [117] (Fig. 9.6), for example, is small enough to observe discrete energy levels. As a rule of thumb, a grain radius of 10 nm suffices to observe size quantization in metals at a temperature of 100 mK. The challenge consists in contacting individual grains in order to avoid ensemble averaging. The length scale is clearly below the resolution limit of conventional lithographic techniques.

In principle, one way to access an individual grain is by contacting it with a scanning tunneling microscope, like in the experiment of Fig. 9.9. Discrete energy levels in individual InAs nanocrystals have in fact been observed with this approach [20]. In [248], the first transport experiment on single metallic quantum dots are reported. The authors used an ingenious fabrication technique, which combines self-assembly of nanometer-sized grains with conventional electron beam lithography (see Fig. 10.21). In a first step, a Si_3N_4 layer is etched. A patterned resist on top of the layer serves as etch mask. The etch is stopped at a point where a tiny hole of a few nanometers in diameter only has been formed. Subsequently, the bowl-shaped hole is filled with Al by thermal evaporation, and the Al is oxidized. Now a granular film of the metallic quantum dots to be investigated is evaporated on the bottom. This layer is subsequently covered by another oxide layer and a homogeneous Al electrode. By chance, one obtains devices this way where one grain sits right below the hole. A combination of self-assembly with the angle evaporation technique (Fig. 4.15) also produced working samples [67].

In none of these schemes could a gate electrode be defined, so that up to now all the information has been collected by current–voltage characteristics or by differential conductance measurements. Typical grain diameters range between 5 and 20 nm. The single-electron charging energies can be estimated from the self-capacitance of a sphere, $C = 4\pi\epsilon\epsilon_0 r$, while measurements give values of $E_C \approx 5\text{--}50$ meV. The single-particle level spacings depend on the energy in three dimensions and scale with the radius and the Fermi wave vector according to

$$\Delta \approx 2\pi^2\hbar^2/(m^*k_Fr^3)$$

Typical values of $\Delta \approx 0.1\text{--}1$ meV have been measured. Gold particles, as well as CdSe particles, have been attached to leads also by a hybrid assembly method, which is based on a combination of angle evaporation with organic layer deposition by wet chemistry. In this scheme, two metal electrodes with a gap of a few nanometers are patterned by electron beam lithography and an-

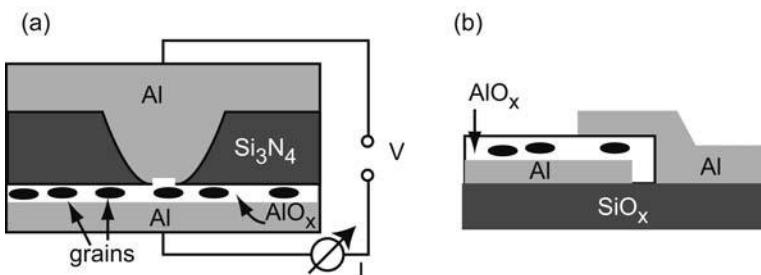


Fig. 10.21 Experimental setups for measuring individual metallic quantum dots. Scheme (a) has been demonstrated in [248], while scheme (b) has been successful in [67].

gle evaporation. Next, an organic molecule is deposited on the electrodes. The molecule 1,6-hexanedithiol can be used for this purpose. It has the property of binding with one end to the electrodes, while the molecules are oriented perpendicular with respect to the surface. Hence, the electrodes are covered with a molecular monolayer this way. In a subsequent step, the nanoparticles are deposited on this substrate from a solution. They bind to the second, dangling endgroup of the monolayer molecules, and there is a good chance that one grain gets deposited in between the two electrodes. In this case, the organic molecules to which the grain is bound serve as tunnel barriers. This approach has been used in [174, 175].

What can we learn from such experiments? First of all, it is no doubt of fundamental interest to investigate size quantization in metals. Up to now, Al [248], Au [67], Cu and Ag [238] as well as Co [130] grains have been investigated. Second, these experiments offer a variety of novel options not readily available in semiconductor quantum dots. The leads can be made superconductive [248], or energy levels of ferromagnetic grains like Co can be studied. In contrast to semiconducting quantum dots, the energy levels in metallic dots have been found to be spin-degenerate at $B = 0$, which suggests that exchange interactions are less important. Energy levels in Al grains show a remarkable clustering, while the effective g -factors in the grains are found to be reduced as compared to the bulk metal.

10.5.2

Molecular quantum dots

Single-electron tunneling transistors have also been built from carbon nanotubes [39, 296], as well as from the C₆₀ fullerene [231]. Since carbon nanotubes (CNs) are usually several micrometers long, it is possible to contact them by wires fabricated by conventional electron beam lithography. In several experiments, a suitable array of gold electrodes was defined on an insulating sub-

strate, and the carbon nanotubes were deposited from a suspension on the surface. By chance, a CN makes contact with two electrodes, while a third metal finger can be used as a gate [296]. Since CNs are quasi-one-dimensional structures, both E_C and Δ scale with $1/L$, where L denotes the length of the CN. Elementary considerations give $\Delta = h\nu_F/(4L)$, while $E_C \approx 5\Delta$ (see [58]). A unique picture of the electronic properties of CNs has not yet emerged. While some experiments indicate deviations from a Fermi gas behavior [40], others show excellent agreement with the constant interaction picture, including spin degeneracy and shell filling [58].

The C_{60} quantum dot measured in [231] shows not only an extremely large single-electron charging energy of 0.27 eV, but also a vibrational excitation of the C_{60} molecule as its charge is altered. These samples were made by depositing a C_{60} in between two electrodes with a separation as small as 1 nm. This electrode geometry has been fabricated by passing a large current through a thin gold wire made by an angle evaporation technique. The current induces migration of the gold atoms in the wire, which finally breaks, and a gap in the regime of 1 nm opens up. This phenomenon is known as electro-migration. The same kind of technique has also been used in [230] to contact individual CdSe clusters. Both types of carbon-based quantum dot devices are the subject of ongoing research, and a lot of further fruitful scientific results can be expected here.

As the size of the molecule becomes smaller, making electrical contact to it gets more and more difficult, and we are well advised to look for schemes in which the molecule does this job for us. This is why thiol-terminated molecules, in particular aromatic ones, have taken a prominent role in the field of molecular electronics. A prominent example is benzene-1,4-dithiol (BDT), i.e. a benzene ring with thiol units (SH^-) attached at carbons 1 and 4. In the presence of a gold surface, the hydrogen atoms of the thiols get desorbed, leaving behind benzene-1,4-dithiolate, whose sulfur atoms can form chemical bonds with Au atoms. Reed et al. [250] managed to sandwich a BDT molecule between two gold wires of a break junction (Fig. 10.22). Measurements of the current–voltage characteristic reveal a Coulomb-type gap of 0.7 eV, qualitatively similar to the one shown in Fig. 9.9, which can be observed very clearly at room temperature.



Fig. 10.22 Scheme of a BDT molecule contacted by two gold electrodes. After [250].

Interpreting such measurements, however, is far from trivial. First of all, the chemical bonds to the gold strongly modify the energy spectrum of the molecule. The geometry and the chemistry of the S–Au connection, as well as the geometry of the gold tips themselves, are not well under control, but influence the conductance of the molecular bridge. Self-consistent simulations of the BDT–gold structure [73] reveal qualitative agreement between the shapes of the I – V traces, but deviate strongly in terms of the absolute value of the current. This discrepancy is most likely determined by the details of the S–Au connection. Moreover, the gap observed in the I – V characteristics cannot be solely attributed to Coulomb blockade, but also reflects the configuration of the molecular orbitals. Heurich et al. [150] have shown that, in similar molecules, the contribution of a molecular level to the current across the molecular bridge depends on three factors: namely, its energy with respect to the Fermi level of the leads, its coupling to the leads, and the degree of localization within the molecule. As a result, many orbitals, even with energies far away from the Fermi level, may give significant contributions to the current.

This brief survey has attempted to show that scientists are about to establish novel fabrication techniques and assembly schemes to overcome the limitations posed by conventional lithographic techniques; and that they are learning how to attach wires and gates to particles as small as a nanometer, as well as to individual molecules. These are important steps toward routine operation of quantum dots at room temperature in electronic circuits, and toward molecular electronics.

10.6

Quantum dots and quantum computation

Quantum computation [219] offers several potential advantages in comparison to conventional computation schemes. Important concepts have been experimentally demonstrated already. The most impressive results so far have been obtained with trapped ions and photon systems – see, for example, [135] and [309], respectively. The application of quantum dot systems for quantum information processing is still in its infancy. Even very elementary operations have not yet been demonstrated experimentally. Quantum dot systems, however, have – at least potentially – some advantages. First of all, it seems desirable that a quantum computer would be defined in a solid state material. While superconductive circuits look promising as well in this respect, they will be limited to temperatures well below 300K. In contrast, at least in principle, it may be possible to realize a room-temperature quantum dot quantum computer. Moreover, a quantum dot quantum processing module embedded in an otherwise conventional semiconductor processor is a concept of potential interest to the semiconductor industry. Theoretical studies

discuss how quantum computation may be brought to reality with quantum dot circuits [298].

A system suited for quantum computation must enable the three key tasks outlined in the following subsections.

Preparation, manipulation, and detection of single qubits The preparation, manipulation, and detection of a single qubit, i.e. a two-level state

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad |\alpha|^2 + |\beta|^2 = 1 \quad (10.8)$$

with a sufficiently long lifetime is considered to be possible in principle with quantum dots.

Imagine that the two eigenstates of the qubit are the two spin states of the first orbital energy level of a quantum dot. Filling one electron in the dot forms a qubit. It has been demonstrated that the spin lifetime in such quantum dots is very large and may exceed 100 μs , much longer than required for a typical computational operation. But how can the dot be prepared in the requested superposition? For a pure state like $|0\rangle$, this could be done by letting an electron tunnel onto the dot from a spin-split edge state in the lead, or from a QPC that transmits only one spin direction.

A unitary transformation (a *quantum gate*) is a manipulation of the qubit, which can also be part of the preparation of the initial state. For example, a very important operation is the *Hadamard gate*, represented by the matrix

$$H \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

It generates states with equal probabilities $|\alpha|^2$ and $|\beta|^2$ from pure states, e.g.

$$H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$$

Such unitary transformations (which can be thought of as rotations on the Bloch sphere – see Exercise E10.3) could be performed by applying a magnetic field pulse of appropriate polarization, strength, and duration to the quantum dot, which causes the spin to precess around the direction of this field about the desired angle. Here, the concept of universal quantum gates is of great help. In quantum computation, like in classical digital electronics, it is possible to find a small number of quantum gates from which any unitary operation can be composed.

A different kind of necessary manipulation is the transport of the qubit in the circuit. While the gate structure defining the quantum dot is fixed, a gate sequence that is able to displace the dot potential as a local potential minimum

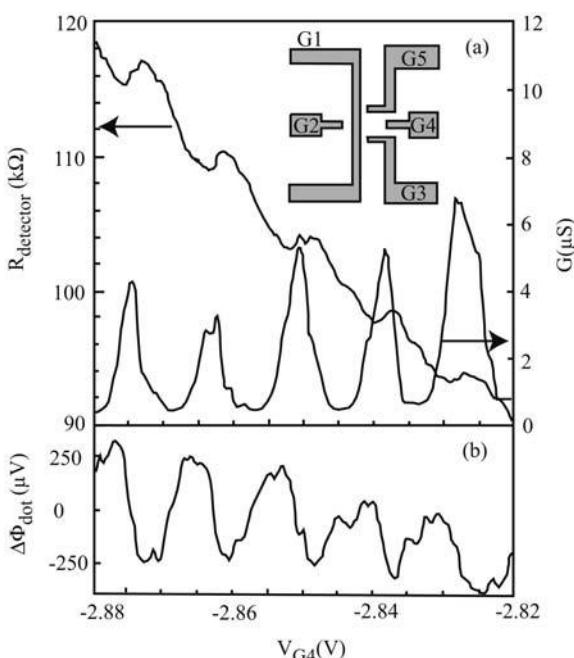


Fig. 10.23 (a) An experimental setup (inset) for the charge readout of a quantum dot. The dot's charge influences the confining potential of the QPC to its left. The main figure shows that the Coulomb blockade oscillations (bottom trace) are reflected in the QPC resistance (top trace), which can be translated into the variation of the dot potential (b). After [92].

over large distances, similar to the readout process in charge coupled devices, may be suited to provide such a tool.

As an example for a detection, the quantum dot could be emptied via an outgoing channel which is constructed like the incoming one just sketched. In order to decide that the electron has actually come from the specific quantum dot, it may be necessary to detect the charge of the dot with a nearby, sensitive detector. This could be a metallic SET transistor, or more simply just a quantum point contact close by (see Fig. 10.23).

Generation of entanglement between at least two qubits Entanglement is a property of some states that exist in (tensor) product spaces, like that formed by the vector spaces of two spins. An entangled state in our example is a state that cannot be written as a product of states of the individual spin spaces. A famous entangled state is the first Bell state

$$|B1\rangle \equiv \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$$

Entangled states have some remarkable properties, one of which is that their elements of the individual spaces have correlations that exceed those possible within classical physics. These properties are key ingredients in almost all quantum computation schemes, and one has therefore to find ways to prepare entangled states with quantum dots. A possible path is to use adjacent quantum dots whose coupling can be tuned by gates in between, in combination with the appropriate quantum gates.

Taking quantum errors into account Quantum errors must be taken into account. All kinds of external influences can, for example via a spin transition, cause the loss of information. Such errors must be corrected. This has developed to a major research topic in quantum information technology. Moreover, one can make the qubit inherently less sensitive to errors by choosing the adequate physical implementation in a given concept. For the quantum dot concept, for example, it has been suggested to use the singlet and triplet states of a double dot as the basis states of a qubit [298].

The implementation of quantum computation circuits will be one of the driving forces of the field in the next few years, and the reader is encouraged to become acquainted with the principles of this fascinating field via the specialized literature. Quantum information not only challenges our understanding of how the world works, but also draws semiconductor physics and technology into new territory.

Papers and Exercises

P10.1 In [297], the density of states is reconstructed from the transmission resonances of a quantum dot. Describe the sample design and the reconstruction of the dot's energy spectrum. How do Hund's rules influence the data?

P10.2 Evans et al. [86] have developed an electrostatic model for a quantum dot with a filling factor below 2. Explain the dot's "phase diagram" within this model.

P10.3 The statistical properties of the transmission resonance *amplitudes* of quantum dots are discussed in [100] and [52]. What are the results?

P10.4 Describe the experiment carried out in [31].

P10.5 In [244], a DNA quantum dot has been measured. Summarize the sample fabrication and the results.

E10.1 This exercise deals with the Fock–Darwin model.

- (a) Show that Eqs. (10.4) and (10.5) are equivalent. What happens for $B \rightarrow \infty$?
- (b) Consider the Fock–Darwin spectrum for $\omega_0 \ll \omega_c/2$. Show that adjacent states with identical m have an energy spacing of

$$\Delta E = E_{m,p+1} - E_{m,p} \approx \hbar \frac{\omega_0^2}{\omega_c}$$

and a spacing of

$$\Delta B \approx B \frac{\omega_0^2}{\omega_c}$$

- (c) Use these approximate expressions to analyze the data of Fig. 10.8.

E10.2 In this exercise, the Wigner surmise will be derived for a 2×2 Hamiltonian matrix in the orthogonal case.

- (a) Consider the Hamiltonian of a chaotic system, which, in some basis, can be written as

$$(H) = \begin{pmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{pmatrix}$$

Assume that the matrix has eigenvalues λ_+ and λ_- . Express these in terms of the matrix elements H_{ij} .

- (b) Each matrix element is supposed to be random, which – by definition – means that the probability for a certain matrix to occur can be written as a product of the probabilities for the matrix elements, i.e.

$$p(H) = p_{11}(H_{11})p_{12}(H_{12})p_{22}(H_{22})$$

Further $p(H)$ has to be invariant under orthogonal basis transformations, which means that

$$p(O^T H O) = p(H)$$

with

$$(O) = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}$$

For our purposes, it suffices to consider only very small transformation angles $\alpha \ll 2\pi$.

Approximate O to first order in α , and derive a system of differential equations for $p_{ij}(H_{ij})$ [use the requirement that $O^T H O$ must be equal to H , independent of α].

Solve the differential equations.

- (c) Show that by a suitable choice of the zero-point of the energy, one can write

$$p(H) = c_1 \exp(-\text{Tr}[H^2])$$

- (d) Substitute the variables in $p(H)$, such that it becomes a function of $\Delta = \lambda_+ - \lambda_-$, plus a second variable. Recall that the transformation law for probability densities, $p(y) = p(x) |\partial y / \partial x|$ generalizes to

$$p(y_1, \dots, y_n) = p(x_1, \dots, x_n) \det \left(\frac{\partial(y_1, \dots, y_n)}{\partial(x_1, \dots, x_n)} \right)$$

where

$$\left(\frac{\partial(y_1, \dots, y_n)}{\partial(x_1, \dots, x_n)} \right)$$

denotes the Jacobian matrix of the transformation. To determine it, consider a suitable transformation that maps H onto its diagonal form.

- (e) Finally, integrate over the second variable that comes into play.

- E10.3** Fig. 10.24 reproduces the Bloch sphere as a graphical representation of a qubit state. In this picture, the qubit state is defined by two angles, θ and ϕ , according to

$$|\psi\rangle = \cos(\frac{1}{2}\theta)|0\rangle + e^{i\phi} \sin(\frac{1}{2}\theta)|1\rangle \quad (10.9)$$

Derive this expression starting from Eq. (10.8).

Note that states that differ only by a global phase factor give identical measurement results.

Note further that after mapping Eq. (10.8) onto spherical coordinates in the most natural way, each qubit state occurs twice on the resulting sphere.

- E10.4** Prove that the first Bell state cannot be written as a tensor product of two arbitrary members of the individual qubits' state spaces.

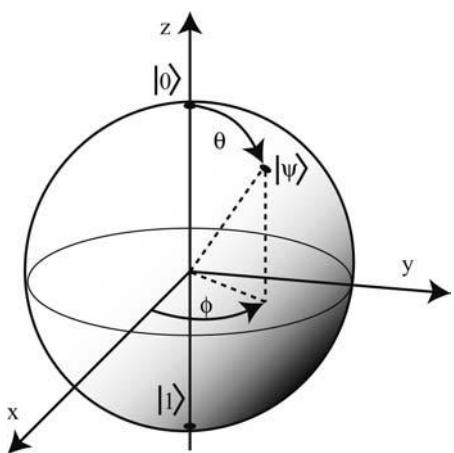


Fig. 10.24 Representation of a qubit state as a point on the Bloch sphere (for Exercise E10.3).

Further Reading

A good introduction to all aspects of quantum dot physics is given in the book by Jacak et al. [165]. A must-read contribution related to the transport properties of quantum dots is [180]. As an introduction to quantum computation and quantum information, the outstanding book [219] is recommended.

This Page Intentionally Left Blank

11

Mesoscopic Superlattices

In the previous chapters, we have been mostly studying individual quantum films, wires, or dots, respectively. A rich phenomenology emerges from packing these structures into periodic arrays. In Section 6.2.3, for example, a stack of quantum films was used to investigate the behavior of the quantum Hall effect as the dimension is gradually changed from two to three. Multilayers of epitaxially grown quantum films are fascinating objects, but beyond our scope here. In fact, the elementary Kronig–Penney potential, i.e. a periodic array of rectangular barriers, can be easily prepared by molecular beam epitaxy. Such samples have been used to demonstrate fundamental effects, like the Wannier–Stark localization, or Bloch oscillations. The reader is referred to textbooks on solid state physics for further information, like e.g. [127].

In this chapter, we focus on the most elementary mesoscopic phenomena that occur in lateral periodic structures, *lateral superlattices*, which are imposed onto a 2DEG by lithographic techniques. Here, the superlattice period is much larger than the lattice constant of the crystal, but comparable to mesoscopic length scales, in particular to the Fermi wavelength and the elastic mean free path.

What, the esteemed reader may ask, can be interesting in such systems? Are we not just rebuilding conventional solids on a different length scale, and moreover with reduced quality? Well, not quite. For example, recall that, during the construction of a localized electronic wave packet in Section 2.3.4, it emerged that the wave packet extends over many lattice constants. In artificial superlattices, however, the wave packet can be localized on the scale of the superlattice constant, and a classical picture of the electron motion is justified. In this picture, the superlattice potential V_{sl} plays the role of scatterer for the electrons, but the random character assumed in the Boltzmann model (via the relaxation time approximation) is absent. Therefore, deviations from Eqs. (2.59) can be expected. Moreover, magnetic fields of moderate strength may force the electrons on cyclotron orbits with diameters comparable to the lattice constant, a scenario that is impossible for natural crystals.

This does not mean that the superlattice band structures are beyond experimental reach, though. Recently, impressive progress has been made in this

respect, and striking signatures of superlattice bands, in particular of their fascinating behavior in quantizing magnetic fields, have been observed. In the future, we may want to design band structures not only in the direction of epitaxial growth, but in all directions, which will most likely require some lateral patterning.

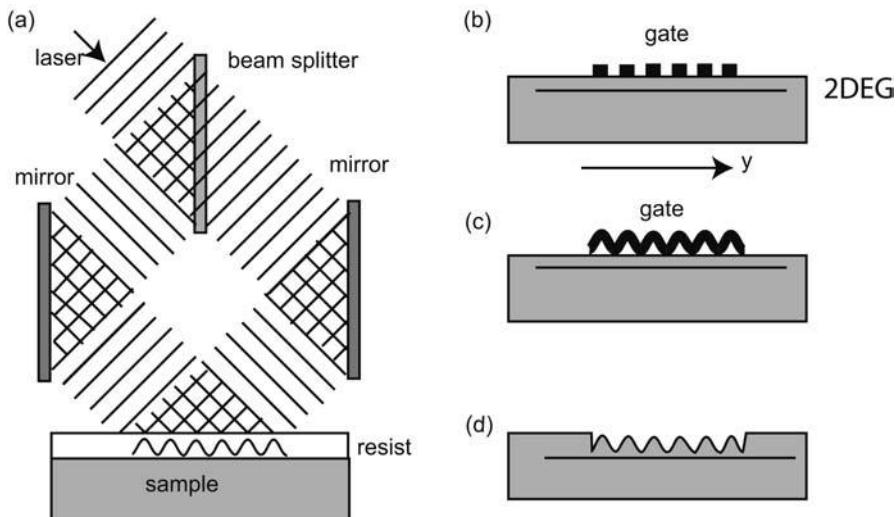


Fig. 11.1 Sketch of various techniques to define a periodic lateral superlattice. (a) A photoresist is illuminated by the periodically modulated intensity of the pattern that emerges from the interference of two partial laser beams. The resist can be used (b) as a mask for a lift-off process, (c) to modulate the distance between the 2DEG and a homogeneous top gate, or (d) simply as an etch mask.

11.1 One-dimensional superlattices

Lateral superlattices can be patterned by holographic schemes (see Fig. 11.1) or by electron beam lithography. While the interference pattern of a laser generates extremely accurate periodicity, the superlattice constants a are subject to the usual limitations, which means $a \geq 200 \text{ nm}$. Smaller periods can be generated by electron beam lithography or scanning probe lithography, but the deviations from perfect periodicity become larger. If the one-dimensional superlattice on the sample surface is used to impose a *weak* density modulation in the 2DEG, novel magneto-resistivity oscillations are observed in ρ_{xx} , i.e. perpendicular to the modulated direction (y -direction) [322, 332]. At first sight, they resemble Shubnikov-de Haas oscillations (see Fig. 11.2). They occur only at small magnetic fields and vanish for cyclotron radii above the su-

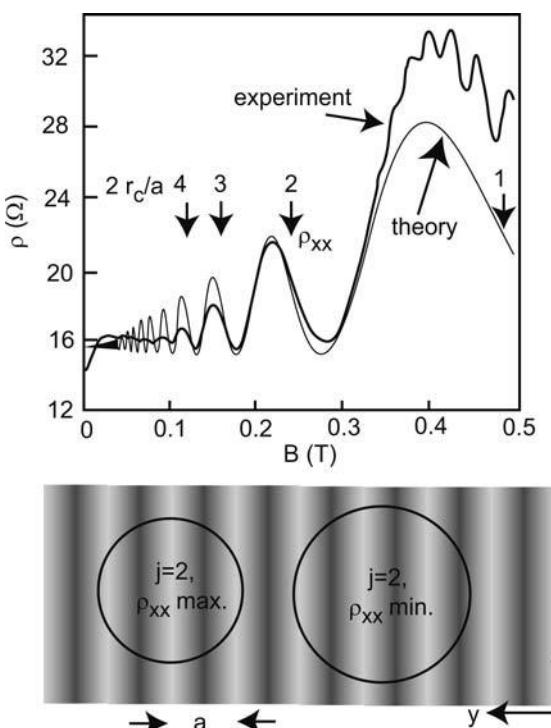


Fig. 11.2 Top: Measured and calculated longitudinal magneto-resistivities ρ_{xx} of a 2DEG with a density modulated in the y -direction. Oscillations are observed in ρ_{xx} at small magnetic fields. The Shubnikov-de Haas oscillations are visible above $B = 0.4$ T. After [22]. The resistivity in the y -direction remains essentially unaffected by the superlattice. Bottom: Cyclotron orbits in a maximum (left) and in a minimum (right) of ρ_{xx} . The gray scale indicates the electric field strength in the y -direction.

perlattice period. Note that in Fig. 11.2 the Shubnikov-de Haas oscillations set in at $B = 0.4$ T. The additional oscillations are periodic in $1/B$ as well. Essentially no effect is observed in the y -direction. These oscillations are known as *Weiss oscillations*.

It seems strange that a density modulation of only a couple of percent is able to produce such a strong modification of ρ_{xx} . Theoretical considerations revealed that the Weiss oscillations can be understood in terms of a resonant drift of the cyclotron orbits induced by the electric field of the superlattice. The drift can be calculated by treating the superlattice electric field as a perturbation to the Hamiltonian for the cyclotron motion of free electrons [115, 332]. The effect can also be understood in a semiclassical picture, by considering the total $\vec{E} \times \vec{B}$ drift that an electron experiences during one complete cyclotron orbit [22]. We assume $r_c \gg a$. At certain cyclotron radii, this drift will average to

zero. At slightly different magnetic fields, however, the drift will average out along those parts of the cyclotron trajectory where the electron moves in the y -direction and thus crosses many modulation periods. A large integrated $\vec{E} \times \vec{B}$ drift is collected for the motion in the x -direction, though. If this drift has the same sign for the electron motion in the positive and negative x -direction, a net drift will remain for a complete cyclotron motion. The result from the calculation in the limit of $r_c \gg a$ reads

$$\rho_{xx} \propto \frac{V_{sl}^2 B}{a E_F^{5/2}} \cos^2 \left(\frac{2\pi r_c}{a} - \frac{\pi}{4} \right) \quad (11.1)$$

This expression predicts minima in ρ_{xx} for $r_c/a = (4j+3)/8$, and maxima at $r_c/a = (4j+1)/8$, where j is an integer, including zero. The corresponding cyclotron orbits are sketched for $j=2$ in Fig. 11.2. Furthermore, it predicts that the oscillations are periodic in $1/B$, that the oscillation amplitude increases as B is increased, and that the oscillations vanish for $r_c < a$, in good agreement with the experiment. This result holds only for $r_c \gg a$, however. It emerges from approximating the Bessel function

$$J_0(x) \approx \sqrt{2/(\pi x)} \cos(x - \pi/4)$$

which is a good approximation for large arguments only. In Fig. 11.2, the correct expression containing the Bessel functions is compared to the experimental data. The agreement is excellent. It should be noted that the theory assumes a sinusoidal superlattice potential. Other potential shapes will give phases that differ from $\pi/4$ in Eq. (11.1), without changing the overall appearance.

11.2

Two-dimensional superlattices

After the previous discussion, it is self-evident to ask what happens to 2DEGs with potential modulations in both directions. Corresponding gate structures can be easily made by adopting the schemes mentioned in relation to Fig. 11.1, for example by performing a second illumination with the interfering laser beam, after rotating the sample by 90° .

11.2.1

Semiclassical effects

We first focus on *antidot lattices*, which are two-dimensional arrays of holes in a 2DEG. In a simple semiclassical picture, one would expect that there exists a set of discrete magnetic fields at which the cyclotron orbits fit in the lattice

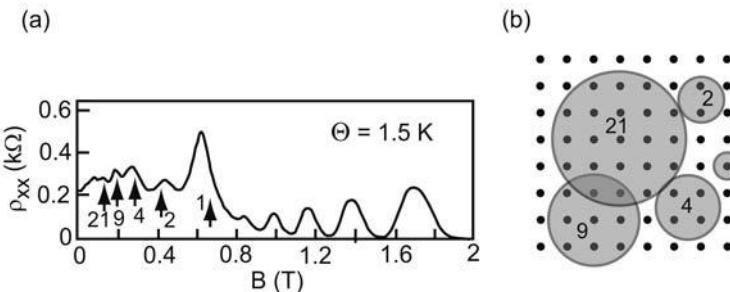


Fig. 11.3 (a) Longitudinal magneto-resistivity of (b) a square antidot lattice (black circles). Peaks in ρ_{xx} are observed if the cyclotron orbit fits in between the antidots. The arrows and numbers in (a) denote the magnetic fields at which the cyclotron orbit is commensurate with the antidot lattice, and the number of antidots enclosed in the cyclotron orbit, respectively. The lattice period in this experiment was 300 nm. Adapted from [323].

without hitting a scatterer. These orbits are called *commensurate* with the lattice. For such magnetic fields, the antidot lattice localizes the electrons, and we expect an increased resistivity. As can be easily verified, for a square lattice, this is possible for cyclotron orbits that enclose 1, 2, 4, 9, 21, ... antidots. Note that this would be in striking contrast to the magneto-resistivities obtained within the Boltzmann model, which predicts ρ_{xx} to be independent of B .

