Lecture Notes in Physics

K. Busch A. Powell C. Röthig
G. Schön J. Weissmüller (Eds.)

CFN Lectures on Functional Nanostructures

Vol. 1

Springer

Editors

Kurt Busch
University of Central Florida
Dept. of Physics and College
of Optics and Photonics: CREOL & FPCE
Orlando, FL 32816, USA

Annie Powell
Universität Karlsruhe (TH)
Institut für Anorganische Chemie
Engesserstr. 15
76131 Karlsruhe, Germany

Christian Röthig
Universität Karlsruhe (TH)
DFG-Centrum für Funktionelle
Nanostrukturen
Wolfgang-Gaede-Str. 1
76131 Karlsruhe, Germany

Gerd Schön
Universität Karlsruhe (TH)
Institut für Theoretische
Festkörperphysik
76128 Karlsruhe, Germany

Jörg Weissmüller
Forschungszentrum Karlsruhe
Institut für Nanotechnologie
Postfach 3640
76021 Karlsruhe, Germany

K. Busch, A. Powell, C. Röthig, G. Schön, J. Weissmüller (Eds.), *Functional Nanostructures*, Lect. Notes Phys. **658** (Springer, Berlin Heidelberg 2004), DOI 10.1007/b100127

Library of Congress Control Number: 2004110450

ISSN 0075-8450
ISBN 3-540-22929-9 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in Germany

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

Typesetting: Camera-ready by the authors/editor
Data conversion: PTP-Berlin Protago-TeX-Production GmbH
Cover design: *design & production*, Heidelberg

Printed on acid-free paper
54/3141/ts - 5 4 3 2 1 0

Lecture Notes in Physics

For information about Vols. 1–609 please contact your bookseller or Springer LNP Online archive: springerlink.com

Vol.610: M. Ristig, K. Gernoth (Eds.), Particle Scattering, X-Ray Diffraction, and Microstructure of Solids and Liquids.

Vol.611: A. Buchleitner, K. Hornberger (Eds.), Coherent Evolution in Noisy Environments.

Vol.612: L. Klein, (Ed.), Energy Conversion and Particle Acceleration in the Solar Corona.

Vol.613: K. Porsezian, V.C. Kuriakose (Eds.), Optical Solitons. Theoretical and Experimental Challenges.

Vol.614: E. Falgarone, T. Passot (Eds.), Turbulence and Magnetic Fields in Astrophysics.

Vol.615: J. Büchner, C.T. Dum, M. Scholer (Eds.), Space Plasma Simulation.

Vol.616: J. Trampetic, J. Wess (Eds.), Particle Physics in the New Millenium.

Vol.617: L. Fernández-Jambrina, L. M. González-Romero (Eds.), Current Trends in Relativistic Astrophysics, Theoretical, Numerical, Observational

Vol.618: M.D. Esposti, S. Graffi (Eds.), The Mathematical Aspects of Quantum Maps

Vol.619: H.M. Antia, A. Bhatnagar, P. Ulmschneider (Eds.), Lectures on Solar Physics

Vol.620: C. Fiolhais, F. Nogueira, M. Marques (Eds.), A Primer in Density Functional Theory

Vol.621: G. Rangarajan, M. Ding (Eds.), Processes with Long-Range Correlations

Vol.622: F. Benatti, R. Floreanini (Eds.), Irreversible Quantum Dynamics

Vol.623: M. Falcke, D. Malchow (Eds.), Understanding Calcium Dynamics, Experiments and Theory

Vol.624: T. Pöschel (Ed.), Granular Gas Dynamics

Vol.625: R. Pastor-Satorras, M. Rubi, A. Diaz-Guilera (Eds.), Statistical Mechanics of Complex Networks

Vol.626: G. Contopoulos, N. Voglis (Eds.), Galaxies and Chaos

Vol.627: S.G. Karshenboim, V.B. Smirnov (Eds.), Precision Physics of Simple Atomic Systems

Vol.628: R. Narayanan, D. Schwabe (Eds.), Interfacial Fluid Dynamics and Transport Processes

Vol.629: U.-G. Meißner, W. Plessas (Eds.), Lectures on Flavor Physics

Vol.630: T. Brandes, S. Kettemann (Eds.), Anderson Localization and Its Ramifications

Vol.631: D. J. W. Giulini, C. Kiefer, C. Lämmerzahl (Eds.), Quantum Gravity, From Theory to Experimental Search

Vol.632: A. M. Greco (Ed.), Direct and Inverse Methods in Nonlinear Evolution Equations

Vol.633: H.-T. Elze (Ed.), Decoherence and Entropy in Complex Systems, Based on Selected Lectures from DICE 2002

Vol.634: R. Haberlandt, D. Michel, A. Pöppl, R. Stannarius (Eds.), Molecules in Interaction with Surfaces and Interfaces

Vol.635: D. Alloin, W. Gieren (Eds.), Stellar Candles for the Extragalactic Distance Scale

Vol.636: R. Livi, A. Vulpiani (Eds.), The Kolmogorov Legacy in Physics, A Century of Turbulence and Complexity

Vol.637: I. Müller, P. Strehlow, Rubber and Rubber Balloons, Paradigms of Thermodynamics

Vol.638: Y. Kosmann-Schwarzbach, B. Grammaticos, K.M. Tamizhmani (Eds.), Integrability of Nonlinear Systems

Vol.639: G. Ripka, Dual Superconductor Models of Color Confinement

Vol.640: M. Karttunen, I. Vattulainen, A. Lukkarinen (Eds.), Novel Methods in Soft Matter Simulations

Vol.641: A. Lalazissis, P. Ring, D. Vretenar (Eds.), Extended Density Functionals in Nuclear Structure Physics

Vol.642: W. Hergert, A. Ernst, M. Däne (Eds.), Computational Materials Science

Vol.643: F. Strocchi, Symmetry Breaking

Vol.644: B. Grammaticos, Y. Kosmann-Schwarzbach, T. Tamizhmani (Eds.) Discrete Integrable Systems

Vol.645: U. Schollwöck, J. Richter, D.J.J. Farnell, R.F. Bishop (Eds.), Quantum Magnetism

Vol.646: N. Bretón, J. L. Cervantes-Cota, M. Salgado (Eds.), The Early Universe and Observational Cosmology

Vol.647: D. Blaschke, M. A. Ivanov, T. Mannel (Eds.), Heavy Quark Physics

Vol.648: S. G. Karshenboim, E. Peik (Eds.), Astrophysics, Clocks and Fundamental Constants

Vol.649: M. Paris, J. Rehacek (Eds.), Quantum State Estimation

Vol.650: E. Ben-Naim, H. Frauenfelder, Z. Toroczkai (Eds.), Complex Networks

Vol.651: J.S. Al-Khalili, E. Roeckl (Eds.), The Euroschool Lectures of Physics with Exotic Beams, Vol.I

Vol.652: J. Arias, M. Lozano (Eds.), Exotic Nuclear Physics

Vol.653: E. Papantonoupoulos (Ed.), The Physics of the Early Universe

Vol.654: G. Cassinelli, A. Levrero, E. de Vito, P. J. Lahti (Eds.), Theory and Appplication to the Galileo Group

Vol.655: M. Shillor, M. Sofonea, J.J. Telega, Models and Analysis of Quasistatic Contact

Vol.656: K. Scherer, H. Fichtner, B. Heber, U. Mall (Eds.), Space Weather

Vol.657: J. Gemmer, M. Michel, G. Mahler (Eds.), Quantum Thermodynamics

Vol.658: K. Busch, A. Powell, C. Röthig, G. Schön, J. Weissmüller (Eds.), Functional Nanostructures

Preface

This book contains a selection of lectures from the first CFN Summer School on Functional Nanostructures which took place from September 24th to September 27th, 2003 in Bad Herrenalb in the Black Forest of Germany. The DFG-funded CFN, or Center for Functional Nanostructures, was founded in July 2001 at the Universität Karlsruhe (TH) and the Forschungszentrum Karlsruhe. Additional funding comes from the State of Baden-Württemberg and from the home institutions, Universität and Forschungszentrum. The mission of the CFN is to investigate nanoscale functional materials within the following broad research areas:

A Nanophotonics
B Nanoelectronics
C Molecular Nanostructures
D Nanostructured Materials

The CFN is made up of a wide range of research groups from 15 different Institutes in Karlsruhe bringing a variety of scientific backgrounds together. The Center thus provides a melting pot where various talents can be combined to address the problems associated with creating functional nanoscale materials. At the same time, the members of the Center are acutely aware of the need to develop a common language to facilitate communication amongst the various disciplines, and thus the idea of holding Summer Schools to bring groups across the four research areas together evolved. The remit of the Summer Schools is to allow members of the CFN and external participants to exchange ideas and explain research methods and strategies through a series of lectures designed both to introduce unfamiliar concepts and discuss the benefits and problems associated with various research methods including many which are highly specialised.

Chapters 1–4 of these Lecture Notes are devoted to research area A (Nanophotonics), Chaps. 5–9 to B (Nanoelectronics) while the last two chapters give a flavor of research areas C (Molecular Nanostructures) and D (Nanostructured Materials).

The lecture notes we have brought together here represent a selection of the presentations made at the Summer School in 2003 and are designed to provide a useful starting point for those interested in learning more about this rapidly developing area of science. It is hoped that they will not only provide a useful working text, but also arouse interest in our activities in Karlsruhe within the CFN.

We would like to take this opportunity to thank all the authors who have contributed to this volume for their valuable input as well as all the participants at the Summer School for helping to make this interdisciplinary venture such a success.

Karlsruhe,
March 2004

Kurt Busch
Annie Powell
Christian Röthig
Gerd Schön
Jörg Weissmüller

Contents

List of Contributors

Wolfgang Belzig
Department of Physics
and Astronomy,
University of Basel,
Klingelbergstraße 82,
4056 Basel, Switzerland

Gunnar Böttger
Technische Universität
Hamburg-Harburg,
Eißendorfer Straße 38,
21073 Hamburg, Germany

Kurt Busch
Department of Physics
and School of Optics:
CREOL & FPCE,
University of Central Florida,
Orlando, FL 32816, USA
and
Institut für Theorie
der Kondensierten Materie,
Universität Karlsruhe (TH),
76128 Karlsruhe, Germany

Manfred Eich
Technische Universität
Hamburg-Harburg,
Eißendorfer Straße 38,
21073 Hamburg, Germany

Igor V. Gornyi
Forschungszentrum Karlsruhe,
Institut für Nanotechnologie (INT),
Postfach 3640,
76021 Karlsruhe, Germany

Eugeniyus L. Ivchenko
Eugeniyus L. Ivchenko
A.F. Ioffe Physico-Technical Institute,
Russian Academy of Sciences,
194021 St. Petersburg, Russia

Heinz Kalt
Institut für Angewandte Physik
and DFG-Center for Functional
Nanostructures,
Universität Karlsruhe (TH),
76128 Karlsruhe, Germany

Jürgen König
Institut für Theoretische Physik III,
Ruhr-Universität Bochum,
44780 Bochum, Germany

Jan Martinek
Institut für Theoretische
Festkörperphysik,
Universität Karlsruhe (TH),
76128 Karlsruhe, Germany
and
Institute for Materials Research,
Tohoku University,
Sendai 980-8577, Japan
and
Institute of Molecular Physics,
Polish Academy of Sciences,
60-179 Poznań, Poland

Sergei Mingaleev
Institut für Theorie
der Kondensierten Materie,
Universität Karlsruhe (TH),
76128 Karlsruhe, Germany
and
Bogolyubov Institute
for Theoretical Physics,
03143 Kiev, Ukraine

John H. Perepezko
University of Wisconsin-Madison,
Department of Materials Science
and Engineering,
1509 University Avenue,
WI 53706, USA

Dmitri G. Polyakov
Forschungszentrum Karlsruhe,
Institut für Nanotechnologie (INT),
Postfach 3640,
76021 Karlsruhe, Germany

Markus Schmidt
Technische Universität
Hamburg-Harburg,
Eißendorfer Straße 38,
21073 Hamburg, Germany

Lasha Tkeshelashvili
Institut für Theorie
der Kondensierten Materie,
Universität Karlsruhe (TH),
76128 Karlsruhe, Germany

Florian Weigend
Forschungszentrum Karlsruhe,
Institut für Nanotechnologie (INT),
Postfach 3640,
76021 Karlsruhe, Germany

Jürgen Weis
Max-Planck-Institut
für Festkörperforschung,
Heisenbergstraße 1,
70569 Stuttgart, Germany

Heiko B. Weber
Forschungszentrum Karlsruhe,
Institut für Nanotechnologie (INT),
Postfach 3640,
76021 Karlsruhe, Germany

Solid State Theory Meets Photonics: The Curious Optical Properties of Photonic Crystals

Kurt Busch[1,2], Sergei F. Mingaleev[1,3], Matthias Schillinger[1],
Daniel Hermann[1], and Lasha Tkeshelashvili[1]

[1] Institut für Theorie der Kondensierten Materie, Universität Karlsruhe,
76128 Karlsruhe, Germany
[2] Department of Physics and School of Optics: CREOL & FPCE, University of
Central Florida, Orlando, FL 32816, USA
[3] Bogolyubov Institute for Theoretical Physics, 03143 Kiev, Ukraine

1 Introduction

The past decades have seen dramatic advances in microstructuring technology. Today, a wide variety of structures with feature sizes ranging from a couple of micrometers all the way down to a few tens of nanometers are routinely fabricated with precision better than ten nanometers. In addition to these improvements in fabrication quality, the variety of materials that can be processed is growing continuously. These advances in materials science are paralleled by the development of novel and improvement of existing laser sources that allows one to generate electromagnetic fields with previously unattainable energy densities as well as temporal and spatial coherences. Bringing together advanced microfabrication technologies with sophisticated laser systems lies at the heart of Nano-Photonics: The control over the flow of light on length scales of the wavelength of light itself through microstructured optical materials ("photonic metamaterials") with carefully designed properties.

A particularly prominent class of metamaterials are the so-called Photonic Crystals (PCs) which consist of a microfabricated array of dielectric materials in two or three spatial dimensions. The resulting combination of microscopic scattering resonances from individual elements of the periodic array and Bragg scattering from the corresponding lattice is very similar to the propagation of electron waves in electronic crystals and, as a result, leads to the formation of an energy bandstructure for electromagnetic waves. The most dramatic modification of the photonic dispersion relation in these systems occurs when the photonic bandstructure of suitably engineered PCs exhibits frequency ranges over which the light propagation is forbidden irrespective of the direction of propagation [1,2]. The corresponding subclass of PCs that exhibit such a Photonic Band Gap (PBG) are commonly referred to as Photonic Band Gap materials and may be regarded as a "Semiconductor for Light" [3]. In fact, this analogy of PBG materials to electronic semiconducting materials may be reaching very far and the current state of PBG research suggests that this field is at a stage comparable to the early years of semiconductor technology shortly before the in-

K. Busch et al., Solid State Theory Meets Photonics: The Curious Optical Properties of Photonic Crystals, Lect. Notes Phys. **658**, 1–22 (2005)
http://www.springerlink.com/

vention of the solid state electronic transistor. If this analogy continues to hold, one may find PBG materials at the heart of a 21st century revolution in optical technologies similar to the revolution in electronics we have witnessed over the latter half of the 20th century.

In this chapter, we want to outline how the vast knowledge about electron propagation in crystalline solids may be employed to determine the optical properties of PCs in general and of PBG materials in particular. In Sect. 2, we introduce photonic bandstructure computations as the central tool for obtaining the photonic dispersion relation, the corresponding eigenmodes (Bloch functions), and related physical quantities such as group velocities, group velocity dispersion as well as total and local density of states. In Sect. 3, we discuss how the existence of a PBG may be utilized for the design of (linear) waveguiding structures through the deliberate incorporation of defects. In addition, we outline the qualitatively new physics that may arise in the case of nonlinear and quantum optical phenomena in PBG materials. Finally, in Sect. 4, we discuss a novel approach to obtain a fully quantitative lattice model for PCs using the solid-state theoretical concept of Wannier functions that allow us to efficiently carry out accurate simulations of PC-based devices. We employ this approach to develop novel concepts and design for functional elements based on the infiltration of individual pores in two-dimensional PBG materials.

2 Photonic Bandstructure Computation

Photonic bandstructure computations determine the dispersion relation of infinitely extended defect-free PCs. In addition, they allow us to design PCs that exhibit PBGs and to accurately interpret measurements on PC samples. As a consequence, photonic bandstructure calculations represent an important predictive as well as interpretative basis for PC research and, therefore, lie at the heart of theoretical investigations of PCs. For instance, the first convincing evidence for the very existence of PBGs has come from the seminal theoretical work of the Iowa State group [4], where it has been reported that certain structures with diamond symmetry exhibit complete three-dimensional (3D) PBGs.

2.1 Photonic Bandstructure and Bloch Functions

More specifically, the goal of photonic bandstructure computations is to find the eigenfrequencies and associated eigenmodes of the wave equation for the perfect PC, i.e., for an infinitely extended periodic array of dielectric material. For the simplicity of presentation, we restrict ourselves in the remainder of this chapter to the case of TM-polarized radiation propagating in the plane of periodicity (x, y)-plane of two-dimensional (2D) PCs. In this case, the wave equation in the frequency domain (harmonic time dependence) for the z-component of the electric field reads

$$\frac{1}{\epsilon_{\mathrm{p}}(\boldsymbol{r})} \left(\partial_x^2 + \partial_y^2\right) E(\boldsymbol{r}) + \frac{\omega^2}{c^2} E(\boldsymbol{r}) = 0\,. \tag{1}$$

Here c denotes the vacuum speed of light and $\boldsymbol{r} = (x, y)$ denotes a two-dimensional position vector. The dielectric constant $\epsilon_{\mathrm{p}}(\boldsymbol{r}) \equiv \epsilon_{\mathrm{p}}(\boldsymbol{r} + \boldsymbol{R})$ is periodic with respect to the set $\mathcal{R} = \{n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2; (n_1, n_2) \in \mathcal{Z}^2\}$ of lattice vectors $\boldsymbol{R}$ generated by the primitive translations $\boldsymbol{a}_i,\ i = 1, 2$ that describe the structure of the PC. Equation (1) represents a differential equation with periodic coefficients and, therefore, its solutions obey the Bloch-Floquet theorem

$$E_{\boldsymbol{k}}(\boldsymbol{r} + \boldsymbol{a}_i) = e^{i\boldsymbol{k}\boldsymbol{a}_i}\, E_{\boldsymbol{k}}(\boldsymbol{r})\,, \tag{2}$$

where $i = 1, 2$. The wave vector $\boldsymbol{k} \in$ 1st BZ that labels the solution is a vector of the first Brillouin zone (BZ) known as the crystal momentum. As a result of this so-called reduced zone scheme, the photonic bandstructure acquires a multi-branch nature that is associated with the backfolding of the dispersion relation into the 1st BZ. This introduces a discrete index n, the so-called band index, that enumerates the distinct eigenfrequencies and eigenfunctions at the same wave vector $\boldsymbol{k}$ [5]. Furthermore, (2) suggests that the Bloch function $E_{n\boldsymbol{k}}(\boldsymbol{r})$ for band n and wave vector $\boldsymbol{k}$ can be written in a form

$$E_{n\boldsymbol{k}}(\boldsymbol{r}) = e^{i\boldsymbol{k}\boldsymbol{r}}\, u_{n\boldsymbol{k}}(\boldsymbol{r})\,, \tag{3}$$

representing a plane wave that is modulated by a lattice periodic function $n\boldsymbol{k}(\boldsymbol{r})$ [5].

A straightforward way of solving (1) is to expand all the periodic functions into a Fourier series over the reciprocal lattice $\mathcal{G}$, thereby transforming the differential equation into an infinite matrix eigenvalue problem, which may be suitably truncated and solved numerically.

For instance, for a PC consisting of pores (radius r, dielectric constant ϵ_b) in a background material (dielectric constant ϵ_b), the periodic dielectric constant $\epsilon_{\mathrm{p}}(\boldsymbol{r})$ may be written as

$$\begin{aligned} \frac{1}{\epsilon_{\mathrm{p}}(\boldsymbol{r})} &= \frac{1}{\epsilon_a} + \left(\frac{1}{\epsilon_b} - \frac{1}{\epsilon_a}\right) \sum_{\boldsymbol{R}} S(\boldsymbol{r} - \boldsymbol{R}) && (4) \\ &= \sum_{\boldsymbol{G}} \eta_{\boldsymbol{G}}\, e^{i\boldsymbol{G}\cdot\boldsymbol{r}}\,, && (5) \end{aligned}$$

where $S(\boldsymbol{r} - \boldsymbol{R})$ takes on the value one if $|\boldsymbol{r}| \leq r$, and is zero elsewhere. The Fourier coefficients $\eta_{\boldsymbol{G}}$ are given by

$$\eta_{\boldsymbol{G}} = \frac{1}{V_{\mathrm{WSC}}} \int_{\mathrm{WSC}} d^2r\, \frac{1}{\epsilon_{\mathrm{p}}(\boldsymbol{r})}\, e^{-i\boldsymbol{G}\cdot\boldsymbol{r}}\,. \tag{6}$$

Here, we designate the volume of the Wigner-Seitz cell (WSC) by V. Similarly, following the Bloch-Floquet theorem we expand $E(\boldsymbol{r})$ for a given wave vector $\boldsymbol{k}$ as

$$E_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} A_{\boldsymbol{G}}^{\boldsymbol{k}}\, e^{i(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{r}}\,. \tag{7}$$

Inserting these expansions, (7) and (4) into (1) and defining the coefficients $B_{\boldsymbol{G}}^{\boldsymbol{k}} \equiv |\boldsymbol{k}+\boldsymbol{G}|\, A_{\boldsymbol{G}}^{\boldsymbol{k}}$, yields a symmetric eigenvalue problem:

$$\sum_{\boldsymbol{G}'} |\boldsymbol{k}+\boldsymbol{G}||\boldsymbol{k}+\boldsymbol{G}'|\, \eta_{\boldsymbol{G}-\boldsymbol{G}'}\, B_{\boldsymbol{G}'}^{\boldsymbol{k}} = \frac{\omega_{\boldsymbol{k}}^2}{c^2}\, B_{\boldsymbol{G}}^{\boldsymbol{k}} \,. \tag{8}$$

The reciprocal lattice sum is then truncated in order to obtain a numerical solution. In our numerical calculations convergence was established by increasing the number of reciprocal lattice vectors used to truncate (8) until the final result was independent of the truncation. We found that using of the order of thousand reciprocal lattice vectors closest to the origin yields a converged band structure for the dielectric contrasts we have considered. For future reference we note that, once the eigenfrequencies $\omega_{n\boldsymbol{k}}$ and associated eigenvectors $B_{\boldsymbol{G}}^{n\boldsymbol{k}}$ (or equivalently $A_{\boldsymbol{G}}^{n\boldsymbol{k}}$) have been found, the eigenfunctions, the so-called Bloch functions, can be recovered using (7). Further Details of this plane wave method (PWM) for isotropic systems can be found, for instance, in [4,6] and for anisotropic systems in [7].

While the PWM provides a straightforward approach to computing the bandstructure of PCs, it also exhibits a number of shortcomings such as slow convergence associated with the truncation of Fourier series in the presence of discontinuous changes in the dielectric constants. In particular, this slow convergence makes the accurate calculation of Bloch functions a formidable and resource-consuming task. Therefore, we have recently developed an efficient real space approach to computing photonic bandstructures [8]. Within this approach, the wave equation, (1), is discretized in a single unit cell in real space (defined through the set of space points $\boldsymbol{r} = r_1\boldsymbol{a}_1 + r_2\boldsymbol{a}_2$ with $r_1, r_2 \in [-1/2, 1/2]$), leading to a sparse matrix problem. The Bloch-Floquet theorem, (2), provides the boundary condition for the elliptic partial differential equation (1). In addition, the eigenvalue is treated as an additional unknown for which the normalization of the Bloch functions provides the additional equation needed for obtaining a well-defined problem. The solution of this algebraic problem is obtained by employing Multi-Grid (MG) methods which guarantee an efficient solution by taking full advantage of the smoothness of the photonic Bloch functions [8,9]. Even for the case of a naive finite difference discretization, the MG-approach easily outperforms the PWM and leads to a substantial reduction in CPU time. For instance, in the present case of 2D systems for which the Bloch functions are required we save one order of magnitude in CPU time as compared to PWM. Additional refinements such as a finite element discretization will further increase the efficiency of the MG-approach.

In Fig. 1b, we show the bandstructure for TM-polarized radiation in a 2D PC consisting of a square lattice (lattice constant a) of cylindrical air pores (radius $r_{\mathrm{pore}} = 0.475\, a$) in a silicon matrix ($\varepsilon_{\mathrm{p}} = 12$). Throughout this chapter, this will serve as a model PC with which to illustrate our results. This structure exhibits two 2D PBGs. The larger, fundamental bandgap (20% of the midgap frequency) extends between $\omega = 0.238 \times 2\pi c/a$ to $\omega = 0.291 \times 2\pi c/a$ and the smaller, higher order bandgap (8% of the midgap frequency) extends from $\omega = 0.425 \times 2\pi c/a$ to $\omega = 0.464 \times 2\pi c/a$.

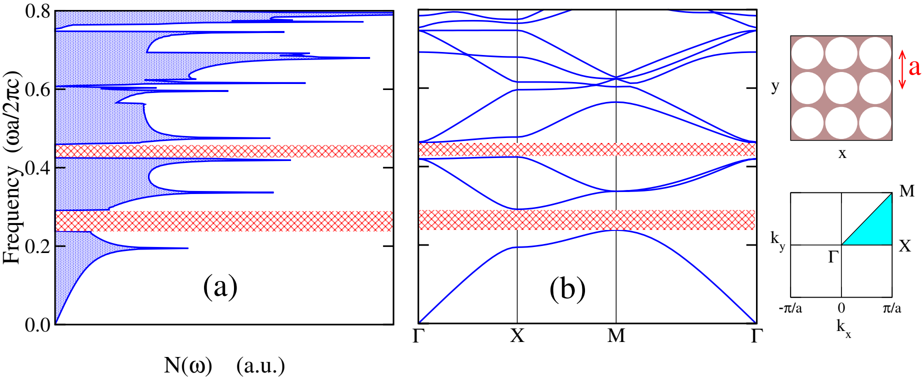


Fig. 1. Density of States (a) and photonic band structure (b) for TM-polarized radiation in a square lattice (lattice constant a) of cylindrical air pores of radius $r = 0.475\ a$ in dielectric with $\varepsilon = 12$ (silicon). This PC exhibits a large fundamental gap extending from $\omega = 0.238 \times 2\pi c/a$ to $\omega = 0.291 \times 2\pi c/a$. A higher order band gap extends from $\omega = 0.425 \times 2\pi c/a$ to $\omega = 0.464 \times 2\pi c/a$

2.2 Photonic Density of States

The photonic dispersion relation $\omega_n(\boldsymbol{k})$ gives rise to a photonic Density of States (DOS), which plays a fundamental role for the understanding of the quantum optical properties of active material embedded in PCs (see Sect. 3). The photonic DOS, $N(\omega)$, is defined by "counting" all allowed states with a given frequency ω

$$N(\omega) = \sum_n \int_{1\mathrm{stBZ}} d^2k\ \delta(\omega - \omega_n(\boldsymbol{k}))\,. \tag{9}$$

In Fig. 1a we depict the DOS for our model system, where the photonic band gaps are manifest as regions of vanishing DOS. Characteristic for 2D systems is the linear behavior for small frequencies, the discontinuity of the DOS at the band edges and the logarithmic singularities, the so-called van Hove singularities, associated with vanishing group velocities for certain frequencies inside the bands (compare with Fig. 1b).

However, for applications to quantum optical experiments in photonic crystals it is necessary to investigate not only the (overall) availability of modes with frequency ω but also the local coupling strength of an emitter at a certain position $\boldsymbol{r}$ in the PC to the electromagnetic environment provided by the PC. Consequently, it is the overlap matrix element of the emitters dipole moment to the eigenmodes (Bloch functions) that is determining quantum optical properties such as decay rates etc. [46]. This may be combined into the local DOS (LDOS), $N(\boldsymbol{r}, \omega)$, defined as

$$N(\boldsymbol{r}, \omega) = \sum_n \int_{\mathrm{BZ}} d^2k\ |E_{n\boldsymbol{k}}(\boldsymbol{r})|^2\, \delta(\omega - \omega_n(\boldsymbol{k}))\,. \tag{10}$$

For an actual calculation, the integrals in (9) and (10) must be suitably discretized and one may again revert to the methods of electronic band structure calculations (see [6]).

2.3 Group Velocity and Group Velocity Dispersion

In order to understand pulse propagation in linear and nonlinear PCs, it is necessary to obtain group velocities and the group velocity dispersion (GVD) from the photonic band structure. In principle, this can be done through a simple numerical differentiation of the band structure, but in particular for the GVD this becomes computationally involved and great care must be exercised in order to avoid numerical instabilities. Therefore, we want to demonstrate how to obtain group velocities and group velocity dispersion through an adaptation of the so-called $\boldsymbol{k}\cdot\boldsymbol{p}$-perturbation theory (kp-PT) of electronic band structure theory. This approach has been applied to systems of arbitrary dimension [10,8,11] and will be particularly useful for the investigation of nonlinear effects in PCs.

With the help of the Bloch-Floquet theorem (3), we may rewrite the wave equation (1) into an equation of motion for the lattice-periodic functions $u_{\boldsymbol{k}}(\boldsymbol{r})$

$$\left(\Delta + 2i\,\nabla\cdot\boldsymbol{k} - \boldsymbol{k}^2\right) u_{\boldsymbol{k}}(\boldsymbol{r}) + \frac{\omega_{\boldsymbol{k}}^2}{c^2}\,\varepsilon_{\mathrm{p}}(\boldsymbol{r})\,u_{\boldsymbol{k}}(\boldsymbol{r}) = 0\ , \tag{11}$$

where, $\Delta = \partial_x^2 + \partial_y^2$. An inspection of (11) for the lattice-periodic $u_{\boldsymbol{k}+\boldsymbol{q}}(\boldsymbol{r})$

$$\begin{aligned}&\left(\Delta + 2i\,\nabla\cdot\boldsymbol{k} - \boldsymbol{k}^2\right)\ u_{\boldsymbol{k}+\boldsymbol{q}}(\boldsymbol{r}) + \boldsymbol{q}\cdot\left(2\hat{\Omega} - \boldsymbol{q}\right) u_{\boldsymbol{k}+\boldsymbol{q}}(\boldsymbol{r}) + \\ &\frac{\omega_{\boldsymbol{k}+\boldsymbol{q}}^2}{c^2}\,\varepsilon_{\mathrm{p}}(\boldsymbol{r})\,u_{\boldsymbol{k}+\boldsymbol{q}}(\boldsymbol{r}) \qquad\qquad = 0\,,\end{aligned} \tag{12}$$

at a nearby wave vector $\boldsymbol{k}+\boldsymbol{q}$ ($|\boldsymbol{q}| \ll \pi/a$) suggests that we treat the second term on the l.h.s. as a perturbation to (11). In writing (12), we have introduced $\hat{\Omega} = i(\nabla + i\boldsymbol{k})$. Comparing the perturbation series with a Taylor-expansion of frequency $\omega_{\boldsymbol{k}+\boldsymbol{q}}$ around $\boldsymbol{k}$ connects group velocities $\boldsymbol{v}_{\boldsymbol{k}} = \partial_{\boldsymbol{k}}\omega_{\boldsymbol{k}}$ and GVD tensor elements $M_{\boldsymbol{k}}^{ij} = \partial_{k_i}\partial_{k_j}\omega_{\boldsymbol{k}}$, $i = 1,2$ to expressions familiar from second order perturbation theory [10,8,11] . Explicitly [8], we obtain for the group velocity

$$\boldsymbol{v}_{n\boldsymbol{k}} = \frac{c^2}{\omega_{n\boldsymbol{k}}}\langle n\boldsymbol{k}|(-i\nabla)|n\boldsymbol{k}\rangle\ , \tag{13}$$

and for the GVD tensor

$$\begin{aligned}\boldsymbol{q}\cdot\mathcal{M}_{n\boldsymbol{k}}\cdot\boldsymbol{q} = |\boldsymbol{q}|^2\frac{c^2}{2\omega_{n\boldsymbol{k}}}\langle n\boldsymbol{k}|n\boldsymbol{k}\rangle - \frac{1}{2\omega_{n\boldsymbol{k}}}\left(\boldsymbol{q}\cdot\boldsymbol{v}_{n\boldsymbol{k}}\right)^2 \\ +\frac{2c^4}{\omega_{n\boldsymbol{k}}}\sum_{m\neq n}\frac{\langle n\boldsymbol{k}|(-i\boldsymbol{q}\cdot\nabla)|m\boldsymbol{k}\rangle\langle m\boldsymbol{k}|(-i\boldsymbol{q}\cdot\nabla)|n\boldsymbol{k}\rangle}{\omega_{n\boldsymbol{k}}^2 - \omega_{m\boldsymbol{k}}^2}\ .\end{aligned} \tag{14}$$

Here, we have used the notation $\int_{\mathrm{WSC}} d^2r\, E_{n\boldsymbol{k}}^*(\boldsymbol{r})\,\hat{O}\,E_{m\boldsymbol{k}}(\boldsymbol{r}) = \langle n\boldsymbol{k}|\hat{O}|m\boldsymbol{k}\rangle$ for matrix elements of the operator $\hat{O}$ between Bloch functions $E_{n\boldsymbol{k}}(\boldsymbol{r})$ and $E_{m\boldsymbol{k}}(\boldsymbol{r})$.

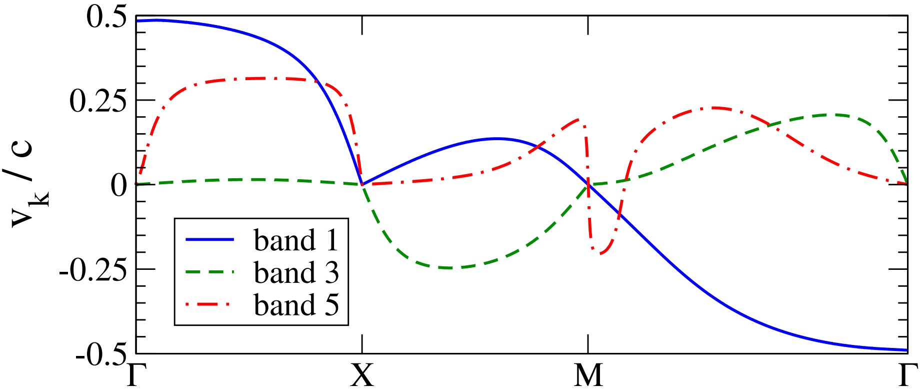


Fig. 2. Group velocities for bands 1, 3, and 5 of our model system (see Fig. 1). These group velocities of these bands exhibit extreme variations which may have numerous application in classical nonlinear optics. This illustrates the huge parameter space of effective parameters accessible with PCs

Despite their complicated appearance, these expressions can be evaluated rather easily using standard PWM and allow very accurate, efficient and numerically stable results. In Fig. 2 we display the variation of the group velocities associated with bands 1, 3, and 5 of our model system. Clearly visible are the extreme variations ranging from 0.5 c for band 1 in the long wavelength (effective medium) limit all the way to the almost vanishing group velocity of band 5 along the entire Γ-X direction. This illustrates the huge parameter space of effective group velocities that can *simultaneously* be realized in PCs.

3 The Physical Significance of Photonic Band Gaps

In the previous section, we have established the basic notions of photonic bandstructure theory and have given illustrations of the basic optical properties of PCs such as PBGs, DOS and group velocities. With this, we are now in the position to give a qualitative account of the physical significance of PBGs and justify the substantial experimental efforts at manufacturing 2D PBG materials in various material systems such as semiconductors [12–17], polymers [18,19], and glasses [20,21] as well as 3D PBG materials in systems that include layer-by-layer structures [22,23], inverse opals [24–26] and the fabrication of templates via laser holography [27,28] and direct laser writing [29–32].

3.1 Linear Waveguiding Structures

In conventional microoptical devices such as straight ridge waveguides and optical fibers, light is guided through the mechanism of total internal reflection inside a material with higher refractive index than the surrounding material. This guiding mechanism is lost when waveguides or fibers or distorted on a microscopic scale. In this case, light propagating in the waveguide or fiber can couple into the leaky modes provided by the background material and will escape from the optical circuit. As a result, microoptical circuits are limited to a few devices as otherwise losses become prohibitively high. This is in stark contrast to electronic microcircuits where electons are guided by thin metal wires. Electrons are bound within the cross section of the wire by the so-called work function (confining potential) of the metal. As a result of this rather different guiding mechanism, electrons follow the path prescribed by the wire without escaping into the background.

PBG materials "emulate" confining potentials for light by removing all the background electromagnetic modes over the relevant band of frequencies. More precisely, the PBG localizes the light and prevents it from escaping an optical microcircuit. This is to say, that light paths can be created inside PBGs material in the form of engineered waveguide channels. As a consequence, this PBG materials offers a viable platform for the creation of large-scale Photonic Integrated Circuits (PICs) which may ultimately result in a seamless all-optical network where communication between nearby computer chips and even within a single computer chip would take place with tiny beams of light rather than electricity.

However, the realization of complex PICs imposes more stringent requirements on the designs than just to have a PBG: In order to achieve acceptable performace and a certain robustness with respect to fabricational tolerances, it is imperative to minimize parasitic Fabry-Perot resonances between connecting elements such as waveguide bends, beamsplitters, and waveguide intersections. Ideally, all these connecting elements should be *non-reflecting* over a *broad* frequency range. In addition, the cross-talk between waveguides in an intersection should be strongly suppressed. We will return to these issues in Sect. 4.

3.2 Nonlinear Excitations

The response of optical materials to external excitations such as laser radiation is generally nonlinear. However, optical nonlinearities are often rather weak and one can argue that it should be possible to essentially simplify the problem by solving it in two steps. The first step is to linearize the governing equations and determine the ground state of the system under consideration together with its linear excitation spectrum. The effects of nonlinearity can subsequently be taken into account by means of appropriate perturbation theories. According to this approach, the effect of nonlinearity appears as an interaction between the linear excitations of the system. Although, this scheme may be useful for certain types of nonlinear problems, it fails to describe a wide range of interesting nonlinear phenomena. The reason for this is that nonlinear wave equations (and many other nonlinear differential equations appearing in other areas of physics) allow novel kinds of stable excitations which cannot be derived from the corresponding linearized equations. These inherently nonlinear solutions are called solitary waves or, in the particular case of integrable models, solitons. They simply vanish in the limit of infinitesimal wave amplitude and, therefore, cannot be considered within a finite order of (linear) perturbation theory. Instead, they should be treated as novel fundamental modes of the system.

For instance, if one of the constituent materials of a PBG structure exhibits a Kerr nonlinearity, i.e., an intensity dependent refractive index, the local intensity of the electromagnetic field can locally tune the PBG. As a consequence nonlinear PBG materials can become transparent to electromagnetic radiation with frequencies in the (linear) PBG. This leads to the formation of a special class of nonlinear excitations, the so-called gap solitons. The fundamental importance of these nonlinear excitations consists in the fact that they have a central frequency inside the PBG where linear excitations do not exist. Therefore, the direct observation of gap solitons would be the experimental proof that solitary waves represent fundamental excitations and cannot be reduced to linear excitations.

Maxwell's equations for TM-polarized light propagating in a nonlinear PC take the form

$$\left(\partial_x^2 + \partial_y^2\right) E(\boldsymbol{r},t) - \frac{\varepsilon_{\mathrm{P}}(\boldsymbol{r})}{c^2}\partial_t^2 E(\boldsymbol{r},t) = \frac{4\pi}{c^2}\partial_t^2 P_{\mathrm{NL}}(\boldsymbol{r},t)\ . \tag{15}$$

In writing this equation we have neglected the linear dispersion of the constituent materials, which is usually negligible compared to the dispersion associated with the photonic band structure.

To date, only a few works have been carried out for Kerr-nonlinearities [33–35] or for $\chi^{(2)}$-nonlinearities [36,37] in PCs. Moreover, the approximations involved in some of these works seriously limit the applicability of these theories to real PCs. For instance, the study of Kerr-nonlinearities in 2D PCs [33] has been limited to weak modulations in the linear index of refraction. Similarly, the recent investigation of second harmonic generation in 2D PCs [36,37] failed to reproduce the well-known results for the limiting case of homogeneous materials.

A systematic approach to quantitative calculations of the optical properties of nonlinear PCs is based on a multi-scale approach [38]. Since optical nonlinearities are generally quite weak, (15) should be solved in a perturbative way taking into account that the effect of the nonlinearity accumulates only on time and spatial scales that are much slower and longer, respectively, than the natural scales of the underlying linear problem. For electromagnetic wave propagation in PCs, these natural scales of the linear problem are determined through the inverse optical period and the associated wavelength of the light. Therefore, key simplifications to (15) arise from separating the fast from slow scales in space and time in the electromagnetic field [10]

$$E(\boldsymbol{r},t) = \mu e_1(\boldsymbol{r}_0, \boldsymbol{r}_1, \cdots; t_0, t_1, \cdots) + \mu^2 e_2(\boldsymbol{r}_0, \boldsymbol{r}_1, \cdots; t_0, t_1, \cdots) + \cdots\ , \tag{16}$$

by formally replacing the space and time variables, $\boldsymbol{r}$ and t, with a set of independent variables $\boldsymbol{r}_n \equiv \mu^n \boldsymbol{r}$ and $t_n \equiv \mu^n t$. Here, we denote the fastest spatial scale corresponding to the wavelength of the electromagnetic waves propagating in the linear PC by $\boldsymbol{r}_0$. Likewise, we denote the associated fastest temporal scale by t_0. Depending on the type of nonlinearity, the hierarchy is suitably truncated and a closed set of equations is obtained by collecting terms of equal order in μ.

To express the results in terms of the original physical variables, at the end of the calculations one has to set $\mu = 1$ [10].

As an illustration, we consider the case of the Kerr-nonlinear material, for which the refractive index depends on the light intensity leading to the nonlinear polarization $P_{\text{NL}}(\boldsymbol{r}, t) = \chi^{(3)}(\boldsymbol{r}) E^3(\boldsymbol{r}, t)$. Here, we have neglected the nonlinear material dispersion. In this case, substituting (16) into (15) and assuming that third-harmonic generation effects are not phase-matched and, hence, can be neglected, we obtain in the third order in μ that

$$e_1(\boldsymbol{r}_0, \boldsymbol{r}_1, \cdots; t_0, t_1, \cdots) = a_{n\boldsymbol{k}}(\boldsymbol{z}_1; \boldsymbol{r}_2, \cdots; t_1, \cdots)\, E_{n\boldsymbol{k}}(\boldsymbol{r}_0)\, e^{i\omega_{n\boldsymbol{k}} t_0} + c.c.\,, \quad (17)$$

where $\boldsymbol{z}_1 \equiv \boldsymbol{r}_1 - \boldsymbol{v}_{n\boldsymbol{k}} t_1$ with the group velocity $\boldsymbol{v}_{n\boldsymbol{k}}$ given by (13), the Bloch function $E_{n\boldsymbol{k}}(\boldsymbol{r}_0)$ represents a carrier wave and the envelope function $a_{n\boldsymbol{k}}(\boldsymbol{r}_1, \cdots; t_1, \cdots)$ has to be determined from the 2D nonlinear Schrödinger equation

$$\begin{aligned} &\left[i\left(\boldsymbol{v}_{n\boldsymbol{k}} \cdot \nabla_{\boldsymbol{r}_2} + \partial_{t_2}\right) + \nabla_{\boldsymbol{z}_1} \cdot \mathcal{M}_{n\boldsymbol{k}} \cdot \nabla_{\boldsymbol{z}_1}\right] a_{n\boldsymbol{k}}(\boldsymbol{z}_1; \boldsymbol{r}_2 \cdots; t_2, \cdots) \\ &\qquad + \alpha_{n\boldsymbol{k}} \left|a_{n\boldsymbol{k}}(\boldsymbol{z}_1; \boldsymbol{r}_2 \cdots; t_2, \cdots)\right|^2 a_{n\boldsymbol{k}}(\boldsymbol{z}_1; \boldsymbol{r}_2 \cdots; t_2, \cdots) = 0, \end{aligned} \quad (18)$$

where the GVD tensor $\mathcal{M}_{n\boldsymbol{k}}$ is given in (14) and the effective nonlinearity

$$\alpha_{n\boldsymbol{k}} = 6\pi\, \omega_{n\boldsymbol{k}} \int_{\text{WSC}} d^2 r\, \chi^{(3)}(\boldsymbol{r})\, |E_{n\boldsymbol{k}}(\boldsymbol{r})|^4 \quad (19)$$

reflects how the carrier wave $E_{n\boldsymbol{k}}(\boldsymbol{r})$ samples the spatial distribution $\chi^{(3)}(\boldsymbol{r})$ of nonlinear material within the PC.

The discussion of the solutions to (18) is outside the scope of the present work and we refer the reader to references on the inverse scattering theory and other methods [39]. However, we would like to note that (18) supports solitary wave solutions and is, therefore, appropriate for the discussion of gap solitons and similar analyses may be employed for discussing the interactions of various types of solitons [40,41]. In addition, we would like to emphasize the multi-scale approach introduced above represents an asymptotic expansion and cannot be reduced to standard perturbation theory. As a result of the foregoing analysis, we have instead obtained a generalization of the slowly varying envelope approximation. Within this approximation, the problem of pulse propagation in nonlinear PCs is mapped onto the problem of an envelope function propagating in an effective homogeneous medium with group velocity $\boldsymbol{v}_{n\boldsymbol{k}}$, GVD tensor $\mathcal{M}_{n\boldsymbol{k}}$, and effective nonlinearity $\alpha_{n\boldsymbol{k}}$ that are determined by the carrier wave, which, in turn, is given by a Bloch function of the linear PC. Therefore, the effective PC parameters can be obtained from band structure theory via (13), (14), and (19) and quantitative investigations become possible. Furthermore, we note that the above considerations are not limited to 2D TM polarized radiation and have recently been extended to 3D systems by Bhat and Sipe [34]. Moreover, the above framework of multi-scale analysis in conjunction with $\boldsymbol{k} \cdot \boldsymbol{p}$-perturbation theory can be applied to other nonlinear PC systems such as PCs consisting of nonresonant $\chi^{(2)}$ [41,42] material and resonant distributed feedback lasing systems [43]. In the present case of Kerr nonlinearities, other effects such as nonresonant soliton interactions can be considered and lead to interesting applications for the detection and manipulation of gap solitons [40,41].