This behavior has in fact been observed (see Fig. 11.3). A closer look, however, reveals that this is not the end of the story. First of all, the resistance maxima are not exactly at the expected positions. Also, a *negative* Hall resistance is observed [323], which cannot be understood in this simple picture [98]. Finally, the experiments have been carried out at non-zero forward bias voltages, which should disturb the commensurate electronic cyclotron motion and destroy its pinning to the superlattice. To resolve this set of questions, we will apply the Kubo formalism mentioned in Chapter 5. Along the way, we will get to know further useful tools and techniques.

In a generalized version of Eq. (5.12), the Kubo formula tells us that the components of the conductivity tensor can be calculated by averaging the velocity correlation functions according to

$$\sigma_{ij} = \frac{m^* e^2}{\pi \hbar^2} \int_0^\infty e^{-t/\tau} C_{v_i v_j}(t) dt \quad (11.2)$$

Here, the term $e^{-t/\tau}$ is included, which models the momentum relaxation after the scattering time τ within the relaxation time approximation. In a numerical simulation, electrons are randomly placed within a unit cell of the antidot lattice. Their velocity vector has the amplitude of the Fermi velocity, and points in an arbitrary direction within the (x, y) plane. Of course, points where

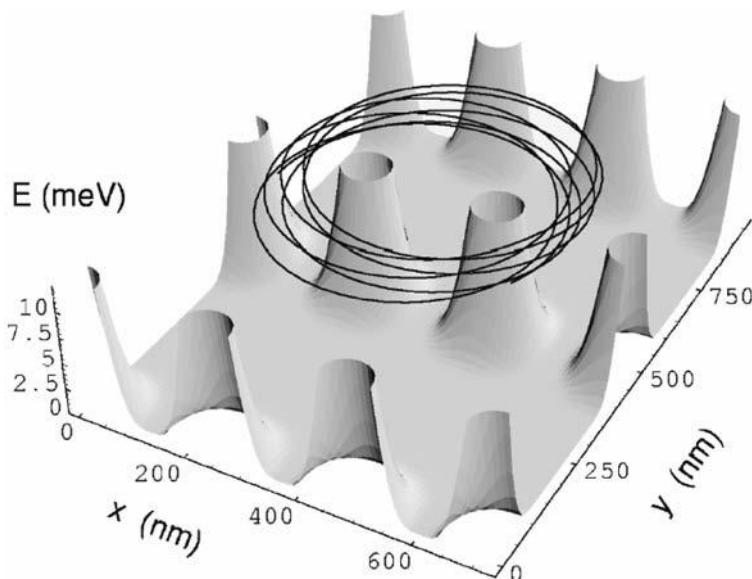


Fig. 11.4 A calculated trajectory of an electron moving in a model potential. After [256].

the antidot potential is above the Fermi level are not starting points. The motion of an ensemble of electrons in the antidot potential and the magnetic field is calculated. The antidot potential of a square lattice can be approximated, for example, by

$$V(x, y) = V_0[\cos(\pi x/a)]^{2\beta}[\cos(\pi y/a)]^{2\beta} \quad (11.3)$$

where β tunes the steepness of the antidot walls. After time τ , the direction of the velocity can be randomized numerically, in order to take residual scattering at random positions into account. If the antidot potential is set to zero, the Drude result is recovered.¹ In Fig. 11.4, a model potential and a sample trajectory are shown.

After a time that is large compared to τ and after the trajectories of a sufficient number of electrons have been simulated, the velocity correlation function, and hence the components of the conductivity tensor, can be determined. The agreement between such simulations and the observations is quite good in many cases, although the walls are soft in real samples, which may displace the commensurability peaks along the magnetic field axis. But why is the electric field due to the source-drain bias voltage apparently not destroying the pinning? This question has been answered in detail in the seminal paper of Fleischmann et al. [97]. The authors incorporated the driving electric field

1) This is demonstrated in Exercise E5.2.

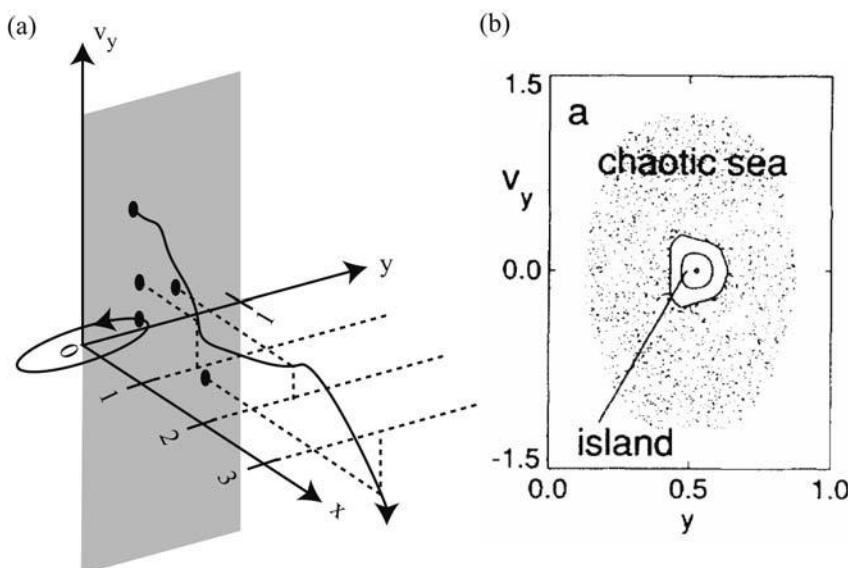


Fig. 11.5 (a) Scheme of a Poincaré section in the (y, v_y) plane.
 (b) Poincaré section of numerically simulated electron trajectories in a square lattice. Taken from [97].

in the simulations and studied the electron motion in the four-dimensional phase space spanned by (x, y, v_x, v_y) . This was done by means of Poincaré sections (see Fig. 11.5). In this representation, the trajectory of the electron motion in phase space is mapped onto the (y, v_y) plane by the following rules (see Fig. 11.5(a)):

1. Define the (y, v_y) plane at $x = 0$ as the monitoring plane; y extends over the unit cell of the lattice in which the electrons were injected, while the relevant range of v_y is limited to $[-v_F, v_F]$.
2. Watch the electron motion in phase space; as soon as it passes one of the $x = ja$ planes, mark the point of intersection in the monitoring plane as its projection parallel to the x -axis.

Interestingly, in this way one observes that, in the absence of driving electric fields, the phase space consists of two separate regions, called the *chaotic* and the *regular* regions (see Fig. 11.5(b)).

Electrons that started inside a certain interval of initial conditions generate spots that are homogeneously distributed across one region, which is called chaotic, since these electrons follow a chaotic motion. Those electrons that started in the remaining interval of initial conditions, on the other hand, generate closed loops or even just single points in the monitoring plane, which indicates a trajectory residing on the surface of a torus in phase space (known

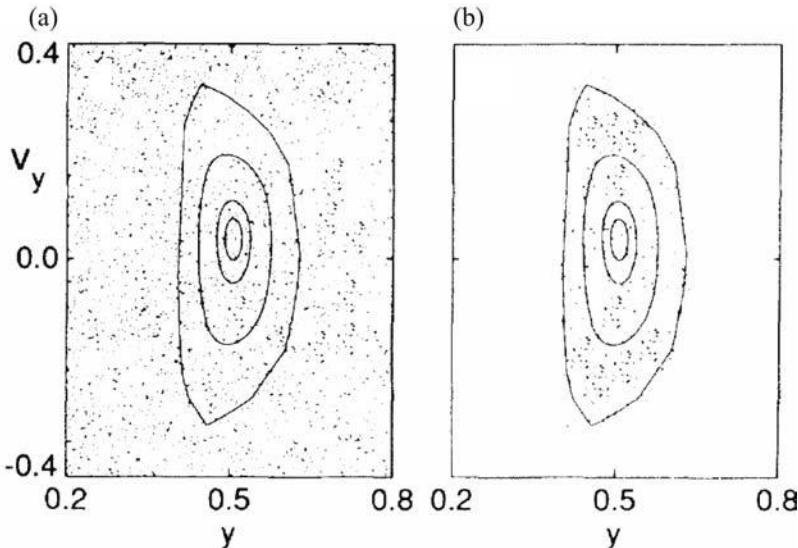


Fig. 11.6 (a) The initial conditions of 5000 electrons in a square lattice and under the influence of a driving electric field in the x -direction, represented in the Poincaré section. (b) Poincaré section of the setting of (a), representing only those electrons that pass through the planes within a time interval with a length of one cyclotron time, long after the initial conditions. Only electrons inside the regular region remain. Taken from [97].

as Kolmogorov–Arnold–Moser tori), or a time-independent cyclotron orbit, respectively. The system is said to have a mixed phase space, where regular islands are embedded in a sea of chaotic motion.

In [97], the Poincaré section mapping has been used to investigate the effect of a driving electric field on the electron dynamics (see Fig. 11.6). An ensemble of electrons was injected in the antidot lattice in a weak electric field, and those electrons were monitored that, after a long time, generated a point monitoring plane within a time window with a length of the cyclotron time. It turned out that, even with a weak electric field applied, the electrons that started out inside a regular region remain therein, while all other electrons have escaped the monitored region (in the y -direction). This numerical result indicates that even a bias voltage does not mix the chaotic and regular regions of the phase space.

Based on this observation, it is clear that the electrons in the regular regions cannot carry current, which is rather carried by electrons with chaotic dynamics. Therefore, the conductance is obtained from the contribution of the chaotic electrons, i.e.

$$\sigma_{ij} = (1 - f_r)\sigma_{ij}^{\text{chaotic}} \quad (11.4)$$

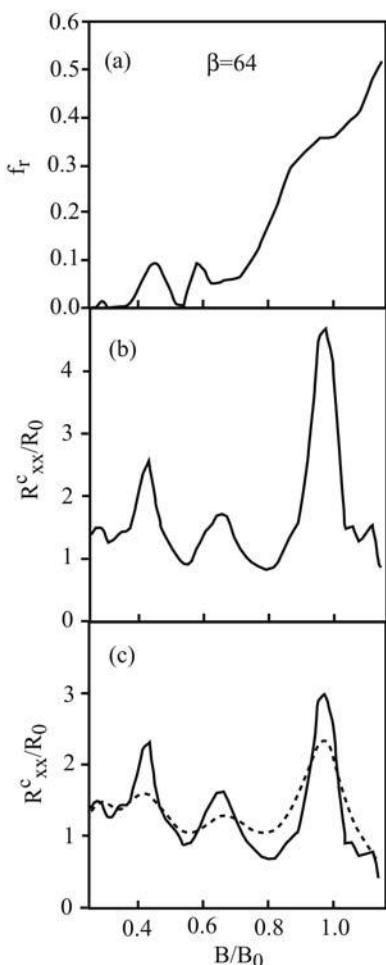


Fig. 11.7 (a) Fraction of the phase space volume of the regular regions of the phase space for a square antidot lattice, (b) the longitudinal resistance of the chaotic electrons, and (c) the calculated resistance (full line) as compared with that measured in [323] (dashed line). Adapted from [97].

where f_r is the volume fraction of the regular phase space and $\sigma_{ij}^{\text{chaotic}}$ denotes the conductivity of the electrons in the chaotic regions. Therefore, the structure observed in the magneto-resistivity can originate either from a varying size of the regular phase space or from the dependence of $\sigma_{ij}^{\text{chaotic}}$ on the magnetic field. By comparison of $f_r(B)$ with $\rho_{xx}(B)$, it can be concluded that the chaotic electron dynamics, and not the size of the regular regions, dominates the magneto-transport in antidot lattices (see Fig. 11.7).

11.2.2

Quantum effects

In hexagonal antidot lattices, additional oscillations can be observed, which are periodic in B (see Fig. 11.8) [220]. The strong temperature dependence of these oscillations is in contrast to the very weak temperature dependence of the commensurability peaks and indicates a phase coherent origin. This observation is thus explained in terms of an enhanced classical backscattering probability in hexagonal lattices as compared to that in square or rectangular lattices. The resulting enhanced backscattering due to phase coherence generates Altshuler–Aronov–Spivak (AAS) oscillations (see Chapter 8) of significant amplitude.

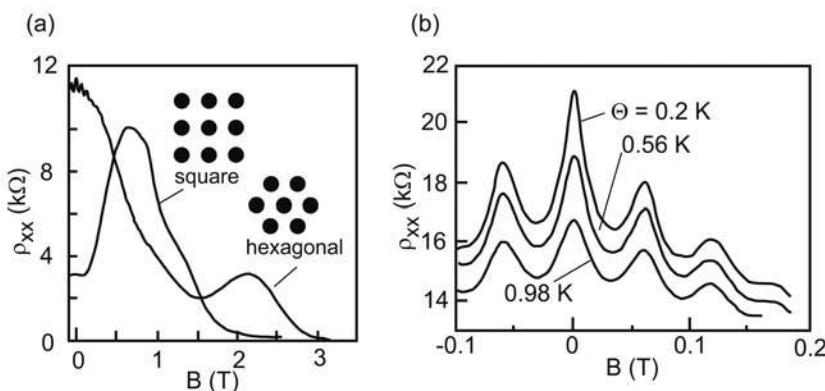


Fig. 11.8 (a) Longitudinal magneto-resistivities in square and hexagonal antidot lattices (the lattice geometries are shown in the inset). (b) Enlargement of ρ_{xx} for the hexagonal lattice around $B = 0$, which oscillates with a period of $\Delta B = h/2eA$, where A is the average area of one antidot. The strong temperature dependence indicates a phase coherent origin. After [220].

We conclude our excursion to lateral superlattices with a glance at the interesting quantum mechanics of a weak, two-dimensional, superlattice with a square geometry. In the limit of a free electron gas in a magnetic field, highly degenerate Landau levels are obtained – see the Landau fan reproduced in Fig. 6.3(b). The electrons move in cyclotron orbits, and the wave functions are thus rotationally invariant. In a periodic potential modulation and in the absence of a magnetic field, the wave functions are Bloch functions and have the corresponding discrete translational invariance. It has been predicted that the combination of periodic potentials and quantizing magnetic fields leads to a very rich phenomenology – see [239] for an introduction to this problem. We consider one interesting limit in more detail, namely the limit of weak periodic potentials in strong magnetic fields. Here “weak” refers to a scenario where the Landau splitting is larger than the amplitude of the modulation potential.

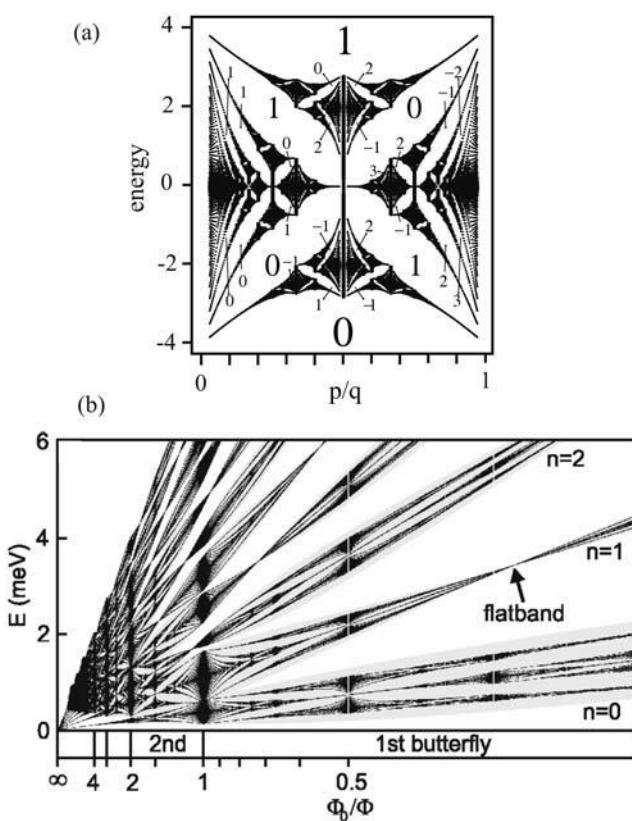


Fig. 11.9 (a) Calculated energy spectrum for one Landau band in a square superlattice. The ratio $p/q \propto 1/B$ measures the number of flux quanta h/e in units of BA , where A denotes the area of the unit cell. The numbers inside the figure indicate the value of the Hall resistance in the corresponding minigaps, in units of e^2/h . After [287]. (b) Effect of the superlattice on the Landau fan. Reprinted from [114].

Theoretically, it is well established that such a weak periodic potential lifts the degeneracy of the Landau levels and induces minibands separated by bandgaps, as you may have expected. This structure has quite unusual properties, though. The essential structure of these minibands is represented for a square lattice in Fig. 11.9(a). The band extends over an energy range that equals the amplitude of the superlattice potential and is periodic in $1/B$, with a period that corresponds to adding one flux quantum $\Phi_0 = h/e$ per superlattice unit cell. It is thus convenient to plot the spectrum as a function of Φ_0/Φ , where $\Phi = a^2B$. This structure is known as a *Hofstadter butterfly* and has been predicted by Hofstadter [151] for natural crystals, where its observation requires extremely strong magnetic fields of thousands of teslas. Scaling up the lattice constant has brought this spectrum within experimental reach.

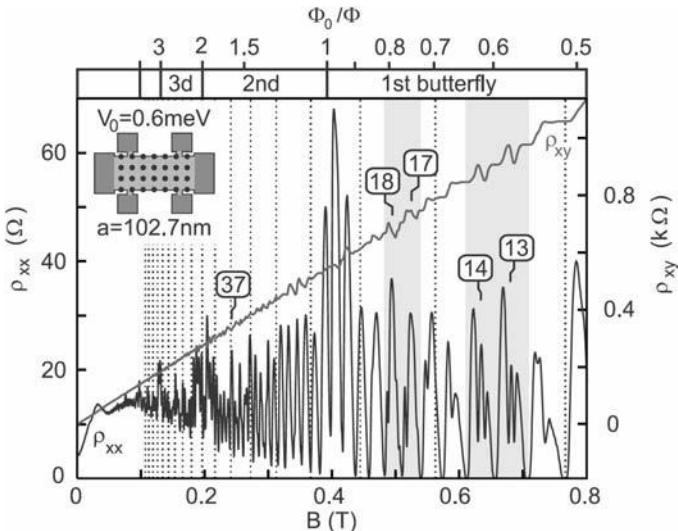


Fig. 11.10 Transport signatures of the Hofstadter butterfly as observed by Geisler et al. [114]. While the Shubnikov–de Haas resonances show a fine structure, the Hall resistance jumps correspondingly between different steps within one quantum Hall plateau (the integer numbers indicate the Landau level index). This fine structure reflects the motion of the Fermi level across a butterfly within a Landau level. Reprinted from [114].

Most interestingly, the spectrum has a fractal structure. For magnetic fields where Φ_0/Φ is a rational number represented by p/q with p and q both integers with no common divisor, the Landau level splits up into p subbands. A more detailed analysis reveals that, in addition, the subbands get modulated by an analytic function, such that the Landau fan of Fig. 6.3(b) evolves into the structure shown in Fig. 11.9(b). Moreover, as has been shown in [302], each minigap inside a Landau level makes a specific contribution to the Hall conductance, which is also quantized in units of $2e^2/h$. Hence, the quantum Hall plateaus should show internal steps according to

$$\sigma_{xy} = \frac{2e^2}{h}(j+k)$$

where the integers j and k denote the Landau level index and the contribution specific to the minigap at which the Fermi energy resides.

This behavior has been clearly observed in a beautiful experiment [114] (see Fig. 11.10). As expected, the minibands generate steps in the Landau plateaus as well as the corresponding minima in the Shubnikov–de Haas oscillations.

In the future, we can expect that the lateral superlattices are driven further into the quantum regime, such that the details of this aspect of mesoscopic

physics can be investigated. Here, probably self-assembly techniques will be required to reduce the lattice constants well below 100 nm.

Papers

- P11.1** In [268], rectangular antidot lattices with $a \ll b$ have been investigated. Discuss the observations and relate them to the experiments on quantum wires with non-specular walls (Section 7.1.2). Compare the results also to those of Section 11.1.
- P11.2** Arrays of antennas in microwave fields are very similar to electrons in periodic superlattices. Discuss this analogy; use [183] as an example.
- P11.3** In [324], Weiss et al. report the finding of magneto-oscillations in antidot lattices that are *periodic* in B . Explain their origin.

This Page Intentionally Left Blank

12 Spintronics

A recent branch of mesoscopic physics investigates the control and manipulation of the electron spin in metals and semiconductors. The name *spintronics* has become established for this field. Spin effects that show up in the resistance are not really new. The anisotropic magneto-resistance (AMR) effect, for example, was used in the magnetic read heads of earlier generations. Meanwhile, however, it has become possible to prepare nanostructures in which a spin polarization of the current adds a dramatic new functionality to devices and raises hopes for a whole family of novel applications, which range from spin-based field effect transistors (FETs), through permanent magnetic storage devices without moving parts like read/write heads, to quantum computation.

One reason for this new functionality is the fact that the spin allows one to establish polarization-based electronic schemes in addition to charge-based schemes, just like the polarization of light widens the field of optics dramatically. A good illustration of these possibilities and the underlying concepts is provided by the Datta–Das spin FET. A second reason is the increased scattering length: under a wide range of circumstances, the spin interacts only weakly with its environment. The majority of the electron scattering events are spin-conserving, and it can therefore be expected that spin is conserved over distances that are much larger than the elastic mean free path. This means that, in principle, spin is superior to charge in terms of coherent effects and for quantum computation applications.

A key parameter in spintronics is the *spin polarization* of the relevant quantity q , which could be the density of states or the current density, for example. It is defined as

$$P_q \equiv \frac{q_\uparrow - q_\downarrow}{q_\uparrow + q_\downarrow} \quad (12.1)$$

where \uparrow and \downarrow denote the majority and the minority spin, respectively. This definition makes a lot of sense: for $q_\downarrow = 0$, we have a polarization $P_q = 1$; while for $q_\downarrow = q_\uparrow$, we have $P_q = 0$.

12.1

Ferromagnetic sandwich structures

In this section, we will look at the properties of layered structures comprising two ferromagnets separated by a non-magnetic metal or insulator. Two milestones in the field of spintronics are the discovery of the *tunneling magneto-resistance* (TMR) and the *giant magneto-resistance* (GMR).

12.1.1

Tunneling magneto-resistance (TMR) and giant magneto-resistance (GMR)

Tunneling magneto-resistance The TMR effect refers to the resistance of a ferromagnet–insulator–ferromagnet (FIF) tunnel junction (see Fig. 12.1). The two ferromagnets are defined in such a way that their coercive magnetic fields differ by a significant amount, which is easily achieved by using different materials or layer thicknesses. In such a sample, the relative orientation of the magnetizations can be changed by sweeping a magnetic field aligned parallel to the layers. It was found that, for the magnetizations of the two layers aligned parallel to each other, the tunnel resistance is lower than for antiparallel alignment. The model of Jullière, who discovered the TMR effect in 1975, provides a simple explanation [171]: it is assumed that the spin does not flip during tunneling. Furthermore, the tunneling rate for each spin direction is proportional to the product of the corresponding densities of states in the two ferromagnets. The total resistance can therefore be thought of as two spin-resolved tunneling resistances in parallel.

We define conductances G_p and G_{ap} for the parallel and antiparallel configurations. The densities of states for the source and drain ferromagnetic electrodes are labeled $D_{\uparrow(\downarrow)S}$ and $D_{\uparrow(\downarrow)D}$. Within the model, we then have

$$\begin{aligned} G_p &\propto D_{\uparrow S}D_{\uparrow D} + D_{\downarrow S}D_{\downarrow D} \\ G_{ap} &\propto D_{\uparrow S}D_{\downarrow D} + D_{\downarrow S}D_{\uparrow D} \end{aligned} \quad (12.2)$$

Since, in a ferromagnet, $D_{\uparrow} > D_{\downarrow}$, we have $G_p > G_{ap}$. The tunneling magneto-resistance (TMR) is usually defined as

$$\text{TMR} \equiv \frac{R_{ap} - R_p}{R_{ap} + R_p}$$

where R_j denotes the resistance of the corresponding configuration. Inserting the conductances gives

$$\text{TMR} = \frac{2P_S P_D}{1 - P_S P_D} \quad (12.3)$$

where P denotes the polarization of the density of states. In cobalt, for example, $P_{Co} = 0.34$, which gives a TMR of 0.26 in an ideal system.

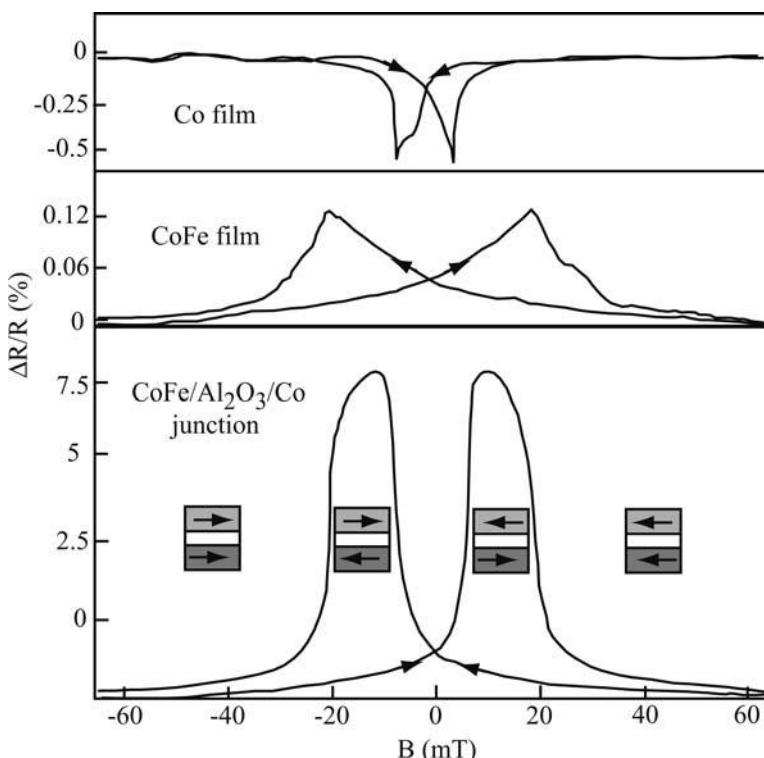


Fig. 12.1 Bottom: Magneto-resistance of a TMR structure formed by a $\text{CoFe}-\text{Al}_2\text{O}_3-\text{Co}$ sandwich, with the magnetization directions of the two films as indicated in the schemes. Also shown in the top two traces are the much weaker anisotropic magneto-resistances of the corresponding individual films. Adapted from [216].

One application of TMR is in memory chips. In a *magnetic random access memory* (MRAM) chip, each bit is stored in a small column of a TMR layer sequence. The lower ferromagnetic layer is *hard*, i.e. not reversible by the magnetic fields acting on the layers. This can be achieved by taking advantage of the exchange bias effect [208, 222], which is the displacement of the hysteresis trace of a ferromagnet along the magnetic field axis when the ferromagnetic layer is sitting on top of an antiferromagnetic film. The TMR column is connected to two wires on its top and bottom. The two states of the bit correspond to the two values of the measured current. Hence, this memory can be read out without a magnetic read head. Rather, square arrays of TMR columns are contacted by one-dimensional arrays of wires on the top and on the bottom, which are rotated by 90° with respect to each other. In this way, each element of the array can be addressed individually (see Fig. 12.2). The writing, which means defining the orientation of the top magnetic layer, can be done by cur-

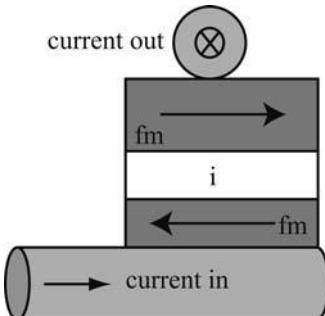


Fig. 12.2 Principle of an MRAM device. Two ferromagnetic layers with different coercive magnetic fields, separated by a tunnel barrier, are electrically accessed by two crossed wires. The resistance (high or low) across the TMR structure is attributed to the stored bit. The bit can be written by a current pulse in the wires that magnetizes the top layer while leaving the bottom layer in its initial state.

rent pulses. The currents have to be chosen such that the individual magnetic fields generated (Oerstedt fields) are insufficient to orientate the magnetization, but their superposition *is* sufficient. At very small column cross sections, it is actually not the Oerstedt field that switches the magnetization, but the torque the conduction electrons exert on the layer [29].

The advantage of this MRAM scheme compared to present memory devices is twofold. First, the delicate read/write head in hard disks becomes obsolete, and no moving parts are required. Second, in comparison to static random access memory (SRAM), the stored data does not get lost when the power is turned off. At present, though, it looks like, despite the beauty of the MRAM concept, it will not become the storage device of the next generation because of limitations regarding the storage density. MRAMs may, however, become important for niche applications.

Giant magneto-resistance The GMR effect, discovered by Grünberg et al. [128] and by Baibich et al. [19], is similar to the TMR effect, although its physical origin is somewhat different: replace the insulator in a TMR layer sequence by a normal metal, and you have a GMR structure. Again, a resistance is observed that depends on the relative orientation of the magnetizations of the ferromagnetic layers (see Fig. 12.3). It has been established that this dependence originates mostly from spin-dependent transmission of the conduction electrons across the ferromagnet–normal metal interfaces, the origin of which we will discuss in the next section. Assuming this as a fact for the moment, we can interpret such a GMR structure qualitatively in the two-current model [91] (Fig. 12.4). If both ferromagnetic layers are magnetized parallel to each other, one spin channel sees two low interface resistances in series, and the second

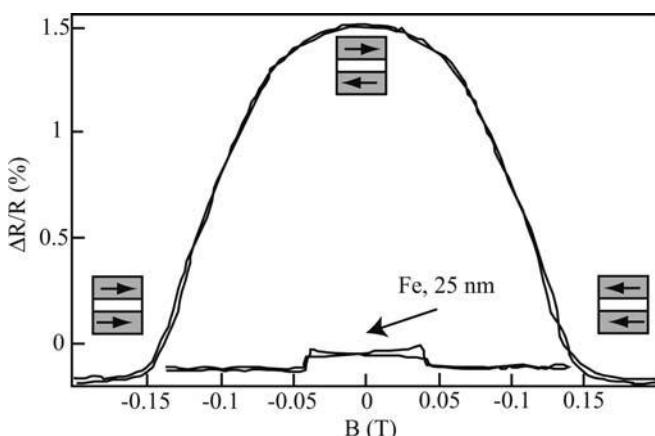


Fig. 12.3 The GMR effect as observed on a Fe–Cr–Fe sandwich structure. The anisotropic magneto-resistance of a ferromagnetic thin film is shown in comparison. Adapted from [35].

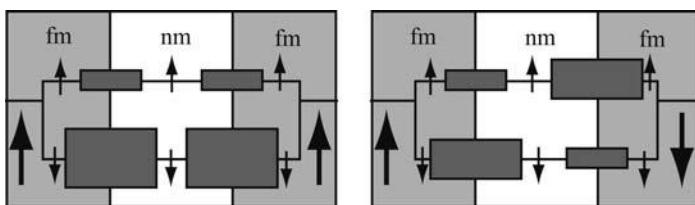


Fig. 12.4 Mott's two-current model.

one sees two large resistances. In the antiparallel configuration, the resistances for the two spin channels are equal. Within this model, the two spin channels can be thought of as being parallel and, if spin-flip processes are neglected, independent of each other. We see that, in the parallel configuration, the highly transmissive spin channel dominates the resistance; whereas, in the antiparallel configuration, both spin channels contribute equally to the resistance. This results in a lower overall resistance for the parallel configuration.

In contrast to the TMR effect, the GMR effect can also be observed when the current flows parallel to the layers. One speaks of the *current perpendicular to plane* (CPP) and the *current in plane* (CIP) configurations. Also, in a CIP setup, the electrons pass through the layers with a certain probability, thereby probing the magnetization configuration. The effect is, however, much weaker. This is nevertheless an important configuration, since using the GMR effect for applications in a CPP structure is tricky because of the very small overall resistance.

The GMR effect is used presently in the read/write heads of magnetic hard disks, where it serves to enhance the sensitivity of the read part (compared to

the anisotropic magneto-resistance effect used previously) to the magnetization of the bits on the disk.

An important point in the interpretation of the GMR effect is the origin of the interface resistance. Since this issue is also highly relevant for spin injection into semiconductors, we discuss it in some detail.

12.1.2

Spin injection into a non-magnetic conductor

In order to inject a spin-polarized current into a conductor, it is self-evident to use a ferromagnetic contact. Consider an interface between a ferromagnet (F) and a normal conductor (N). The current density (flowing parallel to the x -axis) across the interface is composed of two spin components, $j = j_{\uparrow} + j_{\downarrow}$. The spin current density is given by $j_s = j_{\uparrow} - j_{\downarrow}$. Here, we denote by \uparrow the majority spin in the ferromagnet. Please note that, in a normal conductor, there is no spin current associated with a charge current.

In the normal conductor, the conductivities can be assumed to be spin-independent, i.e.

$$\sigma_{N\uparrow} = \sigma_{N\downarrow} = \frac{1}{2}\sigma_N \quad (12.4)$$

which means that the current density spin polarization in the normal conductor far away from the interface is zero. However, close to the interface, the spin currents may be different and we write

$$\begin{aligned} j_{N\uparrow} &= -\frac{\sigma_N}{2} \frac{1}{e} \frac{\partial \mu_{N\uparrow}}{\partial x} \equiv \beta_N j_N \\ j_{N\downarrow} &= -\frac{\sigma_N}{2} \frac{1}{e} \frac{\partial \mu_{N\downarrow}}{\partial x} \equiv (1 - \beta_N) j_N \end{aligned} \quad (12.5)$$

Here, β_N is the fraction of the current density carried by the spin-up channel in the normal metal.

In the ferromagnet, the two spin directions experience different conductivities:

$$\begin{aligned} \sigma_{F\uparrow} &= e^2 D_{F\uparrow}(E_F) D_{F\uparrow} \equiv \alpha_F \sigma_F \\ \sigma_{F\downarrow} &= e^2 D_{F\downarrow}(E_F) D_{F\downarrow} \equiv (1 - \alpha_F) \sigma_F \end{aligned} \quad (12.6)$$

where $D_{F\uparrow(\downarrow)}(E_F)$ and $D_{F\uparrow(\downarrow)}$ denote the spin-resolved densities of states at the Fermi energy and the diffusion constants, respectively, while α_F represents the fraction of the total ferromagnet conductance that is contributed by the spin-up channel. Note from Eq. (12.4) that $\alpha_N = 1/2$. Correspondingly, the current

is split among the spin channels according to

$$\begin{aligned} j_{F\uparrow} &= -\sigma_{F\uparrow} \frac{1}{e} \frac{\partial \mu_{F\uparrow}}{\partial x} \equiv \beta_F j_F \\ j_{F\downarrow} &= -\sigma_{F\downarrow} \frac{1}{e} \frac{\partial \mu_{F\downarrow}}{\partial x} \equiv (1 - \beta_F) j_F \end{aligned} \quad (12.7)$$

where β_F is the fraction of the current density carried by the spin-up channel in the ferromagnet.

The following derivation of the interface resistance is based upon [283]. A spin-polarized current density arrives from the ferromagnet at the interface. In the normal metal, however, both spin channels have equal conductances, to a very good approximation. It is thus clear that the spin accumulates close to the interface, and a spin density gradient builds up in *both* materials. This means that, in the interface region, the two spin directions have different electrochemical potentials. We first calculate this difference $\mu_{SF} \equiv \mu_{F\uparrow} - \mu_{F\downarrow}$ in the ferromagnetic part, where it is given by

$$\frac{\partial \mu_{SF}}{\partial x} = -\frac{e}{\sigma_{F\uparrow}} j_{F\uparrow} + \frac{e}{\sigma_{F\downarrow}} j_{F\downarrow} \quad (12.8)$$

The continuity equation, on the other hand, relates the current densities to the spin-flip scattering times according to

$$\begin{aligned} \frac{\partial j_{F\uparrow}}{\partial x} &= e \left(\frac{n_{F\downarrow}}{T_{\uparrow\downarrow}} - \frac{n_{F\uparrow}}{T_{\downarrow\uparrow}} \right) \\ \frac{\partial j_{F\downarrow}}{\partial x} &= e \left(\frac{n_{F\uparrow}}{T_{\downarrow\uparrow}} - \frac{n_{F\downarrow}}{T_{\uparrow\downarrow}} \right) \end{aligned} \quad (12.9)$$

Here, $n_{F\uparrow(\downarrow)}$ denotes the spin-resolved electron densities; and $T_{\uparrow\downarrow}$ ($T_{\downarrow\uparrow}$) are the spin-flip scattering times from \uparrow into \downarrow (respectively, from \downarrow into \uparrow).

Inserting Eqs. (12.9) in the spatial derivative of Eq. (12.8) results in

$$\frac{\partial^2 \mu_{SF}}{\partial x^2} = \mu_{SF} \left(\frac{D_{F\uparrow} T_{\downarrow\uparrow} + D_{F\downarrow} T_{\uparrow\downarrow}}{D_{F\uparrow} T_{\uparrow\downarrow} D_{F\downarrow} T_{\downarrow\uparrow}} \right) \quad (12.10)$$

The law of detailed balance requires that

$$T_{\uparrow\downarrow} = \frac{D_{F\uparrow}(E_F)}{D_{F\downarrow}(E_F) T_{\downarrow\uparrow}}$$

With the spin relaxation time T_{1F} for the ferromagnet of

$$T_{1F} \equiv \frac{T_{\uparrow\downarrow} T_{\downarrow\uparrow}}{T_{\uparrow\downarrow} + T_{\downarrow\uparrow}}$$

we find the diffusion equation

$$\frac{\partial^2 \mu_{\text{SF}}}{\partial x^2} = \frac{1}{D_{\text{F}}^{\text{eff}} T_{1\text{F}}} \mu_{\text{SF}} \quad (12.11)$$

with the effective diffusion constant for the ferromagnet

$$D_{\text{F}}^{\text{eff}} \equiv \frac{D_{\text{F}\uparrow} D_{\text{F}\downarrow} [\mathcal{D}_{\text{F}\uparrow}(E_{\text{F}}) + \mathcal{D}_{\text{F}\downarrow}(E_{\text{F}})]}{D_{\text{F}\uparrow} \mathcal{D}_{\text{F}\uparrow}(E_{\text{F}}) + \mathcal{D}_{\text{F}\downarrow}(E_{\text{F}}) D_{\text{F}\downarrow}} = \alpha_{\text{F}} D_{\text{F}\uparrow} + (1 - \alpha_{\text{F}}) D_{\text{F}\downarrow} \quad (12.12)$$

Of course, the same type of diffusion equation holds for the normal metal, where the effective density of states is just D_{N} . The solution is

$$\mu_{\text{SF}}(x) = \begin{cases} \mu_{\text{SF}}(0) e^{x\lambda_{\text{F}}} & x \leq 0 \\ \mu_{\text{SN}}(0) e^{-x\lambda_{\text{N}}} & x \geq 0 \end{cases} \quad (12.13)$$

where $\lambda_{\text{F(N)}}$ is the spin relaxation length in the corresponding material, given respectively by

$$\lambda_{\text{F}} = \sqrt{D_{\text{F}}^{\text{eff}} T_{1\text{F}}}, \quad \lambda_{\text{N}} = \sqrt{D_{\text{N}} T_{1\text{N}}}$$

We are now in a position to obtain expressions for $P_{j\text{N}}$ as a function of $P_{j\text{F}}$ and for the emerging interface resistance. As far as the boundary conditions are concerned, first of all, the total current density has to be constant. We also assume that there is neither enhanced spin-flip scattering nor a contact resistance at the interface. In that case, β must be continuous at $x = 0$. Also, the spin-resolved electrochemical potentials must be continuous, and $\mu_{\text{SF}}(0) = \mu_{\text{SN}}(0)$. From Eq. (12.8), we find for the current density at the ferromagnetic side

$$\frac{\partial \mu_{\text{SF}}}{\partial x}(0) = \frac{\mu_{\text{SF}}(0)}{\lambda_{\text{F}}} = -\frac{ej_{\uparrow}}{\sigma_{\uparrow}} + \frac{ej_{\downarrow}}{\sigma_{\downarrow}} = \frac{ej}{\sigma_{\uparrow}\sigma_{\downarrow}} [\sigma_{\uparrow} - (\sigma_{\uparrow} + \sigma_{\downarrow})\beta(0)] \quad (12.14)$$

For the normal metallic side, we find accordingly

$$\frac{\partial \mu_{\text{SN}}}{\partial x}(0) = \frac{\mu_{\text{SN}}(0)}{\lambda_{\text{N}}} = \frac{2ej}{\sigma_{\text{N}}} [1 - 2\beta(0)] \quad (12.15)$$

We solve both equations for the current density and set them equal. We obtain an equation for $\beta(0)$, which we express as

$$2\beta(0) - 1 = \frac{2\alpha_{\text{F}} - 1}{1 + 4\alpha_{\text{F}}(1 - \alpha_{\text{F}})(\lambda_{\text{N}}/\sigma_{\text{N}})(\sigma_{\text{F}}/\lambda_{\text{F}})} \quad (12.16)$$

The corresponding spin-resolved electrochemical potentials close to the interface are depicted in Fig. 12.5.

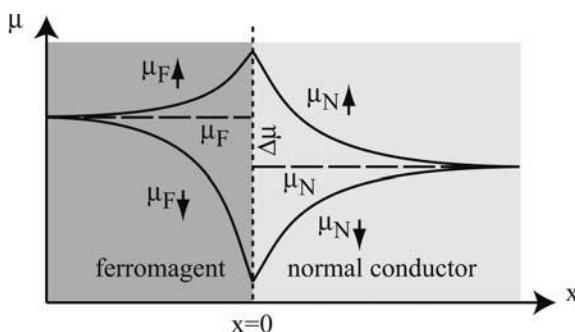


Fig. 12.5 Schematic spin accumulation at the ferromagnet–normal metal interface, expressed in terms of the spin-resolved electrochemical potentials. The step of the averaged electrochemical potentials across the interface is denoted by $\Delta\mu$. Adapted from [283].

Please note that $[2\beta(0) - 1]$ equals P_{jN} , the current spin polarization induced in the normal metal. Note further that $(2\alpha_F - 1)$ is the current spin polarization in the ferromagnet, and we can write

$$P_{jN}(x=0) = \frac{P_{jF}}{1 + (1 - P_{jF}^2)(\lambda_N/\sigma_N)(\sigma_F/\lambda_F)} \quad (12.17)$$

This equation tells us the extent to which the spin polarization in the ferromagnet can be transferred into the normal conductor. Of course, $P_{jN}(x=0) \leq P_{jF}$. It is easily seen that, for a ferromagnet with $P_{jF} = 1$, the spin polarization is perfectly injected into the normal metal. Conventional ferromagnets, however, have a spin polarization significantly smaller than 1, like the value of cobalt mentioned above. We can still aim for a large spin polarization in the normal conductor by making the second term in the denominator as small as possible, in other words by choosing materials with $\lambda_N/\sigma_N \ll \lambda_F/\sigma_F$.

This is not particularly difficult for many conventional metals, as the observation of the GMR effect proves. Moreover, numbers for the quantities of interest have been obtained in various experiments. In [168], for example, a spin-polarized current is injected from permalloy ($\text{Ni}_{80}\text{Fe}_{20}$) into copper, and a value of $P_{j,Cu} = 0.02$ was extracted. A spin-flip time $T_{1,Cu}$ of 42 ps was found at 4.2 K, which corresponds to $\lambda_{Cu} \approx 1 \mu\text{m}$. Since the Drude scattering time in copper is 30 fs at 4.2 K, these results show that the electrons experience 1000 elastic scattering events on average before they experience a spin flip. Even at room temperature, a relatively large value of $\lambda_{Cu} \approx 350 \text{ nm}$ remains.

Deriving the interface resistance R_i from Eq. (12.17) is the topic of Exercise E12.1.

12.2

The Datta–Das spin field effect transistor

In the following, we treat the Datta–Das proposal [66] of a spin field effect transistor (spin FET) as a paradigm and use it as a motivation to discuss various aspects of spintronics.

12.2.1

Concept of the Datta–Das transistor

In a way, the spin FET is a modification of a GMR structure. Suppose we replace the metal of such a structure by a semiconductor with all its advantages, in particular the tunability of the band bending in the conduction channel by a gate voltage (see Fig. 12.6).

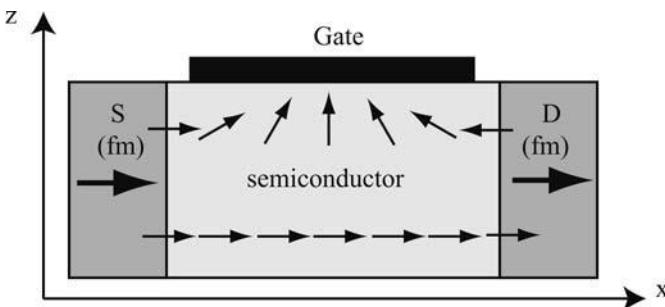


Fig. 12.6 The scheme for a spin field effect transistor. The evolution of the spin in the semiconductor is denoted for two gate voltages by the two sequences of arrows.

The suggestion is based on a close analogy to an electro-optical polarization rotator for linearly polarized light, also known as a Pockels cell [146]: spin-polarized electrons are injected into a semiconductor, a gate is used to rotate the spin direction, and a ferromagnetic drain contact acts as analyzer. This kind of transistor can be switched by rotating the electron spin by 180° . It is thus not necessary to remove the electrons from the conducting channel as in a conventional transistor (see Chapter 3) and it is expected to consume less power.

Up to now, a spin FET has not been realized experimentally. In order to understand the details as well as the problems that have to be solved for its experimental realization, we are going to elaborate the elements of the spin FET. The major issues are injection of a spin-polarized current into the semiconductor, and rotation of the electron spin by a gate voltage.

12.2.2

Spin injection in semiconductors

Equation (12.17) indicates that injecting a spin-polarized current into a semiconductor is not trivial. Typically, not only is the conductivity of the semiconductor much lower than that of the ferromagnet, but also the spin-flip length is much longer. Therefore, low current spin polarizations will be the result. Two approaches have been pursued in order to overcome this difficulty. The idea of the first approach is immediately clear from Eq. (12.17): a ferromagnet with a spin polarization of $P_{fF} = 1$ would generate an equally perfect value for P_{fN} . Such a large spin polarization is not available in conventional ferromagnets, but can be achieved in ferromagnetic semiconductors. Inserting tunnel barriers at the ferromagnet–semiconductor interface is the second approach. In this way, a spin-selective interface resistance is generated which can increase the injected P_{fN} dramatically.

12.2.2.1 Interface tunnel barriers

In the derivation of Eq. (12.17), we have neglected possible influences of interface properties. The interface resistance found stems solely from the differences in the bulk parameters of the two materials. This is not always justified. Instead, interface roughness, interface states or other parameters may generate an additional interface resistance $R_{i\uparrow(\downarrow)}$, which is localized at the interface and may depend upon the spin. We will not discuss the details of these considerations here; the interested reader is referred to [249]. Qualitatively, such an interface resistance makes the spin-resolved electrochemical potentials discontinuous at the interface, i.e.

$$\mu_{F\uparrow(\downarrow)}(x \rightarrow 0) \neq \mu_{N\uparrow(\downarrow)}(x \rightarrow 0)$$

and the spin components of the current densities across the interface are given by

$$j_{\uparrow(\downarrow)} = \frac{1}{eR_{i\uparrow(\downarrow)}} [\mu_{F\uparrow(\downarrow)}(x \rightarrow 0) - \mu_{N\uparrow(\downarrow)}(x \rightarrow 0)] \quad (12.18)$$

Inclusion of a tunnel barrier, which we consider a part of the interface, can increase the difference between j_\uparrow and j_\downarrow : owing to the difference in their electrochemical potentials, the barrier height and thus the transmission probability are not the same for the two spin directions. As a consequence, the injected spin polarization increases.

But how can an injected spin-polarized current be detected? A frequently used technique is based upon conversion of the spin polarization into circular polarization of photons. In close proximity to the spin injector, a LED-type p–n junction can be defined, for example. At the junction, photons are emitted as a

consequence of electron–hole recombination. The dipole selection rules allow only transitions between electron and hole states that emit photons of left or right circular polarization, with a weight given by the corresponding dipole matrix elements (see Fig. 12.7). Light emitted from the $m_s = -1/2$ ($+1/2$) electronic state in bulk GaAs, for example, gives a σ^+ (σ^-) polarization of 0.5.

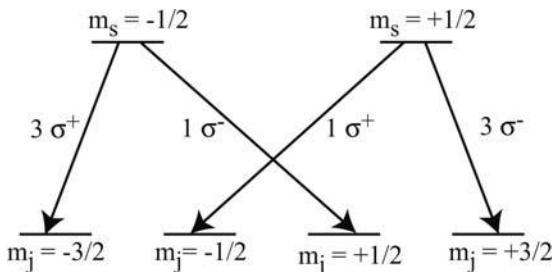


Fig. 12.7 Selection rules for photon emission by electron–hole recombination in GaAs. The relative intensities and the orientation of the circular polarization are indicated at each transition.

Question 12.1: Show that the polarization of the emitted light intensity I , defined as

$$P_\sigma \equiv \frac{I_{\sigma^+} - I_{\sigma^-}}{I_{\sigma^+} + I_{\sigma^-}}$$

depends on the injected current spin polarization according to

$$P_\sigma = -\frac{1}{2}P_j \quad (12.19)$$

The experiment performed by Hanbicki and coworkers [139] is an elegant proof that this technique actually works: the spin current is injected via an $\text{Fe}-\text{Al}_x\text{Ga}_{1-x}\text{As}$ heterolayer. The spin-dependent interface resistance at the interface is provided by the Schottky barrier. In close proximity to this interface, a GaAs quantum well (instead of a p–n junction) is embedded in the $\text{Al}_x\text{Ga}_{1-x}\text{As}$. Here, electron–hole recombination takes place, and the circularly polarized emitted light reflects the spin polarization of the injected current.¹ The authors found a value of $P_j = 0.13$ at the quantum well at a temperature of 4 K, which decayed to 0.04 at 240 K. Taking into account the (temperature-dependent) spin dephasing that occurs between the interface and the quantum well, the injected polarization at the interface was estimated to be as high as $P_j = 0.3$ and independent of the temperature.

1) Note that the selection rules in quantum wells are modified as compared to Fig. 12.7, which results in $P_\sigma = -\frac{1}{2}P_j$.

12.2.2.2 Ferromagnetic semiconductors

The combination of the special properties of semiconductors (like controlling the transport properties by doping or by external parameters) with those of ferromagnets (storage capabilities due to bistable magnetization traces, spin-polarized currents) is a particularly exciting field. Just imagine what kind of devices could be realized. Data could be stored in a non-volatile way on the monolithic processor, with the (maybe TMR-based) storage array made from the same semiconducting material as the processor circuit. If the onset of ferromagnetism depends on the electron density, it could be turned on and off by a gate voltage.

Moreover, and most relevant to our impedance mismatch problem, the Fermi energy in semiconductors is small compared to that in metals, and can become smaller than the spin splitting of the conduction band. This results in a spin polarization of 1. Consequently, the impedance mismatch problem would not occur.

Meanwhile, a variety of ferromagnetic semiconductor systems have in fact been discovered. The most studied one [225] is certainly $\text{Ga}_{1-x}\text{Mn}_x\text{As}$. Here, Mn atoms replace Ga atoms on their lattice sites [274] and act as a deep acceptor [227]. The Mn ions interact with each other and order themselves ferromagnetically, while the interaction is mediated by the conduction holes. The interaction strength, and thereby the Curie temperature, is consequently a function of the hole density [71]. In Fig. 12.8, an experiment is reproduced in which the remanence and the coercive magnetic field could in fact be tuned by a gate voltage.

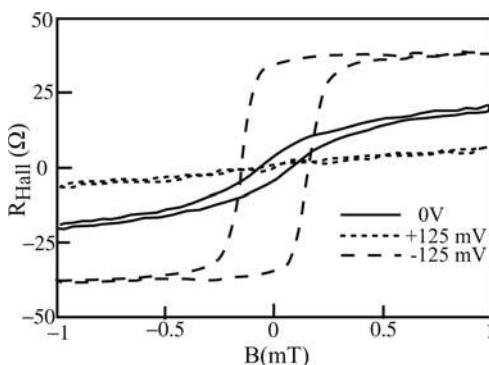


Fig. 12.8 Magnetization traces of $\text{Ga}_{1-x}\text{Mn}_x\text{As}$ as a function of the gate voltage. Adapted from [226].

As expected, a significant degree of spin polarization can be achieved as electrons are injected from a ferromagnetic semiconductor into a non-magnetic one. This has been demonstrated by the electron spin to photon polarization conversion scheme outlined above. In the structure shown

in Fig. 12.9, for example, a spin polarization of the current injected from $\text{Ga}_{1-x}\text{Mn}_x\text{As}$ into GaAs of $P_j = 0.82$ at liquid helium temperatures has been found.

It can thus be concluded that the impedance mismatch problem has been solved, at least conceptually.

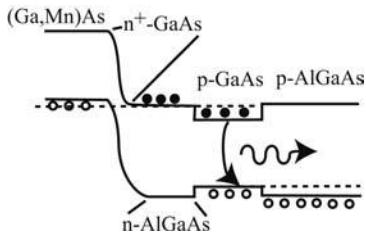


Fig. 12.9 Spin injection from a ferromagnetic semiconductor into a GaAs quantum well. A significant circular polarization of the emitted light is detected. After [76].

12.2.3

Gate-induced spin rotation: The Rashba effect

The Rashba effect denotes the spin-orbit coupling experienced by moving electrons in electric fields. The Rashba Hamiltonian can be written as

$$H = \frac{\vec{p}^2}{2m^*} - \frac{\eta}{\hbar} \vec{\sigma} \cdot (\vec{e}_z \times \vec{p}) \quad (12.20)$$

where $\vec{e}_z = (0, 0, 1)$, $\vec{p} = (p_x, p_y, 0)$, and $\vec{\sigma}$ is the Pauli spin matrix vector. This gives

$$H = \begin{pmatrix} \frac{\vec{p}^2}{2m} & \frac{\eta}{\hbar}(ip_x + p_y) \\ \frac{\eta}{\hbar}(-ip_x + p_y) & \frac{\vec{p}^2}{2m} \end{pmatrix} \quad (12.21)$$

Since for plane electron waves, the spinor is of the form

$$\Psi = e^{i\vec{k}\vec{r}} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = e^{i\vec{k}\vec{r}}(\alpha|\uparrow\rangle + \beta|\downarrow\rangle) \quad (12.22)$$

we can replace p_j with $\hbar k_j$. The problem has cylindrical symmetry, and we can express the Hamiltonian in cylindrical coordinates:

$$H = \begin{pmatrix} \frac{\hbar^2 k^2}{2m} & i\eta k e^{-i\phi} \\ -i\eta k e^{i\phi} & \frac{\hbar^2 k^2}{2m} \end{pmatrix} \quad (12.23)$$

From the characteristic polynomial, we find the eigenvalues

$$\lambda_{\pm} = \frac{\hbar^2 k^2}{2m} \pm \eta k \quad (12.24)$$

This energy dispersion is represented graphically in Fig. 12.10. The corresponding eigenspinor components are

$$\begin{aligned}\lambda_+ : \beta &= -ie^{i\phi}\alpha = e^{i(\phi-\pi/2)}\alpha \\ \lambda_- : \beta &= ie^{i\phi}\alpha = e^{i(\phi+\pi/2)}\alpha\end{aligned} \quad (12.25)$$

From the normalization condition and by choosing the global phase factor such that α is real,² we find the eigenspinors (we omit the real-space plane wave in the following)

$$\begin{aligned}|\lambda_+\rangle &= \frac{1}{\sqrt{2}}|\uparrow\rangle + \frac{1}{\sqrt{2}}e^{i(\phi-\pi/2)}|\downarrow\rangle \\ |\lambda_-\rangle &= \frac{1}{\sqrt{2}}|\uparrow\rangle + \frac{1}{\sqrt{2}}e^{i(\phi+\pi/2)}|\downarrow\rangle\end{aligned} \quad (12.26)$$

These states lie on the equator of the Bloch sphere and are rotated with respect to \vec{k} by $+90^\circ$ for λ_+ and by -90° for λ_- , respectively.

We calculate the momentum vectors at the Fermi energy in a given direction from

$$E_F = \lambda_+ k_{F,+} = \lambda_- k_{F,-}$$

and get

$$\Delta k_F \equiv k_{F,+} - k_{F,-} = \frac{2\eta m^*}{\hbar^2} \quad (12.27)$$

Note that Δk_F does not depend on the energy for $E > 0$. For a Fermi energy of $E_F = 10$ meV and a spin-orbit coupling constant $\eta = 10^{-11}$ eV m, we find

$$\Delta k_F = 1.7 \times 10^7 \text{ m}^{-1} \approx 0.13k_F$$

The energy splitting due to the Rashba term has been observed in several transport experiments, like the one shown in Fig. 12.11. In magneto-transport, the splitting manifests itself in a modulation of the Shubnikov–de Haas oscillation, very similar to a quasi-2DEG with two subbands occupied, as discussed in Section 6.4.1.

The superposition of the two spins propagating with different wave vectors is obtained from

$$|\psi_{\text{spin}}\rangle = \frac{1}{\sqrt{2}}(|\lambda_+\rangle + |\lambda_-\rangle) \quad (12.28)$$

2) In case you wonder why, please go through Exercise E10.3.

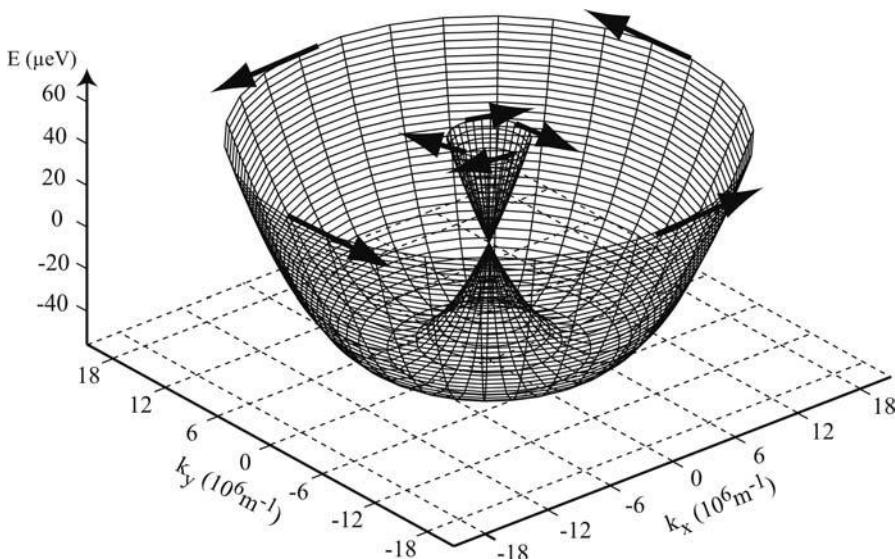


Fig. 12.10 The energy dispersion of a free 2DEG in an $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}-\text{In}_{0.52}\text{Al}_{0.48}\text{As}$ quantum well as used in [221], i.e. with $\eta = 10^{-11} \text{ eV m}$ and $m^* \approx 0.05m_e$, with the Rashba effect taken into account. The arrows depict the corresponding spin directions. Note that the plot range is only a small fraction of a typical Fermi wave vector.

Inserting the expressions and again multiplying the coefficient in front of $|\uparrow\rangle$ by a global phase factor to make it real, we obtain after some algebra

$$|\psi_{\text{spin}}\rangle = \cos(\frac{1}{2}\Delta k_F r)|\uparrow\rangle + \sin(\frac{1}{2}\Delta k_F r)e^{i\phi}|\downarrow\rangle \quad (12.29)$$

This is a rotation along a circle formed by the intersection of the Bloch sphere with the plane given by ϕ . In real space, the spinor rotates around the direction of the magnetic field seen by the electron, i.e. around the axis in the (x, y) plane that is perpendicular to \vec{k} .