3.3 Suppression and Control of Spontaneous Emission

An active material with a free space radiative transition that lies deep inside a PBG will be unable to emit a photon when placed inside a PBG material; instead, a photon–atom bound state is formed [44,46]. For transitions near the edge of a PBG, the emission dynamics will be modified relative to free space, due to the restricted number of photon modes available at the band edge [45,46]. The resulting non–Markovian atom–field interaction has been predicted to give rise to a number of novel quantum optical phenomena, such as laser–like collective atomic emission [47] and atomic states that can be readily generated and protected from processes that would serve to decohere the system [48]. These are but a few of the novel phenomena associated with the suppression and control over active material that can be achieved through PBG materials.

In a rotating wave approximation, the full quantum Hamiltonian for a two-level atom and the electromagnetic field in a PC can be written as [46]

$$\begin{aligned} H = \frac{\hbar}{2}\omega_{21}\sigma_z + \hbar\sum_{\mu}\omega_{\mu}a^{\dagger}_{\mu}a_{\mu} \\ + i\hbar\sum_{\mu}\left(g_{\mu}a^{\dagger}_{\mu}\sigma_{-} - g^{*}_{\mu}\sigma_{+}a_{\mu}\right) . \end{aligned} \tag{20}$$

The index μ labels the energy band and wavevector of a given field mode (Bloch function), $\mu \equiv \{n, \mathbf{k}\}$, and $a^{\dagger}_{\mu}$ and a_{μ} are the corresponding creation and annihilation operators for these modes, respectively. The σ_j $(j = +, -)$ are the usual Pauli operators for a two-level atom with a (bare) atomic resonance frequency ω_{21}. The position–dependent atom-field mode coupling constants, g_{μ}, are given by [46]

$$g_{\mu}(\boldsymbol{d}, \boldsymbol{r}_0) \equiv g_{\mu} = \omega_{21}d_{21}\sqrt{\frac{1}{2\hbar\epsilon_0\omega_{\mu}V_{\mathrm{WSC}}}}\boldsymbol{d}\cdot\boldsymbol{E}^{*}_{\mu}(\boldsymbol{r}_0) , \tag{21}$$

where d_{21} and $\boldsymbol{d}$ are respectively the magnitude and the direction unit vector of the dipole matrix element for the atomic transition.

We wish to analyze the atomic emission in a Schrödinger equation formalism [45,46]. Atom-field interactions that involve more than one photon are more easily (and often necessarily) described by a density matrix or by Heisenberg operator equations, and much of our analysis can be carried over to such systems. In the single photon sector of the atom-field Hilbert space, the wavefunction for a two–level atom with dipole moment $d_{21}\boldsymbol{d}$ is

$$|\Psi\rangle = b_2(\boldsymbol{d}, \boldsymbol{r}_0, t)\,|2, \{0\}\rangle + \sum_{\mu} b_{1,\mu}(\boldsymbol{d}, \boldsymbol{r}_0, t)\,|1, \{\mu\}\rangle\, e^{-i\Delta_{\mu}t} . \tag{22}$$

$b_2(\boldsymbol{d}, \boldsymbol{r}_0, t)$ and $b_{1,\mu}(\boldsymbol{d}, \boldsymbol{r}_0, t)$ label the probability amplitudes for the excited atom plus an electromagnetic vacuum state, and a de-excited atom with a single photon in mode μ, respectively, at a given position $\boldsymbol{r}_0$ of a Wigner-Seitz cell in a

PC; $\varDelta_\mu = \omega_\mu - \omega_{21}$. In a frame that is co-rotating with the bare atomic resonance frequency, ω_{21}, (22) along with the Hamiltonian (20) give the equations of motion for the amplitudes,

$$\frac{d}{dt} b_2(\boldsymbol{d}, \boldsymbol{r}_0, t) = -\sum_\mu g_\mu b_{1,\mu}(\boldsymbol{d}, \boldsymbol{r}_0, t) e^{-i\varDelta_\mu t}, \tag{23}$$

$$\frac{d}{dt} b_1(\boldsymbol{d}, \boldsymbol{r}_0, t) = g_\mu b_2(\boldsymbol{d}, \boldsymbol{r}_0, t) e^{i\varDelta_\mu t}. \tag{24}$$

Formally integrating (24), substituting the solution into (23), and averaging over the dipole orientation, we arrive at an equation for the dipole-averaged excited state amplitude $b_2(\boldsymbol{r}_0, t)$ [46]

$$\frac{d}{dt} b_2(\boldsymbol{r}_0, t) = -\int_0^t G(\boldsymbol{r}_0, t - t')\, b_2(\boldsymbol{r}_0, t') dt'. \tag{25}$$

$G(\boldsymbol{r}_0, t-t')$ is a time delay Green function, or memory kernel, which describes the effect of the modified electromagnetic vacuum on the atomic system at position $\boldsymbol{r}_0$; it is defined as

$$G(\boldsymbol{r}, \tau) = \varTheta(\tau)\, \beta \int_0^\infty d\omega\, \frac{N(\boldsymbol{r}, \omega)}{\omega}\, e^{-i(\omega - \omega_{21})\tau}, \tag{26}$$

where $N(\boldsymbol{r}, \omega)$ is the LDOS of the PC at $\boldsymbol{r}$ and we have absorbed all numerical factors into the prefactor $\beta = \omega_{21}^2 d_{21}^2 / 12\hbar\epsilon_0\pi^2$.

The occurence of the LDOS in (26) can be understood as follows: Consider an excited atom at some specific location within a PBG material. In order for the atom to decay via a single-photon process it needs to emit a photon into a Bloch mode of the PBG material. Consequently, it is the local coupling (overlap matrix element) of the atomic dipole moment to photons in this modes that determines the decay rate of the excited atom. Assuming that an allowed electric dipole transition is the dominant decay channel, we may combine (overall) mode availability, the DOS, and coupling to the mode in the LDOS as a measure of the local coupling strength between the atomic dipole moment and the modes of the PC.

We would like to emphasize that (25) is essentially exact and provides the basis for the fractional localization of the atomic population for atomic transition frequencies near a photonic band edge [44–46] as well as for the anomalous Lamb shift of atomic transition frequencies which can easily lead to corrections of the normal Lamb shift that are several tens of percent in magnitude with both, positive and negative signs [46]. Finally, almost any approximation to (25) rests on shaky foundations: For instance, given the rapidly oscillating exponent in the memory kernel, (26), one is tempted to treat the LDOS near the frequency ω_{21} of interest as constant, take it outside the frequency integral and evaluate this integral to be proportional to a delta-function in time. As a consequence, (25) would take on the form of a simple differential equation which can easily be solved by a decaying exponential function whose decay constant, i.e. rate of spontaneous emission, is proportional to the LDOS at frequency ω_{21}. While this is deceptively simple, it is just as wrong. If the LDOS varies rapidly near the atomic transition frequency ω_{21}, this approximation – generally known as Markov- or Wigner-Weiskopf approximation – cannot be justified. As Fig. 1 demonstrates, such rapid variations of the LDOS occur near the band edges and at van-Hove singularities inside the bands and their exact positions depend strongly on the parameters of the PC.

4 Linear Defect Structures in Photonic Crystals

To date, the overwhelming majority of theoretical investigations of cavities and waveguiding elements in PCs has been carried out using Finite-Difference Time-Domain (FDTD) and/or Finite-Element (FE) techniques. However, applying general purpose methodologies such as FDTD or FE methods to defect structures in PCs largely disregards information about the underlying PC structure which is readily available from photonic bandstructure computation. As a result, only relatively small systems can be investigated and the physical insight remains limited.

4.1 Maximally Localized Photonic Wannier Functions

A more natural description of localized defect modes in PCs consists in an expansion of the electromagnetic field into a set of localized basis functions which have encoded into them all the information of the underlying PC. Therefore, the most natural basis functions for the description of defect structures in PCs are the so-called photonic Wannier functions, $W_{n\boldsymbol{R}}(\boldsymbol{r})$, which are formally defined through a lattice Fourier transform

$$W_{n\boldsymbol{R}}(\boldsymbol{r}) = \frac{V_{\mathrm{WSC}}}{(2\pi)^2} \int_{\mathrm{BZ}} d^2\boldsymbol{k}\, e^{-i\boldsymbol{k}\boldsymbol{R}}\, E_{n\boldsymbol{k}}(\boldsymbol{r}) \tag{27}$$

of the extended Bloch functions, $E_{n\boldsymbol{k}}(\boldsymbol{r})$. The above definition associates the photonic Wannier function $W_{n\boldsymbol{R}}(\boldsymbol{r})$ with the frequency range covered by band n, and centers it around the corresponding lattice site $\boldsymbol{R}$. In addition, the completeness and orthogonality of the Bloch functions translate directly into corresponding properties of the photonic Wannier functions.

Computing the Wannier functions directly from the output of photonic bandstructure programs via (27) leads to functions with poor localization properties and erratic behavior (see, for instance, Fig. 2 in [49]). These problems originate from an indeterminacy of the global phases of the Bloch functions. It is straightforward to show that for a group of N_{W} bands there exists, for every wave vector $\boldsymbol{k}$, a free unitary transformation between the bands which leaves the orthogonality relation of Wannier functions unchanged. A solution to this unfortunate situation is provided by recent advances in electronic bandstructure theory. Marzari and Vanderbilt [50] have outlined an efficient scheme for the

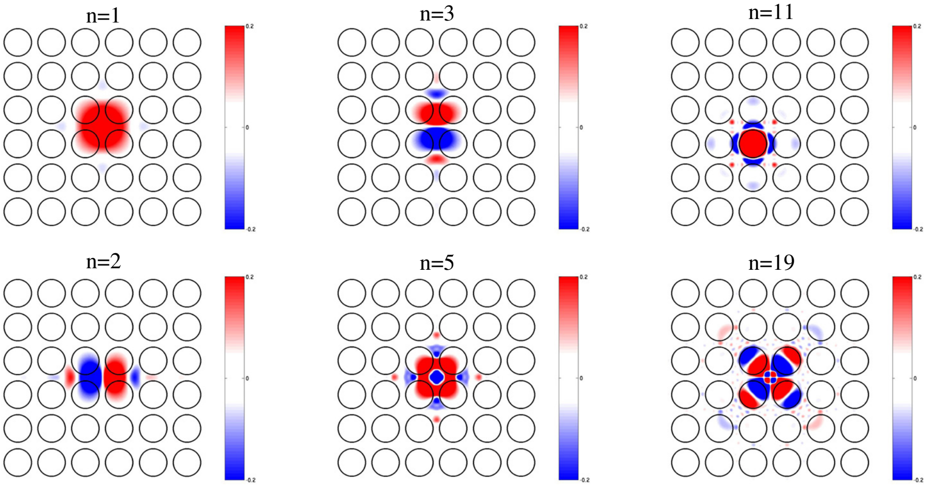


Fig. 3. Photonic Wannier functions, $W_{n\mathbf{0}}(\boldsymbol{r})$, for the six bands that are most relevant for the description of the localized defect mode shown in Fig. 4b. These optimally localized Wannier functions have been obtained by minimizing the corresponding spread functional, (28). Note, that in contrast to the other bands, the Wannier center of the eleventh band is located at the center of the air pore. The parameters of the underlying PBG material are the same as those in Fig. 1

computation of maximally localized Wannier functions by determining numerically a unitary transformation between the bands that minimizes an appropriate spread functional $\mathcal{F}$

$$\mathcal{F} = \sum_{n=1}^{N_{\mathrm{W}}} \left[\langle n\mathbf{0}|\, r^2 \,|n\mathbf{0}\rangle - (\langle n\mathbf{0}|\, \boldsymbol{r} \,|n\mathbf{0}\rangle)^2 \right] = \mathrm{Min}\,. \tag{28}$$

Here we have introduced a shorthand notation for matrix elements according to

$$\langle n\boldsymbol{R}|\, f(\boldsymbol{r})\, |n'\boldsymbol{R}'\rangle = \int_{\mathbf{R}^2} d^2r\, W^*_{n\boldsymbol{R}}(\boldsymbol{r})\, f(\boldsymbol{r})\, \varepsilon_{\mathrm{p}}(\boldsymbol{r})\, W_{n'\boldsymbol{R}'}(\boldsymbol{r})\,, \tag{29}$$

for any function $f(\boldsymbol{r})$. For instance, the orthonormality of the Wannier functions in this notation read as

$$\langle n\boldsymbol{R}|\, |n'\boldsymbol{R}'\rangle = \int_{\mathbf{R}^2} d^2r\, W^*_{n\boldsymbol{R}}(\boldsymbol{r})\, \varepsilon_{\mathrm{p}}(\boldsymbol{r})\, W_{n'\boldsymbol{R}'}(\boldsymbol{r}) = \delta_{nm}\delta_{\boldsymbol{R}\boldsymbol{R}'}\,. \tag{30}$$

The field distributions of the optimized Wannier functions belonging to the six most relevant bands of our model system are depicted in Fig. 3. Their localization properties as well as the symmetries of the underlying PC structure are clearly visible. It should be noted that the Wannier centers of all calculated bands (except of the eleventh band) are located halfway between the air pores, i.e. inside the dielectric (see [50] for more details on the Wannier centers). In this context, most relevant means that the corresponding Wannier functions are sufficient to quantitatively simulate the defect structures considered below to an accuracy better than 0.5%. This is reflected in the so-called V-parameter for a single cavity (see [49]). In addition, we would like to point out that instead of working with the electric field [51,49], (1), one may equally well construct photonic Wannier functions for the magnetic field, as recently demonstrated by Whittaker and Croucher [52].

4.2 Defect Structures via Wannier Functions

The description of defect structures embedded in PCs starts with the corresponding wave equation in the frequency domain

$$\nabla^2 E(\boldsymbol{r}) + \left(\frac{\omega}{c}\right)^2 \left(\varepsilon_{\mathrm{p}}(\boldsymbol{r}) + \delta\varepsilon(\boldsymbol{r})\right) E(\boldsymbol{r}) = 0 \,. \tag{31}$$

Here, we have decomposed the dielectric function into the periodic part, $\varepsilon_{\mathrm{p}}(\boldsymbol{r})$, and the contribution, $\delta\varepsilon(\boldsymbol{r})$, that describes the defect structures. Within the Wannier function approach, we expand the electromagnetic field according to

$$E(\boldsymbol{r}) = \sum_{n,\boldsymbol{R}} E_{n\boldsymbol{R}} \, W_{n\boldsymbol{R}}(\boldsymbol{r}) \,, \tag{32}$$

with unknown amplitudes $E_{n\boldsymbol{R}}$. Inserting this expansion into the wave equation (31) and employing the orthonomality relations, (30), leads to the basic equation for lattice models of defect structures embedded in PCs

$$\sum_{n',\boldsymbol{R}'} \left\{ \delta_{nn'} \delta_{\boldsymbol{R}\boldsymbol{R}'} + D^{nn'}_{\boldsymbol{R}\boldsymbol{R}'} \right\} E_{n'\boldsymbol{R}'} = \left(\frac{c}{\omega}\right)^2 \sum_{n',\boldsymbol{R}'} A^{nn'}_{\boldsymbol{R}\boldsymbol{R}'} E_{n'\boldsymbol{R}'} \,. \tag{33}$$

The matrix $A^{nn'}_{\boldsymbol{R}\boldsymbol{R}'}$ depends only on the Wannier functions of the underlying PC and is defined through

$$A^{nn'}_{\boldsymbol{R}\boldsymbol{R}'} = - \int_{\mathbf{R}^2} d^2\boldsymbol{r} \;\; W^*_{n\boldsymbol{R}}(\boldsymbol{r}) \, \nabla^2 \, W_{n'\boldsymbol{R}'}(\boldsymbol{r}) \,. \tag{34}$$

The localization of the Wannier functions in space leads to a very rapid decay of the magnitude of matrix elements with increasing separation $|\boldsymbol{R}-\boldsymbol{R}'|$ between lattice sites, effectively making the matrix $A^{nn'}_{\boldsymbol{R}\boldsymbol{R}'}$ sparse. Furthermore, it may be shown that the matrix $A^{nn'}_{\boldsymbol{R}\boldsymbol{R}'}$ is Hermitian and positive definite. Similarly, once the Wannier functions of the underlying PC are determined, the matrix $D^{nn'}_{\boldsymbol{R}\boldsymbol{R}'}$ depends solely on the overlap of these functions, mediated by the defect structure:

$$D^{nn'}_{\boldsymbol{R}\boldsymbol{R}'} = \int_{\mathbf{R}^2} d^2\boldsymbol{r} \;\; W^*_{n\boldsymbol{R}}(\boldsymbol{r}) \, \delta\varepsilon(\boldsymbol{r}) \, W_{n'\boldsymbol{R}'}(\boldsymbol{r}) \,. \tag{35}$$

As a consequence of the localization properties of both the Wannier functions and the defect dielectric function, the Hermitian matrix $D^{nn'}_{\boldsymbol{R}\boldsymbol{R}'}$, too, is sparse. In the case of PCs with inversion symmetry, $\varepsilon_{\mathrm{p}}(\boldsymbol{r}) \equiv \varepsilon_{\mathrm{p}}(-\boldsymbol{r})$, the Wannier functions can be chosen to be real. Accordingly, both matrices, $A^{nn'}_{\boldsymbol{R}\boldsymbol{R}'}$ and $D^{nn'}_{\boldsymbol{R}\boldsymbol{R}'}$ become real symmetric ones.

Depending on the nature of the defect structure, we are interested in (i) frequencies of localized cavity modes, (ii) dispersion relations for straight waveguides, or (iii) transmission and reflection through waveguide bends and other, more complex defect structures. Each of these cases can be solved by carefully analyzing the central equation (33) and we would like to refer to [49] for details.

In the following, we take up the discussion of Sect. 2 pertaining to the realization of large-scale PICs using PBG materials.

4.3 PBG Materials and Photonic Integrated Circuits

Research in this area has been initiated in 1996, when it has been realized [53] that removing a single row of rods in a two-dimensional (2D) PC consisting of a square lattice of dielectric rods creates a *broad-band mono-mode* waveguide for TM-polarized light. In addition, reflection from waveguide bends in these systems does not exceed 5% over a wide frequency range [53] and changing the radii of certain rods facilitates the design of relatively broad-band low-reflecting beamsplitters [55] (see [54] for further references). Unfortunately, since the removal of a single rod creates a monopole-type cavity mode, the elimination of parasitic cross-talk between the waveguides in an intersection requires the usage of high-Q resonances. This significantly narrows the free bandwidth for these systems [56]. In addition, any real rod-based structure would consist of finite height rods so that light propagating in the resulting air waveguide cannot be guided in the third dimension. To circumvent these problems, it has recently been suggested to sandwich such structures between properly designed 3D PCs [57]. Clearly, such an approach requires highly advanced 3D fabrication techniques.

As a result, 2D PCs for application at optical wavelengths typically consist of arrays of air pores that have been etched into high-refractive index materials such as macroporous silicon, GaAs, or InP through standard semiconductor processing technologies [12–17] .

The majority of defects studied in such systems consist of missing pores and, therefore, guidance in the third direction may be realized. In these PCs, PBGs for TE-polarized light are typically larger than PBGs for TM-polarized light and substantial efforts have been devoted to the design of efficient functional elements for TE-polarized light. However, in this case a single missing pore creates a *doubly-degenerate* dipole-like cavity mode. Consequently, the resulting PC waveguides are *intrinsically multi-moded*. Therefore, despite the large PBGs, the frequency regions of mono-mode operation of straight waveguides are rather narrow. More importantly, guiding light around bends is rather inefficient in the TE-polarized case because (i) the waveguides may become multi-moded in the vicinity of the bends and undesired cross-coupling between different modes occurs and (ii) there exists a impedance mismatch between the waveguide modes in the different leads due to their asymmetric coupling to the dipole-like cavity modes of the bend. Nevertheless, we would like to note that this asymmetric p-

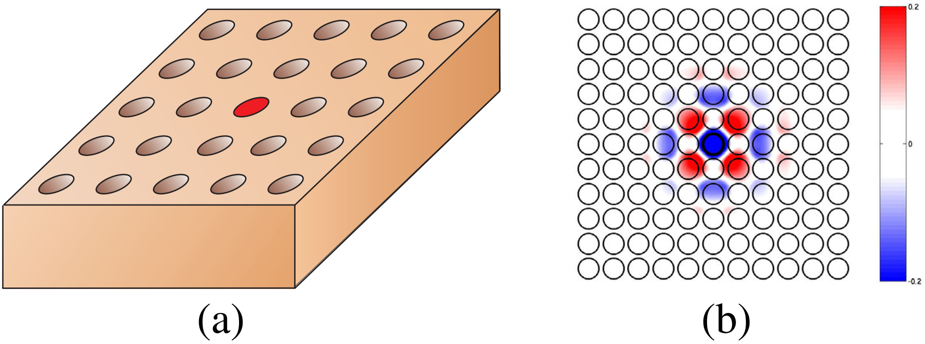


Fig. 4. (a) Schematic of a 2D PBG material consisting of a square lattice of air pores, with a single pore infilled with a liquid crystal or a polymer. **(b)** Electric field distribution for the corresponding non-degenerate cavity mode for TM-polarized light

coupling has been exploited for the design of relatively broad-band low-crosstalk waveguide intersections [58].

In this chapter, we suggest a novel approach to tunable PC circuits that combines several attractive advantages: (i) The PC circuits are based on 2D PBG materials consisting of air pores in high-refractive-index dielectrics and, therefore, are easily fabricated. (ii) They exploit *non-degenerate* defect modes created for *TM-polarized light* by infilling individual pores with appropriate low or moderate refractive index materials such as liquid crystals and/or polymers. By construction, this leads to *essentially mono-mode* PC waveguides. Furthermore, a peculiar symmetry of this cavity modes may be exploited to *simultaneously* obtain design for broad-band non-reflecting waveguides and beamsplitters as well as broad-band low-crosstalk waveguide intersections. (iii) Owing to the tunability of the infilled materials the resulting circuits will be tunable.

Within this context, we would like to note that the idea to infiltrate PCs with liquid crystals to achieve tunable bandstructures has been suggested theoretically [7] and validated experimentally for 2D [61,62] and 3D [59,60] PCs. In addition, a tunable beamsplitter for TE-polarized light in 2D PCs with liquid-crystal infilled pores has recently been suggest theoretically [63]. Unfortunately, the corresponding cavity mode is doubly degenerate and dipole-like. As a result, the PC waveguides are intrinsically multi-moded, and it remains a challenge to design broad-bend non-reflecting waveguide bends, beamsplitters, and intersections.

Infilling a material with refractive index $n_{\text{def}} = \sqrt{\varepsilon_{\text{def}}} = 1.55$ into a single pore of our model PBG material (see Fig. 1) gives rise to a non-degenerate monopole-like cavity mode as depicted in Fig. 4. In Fig. 5a we display the dispersion relation for the propagating guided modes of a straight PC waveguide created by infilling a material with refractive index $n_{\text{def}} = \sqrt{\varepsilon_{\text{def}}} = 1.55$ into a single row of pores. Being based on a non-degenerate cavity mode, this PC waveguides is mono-moded throughout the entire available frequency range. The bandwidth of this mode can be significantly increased either by increasing the radius of the pores

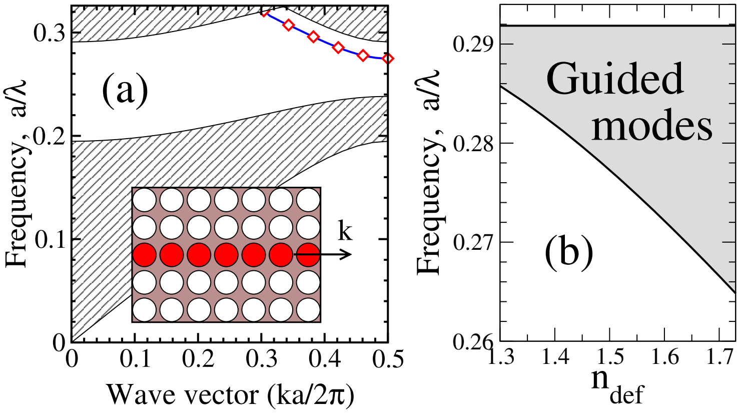


Fig. 5. **(a)** Dispersion relation for a PC waveguide obtained by infilling a single row of pores with a material with refractive index $n_{\mathrm{def}} = \sqrt{\varepsilon_{\mathrm{def}}} = 1.55$. The hatched areas represent the projected band structure of the underlying PC. **(b)** Bandwidth of the guided modes (shaded area) of the same PC waveguide as a function of the refractive index n_{def} of the infilled material

or by increasing the refractive index of the infilled material, as demonstrated in Fig. 5b.

4.4 Functional Elements for Photonic Integrated Circuits

After having introduced the novel concept of cavities and waveguides based on the infilling of pores of high-refractive index PBG materials with low-refractive index liquid crystals or polymers, we now turn to the problem of developing concepts and designs for functional elements based on the same principle. As alluded to in Sect. 2, in order to minimize parasitic Fabry-Perot resonances, these elements should be *non-reflecting* over a *broad* frequency range. In addition, the cross-talk between waveguides in an intersection should be strongly suppressed. In Fig. 6 we present the transmission spectra for three different designs of a 90° waveguide bend. The designs that have been successfully used for TM-polarized light in rod-based PCs [53] appear to be extremely inefficient in our case. For instance, the transmission through the waveguide bend depicted in Fig. 6a is almost zero over the entire frequency range of interest. In fact, our initial attempts at improving these results by "rounding" the waveguide bend have failed completely. However, once we ignored this standard procedures employed in other systems and started to design "photon hopping paths" based on the field distribution of the cavity mode displayed in Fig. 4b, we have been able to arrive at successful designs for broad-band low-reflection waveguide bends as illustrated in Figs. 6b and 6c. As depicted in Fig. 7 for our optimized beamsplitter design, the "photon hopping picture" is an appropriate description for mono-mode waveguide systems in PCs.

To design an efficient waveguide intersection, we have utilized the "inefficient" design of a waveguide bend shown in Fig. 6a. In this case (see Fig. 8), we

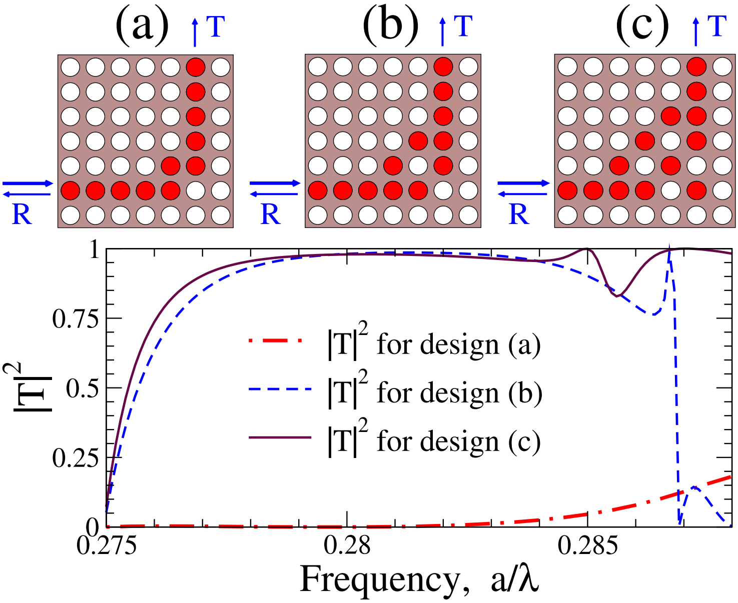


Fig. 6. Transmission spectra for three different designs of waveguide bends as illustrated in the upper drawings.

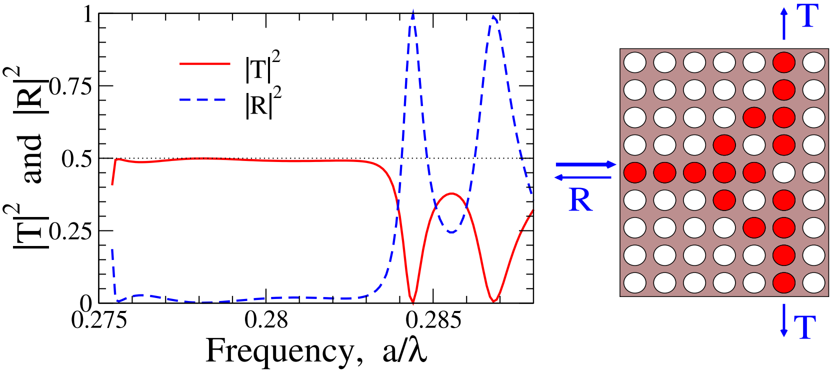


Fig. 7. Transmission and reflection spectra for our optimized design of a the beamsplitter whose schematic is shown in the panel.

have been able to almost completely eliminate parasitic cross-talk between the waveguides without having to take recourse to high-Q resonances as suggested in [56]. As a result, our waveguide intersection operates over a broad range of frequencies. A careful inspection of Fig. 8 reveals that the reflection from the intersection vanishes only in a relatively small frequency range. However, we can easily shift and extend this region by changing the refractive index in the four corner pores. More generally, infilling these pores with materials exhibiting different refractive indices allows us to further improve the characteristics of all

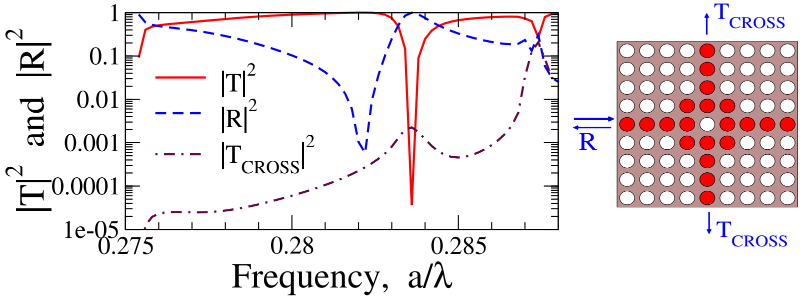


Fig. 8. Transmission and reflection spectra for a broad-band low-crosstalk design of a waveguide intersection whose design is shown in the right panel

the devices discussed above. This suggests that the tunability of the infiltrated materials will provide a corresponding tunability to the devices designed above.

5 Conclusions and Outlook

In summary, we have outlined a framework based on solid-state theoretical methods that allows one to qualitatively and quantitatively treat electromagnetic wave propagation in PCs. Photonic bandstructure computations for infinitely extended PCs provides photonic bandstructures, identifies PBGs and allows us to calculate other physical quantities such as DOS, LDOS, group velocities and group velocity dispersion.

The description of nonlinear PCs through a multi-scale approach facilitates the systematic construction of generalized slowly varying envelope approximations which, in turn, allow us to quantitatively investigate such systems using a limited number of effective parameters with transparent physical meaning. In addition, we have shown how the LDOS determines the radiative characteristics of active material embedded in PCs. This shows that standard approximations of the corresponding equations of motion are generally insufficient and the full non-Markovian problem needs to be solved instead.

Furthermore, the input of bandstructure calculations facilitates the construction of maximally localized photonic Wannier functions which allow us to efficiently obtain the properties of defect structures embedded in PCs. In particular, the efficiency of the Wannier function approach allows us to investigate large-scale PC circuits which, to date, are beyond the reach of standard simulation techniques such as FDTD or FE methods.

In addition, we suggest that the infiltration of low-refractive index materials into air pores of bulk high-index 2D PBG materials provides a novel platform for ultra-compact PICs using TM-polarized radiation. Owing to a non-degenerated cavity mode with a peculiar field distribution, we have designed mono-mode PC waveguides and a number of broad-band non-reflecting functional elements such as bends, beamsplitters, and low-crosstalk waveguide intersections. These functional elements may be realized by infiltrating different types of liquid crystals and/or polymers into appropriate 2D PBG materials (for a recent overview of the current state of microinfiltration see [64]). Our concept opens numerous avenues for tunable PC circuits based on the tunability of the infiltrated materials which may enhance the utility of these composite system over and above the conventional PC circuits.

Future work will be aimed at the design of novel devices with prescribed properties such as folded directional couplers, the investigation of devices based on triangular PBG materials, and the extension to nonlinear defect structures.

Acknowledgments

This work was supported by the Center for Functional Nanostructures (CFN) of the Deutsche Forschungsgemeinschaft (DFG) within project A 1.2. and A 1.3. The research of K.B. and L.T. is further supported by the DFG under grant Bu 1107/2-3 (Emmy-Noether program). The work of M.S. is funded in the framework of the DFG Research Training Group 786 *Mixed Fields and Nonlinear Interactions* at the University of Karlsruhe.

References

1. E. Yablonovitch: Phys. Rev. Lett. **58**, 2059 (1987)
2. S. John: Phys. Rev. Lett. **58**, 2486 (1987)
3. S. John, O. Toader, K. Busch: *Encyclopedia of Science and Technology*, Vol 12 (Academic Press 2001)
4. K.-M. Ho, C.T. Chan, C.M. Soukoulis: Phys. Rev. Lett. **65**, 3152 (1990)
5. N.W. Ashcroft, N.D. Mermin: *Solid State Physics* (Saunders College Publishing 1976)
6. K. Busch, S. John: Phys. Rev. E **58**, 3896 (1998)
7. K. Busch, S. John: Phys. Rev. Lett. **83**, 967 (1999)
8. D. Hermann, M. Frank, K. Busch, P. Wölfle: Optics Express **8**, 167 (2001)
9. A. Brandt, S. McCormick, J. Ruge: SIAM J. Sci. Stat. Comput. **4**, 244 (1983)
10. C. Martijn de Sterke, J.E. Sipe: Phys. Rev. A **38**, 5149 (1988)
11. J.E. Sipe: Phys. Rev. E **62**, 5672 (2000)
12. A. Birner, R.B. Wehrspohn, U.M. Gösele, K. Busch: Adv. Mater. **13**, 377 (2001)
13. T.F. Krauss, R.M. de la Rue: Prog. Quantum Electron. **23**, 51 (1999)
14. A. Forchel et al.: Microelectron. Eng. **53**, 21 (2000)
15. M. Loncar, T. Doll, J. Vuckovic, A. Scherer: J. Lightwave Technol. **18**, 1402 (2000)
16. H. Benisty et al.: IEEE J. Quantum Electron. **38**, 770 (2002)
17. S. Noda, M. Imada, A. Chutinan, N. Yamamoto: Opt. Quantum Electron. **34**, 723 (2002)
18. C. Liguda et al.: Appl. Phys. Lett. **78**, 2434 (2001)
19. A.C. Edrington et al.: Adv. Mater. **13**, 421 (2001)
20. A. Rosenberg, R.J. Tonucci, H.B. Lin, E.L. Shirley: Phys. Rev. B **54**, R5195 (1996)
21. O.J.A. Schueller et al.: Appl. Opt. **38**, 5799 (1999)
22. S.-Y. Lin et al.: Nature **394**, 251 (1998)
23. S. Noda, K. Tomoda, N. Yamamoto, A. Chutinan: Science **289**, 604 (2000)
24. J.E.G.J. Wijnhoven, W.L. Vos: Science **281**, 802 (1998)

25. A. Blanco et al.: Nature **405**, 437 (2000)
26. Y.A. Vlasov, X.Z. Bo, J.C. Sturm, D.J. Norris: Nature **414**, 289 (2001)
27. M. Campbell et al.: Nature **404**, 53 (2000)
28. Y.V. Miklyaev et al.: Appl. Phys. Lett. **82**, 1284 (2003)
29. H.B. Sun, S. Matsuo, H. Misawa: Appl. Phys. Lett. **74**, 786 (1999)
30. H.B. Sun et al.: Appl. Phys. Lett. **79**, 1 (2001)
31. M. Straub, M. Gu: Opt. Lett. **27**, 1824 (2002)
32. M. Deubel, G. Von Freymann, M. Wegener, S. Pereira, K. Busch, C.M. Soukoulis: Nature Materials (in press)
33. N. Aközbek, S. John: Phys. Rev. E **57**, 2287 (1998)
34. N. Bhat, J. Sipe: Phys. Rev. E **64**, 056604 (2001)
35. S.F. Mingaleev, Yu.S. Kivshar: Phys. Rev. Lett. **86**, 5474 (2001)
36. K. Sakoda: *Optical properties of Photonic Crystals*, (Springer, Berlin, Heidelberg, New York 2001)
37. K. Sakoda, K. Ohtaka: Phys. Rev. B **54**, 5742 (1996)
38. A.H. Nayfeh: *Perturbation Methods* (Wiley, New York 1973)
39. R.K. Dodd, J.C. Eilbeck, J.D. Gibbon, H.C. Morris: *Solitons and Nonlinear Wave Equations* (Academic Press, London 1982)
40. L. Tkeshelashvili, S. Pereira, K. Busch: submitted (2004)
41. L. Tkeshelashvili: Interaction of Nonlinear Waves in Photonic Crystals. PhD Thesis, University of Karlsruhe (2003)
42. L. Tkeshelashvili, K. Busch: submitted (2004)
43. L. Florescu, K. Busch, S. John: J. Opt. Soc. Am. B **19**, 2215 (2002)
44. S. John, J. Wang: Phys. Rev. B **43**, 12772 (1991)
45. S. John, T. Quang: Phys. Rev. A **50**, 1764 (1994)
46. N. Vats, S. John, K. Busch: Phys. Rev. A **65**, 043808 (2002)
47. N. Vats, S. John: Phys. Rev. A **58**, 4168 (1998)
48. M. Woldeyohannes, S. John: Phys. Rev. A **60**, 5046 (1999)
49. K. Busch, S.F. Mingaleev, A. Garcia-Martin, M. Schillinger, D. Hermann: J. Phys.: Condens. Matter **15**, R1233 (2003)
50. N. Marzari, D. Vanderbilt: Phys. Rev. B **56**, 12847 (1997)
51. A. Garcia-Martin, D. Hermann, K. Busch, P. Wölfle: Mater. Res. Soc. Symp. Proc. **722**, L 1.1 (2002)
52. D.M. Whittaker, M.P. Croucher: Phys. Rev. B **67**, 085204 (2003)
53. A. Mekis et al.: Phys. Rev. Lett. **77**, 3787 (1996)
54. S.F. Mingaleev, Yu.S. Kivshar: J. Opt. Soc. Am. B **19**, 2241 (2002)
55. S. Fan et al.: J. Opt. Soc. Am. B **18**, 162 (2001)
56. S.G. Johnson et al.: Opt. Lett. **23**, 1855 (1998)
57. A. Chutinan, S. John, O. Toader: Phys. Rev. Lett. **90**, 123901 (2003)
58. S. Lan, H. Ishikawa: Opt. Lett. **27**, 1567 (2002)
59. K. Yoshino et al.: Appl. Phys. Lett. **75**, 932 (1999)
60. G. Mertens et al.: Appl. Phys. Lett. **83**, 3036 (2003)
61. S.W. Leonard et al.: Phys. Rev. B **61**, R2389 (2000)
62. Ch. Schuller et al.: Appl. Phys. Lett. **82**, 2767 (2003)
63. H. Takeda, K. Yoshino: Phys. Rev. B **67**, 073106 (2003)
64. S. Gottardo, D.S. Wiersma, W.L. Vos: Physica B **338**, 143 (2003)

Circular Photo-Galvanic and Spin-Galvanic Effects

Eugeniyus L. Ivchenko

A.F. Ioffe Physico-Technical Institute, RAS, 194021 St. Petersburg, Russia

Introduction

Light propagating through a semiconductor and acting upon mobile carriers can generate a dc electric current, under short-circuit condition, or a voltage, in case of open-circuit samples. In this lecture we consider only the Photo-Galvanic Effects (PGE) which, by definition, appear *not* due to inhomogeneity of optical excitation of electron-hole pairs, as in the Dember and Photo-Electro-Magnetic Effects, and *not* due to inhomogeneity of the sample, as in the conventional Photo-Voltaic Effect in p-n junctions. Phenomenologically, they are described by the following equation

$$j_\lambda = I \left[\gamma_{\lambda\mu} \mathrm{i} \left(\boldsymbol{e} \times \boldsymbol{e}^* \right)_\mu + \chi_{\lambda\mu\nu} \frac{e_\mu e_\nu^* + e_\nu e_\mu^*}{2} + T_{\lambda\mu\nu\eta} q_\mu e_\nu e_\eta^* \right] \tag{1}$$

which relates the dc current density with the light intensity I, polarization $\boldsymbol{e}$ and wave vector $\boldsymbol{q}$. In a bulk semiconductor or superlattice the index λ runs over all three Cartesian coordinates x, y, z. In quantum well (QW) structures the free-carrier motion along the growth direction is quantized and the index λ enumerates two interface coordinates. In quantum wires and nanotubes the free movement is allowed only along one axis, the principal axis of the structure, and the coordinate λ is parallel to this axis. On the other hand, the light polarization unit vector $\boldsymbol{e}$ can be arbitrarily oriented in space and, therefore, $\mu, \nu = x, y, z$. Note that, for linearly polarized light, the complex conjugate vector $\boldsymbol{e}^*$ is parallel to $\boldsymbol{e}$ and the vector product $\boldsymbol{e} \times \boldsymbol{e}^*$ vanishes. For elliptically polarized electromagnetic wave, the vector $i\left(\boldsymbol{e} \times \boldsymbol{e}^*\right)$ is real and proportional to the degree of circular polarization P_c; for a transverse wave it can be presented as a product $P_c \hat{\boldsymbol{o}}$ where $\hat{\boldsymbol{o}}$ is a unit vector in the direction of light propagation.

The tensor $\boldsymbol{\gamma}$ in (1) relates components of the polar vector $\boldsymbol{j}$ and the axial vector $\boldsymbol{e} \times \boldsymbol{e}^*$. It is non-zero for point groups which allow optical activity or gyrotropy. The effect described by this tensor is called the *circular* PGE. It appears only under illumination with circularly polarized light and reverses direction when the sign of circular polarization is changed.

The effect described by the second term in (1) is called the *linear* PGE. The reason is that it is independent on the sign of circular polarization and usually measured under linearly polarized photoexcitation. The third-rank tensor $\boldsymbol{\chi}$ in (1) is invariant under interchange of indices μ and ν. Therefore, the linear PGE can be observed in non-centrosymmetric media of the piezoelectric classes.

E.L. Ivchenko, Circular Photo-Galvanic and Spin-Galvanic Effects, Lect. Notes Phys. **658**, 23–50 (2005)
http://www.springerlink.com/

The third term on the right-hand side of (1) represents the Photon Drag Effect. It is due to momentum transfer from photons to charge carriers and can be induced in both non-centrosymmetric and centrosymmetric systems.

1 Circular Photo-Galvanic Effect in Quantum Well Structures

Physically, the circular PGE can be considered as a transformation of the photon angular momenta into a translational motion of free charge carriers. It is an electronic analog of mechanical systems which transmit rotatory motion to linear one like a screw thread or a plane with propeller. The effect was independently predicted by Ivchenko and Pikus [1] and Belinicher [2] and then studied both theoretically and experimentally in bulk gyrotropic crystals (see the review article [3] and the book [4]), particularly in Tellurium [5,6], and recently in zinc-blende- and diamond-based QW structures [7–10].

Here, we will perform the symmetry analysis of the circular PGE in (001)- and (113)-grown QWs, present demonstrational experimental data and outline the microscopic theory of the effect under interband, intersubband and intrasubband optical transitions in QWs.

The three point groups D_{2d}, C_{2v} and C_s are particularly relevant in connection with the photo-galvanic experiments on zinc-blende-based QW structures; hereafter the Schönflies notation is used to label the point groups. In the international notation they are labelled as $\bar{4}2m, mm2$ and m, respectively. A (001)-grown QW with equivalent normal and inverted interfaces has the D_{2d} point-group symmetry. The point group reduces from D_{2d} to C_{2v} in symmetrical QWs with built-in electric fields or asymmetrical QWs, say compositionally stepped QWs, QWs with different profiles of the left and right interfaces etc. If QWs are grown along the low-symmetry axis $[hhl] \neq [001]$ and $[111]$, the point group becomes C_s and contains only two elements, the identity and one mirror reflection plane $\sigma_{(1\bar{1}0)}$. In the case $h = l = 1$, the QW point symmetry increases up to C_{3v}.

For the point group C_s, in the coordinate system $x \parallel [1\bar{1}0]$, $y \parallel [ll(\overline{2h})]$, $z \parallel [hhl]$ the y- and z-components of a polar vector and x-component of an axial vector are invariants (the representation A^+ of the C_s group), the x-component of a polar vector and y- and z-components of an axial vector transform according to the representation A^-. As a result, the first term in (1) can be rewritten as

$$j_x = (\gamma_{xy} o_y + \gamma_{xz} o_z) I P_c \; , \; j_y = \gamma_{yx} o_x I P_c \; . \tag{2}$$

For the point group C_{2v}, in the coordinate system $x \parallel [1\bar{1}0]$, $y \parallel [110]$, $z \parallel [001]$ the component γ_{xz} is zero and the equations (2) reduce to

$$j_x = \gamma_{xy} o_y I P_c \; , \; j_y = \gamma_{yx} o_x I P_c \; . \tag{3}$$

Finally, for the point group D_{2d}, in the above coordinate system the same equations are also valid but the higher symmetry imposes the condition $\gamma_{xy} = \gamma_{yx} \equiv \gamma$ on the $\boldsymbol{\gamma}$ tensor and one has

$$j_x = \gamma o_y I P_c \; , \; j_y = \gamma o_x I P_c \; . \tag{4}$$

It follows from (2)-(4) that, in QWs of the C_s symmetry, the circular PGE can be observed even under normal incidence of irradiation while, in QWs of the C_{2v} or D_{2d} symmetry, the circular photocurrent can be generated only under oblique incidence.