For the above values, we find from $\Delta k_F L = 2\pi$ a rotation about π over a distance

$$L = \pi \frac{\hbar^2}{\eta m^*} = 350 \text{ nm}$$

The analysis above provides further insight into the design requirements for a Datta–Das transistor. The transistor only works when the spin directions are well defined, which means that the channel should be quasi-one-dimensional. Second, the channel should be ballistic. Even though elastic scattering does not flip the spin, it changes the continuous spin rotation abruptly.

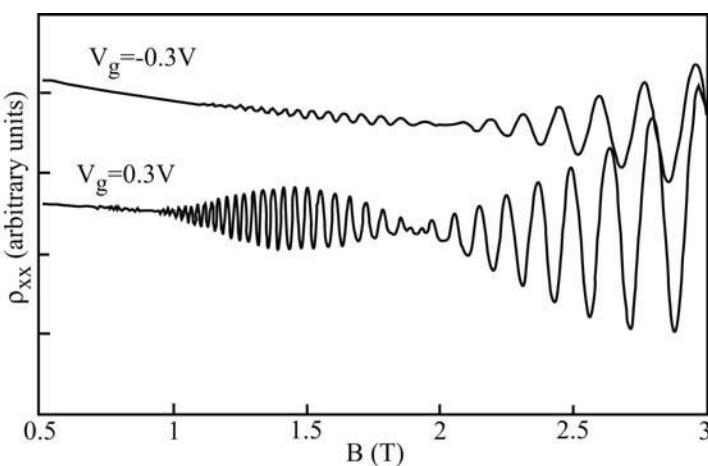


Fig. 12.11 Modulation of Shubnikov–de Haas oscillations due to the Rashba effect. Adapted from [221].

12.2.4

Spin relaxation and spin dephasing

It is clear that the spin relaxation length must be larger than the channel length in a Datta–Das transistor. But how does a spin relax? Here, one has to distinguish between two different time scales. First of all, there is a characteristic time over which a spin-up state experiences a spin-flip scattering event and ends up in a spin-down state (or vice versa). This is the spin relaxation time T_1 encountered already in the discussion of spin injection. A second mechanism is spin dephasing, which refers to the loss of the correlation of the spin precession around the quantization axis. The spins precess with the Larmor frequency $\omega_L = eB/2m^*$ around the effective magnetic field. As the electrons move, they may experience a varying magnetic field and thus their precessional motion loses its coherence. This happens during the dephasing time T_2 . This time is also frequently described as the time over which the superposition of two states, a situation encountered frequently in quantum computational schemes, decays into a pure state.

The following mechanisms can, among more exotic ones, generate spin relaxation and spin dephasing. The *Elliot–Yafet* (EY) mechanism was first pointed out in [84] and [338]. In a system with significant spin–orbit interactions, the conventional Bloch states are not eigenstates of the crystal Hamiltonian. Rather, the eigenstates can be expressed as a linear combination of the outer products of two Bloch wave functions with the two spin eigenstates. As a consequence, even spin-independent interactions can induce transitions between these eigenstates and thereby generate spin dephasing. This type of spin relaxation increases both with the spin–orbit coupling as well as with the

electron scattering rate. In GaAs, for example, Elliot–Yafet spin relaxation is very strong for holes due to the large spin–orbit coupling. The relation between the spin relaxation time τ_s according to the Elliot–Yafet mechanism and the other material parameters has been derived in [243] as

$$\tau_{s,EY} \approx \left(\frac{E_g + \Delta_{SO}}{\Delta_{SO}} \right)^2 \frac{E_g}{E} \tau \quad (12.30)$$

where E is the energy of the electron and Δ_{SO} is the spin–orbit splitting.

The Dyakonov–Perel (DP) mechanism emerges from the fact that, as we have seen in Section 2.2, the spin degeneracy is lifted for $k \neq 0$ in crystals without inversion symmetry, due to the Dresselhaus term. Electrons with identical wave vectors but different spins therefore have different energies in general. Owing to this energy difference, the electrons precess with different frequencies, very similar to the precession caused by the Rashba term. Scattering changes the wave vectors and with it the precession frequencies. In contrast to the Elliot–Yafet mechanism, however, the dephasing occurs not during the scattering but during the electron motion in between scattering events. It turns out that, because of this, the spin relaxation time is inversely proportional to the Drude scattering time [284],

$$\tau_{s,DP} \propto \frac{(k_B T)^3}{\hbar^2 E_g} \frac{1}{\tau} \quad (12.31)$$

This kind of spin relaxation was first discussed in [80]. Note that, by studying the spin relaxation time as a function of the mobility, one can easily distinguish between Elliot–Yafet and Dyakonov–Perel spin relaxation.

Hyperfine interactions can be an important source of spin relaxation as well. The spin-polarized electron gas interacts with the nuclear spins via

$$H \propto I \tilde{S}$$

thereby polarizing the nuclei while experiencing spin relaxation itself. It has been calculated in [95] that

$$\tau_{HF} \propto \sqrt{E_F}$$

which means that hyperfine interactions are particularly important at low carrier densities. Also, note that, in a confined system like a ballistic quantum dot, the first two mechanisms, which rely on extended motion of the electrons, should be of minor importance, while the hyperfine interaction remains relevant. This is an important factor and the reason why extremely long spin relaxation times can be observed in such systems.

Exercises

E12.1 Calculate the interface resistance

$$R_i = \frac{\mu_F(x \rightarrow 0) - \mu_N(x \rightarrow 0)}{ej}$$

at a ferromagnet–normal conductor junction due to spin accumulation.

E12.2 Represent the eigenspinors of Eq. (12.26) graphically on the Bloch sphere. Also indicate the motion of the spinor in Eq. (12.29).

Further Reading

A broad introduction to the field of spintronics is provided in [345]. An excellent overview of the emerging field of semiconductor spintronics can be obtained from the *Special Issue on Semiconductor Spintronics* of “Semiconductor Science and Technology” [285], as well as from [17].

This Page Intentionally Left Blank

A SI and CGS Units

Some people use the cgs (Gaussian) unit system, while others prefer the SI system (also known as the MKSA system). This results in both a constant source of irritation for students as well as an inconvenience for researchers. The following remarks should allow the reader to switch between them with confidence.

The cgs and SI systems originate in a different choice of units in equations containing electrodynamic quantities. Table A.1 lists how the prefactors must be replaced as the unit system is changed.

Tab. A.1 Prefactors in cgs and SI units.

Quantity	cgs	SI
speed of light	c	$1/\sqrt{\epsilon_0\mu_0}$
electric field	\vec{E}	$\sqrt{4\pi\epsilon_0}\vec{E}$
dielectric shift	\vec{D}	$\sqrt{4\pi/\epsilon_0}\vec{D}$
polarization	\vec{P}	$(1/\sqrt{4\pi\epsilon_0})\vec{P}$
magnetic field	\vec{B}	$\sqrt{4\pi/\mu_0}\vec{B}$
magnetizing field	\vec{H}	$\sqrt{4\pi\mu_0}\vec{H}$
magnetization	\vec{M}	$\sqrt{\mu_0/4\pi}\vec{M}$
dielectric constant	ϵ	ϵ/ϵ_0
permeability	μ	μ/μ_0
current	I	$(1/\sqrt{4\pi\epsilon_0})I$
resistance	R	$4\pi\epsilon_0 R$
inductance	L	$4\pi\epsilon_0 L$
capacitance	C	$[1/(4\pi\epsilon_0)]C$

Examples

- In cgs units, the generalized momentum is given by

$$\vec{p} + \frac{e}{c} \vec{A} = \vec{p} - \frac{e}{2c} \vec{r} \times \vec{H}$$

In order to switch to SI units, we make the following replacements:

$$e \longrightarrow e/\sqrt{4\pi\epsilon_0}, \quad \vec{H} \longrightarrow \sqrt{4\pi\mu_0} \vec{H}, \quad c \longrightarrow 1/\sqrt{\epsilon_0\mu_0}$$

such that the generalized momentum is

$$\vec{p} - \frac{e}{2} \vec{r} \times \mu_0 \vec{H}$$

- In cgs units, the Bohr magneton reads

$$\mu_B = \frac{e\hbar}{2mc}$$

This changes to

$$\mu_B = \frac{e\hbar}{2m}$$

on replacing the magnetization, the charge and the speed of light.

Occasionally, quantities have to be transformed as well. Table A.2 lists the most important transformation factors.

Tab. A.2 Numerical factors in cgs and SI units.

Quantity	cgs unit	SI unit
length	1 cm	0.01 m
weight	1 g	10^{-3} kg
force	1 dyn	10^{-5} N
energy	1 erg	10^{-7} J
charge	1 esE	$(1/3) \times 10^{-9}$ C
potential	1 statvolt	300 V
capacitance	1 cm	$(1/9) \times 10^{-11}$ F
magnetic flux	1 Mx	10^{-8} Wb
magnetic flux density	1 G	10^{-4} T
magnetic field	1 Oe	$(1/4\pi) \times 10^3$ A/m

Question A.1: Sodium has an electric polarizability of 0.4×10^{-24} cm³. Express this quantity in SI units.

B

Correlation and Convolution

B.1

Fourier Transformation

The Fourier transform of a function $f(x)$ is the continuous version of its expansion into a Fourier series, namely

$$F(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{i2\pi X x} dx \quad (\text{B.1})$$

and, respectively,

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(X) e^{-i2\pi X x} dX \quad (\text{B.2})$$

The units of the variables are inverse to each other. Fourier transformations are used frequently, owing to their efficiency and versatility in performing certain analytical tasks. Examples can be seen in Chapters 2, 8, and 14, as well as below.

B.2

Convolutions

The convolution of two functions $f(x)$ and $g(x)$ is defined as

$$h(x) = f * g(x) = \int_{-\infty}^{\infty} f(\xi) g(x - \xi) d\xi \quad (\text{B.3})$$

The effect of the convolution is to “smear out” $f(x)$ with $g(x)$. This is illustrated in Fig. B.1, where we convolute

$$f(x) = \theta(x - 1) - \theta(x - 2)$$

with

$$g(x) = \theta(x) - \theta(x - 1)$$

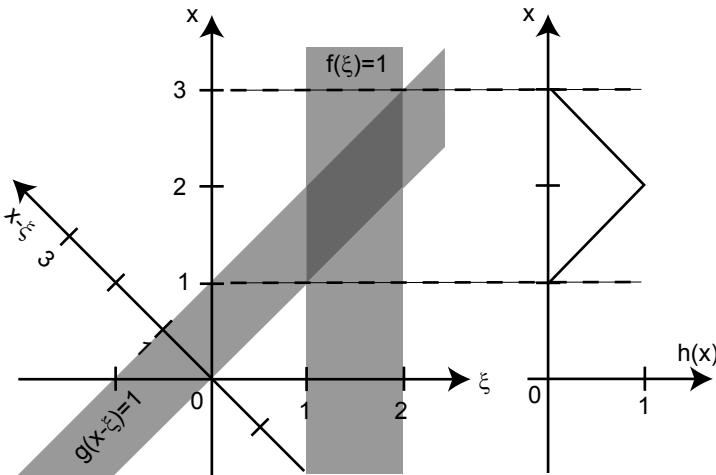


Fig. B.1 Graphical representation of a convolution. The functions $f(x) = \theta(x-1) - \theta(x-2)$ and $g(x-\xi) = \theta(x-\xi) - \theta(x-\xi-1)$ are drawn in the (x, ξ) plane in gray scale. Note that the $x - \xi$ axis is rotated with respect to the ξ -axis by 135° . For these functions, the convolution $h(x)$ is given by the extension of the overlapping area of $f(\xi)$ and $g(x-\xi)$ parallel to the x -axis, as sketched to the right.

Hence, $g(x-\xi) = \theta(x-\xi) - \theta(x-\xi-1)$. One finds for the convolution

$$h(x) = \begin{cases} 0 & x < 1 \\ x-1 & 1 \leq x < 2 \\ 3-x & 2 \leq x < 3 \\ 0 & x \geq 3 \end{cases}$$

The convolution theorem states that

$$H(X) = F(X)G(X) \quad (\text{B.4})$$

i.e. the Fourier transform of the convolved functions is the product of the Fourier transforms of the individual functions. This is useful for a process called *deconvolution*. Suppose we know that a signal, e.g. a QPC characteristic, is thermally smeared. We can then obtain the characteristic at $\Theta = 0$ by numerically Fourier transforming the measured data, dividing it by the Fourier transform of the derivative of the Fermi function, and transforming back the result.

B.3

Correlation Functions

In mathematical terms, the correlation function of the two functions $f(x)$ and $g(x)$ is defined as

$$C_{fg}(x) = \int_{-\infty}^{\infty} f(\xi)g(x + \xi) d\xi \quad (\text{B.5})$$

For $f = g$, we speak of the autocorrelation function $C_f(x)$. An example of an autocorrelation function is shown in Fig. B.2.

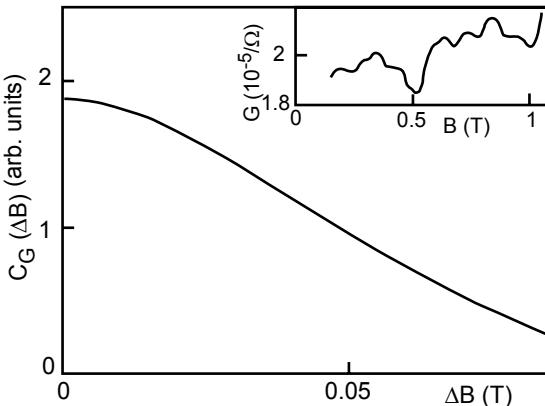


Fig. B.2 The autocorrelation function of the conductance $G(B)$ of a quantum wire as a function of a magnetic field (the raw data are shown in the inset). The shape is typical for autocorrelation functions of experimental parametric fluctuations. The autocorrelation field is $B_c \approx 50$ mT. After [25].

In mesoscopics, this notation is frequently used for the correlation function of the fluctuations around an average, i.e.

$$C_{fg}(x) = \langle \delta f(\xi) \delta g(\xi + x) \rangle \quad (\text{B.6})$$

with $\delta f(\xi) = f(\xi) - \langle f(\xi) \rangle$. The angle brackets denote the ensemble average. This means that

$$C_{fg}(x) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=1}^N \delta f_j(\xi) \delta g_j(\xi + x)$$

with j enumerating the N ensembles. Since, however, we assume that the system is ergodic, we can also average over the parameter. Assuming that the variable is continuous, the correlation function is then obtained by

$$C_{fg}(x) = \lim_{\xi_0 \rightarrow \infty} \frac{1}{\xi_0} \int_0^{\xi_0} \delta f(\xi) \delta g_j(\xi + x) d\xi \quad (\text{B.7})$$

This, by the way, also indicates how the correlation function is obtained for finite measured intervals of the parameter ξ , which can be a magnetic field, a gate voltage, or time, for example.

Effectively, $C_{fg}(x)$ compares f with g and measures the degree of similarity: g is shifted with respect to f along the x -axis, and the product function is integrated. Therefore, the autocorrelation function of a fluctuating function has a characteristic structure. For very small shifts, the original and the shifted function have approximately identical values at each x . Almost everywhere, both functions have the same sign. For large shifts, however, the signs of the two curves are no longer correlated, and the average area under the product function averages to zero. Consequently, $C_f(x)$ will drop to zero within the generalized correlation length x_c . Usually, it is defined as the value of x where $C_f(x)$ has dropped to $1/e$ ($e = 2.71828$) of $C_f(0)$, although sometimes different definitions are used, which, however, do not change the order of magnitude. Note that this continuous drop occurs only for random fluctuations. An oscillatory function, for example, also has an oscillatory autocorrelation function. Generally speaking, x_c becomes smaller as the bandwidth of the fluctuations increases. For $x = 0$, the autocorrelation function is simply the variance:

$$C_f(0) = \langle (\delta f(\xi))^2 \rangle \quad (\text{B.8})$$

The Wiener–Khintchine theorem states that the spectral power $S_f(X)$ is just twice the Fourier transform of the autocorrelation function of $f(x)$:

$$S_f(X) = 2C_f(X) \quad (\text{B.9})$$

Furthermore, the variance must be the spectral power, integrated over all X :

$$\langle (\delta f(\xi))^2 \rangle = \int_{X=0}^{\infty} S_f(X) \quad (\text{B.10})$$

C**Capacitance Matrix and Electrostatic Energy**

In this appendix, the electrostatic energy of a system of conductors is calculated. Consider a system of $n + m$ conductors, with n islands (floating conductors) and m electrodes (connected to voltage sources). See Fig. C.1.

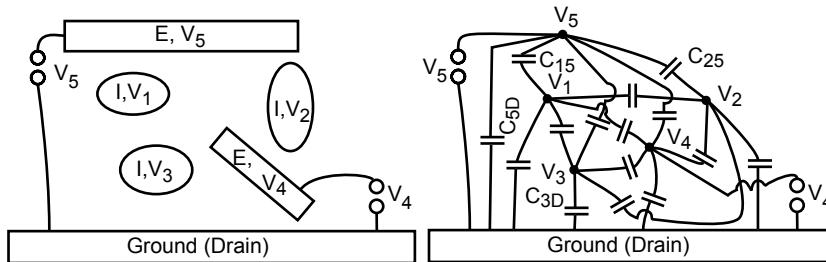


Fig. C.1 Left: A system of islands (floating) and electrodes (connected to voltage sources). Right: The corresponding equivalent circuit composed of potential nodes and mutual capacitances.

As in Chapter 12, the charge distribution is given by the charges at the electrodes, which can be written as a charge vector $\vec{q} = (\vec{q}_L, \vec{q}_E)$ composed of the island charge vector \vec{q}_I and the electrode charge vector \vec{q}_E . Similarly, a potential vector can be constructed: $\vec{V} = (\vec{V}_I, \vec{V}_E)$, with V_i being the potential of conductor i with respect to drain (ground). The charge at each conductor is given by the potentials of all conductors. This can be expressed by

$$q_i = \sum_{j=1}^{n+m} d_{ij} V_j \quad (C.1)$$

with coefficients d_{ij} determined by the electric field distribution [279].

We would like to express this relation in terms of capacitance coefficients, which express the effect of conductor j on the charge at conductor i as a function of the voltage between the two conductors:

$$q_i = \sum_{j=1}^{n+m} C_{ij}(V_i - V_j) + C_{iD}(V_i - V_D) \quad (C.2)$$

Since $V_D = 0$, we can rearrange the sum. In order to express all d_{ij} in terms of C_{ij} , we write Eq. (C.1) as

$$q_i = \sum_{j \neq i}^{n+m} -C_{ij}V_j + \left[\left(\sum_{j \neq i}^{n+m} C_{ij} \right) + C_{iD} \right] V_i$$

Thus, it is immediately obvious that $d_{ij} = -C_{ij}$ for $i \neq j$, and that

$$d_{ii} = C_{iD} + \sum_{k=1, k \neq i}^{n+m} C_{ik}$$

This allows us to define the capacitance matrix \underline{C} by

$$\vec{q} = \underline{C}\vec{V} \quad (\text{C.3})$$

The coefficients of the capacitance matrix are given by

$$(\underline{C})_{ij} = \begin{cases} -C_{ij} & i \neq j \\ C_{iD} + \sum_{k=1, k \neq i}^{n+m} C_{ik} & i = j \end{cases}$$

So far, there has been no distinction between electrodes and islands. The definition of an ideal voltage source requires that the potential of electrodes is constant, no matter what. If, for example, an electron tunnels from electrode k to an island, the potential of the electrode must not change. The voltage source has to do some work to replace the electron. For the islands, this looks as if the electrode has an infinitely large capacitance with drain, $C_{kD} = \infty \Rightarrow C_{kk} = \infty$: no matter how the island potentials change, this will not modify the electrode potential.

We will proceed by applying this formalism to a single-electron tunneling circuit. We are interested in studying how the electrostatic energy changes as \vec{V}_E changes, which may induce charge transfers across “leaky” capacitors, i.e. those capacitors that allow tunneling. Since the voltage sources represent electron reservoirs, the electrostatic energy is in fact a free energy, given by the total energy stored in the system, minus the work W done by the voltage sources. Typically, it is convenient to specify the initial state by \vec{q}_I and \vec{V}_E . A transition to a different state can be characterized by a change of the charge vector, $\Delta\vec{q}$. We therefore write

$$\begin{aligned} \Delta E[\vec{V}_E, \vec{q}, \Delta\vec{q}] &= \frac{1}{2}(\vec{q} + \Delta\vec{q})(\underline{C})^{-1}(\vec{q} + \Delta\vec{q}) - \frac{1}{2}\vec{q}(\underline{C})^{-1}\vec{q} - \Delta W \\ &= \Delta\vec{q}\underline{C}^{-1}\vec{q} + \frac{1}{2}\Delta\vec{q}\underline{C}^{-1}\Delta\vec{q} - \Delta W \end{aligned} \quad (\text{C.4})$$

Inverting \underline{C} results in

$$\underline{C}^{-1} = \begin{pmatrix} 1 & & \\ \vdots & & \\ n & & \\ n+1 & \left(\begin{array}{cc} C_{\text{II}}^{-1} & 0 \\ 0 & 0 \end{array} \right) \\ \vdots & & \\ n+m & & \end{pmatrix}$$

Hence, only the $n \times n$ submatrix of \underline{C}^{-1} that describes inter-island coupling contains non-vanishing elements. Note that $(\underline{C}^{-1})_{\text{II}} = \underline{C}_{\text{II}}^{-1}$. The energy difference in Eq. (C.4) is thus independent of \vec{q}_E . This equation now reads

$$\Delta E[\vec{V}_E, \vec{q}_I, \Delta \vec{q}] = \Delta \vec{q}_I \underline{C}_{\text{II}}^{-1} \vec{q}_I + \frac{1}{2} \Delta \vec{q}_I \underline{C}_{\text{II}}^{-1} \Delta \vec{q}_I - \Delta W \quad (\text{C.5})$$

It remains to calculate the work done by the voltage sources as the charge vector is changed. This work is made up of two components:

- As the charge of one island i changes, all islands that couple to island i will change their potentials accordingly, and their potential difference to an electrode k will change as well. In order to keep V_k constant, the voltage source connected to it has to take care of the charge changes influenced at electrode k . The work done is given by

$$\Delta W_{k,1} = \Delta q_k V_k = \sum_{j=1}^n \Delta V_j C_{jk} V_k$$

where

$$\Delta V_j = \sum_{i=1}^n (\underline{C}_{\text{II}}^{-1})_{ij} \Delta q_i$$

The work done by all voltage sources can therefore be written as

$$\Delta W_1 = \Delta \vec{q}_I \underline{C}_{\text{II}}^{-1} \underline{C}_{IE} \vec{V}_E \quad (\text{C.6})$$

The capacitance coefficients between islands and electrodes form the matrix \underline{C}_{IE} .

- Some electrodes may be connected to some islands via tunnel junctions. In the case when electrons tunnel between electrode k and an island, the voltage source has to neutralize this charge change Δq_k , which requires the work $\Delta W_{k,2} = -\Delta q_k V_k$. Note that such a process will also change the charge configuration at the islands, and the voltage source will therefore

have to perform the corresponding work $\Delta W_{k,1}$ in addition. The contribution to the work done by all voltage sources due to such tunnel processes is given by

$$\Delta W_2 = -\Delta \vec{q}_E \vec{V}_E \quad (C.7)$$

Therefore, the total work done by the voltage sources in response to a change in the island charge configuration is given by

$$\Delta W = \Delta W_1 + \Delta W_2 \quad (C.8)$$

We therefore obtain the final result (Eq. (9.4)):

$$\Delta E[\vec{V}_E, \vec{q}_I, \Delta \vec{q}] = \Delta \vec{q}_I \underline{\mathcal{C}}_{II}^{-1} [\vec{q}_I + \frac{1}{2} \Delta \vec{q}_I - \underline{\mathcal{C}}_{IE} \vec{V}_E] + \Delta \vec{q}_E \vec{V}_E$$

D

The Transfer Hamiltonian

Occasionally, one cannot make the assumption that the energies before and after the tunnel event are identical. This may be due to tunneling induced by photons or phonons, or due to electron-electron interactions, as in Chapter 9. In such cases, the transfer Hamiltonian model is useful, which is based upon time-dependent perturbation theory.

The problem can be elegantly dealt with within the so-called transfer Hamiltonian model. We start from two electron gases separated by an impenetrable barrier. Now, a time-dependent perturbation Hamiltonian is considered, which allows transfer of electrons across the barrier. Time-dependent perturbation theory shows that the transfer rate can be described with Fermi's golden rule, which we consider in the static limit:

$$\Gamma_{i \rightarrow f} = \frac{2\pi}{\hbar} |\langle i | H_t | f \rangle|^2 \delta(E_f - E_i) \quad (\text{D.1})$$

This is just the transmission probability per unit time for a single electron in state $|i\rangle$, with energy E_i , to be transferred into state $|f\rangle$, with energy E_f , on the other side of the barrier. The δ function ensures an elastic event. In order to relate the transfer rate to a current at a voltage drop V across the barrier, we have to consider the following.

1. The electron density in $[E_i, E_i + dE_i]$ is given by the density of states $D_i(E_i)$ times the Fermi-Dirac distribution $f_i(E_i)$. Here the index i denotes the side of the barrier that hosts state i .
2. Since we are dealing with fermions, the electron can tunnel only into an empty state $|f\rangle$. The transfer rate for an electron in $|i\rangle$ will thus be proportional to $D_f(E_f)[1 - f_i(E_f)]$.
3. Electrons can tunnel in both directions. The current is the sum of the two partial currents in opposite directions.

Let us assume that the voltage drop is from left to right (Fig. 9.5). The spectral current at energy E is given by

$$I(E) = e \frac{2\pi}{\hbar} |\langle i|H_t|f\rangle|^2 \{ D_1(E)f(E - E_F)D_r(E + eV)[1 - f(E - E_F - eV)] \\ - D_r(E + eV)f(E - E_F - eV)D_1(E)[1 - f(E - E_F)] \}$$

For large energy barriers, the matrix elements of the perturbation Hamiltonian will be independent of energy.

Second, we assume that the density of states does not depend on energy, either since the electron gas is two-dimensional, or since the voltage drop is sufficiently small. In this approximation, the total current is obtained by integration over all relevant energies:

$$I(V) = e \frac{2\pi}{\hbar} |\langle i|H_t|f\rangle|^2 D^2 \int_{E_{cb},l}^{\infty} [f(E - E_F) - f(E - E_F - eV)] dE$$

If the thermal energy and eV are small compared to E_F , the Fermi functions can both be approximated by step functions, and the integral simply gives eV , resulting in

$$I(V) = \frac{2\pi e^2}{\hbar} |\langle i|H_t|f\rangle|^2 D^2 V$$

Hence, large tunnel barriers show a linear I - V characteristic for small voltages, with a resistance given by

$$R = \frac{\hbar}{2\pi e^2 |\langle i|H_t|f\rangle|^2 D^2}$$

and consequently we can speak of a voltage-independent conductance, which is directly related to transmission $T = 4\pi |\langle i|H_t|f\rangle|^2 D^2$.

Now let us study a tunnel event in an SET device. Here, the electrostatic energy may change, and E_f can differ from E_i . In that case, we have to change Fermi's golden rule accordingly:

$$\Gamma_{i \rightarrow f} = \frac{2\pi}{\hbar} |\langle i|H_t|f\rangle|^2 \delta(E_f - E_i - \Delta E) \quad (\text{D.2})$$

The tunneling rates as a function of the voltage applied now read:

$$\Gamma^{\pm}(V) = \frac{1}{e^2 R} \int_{E_{cb},l}^{\infty} f(E)[1 - f(E + \Delta E^{\pm})] = \frac{1}{R e^2} \frac{\Delta E}{1 - \exp(\Delta E/k_B\Theta)}$$

E Solutions to Selected Exercises

Chapter 2

E2.1 A zinc blende lattice hosts four atoms of each type per unit cell. Therefore, $m_{\text{GaAs}} = 12.12 \text{ g}$. The unit cell of the diamond lattice contains eight atoms, hence $m_{\text{Si}} = 5.58 \text{ g}$.

E2.2

(a) The reciprocal lattice vectors are

$$\vec{b}_1 = \frac{\pi}{6 \text{ nm}} (3, -1) \quad \text{and} \quad \vec{b}_2 = \frac{\pi}{6 \text{ nm}} (0, 4)$$

(b) The electron density is

$$n = \frac{2\pi}{3 \times 12 \text{ nm}^2} \quad \Rightarrow \quad |k_F| = \sqrt{2\pi n} = \frac{\pi}{3} \text{ nm}^{-1}$$

The Fermi circle just touches the edge of the first Brillouin zone in the \vec{b}_2 -direction (see Fig. E.1(a)). Both the first and the second Brillouin zones are partly filled. We do have a metal here, as always for materials where the number of conduction electrons per unit cell is not an even integer.

(c) The repeated zone scheme of Fig. E.1(b) reveals that there is one hole-type de Haas–van Alphen (dHvA) oscillation. From the enclosed area, a dHvA period of $\Delta(1/B) \approx 2\pi e/\hbar A \approx 0.035 \text{ T}^{-1}$ is expected. In addition, the Fermi surface in the second Brillouin zone gives an electron-type orbit with $\Delta(1/B) \approx 0.024 \text{ T}^{-1}$.