Figure 1 shows results of measurements carried out at room temperature on (113)-grown p-GaAs/AlGaAs MQWs under normal incidence (upper panel) and (001)-grown n-InAs/AlGaSb SQW structure under oblique incidence with an angle of incidence in vacuum $\theta_0 = -30°$ (lower panel). Optical excitation was performed by a high power far infrared pulsed NH_3 laser which yields strong linearly polarized emission at wavelengths λ between 35 and 280 μm corresponding to photon energies from 35 to 4.4 meV with power up to 100 kW. The linearly polarized light could be modified to an elliptically polarized radiation by applying a crystalline quartz $\lambda/4$ plate and changing the angle φ between the optical axis of the plate and the polarization plane of the laser radiation. Thus the helicity P_c of the incident light varies from -1 (left handed, σ_-) to $+1$ (right handed, σ_+) according to

$$P_c = \sin 2\varphi \,. \tag{5}$$

One can see from Fig. 1 that the photocurrent direction is reversed when the polarization switches from right-handed circular, $\varphi = 45°$, to left-handed, $\varphi = 135°$. Moreover, the experimental points are well fitted by the equation

$$j_\lambda(\varphi) = j_\lambda^0 \sin 2\varphi \tag{6}$$

with one scaling parameter j_λ^0.

In Fig. 2 a closer look is taken at the dependence of the photocurrent on the angle of incidence θ_0 in configuration with the incidence plane normal to the axis x. According to (2) the photocurrent induced along x in (113)-oriented QWs is given by

$$j_x = (\gamma_{xy} \sin\theta + \gamma_{xz} \cos\theta) t_p t_s I_0 P_c \,, \tag{7}$$

where I_0 is the light intensity in vacuum, t_p and t_s are transmission coefficients after Fresnel's formula for linear p and s polarizations, θ is the refraction angle defined by $\sin\theta = \sin\theta_0/n$, and n is the index of refraction. In this case the circular PGE is observed at normal incidence. The fact that j_x is an even function of θ_0 means that in the sample under study the component γ_{xz} of the $\boldsymbol{\gamma}$ tensor is much larger as compared with γ_{xy}. In (001)-oriented samples where $\gamma_{xz} = 0$ a signal proportional to $\sin 2\varphi$ is only observed under oblique incidence and a variation of θ_0 in the plane of incidence changes the sign of the current j_x exactly at the point $\theta_0 = 0$.

Microscopically, a conversion of photon helicity into a current can be related to $\boldsymbol{k}$-linear terms in the effective Hamiltonian $\mathcal{H}^{(1)} = \beta_{lm}\sigma_l k_m$ (see Appendix for the information concerning these terms). The coefficients β_{lm} form a pseudo-tensor subjected to the same symmetry restriction as the pseudo-tensor $\boldsymbol{\gamma}$. The coupling between the spin Pauli matrices σ_l and the wave vector components k_m as well as spin-dependent selection rules for optical transitions yield a net current sensitive to circularly polarized optical excitation. The circular PGE

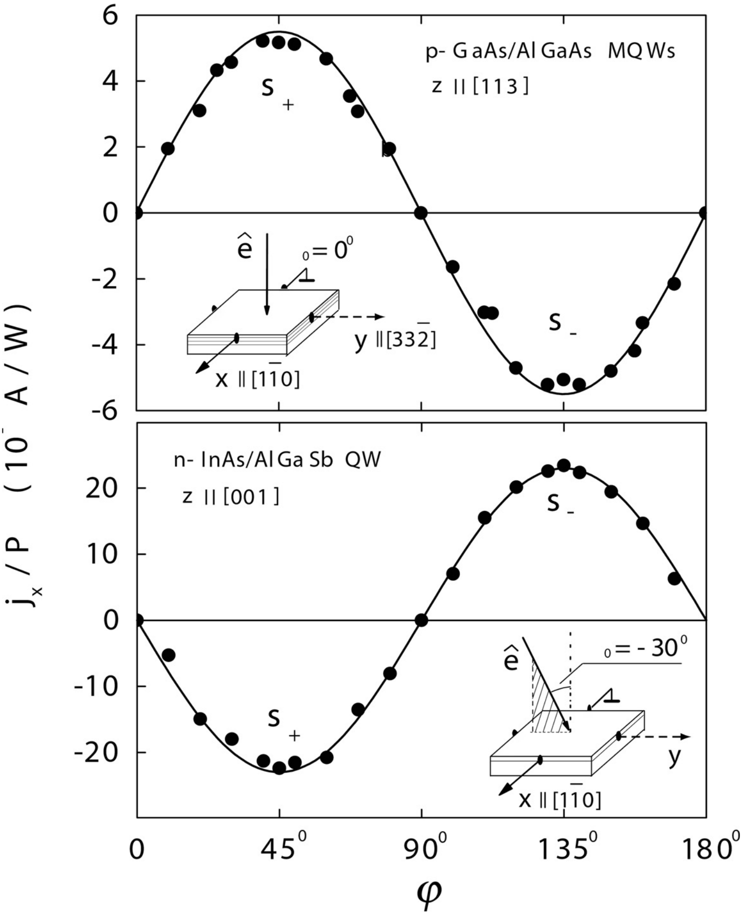


Fig. 1. Photocurrent in QWs normalized by the light power P as a function of the phase angle φ defining helicity. Measurements are presented for T = 300 K and $\lambda = 76$ μm. The insets show the geometry of the experiment. Upper panel: normal incidence of radiation on p-type (113)A-grown GaAs/AlGaAs QWs (symmetry class C_s). The current j_x flows along the $[1\bar{1}0]$ direction perpendicular to the mirror plane. Lower panel: oblique incidence of radiation with an angle of incidence $\theta_0 = -30^\circ$ on n-type (001)-grown InAs/AlGaSb QWs (symmetry class D_{2d} or C_{2v}). Full lines are fitted using one parameter according to (6) (from [8])

is most easily conceivable for direct optical transitions between the heavy-hole valence sub-band $hh1$ and conduction sub-band $e1$ in QWs of the C_s symmetry. For the sake of simplicity we take the linear-$\boldsymbol{k}$ terms into account only in the conduction sub-band assuming the following parabolic dispersion in the $e1$ and $hh1$ subbands

$$E_{e1,\boldsymbol{k},\pm 1/2} = E_g^{QW} + \frac{\hbar^2 k^2}{2m_e} \pm \beta_e k_x \; , \; E^v_{hh1,\boldsymbol{k},\pm 3/2} = -\frac{\hbar^2 k^2}{2m_h} \; , \tag{8}$$

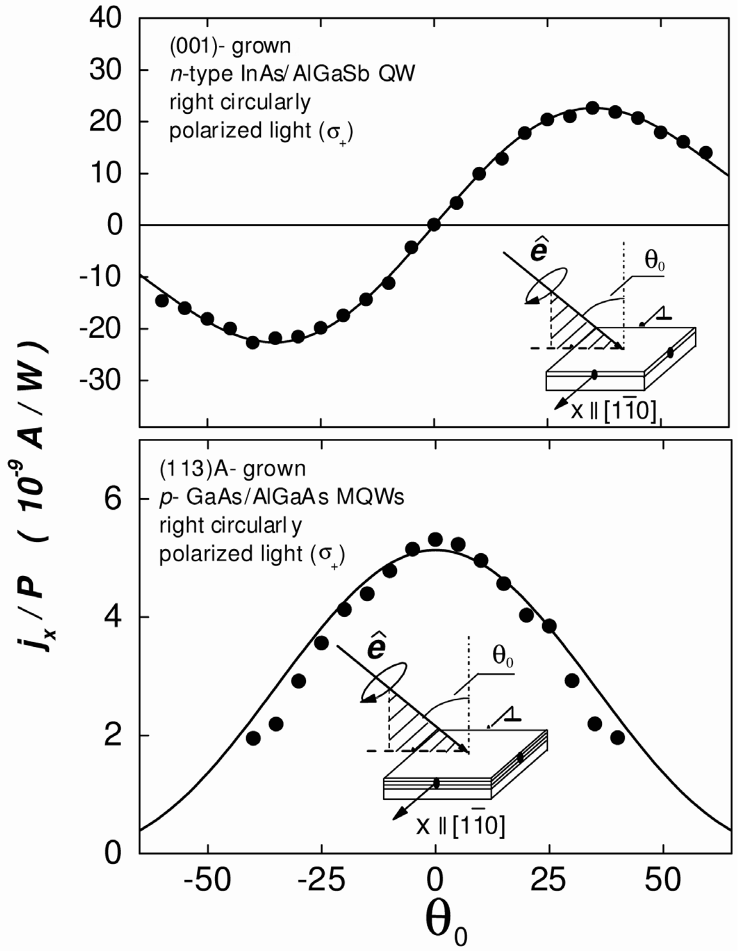


Fig. 2. Photocurrent in QWs normalized by the light power P as a function of the incidence angle θ_0 for right circularly polarized radiation σ_+ measured perpendicularly to light propagation (T = 300 K, λ = 76 μm). Upper panel: n-type (001)-grown InAs/AlGaSb QWs . Lower panel: p-type (113)A-grown GaAs/AlGaAs QWs. Full lines are fitted using (7) (from [8])

where E_g^{QW} is the bandgap renormalized because of the quantum confinement of electrons and holes. In Fig. 3a the allowed optical transitions are from $j = -3/2$ to $s = -1/2$ for the σ_+ polarization and from $j = 3/2$ to $s = 1/2$ for the σ_- polarization. Under circularly polarized radiation with a photon energy $\hbar\omega$ and for a fixed value of k_y, the energy and momentum conservation allow transitions only from two values of k_x. For the σ_+ polarization these particular k_x values of photogenerated electrons are

$$k_x^{\pm} = \frac{\mu}{\hbar^2}\beta_e \pm \left[\frac{2\mu}{\hbar^2}(\hbar\omega - E_g^{QW}) - k_y^2 + \left(\frac{\mu}{\hbar^2}\beta_e\right)^2\right]^{1/2} , \tag{9}$$

where μ is the reduced electron-hole mass $m_e m_h/(m_e + m_h)$. The corresponding transitions are shown in Fig. 3a by the solid vertical arrows with their "center-of-mass" shifted from the point $k_x = 0$ by $\beta_e \mu/\hbar^2$. Thus the average electron velocity in the excited state,

$$\bar{v}_{e,x} = \frac{\hbar(k_x^+ + k_x^-)}{2m_e} - \frac{\beta_e}{\hbar} = -\frac{\mu}{m_h}\frac{\beta_e}{\hbar} ,$$

is non-zero and the contribution of $k_x^{\pm}$ photoelectrons to the current do not cancel as in the case $\beta_e = 0$. Consequently, a spin polarized net current in the x direction results. Changing the photon helicity from +1 to −1 inverts the current because the "center-of-mass" for this transitions is now shifted to $-\beta_e\mu/\hbar^2$. The asymmetric distribution of photoelectrons in the $\boldsymbol{k}$-space decays within the momentum relaxation time τ_p^e. However, under steady-state optical excitation new photocarriers are generated resulting in a dc photocurrent. The photohole contribution in considered in a similar way. Since the average hole velocity $\bar{v}_{h,x}$ coincides with $\bar{v}_{e,x}$, the final result for the interband circular photocurrent can be presented as

$$j_x = e\bar{v}_{e,x}(\tau_p^e - \tau_p^h)\frac{\eta_{cv} I}{\hbar\omega}P_c = -e(\tau_p^e - \tau_p^h)\frac{\beta_e}{\hbar}\frac{\mu}{m_h}\frac{\eta_{cv} I}{\hbar\omega}P_c ,$$

where η_{cv} is the fraction of the energy flux absorbed in the QW due to the $hh1 \to e1$ transitions, different signs of the electron and hole contributions reflect opposite signs of the electron and hole charges. Note that the ratio $I/(\hbar\omega)$ is the flux of photons. If we add the term $\pm\beta_v k_x$ to the electron dispersion $E^v_{hh1,\boldsymbol{k},\pm 3/2}$ in the valence band we obtain

$$j_x = -e(\tau_p^e - \tau_p^h)\left(\frac{\beta_e}{m_h} + \frac{\beta_h}{m_e}\right)\frac{\mu}{\hbar}\frac{\eta_{cv} I}{\hbar\omega}P_c . \tag{10}$$

Above we considered a particular mechanism of the circular PGE. Actually one can use the following general estimation for this effect

$$j_{\text{CPGE}} = e\tau_p \frac{\beta}{\hbar}\frac{\eta I}{\hbar\omega}P_c , \tag{11}$$

where η is the relative absorbance for the considered optical transitions, β is a coefficient in the linear-$\boldsymbol{k}$ spin-dependent Hamiltonian and τ_p is a typical momentum relaxation time.

For more complicated band structures, the previous simple consideration is invalid. Then, one needs to use a sophisticated kinetic theory operating with the electron single-particle density matrix $\rho_{n'n}(\boldsymbol{k})$ and the following general equation of the electron current

$$\boldsymbol{j} = e\sum_{\boldsymbol{k}nn'} \boldsymbol{v}_{nn'}(\boldsymbol{k})\rho_{n'n}(\boldsymbol{k}) . \tag{12}$$

Here the indices n, n' enumerate the electronic states with a given value of $\boldsymbol{k}$ and $\boldsymbol{v}_{nn'}$ is the matrix element of the velocity operator. If the states $|\bar{n}, -\boldsymbol{k}\rangle$ and $|n, \boldsymbol{k}\rangle$ are related by the time inversion operation then one can write

$$\boldsymbol{v}_{\bar{n}\bar{n}'}(-\boldsymbol{k}) = -\boldsymbol{v}_{nn'}(\boldsymbol{k}) \; .$$

This means that the current (12) is contributed only by the anti-symmetrical component of the density matrix

$$\rho^{(-)}_{n'n}(\boldsymbol{k}) = \frac{1}{2}[\rho_{n'n}(\boldsymbol{k}) - \rho_{\bar{n}'\bar{n}}(-\boldsymbol{k})] \; . \tag{13}$$

For photoelectrons excited into the conduction subband $e1$ in a (001)-oriented QW and described by the effective 2×2 Hamiltonian

$$\mathcal{H} = E^0_{e1} + \frac{\hbar^2 k^2}{2m^*} + \beta_{xy}\sigma_x k_y + \beta_{yx}\sigma_y k_x \; , \tag{14}$$

(12) reduces to

$$\boldsymbol{j} = e \sum_{\boldsymbol{k}} \mathrm{Tr} \left\{ \hat{\boldsymbol{v}}(\boldsymbol{k}) \rho^{(e)}(\boldsymbol{k}) \right\} \; , \tag{15}$$

and a similar equation can be written for the photohole contribution. Here, $\rho^{(e)}(\boldsymbol{k})$ is the spin-density matrix and the velocity operator $\hat{\boldsymbol{v}}(\boldsymbol{k}) = \hbar^{(-1)} \partial\mathcal{H}/\partial\boldsymbol{k}$ has the components

$$\hat{v}_x(\boldsymbol{k}) = \frac{\hbar k_x}{m^*} + \frac{\beta_{yx}}{\hbar}\sigma_y \; , \; \hat{v}_y(\boldsymbol{k}) = \frac{\hbar k_y}{m^*} + \frac{\beta_{xy}}{\hbar}\sigma_x \; . \tag{16}$$

In the momentum-relaxation time approximation, one has

$$\boldsymbol{j} = e \sum_{\boldsymbol{k}} \tau^e_p \mathrm{Tr} \left\{ \hat{\boldsymbol{v}}(\boldsymbol{k}) \dot{\rho}^{(e)}(\boldsymbol{k}) \right\} \; , \tag{17}$$

where components of the spin-density generation matrix $\dot{\rho}^{(e)}$ are given by

$$\begin{aligned} \dot{\rho}^{(e)}_{s's}(\boldsymbol{k}) = \frac{\pi}{\hbar} \sum_{v,j} & M_{e1,s';v,j}(\boldsymbol{k}) M^*_{e1,s;v,j}(\boldsymbol{k}) \\ & \times \left[\delta\left(E_{e1,\boldsymbol{k},s'} - E^e_{v\boldsymbol{k}j} - \hbar\omega\right) + \delta\left(E_{e1,\boldsymbol{k},s} - E^e_{v\boldsymbol{k}j} - \hbar\omega\right) \right] \; , \end{aligned} \tag{18}$$

$M_{e1,s;v,j}(\boldsymbol{k})$ is the matrix element of the interband optical transition $(v\boldsymbol{k}j) \to (e1, \boldsymbol{k}, s)$. As soon as linear-$\boldsymbol{k}$ spin-dependent terms are taken into account in the electron Hamiltonian and the light is circularly polarized, the anti-symmetrical component of the generation matrix $\dot{\rho}^{(e)}_{s's}(\boldsymbol{k})$ is non-zero. The photo-hole contribution to the photocurrent is considered in a similar way.

The spectral behavior of the interband circular PGE calculated for (001)-grown QWs is presented in Fig. 3a. The four band edges $j\nu \to e1$ are shown by arrows. As the photon energy approaches the bandgap $e1$-$hh1$ the photocurrent tends to zero. This can be understood taking into account that, for QWs of

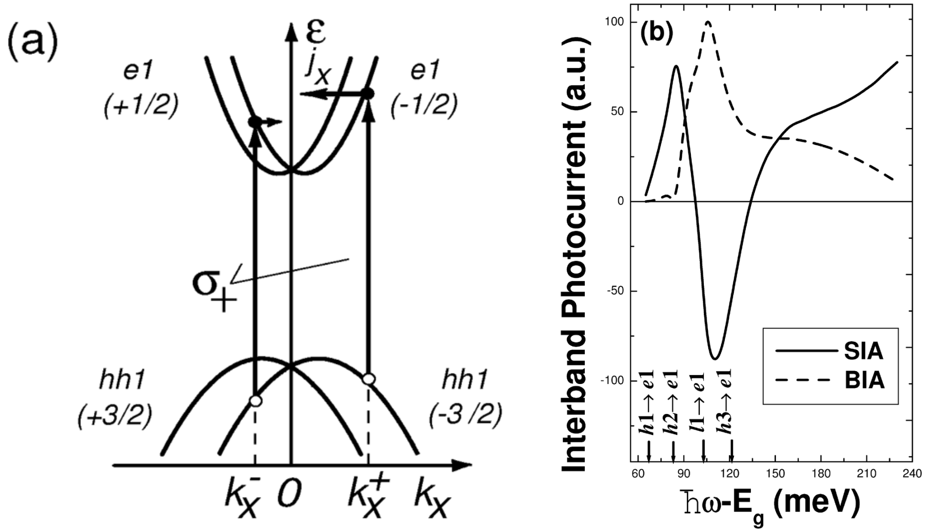


Fig. 3. (**a**) Microscopic picture describing the origin of spin polarized photocurrents. The essential ingredient is the the spin splitting of the electron and/or hole states due to linear-$\boldsymbol{k}$ terms. (**b**) Calculated spectrum of the interband circular photocurrent due to SIA (solid line) and BIA (dashed line) electron spin splittings in a 100-Å wide QW. The arrows indicate the absorption edges for the four optical transitions [11]

the C_{2v} or D_{2d} symmetry, the circular photocurrent appears only under oblique incidence, the optical transitions have to be allowed both in the in-plane and normal-to-plane polarizations, but for purely heavy hole states the interband transitions in the polarization $\boldsymbol{e} \parallel z$ are forbidden. The circular photocurrent due to the $hh1 \to e1$ becomes nonzero because of an admixture of light-hole states in the heavy-hole subband $hh1$ at $\boldsymbol{k} \neq 0$. At small values of $\hbar\omega - E_g^{QW}$ the photocurrent is proportional to $(\hbar\omega - E_g^{QW})^2$ for the BIA linear-$\boldsymbol{k}$ term (63) in the electron Hamiltonian ($\beta_{xy} = \beta_{yx}$) and to the first order of $\hbar\omega - E_g^{QW}$ for the SIA linear-$\boldsymbol{k}$ term ($\beta_{xy} = -\beta_{yx}$) [11]. One can see from Fig. 3b that the spectral variations of the BIA and SIA contributions to the photocurrent differ dramatically in the whole frequency region studied.

Since the characteristic spin splitting is usually small compared to the inhomogeneous broadening and kinetic energy of free carriers, the photocurrents generated under interband, intersubband or intrasubband optical excitation are mainly contributed by terms linear in the coefficients β. In this case one can write the following general relation between the photo-galvanic tensor $\boldsymbol{\gamma}$ and tensor $\boldsymbol{\beta}^{(\nu)}$ describing the linear-$\boldsymbol{k}$ terms in the ν-th conduction or valence subband ν

$$\gamma_{\lambda\mu} \propto \beta_{\mu\lambda}^{(\nu)} \, . \tag{19}$$

In particular, the BIA and SIA terms give rise to independent contributions to the circular PGE and one has

$$j_x \propto I P_c (\beta_{BIA}^{(\nu)} - \beta_{SIA}^{(\nu)}) o_y \, , \; j_y \propto I P_c (\beta_{BIA}^{(\nu)} + \beta_{SIA}^{(\nu)}) o_x \, , \tag{20}$$

where

$$\beta^{(\nu)}_{BIA} = (\beta^{(\nu)}_{xy} + \beta^{(\nu)}_{yx})/2 \, , \; \beta^{(\nu)}_{SIA} = (\beta^{(\nu)}_{xy} - \beta^{(\nu)}_{yx})/2 \, .$$

Note that in Appendix the coefficients β_{BIA} and β_{SIA} are introduced as $-\beta_1$ and β_2, see (63) and (64). It is instructive to rewrite (20) in terms of the current components in the principal axes $x_1 \parallel [100]$, $x_2 \parallel [010]$ and to obtain

$$j_1 \propto IP_c(\beta^{(\nu)}_{BIA} o_1 - \beta^{(\nu)}_{SIA} o_2) \, , \; j_2 \propto IP_c(-\beta^{(\nu)}_{BIA} o_2 + \beta^{(\nu)}_{SIA} o_1) \qquad (21)$$

where o_1, o_2 are the components of the unit vector $\hat{\boldsymbol{o}}$ along x_1 and x_2. It is worth to mention that the BIA and SIA linear-$\boldsymbol{k}$ terms give rise to many spin-dependent phenomena in QWs such as an existence of beats in the Shubnikov-de Haas oscillations, spin relaxation, splitting in polarized Raman scattering spectra, and positive anomalous magnetoresistence. However, in (001)-grown QWs, the BIA and SIA spin-orbit splittings cannot be distinguished in these experiments, particularly if one of the splitting mechanisms is dominating and the electron energy dispersion is uniaxially invariant. On the other hand, the circular PGE suggests a clear and effective way to identify the spin-splitting mechanism in (001)-oriented QWs: under oblique optical excitation by the circularly polarized light with the plane of incidence containing the principal axis 1 or 2, the BIA- and SIA-related circular photocurrents are respectively parallel and perpendicular to the incidence plane.

Next, we turn to a more detailed discussion of the circular PGE for the $e1 \to e2$ intersubband transitions. The circular photocurrent is a sum of two contributions

$$\begin{aligned} \boldsymbol{j} &= \boldsymbol{j}^{(e2)} + \boldsymbol{j}^{(e1)} \\ &= e \sum_{\boldsymbol{k}} \left[\tau_p^{(2)} \mathrm{Tr} \left\{ \hat{\boldsymbol{v}}^{(e2)}(\boldsymbol{k}) \dot{\rho}^{(e2)}(\boldsymbol{k}) \right\} \right. \\ &\quad \left. + \tau_p^{(1)} \mathrm{Tr} \left\{ \hat{\boldsymbol{v}}^{(e1)}(\boldsymbol{k}) \dot{\rho}^{(e1)}(\boldsymbol{k}) \right\} \right] \, , \end{aligned} \qquad (22)$$

respectively, due to the asymmetry of the distribution in $\boldsymbol{k}$-space of the electrons excited to the subband $e2$ and the electrons that stay in the subband $e1$. Here, $\tau_p^{(\nu)}$ is the electron momentum relaxation time in the subband ν. The generation matrix $\dot{\rho}^{(e2)}(\boldsymbol{k})$ for incoming electrons is similar to (18). Therefore, it will suffice to present here the expression for the generation matrix in the $e1$ subband

$$\begin{aligned} \dot{\rho}^{(e1)}_{s's}(\boldsymbol{k}) = -\frac{\pi}{\hbar} \sum_j & M_{e2,j;e1,s}(\boldsymbol{k}) M^*_{e2,j;e1,s'}(\boldsymbol{k}) \\ & \times \left[f^0(E_{e1,\boldsymbol{k},s}) \delta\left(E_{e2,\boldsymbol{k},j} - E_{e1,\boldsymbol{k},s} - \hbar\omega\right) \right. \\ & \left. + f^0(E_{e1,\boldsymbol{k},s'}) \delta\left(E_{e2,\boldsymbol{k},j} - E_{e1,\boldsymbol{k},s'} - \hbar\omega\right) \right] \, , \end{aligned} \qquad (23)$$

where the indices j, s, s' enumerate the spin-split eigenstates and the factor -1 means that the electrons are outgoing from the $e1$ subband. Note that the order of indices s, s' in the product $M_{e2,j;e1,s}(\boldsymbol{k}) M^*_{e2,j;e1,s'}(\boldsymbol{k})$ differs from that for ingoing electrons, see (18).

In order to make the physics more transparent, we will first consider the intersubband circular photocurrent generated under normal incidence in QWs of the C_s symmetry, say in (113)-grown QWs, and use the appropriate coordinate system x, y, z with $z \parallel [113]$. The electron energy spectrum is given by

$$E_{e\nu,\boldsymbol{k},s} = E_\nu^0 + \frac{\hbar^2 k^2}{2m^*} \pm \beta_\nu k_x \, , \tag{24}$$

where $\beta_\nu = \beta_{zx}^{(\nu)}$ and, for the sake of simplicity, we neglect nonparabolicity effects assuming the effective mass m^* to be the same in both subbands. For the direct $e2$-$e1$ transitions, the energy and momentum conservation laws read

$$E_{21} + (s'\beta_2 - s\beta_1)k = \hbar\omega \, .$$

where E_{21} is the $\varGamma$-point gap $E_2^0 - E_1^0$ and $s', s = \pm 1/2$. In the polarization $\boldsymbol{e} \perp z$, the direct intersubband absorption is weakly allowed only for the spin-flip transitions, $(e1, -1/2) \to (e2, 1/2)$ for σ_+ photons and $(e1, 1/2) \to (e2, -1/2)$ for σ_- photons. Particularly, under the σ_+ photoexcitation the electrons involved in the transitions have the fixed x-component of the wave vector

$$k_{21} = \frac{\hbar\omega - E_{21}}{\beta_2 + \beta_1} \tag{25}$$

and velocity

$$v_x^{(e\nu)} = \frac{\hbar k_{21}}{m^*} + \frac{\beta_\nu}{\hbar} \, . \tag{26}$$

It follows then that the circular photocurrent can be written as

$$j_x^{(e1)} = e\left(v_x^{(e2)} \tau_p^{(2)} - v_x^{(e1)} \tau_p^{(1)} \right) \frac{\eta_{21} I}{\hbar\omega} P_c \, , \tag{27}$$

where η_{21} is the absorbance or the fraction of the energy flux absorbed in the QW due to the transitions in consideration, $v_x^{(e\nu)}$ is given by (26) and minus in the right-hand side means that the $e1$-electrons are removed in the optical transitions.

In available QW structures the inhomogeneous broadening of the gap E_{21} exceeds the width of absorption spectrum in an ideal QW. The inhomogeneous broadening is taken into consideration by multiplying the photocurrent $\boldsymbol{j}$ as a function of E_{21} by the distribution function $F(E_{21})$ of the gaps E_{21} and integrating over E_{21}. The convolution of the current (27) with the inhomogeneous distribution function leads to

$$j_x = \frac{e}{\hbar}(\beta_2 + \beta_1) \left[\tau_2 \, \eta_{21}(\hbar\omega) + (\tau_1 - \tau_2) \, \bar{E} \, \frac{d\, \eta_{21}(\hbar\omega)}{d\, \hbar\omega} \right] \frac{I P_c}{\hbar\omega} \, , \tag{28}$$

where $\eta_{21} \propto F(\hbar\omega)$ is the absorbance calculated neglecting the linear-$\boldsymbol{k}$ terms but taking into account the inhomogeneous broadening and $\bar{E}$ is the mean value of the 2D electron energy, a half of the Fermi energy E_F for a degenerate 2D electron gas and $k_B T$ for a non-degenerate gas.

In case of the $e2$-$e1$ transitions in (001)-grown QWs one should start from the spin Hamiltonian

$$\mathcal{H}_\nu = E^0_{e\nu} + \frac{\hbar^2 k^2}{2m^*} + \beta^{(\nu)}_{xy}\sigma_x k_y + \beta^{(\nu)}_{yx}\sigma_y k_x \tag{29}$$

and the intersubband matrix elements of the velocity operator

$$||\boldsymbol{e}\cdot\boldsymbol{v}_{s'_z s_z}|| = v_{21}\begin{bmatrix} e_z & \Lambda(e_x - ie_y) \\ -\Lambda(e_x + ie_y) & e_z \end{bmatrix} \,, \tag{30}$$

$$v_{21} = -i\frac{\hbar}{m^*}\int dz \varphi_{e2}(z)\frac{d}{dz}\varphi_{e1}(z)\,,\; \Lambda = \frac{E_{21}\Delta(2E_g + \Delta)}{2E_g(E_g + \Delta)(3E_g + 2\Delta)}$$

written in the basis of spin states with $s_z = \pm 1/2$. Here $\varphi_{e\nu}(z)$ is the electron envelope function in the subband $e\nu$, E_g and Δ are the band gap and spin-orbit splitting of the valence band in the well material. In order to perform a calculation taking into account all powers of $\beta_{\lambda\mu}$ one needs to use e(22, 23) in a straightforward way. As soon as we are interested in contributions to photocurrents linear in β we can set all β's to zero except for one and proceed similarly to the C_s-symmetry case. For example, we retain the term $\beta^{(\nu)}_{yx}\sigma_y k_x$ in (29) and disregard the term proportional to β_{xy}. The corresponding current is induced in the x-direction perpendicularly to the plane (y, z) of oblique incidence:

$$j_x \propto i(\boldsymbol{e}\times\boldsymbol{e}^*)_y = i(e_z e^*_x - e_x e^*_z) = P_c\, o_y\,. \tag{31}$$

Then, the eigenstates have a fixed spin component on the y axis and the spin split energies are determined by (24) where $\beta_\nu = \beta^{(\nu)}_{zx}$ is changed by $\beta^{(\nu)}_{yx}$ and ± means spin states with $s_y = \pm 1/2$. Since the component e_z is present in (31) and the spin under z-polarized transitions is conserved, see (30), only spin-conserving processes $(e1, s_y) \to (e2, s_y)$ contribute to the circular photocurrent j_x. From (30) one can find the corresponding matrix elements of the velocity operator

$$\langle e2, s_y|\boldsymbol{e}\cdot\hat{\boldsymbol{v}}_{s'_z s_z}|e1, s_y\rangle = v_{21}(e_z + 2i\Lambda s_y e_x)$$

and, hence,

$$|\langle e2, s_y|\boldsymbol{e}\cdot\hat{\boldsymbol{v}}_{s'_z s_z}|e1, s_y\rangle|^2 = |v_{21}|^2(|e_z|^2 - 2\Lambda s_y P_c o_y)\,, \tag{32}$$

where the term quadratic in Λ is neglected. The final result for the circular photocurrent reads

$$j_x = -\Lambda\frac{e}{\hbar}(\beta^{(2)}_{yx} - \beta^{(1)}_{yx})\left[\tau_2\,\eta_{21}(\hbar\omega) + (\tau_1 - \tau_2)\,\bar{E}\,\frac{d\,\eta_{21}(\hbar\omega)}{d\,\hbar\omega}\right]\frac{IP_c}{\hbar\omega}\,o_y\,, \tag{33}$$

where η_{21} is the absorbance in the polarization $\boldsymbol{e} \parallel z$.

An important conclusion is that the photocurrents (28) and (33) change their signs within the resonance absorption spectrum. The sign inversion of the circular photocurrent in the resonant $e2$-$e1$ transition region has been recently observed in n-type GaAs/AlGaAs QW samples [12].

In contrast to electrons in the conduction band, the energy dispersion of holes in the valence band of QWs is essentially nonparabolic and intersubband absorption can involve simultaneously different pairs of subbands $hh\nu$ and $hh\nu'$. However, with some modifications of the theory and complications in calculations the intersubband circular CPE in p-doped samples can be considered similarly to that in n-QWs.

Now we turn to intrasubband optical transitions

$$(e1, \boldsymbol{k}, s) + \hbar\omega \to (e1, \boldsymbol{k}', s')$$

in the lowest electron subband $e1$. They are indirect in the $\boldsymbol{k}$ space, occur due to additional scattering by phonons or static imperfections and involve virtual intermediate states. This situation is realized in n-doped QWs for photon energies to be not high enough in order to excite direct intersubband transitions. The intrasubband photocurrent is given by the general equation (17) where the generation matrix is a sum of contributions due to the ingoing and outgoing electrons. The corresponding generation matrices have the form

$$\begin{aligned}\dot{\rho}^{(\mathrm{out})}_{s's}(\boldsymbol{k}) = -\frac{\pi}{\hbar} \sum_{\boldsymbol{k}'j} & M^{\mathrm{ind}}_{\boldsymbol{k}'j,\boldsymbol{k}s} M^{\mathrm{ind}\,*}_{\boldsymbol{k}'j,\boldsymbol{k}s'} \\ & \times \left[\left(f^0_{\boldsymbol{k}s'} - f^0_{\boldsymbol{k}'j} \right) \delta \left(E_{\boldsymbol{k}',j} - E_{\boldsymbol{k},s'} - \hbar\omega \right) \right. \\ & \left. + \left(f^0_{\boldsymbol{k}s} - f^0_{\boldsymbol{k}'j} \right) \delta \left(E_{\boldsymbol{k}',j} - E_{\boldsymbol{k},s} - \hbar\omega \right) \right] ,\end{aligned} \tag{34}$$

$$\begin{aligned}\dot{\rho}^{(\mathrm{in})}_{j'j}(\boldsymbol{k}') = \frac{\pi}{\hbar} \sum_{\boldsymbol{k}s} & M^{\mathrm{ind}}_{\boldsymbol{k}'j',\boldsymbol{k}s} M^{\mathrm{ind}\,*}_{\boldsymbol{k}'j,\boldsymbol{k}s} \\ & \times \left[\left(f^0_{\boldsymbol{k}s} - f^0_{\boldsymbol{k}'j} \right) \delta \left(E_{\boldsymbol{k}',j} - E_{\boldsymbol{k},s} - \hbar\omega \right) \right. \\ & \left. + \left(f^0_{\boldsymbol{k}s} - f^0_{\boldsymbol{k}'j'} \right) \delta \left(E_{\boldsymbol{k}',j'} - E_{\boldsymbol{k},s} - \hbar\omega \right) \right] .\end{aligned} \tag{35}$$

Here, $E_{\boldsymbol{k},s} \equiv E_{e1,\boldsymbol{k},s}$, $f^0_{\boldsymbol{k}s}$ is the electron distribution function in the $e1$ subband and $M^{\mathrm{ind}}_{\boldsymbol{k}j,\boldsymbol{k}s}$ is the matrix element of the indirect optical transition. In the second order of the perturbation theory it is given by

$$M^{\mathrm{ind}}_{\boldsymbol{k}'j,\boldsymbol{k}s} = \sum_n \left(\frac{V_{e1,\boldsymbol{k}',j;n\boldsymbol{k}} M_{n\boldsymbol{k};e1,\boldsymbol{k},s}}{E_{n\boldsymbol{k}} - E_{e1,\boldsymbol{k},s} - \hbar\omega} + \frac{M_{e1,\boldsymbol{k}',j;n\boldsymbol{k}'} V_{n\boldsymbol{k}';e1\boldsymbol{k}s}}{E_{n\boldsymbol{k}'} - E_{e1,\boldsymbol{k},s} \pm \hbar\Omega} \right) , \tag{36}$$

where the index n enumerates the intermediate states, $M_{n'\boldsymbol{k};n\boldsymbol{k}}$ and $V_{n'\boldsymbol{k}';n\boldsymbol{k}}$ are the matrix elements of the electron-photon and electron-phonon or electron-defect interaction, Ω is the phonon frequency, the sign ± corresponds to emission and absorption of phonons. For the scattering by static defects Ω is set to zero. An important point is that indirect transitions via intermediate states in the same subband do not contribute to the circular PGE. The effect appears if virtual processes involve intermediate states in other bands or subbands, $n \neq e1$.

2 Spin-Galvanic Effect

The mechanisms of the circular PGE discussed so far are linked to the asymmetry in the momentum distribution of carriers excited in optical transitions which are sensitive to the light circular polarization due to selection rules. Now we discuss an additional possibility to generate a photocurrent sensitive to the photon helicity [13,14]. In a system of free carriers with non-equilibrium spin-state occupation but equilibrium energy distribution within each spin branch, the spin relaxation or Larmor precession in an external magnetic field can be accompanied by generation of an electric current. Phenomenologically, this linkage between an electric current and the total electronic spin $\boldsymbol{s}$ is described by

$$j_\lambda = \sum_\mu Q_{\lambda\mu} s_\mu \, . \tag{37}$$

The symmetry of the second-order pseudo-tensor $\mathbf{Q}$ coincides with that of the tensor $\boldsymbol{\gamma}$ describing the circular PGE, see (1). Similarly, its non-vanishing components can exist in non-centrosymmetric systems belonging to one of the gyrotropic classes. In (001)-oriented QWs of the C_{2v} symmetry equation (37) reads

$$j_x = Q_{xy} s_y \, , \, j_y = Q_{yx} s_x \, . \tag{38}$$

If the non-equilibrium spin is produced by optical orientation and the spin s_μ is proportional to the degree of light circular polarization P_c the current generation can be regarded just as another mechanism of the circular PGE. However, the non-equilibrium spin $\boldsymbol{s}$ can be achieved both by optical and non-optical methods, e.g., by electrical spin injection, and in fact (37) presents an independent effect called the Spin-Galvanic Effect. Here, we bear in mind spin-induced electric currents that appear under uniform distribution of the spin polarization in the 3D-, 2D- or 1D space, respectively in a bulk semiconductor, a QW and a quantum wire. In this sense the spin-galvanic effect differs from surface currents induced by inhomogeneous spin orientation [15] and other phenomena where the spin current is caused by gradients of potentials, concentrations etc., like the Spin-Voltaic Effect which occurs in inhomogeneous samples, e.g., the 'paramagnetic metal-ferromagnetic' junction or *p*-*n* junction.

Usually the circular photo-galvanic and spin-galvanic effects are observed simultaneously under illumination by circularly polarized light and do not allow experimental separation. However, they can be separated in time-resolved measurements. Indeed, after removal of light or under pulsed photoexcitation the circular photocurrent decays within the momentum relaxation time τ_p whereas the spin-galvanic current decays with the spin relaxation time τ_s. Next we consider a geometry of experiment under steady-state photoexcitation which allows to observe the spin-galvanic effect and exclude the circular PGE [14]. The geometry is depicted in inset in Fig. 4. The circularly polarized light is incident normally to the interface plane (001) of a QW, the light absorption yields a steady-state spin orientation s_{0z} in the z direction proportional to the spin generation rate $\dot{s}_z$. The symmetry of (001)-grown QWs forbids generation of a current proportional to the normal component of $\boldsymbol{s}$. To obtain an in-plane component of the spins, necessary for the spin-galvanic effect, a magnetic field $\boldsymbol{B} \parallel x$ is applied. Due to

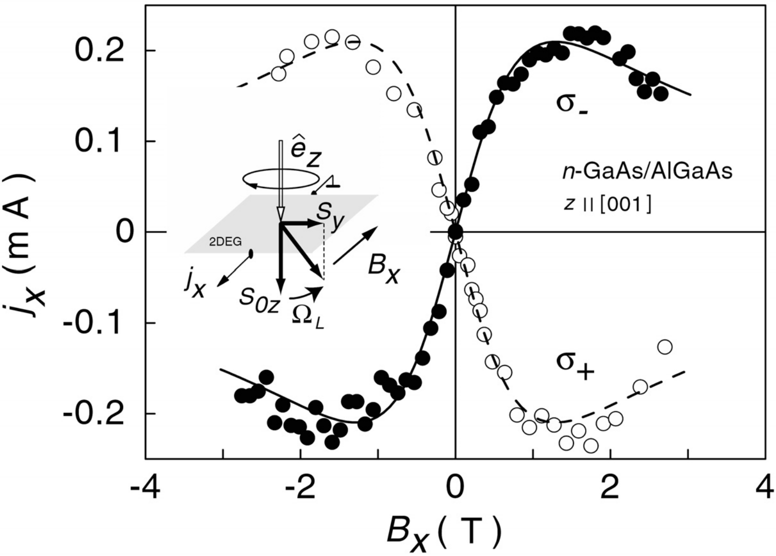


Fig. 4. Current j_x as a function of magnetic field B for normally incident right-handed (open circles) and left-handed (filled circles) circularly polarized radiation at $\lambda = 148$ μm and radiation power 20 kW. Measurements are presented for an n-GaAs/AlGaAs single heterojunction at T = 4.2 K. Curves are fitted from (39) using the same value of the spin relaxation time τ_s and scaling of the j_x value for both the solid and dashed curves (from [14])

Larmor precession a non-equilibrium spin polarization s_y is induced,

$$s_y = -\frac{\Omega_L \tau_{s\perp}}{1 + (\Omega_L \tau_s)^2}\, s_{0z} \,, \tag{39}$$

where $\tau_s = \sqrt{\tau_{s\|}\tau_{s\perp}}$, $\tau_{s\|}$, $\tau_{s\perp}$ are the longitudinal and transverse electron spin relaxation times, Ω_L is the Larmor frequency. The photocurrent measured in the x direction is shown in Fig. 4 as a function of the magnetic field for two opposite circular polarizations of the light. In accordance with the phenomenological equations (38) and (39) the current j_x exhibits a non-monotonous variation with the magnetic field. Comparison with theory allows to find a product $g\tau_s$ and the spin relaxation time if the electron g-factor is known.

There are two different microscopical mechanisms of the spin-galvanic effect, namely, kinetic and relaxational [13]. The experimental data of Fig. 4 can be understood in terms of the kinetic mechanism. It is inherently connected with the spin dependency of matrix elements, $M_{\boldsymbol{k}'s',\boldsymbol{k}s}$, of electron scattering by impurities, other static defects and phonons. It is convenient to represent the 2×2 matrix $\hat{M}_{\boldsymbol{k}'\boldsymbol{k}}$ as a linear combination of the unit matrix $\hat{I}$ and Pauli matrices as follows

$$\hat{M}_{\boldsymbol{k}'\boldsymbol{k}} = A_{\boldsymbol{k}'\boldsymbol{k}}\hat{I} + \boldsymbol{\sigma}\cdot\boldsymbol{B}_{\boldsymbol{k}'\boldsymbol{k}} \,, \tag{40}$$

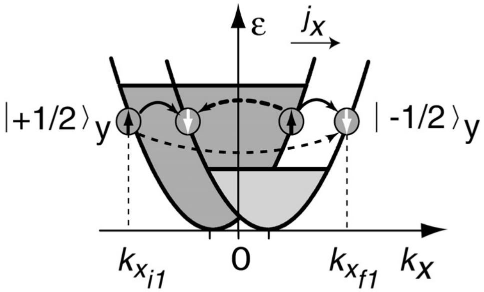


Fig. 5. Microscopic origin of the spin-galvanic current in the presence of $\boldsymbol{k}$-linear terms in the electron Hamiltonian. The $\sigma_y k_x$ term in the Hamiltonian splits the conduction band into two parabolas with the spin ±1/2 in the y direction. If one spin subband is preferentially occupied, asymmetric spin-flip scattering results in a current in the x direction. The rate of spin-flip scattering depends on the value of the initial and final $\boldsymbol{k}$-vectors. There are four distinct spin-flip scattering events possible, indicated by the arrows. The transitions sketched by dashed arrows yield an asymmetric occupation of both subbands and hence a current flow. If, instead of the spin-down subband, the spin-up subband is preferentially occupied the current direction is reversed

where $A^*_{\boldsymbol{k}'\boldsymbol{k}} = A_{\boldsymbol{k}\boldsymbol{k}'}$, $B^*_{\boldsymbol{k}'\boldsymbol{k}} = B_{\boldsymbol{k}\boldsymbol{k}'}$ due to hermiticity of the interaction and $A_{-\boldsymbol{k}',-\boldsymbol{k}} = A_{\boldsymbol{k}\boldsymbol{k}'}$, $B_{-\boldsymbol{k}',-\boldsymbol{k}} = -B_{\boldsymbol{k}\boldsymbol{k}'}$ due to the symmetry under time inversion.

The spin-galvanic current observed in the geometry of Fig. 4 is caused by the asymmetric spin-flip scattering of spin-polarized electrons in the systems with $\boldsymbol{k}$-linear contributions to the effective Hamiltonian. Figure 5 illustrates the electron energy spectrum with the $\beta_{yx}\sigma_y k_x$ term included. Spin orientation in the y direction causes an unbalanced population in the spin-down and spin-up branches. Spins oriented in the y direction are scattered along k_x from the higher filled branch, say the spin-up or $|1/2\rangle_y$ branch, to the less filled branch $|-1/2\rangle_y$. The matrix elements for these spin-flip processes are proportional to the components $B_{\boldsymbol{k}'\boldsymbol{k},x}$ and $B_{\boldsymbol{k}'\boldsymbol{k},z}$ of the vector $\boldsymbol{B}_{\boldsymbol{k}'\boldsymbol{k}}$ in (40).