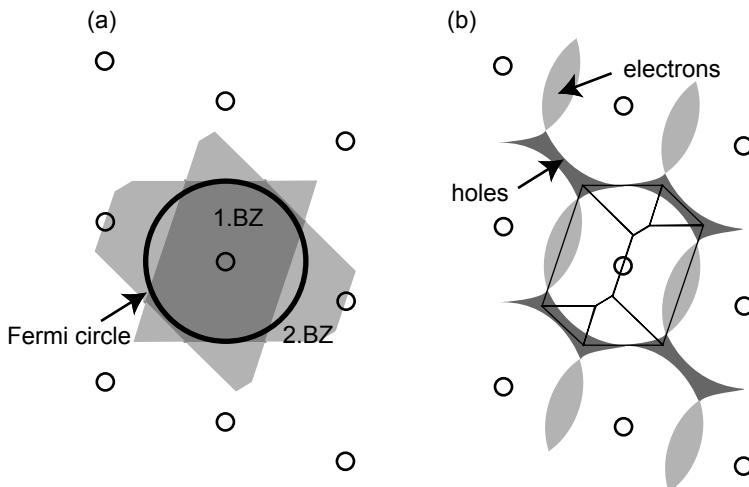


Fig. E.1 (a) Reciprocal lattice, first and second Brillouin zones, and Fermi circle for Exercise E2.3. (b) The repeated zone scheme reveals the regions filled with holes (dark gray) and those filled with electrons (light gray). Also shown is how the elements of the second Brillouin zone combine to the reduced zone scheme.

E2.3

(a) We have

$$\begin{aligned}\langle n_j \rangle &= \frac{2e^{-(E_j - \mu)/k_B\Theta} + 2e^{-2(E_j - \mu)/k_B\Theta}}{1 + 2e^{-(E_j - \mu)/k_B\Theta} + e^{-2(E_j - \mu)/k_B\Theta}} \\ &= \frac{e^{-(E_j - \mu)/k_B\Theta}[1 + e^{-(E_j - \mu)/k_B\Theta}]}{[1 + e^{-(E_j - \mu)/k_B\Theta}][1 + e^{-(E_j - \mu)/k_B\Theta}]} = 2 \frac{1}{1 + e^{(E_j - \mu)/k_B\Theta}}\end{aligned}$$

(b) In analogy to (a) one finds

$$\langle n_j \rangle = \frac{2e^{-(E_j - \mu)/k_B\Theta}}{1 + 2e^{-(E_j - \mu)/k_B\Theta}} = \frac{1}{1 + \frac{1}{2}e^{(E_j - \mu)/k_B\Theta}}$$

(c) The state cannot be occupied by two holes. In other words, the spin-degenerate acceptor state can be occupied by one or by two electrons:

$$\begin{aligned}\langle n_j \rangle &= \frac{2e^{-(E_j - \mu)/k_B\Theta} + 2e^{-2(E_j - \mu)/k_B\Theta}}{2e^{-(E_j - \mu)/k_B\Theta} + e^{-2(E_j - \mu)/k_B\Theta}} \\ &= \dots = 2 - \frac{1}{1 + \frac{1}{2}e^{-(E_j - \mu)/k_B\Theta}}\end{aligned}$$

Defining the average hole occupation number $\langle p_j \rangle \equiv 2 - \langle n_j \rangle$, one obtains

$$\langle p_j \rangle = \frac{1}{1 + \frac{1}{2}e^{(\mu - E_j)/k_B\Theta}}$$

E2.4 Expanding the quantities that vary slowly over the interval of non-vanishing weight function in Eq. (2.30), namely $E(\vec{k})$ and $u_{\vec{k}}(\vec{r})$, in a Taylor series to first order in $\delta\vec{k}$ gives

$$\begin{aligned} |\Phi_e(\vec{k}_e, \vec{r}_e, t)\rangle &\approx |\phi_e(\vec{k}_e, \vec{r}_e, t)\rangle \\ &\times \int_{-\infty}^{\infty} \delta\vec{k} w(\vec{k} - \vec{k}_e) \exp \left[i\delta\vec{k} \left(\vec{r} - \frac{1}{\hbar} \vec{\nabla}_{\vec{k}} E(\vec{k}) \Big|_{\vec{k}_e} t \right) \right] d\vec{k} \end{aligned}$$

which is the Bloch wave for \vec{k}_e , seen through a window function given by the integral. The important point to realize is that the window function moves with a velocity given by Eq. (2.31) across the Bloch wave.

As far as the width of the window function is concerned, please note that $1/a \gg \delta\vec{k} \approx 1/\lambda_{\text{de Broglie}}$, according to the Heisenberg uncertainty relation, where a is the lattice constant in real space.

E2.5 For $\vec{q} \rightarrow 0$, the function $F(s)$ (see Eq. (2.50)) approaches unity, and the dielectric function can be approximated by

$$\epsilon(\vec{q}) \approx 1 + \frac{k_{\text{TF}}^2}{q^2}$$

The potential of the point charge is

$$V_{\text{ext}}(\vec{r}) = -\frac{Ze^2}{r}$$

where Z is the number of protons. Fourier transformation yields

$$V_{\text{ext}}(\vec{r}) = -\frac{Ze^2}{(2\pi)^3} \int_{-\infty}^{\infty} \frac{4\pi}{q^2} e^{i\vec{q}\vec{r}} dq$$

The Fourier components are thus given by

$$V_{\text{ext}}(\vec{q}) = -\frac{4\pi Ze^2}{q^2}$$

The Fourier components of the screened potential are given by

$$V_{\text{eff}}(\vec{q}) = -\frac{4\pi Ze^2}{q^2 \epsilon(\vec{q})}$$

The Fourier transform of the screened potential thus reads

$$V_{\text{eff}}(\vec{r}) = -\frac{Ze^2}{(2\pi)^3} \int_{-\infty}^{\infty} \frac{4\pi}{q^2 + k_{\text{TF}}^2} e^{i\vec{q}\cdot\vec{r}} dq$$

An evaluation of this integral gives the Yukawa potential

$$V_{\text{eff}}(\vec{r}) = -\frac{Ze^2}{r} e^{-ik_{\text{TF}}r}$$

E2.6

(a) The Schrödinger equation reads

$$-\frac{\hbar^2}{2m} \frac{d^2\Phi(x)}{dx^2} - V_0 \delta(x) \Phi(x) = E \Phi(x)$$

$\Phi(x)$ must be continuous at $x = 0$; this requirement can be written as $\lim_{\eta \rightarrow 0} \Phi(\eta) = \lim_{\eta \rightarrow 0} \Phi(-\eta)$. The second necessary condition is obtained from integrating the Schrödinger equation, with the integration limits approaching $x = 0$:

$$\lim_{\eta \rightarrow 0} \left[-\frac{\hbar^2}{2m} \int_{-\eta}^{\eta} \frac{d^2\Phi(x)}{dx^2} dx - V_0 \int_{-\eta}^{\eta} \delta(x) \Phi(x) dx \right] = E \int_{-\eta}^{\eta} \Phi(x) dx$$

thus

$$-\frac{\hbar^2}{2m} [\Phi'(+0) - \Phi'(-0)] - V_0 \Phi(0) = 0$$

Since the wave function is evanescent everywhere except for $x = 0$, the ansatz $\Phi(x) = Ae^{-\kappa|x|}$ makes sense. By inserting this expression in the conditions above, we obtain $\kappa = mV_0/\hbar^2$. From the normalization condition $\int |A|^2 e^{-2\kappa|x|} dx = 1$, the amplitude $A = \sqrt{\kappa}$ is found. Hence,

$$\Phi(x) = \sqrt{\kappa} e^{-\kappa|x|}$$

The eigenvalue is obtained from

$$E_0 = \frac{(\hbar i \kappa)^2}{2m} = -\frac{m V_0^2}{2\hbar^2}$$

(b) $\Psi_k(x)$ satisfies the Bloch theorem if $\Psi_k(x + na) = e^{ikna} \Psi_k(x)$. This is in fact the case:

$$\begin{aligned} \Psi_k(x + na) &= \sum_{j=-\infty}^{\infty} \Phi_0(x + na - ja) e^{ikja} \\ &= e^{ikna} \sum_{j=-\infty}^{\infty} \Phi_0[x - (j - n)a] e^{ik(j-n)a} \\ &= e^{ikna} \Psi_k(x) \end{aligned}$$

(c) We carry out the integration as suggested in the exercise:

$$\langle \Phi_0 | H_0 + \Delta V | \Psi_k \rangle = E(k) \langle \Phi_0 | \Psi_k \rangle$$

Here H_0 denotes the Hamiltonian of a single δ function at $x = 0$, and $\Delta V(x)$ is the residual crystal potential without the δ potential at the origin. The last equation can be rewritten as

$$E_0 I_0(k) + I_1(k) = E(k) I_0(k) \quad \Rightarrow \quad E(k) = E_0 + \frac{I_1(k)}{I_0(k)}$$

We proceed by calculating I_0 :

$$\begin{aligned} I_0(k) &= \langle \Phi_0 | \Psi_k \rangle \\ &= 1 + \sum_{j=1}^{\infty} [e^{ikja} \langle \Phi_0(x) | \Phi_0(x - ja) \rangle + e^{-ikja} \langle \Phi_0(x) | \Phi_0(x + ja) \rangle] \end{aligned}$$

The first term stems from the contribution of $j = 0$. The two integrals entering here are identical, therefore

$$I_0(k) = 1 + \sum_{j=1}^{\infty} 2 \cos(jka) \langle \Phi_0(x) | \Phi_0(x + ja) \rangle$$

where the overlap integral $\langle \Phi_0(x) | \Phi_0(x + ja) \rangle = \alpha_j$ is given by

$$\begin{aligned} \alpha_j &= \kappa \langle e^{-\kappa|x|} | e^{-\kappa|x+ja|} \rangle \\ &= \kappa \left[e^{\kappa ja} \int_{-\infty}^{-ja} dx e^{2\kappa x} + e^{-\kappa ja} \int_{-ja}^0 dx + e^{-\kappa ja} \int_0^{ja} dx e^{-2\kappa x} \right] \\ &= (1 + j\kappa a) e^{-j\kappa a} \end{aligned}$$

We finally obtain

$$I_0(k) = 1 + 2 \sum_{j=1}^{\infty} (1 + j\kappa a) e^{-j\kappa a} \cos(jka)$$

For I_1 , the expression reads

$$I_1(k) = \langle \Phi_0 | \Delta V | \Psi_k \rangle = \beta + 2 \sum_{j=1}^{\infty} \gamma_j \cos(jka)$$

with $\beta = \langle \Phi_0 | \Delta V | \Phi_0 \rangle$ and $\gamma_j = \langle \Phi_0(x) | \Delta V | \Phi_0(x + ja) \rangle$. This is a transfer integral. Inserting the wave function leads, via a geometric series, to

$$\beta = \frac{E_0}{1 - e^{2\kappa a}}$$

and after some algebra, one finds

$$\gamma_j = E_0 \left[-je^{-j\kappa a} + \frac{2 \cosh(\kappa a)}{1 - e^{2\kappa a}} \right]$$

We have determined $E(k)$ for this model potential exactly. For an interpretation, we make two approximations. First of all, only nearest-neighbor transfer integrals are assumed to be non-vanishing. Second, terms of the order $e^{-2\kappa a}$ are neglected, i.e. $1/\kappa \ll a$ is assumed. We obtain

$$\alpha_1 = (1 + \kappa a)e^{-\kappa a} \approx e^{-\kappa a}, \quad \beta = 0, \quad \gamma_1 = E_0 e^{-\kappa a}$$

and from this

$$E(k) = E_0 + \frac{2\gamma_1 \cos(ka)}{1 + 2\alpha_1 \cos(ka)}$$

- (d) Since $1/\kappa \ll a$, $\alpha_1 \ll 1$, the denominator in the dispersion relation can be set to unity. A Taylor expansion to second order gives

$$E(k) = E_0 + \gamma_1 (1 - \frac{1}{2} k^2 a^2)$$

Hence,

$$m^* = -\frac{\hbar^2}{\gamma_1 a^2} = \frac{2\hbar^4}{m V_0^2 a^2}$$

Intuitively, this is a very reassuring result: the effective mass depends exponentially on κa . As the nearest-neighbor overlap increases, it becomes easier for the electron to move from site to site, and its effective mass decreases.

E2.7 The effective densities of states are

$$N_c(T) = g \frac{1}{4} \left[\frac{2m_c^* k_B T}{\pi \hbar^2} \right]^{3/2} \quad \text{and} \quad P_v(T) = g \frac{1}{4} \left[\frac{2m_v^* k_B T}{\pi \hbar^2} \right]^{3/2}$$

where $m_{c,v}^*$ is the geometric mean of the eigenvalues of the effective mass tensor, i.e. $m_{c,v}^* = (\prod_{i=1}^3 m_{c,v}^* i)^{1/3}$. This can be seen as follows: We intend to substitute the Fermi ellipsoid given by

$$\frac{\hbar^2}{2E_F} \left(\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y} + \frac{k_z^2}{m_z} \right) = 1$$

by a Fermi sphere with isotropic effective mass m_{eff} , such that the volume is maintained:

$$\left(\frac{2E_F}{\hbar^2} \right)^{3/2} \frac{4\pi}{3} (m_x m_y m_z)^{1/3} = \left(\frac{2E_F}{\hbar^2} \right)^{3/2} \frac{4\pi}{3} (m_{\text{eff}})^{3/2}$$

so that

$$m_{\text{eff}} = \sqrt{m_x m_y m_z}$$

Furthermore, g is the degeneracy of the band in addition to spin degeneracy. Thus, $g = 6$ for electrons in Si, and $g = 1$ otherwise. For Si, this gives $m_c^* = (m_t^*(m_t^*)^2)^{1/3} = 0.321 m_e$ and thus $N_c = 2.76 \times 10^{25} \text{ m}^{-3}$.

In the case of the holes, we simply have to add up the two effective densities of states: $P_v = P_{v,\text{lh}} + P_{v,\text{hh}} = 1.14 \times 10^{25} \text{ m}^{-3}$, such that an intrinsic carrier concentration of $n_{i,\text{Si}}(300 \text{ K}) = \sqrt{N_c P_v} e^{-E_g/2k_B \times 300 \text{ K}} = 7 \times 10^{15} \text{ m}^{-3}$ is found.

There is no valley degeneracy in GaAs, and one obtains $N_c = 4.35 \times 10^{23} \text{ m}^{-3}$, $P_v(300 \text{ K}) = 9.72 \times 10^{24} \text{ m}^{-3}$, and $n_{i,\text{GaAs}}(300 \text{ K}) = 2.42 \times 10^{12} \text{ m}^{-3}$.

At $\Theta = 77 \text{ K}$, $n_{i,\text{Si}}(77 \text{ K}) \approx 10^{-15} \text{ m}^{-3}$, and $n_{i,\text{GaAs}}(77 \text{ K}) \approx 10^{-26} \text{ m}^{-3}$, which is irrelevant in both cases.

Chapter 3

E3.1 The expectation value for the electron position in the z -direction is given by

$$\langle z \rangle = \int_0^\infty \Phi^*(z) z \Phi(z) dz = \int_0^\infty \frac{b^3}{2} z^3 e^{-bz} dz = \frac{b^3}{2} \frac{\Gamma(4)}{b^4} = \frac{3}{b}$$

Here, we have used

$$\int_0^\infty x^n e^{-ax} dx = \frac{\Gamma(n+1)}{a^{n+1}}$$

with $\Gamma(n+1) = n!$ for integer n . This means that the size quantization removes the electrons from the O-S interface, which increases the mobility.

Chapter 4

E4.1 Consider a gas at low pressure in a vacuum chamber. The number of molecules N_c that hit an area A of the wall within a time interval t is given by

$$N_c = n \bar{v}_x t A$$

Here, n denotes the density of the gas, and \bar{v}_x is their average velocity in the x -direction, which is perpendicular to the wall. The quantity $\bar{v}_x t A$ is the volume that contains all the molecules that hit the area A within time t . On the other hand, the pressure is given by $p = n k_B \Theta$, and thus

$$N_c = \frac{p}{k_B T} \bar{v}_x t A \quad \Rightarrow \quad n_c = \frac{p}{k_B T} \bar{v}_x$$

where n_c is the scattering rate at the wall (number of hits per unit area per unit time). We obtain \bar{v}_x from the Maxwell velocity distribution

$$f(\vec{v}) d\vec{v} = \left(\frac{m}{2\pi k_B T} \right)^{3/2} \exp \left(- \frac{mv^2}{2k_B T} \right) d\vec{v}$$

Since we are only interested in the x -component, we integrate over dv_y and dv_z and find

$$g(v_x) = \left(\frac{m}{2\pi k_B T} \right)^{1/2} \exp \left(- \frac{mv_x^2}{2k_B T} \right)$$

In order to get \bar{v}_x from $g(v_x)$, we have to calculate the expectation value of v_x under the constraint $v_x > 0$:

$$\bar{v}_x = \int_0^\infty v_x g(v_x) dv_x = \dots = \sqrt{\frac{k_B T}{2\pi m}} \quad \Rightarrow \quad n_c = \frac{p}{\sqrt{2\pi m k_B T}}$$

We estimate the time it takes until a monolayer of oxygen has formed at the wall. A sticking coefficient of 1 is assumed, which means that all molecules that hit the wall remain there. We denote the area density of molecules within a monolayer by N_s . The time required to form a monolayer is $t_m = N_s / N_c$. As a simple guess, assume that an oxygen molecule has an effective diameter of $d = 0.36 \text{ nm}$. Suppose further that the molecules form a hexagonal lattice, which means that the area $A_m = d^2 \sqrt{3}/2$ contains one O_2 molecule. Then, $N_s = 1/A_m = 8.7 \times 10^{18} \text{ m}^{-2}$, which means that, at a pressure of $p = 10^{-10} \text{ mbar}$, a monolayer forms within $t_m = 6.5 \text{ h}$; at a pressure of $p = 10^{-6} \text{ mbar}$, this only takes 2.4 s! This simple estimate shows that really high vacuum is needed for molecular beam epitaxy!

E4.2

(a) Inserting gives $D \leq 0.03 \text{ cm}^{-2}$.

(b) $N \leq 9$.

(c) $Y \geq 0.9 \Rightarrow D \leq 44.1 \text{ m}^{-2}$. Within 8 inch^3 , we must have fewer than $6A \times D = 0.0528$ particles. Since $1 \text{ ft} = 12 \text{ inch}$, and the class of a clean room is given by the number of particles per cubic foot with sizes larger than 500 nm, then $R \approx 0.18$ is necessary.

E4.3 The dosage is distributed among 2^{26} points, such that a dosage per point of $d = 3 \times 10^{-16} \text{ C}$ is required. The dwell time is therefore $t_{\text{dwell}} = d/I_{\text{beam}} = 30 \mu\text{s}$. The spots form a square lattice with a lattice constant of 12.2 nm. Hence, each spot must have an illuminating diameter of about $12.2 \text{ nm} \times \sqrt{2} = 17 \text{ nm}$ for complete coverage.

Finally, increasing the current means shorter dwell times, which is limited by the speed of the beam control. This can be circumvented by reducing the bit resolution, but then the spatial resolution is lost as well.

E4.4 Applying the rules for operational amplifiers gives the condition

$$V_{\text{out}}(t) = - \left[\frac{R_2}{R_1} V_{\text{in}}(t) + \frac{1}{R_1 C} \int_0^t V_{\text{in}}(\tau) d\tau \right]$$

For $V_{\text{in}} = 0$, the output voltage is constant. Suppose that the input voltage changes as indicated in the question. The response is

$$V_{\text{out}}(t) = -V_0 \left[\frac{R_2}{R_1} \theta(t - t_0) + \frac{t - t_0}{R_1 C} \right]$$

This circuit is known as a proportional-integral (PI) controller. Suppose some parameter, like the temperature in a gas flow cryostat, has to be held constant. The difference between the measured temperature and the required temperature is translated by some circuit into a voltage, which is applied at the input. The output voltage is then used to adjust the temperature to the desired value by some control function, like, in our example, the He gas flow through a needle valve of a gas flow cryostat. Suppose the temperature is too high. The gas flow must be increased, and the output voltage is used to open the needle valve with a step motor. This opening increases with time, until the input voltage has reached zero again, which means that the temperature is back at its required value. PI controllers are widely used for such tasks.

Chapter 5

E5.1 With the dielectric constants $\epsilon_{\text{GaAs}} \approx 13$ and $\epsilon_{\text{Si}} \approx 11$, one finds the values listed in Table E.1. Note that, in Si MOSFETs, there is an additional valley degeneracy of 2! Apparently, GaAs HEMTs are perfect for investigating ballistic and phase coherence effects, while Si MOSFETs are particularly interesting for studying interaction effects.

E5.2 We have to calculate $\langle v_i(0)v_j(t, B) \rangle$ and insert it into Eq. (5.4). The cyclotron motion causes oscillations of the velocity components given by

$$v_i(t) = v_i(0) \cos(\omega_c t)$$

Therefore, one finds

$$\langle v_x(0)v_x(t, B) \rangle = v_F^2 \langle \cos(\phi) \cos(\phi + \omega_c t) \rangle$$

Tab. E.1 Results for Exercise E5.1.

	GaAs ($T = 4.2\text{ K}$)	Si ($T = 4.2\text{ K}$)
Drude scattering time (10^{-12} s)	38	4.3
Fermi velocity (10^4 m/s)	27	1.3
diffusion constant (m^2/s)	1.43	0.00035
Fermi wavelength (nm)	40	95
phase coherence length (nm)	6500	59
inelastic scattering length (nm)	8200	127
thermal length (nm)	1610	25
interaction parameter	0.87	13.8

$$\langle v_x(0)v_y(t, B) \rangle = v_F^2 \langle \cos(\phi) \sin(\phi + \omega_c t) \rangle$$

Averaging over ϕ by evaluation of the integrals results in

$$D_{xx}(B) = \frac{1}{2}v_F^2 \frac{\tau}{1 + \omega_c^2 \tau^2}$$

$$D_{xy}(B) = -\frac{1}{2}v_F^2 \frac{\omega_c \tau^2}{1 + \omega_c^2 \tau^2}$$

The expressions for $D_{yy}(B)$ and $D_{yx}(B)$ are obtained similarly. Replacing the diffusion coefficients with the conductivity components via the Einstein relation for Fermi gases gives Eq. (2.59).

Chapter 6

E6.1 The filling factor $\nu = 4$ is at $B = 6.4\text{ T}$, as can be seen directly from the position of the Hall plateau. Analyzing the Hall slope at small magnetic fields gives $d\rho_{xy}/dB \approx 1060\Omega/\text{T} = -1/n_{2\text{D}}e$, which corresponds to $n_{2\text{D}} = 5.8 \times 10^{15}\text{ m}^{-2}$. This is close to the upper limit of electron densities possible in Ga[Al]As HEMTs, if the second subband must remain empty. The figure tells us that $\rho_{xx}(B=0) \approx 8\Omega$. Since $\rho_{xx}(B=0) = (n_{2\text{D}}e\mu)^{-1}$, we find an electron mobility of $\mu = 134\text{ m}^2/\text{Vs}$. Because $\mu = e\tau/m^*$ and $\ell_e = v_F\tau$, the elastic mean free path $\ell_e = (\hbar/e)\mu\sqrt{2\pi n_{2\text{D}}} = 16.8\text{ }\mu\text{m}$ is obtained. This corresponds to the elastic scattering time of $\tau = 51\text{ ps}$.

Spin splitting states set in at $B = 1.6\text{ T}$. This allows us to estimate the effective g -factor via

$$2g^*\mu_B B_{\text{split start}} \approx \hbar/\tau_q \quad \Rightarrow \quad g^* \approx 5.5.$$

Here, it is assumed that the peaks in the density of states have width \hbar/τ_q , which is a reasonable approximation.

E6.2

- (a) The effective mass is obtained from the slope of $\ln(A/\Theta)$ vs. Θ as: $m^* = 0.032m_e$. The Dingle plot gives a quantum scattering time of $\tau_q = 0.18 \text{ ps}$. Once we know m^* , a Drude scattering time $\tau = \mu m^*/e = 14 \text{ ps}$ is calculated.
- (b) The ratio $\tau/\tau_q = 78$ is extremely large. This means that the dominant source of scattering is remote scattering centers. In fact, the sample is an InAs quantum well 30 nm below the surface, embedded in an AlSb barrier, and capped with a GaSb layer (see Fig. 3.25). It is known that, in this material system, the charge neutrality level of the GaSb surface states lies above the conduction band bottom of InAs, and thus electrons are transferred from the surface into the quantum well. The remaining space charge region close to the surface represents the scattering potential.

E6.3 The potential is given by $U(z) = -e\epsilon z$, where z denotes the growth direction, and ϵ is the electric field. The Schrödinger equation thus reads

$$\left(-\frac{\hbar^2}{2m} \frac{d^2}{dz^2} + \frac{1}{2}m\omega_0^2 z^2 - e\epsilon z \right) \Psi(z) = E\Psi(z)$$

which, by completing the square and substituting

$$z = u + \frac{e\epsilon}{m\omega_0^2}$$

can be rewritten as

$$\left(-\frac{\hbar^2}{2m} \frac{d^2}{du^2} + \frac{1}{2}m\omega_0^2 u^2 \right) \Psi(u) = E^*\Psi(u)$$

with

$$E^* = E + \frac{q^2\epsilon^2}{2m\omega_0^2}$$

The energy eigenvalues are

$$E_n = (n + \frac{1}{2})\hbar\omega_0 - \frac{e^2\epsilon^2}{2m\omega_0^2}$$

The electric field thus displaces the parabolic potential without modifying its shape. The minimum is given by

$$z_{\min}(\epsilon) = \frac{e\epsilon}{m\omega_0^2}$$

$$E_{\min} = \frac{1}{2}\hbar\omega_0 - \frac{q^2\epsilon^2}{2m\omega_0^2}$$

Chapter 7

E7.1

- (a) The problem is very similar to the discussion of the effects of a parallel magnetic field on a 2DEG given in Chapter 6. Using the gauge given in the question, we obtain

$$\begin{aligned} & \left[-\frac{\hbar^2}{2m^*} \frac{\partial^2}{\partial x^2} + i \frac{eB\hbar y}{m^*} \frac{\partial}{\partial x} + \frac{e^2 B^2}{2m^*} y^2 - \frac{\hbar^2}{2m^*} \frac{\partial^2}{\partial y^2} + \frac{1}{2} m^* \omega_0^2 y^2 \right] \psi(y) e^{ik_x x} \\ &= E\psi(y) e^{ik_x x} \end{aligned}$$

Carrying out the partial differentiation with respect to x gives

$$\left[\frac{\hbar^2}{2m^*} k_x^2 - \frac{eB\hbar y}{m^*} k_x + \frac{e^2 B^2}{2m^*} y^2 - \frac{\hbar^2}{2m^*} \frac{\partial^2}{\partial y^2} + \frac{1}{2} m^* \omega_0^2 y^2 \right] \psi(y) = E\psi(y)$$

With the cyclotron frequency $\omega_c = eB/m^*$, we can write

$$\left[-\frac{\hbar^2}{2m^*} \frac{\partial^2}{\partial y^2} + \frac{1}{2} m^* (\omega_0 + \omega_c)^2 y^2 - \frac{eB\hbar y}{m^*} k_x + \frac{\hbar^2}{2m^*} k_x^2 \right] \psi(y) = E\psi(y)$$

Completing the square by adding and subtracting \bar{y}_0^2 , with

$$\bar{y}_0 = \frac{\hbar k_x \omega_c}{m^* \omega^2} = y_0 \left(\frac{\omega_c}{\omega} \right)^2$$

(this relation holds since $y_0 = \hbar k_x / m^* \omega_c$), it follows that

$$\left[-\frac{\hbar^2}{2m^*} \frac{\partial^2}{\partial y^2} + \frac{1}{2} m^* \omega^2 (y - \bar{y}_0)^2 + \frac{\hbar^2}{2m^*} k_x^2 \left(\frac{\omega_0}{\omega} \right)^2 \right] \psi(y) = E\psi(y)$$

- (b) The third term on the right-hand side of the previous equation represents the energy dispersion in the x -direction. The electron mass now depends on B and is known as the magnetic mass, $m^*(B) = m^*(\omega/\omega_0)^2$. For large B , the solution approaches the Landau quantization, with the magnetic mass going toward infinity.

- (c) With the above solution, the electron density of a quantum wire in a magnetic field can be written as

$$n_{\text{QWR}} = \frac{2}{\pi\hbar} \sum_{j=0}^{\infty} \sqrt{2m^*(\omega/\omega_0)^2(E_F - E_j)} [1 - \Theta(E_F - E_j)]$$

For integer filling factors N , $E_F = \hbar\omega(N + \frac{1}{2})$, such that

$$\begin{aligned} n_{\text{QWR}} &= \frac{2}{\pi\hbar} \sum_{j=0}^N \sqrt{2m^*(\omega/\omega_0)^2\hbar\omega(N-j)} \\ &\approx \frac{2}{\pi} \sqrt{\frac{2m^*}{\hbar}} \left(\frac{\omega^{3/2}}{\omega_0} \right) \int_{l=0}^N l^{1/2} dl \end{aligned}$$

This gives the relation

$$n_{\text{QWR}} = \frac{4}{3\pi} \sqrt{\frac{2m^*}{\hbar}} \left(\frac{\omega^{3/2}}{\omega_0} \right) N^{3/2}$$

The fit parameters are ω_0 and n_{QWR} ($\Rightarrow E_F$). Since $E_F = \frac{1}{2}m^*\omega_0^2(w/2)^2$, we can determine the electronic wire width w .

E7.2 We write down the current that flows at a bias voltage V :

$$\begin{aligned} I &= \int \{ \vec{D}_1(E)v(E)f(E-\mu)[1-f(E-\mu+eV)] \\ &\quad - \overleftarrow{D}_1(E+eV)v(E+eV)f(E-\mu+eV) \\ &\quad \times [1-f(E-\mu+eV)] \} \Theta(E-E_1) dE \\ &= \frac{2e}{h} \int \Theta(E-E_1)[f(E-\mu)-f(E-\mu+eV)] dE \end{aligned}$$

With

$$\begin{aligned} f(E-\mu+eV) &\approx f(E-\mu) + \frac{\partial f}{\partial eV}(eV=0)eV \\ &= f(E-\mu) + \frac{\partial f}{\partial E}(eV=0)eV \end{aligned}$$

we find

$$I = \frac{2e^2 V}{h} \int -\frac{\partial f}{\partial E}(eV=0)\Theta(E-E_1) dE$$

Partial integration gives

$$I = \frac{2e^2}{h} f(E_1 - \mu)V \quad \Rightarrow \quad G = I/V = \frac{2e^2}{h} f(E_1 - \mu)$$

The steps are thermally smeared as soon as the full width at half-maximum of $\partial f / \partial E$ equals Δ . This is the case for $\Delta = 2k_B\Theta \ln(3 + 2\sqrt{2}) \approx 3.52k_B T$.

E7.3

(a) From the Landauer–Büttiker formalism, the system

$$\begin{pmatrix} I_S \\ I_D \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{pmatrix} = \frac{e^2}{h} \begin{pmatrix} N & 0 & -N & 0 & 0 & 0 \\ 0 & N & 0 & 0 & 0 & -N \\ 0 & 0 & N & -M & M-N & 0 \\ 0 & -N & 0 & N & 0 & 0 \\ -N & 0 & 0 & 0 & N & 0 \\ 0 & 0 & 0 & M-N & -M & N \end{pmatrix} \begin{pmatrix} V_S \\ V_D \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix}$$

is obtained. We set the drain potential to zero and use the fact that $I_S + I_D = 0$. The remaining 5×5 matrix equation has the solution

$$\begin{pmatrix} V_S \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix} = \frac{hI_S}{e^2} \begin{pmatrix} 1/M \\ 1/M - 1/N \\ 0 \\ 1/M \\ 1/N \end{pmatrix}$$

(b) The resistances are obtained from (a) as $R_{ij} = (\mu_i - \mu_j)/eI_S$:

$$R_{12} = R_{34} = \frac{h}{e^2} \left(\frac{1}{M} - \frac{1}{N} \right)$$

$$R_{13} = R_{24} = \frac{h}{e^2} \frac{1}{N}$$

$$R_{14} = \frac{h}{e^2} \left(\frac{1}{M} - \frac{2}{N} \right)$$

$$R_{23} = \frac{h}{e^2} \frac{1}{M}$$

By a proper choice of our setup, we can measure just the barrier.

E7.4

(a) The Landauer–Büttiker matrix is

$$\begin{pmatrix} I_S \\ I_D \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{pmatrix} = \frac{e^2}{h} \begin{pmatrix} 2 & 0 & -2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & -2 \\ 0 & 0 & 2 & 0 & -1 & 0 \\ 0 & -2 & 0 & 2 & 0 & 0 \\ -2 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} V_S \\ V_D \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix} + \frac{e^2}{h} \begin{pmatrix} 0 \\ 0 \\ -V_2^* \\ 0 \\ 0 \\ -V_3^* \end{pmatrix}$$

Next, we have to find the V_i^* . They depend on p , V_2 and V_3 . Current conservation and using the definition of p lead to

$$V_3 + V_c^* = V_3^* + V_c$$

$$V_2^* + V_c^* = V_2 + V_c$$

$$p = 1 - \frac{V_2^* - V_c^*}{V_2 - V_c}$$

$$p = 1 - \frac{V_3^* - V_c}{V_3 - V_c^*}$$

and we obtain the system of equations

$$\begin{pmatrix} V_2 \\ V_3 \\ V_2 \\ V_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -1 & 1 \\ 0 & 1 & 1 & -1 \\ 1/(1-p) & 0 & 1 & 1/(p-1) \\ 0 & 1/(1-p) & 1/(p-1) & 1 \end{pmatrix} \begin{pmatrix} V_2^* \\ V_3^* \\ V_c \\ V_c^* \end{pmatrix}$$

with the solution

$$\begin{pmatrix} V_2^* \\ V_3^* \\ V_c \\ V_c^* \end{pmatrix} = \frac{1}{p-4} \begin{pmatrix} (2p-4)V_2 - pV_3 \\ -pV_2 + (2p-4)V_3 \\ (p-2)V_2 - 2V_3 \\ -2V_2 + (p-2)V_3 \end{pmatrix}$$

Inserting this in the above linear system gives

$$\begin{pmatrix} I_S \\ I_D \\ I_1 \\ I_2 \\ I_3 \\ I_4 \end{pmatrix} = \frac{e^2}{h} \begin{pmatrix} 2 & 0 & -2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & -2 \\ 0 & 0 & 2 & -\frac{2p-4}{p-4} & -1 + \frac{p}{p-4} & 0 \\ 0 & -2 & 0 & 2 & 0 & 0 \\ -2 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & -1 + \frac{p}{p-4} & -\frac{2p-4}{p-4} & 2 \end{pmatrix} \begin{pmatrix} V_S \\ V_D \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix}$$

Again, we reduce the matrix by the drain column and row (as in the previous exercise) to

$$\begin{pmatrix} V_S \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix} = \frac{hI_S}{2e^2} \begin{pmatrix} (p-4)/(p-2) \\ 2/(2-p) \\ 0 \\ (p-4)/(p-2) \\ 1 \end{pmatrix}$$

(b) From (a), we find immediately that

$$R_{xx} = \frac{V_1 - V_2}{I_S} = \frac{h}{e^2} \frac{1}{2-p}$$

The other resistances are not of interest here.

Inserting the numerical values given in the question, one gets $R_{12} = 0.53h/e^2 \Rightarrow p = 0.113$. To define the equilibration length L_{eq} , we require that along the distance $L = L_{\text{eq}}$ the potential difference between the edge states has been reduced to $1/e = 0.368$ of its initial value:

$$\frac{\Delta\mu^*}{\Delta\mu} = \frac{1}{e}, \quad \frac{\Delta\mu - \Delta\mu^*}{\Delta\mu} = p \quad \Rightarrow \quad p = 1 - \frac{1}{e}$$

or, more generally,

$$p = 1 - e^{-L/L_{\text{eq}}}$$

For the numbers given, we thus obtain $L_{\text{eq}} = 0.42$ mm. Equilibration between spin-polarized edge states takes place on macroscopic length scales, see e.g. [218].

E7.5

- (a) $\vec{C} = n_1\vec{a}_1 + n_2\vec{a}_2$, with $(n_1, n_2) = (4, 1)$. For symmetry reasons, it suffices to consider $n_1 \geq 0$ and $0 \leq n_2 \leq n_1$. For zigzag tubes, $(n_1, n_2) = (n, 0)$; for armchair tubes, $(n_1, n_2) = (n, n)$.
- (b) We write $\vec{A} = m_1\vec{a}_1 + m_2\vec{a}_2$. From $\vec{A}\vec{C} = 0$, $m_1 = N/(2n_1 + n_2)$ and $m_2 = N/(n_1 + 2n_2)$ are found, where N is the smallest common multiple of $(2n_1 + n_2)$ and $(n_1 + 2n_2)$.
- (c) Since $\vec{A}\vec{B} = 2\pi$, we obtain

$$\vec{B} = \frac{2\pi}{m_1^2 + m_1m_2 + m_2^2} (m_1\vec{a}_1 + m_2\vec{a}_2)$$

For an illustration, we express \vec{B} in terms of the reciprocal lattice vectors of the graphite sheet:

$$\vec{b}_1 = \frac{4\pi}{3a^2} (2\vec{a}_1 - \vec{a}_2)$$

$$\vec{b}_2 = \frac{4\pi}{3a^2} (-\vec{a}_1 + 2\vec{a}_2)$$

where a denotes the lattice constant of the graphite sheet, and find

$$\vec{B} = \frac{a^2}{2(m_1^2 + m_1m_2 + m_2^2)} [(2m_1 + m_2)\vec{b}_1 + (m_1 + 2m_2)\vec{b}_2]$$

The mode spacing in the k_y -direction is found to be

$$\Delta k_y = \frac{a^2}{2(n_1^2 + n_1n_2 + n_2^2)} [(2n_1 + n_2)\vec{b}_1 + (n_1 + 2n_2)\vec{b}_2]$$

- (d) This condition is derived in the further reading on CNs given at the end of Chapter 7. The CN under study here is therefore metallic.
- (e) For these zigzag CNs, we can estimate the distance Γ -X to be about π/a . Assuming a parabolic dispersion around Γ , one estimates $m^* \approx 0.8m$. For the metallic tube, the concept of effective mass is not good, since a parabolic dispersion is a very poor approximation. To calculate the density of states, we set $E(k) = \alpha k$, $\Delta k = \pi/L \Rightarrow D(k) = 2L/\pi \Rightarrow d(k) = 2/\pi$. This is translated in energy via

$$d(E) = d(k) dk/dE = \frac{2}{\pi} \frac{1}{\alpha}$$

From the figure, we estimate $\alpha \approx 4 \text{ eV}/(\pi/a) \Rightarrow d(E_F) \approx 1.25 \times 10^{28} \text{ J}^{-1} \text{ m}^{-1}$. Hence, in 1D, a linear energy dispersion gives a constant density of states. Consequently, the chemical potential does not depend on temperature.

E7.6 Let us add a real ohmic contact to the middle region and connect it to ground, while we connect the collector to a voltmeter. The current is now flowing via m, and the Landauer–Büttiker equations now read

$$\begin{aligned} \frac{h}{2e^2} I_i &= NV_i - T_{ci} V_c \\ \frac{h}{2e^2} I_m &= T_{cm} V_c - T_{im} V_i \\ 0 &= NV_c - T_{ic} V_c \end{aligned}$$

Since we consider a situation where $T_{ic} \ll N$, we can approximate $N \pm T_{ic} \approx N$ and find for $T_{cm} = T_{im}$

$$\frac{V_c}{I_i} = \frac{h}{2e^2} \frac{T_{ic}}{N^2}$$

The quantity of interest is thus no longer a small signal on a large background. This setup is frequently used for measurements on ballistic samples.

Chapter 8

E8.1 For the upper branch, one obtains

$$\begin{aligned} \theta_{\text{upper}} &= -\frac{e}{\hbar} \int_{\Gamma} \vec{A} d\vec{\Gamma} = -\frac{e}{\hbar} \int_{\alpha=0}^{\pi} R \begin{pmatrix} \sin \alpha \\ \cos \alpha \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ -BR \cos \alpha \\ 0 \end{pmatrix} d\alpha \\ &= \pi R^2 B \frac{e}{\hbar} \int_{\alpha=0}^{\pi} \cos^2 \alpha d\alpha = \pi R^2 \frac{e}{2\hbar} B = \pi \frac{\Phi}{\Phi_0} \end{aligned}$$

Correspondingly, an electron collects a phase of $\theta_{\text{lower}} = \pi\Phi/\Phi_0$ as it traverses the lower branch. The interference between these two waves generates the Aharonov–Bohm effect:

$$t = \sqrt{\epsilon}(e^{i\phi} + e^{-i\phi})\sqrt{\epsilon} \quad \implies \quad T = t^*t = 4\epsilon \cos^2 \phi$$

E8.2

- (a) We divide the time t into N intervals of equal length. N is so large that none of the intervals hosts two scattering events. An individual interval is occupied with a probability of $p = \gamma t/N$. The probability for j scattering events follows from the probability that j of the intervals are occupied, times the number of possible arrangements of the occupied intervals among all intervals. Hence,

$$P(j) = p^n(1-p)^{n-j} \times \frac{N!}{j!(N-j)!}$$

In the limit $N \rightarrow \infty$, this probability becomes

$$P(j) = \frac{\gamma^j t^j}{j!} e^{-\gamma t}$$

This is the Poisson distribution of random processes.

- (b) Clearly, $\ell_{\text{e-e}}$ should be the average distance an electron travels before it hits one of its colleagues. Therefore, we require $j = 0$ in the Poisson distribution, which then reads $P(0) = e^{-\gamma t}$. Mapping γ and t on length scales is easy: $t = L/v_F$ and $\gamma = 1/\tau_{\text{e-e}} = v_F/\ell_{\text{e-e}}$. Here, v_F is the Fermi velocity and L is the flight distance under consideration (remember that we are in the ballistic regime). This gives

$$P(0) = e^{-L/(\ell_{\text{e-e}})}$$

Since the amplitude as defined in the text equals $P(0)$, and the assumption has been made that complete dephasing occurs in individual e–e scattering events, Eq. (8.9) follows.

E8.3

- (a) The Schrödinger equation of the system reads

$$\frac{1}{2m^*}(\vec{p} + e\vec{A})^2\Psi(\phi) = E\Psi(\phi)$$

In cylindrical coordinates,

$$\vec{\nabla} \times \vec{A} = \left(\frac{1}{r} \frac{\partial A_z}{\partial \phi} - \frac{\partial A_\phi}{\partial z}, \frac{\partial A_r}{\partial z} - \frac{\partial A_z}{\partial r}, \frac{1}{r} \frac{\partial(rA_\phi)}{\partial r} - \frac{1}{r} \frac{\partial A_r}{\partial z} \right)$$

and

$$\vec{\nabla}\Psi = \left(\frac{\partial}{\partial r}, \frac{1}{r} \frac{\partial}{\partial \phi}, \frac{\partial}{\partial z} \right) \Psi$$

The vector potential in the question gives $\vec{\nabla} \times \vec{A} = (0, 0, B)$. With the ansatz for the wave function

$$\Psi(\phi) = \frac{1}{\sqrt{2\pi r}} e^{i\ell\phi}$$

the Schrödinger equation becomes

$$\frac{1}{2m^*} \left(-i\hbar \frac{1}{r} \frac{\partial}{\partial \phi} + \frac{erB}{2} \right)^2 e^{i\ell\phi} = E e^{i\ell\phi}$$

leading to the energy eigenvalues

$$E_\ell(B) = \frac{\hbar^2}{2m^* r^2} \left(\ell + \frac{eBr^2}{2\hbar} \right)^2$$

Using the magnetic flux quantum $\Phi_0 = h/e$, we can rewrite this as

$$E_{\ell,n} = \frac{\hbar^2}{2m^* r^2} (\ell + n)^2$$

with n being the number of magnetic flux quanta penetrating the ring. Of course, ℓ is the angular momentum quantum number (Fig. E.2). Note that the probability density is independent of ℓ, ϕ and B .

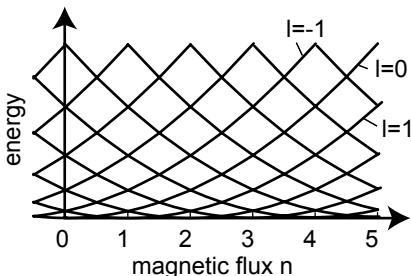


Fig. E.2 Energy spectrum of a one-dimensional quantum ring (Exercise E8.3).

(b) The current is obtained from

$$I_\ell = -\frac{i\hbar e}{2m^*} (\Psi^* \vec{\nabla} \Psi - \Psi \vec{\nabla} \Psi^*)$$

Inserting a wave function gives

$$I_\ell = \frac{\hbar e}{4\pi m^* r^2} \ell$$

Since $\vec{L} = \vec{r} \times \vec{p} = \hbar \sqrt{\ell(\ell+1)} \approx \hbar \ell$, we can write this as

$$I_\ell = \frac{ev_F}{2\pi r} = ev$$

Here, v denotes the circulation frequency of the electron in the ring. States with ℓ and $-\ell$ are degenerate, such that the corresponding currents cancel each other. If, however, the number of electrons in the ring is odd, an *equilibrium current* flows in the ring. This current is known as *persistent current*.

- (c) Suppose that there are about 100 electrons ($\ell = 50$) in the ring with radius $r = 300 \text{ nm}$. A persistent current of 24 nA is found. This is a large current. It can be, and has been, measured by different techniques. One way is to detect the magnetic field generated by the current loop, using a superconductive quantum interference device (SQUID) [200]. Another way becomes apparent as soon as one realizes that

$$I_\ell = \frac{1}{2A_{\text{ring}}} \frac{\partial E_\ell(B=0)}{\partial B}$$

Hence, the magnetic field dispersion of E_ℓ directly measures the persistent current. This can be done in a resonant tunneling experiment (see Chapter 10).

E8.4

- (a) In analogy to Chapter 8, one finds

$$I = \frac{2e}{h} \int -\frac{\partial f}{\partial E}(eV=0) \delta(E - E_r) dE \quad \Rightarrow$$

$$G = I/V = \frac{2e^2}{h} - \frac{\partial f}{\partial E}(E_r - \mu)$$

The peak transmission at $E_r = \mu$ equals $T(E_r) = 1/4k_B\Theta$. For the FWHM, one obtains

$$\frac{1}{8k_B\Theta} = \frac{1}{k_B\Theta} \frac{e^{(E_{1/2}-\mu)/k_B\Theta}}{(1 + e^{(E_{1/2}-\mu)/k_B\Theta})^2} \quad \Rightarrow$$

$$E_{1/2} = k_B \Theta \ln(3 \pm 2\sqrt{2}) = \pm 1.7627 k_B T$$

[Note the remarkable relation $a^2 - b^2 = 1 \Rightarrow -\ln(a - b) = \ln(a + b)$.]

- (b) The general line shape is a convolution of a Lorentzian with the derivative of the Fermi function:

$$G(\mu, E_r) = \frac{2e^2}{h} \frac{\Gamma_a \Gamma_b}{\Gamma} \int -\frac{\partial f}{\partial E}(E - \mu) L(\Gamma, E) dE$$

$L(\Gamma, E)$ is the Lorentzian. Experimentally, either one can fit the data to the above expression, using both Θ and Γ as fit parameters, or one can vary the temperature and plot the FWHM as a function of Θ . The saturation temperature should give a good estimate for Γ .

Chapter 9

- E9.1** The single-electron box consists of one electrode and one island. The capacitance matrix of the circuit in Fig. 9.20 reads

$$\underline{C} = \begin{pmatrix} C_{11} & -C_{1G} \\ -C_{1G} & C_{GG} \end{pmatrix}$$

with $C_{11} = C_{1G} + C_{1D}$ and $C_{GG} = C_{1G}$. The island charge equals $q = q_0 - ne$, such that the charge vector is $\vec{q} = (q_0 - ne, q_G)$. The voltage vector is just V_G . Here, n is the excess number of electrons at the island.

Two charge transfers are possible via the leaky capacitor between island and drain, $\Delta \vec{q} = e(\pm 1, 0)$, which means that the energy relation

$$\Delta E[V, q_0 - ne, e(\pm 1, 0)] \geq 0$$

has to hold. Therefore, n excess electrons are on the island for

$$\frac{1}{C_G} (n(e - \frac{1}{2}) - q_0) < V_G < \frac{1}{C_G} (n(e + \frac{1}{2}) - q_0)$$

Fig. E.3 shows $n(V_G)$ for $q_0 = 0$.

- E9.2** In that case,

$$\Gamma_{i \rightarrow f} = \frac{2\pi}{\hbar} |\langle i | H_t | f \rangle|^2 \delta(E_f - E_i)$$

which means that

$$\Gamma_{1 \rightarrow 2}(\Delta E = 0) = \frac{2\pi}{\hbar} \int_{E_{cb},\max}^{\infty} |\langle i | H_t | f \rangle|^2 D_i(E) D_f(E) f(E) [1 - f(E - eV)] dE$$

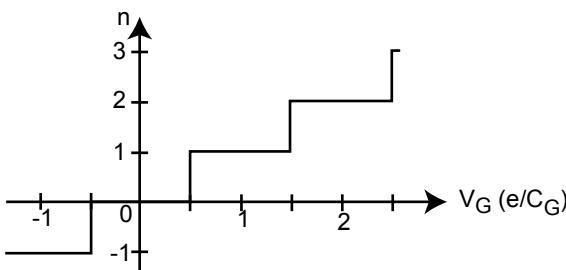


Fig. E.3 Number of excess electrons in a single-electron box (circuit of Fig. 9.20; Exercise E9.1).

For low temperatures, we approximate the Fermi functions by step functions. Furthermore, in the limit of small voltages, the densities of states on both sides of the barrier are identical (in $d = 2$, this is the case anyway), and we obtain

$$I = e\Gamma_{1 \rightarrow 2} = \frac{2\pi}{\hbar} |\langle i | H_t | f \rangle|^2 D^2(E) V$$

E9.3 Although the circuit resembles somewhat the double dot in series, it behaves quite differently. The current through island 1 can be tuned by both gate voltages, although V_B couples to it only via island 2. Since all capacitances are supposed to be equal, we have

$$\underline{C}_{II} = \begin{pmatrix} 4C & -C \\ -C & 2C \end{pmatrix}$$

$$\underline{C}_{IE} = \begin{pmatrix} -C & 0 & -C \\ 0 & -C & 0 \end{pmatrix}$$

For the exchange of electrons between island 1 and the electrodes the charge transfers to be considered are

$$\Delta \vec{q} = (\Delta \vec{q}_I, \Delta \vec{q}_E) = (\Delta q_1, \Delta q_2, \Delta q_A, \Delta q_B, \Delta q_S) = e(\pm 1, 0, 0, 0, \mp 1)$$

for transfers between 1 and S, and

$$\Delta \vec{q} = (\Delta \vec{q}_I, \Delta \vec{q}_E = e(\pm 1, 0, 0, 0, 0)$$

for transfers between 1 and D. Since we consider only the case of $V_S = 0$, transfers between S and 1 should be equivalent to those between D and 1. In other words, $\Delta \vec{q}_E$ will be irrelevant below.

For the exchange of electrons between the islands, we have

$$\Delta \vec{q}_I = e(\pm 1, \mp 1)$$

In addition, the processes

$$\Delta \vec{q}_I = e(0, \pm 1)$$

corresponding to a direct electron transfer between island 2 and S or D are taken into account, for reasons that will become apparent below. The conditions for a stable configuration (n_1, n_2) are obtained by evaluating the system using Eq. (9.4). One gets the stability conditions

$$V_B \leq -2V_A + \frac{e}{C}(2n_1 + n_2 + 1)$$

$$V_B \geq -2V_A + \frac{e}{C}(2n_1 + n_2 - 1)$$

for electron transfers between 1 and S or D, and

$$V_B \leq \frac{1}{3}V_A + \frac{e}{C}(n_2 - \frac{1}{3}n_1 + \frac{2}{3})$$

$$V_B \geq \frac{1}{3}V_A + \frac{e}{C}(n_2 - \frac{1}{3}n_1 - \frac{2}{3})$$

for inter-island transfer. For the direct transfer between 2 and S or D, one finds

$$V_B \leq -\frac{1}{4}V_A + \frac{e}{C}(n_2 + \frac{1}{4}n_1 + \frac{1}{2})$$

$$V_B \geq -\frac{1}{4}V_A + \frac{e}{C}(n_2 + \frac{1}{4}n_1 - \frac{1}{2})$$

respectively. The stability diagram in the (V_A, V_B) plane is shown in Fig. E.4. Leaving the condition for transfers between 2 and S or D aside for the moment, we find a set of intersecting diamonds.

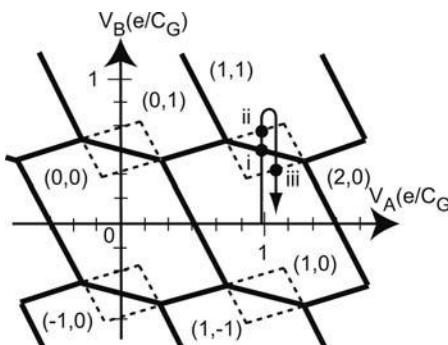


Fig. E.4 Stability diagram of the circuit considered in Exercise E9.3.

As for the relevance of the direct charge transfer between island 2 and the reservoirs, we follow the arrow in Fig. E.4, and increase V_B at constant $V_A \approx 1$ (in units of e/C). As we cross $V_B = 1/2$ (point i), the free energy of the configuration $(1,1)$ becomes smaller than that of the initial $(1,0)$ configuration. But the charge transfer should not be possible. It will nevertheless take place via tunneling with a large time constant, in order to relax the system into the

ground state. If it can be neglected, the system will remain in a metastable state within the region bounded by the dashed lines, until the energies of the configurations $(1, 0)$ and $(0, 1)$ are equal. This is the case at point ii, and finally allows island 2 to obtain an electron from the reservoirs via island 1. As we go back, the system has to wait until the intermediate state $(2, 0)$ becomes accessible, which happens at point iii. Via this state, the electron is transferred back into the reservoir. Hence, if the direct charge transfer between island 2 and the reservoirs is not possible, the system shows hysteresis effects within the diamonds formed by the dashed lines.

Chapter 10

E10.1

- (c) Only states with $m = 1$ couple sufficiently strongly to the leads, such that a current can be detected. In Fig. 10.8, we thus see the fraction of zigzag lines that corresponds to Landau level 1 states. Removing the charging energy gives the discrete spectrum of the island. Suppose it is approximately a Fock–Darwin spectrum. The beginning and the end of each bright line correspond to level crossings of a Landau level 1 state with a Landau level 2 state. One finds $(\Delta B)_{\text{measured}} = 75 \text{ mT}$ at $B \approx 7 \text{ T}$. So far, we have neglected the spin, though. The spin splitting of both Landau levels reduces the average period in B by a factor of 2. Hence, we find

$$\omega_0 = \omega_c \sqrt{\frac{(\Delta B)_{\text{measured}}}{B}} = 1.9 \times 10^{12} \text{ s}^{-1}$$

The reconstructed energy spectrum is shown in Fig. E.5.

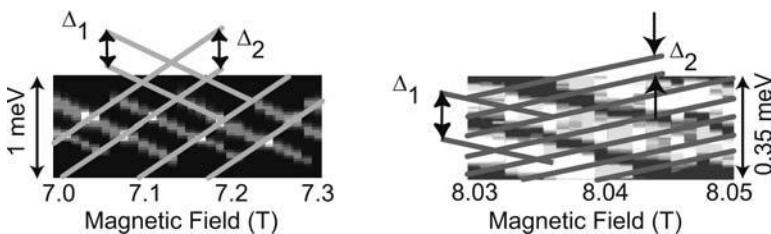


Fig. E.5 Left: The data of Fig. 10.8(c) with the single-electron charging energy removed, and the measured level spacings. Right: The corresponding reconstruction of the data in Fig. 10.9 shown for comparison.

E10.2

(a) From $\det(H - \lambda 1) = 0$, one finds

$$\lambda_{\pm} = \frac{1}{2}(H_{11} + H_{22}) \pm \frac{1}{2}\sqrt{(H_{11} - H_{22})^2 + 4H_{12}^2}$$

(b) For small transformation angles, the transformation matrix becomes

$$(O) = \begin{pmatrix} 1 & \alpha \\ -\alpha & 1 \end{pmatrix}$$

From $p(O^T H O) = p(H)$, the condition

$$p(H) = p(H) \left[1 - \alpha \left(2H_{12} \frac{d \ln p_{11}}{dH_{11}} - 2H_{12} \frac{d \ln p_{22}}{dH_{22}} - (H_{11} - H_{22}) \frac{d \ln p_{12}}{dH_{12}} \right) \right]$$

is obtained. Since α is arbitrary, this requires that the coefficient in front of α vanishes. The set of differential equations has the solution

$$p_{11}(H_{11}) = c_1 e^{-c_2 H_{11}^2 - c_3 H_{11}}$$

$$p_{12}(H_{12}) = c_1 e^{-2c_2 H_{12}^2}$$

$$p_{22}(H_{22}) = c_1 e^{-c_2 H_{22}^2 - c_3 H_{22}}$$

where the c_i are integration constants.

(c) Now $p(H)$ reads

$$p(H) = c_1 e^{-c_2(H_{11}^2 + 2H_{12}^2 + H_{22}^2) - c_3(H_{11} + H_{22})}$$

Choosing the energy reference such that $H_{11} + H_{22} = 0$, we see right away that this can be written as

$$p(H) = c_1 e^{-c_2 \text{Tr}(H^2)}$$

(d) Let β be the transformation angle that maps H onto a diagonal matrix via an orthogonal transformation, namely

$$ODO^T = H$$

with

$$(D) = \begin{pmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{pmatrix}$$

and

$$(O) = \begin{pmatrix} \cos \beta & \sin \beta \\ -\sin \beta & \cos \beta \end{pmatrix}$$

Note that here we cannot assume that a transformation by a small angle will do the job. This transformation gives the functional dependence of H_{ij} on λ_+ , λ_- , and α . Hence, the determinant of the Jacobian transformation matrix

$$J = \frac{\partial(H_{11}, H_{12}, H_{22})}{\partial(\lambda_+, \lambda_-, \beta)}$$

can be calculated, which gives

$$\det(J) = \lambda_+ - \lambda_-$$

In terms of the eigenvalues λ_\pm of H , $\text{Tr}(H^2)$ equals $\lambda_+^2 + \lambda_-^2$.

We can write

$$\begin{aligned} p(H) &= p(H_{11}, H_{12}, H_{22}) \\ &= p(\lambda_+, \lambda_-, \alpha) \det[J(H_{11}, H_{12}, H_{22}, \lambda_+, \lambda_-, \alpha)] \end{aligned}$$

We can therefore write

$$p(\lambda_+, \lambda_-) = c_1(\lambda_+ - \lambda_-) e^{-c_2(\lambda_+^2 + \lambda_-^2)}$$

We transform variables here and write

$$\Delta = \lambda_+ - \lambda_- \quad \text{and} \quad \Sigma = \lambda_+ + \lambda_-$$

such that

$$p(\Delta, \Sigma) = c_1 \Delta e^{-\frac{1}{2}c_2(\Delta^2 + \Sigma^2)}$$

Integration over Σ gives

$$p(\Delta) = \int_{-\infty}^{\infty} p(\Delta, \Sigma) d\Sigma = \sqrt{\frac{2\pi}{c_2}} c_1 \Delta e^{-\frac{1}{2}c_2 \Delta^2}$$

(e) The two normalization conditions read

$$\int_0^{\infty} p(\Delta) d\Delta = 1 \quad \text{and} \quad \int_0^{\infty} \Delta p(\Delta) d\Delta = 1$$

Hence, $c_1 = \pi/4$ and $c_2 = \pi/2$, which gives the corresponding expression in Table 10.1 for normalized peak spacings s .