Four different spin-flip scattering events are schematically sketched in Fig. 5 by arrows. Their probability rates depend on the values of the wave vectors of the initial and final states. Spin-flip transitions shown by solid arrows have the same rate. They preserve the symmetric distribution of carriers in the branches and, thus, do not yield a current. The two processes shown by broken arrows are not equivalent and generate an asymmetric carrier distribution around the branch minima in each spin branch. This asymmetric distribution results in a current flow along the x direction.

In considering the relaxational mechanism of the spin-galvanic effect, we can ignore the spin-dependence of the scattering matrix elements but should retain quantum corrections of the order of $\mathcal{H}^{(1)}/\bar{E}$, where $\bar{E}$ is the average electron kinetic energy. Moreover, we can apply the electron spin density matrix formalism and assume the following hierarchy of relaxation times to be fulfilled

$$\tau_p \ll \tau_\varepsilon \ll \tau_s, \tau_0 \,, \tag{41}$$

where $\tau_p, \tau_\varepsilon, \tau_s$ are the electron momentum, energy and spin relaxation times respectively, and τ_0 is the electron lifetime in case of the interband optical photoexcitation. Then the spin-galvanic current related to the relaxation mechanism may be presented as [13]

$$\boldsymbol{j} = -eN_e\tau_p\nabla_{\boldsymbol{k}}\left(\boldsymbol{\Omega}^{(1)}_{\boldsymbol{k}} \cdot \dot{\boldsymbol{S}}\right) . \tag{42}$$

Under normal incidence of the light on a (001)-grown QW the vector $\dot{\boldsymbol{S}}$ is directed along z and the current is zero. Thus, the relaxational mechanism makes no contribution to the spin-galvanic current in the set-up of Fig. 4 and the latter is completely related to the kinetic mechanism.

The radiation of the CO_2 laser causes direct $e2$-$e1$ transitions in GaAs/AlGaAs MQWs and can induce the resonant spin-galvanic current at normal incidence of radiation in the presence of an in-plane magnetic field as this effect was previously observed under intrasubband transitions, see Fig. 4. Since the spin generation rate $\dot{S}_z \propto K_\perp \propto K_z$, where $K_\perp, K_z$ are the light absorption coefficients in the polarization $\boldsymbol{e} \perp z$ and $\boldsymbol{e} \parallel z$, respectively, the spectral behavior of the spin-galvanic current must coincide with the absorption spectrum. One can see from Fig. 6 that the wavelength dependence of the spin-galvanic effect obtained between 9.2 μm and 10.6 μm indeed repeats the spectrum of the intersubband absorption. The interplay of the Rashba and Dresselhaus spin splitting of the conduction band has been revealed in experiments on resonant intersubband photogalvanic and spin-galvanic effects in n-type GaAs QW structures [17].

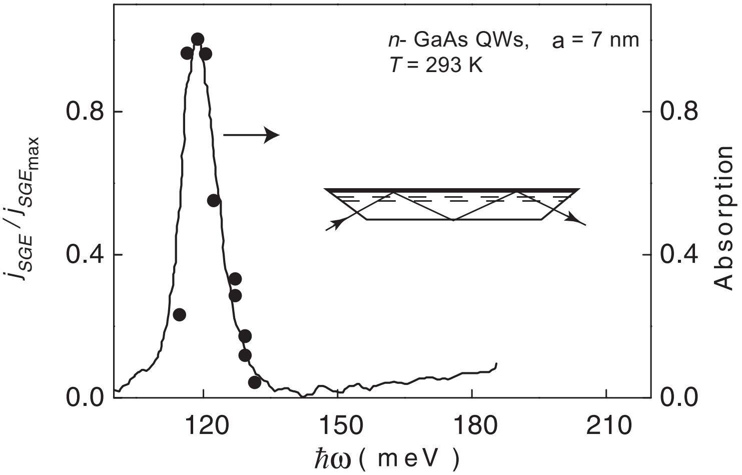


Fig. 6. Absorption spectrum (full line) of an n-doped (001)-grown GaAs/AlGaAs MQW structure ($a = 70$ Å) obtained from transmission in a multiple-reflection waveguide geometry, see the inset. Points show spectral dependence of the spin-galvanic current caused by spin orientation due to direct optical transitions between $e1$ and $e2$ conduction subbands [16]

3 Saturation of Photocurrents at High Light Intensities

Before we turn to discussing the saturation of photocurrents with increasing light intensity, we give a brief information concerning the linear PGE. For this effect, the phenomenological equation (1) in QWs of the C_{2v} symmetry reduces to

$$j_{\mathrm{LPGE},x} = \chi_{xxz}\left(e_x e_z^* + e_z e_x^*\right) I \;,\; j_{\mathrm{LPGE},y} = \chi_{yyz}\left(e_y e_z^* + e_z e_y^*\right) I \;. \qquad (43)$$

In symmetrical QWs, the point-group D_{2d}, the pair of coefficients are linearly dependent, $\chi_{xxz} = -\chi_{yyz}$.

The C_s symmetry allows both circular and linear PGEs for normal incidence because in this case the tensors $\boldsymbol{\gamma}$ and $\boldsymbol{\chi}$ have the additional non-zero components γ_{xz}, see (2), $\chi_{xxy} = \chi_{xyx}$, χ_{yxx} and χ_{yyy}. As a result, under normal incidence one has

$$j_x = \left[\gamma_{xz} P_c + \chi_{xxy}\left(e_x e_y^* + e_y e_x^*\right)\right] I \;,\; j_y = \left(\chi_{yxx}|e_x|^2 + \chi_{yyy}|e_y|^2\right) I \;. \qquad (44)$$

In particular, for linearly polarized light

$$j_{\mathrm{LPGE},x} = I\chi_{xxy}\sin 2\alpha \;,\; j_{\mathrm{LPGE},y} = I\left(\chi_+ + \chi_- \cos 2\alpha\right) \;, \qquad (45)$$

where $\chi_\pm = (\chi_{yxx} \pm \chi_{yyy})/2$ and α is the angle between the plane of polarization and x. Figure 7 presents the measured dependence of j_x and j_y as a function of the angle α and the fit to (45) for a p-type SiGe (113)-grown asymmetrical QW

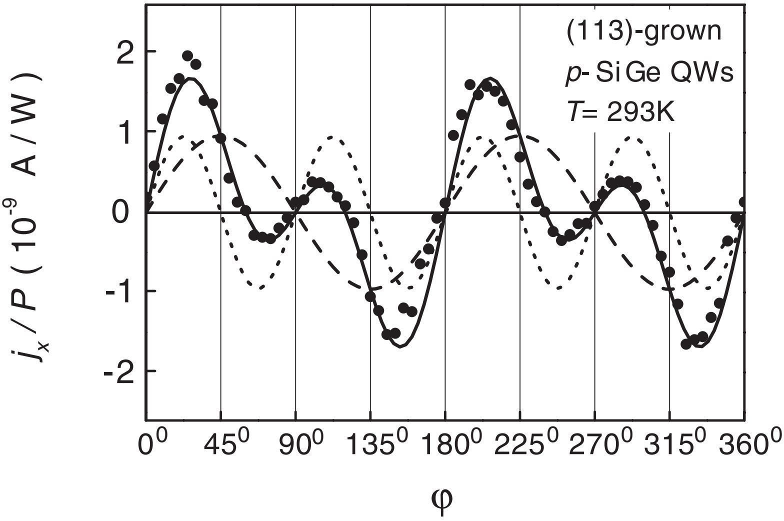


Fig. 7. Photogalvanic current in a (113)-grown $\mathrm{Si}_{0.75}\mathrm{Ge}_{0.25}$(5 nm)/Si single QW normalized by the light power P as a function of the phase angle φ. The results are obtained under normal incidence of irradiation at $\lambda = 280$ μm at room temperature. The full line is fitted after (46). Broken and dotted lines show $j_x \propto \sin 2\varphi$ and $j_x \propto \sin 2\varphi \cos 2\varphi$, respectively (from [9])

structure. In the experimental setup, where the laser light is linearly polarized along x and a $\lambda/4$ plate is placed between the laser and the sample, (44) takes the form

$$j_x = I\left(\gamma_{xz} + \chi_{xxy}\cos 2\varphi\right)\sin 2\varphi\ ,\ j_y = I\left(\chi_+ + \chi_-\cos 2\varphi\right)\ , \tag{46}$$

where φ is the angle between the initial plane of linear polarization and the optical axis of the polarizer. The circular and linear polarizations of the incident light vary with φ in accordance to $P_c = \cos\varphi$, see (5), and $P_l = \sin\varphi$. In Fig. 7 experimental data and a fit to these functions are presented for the same p-type SiGe (113)-grown QW structure.

The linear PGE was observed in some insulators as early as the 1950s, and possibly even earlier, but was correctly identified as a novel phenomenon only in 1974-75 [18,19]. In semiconductors, the linear PGE was first observed on Tellurium [20,21] and then studied in detail on p-GaAs [22].

Microscopically, a current of the linear PGE consists of the so-called ballistic and shift contributions [23–26]. The first of them is described by the conventional equation

$$\boldsymbol{j} = e\sum_{nn'} W_{n'n}(\boldsymbol{v}_{n'}\tau_p^{(n')} - \boldsymbol{v}_n\tau_p^{(n)})\ . \tag{47}$$

Here the index n describes all quantum numbers characterizing the electron eigenstates, namely the band and subband labels, spin sublevel and wave vector $\boldsymbol{k}$; the probability transition rate from the state n to n' is given by Fermi's golden rule

$$W_{n'n} = \frac{2\pi}{\hbar}\left|M_{n'n}\right|^2(f_n - f_{n'})\delta(E_{n'} - E_n)\ , \tag{48}$$

$M_{n'n}$ is the transition matrix element, $\boldsymbol{v}_n$ and $\tau_p^{(n)}$ are the electron velocity and momentum relaxation time in the state n, f_n is the distribution function, or the occupation, of the state n. The energy E_n includes the photon or phonon energy in the initial or final state. Equation (47) is a contribution to the general expression for the current (12) of diagonal components $\rho_{nn} = f_n$ of the electron density matrix and of the velocity $\boldsymbol{v}_{nn} \equiv \boldsymbol{v}_n$. The ballistic current is non-zero only if one simultaneously includes in $M_{n'n}$ carrier interaction both with a photon and with another particle, a phonon, impurity or static defect, another electron or hole, including a geminate partner photocreated in the same photoabsorption process. In other words one needs to go beyond the Born approximation in calculating $M_{n'n}$.

The second contribution to the linear PGE current comes from inclusion in (12) of the non-diagonal components $\rho_{nn'}$ and $\boldsymbol{v}_{nn'}$ with $n' \neq n$. This current was shown [25] to originate from the shift of the wave packet's center-of-mass in quantum transitions and can be written as

$$\boldsymbol{j} = e\sum_{nn'} W_{n'n}\boldsymbol{R}_{n'n}\ . \tag{49}$$

For the shift in the real space we have

$$\boldsymbol{R}_{n'n} = -\left(\nabla_{\boldsymbol{k}} + \nabla_{\boldsymbol{k}'}\right)\varPhi_{n'n} + \boldsymbol{\varOmega}_{n'} - \boldsymbol{\varOmega}_n\ , \tag{50}$$

where $\Phi_{n'n}$ is the phase of the transition matrix element, $\boldsymbol{k}$ and $\boldsymbol{k}'$ are the wave vectors in the states n and n', $\boldsymbol{\Omega}_n$ is the diagonal matrix element of the coordinate

$$\boldsymbol{\Omega}_n = i \int u_n^* \nabla_{\boldsymbol{k}} u_n \, d\boldsymbol{r} \; ,$$

and $u_n(\boldsymbol{r})$ is the Bloch periodical amplitude. In a steady-state regime, when the processes of generation, scattering and recombination are taken altogether into consideration, the contributions associated with $\boldsymbol{\Omega}_n$ vanish since they describe the charge static redistribution. The first term in the right-hand side of (50) can be rewritten as

$$\boldsymbol{R}_{n'n} = - \frac{\operatorname{Im} \{ M_{n'n}^* \left(\nabla_{\boldsymbol{k}} + \nabla_{\boldsymbol{k}'} \right) M_{n'n} \}}{|M_{n'n}|^2} \; . \tag{51}$$

This form is useful in practical calculations.

The linear PGE can be also induced in non-centrosymmetric superlattices (SLs), i.e. in a saw-tooth SL, and multiple quantum well (MQW) structures, i.e. in MQWs with asymmetric double wells, under illumination with unpolarized light [27–30]. The photocurrent is generated along the growth direction z because of the lack of reflection symmetry $z \to -z$. Note that in MQWs the effect has a threshold at the edge of transitions between quantized and continuum states, the so-called bound-to-continuum or above-barrier transitions.

Now, we concentrate on non-linear behavior of the linear and circular PGE with increasing the light intensity due to saturation or bleaching of the absorption. Since the saturation effect was observed on p-doped QW structures [31] we consider direct intersubband optical transitions from the heavy-hole subband $hh1$ to higher subbands, say the $lh1$ subband.

Spin sensitive bleaching can be analyzed in terms of the following simple model taking into account both optical excitation and non-radiative relaxation processes. The probability rates for direct optical transitions from the $hh1$ states with $m = \pm 3/2$ to higher subbands are denoted as $W_\pm$. For linearly polarized light, W_+ and W_- are equal. For the circular polarization, right-handed, σ_+, or left-handed, σ_-, the rates $W_\pm$ are different but, due to time inversion symmetry, satisfy the condition $W_+(\sigma_\pm) = W_-(\sigma_\mp)$. The photo-excited holes are assumed to lose their spin orientation in the course of energy relaxation to the bottom of the $hh1$ subband, due to rapid spin relaxation in hot states. Thus, spin orientation occurs only in this subband. If p_+ and p_- are the 2D densities of heavy holes occupying the subbands $(hh1, +3/2)$ and $(hh1, -3/2)$, respectively, then the rate equations for $p_\pm$ can be written as

$$\begin{aligned} \frac{\partial p_+}{\partial t} + \frac{p_+ - p_-}{2\tau_s} &= -W_+ + \frac{1}{2}(W_+ + W_-) \, , \\ \frac{\partial p_-}{\partial t} + \frac{p_- - p_+}{2\tau_s} &= -W_- + \frac{1}{2}(W_+ + W_-) \, . \end{aligned} \tag{52}$$

The second terms on the left-hand side describe the spin relaxation trying to equalize the population in the $(hh1, \pm 3/2)$ spin branches. The first terms on the right-hand side describe the removal of holes from the $hh1$ subband due to photo-excitation while the second terms characterize the relaxation of holes which come down to the $(hh1, +3/2)$ and $(hh1, -3/2)$ states with equal rates. If the laser pulse duration is longer than any relaxation time, the time derivatives in (52) can be omitted and, instead of this equation, we have

$$\frac{p_+ - p_-}{\tau_s} = -(W_+ - W_-) \, . \tag{53}$$

The hole-removal rates can be presented in the form

$$W_+ = \frac{1}{2} \frac{\eta I}{\hbar\omega} (1 - \rho_0 P_c)(1 + \rho) \, , \; W_- = \frac{1}{2} \frac{\eta I}{\hbar\omega} (1 + \rho_0 P_c)(1 - \rho) \, , \tag{54}$$

where ρ is the hole spin polarization degree $(p_+ - p_-)/(p_+ + p_-)$, η is a function of the light intensity I, the parameter ρ_0 is defined as the ratio $(W_- - W_+)/(W_- + W_+)$ for the σ_+-polarized radiation of low enough intensity where η is constant and $W_\pm$ is proportional to I. The factors $1 \pm \rho_0 P_c$ take into account the sensitivity of optical transitions to the circular polarization of light and spin of the involved particle. The factors $1 \pm \rho$ take into account that the transition probability rate depends on the occupation number of the initial state and, hence, on the hole spin polarization. Substitution of (54) into (53) leads to the linear equation for ρ

$$\frac{p_s}{\tau_s} \rho = \frac{\eta I}{\hbar\omega} (\rho_0 P_c - \rho) \, , \tag{55}$$

where p_s is the hole density and we rewrote $p_+ - p_-$ as $p_s \rho$. The solution reads

$$\rho = \rho_0 P_c \frac{\tau_s \eta I/(p_s \hbar\omega)}{1 + [\tau_s \eta I/(p_s \hbar\omega)]} \, . \tag{56}$$

Bleaching of absorption with increasing intensity of linearly-polarized light is described phenomenologically by the function

$$\eta(I) = \frac{\eta_0}{1 + \frac{I}{I_{\mathrm{se}}}} \, , \tag{57}$$

where $\eta_0 = \eta(I \to 0)$ and I_{se} is the characteristic saturation intensity controlled by energy relaxation of the 2D hole gas. Since the photocurrent of linear PGE, j_{LPGE}, induced by the linearly polarized light is proportional to ηI, one has

$$\frac{j_{\mathrm{LPGE}}}{I} \propto \frac{1}{1 + \frac{I}{I_{\mathrm{se}}}} \, . \tag{58}$$

The circular current j_{CPGE} induced by the circular polarized radiation is proportional to $W_+ - W_- \propto \rho$. Substituting $\eta(I)$ from (57) into (56) we find

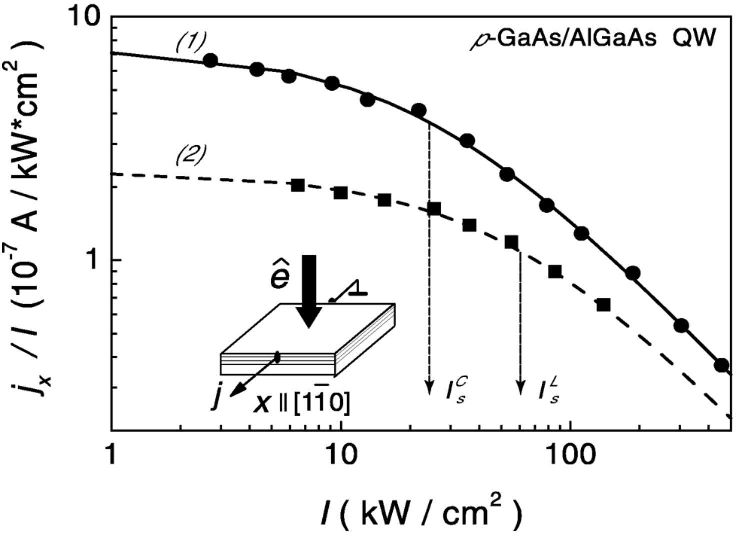


Fig. 8. Photogalvanic current j_x normalized by the intensity I as a function of I for circularly and linearly polarized radiation at T = 20 K. The inset shows the geometry of the experiment; $\hat{\boldsymbol{e}}$ indicates the direction of the incoming light. The current j_x flows along [1$\bar{1}$0] direction at normal incidence of radiation on p-type (113)A-grown GaAs/AlGaAs QWs. In order to obtain the circular PGE right or left circularly polarized light has been applied. To obtain the linear PGE linearly polarized radiation with the electric field vector $\boldsymbol{E}$ oriented at 45° to the x direction was used. The measurements are fitted to $j_x/I \propto 1/(I + I/I_s)$ with one parameter I_s for each state of polarization (full line: circular, broken line: linear) (from [31])

after some development

$$\frac{j_{\mathrm{LPGE}}}{I} \propto \frac{1}{1 + I\left(\frac{1}{I_{\mathrm{se}}} + \frac{1}{I_{\mathrm{ss}}}\right)} , \tag{59}$$

where $I_{\mathrm{ss}} = p_s \hbar\omega / (\eta_0 \tau_s)$.

The measurements illustrated in Fig. 8 indicate that the photocurrent j_x at a low power level depends linearly on the light intensity and gradually saturates with increasing intensity, $j_x \propto I/(1 + I/I_s^{L,C})$, where $I_s^{L,C}$ is the saturation parameter for linearly and circularly polarized radiation. This corresponds to a constant absorbance at low values of I and decreasing absorption with rising I. From (58, 59) we obtain

$$I_s^L = I_{\mathrm{se}} \; , \; I_s^C = \frac{I_{\mathrm{se}} I_{\mathrm{ss}}}{I_{\mathrm{se}} + I_{\mathrm{ss}}} \; . \tag{60}$$

One can see from Fig. 8 that the measured saturation intensities $I_s^{L,C}$ are different, namely $I_s^C < I_s^L$. This is in agreement with the theory. The saturation of the absorption of linearly polarized radiation is governed by the energy relaxation time τ_ε whereas in case of the circular polarization it is governed by both τ_ε

and τ_s. If τ_s is of the order of τ_ε or larger, the saturation becomes spin sensitive and the saturation intensity of circularly polarized radiation drops below that for the linear polarization.

Taking into account that $I_{\mathrm{ss}} = I_s^L I_s^C / (I_s^L - I_s^C)$ and using the measured values of I_s^L, I_s^C one can estimate the parameter $I_{\mathrm{ss}} = p_s \hbar\omega / (\eta_0 \tau_s)$ and even the time τ_s. The latter is possible if the absorbance η_0 is known from an independent experiment or theoretical calculation, see details in [31,33].

It is worth to mention effects which are inverse to the circular PGE. In the absence of magnetic field the dc current in a QW should induce the circular dichroism of the optical absorption or the circular polarization of the photoluminescence. Up to now the effect of the electric current on the optical activity has been observed in bulk gyrotropic Tellurium [32].

4 Summary

A non-equilibrium uniform spin polarization obtained by optical orientation induces an electric current in QW structures of the point symmetry D_{2d}, C_{2v} or C_s belonging to gyrotropic crystal classes. In symmetrical zinc-blende-based QWs the gyrotropy naturally appears due to the lack of inversion centers in the bulk basic material as well as the quantum confinement effect. In QWs grown from diamond-lattice materials, like Si and Ge, which possess a center of inversion, the gyrotropy appears due to artificial asymmetry of the grown structures.

The transformation of spin polarization into an electric current occurs due to the spin-dependent odd-in-$\boldsymbol{k}$ contribution to the electron or hole effective Hamiltonian. There are two different microscopic mechanisms of spin photocurrents, namely, circular photo-galvanic effect and spin-galvanic effect. Usually both effects appear simultaneously and the measured current is a sum of the two contributions. However, in particular cases they can be separated.

The experimental results on spin photocurrents due to homogeneous spin polarization are in good agreement with the phenomenological theory. Both mechanisms of spin photocurrents as well as the removal of spin degeneracy in the $\boldsymbol{k}$-space are described by second rank pseudo-tensors. Therefore macroscopic measurements of photocurrents in different geometric configurations allow to conclude on details of the microscopic spin-orbit interaction. In particular, the relation between Dresselhaus- and Rashba-like terms (including interface inversion asymmetry) may be estimated. Furthermore, the macroscopic symmetry of QWs may easily be determined.

Most recently the circular PGE has been predicted for gyrotropic 1D systems like carbon nanotubes of spiral symmetry [51]. The effect is caused by coupling between the electron wave vector along the tube's principal axis and the *orbital* momentum around the tube circumference.

Spin photocurrents were applied to investigate the mechanisms of free-carrier spin relaxation under monopolar spin orientation where only one type of charge carriers is involved.

In above-cited experimental works the circular PGE has been detected in QWs only under intraband transitions by excitation with infrared radiation.

Recently the first observation of this effect has been reported at interband excitation in GaAs-based QWs [52]. Further experiments on interband circular PGE will open new possibilities in studies of interband optical spin orientation of free carriers, spin relaxation mechanisms of hot and thermalized photocarriers, the role of bulk, structure and interface asymmetries in the spin splitting of conduction- and valence-band subbands. Another subsequent important step will be to perform experiments under short-pulse excitation and reveal the time-resolved kinetics of the photocurrents related to the momentum, energy and spin relaxation of photocarriers. The future work can also be directed towards extension of studies on one-dimensional objects (quantum wires and nanotubes), and even on quantum-dot structures taking into account optical transitions between zero-dimensional states in the dots and two-dimensional states in the wetting layer.

More details concernning photo-galvanic effects in semiconductor nanostructures are presented in the book [53].

Acknowledgments

The work is financially supported by the RFBR and the Programmes of RAS.

Appendix. Spin Splitting of Electron Subbands: Bulk- and Structure-Inversion Asymmetry

It is well-known that, in centrosymmetric crystals, the electronic state at each band n and wave vector $\boldsymbol{k}$ is at least twofold degenerate (the so-called Kramers degeneracy, or spin degeneracy). In crystals lacking a center of inversion, the Kramers degeneracy of the Bloch states is lifted except for special points or lines in the Brillouin zone. Particularly, in zinc-blende-lattice semiconductors like GaAs, InAs, ZnSe, CdTe etc. (the class T_d) the conduction band Γ_6 and the valence band Γ_8 are respectively twofold and fourfold degenerate at the Γ point. However, away from this point the conduction and valence bands are split into non-degenerate spin branches, even at zero magnetic field. The spin-dependent Hamiltonian can be constructed by expanding the effective Hamiltonian in powers of $\boldsymbol{k}$ and applying the method of invariants. For the T_d-symmetry crystals, one can show that the spin-dependent term in the conduction-band Hamiltonian appears starting with k^3 [34]

$$\mathcal{H}_{c3} = \gamma_c \left[\sigma_x k_x (k_y^2 - k_z^2) + \sigma_y k_y (k_z^2 - k_x^2) + \sigma_z k_z (k_x^2 - k_y^2) \right] \,, \qquad (61)$$

where σ_i are the Pauli matrices and $x \parallel [100], y \parallel [010], z \parallel [001]$. For the band Γ_8 the main contribution is given by a similar expression

$$\mathcal{H}_{v3} = \gamma_v \left[J_x k_x (k_y^2 - k_z^2) + J_y k_y (k_z^2 - k_x^2) + J_z k_z (k_x^2 - k_y^2) \right] \,, \sum_i J_i \kappa_i \,, \qquad (62)$$

where J_i are the 4×4 matrices of the angular-momentum operators $\hat{J}_i$ in the basis $\mathrm{Y}_{3/2,m}$. In contrast to the conduction band, the constant γ_v has a non-relativistic nature. Indeed, this constant equals 2/3 of the similar constant in the 3×3 Hamiltonian for the non-relativistic valence band Γ_{15}.

In heterostructures, QWs or SLs, grown from zinc-blende-lattice semiconductors, the spin-dependent Hamiltonians contains both linear and cubic terms. Particularly, in (001)-grown QWs with symmetrical interfaces (the D_{2d} point-group symmetry) the linear-$\boldsymbol{k}$ spin-dependent term in the conduction subband $e1$ has the form [35]

$$\mathcal{H}_{\mathrm{BIA}} = \beta_1(\sigma_y k_y - \sigma_x k_x), \tag{63}$$

where β_1 is a constant. This term can be obtained from the cubic-$\boldsymbol{k}$ term (61) describing the removal of spin degeneracy of the conduction-band states in a bulk semiconductor. Really, taking into account the quantum confinement effect we can replace in (61) k_z and k_z^2 by the average values $\langle k_z \rangle = 0$ and $\langle k_z^2 \rangle \neq 0$, respectively, and arrive at (63) with $\beta_1 = \gamma_c \langle k_z^2 \rangle$. Here, the symbol $\langle k_z^l \rangle$ means the expectation value $\int dz \varphi_{e1}(z) \hat{k}_z^l \varphi_{e1}(z)$ with $\varphi_{e1}(z)$ being the electron envelope function at the lowest subband $e1$ and $\hat{k}_z = -i\partial/\partial z$. Since the term $\mathcal{H}_{\mathrm{BIA}}$ is due to the lack of inversion symmetry in the bulk material it is called the Bulk Inversion Asymmetry (BIA) term which explains the subscript BIA. Sometimes it is also called the Dresselhaus term [36].

In heterostructures with asymmetrical superstructural potential (the C_{2v} point group) there exists another spin-dependent contribution

$$\mathcal{H}_{\mathrm{SIA}} = \beta_2(\sigma_x k_y - \sigma_y k_x) \tag{64}$$

which is called the Structure-Inversion Asymmetry (SIA) term, or the Rashba term. The structure asymmetry can be related with non-equivalent normal and inverted interfaces, external or built-in electric fields, compositionally stepped QWs etc. The spin-orbit interaction term in the form of (64) was first predicted in [37,38] for bulk polar hexagonal crystals with the wurtzite structure (the C_{6v} point symmetry). Nature of the similar term in an asymmetrical 2D system has been analyzed by different authors [39–45] (see also [46] and references therein). Note that anisotropic orientation of chemical bonds at the interfaces gives rise to an additional contribution to the linear-$\boldsymbol{k}$ Hamiltonian, even in symmetrical QWs [47,48]. This is the so-called Interface-Inversion Asymmetry (IIA) term. Since it has the same structure as the contribution (63), they can be described together by one inseparable Dresselhaus term with the common parameter β_1.

By using the Cartesian coordinates $x' \| [1\bar{1}0]$, $y' \| [110]$, $z \| [001]$ one can write a sum of the BIA and SIA terms in the form

$$\mathcal{H}_{\mathrm{c1}}(\boldsymbol{k}) = \frac{1}{2}\,(\beta_- \sigma_{x'} k_{y'} - \beta_+ \sigma_{y'} k_{x'})\ , \tag{65}$$

where $\beta_\pm = 2(\beta_2 \pm \beta_1)$ and from now on $\boldsymbol{k} \equiv \boldsymbol{k}_\parallel$. Introducing the effective Larmor frequency $\boldsymbol{\Omega}_{\boldsymbol{k}}$ (at zero magnetic field) by

$$\mathcal{H}_{\mathrm{c1}} = \frac{\hbar}{2}\,\boldsymbol{\sigma} \cdot \boldsymbol{\Omega}_{\boldsymbol{k}} \tag{66}$$

and comparing with (65) we obtain

$$\Omega_{\boldsymbol{k},x'} = \beta_- k_{y'}/\hbar\,, \quad \Omega_{\boldsymbol{k},y'} = \beta_+ k_{x'}/\hbar\,, \quad \Omega_{\boldsymbol{k},z} = 0. \tag{67}$$

Thus, in the parabolic approximation the resulting energy dispersion is

$$E_{e1\boldsymbol{k}} = E^0_{e1} + \frac{\hbar^2 k^2}{2m^*} \pm \frac{1}{2}\,\hbar\Omega_{\boldsymbol{k}}\,,$$

with the spin splitting given by

$$\Delta E = \hbar\Omega_{\boldsymbol{k}} = \sqrt{\beta_+^2 k_{x'}^2 + \beta_-^2 k_{y'}^2}. \tag{68}$$

If only one of the constants β_1, β_2 is nonzero then $\beta_-^2 = \beta_+^2 \equiv \beta^2$ and the splitting $\hbar\Omega_{\boldsymbol{k}} = \beta k$ is angular independent.

At low electron energies the spin splitting due to the linear-in-$\boldsymbol{k}$ term dominates. At higher energies relevant to high temperatures or large concentrations, the cubic-in-$\boldsymbol{k}$ term can be important as well.

The Rashba term can be presented in an invariant form as

$$\mathcal{H}_{\mathrm{SIA}} = \beta_2 [\boldsymbol{\sigma} \times \boldsymbol{k}] \cdot \boldsymbol{N}\,, \tag{69}$$

where $\boldsymbol{N}$ is the unit vector directed along the normal to the interface. In a symmetrical QW subject to a homogeneous electric field $\boldsymbol{F} \parallel z$, the constant β_2 can be crudely estimated as

$$\beta_2 = \frac{P^2}{3} \frac{\Delta(2E_g + \Delta)}{E_g^2(E_g + \Delta)^2} \left\{ |e|F + V\left[\varphi_{e1}^2(a/2) - \varphi_{e1}^2(-a/2)\right]\right\}\,,$$

where V is the conduction-band offset, the parameter P is defined by

$$P = i\frac{\hbar p_{cv}}{m_0}\,,\ p_{cv} = \langle S|\hat{p}_z|Z\rangle\,,$$

$\varphi_{e1}(z)$ is the envelope calculated in the presence of electric field, and $\pm a/2$ are the interface coordinates.

The BIA contribution to the valence band effective Hamiltonian responsible for removal of spin degeneracy of the hole subbands can be obtained by averaging the odd-in-$\boldsymbol{k}$ terms over the quantum-confined states $hh\nu$ or $lh\nu$ calculated neglecting the spin splitting. Note that this procedure applied to the non-relativistic term (62) does not lead to linear-$\boldsymbol{k}$ splitting of heavy-hole states because the off-diagonal components $J_{i;\pm 3/2,\mp 3/2}$ equal zero. The Rashba spin splitting in 2D hole systems was analyzed by Winkler [49]. At small values of $\boldsymbol{k}$, the heavy-hole spin splitting is of third order, in qualitative difference with the conduction-band and light-hole states in QWs.

Linear-in-$\boldsymbol{k}$ spin splitting of electron subbands in QWRs is described by a Hamiltonian

$$\mathcal{H}_1 = (\boldsymbol{\sigma} \cdot \boldsymbol{\beta})\, k_z\,, \tag{70}$$

where z is the principal axis of the wire and $\boldsymbol{\beta}$ is a constant vector, its non-zero components are determined by symmetry of the structure. The effective Larmor frequency defined according to (66) is equal to $2\boldsymbol{\beta}k_z/\hbar$. Depending on the sign of k_z it is directed parallel or anti-parallel to the fixed direction of $\boldsymbol{\beta}$. This is the main difference with QWs where the direction of $\boldsymbol{\Omega_k}$ is independent of $\boldsymbol{k}$ only in asymmetrical QWs in the particular case $\beta_1 = \pm\beta_2$ when either β_+ or β_- vanish. In [50] a simple 3D model of an asymmetric QWR is introduced in which the Rashba spin-orbit coupling (70) is derived from a realistic description of the bulk semiconductor electronic structure.

References

1. E.L. Ivchenko, G.E. Pikus: Pis'ma ZhETF **27**, 640 (1978) [JETP Lett. **27**, 604 (1978)]
2. V.I. Belinicher: Phys. Lett. A **66**, 213 (1978)
3. V.I. Belinicher, B.I. Sturman: Usp. Fiz. Nauk **130**, 415 (1980) [Sov. Phys. Usp. **23**, 199 (1980)]
4. B.I. Sturman, V.M. Fridkin: *The Photovoltaic and Photorefractive Effects in Non-Centrosymmetric Materials* (Gordon and Breach Science Publishers, New York 1992)
5. V.M. Asnin, A.A. Bakun, A.M. Danishevskii, E.L. Ivchenko, G.E. Pikus, A.A. Rogachev: Pis'ma ZhETF **28**, 80 (1978); Solid State Commun. **30**, 565 (1979)
6. N.S. Averkiev, V.M. Asnin, A.A. Bakun, A.M. Danishevskii, E.L. Ivchenko, G.E. Pikus, A.A. Rogachev: Fiz. Tekh. Poluprovodn. **18**, 639 (1984); *ibid* **18**, 648 (1984)
7. S.D. Ganichev, H. Ketterl, W. Prettl, E.L. Ivchenko, L.E. Vorobjev: Appl. Phys. Lett. **77**, 3146 (2000)
8. S.D. Ganichev, E.L. Ivchenko, S.N. Danilov, J. Eroms, W. Wegscheider, D. Weiss, W. Prettl: Phys. Rev. Lett. **86**, 4358 (2001)
9. S.D. Ganichev, U. Rössler, W. Prettl, E.L. Ivchenko, V.V. Bel'kov, R. Neumann, K. Brunner, G. Abstreiter: Phys. Rev. **B 66**, 75328 (2002)
10. S.D. Ganichev, W. Prettl: J. Phys.: Condens. Matter **15**, 935 (2003)
11. L.E. Golub: Physica **E 17**, 342 (2003); Phys. Rev. **B 67**, 235320 (2003)
12. S.D. Ganichev, V.V. Bel'kov, P. Schneider, E.L. Ivchenko, S.A. Tarasenko, W. Wegscheider, D. Weiss, D. Schuh, E.V. Beregulin, W. Prettl: Phys. Rev. **B 68**, 35319 (2003)
13. E.L. Ivchenko, Yu.B. Lyanda-Geller, G.E. Pikus: JETP Lett. **50**, 175 (1989); Zh. Eksp. Teor. Fiz. **98**, 989 (1990) [Sov. Phys. JETP **71**, 550 (1990)]
14. S.D. Ganichev, E.L. Ivchenko, V.V. Bel'kov, S.A. Tarasenko, M. Sollinger, D. Weiss, W. Wegscheider, W. Prettl: Nature **417**, 153 (2002)
15. N.S. Averkiev, M.I. D'yakonov: Sov. Phys. Semicond. **17**, 395 (1983)
16. S.A. Tarasenko, E.L. Ivchenko, V.V. Bel'kov, S.D. Ganichev, D. Schowalter, P. Schneider, M. Sollinger, W. Prettl, V.M. Ustinov, A.E. Zhukov, L. E. Vorobjev: J. Superconductivity: Incorporating Novel Magnetism **16**, 419 (2003)
17. S.D. Ganichev, P. Schneider, V.V. Bel'kov, E.L. Ivchenko, S.A. Tarasenko, W. Wegscheider, D. Weiss, D. Schuh, B.N. Murdin, P.J. Phillips, C.R. Pidgeon, D.G. Clarke, M. Merrick, P. Murzyn, E.V. Beregulin, W. Prettl: Phys. Rev. B **68**, 081302 (2003)

18. A.M. Glass, D. von der Linde, T.J. Negran: Appl. Phys. Lett. **25**, 233 (1974)
19. A.M. Glass, D. von der Linde, D.H. Auston, T.J. Negran: J. Electr. Mater. **4**, 915 (1975)
20. K.H. Herrmann, R. Vogel: Proc. XI Int. Conf. Phys. Semicond, Warsaw, Poland, 1972 (Elsevier, Amsterdam, 1972) p. 870
21. C.R. Hammond, J.R. Jenkins, C.R. Stanley: Optoelectronics **4**, 189 (1972)
22. A.V. Andrianov, E.L. Ivchenko, G.E. Pikus, R.Ya. Rasulov, I.D. Yaroshetskii: Zh. Eksp. Teor. Fiz. **81**, 2080 (1981) [Sov. Phys. JETP **54**, 1105 (1981)]
23. N. Kristoffel, A. Gulbis: Izv. A.N. Est. SSR **28**, 268 (1979)
24. R. von Baltz, W. Kraut: Phys. Lett. A **79**, 364 (1980)
25. V.I. Belinicher, E.L. Ivchenko, B.I. Sturman: Zh. Eksp. Teor. Fiz. **83**, 649 (1982) [Sov. Phys. JETP **56**, 359 (1982)]
26. E.L. Ivchenko, Yu.B. Lyanda-Geller, G.E. Pikus: Fiz. Tekh. Poluprovodn. **18**, 93 (1984) [Sov. Phys. Semicond. **18**, 55 (1984)]
27. F. Capasso, S. Luryi, W.T. Tzang, C.G. Bethea, B.F. Levine: Phys. Rev. Lett. **51**, 2318 (1983)
28. F.G. Pikus: Fiz. Tekh. Poluprovodn. **22**, 940 (1988) [Sov. Phys. Semicond. **22**, 594 (1988)]
29. C. Schönbein, H. Schneider, G. Bihlmann, K. Schwarz, P. Koidl: Appl. Phys. Lett. **68**, 973 (1995)
30. H. Schneider, S. Ehret, C. Schönbein, K. Schwarz, G. Bihlmann, J. Fleissner, G. Ränkle: Superlatt. Microstruct. **23**, 1289 (1998)
31. S.D. Ganichev, S.N. Danilov, V.V. Bel'kov, E.L. Ivchenko, M. Bichler, W. Wegscheider, D. Weiss, W. Prettl: Phys. Rev. Lett. **88**, 057401 (2002)
32. L.E. Vorobjev, E.L. Ivchenko, G.E. Pikus, I.I. Farbshtein, V.A. Shalygin, A.V. Shturbin: JETP Lett. **29**, 441 (1979)
33. P. Schneider, S.D. Ganichev, J. Kainz, U. Rössler, W. Wegscheider, D. Weiss, W. Prettl, V.V. Bel'kov, L.E. Golub, D. Schuh: phys. stat. sol. (b) **238**, 533 (2003)
34. G. Dresselhaus: Phys. Rev. **100**, 580 (1955)
35. M.I. D'yakonov, V.Yu. Kachorovskii: Fiz. Tekhn. Poluprov. **20**, 178 (1986) [Sov. Phys. Semicond. **20**, 110 (1986)]
36. W. Knap, C. Skierbiszewski, A. Zduniak, E. Litwin-Staszewska, D. Bertho, F. Kobbi, J.L. Robert, G.E. Pikus, F.G. Pikus, S.V. Iordanskii, V. Mosser, K. Zekentes, Yu.B. Lyanda-Geller: Phys. Rev. **B 53**, 3912 (1996)
37. E.I. Rashba, V.I. Sheka: Fiz. Tverd. Tela, Collected Papers, vol. 2, 162 (1959)
38. E.I. Rashba: Fiz. Tverd. Tela **2**, 1224 (1960) [Sov. Phys. Solid State **2**, 1109 (1960)]
39. F.G. Ohkawa, Y. Uemura: J. Phys. Soc. Japan **37**, 1325 (1974)
40. F.T. Vas'ko: Pis'ma Zh. Eksp. Teor. Fiz. **30**, 574 (1979) [JETP Lett. **30**, 541 (1979)]
41. Yu.A. Bychkov, E.I. Rashba: Pis'ma Zh. Eksp. Teor. Fiz. **9**, 66 (1984) [JETP Lett. **39**, 78 (1984)]
42. B. Das, S. Datta, R. Reifenberger: Phys. Rev. **B 41** 8278 (1989)
43. E.A. de Andrada e Silva, G.C. Rocca, F. Bassani: Phys. Rev. **B 55**, 16293 (1997)
44. L. Wissinger, U. Rössler, R. Winkler, B. Jusserand, D. Richards: Phys. Rev. **B 58**, 15375 (1998)
45. P. Pfeffer, W. Zavadzki: Phys. Rev. **B 59**, 5312 (1999)
46. N.S. Averkiev, L.E. Golub, M. Willander: Fiz. Tekhn. Poluprov. **36**, 97 (2002) [Semiconductors **36**, 91 (2002)]
47. L. Vervoort, R. Ferreira, P. Voisin: Phys. Rev. **B 56**, 12744 (1997)
48. U. Rössler, J. Kainz: Solid State Commun. **121**, 313 (2002)

49. R. Winkler: Phys. Rev. **B 62**, 4245 (2000)
50. E.A. de Andrada e Silva, G.C. La Rocca: Phys. Rev. **B 67**, 165318 (2003)
51. E.L. Ivchenko, B. Spivak: Phys. Rev. **B 66**, 155404 (2002); Physica **E 17**, 376 (2003)
52. V.V. Bel'kov, S.D. Ganichev, P. Schneider, C. Back, M. Oestreich, J. Rudolph, D. Hägele, L.E. Golub, W. Wegscheider, W. Prettl, Solid State Commun. **128**, 283 (2003)
53. E.L. Ivchenko: *Optical Spectroscopy of Semiconductor Nanostructures* (Alpha Science International Ltd., Pangbourne, UK 2004)

Nano-Photoluminescence

Heinz Kalt

Institut für Angewandte Physik and DFG-Center for Functional Nanostructures,
Universität Karlsruhe (TH), 76128 Karlsruhe (Germany)

Abstract. We discuss various modifications of spatially resolved photoluminescence and their application in semiconductor spectroscopy. The methods described are micro-photoluminescence (μ-PL) under global excitation, μ-PL imaging, and their extension to nano-photoluminescence by using a confocal setup in combination with solid immersion lenses (SIL). In particular we outline the implementation of SILs into a μ-PL experiment and demonstrate their merits in terms of enhanced resolution and collection efficiency. We demonstrate the huge potential of spatially resolved photoluminescence by its application in the characterization of extended defects, in the ultra-high precision spectroscopy of electron–phonon coupling, in the detection of coherent and non-thermal transport of excitons on the length scale of the light wavelength, and in single-dot spectroscopy.

1 Introduction

Photoluminescence (PL) is one of the most important methods to characterize semiconductor structures and to investigate their electronic states. Since both fundamental research as well as optoelectronic applications increasingly focus on semiconductor nanostructures and on a length scale well below the wavelength of light, PL has been developed into a local spectroscopy method.

There are several ways to achieve a spatial resolution of one micrometer or less. Single nano-objects like quantum dots or wires can be addressed when they are sufficiently separated in space or when they can be distinguished by differing energies of their optical transitions. The number of nano-objects in dense ensembles - like quantum dots laterally embedded in a wetting layer or a quantum well [1] or interfacial islands [2] - can be significantly reduced e.g. by the etching of mesas or by coating the sample surface leaving only nano-apertures of typical sizes of 100 nm [3].

One can also realize sub-micron resolution in PL experiments by local excitation, local detection, or a combination of both techniques. In far-field optics, the resolution is limited by diffraction to the order of the light wavelength. This limitation can be overcome by working in the near-field regime, where the diffraction limit is not yet established. Scanning near-field optical microscopy (SNOM) [4], designed according to this idea, realizes a resolution of less than 100 nm. These conditions are typically achieved for either global excitation and local detection with the SNOM or local excitation and global detection. For simultaneous local excitation and detection one employs uncoated fiber tips for sufficient signal strength [5]. The spatial resolution is then in the range of 100-200 nm. SNOM

H. Kalt, Nano-Photoluminescence, Lect. Notes Phys. **658**, 51–70 (2005)
http://www.springerlink.com/

set-ups can be integrated into cryostats to work at low temperatures. The use of picosecond laser pulses together with time-resolved detection schemes like a streak camera allows the combination of high spatial and temporal resolution [6]. Finally, SNOM systems are intrinsically coupled to scanning surface microscopy providing additional information on the surface morphology.

Despite these large number of merits of the SNOM one finds an attractive alternative for the investigation of semiconductors in PL set-ups based on microscopes, the so-called micro-photoluminescence (μ-PL). These set-ups are much more flexible and easier to implement, have a high collection efficiency and achieve a spatial resolution (without any sample structuring) down to 200 nm. The typical resolution of a confocal μ-PL (for a review, see e.g. [7]) can be enhanced to about 0.5 μm by increasing the effective numerical aperture ($\mathrm{NA_{eff}}$) of the optical system. This can be realized for one by using reflective optics, in particular parabolic mirrors resulting in a numerical aperture NA ≈ 1 [8]. A second approach are solid immersion lenses (SILs) placed on the surface of the sample [9]. Both systems have the possibility, not given for an ordinary SNOM, to individually shift the position of the high resolution detection spot with respect to the excitation spot.

We will discuss in the following various implementations of μ-PL imaging and spectroscopy and its enhancement by the use of solid immersion lenses to nano-PL. We will focuss on the technical details of these methods and illustrate their merits and limitations by giving typical examples of spectroscopic results.

2 Definition of Spatial Resolution

In a far-field optical system, the spatial resolution is limited by diffraction of the light at the aperture of the imaging optics. This diffraction results in the fact that a point source produces an Airy pattern in the image plane. The spatial resolution is then often defined by the Rayleigh criterion. Two adjacent point sources are here considerd to be spatially resolvable when the maximum of the Airy pattern of the first light source coincides with the first minimum of the Airy pattern of the second one. For practical purposes this criterion is difficult to verify, because the minimum of an Airy pattern is typically not well defined since its is hidden within the noise of the detection. A much easier way to measure spatial resolution is to determine the half width at half maximum (HWHM) of the maximum of the Airy pattern (Sparrow criterion). This definition of the spatial resolution differs from the Rayleigh criterion by a factor of 2.

Taking into account not only the geometrical aperture $\mathrm{NA_{obj}}$ of the imaging objective, but also the fact that the refractive index n of the medium surrounding the object affects the light diffraction one finds for the HWHM:

$$\mathrm{HWHM} = \frac{0.26\lambda}{n\mathrm{NA_{obj}}}. \tag{1}$$

This definition of the spatial resolution directly points out where to start with its optimization. Of course one can, to some extent, increase the geometrical

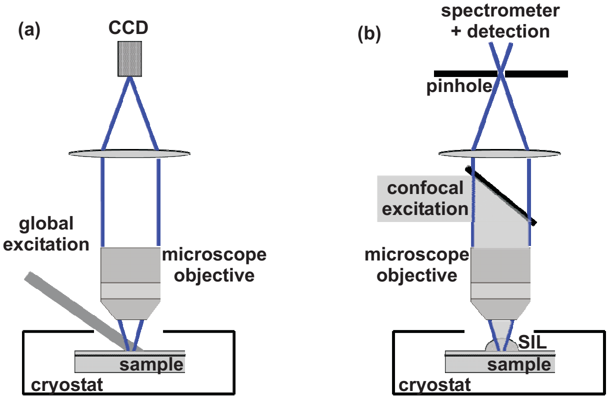


Fig. 1. (**a**) μ-PL imaging under global excitation, (**b**) SIL-enhanced nano-PL based on a confocal microscope

aperture $\mathrm{NA}_{\mathrm{obj}}$. This is limited by the size of the aperture achievable without introducing aberrations and by the working distance (distance of the objective from the object). The latter is typically of the order of 10 mm when the object is placed inside a cryostat and the objective is situated outside separated by a thin window. Smaller working distances of a few mm are possible when the ojective is also placed within the dewar. Then the sample can be immersed within an immersion fluid to enhance the refractive index. But this is not realistic in the case of semiconductors kept at He-temperature. Here on can use solid immersion lenses (SILs) as will be described in Sect. 4. Finally one has to consider the influence of the detection scheme (pinhole in the image plane, CCD array detector) and its interplay with the excitation scheme (e.g. in a confocal arrangement).

Figure 1 shows two modifications of spatially resolved photoluminescence, namely a μ-PL set-up with global excitation (Fig. 1a) and a confocal set-up enhanced by a SIL (Fig. 1b). With the set-up in (a) one can typically achieve a spatial resolution of 350 nm (at a wavelength of 440 nm) while 220-250 nm are typical for the set-up in (b). The objective used here has a working distance of 10 mm, a magnification of 20 and a numerical aperture of $\mathrm{NA}_{\mathrm{obj}} = 0.4$. In the following we want to discuss the merits of the different modifications and give examples for their application in semiconductor spectroscopy.

3 μ-PL Under Global Excitation and μ-PL Imaging

Figures 2 and 3 demonstrate the use of μ-PL (set-up as in Fig. 1a) for PL imaging and spectroscopy with high spatial resolution. The excitation source for the μ–PL experiments is a continuous wave (cw) laser. A set of pinholes with different sizes is installed in the image plane of the microscope to select the detection area on the sample surface. By scanning the pinhole in the image plane (for distances exceeding several 10 μm one has to scan the sample), one can detect luminescence from different positions on the sample surface. The pinhole is then

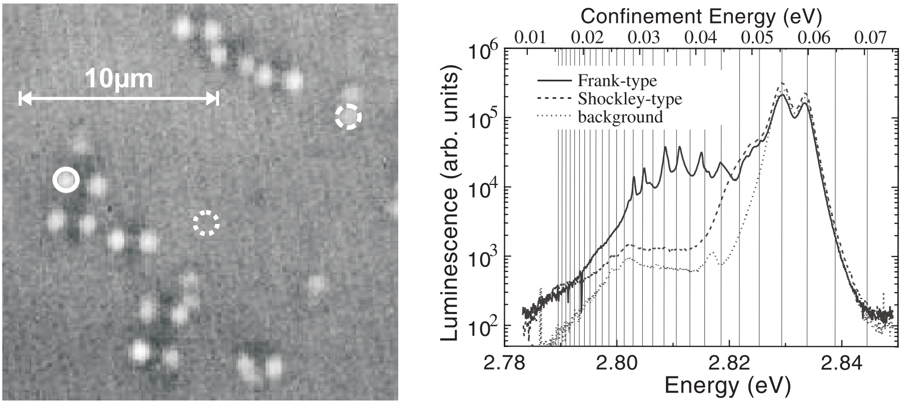


Fig. 2. (**a**) μ-PL image of a ZnSe QW showing aligned pairs of bright-spot defects, (**b**) μ-PL spectra taken at various positions of the PL landscape in (a) [10]

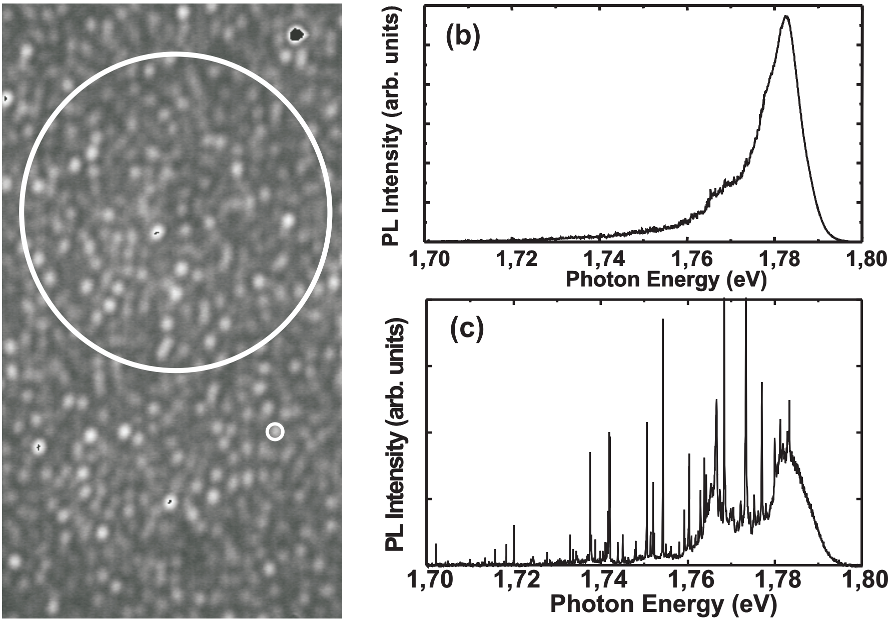


Fig. 3. (**a**) μ-PL image of a type-II GaAs/AlAs quantum well showing local type-I band alignment. (**b**) macro-PL and (**c**) μ-PL of the sample [11]

imaged onto the entrance slit of a spectrometer. A shiftable set of a mirror and a lens is installed in front of the spectrometer, reflecting and focusing the light onto a CCD camera connected to a monitor. This configuration achieves a direct imaging of the sample surface on the monitor, thus ensures fine alignments of the sample, the SIL, the objective, and the pinhole. Also, by removing the pinhole, one can take PL intensity maps (PL imaging). No scanning is needed here to acquire an intensity map with an area of several 100 μm^2. When shifting the mirror out of the optical path, the signal is sent into the spectrometer with a spectral resolution of 30 μeV and finally recorded by a cooled CCD camera.

This combination of imaging and spectroscopy is very useful for samples with local structural features like extended defects or single embedded nanostructures. The first example shown here is the characterization of bright-spot defects induced by stacking–fault pairs in ZnSe-based quantum wells [10]. The PL landscape (Fig. 2a) displays pairs of bright spots well aligned either in [110] or [$\bar{1}$10] direction, respectively. Statistical analysis of the spot separation within each pair gives a fixed distance for the [110] pairs, while a distribution of separations with a standard deviation of 30% is found for the other orientation. The separation within the [110] pairs directly scales with the overall thickness of the epitaxial structure roughly like $\sqrt{2}\times$thickness. The enhancement of the emission in the region of the bright-spot defects with respect to the background PL is found to be up to 20%.

Additional structural investigation by atomic-force microscopy and transmission electron microscopy (plain-view as well as cross-section) reveal that a widening and bending of quantum wells occurs during growth, when they are intersected by Frank-type stacking faults. These stacking faults originate at the interface between the GaAs substrate and the (Zn,Mg)Se barrier layer. The areal defects come in pairs oriented along [110] with an increasing spatial separation within the pair as a function of the thickness of the epitaxial layer. The enlargement of the well width by up to 12 bilayers evokes an efficient localization of excitons. The localizing potential related to Shockley-type stacking fault pairs aligned in [$\bar{1}$10] direction is found to be much shallower.

The results of the exciton localization can be investigated in more detail by local μ-PL spectroscopy (Fig. 2b). Here, the pinhole in the image plane is placed such as to detect the local signal from the different positions on the sample (see circles in Fig. 2a). The localization of the excitons occurs in areas of the quantum well enlarged in consecutive steps of one bilayer each. This is obvious from the peaked PL emission of the Frank-type defects in comparison to calculations of the expected exciton energy (see vertical lines in Fig. 2b).

These investigations revealed that stacking faults - in contrast to a common prejudice - do not lead to non-radiative centers but can even enhance the PL intensity due to localization. However, further PL imaging at different wavelength using a low-temperature SNOM showed that the line defects bounding the stacking faults are sources for non-radiative recombination. For more details of these studies see [5,10].

The second example in this section is related to local type-I centers in type-II GaAs/AlAs multiple quantum wells. Depending on the thickness of the GaAs wells these samples are of a type-I or a type-II band alignment (see e.g. [12]). In wide quantum wells one finds the lowest electron states within the GaAs layers and related to the conduction band minimum at the Γ point. Due to the increasing confinement energy the alignment changes to type-II for narrow quantum wells. Now, the lowest electron states are within the AlAs layers and related to the conduction-band minimum at the X point. Since the electrons are at high momentum and separated from the holes, which are situated in the GaAs layers, it is obvious that the oscillator strength of excitons in the type-II structures is much smaller as compared to those in the type-I quantum wells.

In type-II GaAs/AlAs samples with a quantum well thickness close to the transition of the band alignment one finds local areas with type-I character. In the μ-PL image these areas show up as bright spots (see Fig. 3a). Looking at the luminescence spectra from this sample one sees a splitting of the PL band using a macroscopic spot (Fig. 3b) into an ensemble of sharp lines (Fig. 3c) when chosing a bright spot with the pinhole (small circle in Fig. 3a). Due to the high spectral resolution of the setup of 30 μeV one can now address a large number of individual lines related to excitons in local type-I areas for high-resolution spectroscopy. It was e.g. possible from temperature dependent studies of the line positions to determine the shift of the band gap in GaAs due to electron-phonon coupling with an ultra-high precision of 5 μeV [11].

4 SIL-Enhanced Nano-Photoluminescence

4.1 Implementation of a SIL

Solid immersion lenses come in the form of h-SILs (hemispheres) and s-SILs (superspheres) as illustrated in Fig. 4. During the last decade, SILs have been used in solid-immersion microscopes [13] and (magneto-)optical data storage [14] for high spatial resolution or high storage density, respectively. Recently, SILs have also been introduced in spatially resolved pump-probe experiments [15,16] and photoluminescence experiments [9,17,18] on semiconductors. By including a superspherical SIL [19] in a microscope system, an improved spatial resolution at room temperature [17] as well as at low temperatures [18] has been demonstrated by PL-imaging measurements of GaAs quantum wells (QW). The high spatial resolution has allowed to study carrier migration under global [18] or local [20] excitation conditions. Besides, an s-SIL has also been used in a spatially resolved PL setup to investigate exciton localization in GaAs QW [21,22].

Up to now, mostly s-SILs have been applied in PL systems. But, the thickness of an s-SIL is designed for one particular wavelength since the incident parallel beam to the objective is focused at the distance $r(1+1/n_{\mathrm{SIL}})$ away from the top

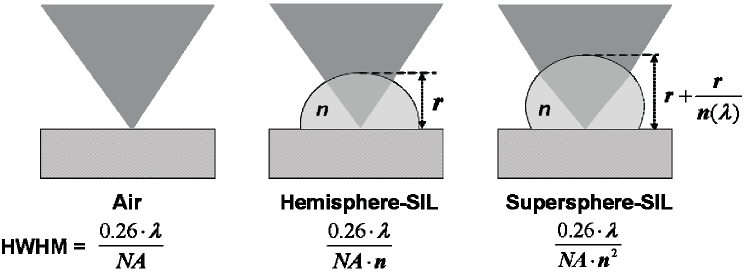

n
r
n
$r+\frac{r}{n(\lambda)}$
Air
Hemisphere-SIL
Supersphere-SIL
HWHM = $\frac{0.26\cdot\lambda}{NA}$
$\frac{0.26\cdot\lambda}{NA\cdot n}$
$\frac{0.26\cdot\lambda}{NA\cdot n^2}$

Fig. 4. Sketch of the implementation of SILs and the resulting spatial resolution

of the s-SIL, where r is the radius of the SIL and n_{SIL} is the refractive index of the SIL material. Consequently, the focus of an s-SIL is wavelength-dependent since n_{SIL} depends on the wavelength of light, λ. In contrast, a hemispherical SIL (h-SIL) is universal for any wavelength. In a PL experiment, one typically deals with different wavelengths for excitation and detection. Thus, although an s-SIL can improve the resolution n_{SIL}^2 times while an h-SIL can only improve it n_{SIL} times, the latter is more appropriate for PL studies.

Our nano-PL set-up (see Fig. 1b) uses an h-SIL made of $\mathrm{ZrO_2}$ with $n_{\mathrm{SIL}} = 2.16$ at $\lambda = 600$ nm which is adhesively fixed to the sample surface. The sample with the SIL is vertically mounted inside a helium-flow cryostat. The SIL can be used in the temperature range of 6–300 K for an unlimited number of cooling cycles. The diameter of the SIL is chosen to be 1 mm, which is large enough for giving a sufficently large working area for spectroscopy and for being handled without any special equipment, but is still small enough to be stuck on the sample adhesively even in a vertical configuration.

The excitation source for the nano-PL experiments can be a continuous wave (cw) or a pulsed laser. The laser beam is expanded to fit the diameter of the objective, then reflected by a beam-splitter and focused onto the sample surface through the microscope objective as seen in Fig. 1b. The same objective is used for collecting the PL from the sample. The signal passes the beam-splitter and is focused by the tube lens onto the image plane of the microscope. The detection scheme is similar as in Sect. 3. The pinhole installed in the image plane allows to separately move the detection and excitation spots. Besides cw measurements, where the PL is recorded by a cooled CCD camera, we can perform time-resolved measurements using a streak camera with a temporal resolution of 2 ps in combination with a CCD camera in photon-counting mode.

4.2 Spatial Resolution of the SIL-Enhanced Nano-PL

The spatial resolution of an optical system was defined in Sect. 2. Without SIL, $n \approx 1$ (air) and with the h-SIL we have $n = n_{\mathrm{SIL}} = 2.16$. Thus, by introducing the h-SIL, we can improve the resolution by more than two times in diameter, thus four times in spot area. In order to confirm the achieved resolution, we install the SIL onto a sample with a flat surface and focus the incident laser beam of a He–Ne laser onto the sample surface (i)underneath or (ii)outside the SIL, respectively. We measure two-dimensional intensity maps of the laser spots in both cases, as shown in Fig. 5a with the same gray scale encoding. The length-scale calibration in these maps is obtained by imaging an optical grating with known parameters. The spatial intensity profiles in Fig. 5b are obtained by taking a line-scan across the laser spots. As expected, the profile with SIL (i) is about two times narrower than that obtained without SIL (ii). In (i) the realized spatial resolution (HWHM of the laser spot) is 0.4 λ [corresponding to 260 nm for the He–Ne laser (633 nm)] in contrast to 0.8 λ for (ii). We note that the achieved values of HWHM in both cases are larger than that calculated from (1). This is consistent with theoretical estimations [23], and can be attributed

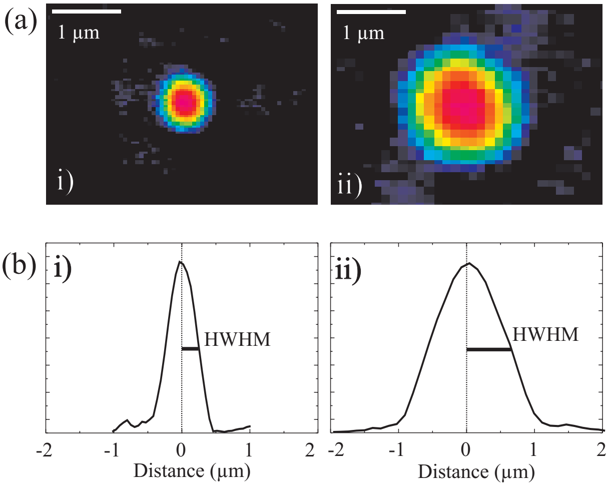


Fig. 5. Intensity maps (a) and cross-sections (b) of the focused laser spot onto an arbitrary sample. The width of the spot obtained with SIL (i) is n_{SIL} times narrower than that without SIL (ii) [9]

to the Gaussian profile of the laser beam used in the experiment rather than a plane wave [24], and the high NA of the system [23].

In a confocal microscope system, the resolution can be further improved by introducing a pinhole with a suitable size to the image plane of the microscope [24]. In the following, we present a quantitative analysis of this further improvement. The illumination function of the laser excitation can be described by a Gaussian function,

$$i_{\text{ill}}(q) = \exp\left(-2\frac{q^2}{w_{\text{laser}}^2}\right) . \tag{2}$$

Here, q is the coordinate inside the focal plane, i.e. on the sample surface, and w_{laser} is the spot radius at $1/e^2$. The detection function i_{det} can also be described by a Gaussian function, but with a different radius w_{lumi} since generally the wavelength of the luminescence is different from that of the excitation laser in a PL experiment. The transmission function of the pinhole is

$$t_{\text{p}}(q) = \text{rect}\left(\frac{q}{q_0}\right) = \begin{cases} 1 & |q| < q_0 \\ 0 & |q| > q_0 \end{cases} \tag{3}$$

with q_0 being the radius of the pinhole image. Thus, the detection probability, i.e. the probability of a photon emitted at point p to be transmitted through the pinhole, thus to be detected, is given by

$$\begin{aligned} c(q) &= t_{\text{p}}(q) * i_{\text{det}}(q) \\ &= \int_0^{q_0} q' dq' \int_0^{2\pi} d\Phi' \exp\left(-2\frac{q^2 - 2qq'\cos\Phi' + q'^2}{w_{\text{det}}^2}\right) . \end{aligned} \tag{4}$$

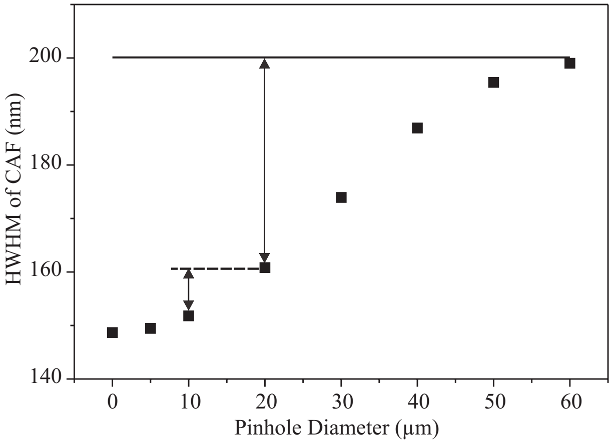


Fig. 6. Calculated HWHM of the confocal acceptance function (CAF) as function of the pinhole diameter. The horizontal line represents the HWHM obtained with an infinite large pinhole. In the calculation, the excitation and detection wavelengths are 476.5 nm and 529 nm, respectively, which are consistent with the experimental conditions [9]

The confocal acceptance function (CAF) is then given by

$$p_{\text{conf}}(q) = i_{\text{ill}}(q) \cdot c(q). \tag{5}$$

Based on the above analysis, we calculate p_{conf} of our SIL-enhanced nano-PL system. Figure 6 shows the calculated HWHM of the CAF, which defines the confocal resolution, as a function of the pinhole size. The horizontal line represents the resolution obtained without pinhole. We find that a pinhole of 60 μm has no effect on the resolution, but decreasing the pinhole size from that value the resolution is enhanced. Below 10 μm, the enhancement saturates when further decreasing the pinhole size.

In order to confirm that the enhancement of resolution by the SIL and pinhole can be achieved in a realistic PL measurement, we measure the spectra from a ZnCdSe/ZnSe quantum-dot sample with different SIL-pinhole configurations. In this sample, a ZnCdSe layer with a thickness of 2.9 monolayers is embedded between two ZnSe barriers, including Cd-rich quantum dots with an average size of about 10 nm. Excitonic transitions in individual dots lead to sharp lines observed in the PL spectrum. The variations in size, shape and composition of these dots result in a wide spectral distribution of the lines. Hence, in a macroscopic PL spectrum (not shown here) one observes a broad smooth emission band due to the large number of contributing dots. When decreasing the detection area, hence the number of the dots, individual sharp lines can be resolved on top of the unresolved smooth background. The resolved part becomes more and more pronounced with decreasing detection area. This kind of sample with a rather large dot density can be used to prove qualitatively the enhancement of the spatial resolution by introducing the SIL.

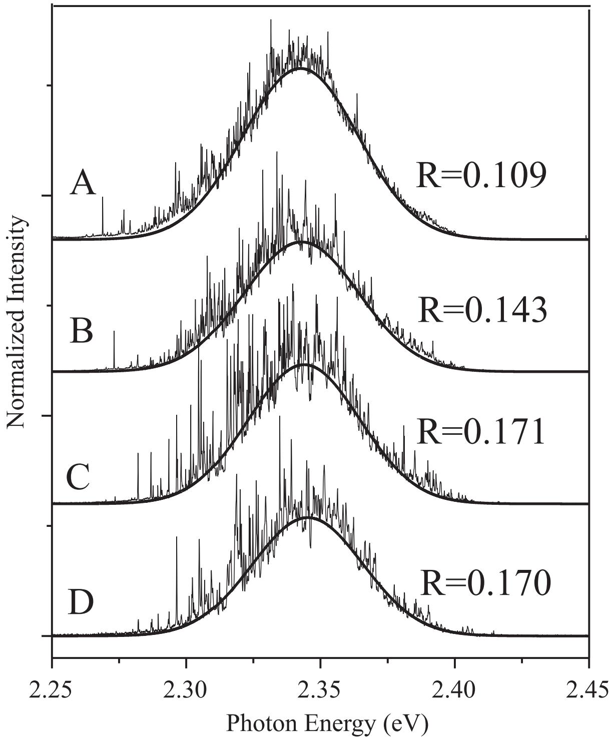


Fig. 7. PL spectra of a ZnCdSe/ZnSe quantum-dot sample measured with different configurations. A: without SIL and pinhole; B: with SIL but without pinhole; C: with SIL and 20 μm pinhole; D: with SIL and 10 μm pinhole. The low-energy side of curve C is also shown in Fig. 12b for a closer look on the sharp lines [9]

Figure 7 shows four spectra detected at a sample temperature of 6 K with different SIL-pinhole configurations, i.e. without SIL and pinhole (A), with SIL but without pinhole (B), with SIL and a pinhole of 20 μm diameter (C), with SIL and a pinhole of 10 μm diameter, respectively. The sample is excited by the 476.5 nm line of an Ar-ion laser. All spectra are composed of a resolved and an unresolved part, but the resolved sharp lines in the spectrum are more pronounced as we go from (A) to (D). We fit the background by a Gaussian function in order to separate the resolved and the unresolved part. The choice of a Gaussian is legitimate because of the inhomogeneous distribution of a large number of quantum dots contributing to the spectra. For each spectrum, we calculate the ratio, R, of the spectrally integrated intensities of the resolved part to the unresolved smooth background. This ratio increases when enhancing the resolution of the system, as we discussed above. From Fig. 4 we obtain an increase of R by 30 % by introducing the SIL [compare 0.109 for (A) to 0.143 for (B)]. By introducing a 20 μm pinhole, R is further increased by 20 % [0.171 for (C)]. In case D, a 10 μm pinhole is used instead of the 20 μm one. But we don't find a further increase of R [0.170 for (D)]. This is consistent with our analysis discussed above. As shown in Fig. 6, the enhancement of the resolution introduced by changing a 20 μm pinhole to a 10 μm one is much smaller than that from no pinhole to a 20 μm pinhole (vertical lines). In practice, the signal level drops significantly when decreasing the pinhole size from 20 μm, and the alignment becomes more difficult. Thus, a pinhole size of 20 μm is the optimal choice in our system. With this configuration, individual sharp lines can be resolved even for this kind of sample with a rather high dot density. [See the low-energy side of curve C in Fig. 7, Fig. 12b for a closer look-up, and discussions below.]

4.3 Collection Efficiency

In a PL experiment, only part of the luminescence from the sample can be collected due to the reflection losses and the finite size of the optics. The collection efficiency of a spectroscopy system is of crucial importance, especially in the cases of low-excitation conditions or low signal level. Generally, the SNOM experiments with high spatial resolution yield rather low collection efficiency. By using an uncoated tip, the collection efficiency can be significantly improved, but simultaneously the spatial resolution is limited to about 200 nm [5]. In contrast, the μ-PL is operated in the far-field regime, thus has a high collection efficiency. By introducing a SIL into a μ-PL system, the collection efficiency can be further improved [25–28]. By comparing the luminescence intensities measured with SIL and without SIL, we find an enhancement of the collection efficiency by about a factor of five. Small variations of typically less than 20 % with respect to this factor depend mainly on the cleaning process of both the sample and the SIL.

Here, we present a quantitative analysis on the enhancement of the collection efficiency introduced by using the SIL [9]. Since the n_{SIL} is smaller than the refraction index of the sample, n_{samp}, but larger than that of air, the SIL has the property to reduce the reflection losses, i.e. to enhance the transmission of both the luminescence and the laser. The enhancement of the collection efficiency by this factor, k_{T}, can be calculated by using the Fresnel formula. Figure 8a shows the configurations for our calculation of k_{T} by comparing the transmissions when the SIL is used (i) or not (ii). In case (i), since the light enters perpendicularly through the top of the SIL, the transmission coefficient from air to the SIL is given by $4n_{\mathrm{SIL}}/(1+n_{\mathrm{SIL}})^2$ for all rays.

However, when entering the sample, the transmission coefficient depends on the angle of incidence and the polarization of the ray. This angle dependence is weak in the range of angles given by the microscope objective. Thus, for average, we calculate for each polarization the transmission of a ray with an angle to the optical axis of $\theta/2$. Furthermore, the transmissions of s and p polarizations are averaged to get the total transmission. Considering the reflection losses of both the laser and the luminescence, we obtain an enhancement factor $k_{\mathrm{T}} = 1.2$.

But a more important effect than this transmission enhancement is, that the SIL enlarges the collection angle of the μ-PL system, as shown in Fig. 8b. The solid angle outside of the sample is independent of whether the SIL is used (i) or not (ii) and is directly given by $\mathrm{NA}_{\mathrm{obj}}$. However, the solid angle inside the sample

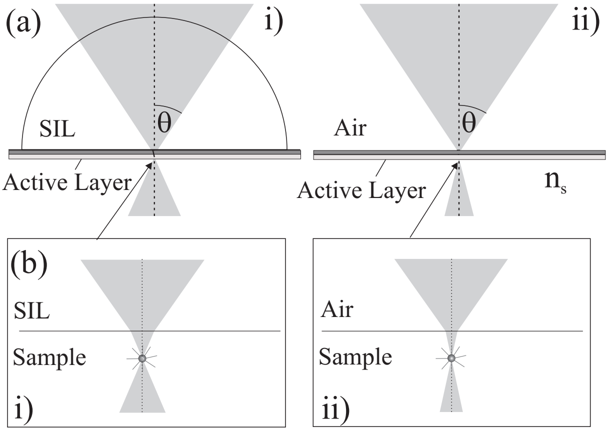


Fig. 8. Schematic drawing of the SIL-sample configuration. The enhancement of collection efficiency is explained by higher transmission (a) and larger collection angle in the sample (b) [9]

increases when the SIL is introduced. This is due to the smaller refraction of the light at the sample surface since the material on top of the sample has now a refractive index higher than that of air. As a result, a point source emitting light in all directions as shown in Fig. 8b experiences a larger solid angle in which the emitted photons can be collected by the objective. A ray emitted outside of this angle will miss the objective and will not contribute to the signal, even if its angle to the optical axis is smaller than the critical angle of total internal reflection. In the approximation that the photons are emitted from the point source homogeneously in all directions, the enhancement of collection efficiency due to the larger PL collection angle, k_Ω, is given by the ratio of the solid angles:

$$k_\Omega = \frac{\Omega_{\mathrm{SIL}}}{\Omega_{\mathrm{air}}} \approx \frac{1-\cos\left(\frac{n_{\mathrm{SIL}}}{n_{\mathrm{samp}}}\sin\theta\right)}{1-\cos\left(\frac{1}{n_{\mathrm{samp}}}\sin\theta\right)} \approx \frac{1-\left(1-\frac{1}{2}\left(\frac{n_{\mathrm{SIL}}}{n_{\mathrm{samp}}}\right)^2\sin^2\theta\right)}{1-\left(1-\frac{1}{2}\left(\frac{1}{n_{\mathrm{samp}}}\right)^2\sin^2\theta\right)} = n_{\mathrm{SIL}}^2\,. \quad (6)$$

Thus, the total enhancement of the collection efficiency by SIL is simply

$$k_{\mathrm{total}} = k_{\mathrm{T}} \cdot k_\Omega \approx k_{\mathrm{T}} \cdot n_{\mathrm{SIL}}^2\,. \quad (7)$$

In our set-up, we have $k_{\mathrm{T}} = 1.2$, $k_\Omega = 4.8$ so $k_{\mathrm{total}} = 5.76$. This calculated value is consistent with out experimental results.

4.4 Influence of an Air Gap

In the analysis of the previous section, we assumed that the SIL is ideally attached to the sample surface. In a realistic experiment, an air gap exists between the flat surface of the SIL and the sample surface due to the fluctuations of both surfaces as well as due to particles between them. In this section, we discuss the influence of such an air gap on the resolution and collection efficiency of the SIL-enhanced nano-PL system.

As discussed above, the $\mathrm{NA}_{\mathrm{eff}}$ of a SIL-enhanced nano-PL system is determined by the $\mathrm{NA}_{\mathrm{obj}}$ and n_{SIL}, i.e. $\mathrm{NA}_{\mathrm{obj}} \cdot n_{\mathrm{SIL}}$ for an h-SIL and $\mathrm{NA}_{\mathrm{obj}} \cdot n_{\mathrm{SIL}}^2$ for an s-SIL. The influence of the air gap on the resolution depends strongly on whether $\mathrm{NA}_{\mathrm{eff}} > 1$ or not. In a system with $\mathrm{NA}_{\mathrm{eff}} > 1$, a description within the near-field regime is required. This situation applies for the evanescent coupling schemes developed for technical applications [29]. Theoretical analysis has shown that even an air gap with a thickness of one fifth of the wavelength can deteriorate the resolution seriously [30]. In contrast, a system with $\mathrm{NA}_{\mathrm{eff}} < 1$ is still in the far-field regime, and it has been shown theoretically that an air gap of several micrometers does not influence the resolution [31]. In our set-up, we have $\mathrm{NA}_{\mathrm{eff}} = 0.87 < 1$ and thus far-field coupling. To check the influence of an air gap on the resolution of our system, we attach the SIL to the sample with and without cleaning procedure, respectively. In the latter case, an air gap of several micrometers is anticipated (we will prove this fact later). We focus the laser beam onto the sample surface, and in both cases we get the same size of the laser spots. Thus, we confirm that in a system with $\mathrm{NA}_{\mathrm{eff}} < 1$, an air gap of several micrometers has no influence on the resolution.

Generally, an air gap introduces additional reflection losses between the sample and the SIL, thus reducing the collection efficiency. In the near-field regime with $\mathrm{NA}_{\mathrm{eff}} > 1$, the collection efficiency can be deteriorated seriously by an air gap of several hundred nanometers, i.e. comparable to the light wavelength [31]. In contrast, a system with $\mathrm{NA}_{\mathrm{eff}} < 1$ is anticipated to be more robust due to the far-field conditions. In order to investigate the tolerance of our system to the air gap, we attach the SIL onto a ZnCdSe/ZnSe quantum-dot sample without any cleaning procedure. By comparing the spectra measured beneath or outside the SIL at a sample temperature of 6 K, we find an *enhancement* of the collection efficiency by a factor of 2.

To explain the observed enhancement, we calculate the collection efficiency of the system with an air gap. Figure 9 shows the configuration of the objective, the SIL, the air gap, and the sample. In the measurement we focus the laser beam onto the sample surface. The ray-path is shown as a solid line in Fig. 9. The dotted lines show the situation when the laser beam is focused onto the flat surface of the SIL. Using the monitor CCD (Fig. 1) we can clearly observe the images of both surfaces, thus accurately measure the difference between the two focal lengths, d. In this experiment, we have $d = 40\ \mu$m. By some simple geometrical considerations, we deduce the thickness of the air gap to be 5 μm from the measured d. Based on Fig. 9, we calculate the collection efficiency of this configuration by the method discussed in the previous section.

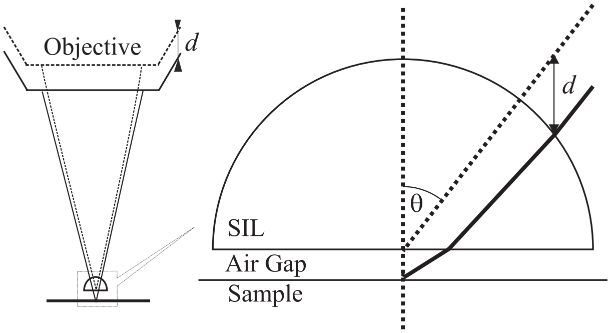


Fig. 9. Left: Schematic drawing of the objective-SIL-sample configuration when the laser beam is focused onto the sample surface (solid line) and the flat surface of the SIL (dotted line); Right: Details close to the sample surface. The angle θ is defined by the NA of the objective [9]

We obtain $k_{\mathrm{T}} = 0.55$ and $k_{\Omega} = 4.27$ so that $k_{\mathrm{total}} = 2.36$. The calculation is well consistent with the experiment. We note that the deterioration of enhancement from 5.76 to 2.36 by the 5 μm air gap originates mainly from the increase of the reflection losses (k_{T} drops from 1.2 to 0.55). The enhancement due to the enlarged collection angle, k_{Ω}, is not sensitive to the presence of the air gap.

In summary, we prove the tolerance of the SIL-enhanced nano-PL system to an air gap of several micrometers. In a typical measurement, an air gap of about 1 μm exists between the SIL and the sample surface after a regular cleaning procedure. The enhancement factor of the collection efficiency is then about 5 in our experiments, as mentioned above. In principle, by increasing the $\mathrm{NA}_{\mathrm{obj}}$ or n_{SIL}, or using an s-SIL, one can further improve the resolution of a SIL-enhanced nano-PL system. But, if the $\mathrm{NA}_{\mathrm{eff}}$ is increased beyond 1, the near-field regime is reached, and the tolerance to the air gap drops seriously. In this sense, our choice of $\mathrm{NA}_{\mathrm{eff}} = 0.864$ is a good compromise between the enhancement of spatial resolution and collection efficiency as well as the feasibility in practical operations. Still, it is worth noting that the described configuration can be used even for samples with rough surfaces with fluctuations of several micrometers.

4.5 Application of Nano-PL: Single-Dot Spectroscopy

The investigation of the properties of individual quantum dots requires single-dot spectroscopy. This can be achieved by reducing the number of dots (selected e.g. by a nano-aperture or a mesa) or high spatial resolution (e.g. SNOM). In SIL-enhanced nano-PL, single dot spectroscopy can also be achieved for samples with a low dot density. In such a system, the choice of dots is more flexible. One can address a large number of individual dots and thus judge whether the observed single-dot properties are typical for the whole ensemble. Since there is no patterning required, it is a non-destructive method.

The high spatial and spectral resolutions of the SIL-enhanced nano-PL system enable us to detect isolated narrow lines from single quantum dots undistur-

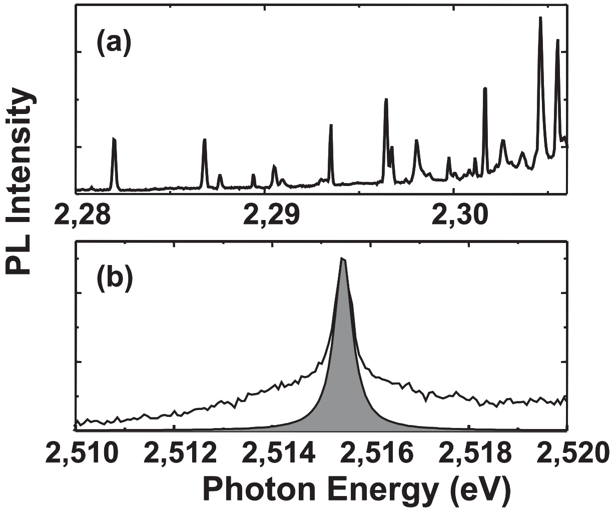


Fig. 10. Single dot spectroscopy achieved by a SIL-enhanced nano-PL set-up. **(a)**Low-energy side of the spectrum of the ZnCdSe/ZnSe quantum-dot sample in Fig. 7C. **(b)** Details of an individual sharp line of a ZnCdSe/ZnSe quantum-dot sample measured at a sample temperature of 60 K. The shaded area shows a Lorentzian fit to the central part of the peak

bed by the luminescence from other dots. The high collection efficiency makes it an ideal system for weak-signal detection. For example, under low-density excitation the spectrum of sharp lines can still be measured with a reasonable integration time at high temperatures up to 120 K. In Fig. 3c we have shown the spectrum of the type-I centers in a type-II GaAs/AlAs [11]. A He–Ne laser is used for excitation with an intensity of 0.5 W/cm^2. By introducing the SIL and the 20 μm pinhole, a large number of isolated sharp lines from individual localization centers are well resolved. This allows us to measure accurately the temperature dependence of the homogeneous line width, thus to study the exciton–phonon interactions in this kind of structures [32].

With this set-up, isolated sharp lines can also be observed for ZnCdSe/ZnSe quantum-dot samples with a rather high dot density. In Fig. 10a, we plot a small part of the spectrum C in Fig. 7 for a closer look at the well-separated lines. The spectral line shape of the individual lines can be studied with high spectral resolution. Increasing the sample temperature we observe a strong deviation from the Lorentzian line shape. Figure 10b shows one example of such a line shape measured at 60 K. The observed deviation is consistent with previous results obtained on similar structures, and can be attributed to the strong coupling regime of excitonic states to acoustic phonons [33]. We note that with this set-up, we are able to study a rather large number of the lines simultaneously. We do find that this kind of deviation is a general feature of all measured lines [34].

Furthermore, we confirm that the polarization information of the luminescence, which is of crucial importance, e.g. in the investigations of spin dynamics,

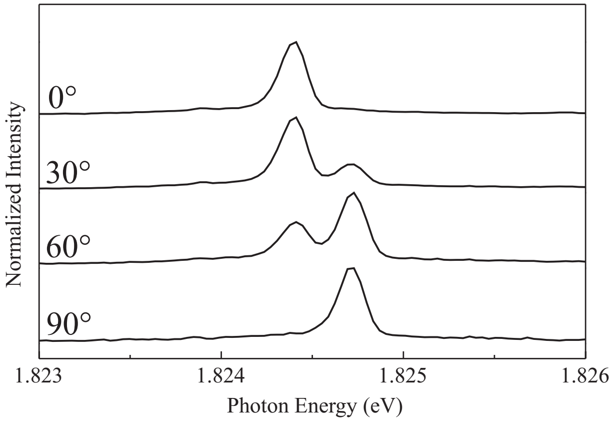


Fig. 11. SIL-enhanced nano-PL spectra of a single ZnCdTe quantum dot embedded in a ZnTe matrix as function of the angle of a linear polarizer in the detection path [9]

is preserved in this set-up. Figure 11 shows the spectra of a single quantum island in a quantum well of 6.5 monolayers ZnCdTe embedded in ZnTe barriers. The sample is excited with the 488 nm line from an Ar-ion laser. To determine the polarization of luminescence, a linear polarizer is placed in the detection path. The spectra show a doublet structure composed of two lines which are linearly polarized along two orthogonal directions. Such line doublets are ascribed to fine-structure splitting of excitons in asymmetric quantum dots [35]. It is typically very difficult to extract the polarization information from other spatially resolving techniques like near-field spectroscopy [36]. Our measurement demonstrates that a SIL can be applied to nano-PL when the polarization of the light is of interest.

4.6 Application of Nano-PL: Spatio-temporal Dynamics of Excitons in Quantum Wells

Lateral transport of excitons is an important aspect of exciton dynamics in quantum wells (QWs). Due to the continuing miniaturization of electronic and optical devices and thus the increasing importance of nanostructures, this transport has to be understood on a length scale comparable to the optical wavelength. It turns out that excitonic transport on such a length scale is far from the classical behavior [37–42]. The SIL-enhanced nano-PL can be used to investigate the transport behavior in a rather direct way with sub-μm resolution. By scanning the pinhole in the image plane of the objective, one can detect luminescence from positions which are different from the position where the sample is locally excited. This enables one e.g. to get the spatial profile of the luminescence intensity which is related to the spatial distributions of the exciton density. The field of view of the SIL [31] (35 μm in our set-up) is far beyond the typical transport length of excitons.

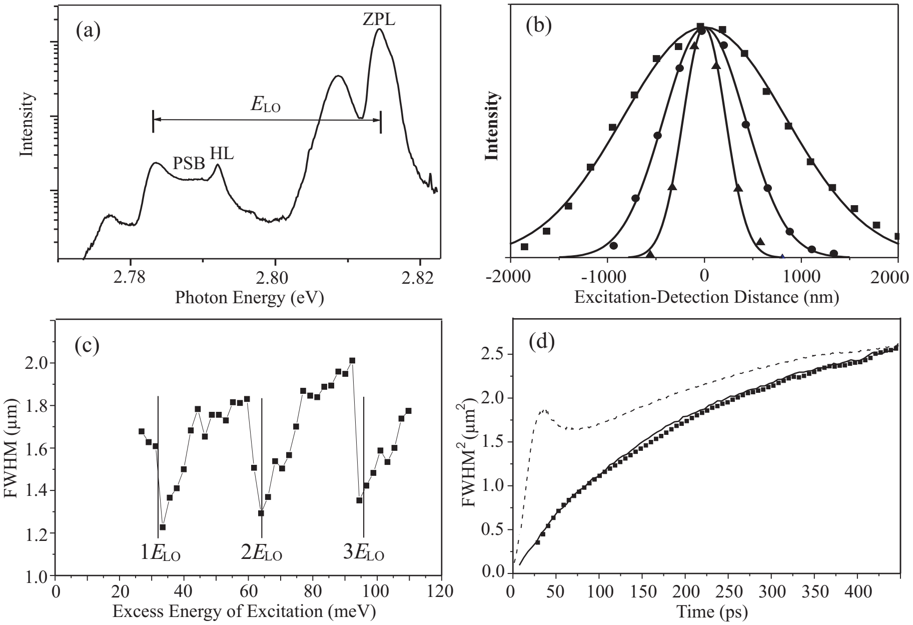


Fig. 12. **(a)** A PL spectrum of a ZnSe QW measured at 7 K; **(b)** Spatial profiles of the ZPL intensity (squares), of the laser spot (triangles), and of the HL peak (circles); **(c)** FWHM of the spatial distribution of ZPL intensity as function of the excitation excess energy in cw experiments; **(d)** Temporal evolution of the squared FWHM of the ZPL spatial distribution measured under pulsed excitation (squares), simulated by a Monte Carlo method (solid line) and the simulated expansion of the total exciton population including hot excitons (dashed line) [41]

We summarize in Fig. 12 some of the most important results found for the transport of hot excitons in ZnSe QWs. These excitons can be generated via the ultra-fast emission of LO phonons with well defined excess kinetic energy. This generation is followed by an energy relaxation due to emission of acoustic phonons which continues over several hundred picoseconds. The relaxed excitons couple to photons directly, resulting in a zero-phonon line (ZPL) in the PL spectrum, as shown in Fig. 12a. In cw experiments, we fix the excitation laser spot and scan the detection spot with respect to the former spot to measure the spatial distribution of the ZPL intensity. An example of the measured profiles is shown in Fig. 12b by the squares. The profile of the laser spot is given in the same panel by the triangles for comparison. By Gaussian fits (solid lines in Fig. 12b), we obtain the full width at half maximum (FWHM) of the distribution, which reflects the transport length of the excitons. The measured FWHM is plotted in Fig. 12c as a function of excitation excess energy. We find a pronounced periodic quenching of the transport length with a period equal to the LO-phonon energy. This effect is similar to the LO-phonon cascade typically observed in PL excitation spectra. These cascades result from the fact that the excitonic formation and relaxation processes assisted by LO phonons are much faster than the acoustic-phonon scattering. The periodic feature reveals the importance of the exciton kinetic energy in the transport process and is not concurrent with a diffusive transport [38].