E10.3 The qubit is an element of the two-dimensional, complex vector space with coordinates α and β . Each of them can be written in polar coordinates of a complex plane: $\alpha = |\alpha| \exp(i\phi_\alpha)$ and $\beta = |\beta| \exp(i\phi_\beta)$. We multiply the qubit by the global phase factor $\exp(-i\phi_\alpha)$ and obtain

$$|\psi\rangle = |\alpha| |0\rangle + e^{i\phi} |\beta| |1\rangle$$

with $\phi \equiv \phi_\beta - \phi_\alpha$. This can be rewritten as

$$|\psi\rangle = x|1\rangle + iy|1\rangle + z|0\rangle$$

with

$$x = |\beta| \cos \phi \equiv \sin \theta \cos \phi$$

$$y = |\beta| \sin \phi \equiv \sin \theta \sin \phi$$

$$z = |\alpha| \equiv \cos \theta$$

where $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$. The qubit is a point on the sphere given by $|x|^2 + |y|^2 + |z|^2 = 1$, and can be written as

$$|\psi\rangle = \cos \theta |0\rangle + \sin \theta e^{i\phi} |1\rangle$$

which is not quite the Bloch sphere representation yet. Note that, now, the equator consists of pure $|1\rangle$ states, up to a global phase factor, while north and south poles are pure $|0\rangle$ states. In fact, all states on the northern hemisphere can be found also on the southern hemisphere up to global phase factors. To remove this double representation, the convention is to map the sphere onto the Bloch sphere by the transformation $\theta \rightarrow \theta/2$ with the new θ running from 0 to π . This mapping is unique even for the equator since global phase factors do not matter.

E10.4 Suppose that the first Bell state $|B1\rangle$ can be constructed by tensor multiplication as $|a_1 a_2\rangle$ from

$$|a_i\rangle = \cos(\frac{1}{2}\theta_i) |0\rangle + e^{i\phi_i} \sin(\frac{1}{2}\theta_i) |1\rangle$$

with $i = 1, 2$. The coefficients have to obey

$$\cos(\frac{1}{2}\theta_1) \cos(\frac{1}{2}\theta_2) = 1/\sqrt{2} \quad (\text{E.1})$$

$$\cos(\frac{1}{2}\theta_1) \sin(\frac{1}{2}\theta_2) e^{i\phi_2} = 0 \quad (\text{E.2})$$

$$\cos(\frac{1}{2}\theta_2) \sin(\frac{1}{2}\theta_1) e^{i\phi_1} = 0 \quad (\text{E.3})$$

$$\sin(\frac{1}{2}\theta_1) \sin(\frac{1}{2}\theta_2) e^{i(\phi_1+\phi_2)} = 1/\sqrt{2} \quad (\text{E.4})$$

Since $e^{i\phi} \neq 0$ for all ϕ , we require from (E.2) and (E.3) that

$$\cos(\frac{1}{2}\theta_1) \sin(\frac{1}{2}\theta_2) = 0$$

and

$$\cos\left(\frac{1}{2}\theta_2\right) \sin\left(\frac{1}{2}\theta_1\right) = 0$$

However, none of the factors can be zero because of (E.1) or (E.4).

Chapter 12

E12.1 Our task now is to determine the difference of the spin-averaged electrochemical potentials $\mu_F(x \rightarrow 0)$ and $\mu_N(x \rightarrow 0)$ in both materials at the interface, from which we obtain the interface resistance R_I from

$$R_I = \frac{\mu_F(x \rightarrow 0) - \mu_N(x \rightarrow 0)}{ej}$$

Since

$$\begin{aligned} \mu_F(x \rightarrow 0) &= \frac{1}{\sigma_F} (\sigma_{F,\uparrow} \mu_{F,\uparrow} + \sigma_{F,\downarrow} \mu_{F,\downarrow}) \\ &= \alpha_F \mu_{F,\uparrow} + (1 - \alpha_F) \mu_{F,\downarrow} = \alpha_F \mu_S(0) + \mu_{F,\downarrow} \end{aligned}$$

and similarly

$$\mu_N(x \rightarrow 0) = \frac{1}{2} \mu_S(0) + \mu_{N,\downarrow}$$

we find, using the continuity of μ_\downarrow at the interface, that

$$\mu_F(x \rightarrow 0) - \mu_N(x \rightarrow 0) = ej \frac{\lambda_N}{\sigma_N} [1 - 2\beta(0)] (1 - 2\alpha_F)$$

Inserting $\beta(0)$ gives an interface resistance of

$$R_I = \frac{(2\alpha_F - 1)^2 \lambda_N / \sigma_N}{1 + 4\alpha_F (1 - \alpha_F) (\lambda_N / \sigma_N) (\sigma_F / \lambda_F)}$$

References

- 1** Y. Aharonov and D. Bohm, Phys. Rev. 115, 485 (1959)
- 2** B. L. Altshuler, A. G. Aronov, and B. Z. Spivak, JETP Lett. 33, 94 (1981)
- 3** B. L. Altshuler, A. G. Aronov, and D. E. Khmelnitsky, J. Phys. C 15, 7367 (1982)
- 4** V. Ambegaokar, B. I. Halperin, and J. S. Lange, Phys. Rev. B 4, 2612 (1971)
- 5** I. Amlani, A. O. Orlov, G. L. Snider, C. S. Lent, and G. H. Bernstein, Appl. Phys. Lett. 71, 1730 (1997)
- 6** M. Amman, R. Wilkins, E. Ben Jacob, P. D. Maker, and R. C. Jaklevic, Phys. Rev. B 43, 1146 (1991)
- 7** M. G. Ancona, J. Appl. Phys. 79, 526 (1996)
- 8** T. Ando, J. Phys. Soc. Jpn. 36, 959 (1974); ibid. 36, 1521 (1974); ibid. 37, 1044 (1974); Jpn. J. Appl. Phys. 2, 329 (1974)
- 9** T. Ando, Phys. Rev. B 13, 3468 (1976)
- 10** T. Ando, A. B. Fowler, and F. Stern, Rev. Mod. Phys. 54, 437 (1982)
- 11** E. L. Andronikashvili, J. Phys. USSR 10, 201 (1946) (in Russian); translated in *Hélium 4*, ed. A. Galasiewicz, Pergamon Press (1971)
- 12** N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing (1975)
- 13** D. V. Averin and K. K. Likharev, in *Mesoscopic Phenomena in Solids*, eds. B. L. Altshuler, P. A. Lee, and R. A. Webb, Elsevier, Oxford (1991)
- 14** D. V. Averin and Y. V. Nazarov, Phys. Rev. Lett. 65, 2446 (1990)
- 15** D. V. Averin and A. A. Odintsov, Phys. Lett. A 140, 251 (1989)
- 16** D. V. Averin, A. N. Korotkov, and K. K. Likharev, Phys. Rev. B 44, 6199 (1991)
- 17** D. D. Awschalom, D. Loss, and N. Samarth (eds.), *Semiconductor Spintronics and Quantum Computation*, Springer Series in NanoScience and Technology (2002)
- 18** A. Bächtold, M. Henny, C. Terrier, C. Strunk, C. Schönenberger, J.-P. Salvetat, J.-M. Bonard, and L. Forro, Appl. Phys. Lett. 73, 274 (1998)
- 19** M. N. Baibich, J. M. Broto, A. Fert, Van Dau Nguyen, F. Petroff, P. Etienne, G. Creuzet, A. Friedrich, and J. Chazelas, Phys. Rev. Lett. 61, 2472 (1988)
- 20** U. Banin, Y. Cao, D. Katz, and O. Millo, Nature 400, 542 (1999)
- 21** G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Les Ulis, France (1989)
- 22** C. W. J. Beenakker, Phys. Rev. Lett. 62, 2020 (1989)
- 23** C. W. J. Beenakker, Phys. Rev. B 44, 1646 (1991)
- 24** C. W. J. Beenakker, Rev. Mod. Phys. 69, 732 (1997)
- 25** C. W. J. Beenakker and H. van Houten, Phys. Rev. B 37, 6544 (1988)
- 26** C. W. J. Beenakker and H. van Houten, Phys. Rev. Lett. 63, 1857 (1989)
- 27** C. W. J. Beenakker and H. van Houten, in *Solid State Physics*, Vol. 44, pp. 1–228, eds. H. Ehrenreich and D. Turnbull, Academic Press (1991)
- 28** B. R. Bennett, M. J. Yang, B. V. Shanabrook, J. B. Boos, and D. Park, Appl. Phys. Lett. 72, 1193 (1998)
- 29** L. Berger, Phys. Rev. B 54, 9353 (1996)
- 30** K. F. Berggren, G. Roos, and H. van Houten, Phys. Rev. B 37, 10118 (1988)
- 31** D. Berman, N. B. Zhitenev, R. C. Ashoori, and M. Shayegan, Phys. Rev. Lett. 82, 161 (1999)
- 32** G. Bernstein and D. K. Ferry, J. Vac. Sci. Technol. B 5, 964 (1987)
- 33** C. Berthod, N. Binggeli, and A. Baldereschi, Phys. Rev. B 57, 9757 (1998)
- 34** S. K. Bhattacharya, Phys. Rev. B 25, 3756 (1982)
- 35** G. Binasch, P. Grünberg, F. Saurenbach, and W. Zinn, Phys. Rev. B 39, 4828 (1989)
- 36** G. Binnig and H. Rohrer, Helv. Phys. Acta 55, 726 (1982)
- 37** G. Binnig and H. Rohrer, Rev. Mod. Phys. 59, 615 (1987)
- 38** G. Binnig, H. Rohrer, Ch. Gerber, and E. Weibel, Phys. Rev. Lett. 50, 120 (1983)

- 39** M. Bockrath, D. H. Cobden, P. L. McEuen, N. G. Chopra, A. Zettl, A. Thess, and R. E. Smalley, *Science* 275, 1922 (1997)
- 40** M. Bockrath, D. H. Cobden, J. Lu, A. G. Rinzler, R. E. Smalley, L. Balents, and P. L. McEuen, *Nature* 397, 598 (1999)
- 41** O. Bohigas, M. J. Giannoni, and C. Schmit, *Phys. Rev. Lett.* 52, 1 (1984)
- 42** A. Bohr and B. R. Mottelson, in *Nuclear Structure*, W. A. Benjamin (1969)
- 43** C. R. Bolognesi, J. E. Bryce, and D. H. Chow, *Appl. Phys. Lett.* 69, 3531 (1996)
- 44** M. G. Burt, *J. Phys. Cond. Matter* 4, 6651 (1992)
- 45** M. Büttiker, *Phys. Rev. Lett.* 57, 1761 (1986)
- 46** M. Büttiker, *Phys. Rev. B* 33, 3020 (1986)
- 47** M. Büttiker, *IBM J. Res. Dev.* 32, 317 (1988)
- 48** M. Büttiker, *Phys. Rev. B* 41, 7906 (1990)
- 49** M. Büttiker, Y. Imry, R. Landauer, and S. Pinhas, *Phys. Rev. B* 31, 6207 (1985)
- 50** M. Cahay, M. McLennan, and S. Datta, *Phys. Rev. B* 37, 10125 (1988)
- 51** D. J. Chadi, *Phys. Rev. Lett.* 43, 43 (1979)
- 52** A. M. Chang, H. U. Baranger, L. N. Pfeiffer, K. W. West, and T. Y. Chang, *Phys. Rev. Lett.* 76, 1695 (1996)
- 53** D. B. Chklovskii, B. I. Chklovskii, and L. I. Glazman, *Phys. Rev. B* 46, 4026 (1992)
- 54** K. K. Choi, D. C. Tsui, and S. C. Palmateer, *Phys. Rev. B* 33, 8216 (1986)
- 55** K. K. Choi, D. C. Tsui, and K. Alavi, *Phys. Rev. B* 36, 7751 (1987)
- 56** M. Ciorga, A. S. Sachrajda, P. Hawrylak, C. Gould, P. Zawadzki, S. Jullian, Y. Feng, and Z. Wasilewski, *Phys. Rev. B* 61, R16315 (2000)
- 57** A. N. Cleland, J. M. Schmidt, and J. Clarke, *Phys. Rev. Lett.* 64, 1565 (1990)
- 58** D. H. Cobden and J. Nygard, *Phys. Rev. Lett.* 89, 046803 (2002)
- 59** C. P. Collier, G. Mattersteig, E. W. Wong, Y. Luo, K. Beverly, J. Sampaio, F. M. Raymo, J. F. Stoddart, and J. R. Heath, *Science* 289, 1172 (2000)
- 60** J. L. Costa-Kramer, *Phys. Today* 49, 9 (1996)
- 61** J. L. Costa-Kramer, *Phys. Rev. B* 55, 4875 (1997)
- 62** S. M. Cronenwett, H. J. Lynch, D. Goldhaber-Gordon, L. P. Kouwenhoven, C. M. Marcus, K. Hirose, N. S. Wingreen, and V. Umansky, *Phys. Rev. Lett.* 88, 226805 (2002)
- 63** J. A. Dagata, *Science* 270, 1625 (1990), and references therein
- 64** C. G. Darwin, *Proc. Cambr. Philos. Soc.* 27, 86 (1931)
- 65** S. Datta, *Electronic Transport in Mesoscopic Systems*, Cambridge University Press (1997)
- 66** S. Datta and B. Das, *Appl. Phys. Lett.* 56, 665 (1990)
- 67** D. Davidovic and M. Tinkham, *Phys. Rev. Lett.* 83, 1644 (1999)
- 68** S. G. Davidson and M. Steslicka, *Basic Theory of Surface States*, Oxford University Press (1992)
- 69** M. C. Desjonquieres and D. Spanjaard, *Concepts in Surface Physics*, 2nd edn., Springer (1998)
- 70** M. H. Devoret, D. Esteve, H. Grabert, G.-L. Ingold, H. Pothier, and C. Urbina, *Phys. Rev. Lett.* 64, 1824 (1990)
- 71** T. Dietl, H. Ohno, F. Matsukura, J. Cibert, and D. Ferrand, *Science* 287, 1019 (2000)
- 72** R. Dingle, H. L. Störmer, A. C. Gossard, and W. Wiegmann, *Appl. Phys. Lett.* 33, 665 (1978)
- 73** M. Di Ventra, S. T. Pantelides, and N. D. Lang, *Phys. Rev. Lett.* 84, 979 (2000)
- 74** G. Dolan, *Appl. Phys. Lett.* 31, 337 (1977)
- 75** Z. J. Donhauser, B. A. Mantooth, K. F. Kelly, L. A. Bumm, J. D. Monnell, J. J. Stapleton, D. W. Price, A. M. Rawlett, D. L. Allara, J. M. Zour, and P. S. Weiss, *Science* 292, 2303 (2001)
- 76** P. van Dorpe, Z. Liu, W. Van Roy, V. F. Motsnyi, M. Sawicki, G. Borghs, and J. De Boeck, *Appl. Phys. Lett.* 84, 3495 (2004)
- 77** P. D. Dresselhaus, C. M. A. Papavassiliou, R. G. Wheeler, and R. N. Sacks, *Phys. Rev. Lett.* 68, 106 (1992)
- 78** H. Drexler, PhD Thesis, LMU Munich (1994)
- 79** H. Drexler, D. Leonhard, W. Hansen, J. P. Kotthaus, and P. M. Petroff, *Phys. Rev. Lett.* 73, 2252 (1994)

- 80** M. I. Dyakonov and V. I. Perel, Sov. Phys. JETP 33, 1053 (1971)
- 81** D. J. Eaglesham and M. Cerullo, Phys. Rev. Lett. 64, 1943 (1990)
- 82** D. M. Eigler and E. K. Schweizer, Nature 344, 524 (1990)
- 83** G. T. Einevoll and L. J. Sham, Phys. Rev. B 49, 10533 (1994)
- 84** R. J. Elliot, Phys. Rev. 96, 266 (1954)
- 85** L. Esaki and L. L. Chang, Phys. Rev. Lett. 33, 495 (1974)
- 86** A. K. Evans, L. I. Glazman, and B. I. Shklovskii, Phys. Rev. B 48, 11120 (1993)
- 87** Z. F. Ezawa, *Quantum Hall Effects: Field Theoretical Approach and Related Topics*, World Scientific (2000)
- 88** G. R. Facer, B. E. Kane, R. G. Clark, L. N. Pfeiffer, and K. W. West, Phys. Rev. B 56, 10036 (1997)
- 89** F. F. Fang and W. E. Howard, Phys. Rev. Lett. 16, 797 (1966)
- 90** D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures*, Cambridge University Press (1997)
- 91** A. Fert and I. A. Campbell, J. Phys. F 6, 849 (1976)
- 92** M. Field, C. Smith, M. Pepper, D. A. Ritchie, J. E. F. Frost, G. A. C. Jones, and D. G. Hasko, Phys. Rev. Lett. 70, 1311 (1993)
- 93** C. M. Fischer, M. Burghard, S. Roth, and K. von Klitzing, Europhys. Lett. 28, 129 (1994)
- 94** D. S. Fisher and P. A. Lee, Phys. Rev. B 23, 6851 (1981)
- 95** G. Fishman and G. Lampel, Phys. Rev. B 16, 820 (1977)
- 96** R. J. Fitzgerald, S. L. Pohlen, and M. Tinkham, Phys. Rev. B 57, 11073 (1998)
- 97** R. Fleischmann, T. Geisel, and R. Ketzmerick, Phys. Rev. Lett. 68, 1367 (1992)
- 98** R. Fleischmann, T. Geisel, and R. Ketzmerick, Europhys. Lett. 25, 219 (1994)
- 99** V. Fock, Z. Phys. 47, 446 (1928)
- 100** J. A. Folk, S. R. Patel, S. F. Godijn, A. G. Huibers, S. M. Cronenwett, C. M. Marcus, K. Campman, and A. C. Gossard, Phys. Rev. Lett. 76, 1699 (1996)
- 101** S. Franco, *Design with Operational Amplifiers and Analog Integrated Circuits*, McGraw-Hill (1997)
- 102** F. C. Frank and J. H. van der Merve, Proc. R. Soc. London A 198, 205 (1949)
- 103** S. Frank, P. Poncharal, Z. L. Wang, and W. de Heer, Science 280, 1744 (1998)
- 104** W. R. Frensley and H. Kroemer, Phys. Rev. B 16, 2642 (1977)
- 105** T. M. Fromhold, P. B. Wilkinson, F. W. Sheard, L. Eaves, J. Miao, and G. Edwards, Phys. Rev. Lett. 75, 1142 (1995)
- 106** A. Fuhrer, S. Lüscher, T. Heinzel, K. Ensslin, W. Wegscheider, and M. Bichler, Phys. Rev. B 63, 125309 (2001)
- 107** A. Fuhrer, S. Lüscher, T. Ihn, T. Heinzel, K. Ensslin, W. Wegscheider, and M. Bichler, Nature 413, 822 (2001)
- 108** T. A. Fulton and G. J. Dolan, Appl. Phys. Lett. 42, 752 (1983)
- 109** T. A. Fulton and G. J. Dolan, Phys. Rev. Lett. 59, 109 (1987)
- 110** M. Furlan, T. Heinzel, B. Jeanneret, S. V. Lothkov, and K. Ensslin, Europhys. Lett. 49, 369 (2000)
- 111** M. Furlan, T. Heinzel, B. Jeanneret, and S. V. Lothkov, J. Low Temp. Phys. 118, 297 (2000)
- 112** F. Geerinckx, F. M. Peeters, and J. T. Devreese, J. Appl. Phys. 68, 3435 (1990)
- 113** L. J. Geerligs, V. F. Anderegg, P. A. M. Holweg, J. E. Mooij, H. Pothier, D. Esteve, C. Urbina, and M. H. Devoret, Phys. Rev. Lett. 64, 2691 (1990)
- 114** M. C. Geisler, J. H. Smet, V. Umansky, K. von Klitzing, B. Naundorf, R. Ketzmerick, and H. Schweizer, Phys. Rev. Lett. 92, 256801 (2004)
- 115** R. R. Gerhardts, D. Weiss, and K. von Klitzing, Phys. Rev. Lett. 62, 1173 (1989)
- 116** S. K. Ghandhi, *VLSI Fabrication Principles: Silicon and Gallium Arsenide*, Wiley-Interscience (1994)
- 117** I. Giaever and H. R. Zeller, Phys. Rev. Lett. 20, 1504 (1968)
- 118** P. Giannozzi, S. de Gironcoli, P. Pavone, and S. Baroni, Phys. Rev. B 43, 7231 (1991)
- 119** G. F. Giuliani and J. J. Quinn, Phys. Rev. B 26, 4421 (1982)

- 120** L. I. Glazman, G. B. Lesovik, D. E. Khmel'nitsky, and R. I. Shekter, JETP Lett. 48, 238 (1988)
- 121** E. T. Goodwin, Proc. Cambr. Philos. Soc. 35, 205 (1939)
- 122** E. T. Goodwin, Proc. Cambr. Philos. Soc. 35, 221 (1939)
- 123** C. J. Gorter, Physica 17, 777 (1951)
- 124** H. Grabert and M. Devoret, *Single Charge Tunneling*, NATO ASI, Ser. B, Vol. 294, Plenum Press (1992)
- 125** H. Grabert, G.-L. Ingold, M. H. Devoret, D. Esteve, H. Pothier, and C. Urbina, Z. Phys. B, Cond. Matter 84, 143 (1991)
- 126** H. Grahn, *Introduction to Semiconductor Physics*, World Scientific (1999)
- 127** G. Grossi and G. Parravicini, *Solid State Physics*, Academic Press (2000)
- 128** P. Grünberg, R. Schreiber, Y. Pang, M. B. Brodsky, and H. Sower, Phys. Rev. Lett. 57, 2442 (1986)
- 129** G. Grüner and M. Minier, Adv. Phys. 26, 231 (1977)
- 130** S. Gueron, M. M. Desmukh, E. B. Myers, and D. C. Ralph, Phys. Rev. Lett. 83, 4148 (1999)
- 131** L. Gurevich, L. Canali, and L. P. Kouwenhoven, Appl. Phys. Lett. 76, 384 (2000)
- 132** M. C. Gutzwiller, J. Math. Phys. 12, 343 (1971); see also W. H. Miller, J. Chem. Phys. 63, 996 (1975)
- 133** M. C. Gutzwiller, *Chaos in Quantum and Classical Mechanics*, Springer (1990)
- 134** F. Haake, *Quantum Signatures of Chaos*, Springer (1991)
- 135** H. Häffner, W. Hänsel, C. F. Roos, J. Benhelm, D. Chek-al-kar, M. Chwalla, T. Köber, U. D. Rapol, M. Riebe, P. O. Schmidt, C. Becher, O. Gühne, W. Dür, and R. Blatt, Nature 438, 643 (2005)
- 136** J. Hajdu (ed.), *Introduction to the Theory of the Integer Quantum Hall Effect*, Wiley-VCH (1994)
- 137** B. I. Halperin, Phys. Rev. B 25, 2185 (1982)
- 138** N. Hamada, S. Sawada, and A. Oshiyama, Phys. Rev. Lett. 68, 1579 (1992)
- 139** A. T. Hanbicki, B. T. Jonker, G. Itskos, G. Kioseglou, and A. Petrou, Appl. Phys. Lett. 80, 1240 (2002)
- 140** A. E. Hanna and M. Tinkham, Phys. Rev. B 44, 5919 (1991)
- 141** P. Harris, *Carbon Nanotubes and Related Structures*, Cambridge University Press (1999)
- 142** W. A. Harrison, *Elementary Electronic Structure*, World Scientific, Singapore (1999)
- 143** T. Hattori (ed.), *Ultraclean Surface Processing of Silicon Wafers*, Springer (1998)
- 144** R. J. Haug, R. R. Gerhardts, K. von Klitzing, and K. Ploog, Phys. Rev. Lett. 59, 1349 (1987)
- 145** R. J. Haug, J. Kucera, P. Streda, and K. von Klitzing, Phys. Rev. B 39, 10892 (1989)
- 146** E. Hecht, *Optics*, Addison Wesley Longman (2002)
- 147** V. Heine, Phys. Rev. 138, A1689 (1965)
- 148** T. Heinzel, G. Salis, R. Held, S. Lüscher, K. Ensslin, W. Wegscheider, and M. Bichler, Phys. Rev. B 61, 13353 (2000)
- 149** R. Held, T. Vancura, T. Heinzel, K. Ensslin, M. Holland, and W. Wegscheider, Appl. Phys. Lett. 73, 262 (1998)
- 150** J. Heurich, J. C. Cuevas, W. Wenzel, and G. Schön, Phys. Rev. Lett. 88, 256803 (2002)
- 151** D. Hofstadter, Phys. Rev. B 14, 2239 (1976)
- 152** D. S. Hopkins, D. Pekker, P. M. Goldbart, and A. Bezryadin, Science 308, 1762 (2005)
- 153** G. Horowitz, R. Hajlaoui, D. Fichou, and A. El Kassmi, J. Appl. Phys. 85, 3202 (1999)
- 154** P. Horowitz and W. Hill, *The Art of Electronics*, Cambridge University Press (1989)
- 155** H. van Houten and C. W. J. Beenakker, Phys. Rev. Lett. 63, 1893 (1989)
- 156** H. van Houten, J. G. Williamson, M. E. I. Broekaart, C. T. Foxon, and J. J. Harris, Phys. Rev. B 37, 2756 (1988)
- 157** H. van Houten, C. W. J. Beenakker, P. H. M. van Loosdrecht, T. J. Thornton, H. Ahmed, M. Pepper, C. T. Foxon, and J. J. Harris, Phys. Rev. B 37, 8534 (1988)
- 158** H. van Houten, C. W. J. Beenakker, J. G. Williamson, M. E. I. Broekaart, P. H. M. van Loosdrecht, B. J. van Wees, J. E. Mooij, C. T. Foxon, and J. J. Harris, Phys. Rev. B 39, 8556 (1989)

- 159** H. van Houten, C. W. J. Beenakker, and B. J. van Wees, in *Semiconductors and Semimetals*, Vol. 35, pp. 9–112, eds. H. Ehrenreich and H. Turnbull, Academic Press (1992)
- 160** S. Iijima, Nature 354, 56 (1991)
- 161** S. Ilani, A. Yacoby, D. Mahalu, and H. Shtrikman, Science 292, 1354 (2001)
- 162** B. Irmer, R. H. Blick, F. Simmel, W. Gödel, H. Lorenz, and J. P. Kotthaus, Appl. Phys. Lett. 73, 2051 (1998)
- 163** K. Ismail, M. Arafa, K. L. Saenger, J. O. Chu, and B. S. Meyerson, Appl. Phys. Lett. 66, 1077 (1995)
- 164** A. van Itterbeck and L. de Greeve, Experimentia 3, No. 7 (1947)
- 165** L. Jacak, P. Hawrylak, and A. Wojs, *Quantum Dots*, Springer (1998)
- 166** C. Jacoboni and L. Reggiani, Rev. Mod. Phys. 55, 645 (1983)
- 167** B. Jeckelmann, B. Jeanneret, and D. Inglis, Phys. Rev. B 55, 13124 (1997)
- 168** F. Jedema, A. T. Filip, and B. J. van Wees, Nature 410, 345 (2001)
- 169** A. T. Johnson, L. P. Kouwenhoven, W. de Jong, N. C. van der Vaart, C. J. P. M. Harmans, and C. T. Foxon, Phys. Rev. Lett. 69, 1592 (1992)
- 170** M. J. M. de Jong, Phys. Rev. B 49, 7778 (1994)
- 171** M. Jullière, Phys. Lett. 54A, 225 (1975)
- 172** M. W. Keller, A. L. Eichenberger, J. M. Martinis, and N. W. Zimmerman, Science 285, 1706 (1999)
- 173** U. Klass, Z. Phys. B 82, 351 (1991)
- 174** D. L. Klein, P. L. McEuen, J. E. Bowen Katari, R. Roth, and A. P. Alivisatos, Appl. Phys. Lett. 68, 2574 (1996)
- 175** D. L. Klein, R. Roth, A. P. Alivisatos, A. K. Lim, and P. L. McEuen, Nature 389, 699 (1997)
- 176** K. von Klitzing, G. Dorda, and M. Pepper, Phys. Rev. Lett. 45, 494 (1980)
- 177** R. Knott, Solid State Electron. 37, 689 (1994)
- 178** A. N. Korotkov, in *Molecular Electronics*, eds. J. Jortner and M. A. Ratner, Blackwell (1996)
- 179** A. N. Korotkov, Appl. Phys. Lett. 72, 3226 (1998)
- 180** L. P. Kouwenhoven, C. M. Marcus, P. L. McEuen, S. Tarucha, R. M. Westervelt, and N. S. Wingreen, in *Mesoscopic Electron Transport*, pp. 105–214, NATO ASI, Ser. E, Vol. 345, eds. L. P. Kouwenhoven, G. Schön, and L. L. Sohn, Kluwer (1997)
- 181** R. Kubo, J. Phys. Soc. Jpn. 12, 570 (1957)
- 182** R. Kubrak, A. Neumann, B. L. Gallagher, P. C. Main, M. Henini, C. H. Marrows, and B. J. Hickey, J. Appl. Phys. 87, 5986 (2000)
- 183** U. Kuhl and H.-J. Stöckmann, Phys. Rev. Lett. 80, 3232 (1998)
- 184** I. O. Kulik and R. I. Shekter, Sov. Phys. JETP 41, 308 (1975)
- 185** A. Kumar, S. E. Laux, and F. Stern, Phys. Rev. B 42, 5166 (1990)
- 186** R. Landauer, IBM J. Res. Dev. 1, 223 (1957)
- 187** R. B. Laughlin, Phys. Rev. B 23, 5632 (1981)
- 188** M. L. Leadbeater, C. L. Foden, J. H. Burroughes, M. Pepper, T. M. Burke, L. L. Wang, M. P. Grimshaw, and D. A. Ritchie, Phys. Rev. B 52, R8629 (1995)
- 189** P. Lee, Physica 140A, 169 (1986)
- 190** D. Leonard, M. Krishnamurthy, C. M. Reaves, S. P. Denbaars, and P. M. Petroff, Appl. Phys. Lett. 63, 3203 (1993)
- 191** K. K. Likharev, IBM J. Res. Dev. 32, 144 (1988)
- 192** F. London, Nature 141, 643 (1938)
- 193** S. V. Lotkhov, S. A. Bogoslovsky, A. B. Zorin, and J. Niemeyer, Appl. Phys. Lett. 78, 946 (2001)
- 194** S. G. Louie and M. L. Cohen, Phys. Rev. B 15, 2154 (1977)
- 195** O. V. Lounasmaa, *Experimental Principles and Methods Below 1 K*, Academic Press (1974)
- 196** B. Ludoph and J. M. van Ruitenbeek, Phys. Rev. B 61, 2273 (2000)
- 197** S. Lüscher, T. Heinzel, K. Ensslin, W. Wegscheider, and M. Bichler, Phys. Rev. Lett. 86, 2118 (2001)
- 198** J. C. Maan, in *Two Dimensional Systems, Heterostructures, and Superlattices*, p. 183, eds. G. Bauer, F. Kuchar, and H. Heinrich, Springer Verlag, Berlin (1984)