Time-resolved measurements provide additional insight into the dynamics of these processes. In this kind of experiments, a laser pulse is used to generate excitons locally. In contrast to non-linear optical experiments with high spatial resolution, these studies can be performed in the low excitation regime where many-body interactions are negligible. The time-resolved spectra of the ZPL are measured at different positions with respect to the excitation spot. From these spectra, we get the temporal evolution of the spatial distribution of the ZPL intensity. The squares in Fig. 12d show the squared FWHM of this distribution as a function of time. Anticipating exciton diffusion, one expects a linear increase of the squared FWHM. The observed sub-linear expansion shows again that the diffusion model is not valid for the description of the exciton transport [40].

In order to model the exciton dynamics, we apply a Monte Carlo simulation. The processes considered in the simulation include laser excitation, exciton formation, relaxation, transport and recombination. These processes are controlled by acoustic-phonon and interface-roughness scattering. Using the correlation length of the interface roughness as the only fitting parameter, the simulation (solid curve in Fig. 12d) reproduces the experimental result perfectly.

These simulations also confirm that the spatial profile of the ZPL does not directly reflect the spatial distribution of the excitons [40]. Since the photon momentum is negligible, only cold excitons with small momentum can couple to a photon. Thus, the ZPL only monitors the presence of cold excitons while hot excitons are not visible. A large portion of the exciton ensemble, however, populates high-momentum (i.e. dark) states. While energy relaxation continues, the spatial distribution of the total exciton density including hot excitons can be significantly different from the measured ZPL intensity distribution. The temporal evolution of the total exciton density obtained in our simulation is shown by the dashed line in Fig. 12d. The difference is clearly visible up to 400 ps. In Fig. 12d we also see a striking peak around 30 ps, indicating a breathing-like transport of hot excitons. The excitons start with a quasi-ballistic motion away from their excitation sides. Due to the fact, that acoustic-phonon emission most likely results in a backward scattering, the first emission reverses the exciton motion towards the spot center. After several scattering events the density distribution expands in a nearly diffusive fashion.

Such a spatial oscillation of the exciton distribution can be monitored directly by the ZPL when the excitation is quasi-resonant. From these oscillations one can simultaneously deduce the coherence length and time of the exciton transport [42]. Finally, analysis of the phonon-side band emission (PSB in Fig. 12a) with cw or time-resolved nano-PL yields the dynamics of the energy relaxation during the excitonic transport [39]. The spatial dependence ot the intensity of the hot exciton PL line (HL in Fig. 12a) is a direct measure of the exciton coherence length [37].

Acknowledgments

These investigations on nano-PL were performed by my coworkers Hui Zhao, D. Lüerßen, B. Dal Don, S. Wachter, S. Moehl, R. Bleher, and G. Schwartz. We gratefully acknowledge the growth of excellent samples by the groups of M. Heuken (Aachen), D. Hommel (Bremen), C. Klingshirn (Karlsruhe), K. Ploog (Berlin), and S. Tatarenko (Grenoble). We also acknowledge the support by the Deutsche Forschungsgemeinschaft.

References

1. D. Bimberg, M. Grundmann, N.N. Ledentsov: *Quantum Dot Heterostructures* (Wiley, New York 1998)
2. D. Gammon, E.S. Snow, B.V. Shanabrook, D.S. Katzer, D. Park: Science **273**, 87 (1996)
3. S. Kaiser, T. Mensing, L. Worschech, F. Klopf, J.P. Reithmaier, A. Forchel: Appl. Phys. Lett. **81**, 4898 (2002)
4. E. Betzig, P.L. Finn, J.S. Weiner: Appl. Phys. Lett. **60**, 2484 (1992)
5. G. von Freymann, D. Lüerßen, C. Rabenstein, M. Mikolaiczyk, H. Richter, H. Kalt, Th. Schimmel, M. Wegener, K. Ohkawa, D. Hommel: Appl. Phys. Lett. **76**, 203 (2000)
6. U. Neuberth, L. Walter, G. von Freymann, B. Dal Don, H. Kalt, M. Wegener, G. Khitrova, H.M. Gibbs: Appl. Phys. Lett. **80**, 3340 (2002)
7. A. Gustafsson, M.-E. Pistol, L. Montelius, L. Samuelson: J. Appl. Phys. **84**, 1715 (1998)
8. A. Drechsler, M.A. Lieb, C. Debus, A.J. Meixner, G. Tarrach: Optics Express **9**, 637 (2001)
9. S. Moehl, H. Zhao, B. Dal Don, S. Wachter, H. Kalt: J. Appl. Phys. **93**, 6255 (2003)
10. D. Lüerßen, R. Bleher, H. Richter, Th. Schimmel, H. Kalt, A. Rosenauer, D. Litvinov, A. Kamilli, D. Gerthsen, K. Ohkawa, B. Jobst, D. Hommel: Appl. Phys. Lett. **75**, 3944 (1999)
11. D. Lüerßen, R. Bleher, H. Kalt: Phys. Rev. B **61**, 15812 (2000)
12. Heinz Kalt: 'Optical properties of III-V semiconductors: the influence of multivalley bandstructures'. In: *Springer Series in Solid State Science* **120**, ed. by H.-J. Queisser (Springer, Berlin 1995)
13. S.M. Mansfield, G.S. Kino: Appl. Phys. Lett. **57**, 2615 (1990)
14. B.D. Terris, H.J. Mamin, D. Rugar, W.R. Studenmund, G.S. Kino: Appl. Phys. Lett. **65**, 388 (1994)
15. M. Vollmer, H. Giessen, W. Stolz, W.W. Rühle, L. Ghislain, V. Elings: Appl. Phys. Lett. **74**, 1791 (1999).
16. M. Vollmer, H. Giessen, W. Stolz, W.W. Rühle, A. Knorr, S.W. Koch, L. Ghislain, V. Elings: J. Microscopy **194**, 523 (1999).
17. T. Sasaki, M. Baba, M. Yoshita, H. Akiyama: Jpn. J. Appl. Phys. Part 2 **36**, L962 (1997)
18. M. Yoshita, T. Sasaki, M. Baba, H. Akiyama: Appl. Phys. Lett. **73**, 635 (1998)
19. M. Born, E. Wolf: *Principles of Optics* (Pergamon, Oxford 1970)
20. M. Yoshita, M. Baba, S. Koshiba, H. Sakaki, H. Akiyama: Appl. Phys. Lett. **73**, 2965 (1998)

21. Q. Wu, R.D. Grober, D. Gammon, D.S. Katzer: Phys. Rev. Lett. **83**, 2652 (1999)
22. Q. Wu, R.D. Grober, D. Gammon, D.S. Katzer: Phys. Stat. Sol. B **221**, 505 (2000)
23. B. Richards, E. Wolf: Proc. R. Soc. London A **253**, 358 (1959)
24. R.H. Webb: Rep. Prog. Phys. **59**, 427 (1996)
25. K. Koyama, M. Yoshita, M. Baba, T. Tohru, H. Akiyama: Appl. Phys. Lett. **75**, 1667 (1999)
26. M. Yoshita, K. Koyama, M. Baba, H. Akiyama: J. Appl. Phys. **92**, 862 (2002)
27. M. Yoshita, Y. Hayamizu, K. Koyama, M. Baba, H. Akiyama: Jpn. J. Appl. Phys. **41**, L858 (2002)
28. V. Zwiller, G. Björk: J. Appl. Phys. **92**, 660 (2002)
29. see e.g. www.terastor.com
30. M. Baba, T. Sasaki, M. Yoshita, H. Akiyama: J. Appl. Phys. **85**, 6923 (1999)
31. G.S. Kino. In: *Optical pulse and beam propagation, Vol.* **3609** *of* Proceedings of the SPIE (SPIE, Washington 1999) p.56
32. H. Zhao, S. Wachter, H. Kalt: Phys. Rev. B **66**, 085337 (2002)
33. L. Besombes, K. Kheng, L. Marsal, H. Mariette: Phys. Rev. B **63**, 144307 (2001)
34. B. Dal Don, H. Zhao, S. Moehl, C. Ziegler, H. Kalt: Phys. Stat. Sol. (c) **0**, 1237 (2003)
35. L. Besombes, L. Marsal, K. Kheng, T. Charvolin, L.S. Dang, A. Wasiela, H. Mariette: J. Cryst. Growth **214/215**, 742 (2000)
36. G. Eggers, A. Rosenberger, N. Held, G. Güntherodt, P. Fumagalli: Appl. Phys. Lett. **79**, 3929 (2001)
37. H. Zhao, S. Moehl, H. Kalt: Phys. Rev. Lett. **89**, 097401 (2002)
38. H. Zhao, S. Moehl, S. Wachter, H. Kalt: Appl. Phys. Lett. **80**, 1391 (2002)
39. H. Zhao, S. Moehl, H. Kalt: Appl. Phys. Lett. **81**, 2794 (2002)
40. H. Zhao, B. Dal Don, S. Moehl, H. Kalt, K. Ohkawa, D. Hommel: Phys. Rev. B **67**, 035306 (2003)
41. H. Zhao, B. Dal Don, S. Moehl, H. Kalt: phys. stat. solidi (b) **238**, 529 (2003)
42. B. Dal Don, H. Zhao, G. Schwartz, H. Kalt. In: *Proc. of Optics of Excitons in Confined Systems (OECS 8), Lecce 2003*, in press

Spectral Trimming of Photonic Crystals

Markus Schmidt[1], Gunnar Böttger[1], Manfred Eich[1], Uwe Hübner[2],
Wolfgang Morgenroth[2], and Hans-Georg Meyer[2]

[1] Technische Universität Hamburg-Harburg, Eißendorfer Straße 38, 21073 Hamburg, Germany

[2] Institut für Physikalische Hochtechnologie e.V., Abt. Kryoelektronik,
A.-Einstein-Str. 9, 07745 Jena, Germany

Abstract. We present a novel concept to trim the transmission properties of finite two dimensional photonic crystal slab waveguide structures by UV photobleaching. Systematic fabrication inaccuracies may be compensated due to the shift of the spectral properties during the bleaching process. To prove our concept experimentally, we measured the transmission of UV sensitive photonic crystal structures for different doses. A shift of band edges and defect resonance peaks depending on UV dose is observed due to changes in refractive index and geometry.

1 Introduction

Photonic crystals (PCs) are a current research topic in applied physics since they offer a significant potential in the area of ultracompact integrated optics devices. PCs apply the concept of using periodic dielectric functions in space in more than one dimension and were introduced by Yablonovitch [1] and John [2]. Dielectric periodicities lead to bandstructures of the same dimensionality, comparable to the opening of band gaps for electrons due to periodic atomic potentials. In both cases the wave functions are represented by Bloch functions with wavelengths on the scale of the periodicity. If the dielectric lattice geometry is chosen appropriately, a frequency range opens in the bandstructure blocking light of respective wavelengths [3,4]. This range is called photonic band gap (PBG) and is one of the most important characteristics of PCs because it allows to manipulate the propagation of light within a few lattice constants, on a much smaller scale than conventional gratings. The higher the dielectric contrast, the wider the PBGs. By doping PCs with isolated defects it is possible to create allowed states inside the PBG, resulting in high-Q cavities [5–7]. Also sharp bends [8–10] using connected defects in PC waveguide structures, and superprisms [11–13] based on the highly anisotropic dispersion surface of a PC can be realized.

For integrated optics finite two dimensional (2d) PCs play a key role because planar slab structures may fairly easily be structured using current fabrication methods. The in-plane propagation of light is governed by its interaction with a 2d lattice of air holes etched through the slab. Vertical confinement is facilitated by total internal reflection. Slab waveguide cores thus are required to have a higher refractive index than substrate and cladding, which is an important issue also discussed in this article. Defects resulting in allowed states inside the band gap are for example realized by leaving out air holes in the otherwise regular

M. Schmidt et al., Spectral Trimming of Photonic Crystals, Lect. Notes Phys. **658**, 71–86 (2005)
http://www.springerlink.com/

PC. For stability reasons and ease of fabrication planar PCs should have non-air substrates and hence are in most cases vertically asymmetric. Choosing an air hole lattice the functionality of the PC is limited to one polarization, nevertheless showing an in-plane, direction independent stop gap for triangular lattices [14–17].

Most 2d PCs realized today are based on a semiconductor core material (GaAs, InP or Si) in order to achieve a high horizontal refractive index contrast and thus a wide bandgap [18]. Waveguide cores for single mode operation are around $0.25\,\mu m$ thick, resulting in a strong coupling mismatch between standard optical single mode fibers and 2d slab PCs. Another approach is using materials with moderate refractive indices in the range of n=1.54 to n=2.30, increasing waveguide core thicknesses to typically between $0.5\,\mu m$ and $1.5\,\mu m$, in this way reducing mode mismatch and Fresnel reflection on coupling interfaces. Polymers and inorganic glasses have indices in this region and show low intrinsic slab waveguide losses [19–22]. For reasons of longer material wavelengths statistical fabrication errors in the PC lattice are less critical in moderate refractive index materials than in high index ones. Using chromophores in the material matrix, one may also to some extent alter the refractive index in irreversible and reversible ways, directly shifting the spectral properties of a PC device.

In the first part of this chapter we study the influence of the air hole etching depths on overall transmission characteristics of planar 2d PC. Etching air holes into the core material reduces the effective refractive index at the same time, decreasing the vertical contrast required for guiding by total internal reflection. It is shown that waveguiding conditions may be improved by etching into the material substrate to sufficient depths. Finite difference time domain (FDTD) simulations and experimental data prove that the optical transmission properties of a moderate refractive index PC are indeed dramatically enhanced by air holes reaching sufficiently far into the substrate [23,24].

PCs are very sensitive to deviations from the ideal design, especially in lattice constant and air hole radius. Section 3 introduces a novel concept of photobleaching chromophores in PCs to compensate systematic fabrication errors and furthermore shift optical properties to desired wavelengths used in dense wavelength division multiplexing (DWDM). By photobleaching one can irreversibly adjust the refractive index of the polymeric waveguide core, and consequently trim the spectral transmission properties of the PC. The concept is proved both in FDTD simulations and experimental measurements. Section 5 concludes this contribution.

2 Importance of Etching Depths in 2D Planar PCs

The slab waveguide considered in this section consists of a $1.5\,\mu m$ thick polymer slab waveguide core with a refractive index of $n = 1.54\,@\,1.3\,\mu m$ on a Teflon substrate with $n = 1.30\,@\,1.3\,\mu m$. To scale the PC stop gaps to near infrared wavelengths, a lattice constant a of $500\,nm$ and a hole radius r of $150\,nm$ are chosen. The three innermost lines of air holes are omitted, forming a 2d PC

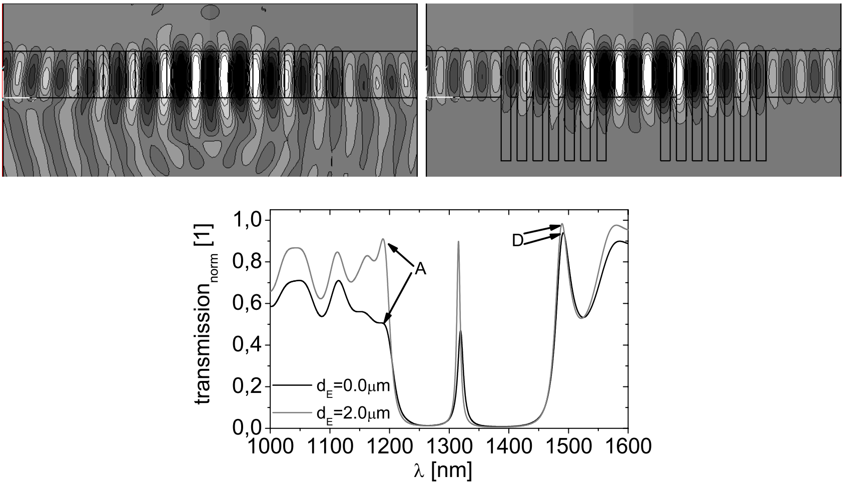


Fig. 1. Simulations of the electric field distribution inside a finite 7_3_7 2d PC line defect resonator (TE-polarization). $n_{\mathrm{CORE}} = 1.54$, $n_{\mathrm{SUBSTRATE}} = 1.30$, $r = 150\,nm$, $a = 500\,nm$, $d_{\mathrm{SLAB}} = 1.5\,\mu m$. Left: Holes not etched into the substrate ($d_{\mathrm{ETCH}} = 0.0\,\mu m$). Substantial radiation losses to the substrate are observed. Right: Sufficient etching depth ($d_{\mathrm{ETCH}} = 2.0\,\mu m$), reaching nearly optimal confinement and transmission. Bottom: Transmission spectra of both resonator structures as function of vacuum wavelength (A: air band edge, D: Dielectric band edge) in TE polarization

line defect resonator, introducing a defect state inside the stop gap. The PC consists of a square lattice of air holes with a total length of 16 lattice constants, equivalent to $8\,\mu m$ only. The notation n_m_n used in this article refers to a resonator which consists of a defect of width $m \cdot a$ embedded in a hole lattice of width $n \cdot a$ from both sides. Figure 1 gives side views of the simulated PCs, Fig. 2 shows a scanning electron micrograph (SEM) picture. A square lattice of air holes in theory does not show a direction independent stop gap. The 2d line defect resonator, however, is just driven in one direction, perpendicular to the geometric line defect. Furthermore, stop gaps appear for both TE and TM polarization. Such resonator structures may be used as optical filters because of a narrow transmission bandwidth at the resonance wavelength, reaching high quality factors within very compact devices.

2.1 Simulation of Varied Etching Depths

To analyze the influence of etching depths we performed three dimensional (3d) finite integration simulations (an FDTD scheme as implemented in CST Microwave Studio, [25]). The structure is excited with the fundamental slab waveguide mode from the left and the transmitted signal is recorded at the right output port (Fig. 1).

The resonance peak observed inside the stop gap is directly related to the defect geometry of the 2d PC lattice. The transmission at the resonance wavelength

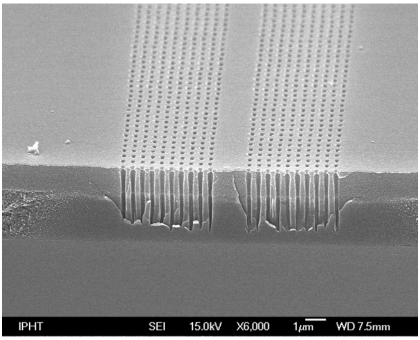


Fig. 2. SEM picture displaying the cleaved edge of a 10_3_10 2d PC line defect resonator made of P(MMA/DR-1)-Teflon. The holes of the square lattice were etched through the waveguide core layer into the Teflon substrate. This structure is excited from the left and the transmission is measured on the right side of the 2d PC

is very high for the sufficiently etched system ($d_{\mathrm{ETCH}} = 2.0\,\mu m$, $T = 90.0\%$) compared with the slab waveguide where just the core material is perforated by the air holes ($d_{\mathrm{ETCH}} = 0.0\,\mu m$, $T = 47.1\%$). Strong radiation losses are clearly visible for the insufficiently etched 2d PC line defect resonator due to a lack of total internal reflection at the core/substrate boundary. In the sufficiently etched resonator, the total internal reflection condition is recovered by decreasing the average refractive index of the substrate as well. At shorter wavelengths to the left of the stop gap electrical fields are concentrated inside the air holes, this part of the spectrum is called the air band edge. It is much more sensitive to radiation losses compared to the side with longer wavelengths (dielectric band edge), where the electric field is concentrated in the dielectric material. For the sufficiently etched 2d PCs radiation losses are minimized, resulting in a high transmission namely at the air band but to some extent also at the dielectric band edge.

2.2 Experimental Characterization of PCs with Different Etching Depths

To investigate the prerequisite of sufficient etching depths we fabricated a square lattice of air holes in a polymeric slab waveguide [26]. The substrate of the slab waveguides consists of a low refractive index polymer Teflon AF (DuPont, refractive index $1.30\,@\,1300\,nm$) with a thickness of about $2\,\mu m$ coated on a 3 inch oxidized silicon wafer. As slab waveguide core material we use a side chain polymethylmethacrylate (PMMA) polymer covalently functionalized with 10 mol % of nonlinear dye molecules Disperse Red 1 (DR-1).

This material has a refractive index of $n = 1.54\,@\,1300nm$ and exhibits low optical losses ($1dB/cm\,@\,1.3\,\mu m$). In order to achieve single mode operation in the 2d PC at 1300 nm excitation wavelength a core layer thickness of $1.45\,\mu m$ is chosen. The 2d PC structure is fabricated by electron beam lithography and reactive ion etching (RIE) employing a standard electron beam lithography resist

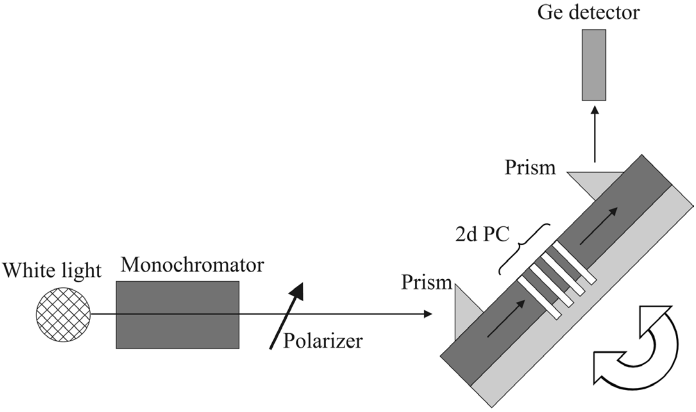


Fig. 3. Experimental setup to measure the transmission properties of finite 2d PC slab waveguide structures

and a NiCr - metal mask for the etching process. The processing steps are as follows: After the hole array is written into the resist and the mask is opened by ion etching with Argon, the PC structure is transferred into the waveguide core by reactive ion etching using a electron - cyclotron - resonance (ECR) plasma source with a mixture of O_2 and argon. As a final step, the remaining metal mask is removed by wet etching. The realized 2d PC structure corresponds to the simulated 2d PC described above (Fig. 1). Figure 2 shows a scanning electron micrograph (SEM) image of an example of a finite 10_3_10 2d PC in the P(MMA/DR-1)-Teflon system with an etching depth of $1.8\,\mu m$ to $2.0\,\mu m$ into the substrate.

For the optical characterization of our 2D PC waveguides we use an arrangement of a white light source (Oriel, 1000 W), a monochromator (1/4 m, Yvon Jobin), a polarizer, and a waveguide prism coupler setup. The PC structure is positioned between two prisms to couple light into and out of the waveguide. The wavelength dependent optimal coupling angle is adjusted by rotating the incoupling prism and the sample holder for optimum coupling. The optical signal is detected by a Ge photodiode and recorded by a lock in amplifier. The experimental setup (Fig. 3) allows mode selective polarization and wavelength dependent transmission measurements in the near infrared regime.

To investigate the influence of the etching depth on the transmission properties of the 2d PC we compare the spectra of a deeply etched 7_3_7 resonator made of P(MMA/DR-1)-Teflon ($d_{\mathrm{ETCH}} = 2.0\,\mu m$) with the spectra of a shallowly etched 7_3_7 BCB-Teflon (bencocyclobutene polymer [20], $d_{\mathrm{ETCH}} = 0.2\,\mu m$) resonator. Both polymers have refractive indices of $n = 1.54$ in the near infrared regime hence behave identically as PC waveguide slab materials. Figure 4 shows the transmission spectra of the two PC resonators.

Again the spectrum of the sufficiently etched 2d PC on the right shows a much better transmission at the defect wavelength as well as at the air and

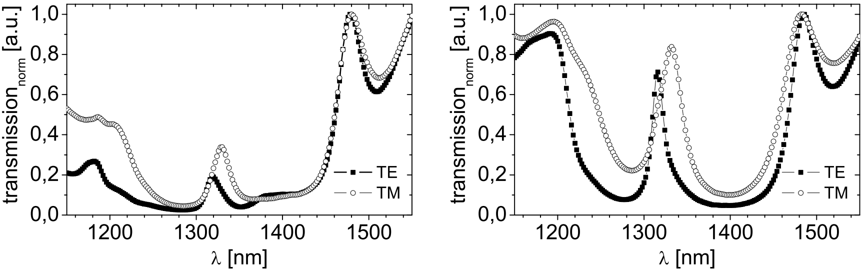


Fig. 4. Transmission of 7_3_7 2d PC line defect resonators for TE$_0$-like and TM$_0$-like polarizations. Left: 7_3_7 BCB resonator, etching depth $200\,nm$. Right: P(MMA/DR-1) resonator, etching depth $2\,\mu m$

dielectric band edges. Especially the transmission on the air band side benefits from the large etching depth compared with the curves in the left diagram.

To conclude this section, we showed both experimentally and theoretically that finite 2d PC waveguides based on core and cladding dielectrics of comparatively small vertical refractive index contrast show good transmission performance only for high aspect ratio air holes. To investigate the influence of the etching depth experimentally, we realized finite 2d PC resonators of different low loss optical polymers. These structures show high transmission at the resonance and at the air and dielectric band edges only when the holes are deeply etched into the substrate, significantly reducing radiation losses both on the defect wavelength and at the air band edge.

3 Trimming of 2D PC Optical Spectra by UV Photobleaching

For applications in integrated optics and wavelength division multiplexing (WDM) PC structures like high Q cavities are of great interest for spectral channel separation in the order of $0.8\,nm$ $(100\,GHz)$.

Fine tuning of working frequencies is also an important issue in quantum optical applications because defect frequencies or band edges have to match atomic transition energies. Therefore new concepts have to be developed to compensate for systematic fabrication deviations.

In this section, we present a new concept for trimming spectral transmission properties of 2d PCs after fabrication. We experimentally and theoretically show that an irreversible reduction of the refractive index and layer thickness of the waveguide core is achievable. Chromophores in a polymer slab PC may be UV bleached, leading to such an irreversible shift in the spectral position of the defect frequency and the dielectric band edge. Our 2d PC devices are made of a polymeric waveguide core as described in Sect. 2.2, the polymer is containing Disperse Red 1 (DR-1, Fig. 5) dye molecules, which strongly absorb UV-photons due to their $\pi\pi^*$ transition.

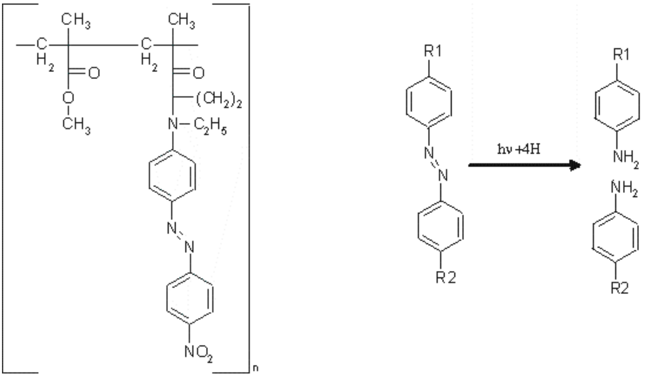


Fig. 5. Left: Polymethylmethacrylate (PMMA) covalently functionalized with the Disperse Red 1 (DR-1) chromophore. The dye molecule is attached to the PMMA polymer backbone. Right: UV photobleaching process of the DR-1 molecule

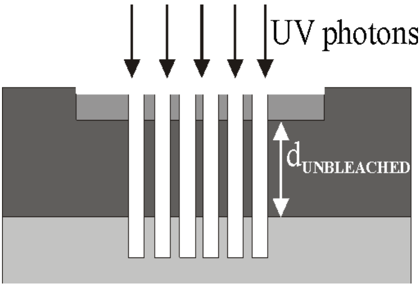


Fig. 6. Principle of photobleaching a finite 2d PC structure. A spatial refractive index distribution results with the shape of a step function. Due to the high optical density of the unbleached material most of the UV photons are absorbed at the surface. Hence, the step function moves vertically through the slab waveguide with increasing illumination time i.e. UV dose. $d_{\mathrm{UNBLEACHED}}$ represents the thickness of the unbleached waveguide

The UV-deposited energy leads to a decomposition of the chromophore into two smaller fragments which no longer absorb into the $\pi\pi^*$ band because of the destruction of the conjugated π-electron system. This effect lowers the linear polarizability as well and typically decreases the refractive index in the visible and near infrared regime by a factor of a few 10^{-2} [27]. The volatile molecule fragments may diffuse out of the polymer matrix leading to an effective reduction of the core film thickness after photobleaching. Figure 6 shows the principle of photobleaching finite 2d PCs.

3.1 3D Simulations of Bleached 2D PCs

In order to estimate the maximally achievable shift we performed 3d finite integration simulations of a bulk 2d PC with a length of 19 lattice constants. The transmission and the spectral position of the first Fabry-Perot (FP) maximum of the dielectric band edge in TE polarization is calculated for different thickn-

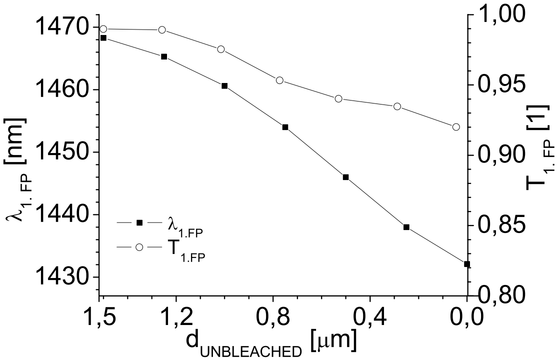


Fig. 7. Calculated spectral position and transmission of the dielectric band edge (first FP maximum) as a function of the unbleached slab thickness $d_{\mathrm{UNBLEACHED}}$ (see Fig. 6) corresponding to different bleaching times i.e. UV doses

esses for the unbleached slab, corresponding to various UV doses i.e. bleaching times (Fig. 7). In this model, we assume a reduction of the refractive index from $n = 1.54$ to $n = 1.51$ which is confirmed by experimental results. A decreased slab waveguide thickness of the bleaching region of 16% is included [27]. Other simulation parameters are as in Sect. 2.1.

We observe a shift to shorter wavelengths because a reduced refractive index and a decreased slab waveguide thickness lead to a lower effective slab waveguide mode index. From simulations we get a maximum shift of the dielectric band edge of 36 nm in TE polarization. The transmission decreases for this large shift by just 8%. When the dielectric band edge is shifted by a wavelength interval equivalent to 10 DWDM channels (0.8 nm each), the transmission is lowered by only 5%. A shift of only one DWDM channel is accompanied by hardly any transmission reduction.

3.2 Experimental Bleaching of Regular 2D PCs

To verify our concept experimentally we fabricated a finite 2d PC slab consisting of PMMA covalently functionalized with 10 mol% DR-1 chromophores as core and Teflon as substrate materials. The fabrication process was described in Sect. 2.2. The 2d PC has 20 air holes along and 8000 holes perpendicular to the direction of propagation ($a = 500\,nm$, $r = 150\,nm$), resulting in an optical stop gap at around 1.3 μm vacuum wavelength. In order to reduce radiation losses into the substrate an etching depth of 2.5 μm into the substrate was chosen (sec. 2).

For photobleaching the polymeric waveguide core material, a 1 kW Xe high pressure lamp (Oriel) with a spectral emission from 220 nm to the IR regime is used. To cut off light with wavelengths shorter than 290 nm, which would destroy the PMMA main chain, a Schott WG 295 filter is inserted into the beam. At the other end of the spectrum, light with wavelengths larger than 700 nm is filtered out by a H_2O filter, reducing the heat transferred to the sample. The light is

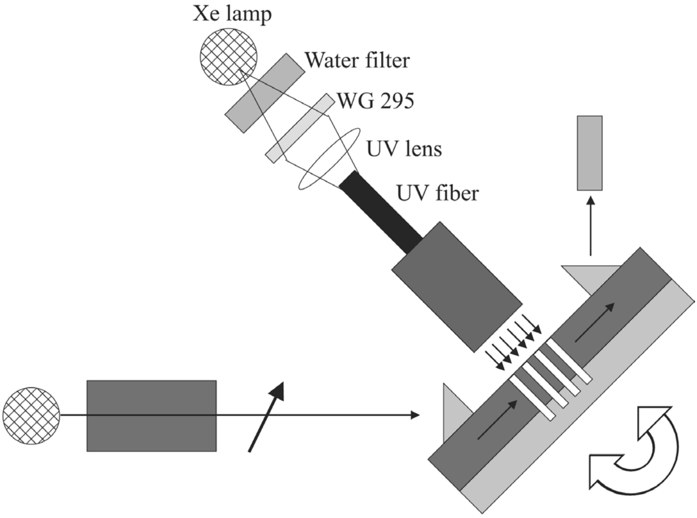


Fig. 8. Experimental setup for photobleaching 2d PC structures and measuring the transmission as a function of wavelength (see Fig. 3)

imaged into a UV fiber (Oriel $d_{\mathrm{CORE}} = 3\,mm$), guiding the UV photons directly onto the 2d PC structure mounted between two prism. The light finally emitted from the fiber has a spectral range of $290\,nm$ to $700\,nm$ and an integral power density of approximately $I = 440\,mW/cm^2$. The combination of this set-up with the prism coupling set-up described in Sect. 2.2 allows in situ photobleaching of the 2d PC slab waveguide (Fig. 8). The samples are not measured directly after the exposure to UV radiation, but rather after 8 hours, to allow the cis-isomers, which are the result of a different absorption into the chromophores, to convert back to the thermodynamically stable trans state. Typical exposure times are 10 minutes only. The bleached spot on the sample has a round shape of $5\,mm$ diameter visible to the naked eye by its faded color. Therefore, the complete 2d PC structure was exposed.

The left image of Fig. 9 shows a cleaved edge of an unbleached bulk 2d PC (without line defect) consisting of 20 layers of air holes perpendicular to the direction of propagation.

There was no further shift of the dielectric band edge after 80 minutes of bleaching, amounting to a total UV dose of $2.1\,kJ/cm^2$.We observe a maximum shift of the dielectric band edge of approximately $35\,nm$ toward shorter wavelengths (Fig. 9 right) which is in agreement with our 3d FDTD simulations. In TM polarization a slightly smaller shift of $27\,nm$ was measured. This difference may be explained by simulations (plane wave calculations, MPB package [28]) where an induced birefringence of $n_{TE} = 1.51$ and $n_{TM} = 1.532$ is assumed at the X-point of the reciprocal square lattice. Chromophores in the waveguide core oriented parallel to the film surface absorb the UV photons better than perpendicular aligned dye molecules due to the mismatch between polarization and

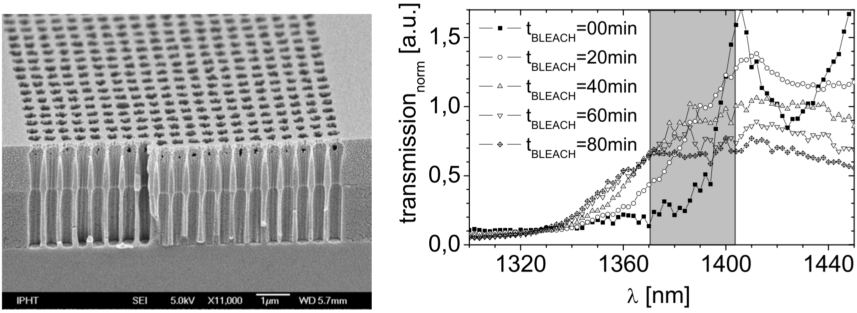


Fig. 9. Left: SEM picture of a cleaved edge of a 2d PC crystal containing 20 layers of air holes used for bleaching experiments. Right: Spectral transmission taken at the dielectric band side of the stop gap, each curve corresponding to a different bleaching time. The shaded region indicates the total shifting of the dielectric band edge up to 80 minutes of photobleaching

transition dipole moment orientations. Therefore, the refractive index change in the TM direction is expected to be smaller [29].

The transmission on the dielectric band edge is reduced for very long bleaching times thus deposited UV doses. This may be due to two reasons: First, the reduced transmission of the dielectric band edge may be related to a stronger induced surface roughness, which is a result of chromophore fragments gassing out of the waveguide core. Second, the SEM picture shows that, due to deficiencies in our nanostructuring process, the air holes are of a cone shape and interconnected at the top of the waveguide. Simulations precisely of this structure show a strong mode mismatch between the fundamental slab waveguide mode and the 2d PC mode when bleaching the core material. This results in back reflection into the slab waveguide and strong radiation losses into the substrate. By optimizing the etching parameters the air holes will have a more cylindrical shape and the transmission is expected to stay at a high level also for longer bleaching times. However, to make spectral adjustments on the scale of a few DWDM channels, only shifts of a few nanometers or even fractions of a nanometer are needed. This requires only a relatively small UV dose or short bleaching times below one minute, not affecting transmission substantially [22].

3.3 Experimental Bleaching of 2D PC Line Defect Resonators

We also performed first experiments on trimming PC resonator structures. A 10_3_10 2d PC line defect resonator consisting of P(MMA/DR-1) as core and Teflon as substrate material is fabricated ($d_{\mathrm{ETCH}} = 2\,\mu m$). The left diagram of Fig. 10 shows the spectral transmission of this resonator structure at the resonance wavelength for different UV doses, i.e. different bleaching times in TM polarization.

For bleaching times up to 60 minutes the defect peak, against physical intuition, moves to longer wavelengths. This can be explained by the following model:

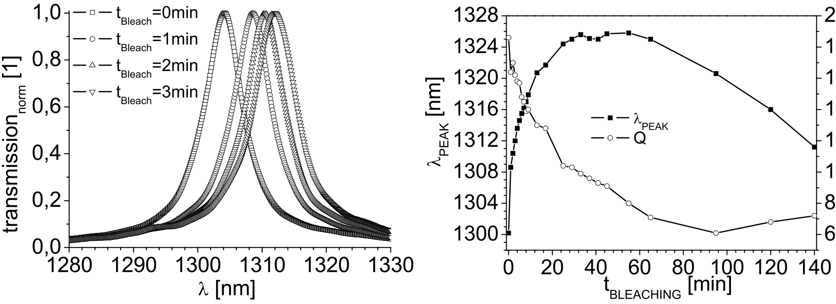


Fig. 10. Left: Normalized transmission of the 10_3_10 2d PC line defect resonator around the resonance wavelength for different bleaching times in TM polarization. Right: Resonance wavelength of the resonator and Q factor of the cavity as function of bleaching time, i.e. deposited UV dose.

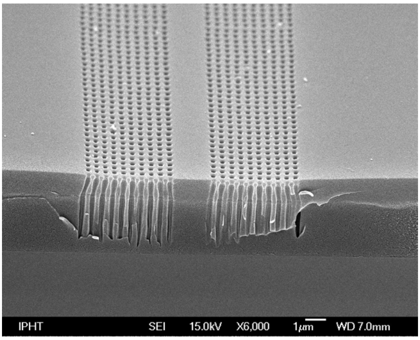


Fig. 11. SEM picture of the 10_3_10 finite 2d PC line defect resonator after a 140 minutes exposure of UV light

In principle, the 2d PC line defect resonator consists of two Bragg mirrors spatially located around the defect. When photobleaching the structure from the top (Fig. 6), UV photons penetrate into the holes, and light is primarily absorbed at the top of the material between the holes but also on the side walls. Due to this penetration the Bragg mirror parts of the 2d PC absorb more UV light than the defect. As shown by Vydra et al. [27], the volume density of the material does not change during the bleaching process. This necessarily results in a shrinking of the Bragg mirrors while photobleaching. Therefore, the defect is expected to effectively elongate (Fig. 11). A careful examination of the center region in the SEM picture of the bleached PC waveguide exactly confirms this mechanism.

These larger defect volumes directly cause a shift of the resonance peak to higher wavelengths as observed in the transmission spectra. For bleaching times up to 60 minutes, the change in geometry dominates the spectral behavior of the structure. After 60 minutes of photobleaching, leading to a shift of $22\,nm$ to longer wavelengths, no further change in geometry is observed. For exposures longer than 60 minutes the reduction of the refractive index from $n = 1.54$ to $n = 1.532$ becomes dominant. The faster geometry change, i.e. the strain of the central defect region, mainly stems from the UV conversion and extraction of molecular fragments from the adjacent perforated regions which are highly accessible for the UV photons due to their large surface. The refractive index of the solid center region, on the other hand, reacts more slowly since this unperforated part of the PC is less open to the UV photon attacks and provides less surface for diffusion process. As shown in Sect. 3.2, the refractive index change leads to a shift to shorter wavelengths related to a reduced effective index of the slab waveguide mode. Only for very high UV doses, the quality factor of the cavity (Q-factor) is substantially reduced during the bleaching process (Fig. 10 right). However, the maximum shift to longer wavelengths is $22\,nm$. To address a few DWDM channels, i.e. to shift the transmission peak a few nanometers, only a very small UV dose or a bleaching time below one minute is needed. Therefore, the initial Q-factor is expected to remain unchanged. The Q-factor mainly depends on the number of lattice constants of the PC regions adjacent to the defect. Maximizing Q therefore requires larger PC structures and will be addressed in a different publication. All results presented here can be applied also to these larger PC structures.

4 Photonic Crystal Waveguides

To use integrated optics components on an all optical chip, concepts have to be developed to connect the different structures. In today's applications in optical circuits, ridge or channel waveguides are widely used whereas different prerequisites have to be fulfilled. The mode sizes of the functional units and of the connecting waveguides have to be matched preferably with low coupling losses below $1\,dB$. Typically, this requirement is fulfilled if the waveguide and the functional unit are of the same order of size. Waveguide transmission losses in the range of $10\,dB/mm$ are possible today in high refractive index systems [30]. Another important factor to increase the number of components on one chip is the realization of sharp waveguide bends with high and broadband transmission connecting the optical elements in two dimensions.

Photonic crystals are one of the most feasible concepts to satisfy the criteria above mentioned. As mentioned in the beginning of the article, photonic crystal structures have dimensions in the order of the operation wavelength, hence fulfilling the requirement of small sizes.

Doping of photonic crystal structures results in allowed states inside the photonic band gap. This kind of defect is typically realized by leaving out one complete line in a bulk PC hole array. Excitation of a defect mode along the direction of the defect results in a propagation waveguide mode which is laterally confined by Bragg reflection. Vertically, the light is guided through the structure by total internal reflection.

4.1 Straight Waveguides

Due to the fact that the photonic crystal waveguide mode which results from the line defect inside the PC contains k vectors (in plane wave base) pointing

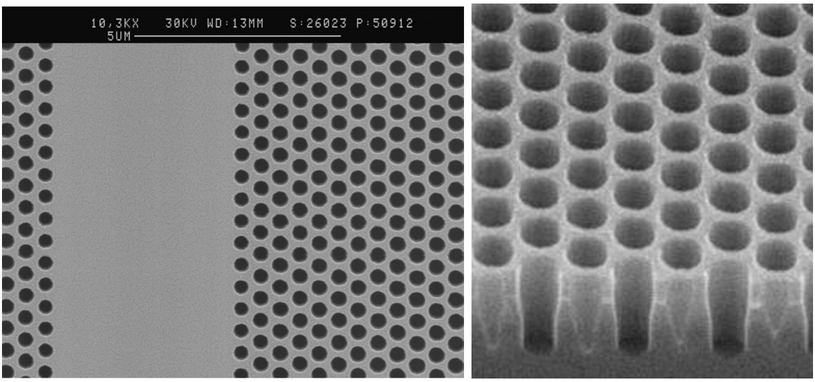


Fig. 12. SEM - pictures of W9 photonic crystal channel waveguide consisting of Nb_2O_5 as core material [31]

to all horizontal directions, a complete in-plane band gap for one polarization is required. Typically, triangular lattices which Bragg reflect light in all directions, if the incoming light has a frequency inside the photonic band gap, are used for waveguiding applications. Therefore, it can be shown theoretically that a refractive index larger than 1.8 is needed to open a complete band gap for the TE polarization.

Augustin et al. [31] have realized straight photonic crystal waveguides made of a triangular hole lattice in moderate refractive index Nb_2O_5 with $n = 2.17 @ 1.5\,\mu m$. This relatively moderate refractive index opens a complete band gap for the TE polarization and keeps the Rayleigh scattering losses relatively low. These losses result from the etching process inducing surface roughness at the side walls of the holes.

Also, the pentoxide material shows very low intrinsic slab waveguide losses below $1\,dB/cm @ 1.3\,\mu m$ operation wavelength. Hole diameters of $370\,nm$ and lattice constants of $595\,nm$ result in a photonic band gap around the operation wavelength of $1.3\,\mu m$. The thickness of the slab waveguide core which was surrounded by upper and lower SiO_2 cladding layers was $500\,nm$. Figure 12 shows the realized structure where 9 lines inside the photonic crystal have been left out (W9 - waveguide).

The cut back method was used to determine the photonic crystal waveguide losses. Therefore, the photonic crystal channel waveguide was successively shortened by cleaving the end of the wafer, resulting in a exponentially increase of the transmission. The slope of the decay in logarithmic scale is correlated with the waveguide losses. Augustin et al. measured $1.7\,dB/mm @ 1.50\,\mu m$ wavelength for the W9 and $8.5\,dB/mm @ 1.50\,\mu m$ for the W3 waveguide which is in the range of the prerequisites to use photonic crystals in telecommunications.

4.2 Photonic Crystal Waveguide Bends

An all optical chip design requires a two dimensional interconnection between the different optical components. Due to the fact that the size of the components is in the 10s of micrometer regime, small waveguide bends with relatively low optical losses are required to reach a high integration density. In today's telecommunication technology, ridge waveguide bends are commonly used with bending radii of about $10\,\mu m$ which is still large compared to the component dimension. Modification of the bend may result in an improved transmission [32].

Photonic crystal waveguides consisting of a triangular lattice provide the possibility of 60 degree bends. This kind of photonic crystal bend is compatible with the lattice of the photonic crystal. Different bend geometries in various material systems have been manufactured to evaluate the bending losses [33,34].

In a first order approximation, the bend can be regarded as a cavity. The modes of this cavity have lossy k - vector components in the vertical direction, resulting in radiation losses which decrease the bend performance. Therefore, lattice tuning and impedance matching concepts at the location of the bend are promising options to keep the vertical losses small. Most of the concepts are based on the fact that by introducing holes inside the bend, the effective mode volume decreases. This results in a smaller number of modes in the bend decreasing the radiation losses because lossy higher order modes have been cancelled out [35]. Boscolo et al. showed that $0.06\,dB$ at a Y- junction are theoretically achievable in a $n = 3.4$ material system [10].

Augustin et al. have realized photonic crystal channel waveguides with two double bends using Nb_2O_5 as core material which operate in the telecommunication wavelength regime at $1.50\,\mu m$ (Fig. 13). To couple light into the photonic crystal structure, ridge waveguides have been connected to the photonic crystal structure. By referencing the transmission of the double bend to a single photonic crystal waveguide, a low loss of $1.2\,dB/bend$ was achieved if the bend had been optimized by adding holes into the bend [36].

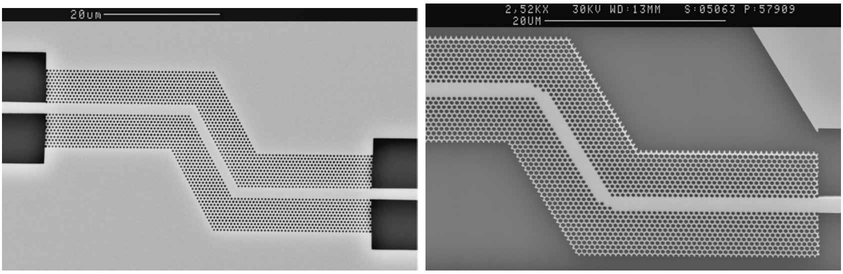


Fig. 13. SEM pictures of double 60° photonic crystal channel waveguide bends consisting of Nb_2O_5 as core material. Left: non-modified double bend; right: optimized bend by additional holes inside the waveguide [36]

5 Conclusion

In this article, we showed 2d PC structures from moderate refractive index materials operated in the near infrared. Square lattices of holes with a lattice constant of 500 *nm* and a radius of 150 *nm* were deeply etched into polymeric slab waveguides. High aspect ratios (>10) were shown to restore improving total internal reflection at the core/substrate interfaces. As a consequence, high transmission at the air band edge and at resonance wavelengths is theoretically expected and experimentally observed.

To compensate for systematic fabrication deviations, a novel concept to trim PC structures by UV photobleaching was introduced. Simulations show a maximal irreversible shift of the band edges of 35 *nm* in a bulk 2d PC, which was verified experimentally. In a second step, a finite 2d PC line defect resonator was photobleached resulting in a maximum shift of 22 *nm* to longer wavelengths for the first 60 minutes of the UV exposure followed by a reduction of 15 *nm* within the last 80 minutes. The initial increase can be explained by a strain from the photosensitive polymer P(MMA-DR-1) in the defect geometry of the resonator. At longer illumination times, the change in refractive index dominates and the resonance peak shifts to shorter wavelengths. In order to trim on nanometer scales, corresponding to one or a few DWDM channels, only very small UV doses or bleaching times below one minute are necessary. We think that this concept may be applied to level out systematic fabrication inaccuracies in PC structures.

Acknowledgments

We thank CST, Darmstadt for their support with CST Microwave software products and the Federal Ministry of Education and Research (BMBF) and the Deutsche Forschungsgemeinschaft (DFG) for their financial support.

References

1. E. Yablonovitch: Phys. Rev. Lett. **58**, 2029 (1987)
2. S. John: Phys. Rev. Lett. **58**, 2486 (1987)
3. K.M. Ho, C.T. Chan, C.M. Soukoulis: Phys. Rev. Lett. **65**, 3152 (1990)
4. R.D. Meade, A.M. Rappe, K.D. Brommer, J.D. Joannopoulos: J. Opt. Soc. Am. B **10**, 328 (1993)
5. E. Miyai, K. Sakoda: Opt. Lett. **26**, 740 (2001)
6. Y. Akahane, M. Mochizuki, T. Asano, Y. Tanaka, S. Noda: Appl. Phys. Lett. **82**, 1341 (2003)
7. Y. Akahane, T. Asano, B. S. Song, S. Noda: Nature (London) **425**, 944 (2003)
8. J.D. Joannopoulos, P.R. Villeneuve, S. Fan: Nature (London) **386**, 143 (1997)
9. A. Chutinan, M. Okano, S. Noda: Appl. Phys. Lett. **80**, 1698 (2002)
10. S. Boscolo, M. Midrio, T.F. Krauss: Opt. Lett. **27**, 1001 (2002)
11. S.Y. Lin, V.M. Hietala, L. Wang, E.D. Jones: Opt. Lett. **21**, 1771 (1996)
12. H. Kosaka, T. Kawashima, A. Tomita, M. Notomi, T. Tamamura, T. Sato, S. Kawakami: Phys. Rev. B **58**, R10096 (1998)

13. M. Notomi: Phys. Rev. B **32**, 10696 (2000)
14. P.E. Barclay, K. Srinivasan, M. Borselli, O. Painter: Elec. Lett. **39**, 842 (2003)
15. P. Bienstman, S. Assefa, S.G. Johnson, J.D. Joannopoulos, G.D. Petrich, L.A. Kolodziejski: J. Opt. Soc. Am. B **20**, 1817 (2003)
16. M. Qui: Phys. Rev. B **66**, 033103 (2002)
17. S.G. Johnson, S. Fan, P.R. Villeneuve, J.D. Joannopoulos: Phys. Rev. B **60**, 5751 (1999)
18. S.I. Bozhevolnyi, V.S. Volkov, T. Sondergaard, A. Boltasseva, P.I. Borel, M. Kristensen: Phys. Rev. B **66**, 235204 (2002)
19. H. Benisty, D. Labilloy, C. Weisbuch, C.J.M. Smith, T.F. Krauss, D. Cassagne, A. Beraud, C. Jouanin: Appl. Phys. Lett. **76**, 532 (2000)
20. C. Liguda, G. Boettger, A. Kuligk, M. Eich, H. Roth, J. Kunert, W. Morgenroth, H. Elsner, H. G. Meyer: Appl. Phys. Lett. **78**, 2434 (2001)
21. S. Foteinopoulou, A. Rosenberg, M.M. Sigalas, C.M. Soukoulis: J. Appl. Phys. **89**, 824 (2001)
22. K. Busch, S. Loelkes, R.B. Wehrspohn, H. Föll: *Photonic Crystals - Advances in Design, Fabrication, and Characterization* (Wiley-VCH, Weinheim 2004)
23. H. Benisty, P. Lalanne, S. Olivier, M. Rattier, C. Weisbuch, C.J.M. Smith, T.F. Krauss, C. Jouanin, C. Cassagne: Optical and Quantum Electronics **34**, 205 (2002)
24. G. Boettger, C. Liguda, M. Schmidt, M. Eich: Appl. Phys. Lett. **81**, 2517 (2002)
25. M. Clemens, T. Weiland: Progress in Electromagnetic Research PIER **32**, 65 (2001)
26. R. Boucher, U. Huebner, W. Morgenroth, H. Roth, H.G. Meyer, M. Schmidt, M. Eich: Micro & Nano Engineering (MNE), (2003), in press
27. J. Vydra, H. Beisinghoff, T. Tschudi, M. Eich: Appl. Phys. Lett. **69**, 1035 (1996)
28. S.G. Johnson and J.D. Joannopoulos: Optics Express **8**, 173 (2001)
29. W. Feng, S. Lin, B. Hooker, A.R. Mickelson: Applied Optics **34**, 6885 (1995)
30. A. Sakai, T. Fukazawa,T. Baba: IEICE Trans. Electron. **E85-C**, 1033 (2002)
31. M. Augustin, H.-J. Fuchs, D. Schelle, E.-B. Kley, S. Nolte, A. Tünnermann, R. Iliew, C. Etrich, U. Peschel,F. Lederer: Appl. Phys. Lett. **84**, 663 (2004)
32. C. Manolatou, S.G. Johnson, S. Fan, P.R. Villeneuve, H.A. Haus, J.D. Joannopoulos: J. Lightwave Tech. **17**, 1682 (1999)
33. W.J. Kim, J.D. O'Brien: J. Opt. Soc. Am. B **21**, 289 (2004)
34. P.I. Borel, A. Harpoth, L.H. Frandsen, M. Kristensen, P. Shi, J.S. Jensen, O. Sigmund: Opt. Express **12**, 1996 (2004)
35. H. Benisty, S. Olivier, C. Weisbuch, M. Agio, M. Kafesaki, C. Soukoulis, M. Qiu, M. Swillo, A. Karlsson, B. Jaskorzynska, A. Talneau, J. Moosburger, M. Kamp, A. Forchel, R. Ferrini, R. Houdre, U. Oesterle: IEEE J. Quant. Electron. **38**, 770 (2002)
36. M. Augustin, H.-J. Fuchs, D. Schelle, E.-B. Kley, S. Nolte, A. Tünnermann, R. Iliew, C. Etrich, U. Peschel, F. Lederer: Opt. Express **11**, 3284 (2003)

Single-Electron Devices

Jürgen Weis

Max-Planck-Institut für Festkörperforschung, Heisenbergstr. 1, 70569 Stuttgart, Germany

1 Introduction

The electrical charge is quantized in the elementary quantum $-e$ carried by single electrons. In mesoscopic systems at sufficiently low temperature, this discrete elementary charge can give rise to peculiar electrostatic effects. With achieving the ability of making small devices on the scale of less than few hundred nanometers, devices based on single-electron charging effects have been proposed and realized in the last 15 years.

After a brief introduction to the concepts of Coulomb blockade and single-electron charging, some device concepts for applications are presented, but also arrangements for studying basic physics of electrical transport relevant for molecular electronics are discussed. The presented picture for electrical transport through conducting mesoscopic particles ('island') by single-electron tunneling breaks down if correlated electron tunneling takes place. Under certain circumstances, correlated electron tunneling leads even to the conductance of a one-dimensional channel although Coulomb blockade is expected.

For historical reviews, further readings and other approaches to the topic of single-electron devices, the articles [1–8] are recommended. Especially for superconducting devices not treated here, we refer to [9,10], for proposals of using single-electron devices as qubits to [11] (quantum dots as islands), and [12] (superconducting devices).

The experimental data presented here have been collected during the last ten years in our institute. For their contributions I would like to thank my coworkers on this topic during that time – Jan Hüls, Matthias Keller, David Quirion, Jörg Schmid, Yayi Wei, Armin Welker, Ulf Wilhelm, and Klaus v. Klitzing. Of course, similar data can be found in literature published by other groups.

2 Single-Electron Charging Energy and Coulomb Blockade Effect

Figure 1 shows an arrangement of an electrically uncharged *metal island* embedded in a dielectric medium and surrounded by other metal electrodes which are electrically connected. By transfering a single electron from the electrodes to the island, the island is charged negative to $q = -e$ and positive image charges q_1, q_2 spread over the electrodes (see Fig. 1b). Note, the overall charge of the system compensates to zero: $-e + q_1 + q_2 = 0$. Similarly, by transferring an electron

J. Weis, Single-Electron Devices, Lect. Notes Phys. **658**, 87–121 (2005)
http://www.springerlink.com/

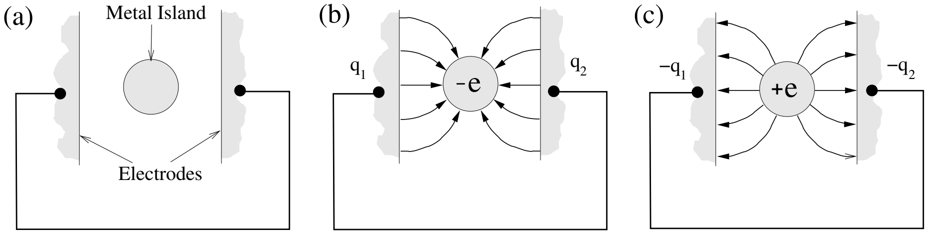


Fig. 1. (a) A metal island embedded between electrodes which are electrically connected. Transfering an electron onto the island (b) or taking off the electron from the island (c) charges the capacitor formed by the island and the electrodes.

from the electrically uncharged island to the electrodes, the island is charged positively and negative image charges are induced on the surrounding electrodes (see Fig. 1c). The arrangement resembles a capacitor configuration with the capacitance C_Σ where the island reflects one electrode of this capacitor and the others form all together the counter electrode. For both charge configurations ($q = -e$ and $q = e$), the electrostatic energy E_C

$$E_\mathrm{C} = \frac{e^2}{2\,C_\Sigma} \tag{1}$$

is stored in the arrangement. The quantity E_C is usually denoted as *single-electron charging energy*[1]. This energy is required for the separation of a single electron from its positive counter charge spread over the other conductors. It is the electrostatic energy barrier felt by the single electron moving onto or from the electrically neutral island.

Usually this energy E_C is not noticeable since the island size and therefore C_Σ is large. However, for $C_\Sigma < 10^{-15}$ F which corresponds[2] to the 'self-capacitance' $C_\Sigma = 4\pi\epsilon_0\epsilon\, R$ of a metallic sphere with radius $R < 1\,\mu$m embedded in a dielectric medium with $\epsilon = 10$, E_C exceeds the thermal energy $k_\mathrm{B}T$ at $T = 4$ K. For $C_\Sigma < 3\cdot 10^{-18}$ F which is fulfilled for $R < 2.8$ nm, even $k_\mathrm{B}T$ at room temperature ($T = 300$ K) is exceeded. From this, we have to conclude that the single-electron charging energy E_C is of importance to describe single-electron movements in systems from mesoscopic size down to atomic size.

A simple two-terminal arrangement for discussing the consequence is shown in Fig. 2a. A small island is embedded between two lead electrodes denoted as *source* S and *drain* D. Thin insulators separate the island from the two leads. These layers should be thin enough that – due to quantum mechanics – tunneling of electrons through the insulator layers is possible, thick enough that it is plausible to describe single electrons in the system as being localized either on the metal island or the lead electrodes. Since the metal island is *almost isolated*, the total charge on the metal electrodes is considered as being quantized in the elementary charge e. Due to E_C which is required for recharging the island by

[1] Sometimes [2] the quantity e^2/C_Σ is denoted by the same name.

[2] Counter electrode at infinite distance.

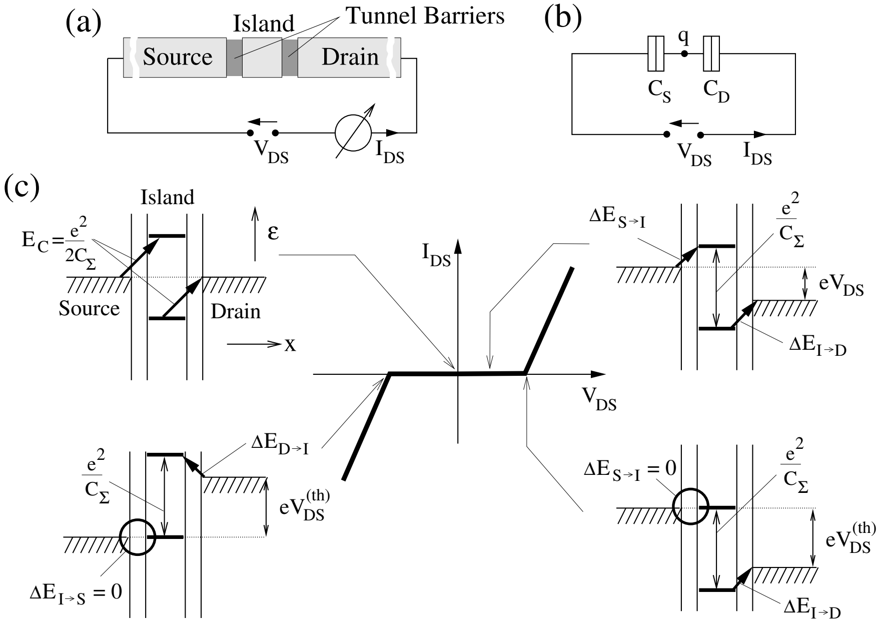


Fig. 2. (a) Two-terminal arrangement for discussing the Coulomb blockade effect in electrical transport. (b) The respective capacitance circuit. Note $C_\Sigma = C_S + C_D$. (c) Sketch of the expected non-linear $I_{DS}(V_{DS})$ characteristic with energy schemes for distinct V_{DS} values reflecting the energetical position of the Fermi levels of the island for charge states $q = -e$ and $q = e$ relatively to the Fermi level of source and drain.

a single electron entering or leaving, electrical transport is suppressed around $V_{DS} = 0$ if $E_C \gg k_B T$ (*Coulomb blockade effect* of electrical transport).

With increasing bias voltage $V_{DS} > 0$, the electrostatic energy barriers for adding an electron from source

$$\Delta E_{S\to I} = E_C - e\,\frac{C_D}{C_\Sigma}\,V_{DS} \tag{2}$$

and the electrostatic energy barrier for an electron leaving to drain

$$\Delta E_{I\to D} = E_C + e\,\frac{C_D}{C_\Sigma}\,V_{DS} - e\,V_{DS} \tag{3}$$

are reduced as a consequence of the applied voltage V_{DS} (The respective capacitance circuit is given in Fig. 2b). Similar happens for $V_{DS} < 0$. The suppression of current is finally overcome for

$$|V_{DS}| \geq V_{DS}^{(th)} \equiv \min\left(\frac{e}{2\,C_S}; \frac{e}{2\,C_D}\right) , \tag{4}$$

and the drain-source current $|I_{DS}|$ rises rapidly with increasing $|V_{DS}|$. If $E_C \gg k_B T$, for such a two-terminal device a non-linear current-voltage characteristic with threshold values lying symmetrically around $V_{DS} = 0$ is obtained.

3 Concept of a Single-Electron Transistor (SET)

Instead of overcoming the Coulomb blockade by increasing V_{DS}, a *gate electrode* G with variable gate-voltage V_{GS} can be added to the arrangement (see Fig. 3a). With increasing gate voltage V_{GS}, the electrostatic potential of the island is shifted due to the capacitance circuit sketched in Fig. 3b. With increasing $V_{\mathrm{GS}} > 0$, negative charge is accumulated on the island – not in a continuous but in a step-like manner as sketched in Fig. 3c (*single-electron charging*). The first electron is charged at $V_{\mathrm{GS}} = V_{\mathrm{GS}}^{(\mathrm{th})}$ when the electrostatic energy for an electron on the island is lowered just compensating for E_{C}, i.e.,

$$\Delta E_{\mathrm{S}\to\mathrm{I}} = E_{\mathrm{C}} - e\,\frac{C_{\mathrm{G}}}{C_{\Sigma}}\,V_{\mathrm{GS}} \overset{!}{=} 0\ , \tag{5}$$

leading to the threshold voltage

$$V_{\mathrm{GS}}^{(\mathrm{th})} = \frac{E_{\mathrm{C}}}{e\,C_{\mathrm{G}}/C_{\Sigma}} = \frac{e}{2\,C_{\mathrm{G}}}\ . \tag{6}$$

At this gate-voltage value, the charge state of the island fluctuates by e. Applying a small drain-source voltage V_{DS}, a directed current is measured between source and drain – carried by single electrons passing one after the other the island.

What about charging the electrically neutral island by ΔN electrons from the source lead? The electrostatic energy stored in such a charge configuration ($q = -\Delta N\,e$) – under the condition that V_{DS} and V_{GS} are fixed – is given by

$$E_{\mathrm{elst}}(\Delta N; V_{\mathrm{GS}}, V_{\mathrm{DS}}) = -\Delta N\,e\left(\frac{C_{\mathrm{G}}}{C_{\Sigma}}\,V_{\mathrm{GS}} + \frac{C_{\mathrm{D}}}{C_{\Sigma}}\,V_{\mathrm{DS}}\right) + \frac{(\Delta N\,e)^2}{2\,C_{\Sigma}}\ . \tag{7}$$

The first term describes the potential energy of ΔN electrons at the electrostatic potential which is found due to the capacitance divider at the electrically neutral island. The second term takes into account the work which has to be done to separate the charge $q = -\Delta N\,e$ from its counter charge spread over the electrodes source S, drain D and gate G.

Having already charged the island with ΔN electrons, the next electron '$\Delta N+1$' moving from source to the charged island feels at fixed applied V_{GS} and V_{DS} the electrostatic energy difference

$$\begin{aligned}\Delta E_{\mathrm{S}\to\mathrm{I}}(\Delta N{+}1; V_{\mathrm{GS}}, V_{\mathrm{DS}}) &= E_{\mathrm{elst}}(\Delta N{+}1; V_{\mathrm{GS}}, V_{\mathrm{DS}}) - E_{\mathrm{elst}}(\Delta N; V_{\mathrm{GS}}, V_{\mathrm{DS}}) \\ &= \left(\Delta N + \tfrac{1}{2}\right)\frac{e^2}{C_{\Sigma}} - e\,\frac{C_{\mathrm{G}}}{C_{\Sigma}}\,V_{\mathrm{GS}} - e\,\frac{C_{\mathrm{D}}}{C_{\Sigma}}\,V_{\mathrm{DS}}\ . \end{aligned}\tag{8}$$

Similarly, having ΔN electrons on the island, the electron 'ΔN' feels for moving towards drain the electrostatic energy difference

$$\begin{aligned}\Delta E_{\mathrm{I}\to\mathrm{D}}(\Delta N; V_{\mathrm{GS}}, V_{\mathrm{DS}}) &= E_{\mathrm{elst}}(\Delta N{-}1; V_{\mathrm{GS}}, V_{\mathrm{DS}}) - e\,V_{\mathrm{DS}} - E_{\mathrm{elst}}(\Delta N; V_{\mathrm{GS}}, V_{\mathrm{DS}}) \\ &= -\left(\Delta N - \tfrac{1}{2}\right)\frac{e^2}{C_{\Sigma}} + e\,\frac{C_{\mathrm{G}}}{C_{\Sigma}}\,V_{\mathrm{GS}} - e\left(1 - \frac{C_{\mathrm{D}}}{C_{\Sigma}}\right)V_{\mathrm{DS}}\ . \end{aligned}\tag{9}$$

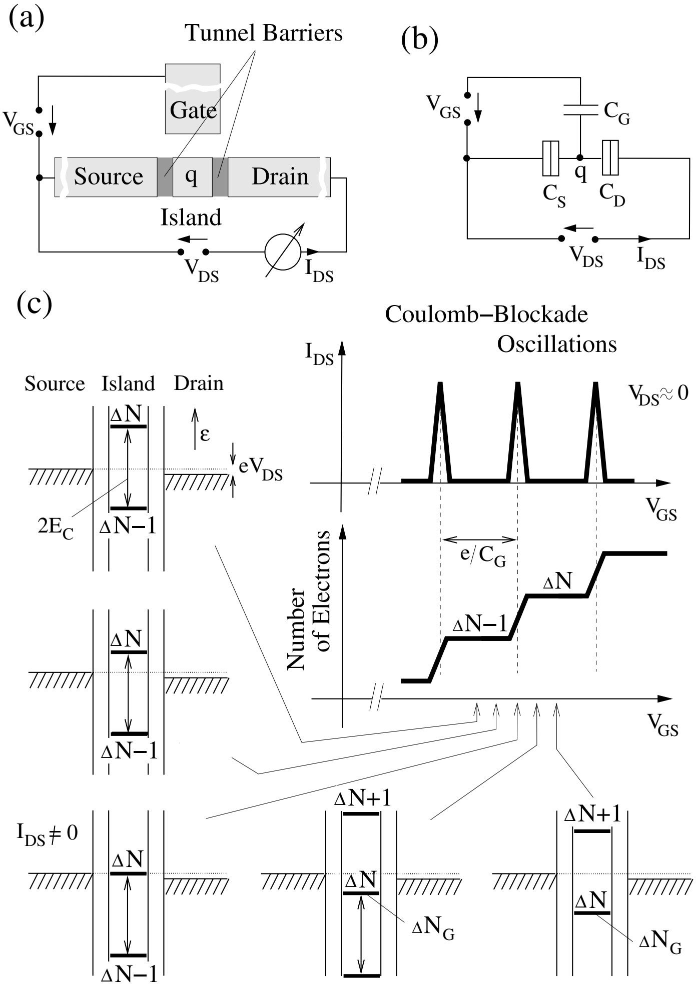


Fig. 3. (a) Three-terminal arrangement of a single-electron transistor. (b) The respective capacitance circuit. Note $C_\Sigma = C_\mathrm{S} + C_\mathrm{D} + C_\mathrm{G}$. (c) With increasing gate voltage V_GS, electrons are accumulated on the island. Whenever the charge state can energetically fluctuate by e, i.e., the energy for two charge states is degenerate, current I_DS flows for small applied V_DS through the island, leading to a periodically modulated $I_\mathrm{DS}(V_\mathrm{GS})$-characteristic – the Coulomb blockade oscillations. For distinct V_GS values, the respective energy schemes are given.

It contains the final electrostatic energy $-e\,V_\mathrm{DS}$ of the electron on the drain site.

The energy differences $E_\mathrm{elst}(\Delta n; V_\mathrm{GS}, V_\mathrm{DS}) - E_\mathrm{elst}(\Delta n{-}1; V_\mathrm{GS}, V_\mathrm{DS})$ with $n \in \{\cdots, N-1, N, N+1, \cdots\}$ define an energy ladder with fixed energy level spacing $2\,E_\mathrm{C} = e^2/C_\Sigma$ which shifts linearly with V_DS and V_GS: For given V_DS and V_GS the level 'Δn' reflects the energetical position of the Fermi level on the island relatively to the Fermi levels of the two leads if the island is charged to $q = -\Delta n\, e$.

The relative position of this energy ladder are given for distinct parameters $(V_{\mathrm{GS}}; V_{\mathrm{DS}} \approx 0)$ in the energy schemes of Fig. 3c. In thermodynamic equilibrium, $\Delta n = \Delta N_{\mathrm{G}}$ additional electrons are trapped on the island if

$$\begin{aligned}
&\text{for } V_{\mathrm{DS}} \geq 0 \\
&\qquad \Delta E_{\mathrm{S}\to\mathrm{I}}(\Delta N_{\mathrm{G}}{+}1; V_{\mathrm{GS}}, V_{\mathrm{DS}}) > 0 \ \text{ and } \ \Delta E_{\mathrm{I}\to\mathrm{D}}(\Delta N_{\mathrm{G}}; V_{\mathrm{GS}}, V_{\mathrm{DS}}) > 0\ , \qquad (10) \\
&\text{for } V_{\mathrm{DS}} \leq 0 \\
&\qquad \Delta E_{\mathrm{I}\to\mathrm{S}}(\Delta N_{\mathrm{G}}; V_{\mathrm{GS}}, V_{\mathrm{DS}}) > 0 \ \text{ and } \ \Delta E_{\mathrm{D}\to\mathrm{I}}(\Delta N_{\mathrm{G}}{+}1; V_{\mathrm{GS}}, V_{\mathrm{DS}}) > 0\ . \qquad (11)
\end{aligned}$$

Whenever $\Delta E_{\mathrm{S}\to\mathrm{I}} = 0$ or $\Delta E_{\mathrm{I}\to\mathrm{D}} = 0$, the charge state of the island can fluctuate by e. Applying a small drain-source voltage V_{DS}, a directed current is measured between source and drain. With changing the gate voltage V_{GS} at small V_{DS}, the current is modulated with the gate voltage period

$$\Delta V_{\mathrm{GS}} = \frac{e}{C_{\mathrm{G}}} \qquad (12)$$

as sketched in Fig. 3c. This characteristic is denoted as *Coulomb blockade oscillations* (CBOs). Since the current is carried by single electrons passing the island one-by-one, the three-terminal device with such a characteristic is named *single-electron transistor* (SET) [13,14].

Evaluating (10) and (11) allows to define transport regions for a single-electron transistor as a function of the drain-source voltage V_{DS} and the gate voltage V_{GS}. The result is sketched in Fig. 4: Light grey shaded are the *regions of Coulomb blockade* (fulfilling (10) and (11)) at low temperature where the electron number is fixed. Fluctuations by only one electron charge $-e$ are possible in the adjacent regions. These are the *regions of single-electron tunneling* since there the electrons are passing the island one after the other. Along the gate voltage axis with $V_{\mathrm{DS}} \approx 0$, the Coulomb blockade oscillations are obtained. With further increasing $|V_{\mathrm{DS}}|$, more and more charge configurations become energetically possible. For distinct parameter configurations $(V_{\mathrm{DS}}, V_{\mathrm{GS}})$, the respective energy scheme are depicted. For the metal single-electron transistor, the transport characteristics are periodic in V_{GS}: With each gate voltage change $\Delta V_{\mathrm{GS}} = e/C_{\mathrm{G}}$, the same electrostatic energy barriers for recharging the island are present – only with one electron more trapped on the island.

The borderlines between Coloumb blockade and single-electron tunneling regime have the slopes

$$\left.\frac{\mathrm{d}V_{\mathrm{GS}}}{\mathrm{d}V_{\mathrm{DS}}}\right|_{\Delta E_{\mathrm{S}\to\mathrm{I}}=0} = -\frac{C_{\mathrm{D}}}{C_{\mathrm{G}}} \qquad \text{and} \qquad \left.\frac{\mathrm{d}V_{\mathrm{GS}}}{\mathrm{d}V_{\mathrm{DS}}}\right|_{\Delta E_{\mathrm{I}\to\mathrm{D}}=0} = \frac{C_{\Sigma} - C_{\mathrm{D}}}{C_{\mathrm{G}}}\ . \qquad (13)$$

Note, these relations are valid for the special choice of the source electrode as the reference electrode for all applied voltages.

One should also realize that the notation of the two different transport regions of a single-electron transistor – Coulomb blockade and single-electron tunneling regime – as a function of V_{DS} and V_{GS} are obtained due to energy considerations. Multi-electron transport is predicted at higher $|V_{\mathrm{DS}}|$ values where regions of

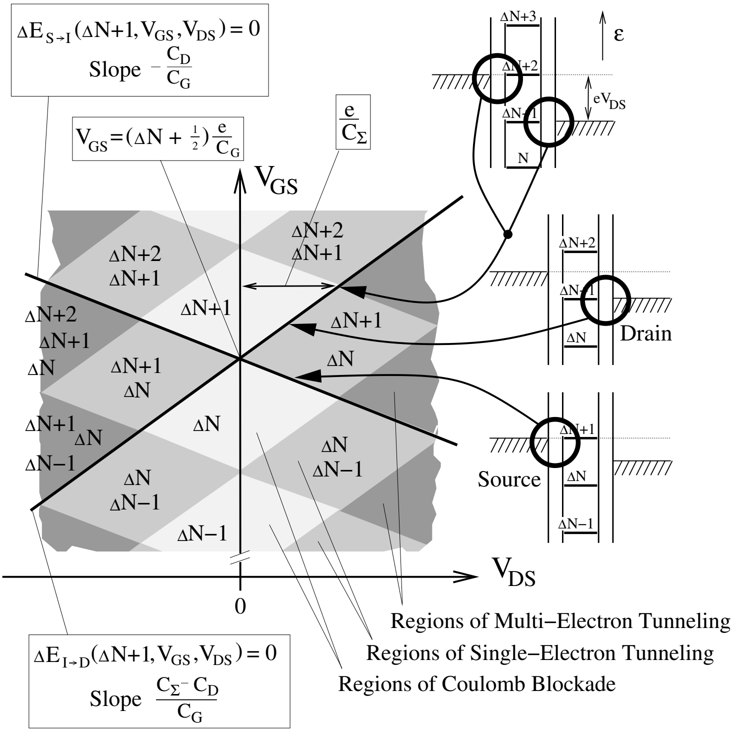


Fig. 4. Transport regions of a single-electron transistor as a function of V_{DS} and V_{GS}. This pattern is usually referred to as 'diamond-like'.

more than two charge states could coexist. However, SETs with strongly asymmetric tunnel barriers also show single-electron transport at these higher $|V_{\mathrm{DS}}|$ values: An electron leaving via the thicker tunnel barrier is almost immediately replaced by an electron tunneling through the thinner tunnel barrier; for opposite drain-source voltage an electron entering the island via the thicker tunnel barrier leaves usually faster via the thinner barrier than another electron can enter via the thicker barrier. The dynamics of the system restrict the charge fluctuations on the island to e. Under such conditions the current I_{DS} increases in a step-like manner with increasing $|V_{\mathrm{DS}}|$ whenever another charge state has become energetically available, i.e., a boundary line in Fig. 4 is crossed with increasing $|V_{\mathrm{DS}}|$. The so-called *Coulomb-staircase* characteristic in $I_{\mathrm{DS}}(V_{\mathrm{DS}})$ is obtained [15].

4 Examples for the Realization of Single-Electron Transistors

Two examples for the realization of a single-electron transistor are discussed in this section. First, a device made from metal is shown to demonstrate that small metal islands indeed offer transport characteristics dominated by Coulomb blockade and single-electron charging effects although more than 10^9 electrons are actually present in the condcution band of the island. In contrast, as the second realization, a SET made from a semiconductor material is presented. It contains a *quantum dot* as the island with a small number of trapped electrons (about 10 to 20) and a discrete excitation spectrum, and even allows the in-situ control over the tunnel coupling between island and leads. Due to their *in-situ tunability*, such quantum dot systems can act as model systems for studying basic phenomena in electrical transport through single molecules or atoms embedded between lead electrodes.

Other arrangements and realizations of single-electron transistors can be found in the literature cited in the introduction.

4.1 Single-Electron Transistor Made from Metal

An example for a metal single-electron transistor made from aluminum is shown in 'Cross Section 1' of Fig. 5a and as a scanning-electron microscope image in Fig. 5b. The devices is fabricated by using a two-angle evaporation technique also used to fabricate the first SET [16]: With electron-beam lithography, a two-layer organic resist is patterned resulting in openings to the substrate with large undercut (see 'Cross Section 2'). In vacuum, aluminum layers are evaporated twice under different angles through the openings onto the substrate. By an in-situ oxidation *between* first and second evaporation process, a thin aluminum oxide of few nanometers is formed on the first aluminum layer. The resist is lifted off and a metal structure remains on the substrate. Due to the two different evaporation angles, the metal patterns of the first and second evaporation process are slightly shifted against each other leading to an overlap in certain regions. In the overlap regions, the thin aluminum oxide acts as tunnel barriers between both aluminum layers, whereas the uncovered aluminum is unavoidable oxidized further in air. The island has a length of $1\,\mu$m and a width of $0.1\,\mu$m. The overlap region defining the tunnel barriers towards the leads are $0.1\,\mu$m by $0.1\,\mu$m in size. Coulomb blockade oscillations measured on this device at $T = 0.1\,$K for $V_{\rm DS} = 80\,\mu$V are shown in Fig. 5c. As the gate electrode, a conductive layer in the substrate 86 nm below the surface is used. Due to the small size of the device, the total capacitance C_Σ – dominated by the overlap regions of the tunnel junctions – is small leading to $E_{\rm C} \approx 0.1\,$meV. In Fig. 5d the measured $I_{\rm DS}(V_{\rm DS}, V_{\rm GS})$ characteristics of a similar metal single-electron transistor ($E_{\rm C}$ slightly smaller) are shown. Clearly the Coulomb blockade regions are visible. Beyond the respective threshold in $V_{\rm DS}$, the current $I_{\rm DS}$ increases.

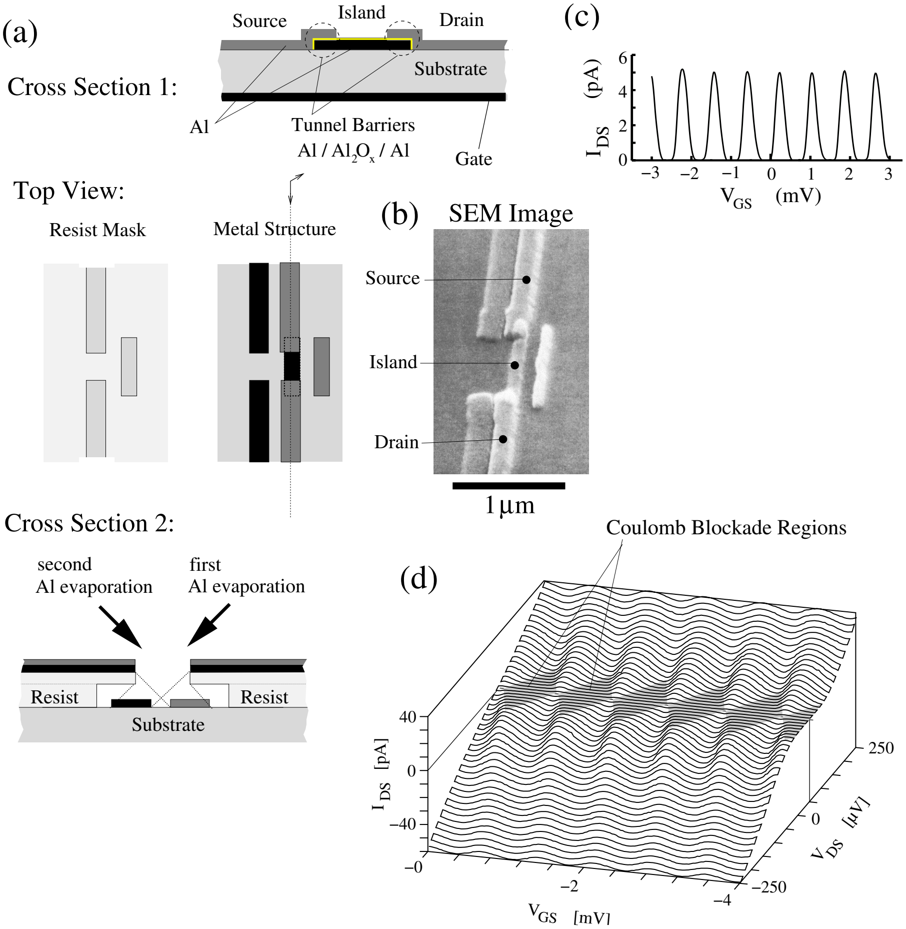


Fig. 5. SET made from metal: (a) Fabrication process (see text). (b) Scanning electron microscope image. (c) Coulomb-blockade oscillations. (d) $I_{\rm DS}(V_{\rm DS}, V_{\rm GS})$ characteristics measured at $T = 0.1\,{\rm K}$. (from Y.Y. Wei, J. Hüls et al., MPI-FKF)

4.2 Single-Electron Transistor Containing a Quantum Dot as Island

Quantum dots or *zero-dimensional electron systems* are objects where electrons are confined in a small enclosure allowing the single electron only certain eigenvalues for its energy due to the wave character of electrons as quantum mechanical particles. As sketched in Fig. 6, with decreasing the size of the island, the quasi-continuous single-particle energy spectrum (like that of a metal) turns into a discrete one (like that of an atom) if the deBroglie wavelength $\lambda_{\rm F} = h/\sqrt{2m\,\varepsilon_{\rm F}}$ of an electron at the Fermi energy $\varepsilon_{\rm F}$ of the respective bulk material becomes comparable to the island diameter D.

A realization of a single-electron transistor with a quantum dot as island is shown in Fig. 7 – denoted as *split-gate quantum dot system*: Base is a GaAs/

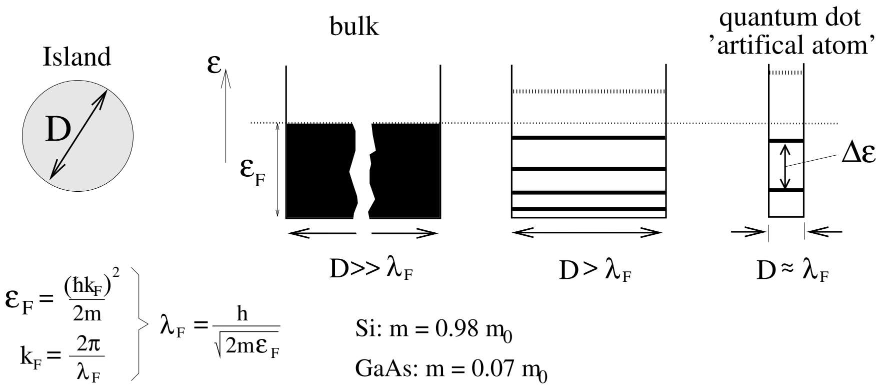


Fig. 6. Enclosing electrons to a smaller space, only certain eigenvalues for their kinetic energy become possible (Sketch!). Spatial enclosures with a discrete single-particle spectrum are denoted as quantum dots.

$\mathrm{Al}_{0.33}$ $\mathrm{Ga}_{0.67}\mathrm{As}$ heterostructure containing a *two-dimensional electron system* at the GaAs/AlGaAs heterojunction interface 86 nm below the surface. In GaAs, the effective mass of an electron in the conduction band is rather small, $m = 0.07\,m_0$ where m_0 is the free electron mass. Therefore, single-particle energy level spacing $\Delta\varepsilon$ of several meV are achieved for GaAs islands of few tens of nanometers – large enough to be resolved at low temperature ($k_BT = 1$ meV at $T = 12$ K). To define the quantum dot system, metallic gates were deposited on top of a mesa remained after partially etching the surface of the heterostructure. The 2DES is electrically contacted by alloying metal at certain regions of the mesa. The diameter of the area between the tips of the gate fingers is here about 0.35 μm. With applying negative voltages to the gate electrodes, the 2DES is divided in parts, defining the quantum dot of about 0.2 μm in diameter between the gate fingers, coupled by tunnel barriers to parts of the 2DES acting as source and drain leads. In addition to these topgates, a metallic backgate electrode on the reverse side of the undoped substrate (0.5 mm thick) is used to change the electrostatic potential of the quantum dot by changing the applied voltage V_{BS}. In Fig. 7b, a typical curve of the conductance I_{DS}/V_{DS} versus the backgate voltage for small drain-source voltage ($V_{DS} \approx 5\,\mu$V) is shown – the Coulomb blockade oscillations ($T = 0.1$ K). In contrast to the CBO characteristic shown for the metal single-electron transistor, the peak heights are strongly modulated and the peak distances are not exactly periodic. Both effects are even emphasized by applying a magnetic field as shown in Fig. 7c. This indicates that the character of the electronic states of the quantum dot – changed by the magnetic field – affects the electrical transport.

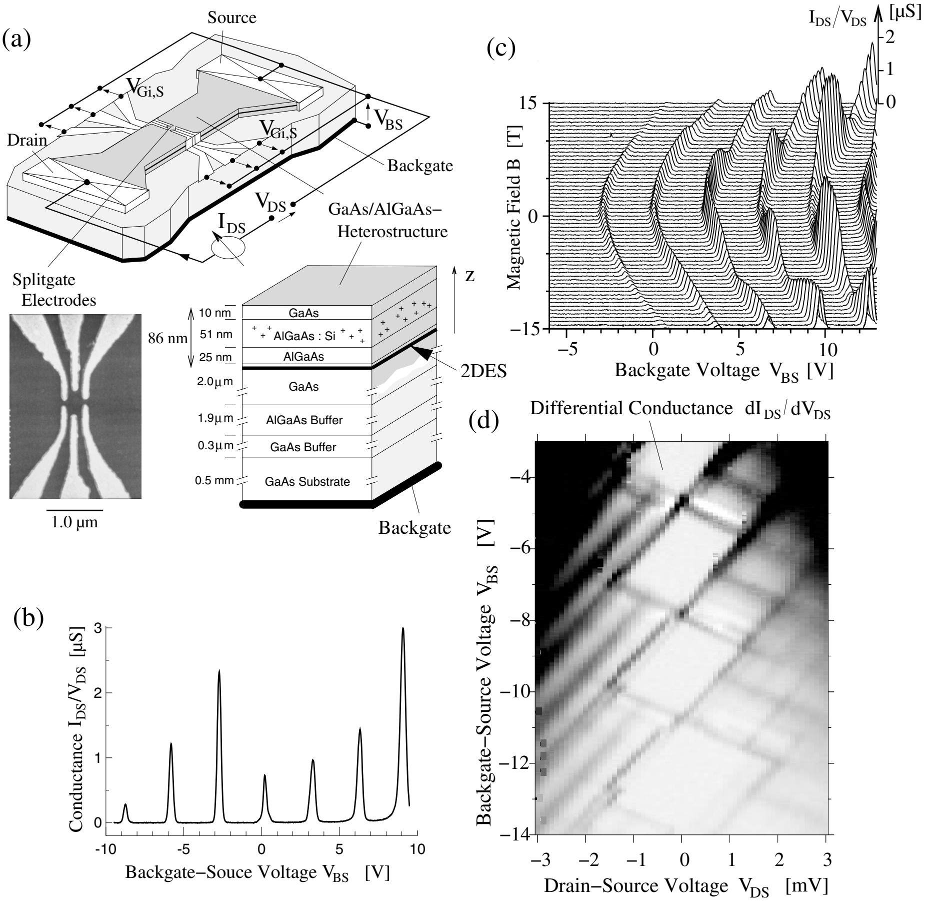


Fig. 7. SET with quantum dot as island: (a) Metallic gates on top of a GaAs-AlGaAs heterostructure are used to define a quantum dot system by partially electrostatically depleting a two-dimensional electron system (2DES). (b) Coulomb blockade oscillations as a function of the backgate voltage V_{BS}. (c) Coulomb blockade oscillations $I_{\mathrm{DS}}(V_{\mathrm{BS}})$ for different magnetic fields applied in parallel to the plane of the 2DES. (d) Differential conductance $\mathrm{d}I_{\mathrm{DS}}/\mathrm{d}V_{\mathrm{DS}}$ in greyscale as a function of V_{DS} and V_{GS}. (from J. Weis et al., MPI-FKF)

5 Quantum Dot as an Interacting N-Electron System: An Artifical Atom with Tunable Properties

Obviously the electrostatic model is not sufficient, i.e., the description has to be extended. A better approach is to ask which is the energy necessary for adding an electron into a given confining potential (defined by gate electrodes with electrostatic potentials $\{V_i\}$, material composition and fixed charges due to donors and acceptors) when already the number N of electrons is present. To answer this, N and $N+1$ electrons have to be treated quantum-mechanically as interacting N and $N+1$ electron systems in the confining potential. A Hamiltonian $\hat{H}(n;\{V_i\})$ of n electrons modeling the electrostatics of realistic quantum dots has the form [17]

$$\hat{H}(n;\{V_i\}) = \sum_{s=1}^{n} \frac{\hat{p}_s^2}{2\,m} - \sum_{s=1}^{n} e\,\Phi_{\text{ext}}(\hat{r}_s;\{V_i\}) + \tfrac{1}{2}\sum_{s=1}^{n}\sum_{\substack{s'=1\\ s'\neq s}}^{n} e^2\,G(\hat{r}_s,\hat{r}_{s'}) \tag{14}$$

where $\hat{p}_s$ and $\hat{r}_s$ denote the momentum and position operator for electron s, respectively. The quantity $G(\boldsymbol{r},\boldsymbol{r}\,')$ is the electrostatic Green's function for describing the electrostatics of the system without the presence of the n electrons [17]. The physical meaning of $q\,G(\boldsymbol{r},\boldsymbol{r}\,')$ is the electrostatic potential contribution at position $\boldsymbol{r}$ caused by a point charge q located at $\boldsymbol{r}\,'$ in the given arrangement. In particular, it describes the electrostatic electron-electron interaction in the quantum dot taking into account the electrostatic screening effect by the electrodes and the dielectric medium. Comparing (14) with (7), it becomes clear that the effective electron-electron interaction (last term in (14)) is responsible for the Coulomb blockade effect in quantum dots. The confining potential $\Phi_{\text{ext}}(\boldsymbol{r};\{V_i\})$ is given by the fixed charge distribution, the arrangement of the electrodes and conduction band offsets due to the use of different materials (see Fig. 8a). It is *independent* of the electron number confined in the quantum dot. The electrostatic contributions to $\Phi_{\text{ext}}(\boldsymbol{r};\{V_i\})$ can all be expressed by $G(\boldsymbol{r},\boldsymbol{r}\,')$ [17]. One should note that the confining potential depends linearly on the electrostatic potentials $\{V_i\}$ of the electrodes, i.e., the electrostatic potential at position $\boldsymbol{r}$ is linearly shifted with changing V_i, i.e.,

$$\Phi_{\text{ext}}(\boldsymbol{r};\{V_i\}) \propto \sum_i \alpha_i(\boldsymbol{r}\,)\,V_i \tag{15}$$

where the quantity $\alpha_i(\boldsymbol{r}\,)$ reflects the fraction of image charge induced by a point charge at position $\boldsymbol{r}$ in the arrangment on electrode i.