- 199** F. A. Maaø, I. V. Zozulenko, and E. H. Hauge, Phys. Rev. B 50, 17320 (1994)
- 200** D. Mailly, C. Chapelier, and A. Benoit, Phys. Rev. Lett. 70, 2020 (1993)
- 201** H. J. Mamin, P. H. Guethner, and D. Rugar, Phys. Rev. Lett. 65, 2418 (1990)
- 202** C. R. K. Marrian (ed.), *Technology of Proximal Probe Lithography*, SPIE Optical Engineering Press, Bellingham, WA (1993)
- 203** J. M. Martinis, M. Nahum, and H. D. Jensen, Phys. Rev. Lett. 72, 904 (1994)
- 204** J.-Y. Marzin and J.-M. Gerard, Phys. Rev. Lett. 62, 2172 (1989)
- 205** A. W. Maue, Z. Phys. 94, 717 (1935)
- 206** P. V. E. McClintock, D. J. Meredith, and J. K. Wigmore, *Matter at Low Temperatures*, Blackie and Son (1984)
- 207** P. L. McEuen, E. B. Foxman, J. Kinaret, U. Meirav, M. A. Kastner, N. S. Wingreen, and S. J. Wind, Phys. Rev. B 45, 11419 (1992)
- 208** W. H. Meiklejohn and C. P. Bean, Phys. Rev. 105, 904 (1957)
- 209** U. Meirav, M. A. Kastner, M. Heiblum, and S. J. Wind, Phys. Rev. B 40, 5871 (1989)
- 210** U. Meirav, M. A. Kastner, and S. J. Wind, Phys. Rev. Lett. 65, 771 (1990)
- 211** M. L. Mehta, *Random Matrices*, Academic Press (1991)
- 212** L. W. Molenkamp, A. A. M. Staring, C. W. J. Beenakker, R. Eppenga, C. E. Timmering, J. G. Williamson, C. J. P. M. Harmans, and C. T. Foxon, Phys. Rev. B 41, 1274 (1990)
- 213** W. Mönch, *Semiconductor Surfaces and Interfaces*, Springer (2001)
- 214** E. A. Montie, E. C. Cosman, G. W. 't Hooft, M. B. van der Mark, and C. W. J. Beenakker, Nature 350, 594 (1991)
- 215** F. G. Monzon, M. Johnson, and M. L. Roukes, Appl. Phys. Lett. 71, 3087 (1997)
- 216** J. S. Moodera, L. R. Kinder, T. M. Wong, and R. Meservey, Phys. Rev. Lett. 74, 3273 (1995)
- 217** J. S. Moon, J. A. Simmons, and J. L. Reno, Appl. Phys. Lett. 71, 656 (1997)
- 218** G. Müller, D. Weiss, A. V. Khaetskii, K. von Klitzing, S. Koch, H. Nickel, W. Schlapp, and R. Lösch, Phys. Rev. B 45, 3932 (1992)
- 219** M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press (2000)
- 220** F. Nihey, S. W. Hwang, and K. Nakamura, Phys. Rev. B 51, 4649 (1995)
- 221** J. Nitta, T. Akazaki, H. Takayanagi, and T. Enoki, Phys. Rev. Lett. 78, 1336 (1997)
- 222** J. Nogués and I. K. Schuller, J. Magn. Magn. Mater. 192, 203 (1999)
- 223** K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, M. I. Katsnelson, and I. V. Grigorieva, Nature 438, 197 (2005)
- 224** T. W. Odom, J. Huang, P. Kim, and C. M. Lieber, Nature 391, 62 (1998)
- 225** H. Ohno, Science 281, 951 (1998)
- 226** H. Ohno, D. Chiba, F. Matsukura, T. Omiya, E. Abe, T. Dietl, Y. Ohno and K. Ohtani, Nature 408, 944 (2000)
- 227** J. Okabayashi, A. Kimura, O. Rader, T. Mizokawa, A. Fujimori, T. Hayashi, and M. Tanaka, Phys. Rev. B 58, R4211 (1998)
- 228** G. S. Painter and D. E. Ellis, Phys. Rev. B 1, 4747 (1970)
- 229** R. G. Palmer, Adv. Phys. 31, 669 (1982)
- 230** H. Park, A. K. Lim, A. P. Alivisatos, J. Park, and P. L. McEuen, Appl. Phys. Lett. 75, 301 (1999)
- 231** H. Park, J. Park, A. K. Lim, E. H. Anderson, A. P. Alivisatos, and P. L. McEuen, Nature 407, 57 (2000)
- 232** G. H. Parker and C. A. Mead, Phys. Rev. Lett. 21, 605 (1968)
- 233** S. R. Patel, S. M. Cronenwett, D. R. Stewart, A. G. Huibers, C. M. Marcus, C. I. Duruöz, J. S. Harris, K. Campman, and A. C. Gossard, Phys. Rev. Lett. 80, 4522 (1998)
- 234** L. Pauling, *The Nature of the Chemical Bond*, Cornell University Press (1960)
- 235** L. J. van der Pauw, Philips Res. Rep. No. 13, 1–9 (1958)
- 236** S. Pedersen, A. E. Hansen, A. Kristensen, C. B. Sørensen, and P. E. Lindelof, Phys. Rev. B 61, 5457 (2000)
- 237** P. M. Petroff, R. C. Miller, A. C. Gossard, and W. Wiegmann, Appl. Phys. Lett. 44, 217 (1984)

- 238** J. R. Petta and D. C. Ralph, Phys. Rev. Lett. 87, 266801 (2001)
- 239** D. Pfannkuche and R. R. Gerhardts, Phys. Rev. B 46, 12606 (1992)
- 240** L. Pfeiffer, K. W. West, H. L. Stormer, and K. W. Baldwin, Appl. Phys. Lett. 55, 1888 (1989)
- 241** R. de Picciotto, H. L. Stormer, L. N. Pfeiffer, K. W. Baldwin, and K. W. West, Nature 411, 52 (2001)
- 242** M. Pinczolits, G. Springholz, and G. Bauer, Appl. Phys. Lett. 73, 250 (1998)
- 243** G. Pikus and A. Titkov, in *Optical Orientation, Modern Problems in Condensed Matter Science*, Vol. 8, eds. F. Meier and B. Zakharchenya, North-Holland, Amsterdam (1984)
- 244** D. Porath, A. Bezryadin, S. de Vries, and C. Dekker, Nature 403, 635 (2000)
- 245** H. Pothier, P. Lafarge, C. Urbina, D. Esteve, and M. H. Devoret, Europhys. Lett. 17, 249 (1992)
- 246** R. E. Prange and S. M. Girvin (eds.), *The Quantum Hall Effect*, Springer (1990)
- 247** T. Quinn, Metrologia 26, 69 (1989)
- 248** D. C. Ralph, C. T. Black, and M. Tinkham, Phys. Rev. Lett. 74, 3241 (1995)
- 249** E. I. Rashba, Phys. Rev. B 62, 16267 (2000)
- 250** M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour, Science 278, 252 (1997)
- 251** M. A. Reed, J. Chen, A. M. Rawlett, D. W. Price, and J. M. Tour, Appl. Phys. Lett. 78, 3735 (2001)
- 252** F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw-Hill (1985)
- 253** R. C. Richardson and E. N. Smith, *Experimental Techniques in Condensed Matter Physics at Low Temperatures*, Addison-Wesley (1988)
- 254** B. K. Ridley, *Quantum Processes in Semiconductors*, Oxford University Press (1999)
- 255** A. C. Rose-Innes, *Low Temperature Laboratory Techniques*, English University Press (1973)
- 256** J. Rychen, T. Vancura, T. Heinzel, R. Schuster, and K. Ensslin, Phys. Rev. B 58, 3568 (1998)
- 257** R. Saito, M. Fujita, G. Dresselhaus, and M. S. Dresselhaus, Appl. Phys. Lett. 60, 2204 (1992)
- 258** R. Saito, G. Dresselhaus, and M. S. Dresselhaus, *Physical Properties of Carbon Nanotubes*, Imperial College Press (1998)
- 259** G. Salis, B. Graf, K. Ensslin, K. Campman, K. Maranowski, and A. C. Gossard, Phys. Rev. Lett. 79, 5106 (1997)
- 260** G. Salis, B. Ruhstaller, K. Ensslin, K. Campman, K. Maranowski, and A. C. Gossard, Phys. Rev. B 58, 1436 (1998)
- 261** G. Salis, P. Wirth, T. Heinzel, T. Ihn, K. Ensslin, K. Maranowski, and A. C. Gossard, Phys. Rev. B 59, R5304 (1999)
- 262** A. Saxler, P. Debray, R. Perrin, S. Elhamri, W. C. Mitchel, C. R. Elsass, I. P. Smorchkova, B. Heying, E. Haus, P. Fini, J. P. Ibbetson, S. Keller, P. M. Petroff, S. P. DenBaars, U. K. Mishra, and J. S. Speck, J. Appl. Phys. 87, 369 (2000)
- 263** E. Scheer, P. Joyez, D. Esteve, C. Urbina, and M. H. Devoret, Phys. Rev. Lett. 78, 3535 (1997)
- 264** A. G. Scherbakov, E. N. Bogachev, and U. Landman, Phys. Rev. B 53, 4054 (1996)
- 265** K. Schönhammer and V. Meden, Am. J. Phys. 64, 1168 (1996)
- 266** W. Schottky, Naturwissenschaften 26, 843 (1938)
- 267** B. Schuh, J. Phys. A 18, 803 (1985)
- 268** R. Schuster, K. Ensslin, J. P. Kotthaus, G. Böhm, and W. Klein, Phys. Rev. B 55, 2237 (1997)
- 269** J. H. F. Scott-Thomas, S. B. Field, M. A. Kastner, H. I. Smith, and D. A. Antoniadis, Phys. Rev. Lett. 62, 583 (1989)
- 270** K. Seeger, *Semiconductor Physics, An Introduction*, Springer Series in Solid State Sciences (1997)
- 271** V. Senz, PhD Thesis, ETH Zürich (2002)
- 272** D. Y. Sharvin and Y. V. Sharvin, JETP Lett. 34, 272 (1981)
- 273** K. L. Shepard, M. L. Roukes, and B. P. van der Gaag, Phys. Rev. Lett. 68, 2660 (1992)
- 274** R. Shioda, K. Ando, T. Hayashi, and M. Tanaka, Phys. Rev. B 58, 1100 (1998)
- 275** J. Shirakashi, K. Matsumoto, N. Miura, and M. Konagai, Appl. Phys. Lett. 72, 1893 (1998)

- 276** W. Shockley, Phys. Rev. 56, 3175 (1939)
- 277** F. Simmel, T. Heinzel, and D. A. Wharam, Europhys. Lett. 38, 123 (1997)
- 278** S. Simon and B. Halperin, Phys. Rev. Lett. 73, 3278 (1994)
- 279** K. Simonyi, *Foundations of Electrical Engineering*, Pergamon Press (1963)
- 280** U. Sivan, R. Berkovits, Y. Aloni, O. Prus, A. Auerbach, and G. Ben-Yoseph, Phys. Rev. Lett. 77, 1123 (1996)
- 281** R. M. H. Smit, Y. Noat, C. Untiedt, N. D. Lang, M. C. van Hemert, and J. M. van Ruitenbeek, Nature 419, 906 (2002)
- 282** T. P. Smith, W. I. Wang, and P. J. Stiles, Phys. Rev. B 34, 2995 (1986)
- 283** P. C. van Son, H. van Kempen, and P. Wyder, Phys. Rev. Lett. 58, 2271 (1987)
- 284** P. Song and K. Kim, Phys. Rev. B 66, 035207 (2002)
- 285** Special Issue on Semiconductor Spintronics, Semicond. Sci. Technol. 17, No. 4 (2002)
- 286** G. Springholz, V. Holy, M. Pinczolits, and G. Bauer, Science 282, 734 (1998)
- 287** D. Springsguth, R. Ketzmerick, and T. Geisel, Phys. Rev. B 56, 2036 (1997)
- 288** D. Stauffer and A. Aharony, *Introduction to Percolation Theory*, Wiley-VCH (1995)
- 289** G. E. Stillman and C. M. Wolfe, Thin Solid Films 31, 69 (1976)
- 290** H. Stormer, Solid State Commun. 84, 95 (1992)
- 291** H. Stormer, A. C. Gossard, and W. Wiegmann, Solid State Commun. 41, 707 (1982)
- 292** H. L. Stormer, J. P. Eisenstein, A. C. Gossard, W. Wiegmann, and K. Baldwin, Phys. Rev. Lett. 56, 85 (1986)
- 293** I. N. Stranski and L. Krastanov, Akad. Wiss. Lit. Mainz Math.-Naturwiss. Kl. IIb 146, 797 (1939)
- 294** S. M. Sze, *Semiconductor Devices: Physics and Technology*, Wiley, New York (1985)
- 295** I. E. Tamm, Phys. Z. Sowjetunion 1, 733 (1932)
- 296** S. J. Tans, M. H. Devoret, H. Dai, A. Thess, R. E. Smalley, L. J. Geerligs, and C. Dekker, Nature 386, 474 (1997)
- 297** S. Tarucha, D. G. Austing, T. Honda, R. J. van der Hage, and L. P. Kouwenhoven, Phys. Rev. Lett. 77, 3613 (1996)
- 298** J. M. Taylor, H.-A. Engel, W. Dür, A. Yacobi, P. Zoller, and M. D. Lukin, Nature Phys. 1, 177 (2005)
- 299** K. J. Thomas, J. T. Nicholls, M. Y. Simmons, M. Pepper, D. R. Mace, and D. A. Ritchie, Phys. Rev. Lett. 77, 135 (1996)
- 300** T. J. Thornton, M. L. Roukes, A. Scherer, and B. P. van de Gaag, Phys. Rev. Lett. 63, 2128 (1989)
- 301** D. J. Thouless, J. Phys. C 14, 3475 (1981)
- 302** D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. de Nijs, Phys. Rev. Lett. 49, 405 (1982)
- 303** G. Timp, Surf. Sci. 196, 68 (1988)
- 304** M. A. Topinka, B. J. LeRoy, S. E. J. Shaw, E. J. Heller, R. M. Westervelt, K. D. Maranowski, and A. C. Gossard, Science 289, 2323 (2000)
- 305** D. C. Tsui, H. L. Stormer, and A. C. Gossard, Phys. Rev. Lett. 48, 1559 (1982)
- 306** D. Ullmo and H. Baranger, Phys. Rev. B 64, 245324 (2001), and references therein
- 307** V. Umansky, R. de Picciotto, and M. Heiblum, Appl. Phys. Lett. 71, 683 (1997)
- 308** C. P. Umbach, C. Van Haesendonck, R. B. Laibowitz, S. Washburn, and R. A. Webb, Phys. Rev. Lett. 56, 386 (1986)
- 309** R. Ursin, T. Jennewein, M. Aspelmeyer, R. Kaltenbaek, M. Lindenthal, P. Walther, and A. Zeilinger, Nature 430, 849 (2004)
- 310** T. Vancura, T. Ihn, S. Broderick, K. Ensslin, W. Wegscheider, and M. Bichler, Phys. Rev. B 62, 5074 (2000)
- 311** M. Volmer and A. Weber, Z. Phys. Chem. 119, 277 (1926)
- 312** D. K. de Vries and A. D. Wieck, Am. J. Phys. 63, 1074 (1995)
- 313** P. R. Wallace, Phys. Rev. 71, 622 (1947)
- 314** W. Walukiewicz, Phys. Rev. B 30, 4571 (1984)
- 315** A. C. Warren, IEEE Electron. Dev. Lett. 6, 294 (1985)
- 316** R. A. Webb, S. Washburn, C. P. Umbach, and R. B. Laibowitz, Phys. Rev. Lett. 54, 2696 (1985)
- 317** B. J. van Wees, H. van Houten, C. W. J. Beenakker, J. G. Williamson, L. P. Kouwenhoven, D. van der Marel, and C. T. Foxon, Phys. Rev. Lett. 60, 848 (1988)

- 318** B. J. van Wees, L. P. Kouwenhoven, H. van Houten, C. W. J. Beenakker, J. E. Mooij, C. T. Foxon, and J. J. Harris, Phys. Rev. B 38, 3625 (1988)
- 319** W. Wegscheider, in *Optics of Semiconductor Quantum Wires and Dots*, ed. G. W. Bryant, Gordon and Breach Science Publishing (1998)
- 320** W. Wegscheider, G. Schedelbeck, G. Abstreiter, M. Rother, and M. Bichler, Phys. Rev. Lett. 79, 1917 (1997)
- 321** Y. Y. Wei, J. Weis, K. von Klitzing, and K. Eberl, Appl. Phys. Lett. 71, 2514 (1997)
- 322** D. Weiss, K. von Klitzing, K. Ploog, and G. Weimann, Europhys. Lett. 8, 179 (1989)
- 323** D. Weiss, M. L. Roukes, A. Menschig, P. Grambow, K. von Klitzing, and G. Weimann, Phys. Rev. Lett. 66, 2790 (1991)
- 324** D. Weiss, K. Richter, A. Menschig, R. Bergmann, H. Schweizer, K. von Klitzing, and G. Weimann, Phys. Rev. Lett. 70, 4118 (1993)
- 325** N. H. Weste and K. Eshraghian, *Principles of VLSI Design*, Addison Wesley, Reading, MA (1994)
- 326** D. A. Wharam, T. J. Thornton, R. Newbury, M. Pepper, H. Ahmed, J. E. F. Frost, D. G. Hasko, D. C. Peacock, D. A. Ritchie, and G. A. C. Jones, J. Phys. C 21, L209 (1988)
- 327** D. A. Wharam, M. Pepper, H. Ahmed, J. E. F. Frost, D. G. Hasko, D. C. Peacock, D. A. Ritchie, and G. A. C. Jones, J. Phys. C 21, L887 (1988)
- 328** J. W. G. Wildoer, L. C. Venema, A. G. Rinzel, and R. E. Smalley, and C. Dekker, Nature 391, 59 (1998)
- 329** J. Wilks, *An Introduction to Liquid Helium*, Clarendon Press (1970)
- 330** R. Willett, J. P. Eisenstein, H. L. Störmer, D. C. Tsui, A. C. Gossard, and J. H. English, Phys. Rev. Lett. 59, 1776 (1987)
- 331** R. E. Williams, *Modern GaAs Processing Methods*, Artech House (1990)
- 332** R. W. Winkler, J. P. Kotthaus, and K. Ploog, Phys. Rev. Lett. 62, 1177 (1989)
- 333** L. Worschech, F. Beuscher, and A. Forchel, Appl. Phys. Lett. 75, 578 (1999)
- 334** Q. Xie, A. Madhukar, P. Chen, and N. P. Kobayashi, Phys. Rev. Lett. 75, 2542 (1995)
- 335** A. Yacoby and Y. Imry, Phys. Rev. B 41, 5341 (1990)
- 336** A. Yacoby, U. Sivan, C. P. Umbach, and J. M. Hong, Phys. Rev. Lett. 66, 1938 (1991)
- 337** A. Yacoby, H. L. Stormer, N. S. Wingreen, L. N. Pfeiffer, K. W. Baldwin, and K. W. West, Phys. Rev. Lett. 77, 4612 (1996)
- 338** Y. Yafet, in *Solid State Physics*, Vol. 14, eds. F. Seitz and D. Turnbull, Academic Press (1963)
- 339** H. Yan, S. H. Park, G. Finkelstein, J. H. Reif, and T. H. LaBean, Science 301, 1882 (2003)
- 340** M. J. Yoo, T. A. Fulton, H. F. Hess, R. L. Willett, L. N. Dunkelberger, R. J. Chichester, L. N. Pfeiffer, and K. W. West, Science 276, 579 (1997)
- 341** D. Yoshioka, *The Quantum Hall Effect*, Springer (2002)
- 342** P. Yu and M. Cardona, *Fundamentals of Semiconductors*, Springer, Berlin (1999)
- 343** J. Zak, Phys. Rev. B 32, 2218 (1985)
- 344** E. Zaremba, Phys. Rev. B 45, 14143 (1992)
- 345** M. Ziese and T. M. Thornton (eds.) *Spin Electronics*, Cambridge University Press (1992)
- 346** J. M. Ziman, *Principles of the Theory of Solids*, Cambridge University Press (1995)
- 347** J. M. Ziman, *Elements of Advanced Quantum Theory*, Springer (2001)
- 348** G. Zimmerli, T. M. Eiles, R. L. Kautz, and J. M. Martinis, Appl. Phys. Lett. 61, 237 (1992)
- 349** N. M. Zimmerman, W. H. Huber, A. Fujiwara, and Y. Takahashi, Appl. Phys. Lett. 79, 3188 (2001)

This Page Intentionally Left Blank

Index

π conjugation 90
 s -matrix 237
0.7 structure 195
2DEG 79

a
acceptor 37
accumulation 78
addition energy 280
addition spectrum 263
Aharonov–Bohm effect 8, 223
Al_xGa_{1-x}As 26
Altshuler–Aronov–Spivak oscillations 225, 318
ambipolar devices 78
angle evaporation technique 113
antidots 312
antilocalization 245
artificial atom 3, 273

b
ballistic transport 3
ballistic wires 182
band alignment 69, 75
band bending 67
bandgaps 18
bending of energy bands 67
Berggren model 181
Bloch theorem 18
Boltzmann distribution 30
Boltzmann equation 41
Boltzmann model 15, 40
bonding 115
boundary scattering 181
break junction 205
Bridgman growth technique 100
Brillouin zone 17
Brownian motion 140

c
capacitance spectroscopy 152

carbon nanotubes 206, 299
charge neutrality level 66
chemical potential 29
cleaved edge overgrowth 102
coherence 8
conductance quantization 183
constant interaction model 279
contact resistance 188
correlation function 141
cotunneling 269
Coulomb blockade 247
Coulomb staircase 255
cryostats 122
crystal growth 98
crystal structure: Si, GaAs 16
current sources 129
current-to-voltage converter 133
cyclotron radius 145
Czochralski growth 98

d
density of states 27
dephasing time 143
diamagnetic shift 170
differential conductance 135
diffusion constant 140
diffusion equation 140
dilution refrigerator 125
direct bandgap 26
donor 37
doping 36
double barrier 238
Drude model 41
dry etching 112

e
edge channels 202
edge states 177, 199
effective density of states 30
effective mass 2
effective mass in heterostructures 91
effective mass in MOSFETs 80
Einstein relation 48, 142

- elastic mean free path 140
 electrometer 267
 electron beam lithography 1, 109
 electron waveguide 193
 electron–phonon scattering 49
 energy bands 18
 envelope wave equation 35
 envelope wave functions 32
 ergodicity 231
 etching 112
- f**
 fcc lattice 16
 Fermi energy 29
 Fermi wavelength 2, 139
 Fermi–Dirac distribution 29
 field effect transistor 77
 filling factor 150
 flux cancellation 245
 Fock–Darwin model 195, 282
 focused ion beam writing 110
 four-probe measurement 134
 Fourier transformation 345
 fractional quantum Hall effect 155
 freezeout regime 38
 Friedel oscillations 53
 fullerene 299
- g**
 Ga[Al]As 26
 Ga[Al]As HEMT 84
 GaAs 16
 GaAs: band structure 21
 GaAs: holes 25
 Gaussian ensembles 292
 giant magneto-resistance 324
 GMR 324
 graphite 17, 21
 guiding center 159
- h**
 Hall bar 46
 Hall effect 47
 He II 117
 He II osmosis 119
 heavy holes 25
 helium cryostats 122
 HEMT 84
 heterointerface 74, 84
 holes 15
 Hund's rules 287
- i**
 impurity scattering 49
 InAs–AlSb quantum well 87
 indirect bandgap 26
- induced gap states 69, 75
 interaction parameter 145
 interfaces 68
 intrinsic carrier concentration 29
 intrinsic regime 39
 inversion 78
- k**
 kp model 24
 Kramers degeneracy 25
 Kubo formula 142, 313
- l**
 Landau level 149
 Landau quantization 148
 Landauer formula 189
 Landauer–Büttiker formalism 198
 Langevin equation 141
 law of mass action 30
 LEC growth technique 100
 lift-off 109
 light holes 25
 liquid helium 116
 lithography 1, 107
 localization length 145
 Luttinger parameters 25
- m**
 magnetic length 145, 149
 magnetic mass 180
 magneto-resistivity tensor 46
 mask aligner 107
 MBE 101
 mesoscopic transport 2
 metal organic chemical vapor deposition 101
 metal–insulator transition 151
 metallization 113
 microchip 2
 mobility in HEMTs 86
 MOCVD 101
 modulation doping 11, 84
 molecular beam epitaxy 101
 molecular electronics 13
 Moore's law 1
 MOS interface 78
 MOSFET 12, 77
- n**
 nearly free electron model 18
- o**
 ohmic contact 73
 Onsager–Casimir symmetry relation 47
 operational amplifier 131
 optical lithography 1, 107

overcut profile 109

p

pentacene 89
persistent current 374
phase coherence 8
phase coherence length 143
photoresist 108
PI controller 136
pinning 40
pinning of the Fermi level 67, 68
plastic transistors 90
Poisson distribution 236
polythiophene 89

q

QHE 154
quantized chaos 290
quantum computation 301
quantum dot 3, 9, 273
quantum film 3
quantum Hall effect 5, 154
quantum point contact 4, 177, 210
quantum scattering length 139
quantum scattering time 139
quantum well 77
quantum wire 3, 177, 229
quasi-2DEG 165
qubit 302

r

random matrix theory 292
RHEED 102

s

sample holders 128
saturation regime 39
scanning probe lithography 110
scattering in heterostructures 92
scattering mechanisms 48
Schottky barrier 68
Schottky diode 73
screening 50
screening in two dimensions 92
self-assembled quantum dots 103
semi-insulating 40
semiconductor: definition 21
SET transistor 259
short-period superlattice 102
Shubnikov-de Haas oscillations 5, 156
Si 16
Si MOSFET 77

Si: band structure 21

Si: holes 25
Si[Ge] quantum well 87
silicon 12
single-electron box 271
single-electron pump 267
single-electron tunneling 9, 247
single-electron turnstile 270
size quantization 7
skipping orbits 160, 199
spin polarization 323
spin-orbit splitting 24
spintronics 11, 323
split gates 184
subband 79
superfluid film creeping 119
superlattice 10
superlattices 309
superleak 119
surface bands 65
surface recombinations 65
surface states 59

t

thermal length 144
Thomas-Fermi screening 52
tight binding model 18
TMR 324
transconductance 131, 135
transistor amplifier 130
triangular interface potential 79
tunneling magneto-resistance 324
two-dimensional electron gas 79
two-probe measurement 134

u

undercut profile 108
universal conductance fluctuations 229

v

valley degeneracy 23
velocity autocorrelation function 141
virtual ground 133
voltage amplifier 132
voltage divider 129
voltage sources 129

w

warped spheres 25
weak localization 226
Weiss oscillations 311
wet chemical etching 112

Related Titles

Wilkening, G., Koenders, L.

Nanoscale Calibration Standards and Methods

Dimensional and Related Measurements in the Micro- and Nanometer Range

541 pages with 380 figures

2005, Hardcover

ISBN-13: 978-3-527-40502-2

ISBN-10: 3-527-40502-X

Waser, R. (ed.)

Nanoelectronics and Information Technology

995 pages with 1148 figures and 47 tables

2005, Hardcover

ISBN-13: 978-3-527-40542-8

ISBN-10: 3-527-40542-9

Baltes, H., Brand, O., Fedder, G. K., Hierold, C., Korvink, J. G., Tabata, O. (eds.)

CMOS-MEMS

608 pages with 312 figures and 32 tables

2005, Hardcover

ISBN-13: 978-3-527-31080-7

ISBN-10: 3-527-31080-0

Fecht, H.-J., Werner, M. (eds.)

The Nano-Micro Interface

Bridging the Micro and Nano Worlds

351 pages with 102 figures and 27 tables

2004, Hardcover

ISBN-13: 978-3-527-30978-8

ISBN-10: 3-527-30978-0

Champion, Y., Fecht, H.-J. (eds.)

Nano-Architected and Nanostructured Materials

Fabrication, Control and Properties

166 pages with 101 figures and 16 tables

2004, Hardcover

ISBN-13: 978-3-527-31008-1

ISBN-10: 3-527-31008-8

Wolf, E. L.

Nanophysics and Nanotechnology

An Introduction to Modern Concepts in Nanoscience

187 pages with 69 figures and 5 tables

2004, Softcover

ISBN-13: 978-3-527-40407-0

ISBN-10: 3-527-40407-4

Köhler, M., Fritzsche, W.

Nanotechnology

An Introduction to Nanostructuring Techniques

284 pages with 143 figures and 9 tables

2004, Hardcover

ISBN-13: 978-3-527-30750-0

ISBN-10: 3-527-30750-8

Reich, S., Thomsen, C., Maultzsch, J.

Carbon Nanotubes

Basic Concepts and Physical Properties

224 pages with 126 figures

2004, Hardcover

ISBN-13: 978-3-527-40386-8

ISBN-10: 3-527-40386-8