By solving the Schrödinger equation

$$\hat{H}(n;\{V_i\})\,|n,l;\{V_i\}\rangle = E(n,l;\{V_i\})\,|n,l;\{V_i\}\rangle \tag{16}$$

a total energy spectrum $E(n,l;\{V_i\})$ for the confined n-electron system is obtained for a certain set of applied voltages $\{V_i\}$. For convenience, the index l represents a set of quantum numbers that characterizes the different n-electron states $|n,l;\{V_i\}\rangle$ starting from $l=0$ for the groundstate, and numbering the excited states unambiguously further with increasing energy $E(n,l;\{V_i\})$.

Looking at the Hamiltonian (14), it becomes clear why quantum dots have sometimes been denoted as *artifical atoms* [19,20] with tunable properties: The confining potential for electrons in an atom (the Coulomb potential of the bare nucleus) is replaced by $\Phi_{\text{ext}}(\boldsymbol{r};\{V_i\})$. The pure Coulomb interaction between electrons in atoms has to be replaced by $e^2\,G(\boldsymbol{r},\boldsymbol{r}\,')$ if electrostatic screening due to the dielectric medium or surrounding electrodes is present. In principle, both

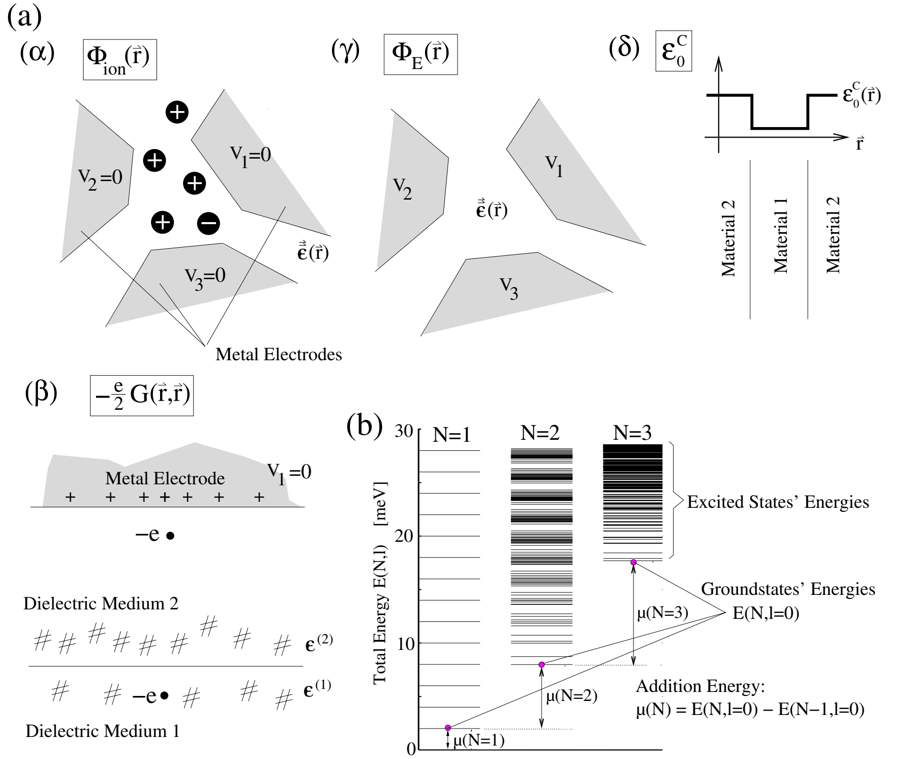


Fig. 8. (a) Ingredients defining the confining potential for the electrons: (α) Fixed charge distribution with its image charges induced on the electrodes. (β) Own image charges induced on the electrodes or dielectric interfaces. (γ) Voltages applied to the electrodes. (δ) Conduction band offsets by using different materials. (b) Total energy spectra for one, two and three electrons confined in a parabolic confining potential ($\hbar\omega_0 = 2$ meV) (adopted from D. Pfannkuche et al. [18]).

$\Phi_{\text{ext}}(\boldsymbol{r};\{V_i\})$ and $G(\boldsymbol{r},\boldsymbol{r}')$ can be designed to purpose. If the confining potential obeys spatial symmetries, certain degeneracies in the electronic spectrum can be expected. On the other hand, certain shapes of the confining potential allow to consider the quantum dot as a chaotic system.

In a very popular model – the Constant Interaction Model (CIM) [21,15] – the total energy $E(n;\{V_i\})$ is written as

$$E(n;\{V_i\}) = \sum_{s=1}^{n} \varepsilon_s - n\,e \sum_i \frac{C_i}{C_\Sigma}\,V_i + \frac{(n\,e)^2}{2\,C_\Sigma} - n\,e \cdot \text{const} \tag{17}$$

where ε_s is the eigenenergy of the single electron 's' in the (effective) confining potential of the quantum dot. Due to Pauli's principle, single-particle states are sequentially occupied with increasing electron number n and the electron-electron interaction is treated by the constant C_Σ. This description is not generally valid: One should note that – different to atoms – due to the larger size, in quantum dots usually the electron-electron interaction is dominating the electronic properties and not the quantization effect on the kinetic energy due to the confining of the electrons. The total energy spectrum becomes complex as shown as an example in Fig. 8b. The electrons in the quantum dot feel each other and behave correlated (which is an exciting subject on its own (see for recent review [22])).

6 Transport Spectroscopy on Quantum Dot Systems

Having N electrons confined, they will end up in the groundstate $|N, 0; \{V_i\}\rangle$ at low temperature. The minimum in energy required for adding another electron to the system is achieved when ending in the groundstate $|N+1, 0; \{V_i\}\rangle$ of the $N+1$ electron system. The energy ladder

$$\begin{aligned} \mu(n;\{V_i\}) \equiv E(n,0;\{V_i\}) - E(n-1,0;\{V_i\}) \ , \ \text{where} \\ n \in \{\cdots, N-1, N, N+1, \cdots\} \end{aligned} \tag{18}$$

gives for fixed potentials $\{V_i\}$ by its position relatively to the electrochemical potentials (Fermi levels) μ_{S} and μ_{D} of source and drain the energy barriers for recharging the quantum dot by a single electron. Under circumstances this energy ladder is *linearly* shifted with changing one of the applied voltages V_{GS} and V_{DS}: The characteristic 'diamond-like' transport regions of a single-electron transistor as shown in Fig. 4 are recovered – although not that regular in size. The boundaries between the different charge states in the V_{DS} vs. V_{GS} are obtained with $\mu_{\mathrm{S}} - \mu_{\mathrm{D}} = e\,V_{\mathrm{DS}}$ from

$$\begin{aligned} \mu(n;\{V_i\}) = \mu_{\mathrm{S}} \quad \text{and} \quad & \mu(n;\{V_i\}) = \mu_{\mathrm{D}} \\ \text{with} \quad & n \in \{\cdots, N-1, N, N+1, \cdots\} \ . \end{aligned} \tag{19}$$

In Fig. 7d, the differential conductance $\mathrm{d}I_{\mathrm{DS}}/\mathrm{d}V_{\mathrm{DS}}$ of the quantum dot system is shown measured as a function of V_{DS} and V_{BS}. In the linear greyscale plot, white regions correspond to $\mathrm{d}I_{\mathrm{DS}}/\mathrm{d}V_{\mathrm{DS}} < -0.1\,\mu\mathrm{S}$ and black ones to $\mathrm{d}I_{\mathrm{DS}}/\mathrm{d}V_{\mathrm{DS}} > 2\,\mu\mathrm{S}$. Positive peaks in the differential conductance indicate a step-like increase in the current I_{DS} with increasing $|V_{\mathrm{DS}}|$, negative ones a step-like decrease. Clearly the Coulomb-blockade regions are identified. In the adjacent single-electron tunneling regions, additional peaks in the differential conductance are observed indicating the opening of other transport channels although the charge state of the quantum dot can only fluctuate by one elementary charge. These can be attributed to electrical transport using in competition excited states of the quantum dot system [23–26].

What is the link between the total energy spectra of n and $n+1$ electron systems and that what is seen in the single-electron tunneling regime ('transport spectrum')? In Fig. 9a, the fictitious total energy spectra for N and $N+1$ electrons are given which lead to the energy ladder defined by the transistion

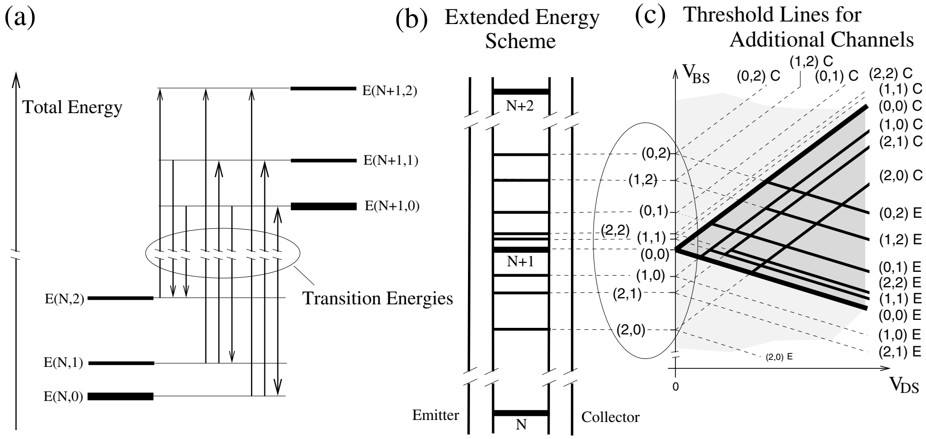


Fig. 9. (a) Fictitious total energy spectra of N and $N+1$ electrons confined in the quantum dot. Bold are the groundstate energies. (b) Plot of the transition energies $E(n,k;\{V_i\}) - E(n-1,l;\{V_i\})$ as energy levels. Energy levels representing differences between groundstate energies are bold and marked the respective $n \in \{\cdots, N-1, N, N+1, \cdots\}$. (b) Threshold lines for additional channels extracted from (a). Whether all are visible depends in detail on (quasi-)selection rules and the dynamic of the system.

energies $E(N{+}1,k;\{V_i\}) - E(N,l;\{V_i\})$ and plotted in Fig. 9b. It includes the transition energy $\mu(N;\{V_i\}) = E(N{+}1,0;\{V_i\}) - E(N,0;\{V_i\})$ between the groundstates. With changing a gate voltage V_{GS} or the drain-source voltage V_{DS}, the energy ladder is shifted, i.e., these levels come in resonance with μ_{S} or μ_{D} for certain $(V_{\mathrm{GS}}, V_{\mathrm{DS}})$ values,

$$\begin{aligned} E(N{+}1,k;\{V_i\}) - E(N,l;\{V_i\}) &= \mu_{\mathrm{S}} \quad \text{or} \\ E(N{+}1,k';\{V_i\}) - E(N,l';\{V_i\}) &= \mu_{\mathrm{D}} \ . \end{aligned} \tag{20}$$

By this, an additional transport channel might be opened on source or drain side, respectively. However, it requires that the electron system of the quantum dot is not captured in one of the groundstates and remains there, but allows for fluctuations between N and $N+1$, i.e., besides (20) at the same time

$$\begin{aligned} \mu_{\mathrm{S}} \geq \mu(N{+}1;\{V_i\}) &\geq \mu_{\mathrm{D}} \quad (V_{\mathrm{DS}} > 0) \quad \text{or} \\ \mu_{\mathrm{D}} \geq \mu(N{+}1;\{V_i\}) &\geq \mu_{\mathrm{S}} \quad (V_{\mathrm{DS}} < 0) \end{aligned} \tag{21}$$

has to be fulfilled. Condition (20) defines for diverse l and k (l' and k') threshold lines for additional transport channels in the V_{GS} versus V_{DS} plane. Fulfilling this requirement, the transition $|N+1;k\rangle \to |N;l\rangle$ ($|N;l'\rangle \to |N+1;k'\rangle$) might be usable for transport at these $\{V_i\}$ if the initial state $|N+1;k\rangle$ ($|N;l'\rangle$) for this transition is reached regularly via other transitions. It leads to the pattern depicted in Fig. 9b.

With decreasing the size of a quantum dot, the single-particle eigenenergy spacing $\Delta\varepsilon = \varepsilon_i - \varepsilon_j$ increases and might even exceed the electron charging

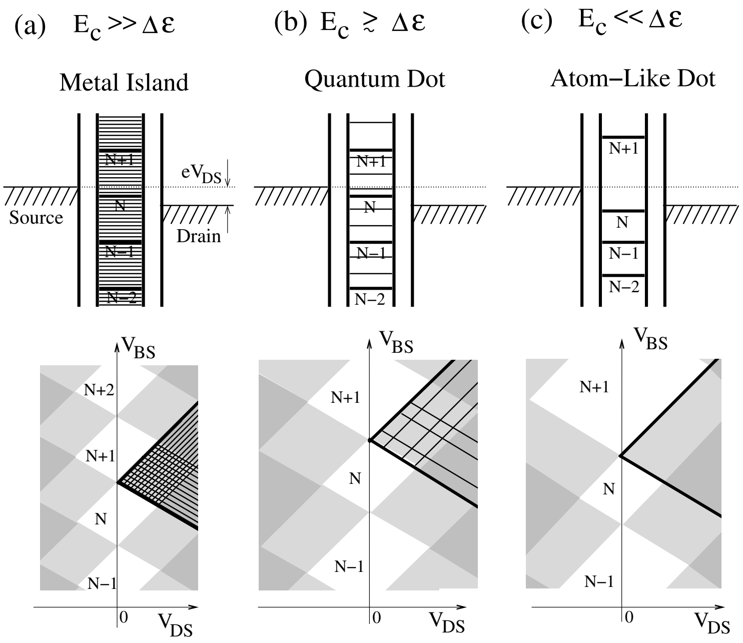

(a) $E_C >> \Delta\varepsilon$
Metal Island
(b) $E_C \gtrsim \Delta\varepsilon$
Quantum Dot
(c) $E_C << \Delta\varepsilon$
Atom–Like Dot
N+1
N
N−1
N−2
eV_{DS}
Source
Drain
V_{BS}
N+2
N+1
N
N−1
0
V_{DS}

Fig. 10. For increasing the ratio $\Delta\varepsilon/E_C$, the energy level scheme shows less transition energies. Therefore less additional transport channels due to (single-particle-) excitations of the quantum dot are expected in the single-electron tunneling regime.

energy E_C due to the electron-electron interaction on the quantum dot: The single-electron charging energy – being a consequence of the unscreened electron-electron interaction on the island – scales like $E_C \propto 1/\epsilon D$ with the island diameter D. The level spacing in a parabolic confining potential (taken as the simplest example) scales like $\Delta\varepsilon = \hbar\,\omega_0 = h^2/(2\,m\,D^2)$. As shown in Fig. 10, with increasing ratio $\Delta\varepsilon/E_C$, the Coulomb blockade regions in the (V_{GS}, V_{DS}) plane vary more and more in size with the electron number, and a less number of additional channels due to transitions to excited states occur in the single-electron tunneling regime.[3]

In a first approach, the dynamics of electron transport can be described by tunneling rates included in a master equation ansatz. The rate is proportional to the tunneling probability for an electron leading to the transition $|N+1;k\rangle \rightarrow |N;l\rangle$ ($|N;l'\rangle \rightarrow |N+1;k'\rangle$). Obviously such a transition is weighted by the strength of the spatial overlap of the wavefunction of the quantum dot and the respective reservoir. However, such a transition might also obey certain (quasi-)selection rules due to spin conservation or correlation effects of the n-electron system in the quantum dot [27–31]. Therefore, the properties of the $N+1$ and N-electron state are of importance. It might even occur that the occupation of

[3] This should be understood as a trend. Indeed, low lying excitations might be possible in a correlated electron system.

certain excited states blocks the electron transport through the quantum dot [24] – as visible by the negative differential conductance in Fig. 7d.

7 Summarizing the Conditions for Coulomb Blockade

To summarize, the Coulomb blockade effect is observable in electrical transport through small islands if

- the single-electron charging energy exceeds significantly the thermal energy,

$$\frac{\mu(N{+}1;\{V_i\}) - \mu(N;\{V_i\})}{2} \gg k_\mathrm{B}T \qquad (E_\mathrm{C} \gg k_\mathrm{B}T)\ , \tag{22}$$

- the applied drain-source voltage V_DS is not too large,

$$e\,|V_\mathrm{DS}| < \mu(N{+}1;\{V_i\}) - \mu(N;\{V_i\}) \qquad (e\,|V_\mathrm{DS}| < 2\,E_\mathrm{C})\ , \tag{23}$$

- the tunnel coupling to the leads is small, i.e., the island can be considered as (quasi-)isolated. Due to Heisenberg's uncertainty relation, the dwell time τ_H of an electron on the island has to be so long that the uncertainty $\varDelta\varepsilon_\mathrm{H} \approx h/\tau_\mathrm{H}$ for the energy of an electron on the island does not exceed the single-electron charging energy, i.e.,

$$\tau_\mathrm{H} > \frac{2\,h}{\mu(N{+}1;\{V_i\}) - \mu(N;\{V_i\})} \qquad (\tau_\mathrm{H} > h/E_\mathrm{C})\ . \tag{24}$$

  This is usually achieved if the tunnel barriers to the lead electrodes have a conductance which is much less than $e^2/h \approx (26\,\mathrm{k}\Omega)^{-1}$ – the conductance of a ballistic (one-mode) one-dimensional channel.

Since the Coulomb blockade is based on an electrostatic effect, Coulomb blockade and single-electron charging effect can be observed for tunneling through quasi-isolated

- mesoscopic metal islands,
- mesoscopic superconducting islands,
- mesoscopic quantum dots,
- molecules and atom clusters, and
- bounded electron states to impurities.

Several examples will be given in the course of this school, for instance, single-electron transistors containing a carbon nanotube as the island.

Depending on the confined electron number, size and effective mass of the electrons, quantum dots resemble in one limit metal-like islands, in the other limit they mimic atom-like properties. Furthermore, the electronic structure of quantum dots can be affected by an applied magnetic field which allows to study the character and degeneracy of electronic states and its influence on electrical transport. Due to their tunability, such quantum dot systems have been used as model systems for investigating interacting N-electron systems and for approaching an understanding of electrical transport through single molecules or single atoms weakly coupled to leads.

8 Some Applications of Single-Electron Transistors

8.1 SET as a Voltage Signal Amplifier

The single-electron transistor can be used to amplify a voltage signal. Biasing the SET with a constant current I_{DS} as shown in Fig. 11, the voltage V_{DS} drops between the source and drain contact which depends in its magnitude on the applied gate voltage V_{GS}. Contour lines of constant current I_{DS} are obtained in the V_{DS} vs. V_{GS} plane parallel to the borderlines defining the different transport regions of the SET as sketched in Fig. 4: A change $\mathrm{d}V_{\mathrm{GS}}$ causes due to (13) the change

$$\mathrm{d}V_{\mathrm{DS}} = -\frac{C_{\mathrm{G}}}{C_{\mathrm{D}}}\,\mathrm{d}V_{\mathrm{GS}} \qquad \text{or} \qquad \mathrm{d}V_{\mathrm{DS}} = \frac{C_{\mathrm{G}}}{C_{\Sigma}-C_{\mathrm{D}}}\,\mathrm{d}V_{\mathrm{GS}}\,. \tag{25}$$

The voltage signal $\mathrm{d}V_{\mathrm{GS}}$ is amplified in $\mathrm{d}V_{\mathrm{DS}}$ if

$$\left|\frac{\mathrm{d}V_{\mathrm{DS}}}{\mathrm{d}V_{\mathrm{GS}}}\right|_{I_{\mathrm{DS}}=\mathrm{const}} > 1\,, \tag{26}$$

i.e., *voltage gain* is present. For the SET this can only be obtained for the gate voltage regime where the first relation of (25) is valid. That means $C_{\mathrm{G}} > C_{\mathrm{D}}$ [32]. Thus the capacitive coupling of the SET island to the gate electrode where the voltage signal is applied has to be chosen larger than the capacitive coupling to the drain electrode where the output voltage $\mathrm{d}V_{\mathrm{DS}}$ arises. The same can be expressed more general in other words: The SET has to be designed in such a way that the electron charge added to the island induces a larger fraction α_{G} of its image charge on the gate electrode than α_{D} on the drain electrode,

$$\alpha_{\mathrm{G}} > \alpha_{\mathrm{D}} \qquad \text{where} \quad \alpha_{\mathrm{G}} = \frac{C_{\mathrm{G}}}{C_{\Sigma}} \quad \text{and} \quad \alpha_{\mathrm{D}} = \frac{C_{\mathrm{D}}}{C_{\Sigma}} \quad \text{for metal SETs.} \tag{27}$$

This is at least required to obtain a voltage gain described by relation (26).

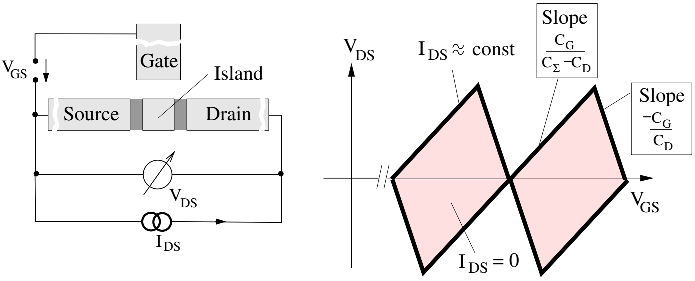


Fig. 11. SET as voltage signal amplifier.

8.2 SET as an Electrometer Sensitive to a Fraction of the Elementary Charge

The electrostatic potential of electrons on the SET island might not only be changed by voltages applied to adjacent electrodes, but also by putting a charge close to the SET island. As sketched in Fig. 12, adding a negative (positive) charge Q shifts the CBO characteristic towards positive (negative) values of V_{GS}. How sensitive is the single-electron transistor to charges? If the charge $Q = \pm e$ would be added directly to the island, then the CBO characteristic is shifted by one period along the gate voltage axis. In this sense, the SET is a highly sensitive electrometer which is even able to detect easily a fraction of the elementary charge e by the change in its characteristics if the charge is added closely to the island [33]. SETs have been demonstated as electrometers with a charge sensitivity down to $8 \cdot 10^{-8}\, e/\sqrt{\mathrm{Hz}}$ at 10 Hz [34]. Incorporating the SET into a radio-frequency resonance circuit – denoted as *RF-SET* [35] – fast charge fluctuations are detectable ($1.2 \cdot 10^{-5}\, e/\sqrt{\mathrm{Hz}}$ at 1.1 MHz). This high charge sensitivity offers on one hand a *ultrasensitive electrometer*, on the other hand it is a disadvantage for applications where a stable and reproducible SET characteristic is required for a large number of SET devices – like in very-large scale integration (VLSI) of digital circuits. Telegraph noise due to charge fluctuations in the SET surroundings makes them almost useless for this purpose.

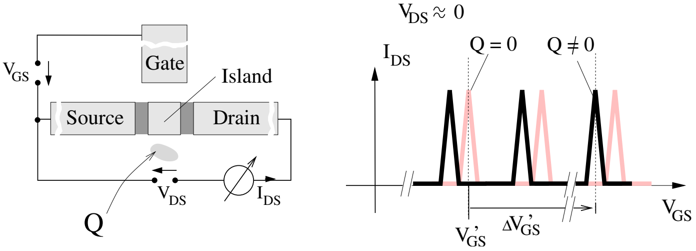


Fig. 12. SET as ultrasensitive electrometer.

8.3 SET as an Electrostatic Sensor in a Scanning Probe Microscope

The sensitivity of a single-electron transistor to the electrostatic environment can be used to measure chemical potential variations of conducting materials affected by external parameters [36]. A SET can even be incorporated into a scanning probe microscope [37]: As sketched in Fig. 13, a SET is fabricated on a microscopic glass tip which is then scanned over a substrate. Monitoring the changes in the SET characteristics as a function of position, the SET can be used as a local probe for the local electrostatic potential variations along the substrate surface. With reducing the distance d between SET and substrate, the capacitance between substrate and SET island reduces roughly like $1/d$. Therefore the CBOs, observed as a function of the voltage applied to the substrate,

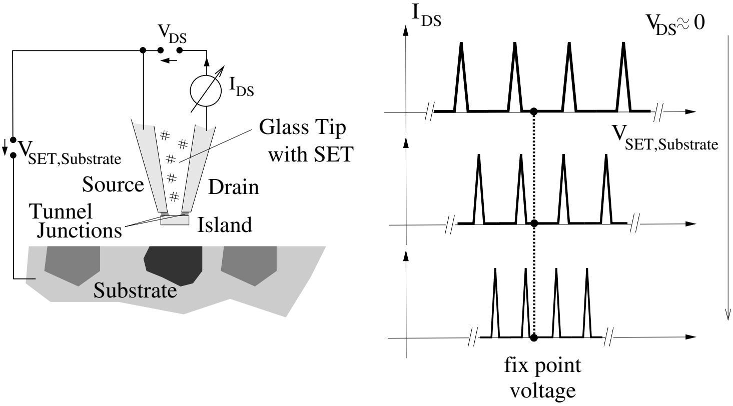


Fig. 13. SET as electrostatic sensor on a tip of a scanning probe microscope.

decrease in their periodicity, squeezing to a fix point on the substrate-SET voltage axis just compensating for the instrinsic contact voltage between SET and substrate. Such an SET on a scanning tip can be considered as an alternative to a scanning force microscope running in the Kelvin probe mode [38] where the local electrostatic force between tip and substrate is minimized by tuning the substrate-tip voltage.

8.4 SET as a Current Rectifier

As shown in Fig. 4, the capacitance ratios $-C_{\mathrm{D}}/C_{\mathrm{G}}$ and $(C_{\Sigma} - C_{\mathrm{D}})/C_{\mathrm{G}}$ are responsible for the slopes of the boundary lines between Coulomb blockade and single-electron transport regions in the V_{GS} vs. V_{DS} plane. Therefore, threshold values $V_{\mathrm{DS}}^{(\mathrm{th})}$ at fixed V_{GS} lie usually asymmetrically with respect to $V_{\mathrm{DS}} = 0$. Therefore, SETs display a non-linear $I_{\mathrm{DS}}(V_{\mathrm{DS}})$ characteristics where the asymmetry of the characteristics is tunable by V_{GS}. Due to the non-linearity of such devices around $V_{\mathrm{DS}} = 0$, frequency mixing of ac voltage signals is possible around $V_{\mathrm{DS}} = 0$. Especially a rectification process can occur: An applied ac bias voltage $V_{\mathrm{DS}}(t)$ results in a time-averaged net dc current [39]. Depending on the ratio $C_{\mathrm{D}}/C_{\Sigma}$, three different behaviours are expected (see Fig. 14): In the case of $C_{\mathrm{D}}/C_{\Sigma} > \frac{1}{2}$, for a fixed ac bias modulation with $|V_{\mathrm{DS}}(t)| \ll e/C_{\Sigma}$, the sequence in the dc current polarity is zero/positive/negative/zero with increasing V_{GS} from one Coulomb blockade region to the next. In the case $C_{\mathrm{D}}/C_{\Sigma} < \frac{1}{2}$ the sequence is zero/negative/positive/zero. Only in the case $C_{\mathrm{D}}/C_{\Sigma} = \frac{1}{2}$, the net current is basically zero over the whole V_{GS} range.

9 The SET for Very-Large Scale Integration (VLSI) of Digital Circuits?

Carrying the current by electrons passing the island one-by-one and being switched on and off by the elementary charge, the single-electron transistor can

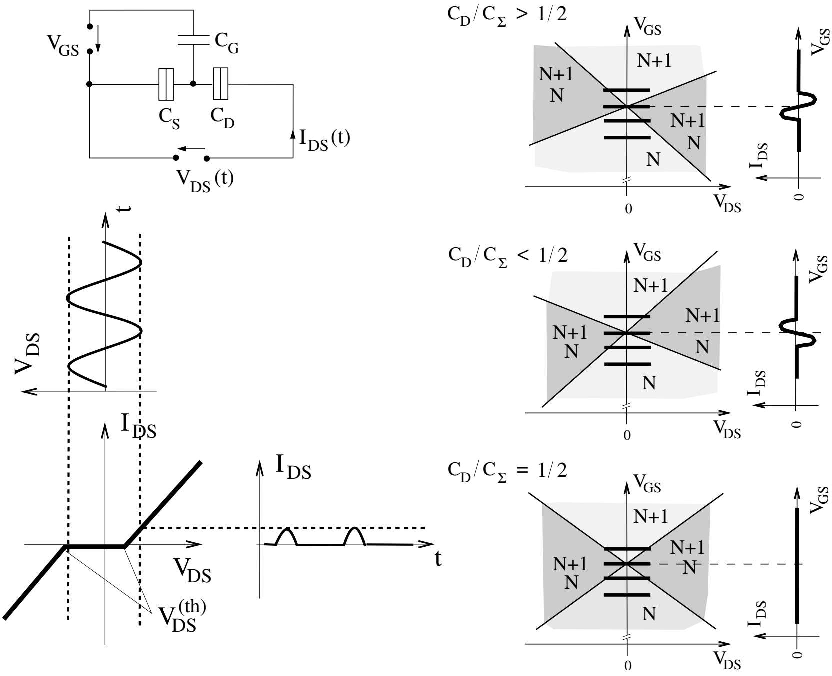
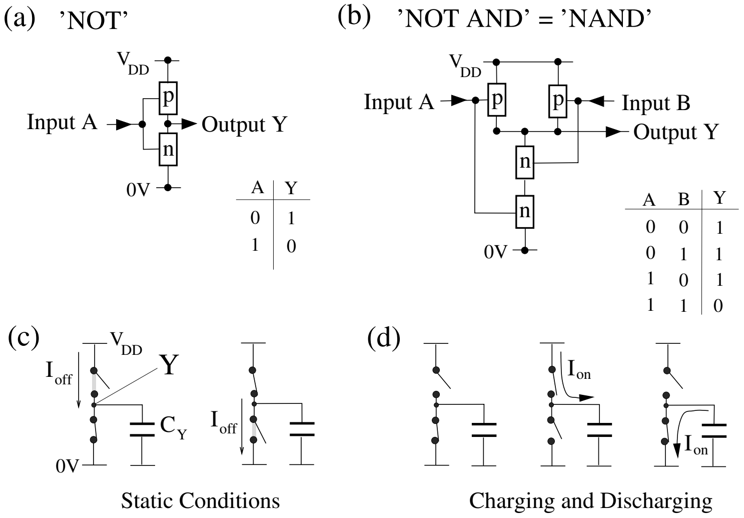


Fig. 14. SET as a potential-controlled current rectifier.

be considered as the ultimate transistor. Dealing with the smallest amount of charge, it has been suggested with presenting the concept of a SET in the mid 1980´s that integrated circuits based on SETs would lead to lowest power consumption.

It was already pointed out, the sensitivity of a SET on single-electron charge fluctuations is a strong disadvantage in this context [40]. Despite of this, the question arises: Is the SET conceptionally a severe candidate for replacing the MOSFET (Metal-Oxide-Semiconductor Field-Effect Transistor) which is used nowadays as the electronic switch in digital circuits? Both transistor concepts belong to the same class of electrostatically controlled switches and obey therefore the same electrostatic requirements for being a good switch for this application. The answer is basically ‘no’ [41,40] which will be further explained in the following.

The overall power dissipation is a severe problem of nowadays microprocessor chips. The only known concept for logical circuits, fulfilling the requirement of reliable computation [42] and thereby strongly suppressing the standby power dissipation, is based on two complementary working switches (see Fig. 15). It has lead to what is known as CMOS technology. Single-electron transistors can be biased to different working points and then act complementary (one turns on and the other off, controlled by the same voltage signal) [43]. However the circuit concept requires that the transistors have voltage gain. This is hardly to achieve for a single-electron transistor working at room temperature: The island



A	Y
0	1
1	0

A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

Fig. 15. (a), (b) Circuits for logic gates using complementary working switches 'n' and 'p' . (c) To keep the stand-by power dissipation small under static condition, the leakage current I_{off} has to be small. (d) Fast charging and discharging of the output node requires a large I_{on}.

size has to be only few nanometers to reach the high single-electron charging energy, and at the same time the island has to be coupled capacitively stronger to the gate electrode than to the leads ($\alpha_{\text{G}} > \alpha_{\text{D}}$)!

The voltage swing ΔV defines the difference in the voltage levels representing logic '0' and '1'. These are almost given by the positive and negative supply terminals denoted by '0' and 'V_{DD}' in Fig. 15. The voltage $\mathit{\Delta V}$ drops as the drain-source voltage over the transistor (see Fig. 15c and d): The 'on'-current driven through the transistor determines the speed by which the logic gates can switch. The 'off'-current is a leakage causing power dissipation even when the circuit is not doing useful computation (static condition). VLSI requires typically $I_{\text{on}}/I_{\text{off}} > 10^8$ for fulfilling the required performance.

A switch based on tuning an energy barrier electrostatically via a gate voltage leads to the superior characteristic

$$\frac{I_{\text{on}}}{I_{\text{off}}} = \exp \frac{\alpha_{\text{G}}\, e\, \Delta V_{\text{GS}}}{k_{\text{B}} T} \; . \tag{28}$$

The ratio between 'on' and 'off' current depends exponentially on the gate voltage swing ΔV_{GS} which is at the same time ΔV – the difference between the voltage levels representing the logic '0' and '1' state. The quantity α_{G} is limited by $0 \leq \alpha_{\text{G}} \leq 1$ and gives the fraction of image charge which is induced on the controlling gate electrode by a charge in the channel of the electrostatic switch.

MOSFETs offer such an exponential characteristic where α_{G} is close to one. Actually this electrostatic requirement ($\alpha_{\text{G}} \to 1$) is mainly the reason why MOSFET have to shrink in *all* spatial dimensions, and therefore the gate oxide of a

0.1 μm MOSFET has been reduced already to 4 nm thickness! For SETs working at room temperature, again, the request on the electrostatics α_{G} close to one is hardly to achieve.

MOSFETs offer for the 'on'-current 0.5 mA per μm channel width – a value which has remained constant over the last decades. Conceptionally, SETs are limited in their capability in driving a current since electrons are passing the island one-by-one. To have a large I_{on}, the dwell time of an electron on the island has to be short. Therefore, the tunnel coupling has to be enhanced which leads to a stronger leakage I_{off} in the 'off'-state. The ratio $I_{\mathrm{on}}/I_{\mathrm{off}}$ cannot follow an exponential dependence on the gate voltage which make SETs worse: For a certain 'on'-current – required for recharging the connections and the inputs of the following logic gates –, the 'off'-current gets too high. This might be compensated by increasing ΔV which again requires that the single-electron charging energy is enlarged, i.e., the island size has to be shrinked even more. We have to state [41]: Single-electron transistor circuits cannot fulfill the expectation of low power dissipation at reasonable speed performance.

Note, these electrostatic constraints are also valid for using molecules as islands as long as their switching mechanism is purely based on tuning an energy barrier electrostatically. In conclusion, to overcome the severe problems of VLSI, either new circuit design concepts are required – which have not been invented up to now – or a switch has to be found which offers $\alpha_{\mathrm{G}} > 1$ in relation (28). Here is indeed potential for molecules if the switching of the electrical path is controlled by the conformation change of the molecule, induced by an applied electrical field.

10 Charge-Stability Diagram of Two-Island Devices

Up to now we have considered only devices with one island embedded between electrodes of defined electrostatic potentials. Examples for two-island arrangements are depicted in Fig. 16. Both islands are directly or indirectly connected via tunnel barriers to electrodes. Without a capacitive coupling, the islands do not feel each other. Therefore, in the ideal case, two gate electrode can be used to control independently the charge state of the two islands. As a function of the two gate voltages $V_{\mathrm{G1,S}}$ and $V_{\mathrm{G2,S}}$, the charge configuration of the two-island arrangement is stable within rectangular regions (indicated by dashed lines in Fig. 16). Allowing capacitive interaction between both islands, the gate voltage variations shifts the electrostatic potentials of *both* islands, and the charge states of the islands affect each other. The charge stability diagram divides under such a capacitive coupling between the islands into a *honeycomb pattern* as depicted in Fig. 16.

All the two-island arrangements depicted in Fig. 16 have this charge stability diagram. Which of these borderlines between the stable regions are actually seen in electrical transport depends on how source and drain electrodes are connected. For the arrangement (I), for instance, only the triple points are visible.

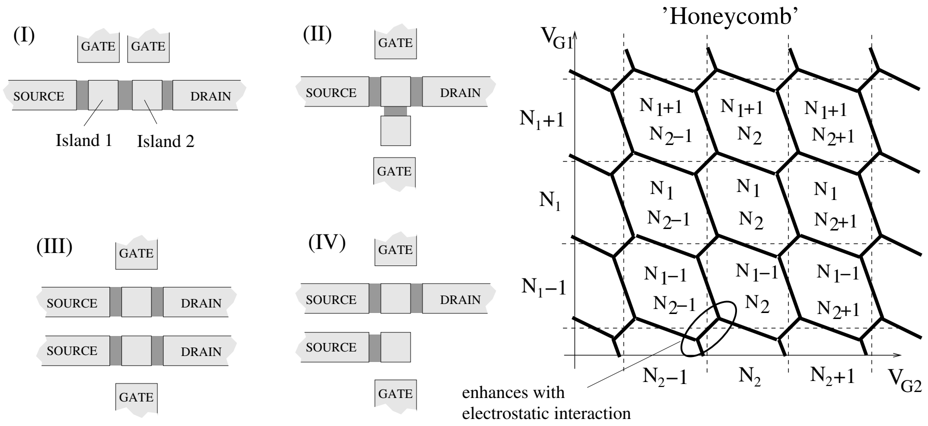


Fig. 16. Charge stability diagram valid for the two-island arrangements (I) to (IV) for $V_{\mathrm{DS}} = 0$ – denoted as 'honeycomb' pattern.

By using quantum dots as islands, molecule-like states can be formed [44,45] by increasing the tunnel coupling between these 'artifical atoms'. The charge stability diagram pattern deviates at the triple points.

11 Single-Electron Turnstile and Single-Electron Pump

Having control over single electrons, why not creating a device which transfers a single electron within a cycle – controlled by external ac voltage signals – from source to drain? The current passing such a device is determined by the cycle frequency f,

$$I_{\mathrm{DS}} = e\,f\ . \tag{29}$$

Such a devices would allow to define a *current standard* and to close the *quantum metrological triangle* [13,1] depicted in Fig. 17a: Three basic physical quantities – current I, voltage V and frequency f – are linked by three fundamental effects – the Jospshon effect connects V with f, the quantum Hall effect V with I, and perhaps a single-electron device obeying (29) connects I with f. Closing this triangle would allow to represent their units with higher precision and even to check whether the fundamental relations given in Fig. 17 are indeed valid.

One version of such a single-electron device is sketched in Fig. 17b denoted as *single-electron turnstile*: The tunnel barriers of a single-electron transistor are tuned similarly to the cycle which the gates of a water lock have to follow to transfer a ship between two water levels through the lock. The Coulomb blockade effect ensures that the island is charged each cycle only with one electron. Such a turnstile with tunable tunnel barriers has been realized by using a split-gate quantum dot [46].

Another version of such a single-electron device obeying (29) is shown in Fig. 17a: By changing the gate voltages in time in the way sketched in Fig. 17c,

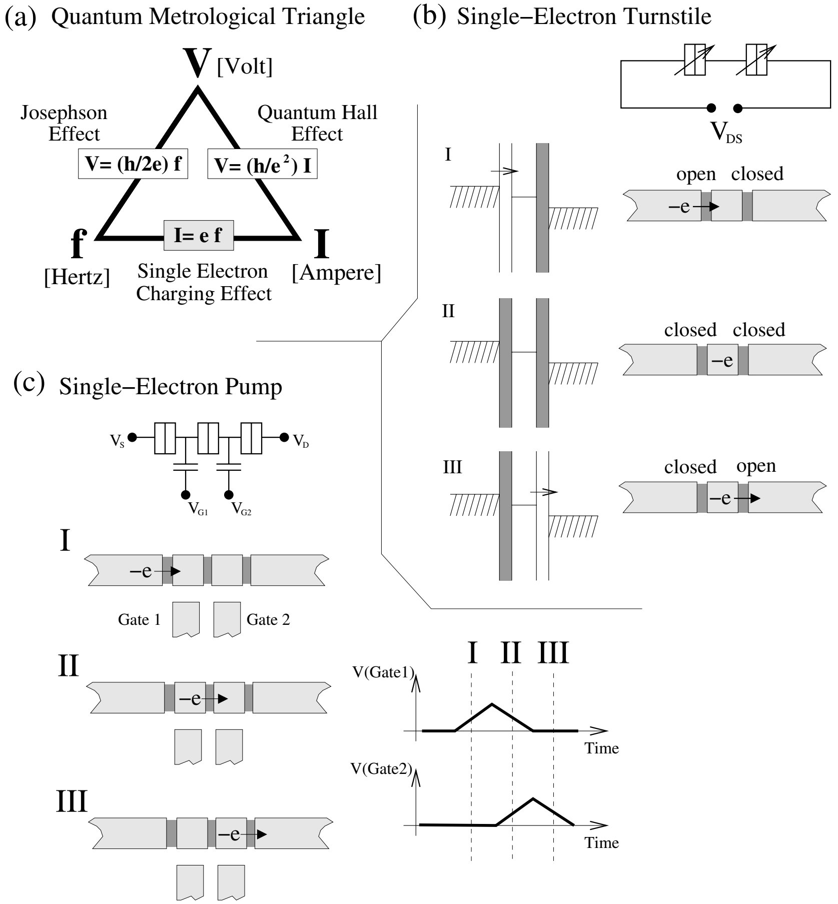


Fig. 17. (a) Quantum Metrological Triangle. (b) Single-electron turnstile. (c) Single-electron pump.

one electron is transferred within such a cycle via the islands from source and drain. These phase-locked variations of the gate voltages decribe a path which encircles one triple point in the charge stability diagram of Fig. 16. The two-terminal arrangement of Fig. 17c behaves as a *single-electron pump* [47].

Two islands are enough to perform single-electron pumping. However, several islands in series are required to obtain a high accuracy: Correlated tunneling (co-tunneling) of electrons through the device has to be suppressed because such processes lead to a leakage. Correlated electron tunneling is the topic of section 13. An accuracy of $\Delta I_{\mathrm{DS}}/I_{\mathrm{DS}} \approx 10^{-8}$ has been achieved [48] in single-electron pumps with seven islands in series, i.e., one electron is missed within 10^8 cycles. Unfortunately the current which is driven through a single pump is too small (f about few MHz) for allowing to close the quantum metrological triangle.

Another approach [49,50] uses surface acoustic waves (SAW) to confine electrons which then have to pass – traveling with this SAW – a small contriction. In another proposal, a certain amount of electrons is shuttled mechanically between source and drain [51].

12 Single-Electron Devices as Primary Thermometer

One-dimensional arrays of M small metal islands of almost same size and tunnel junctions offer at low temperature a pronounced nonlinear $I_{\mathrm{DS}}(V_{\mathrm{DS}})$ characteristic which is rather similar to the one of the single-island arrangement shown in Fig. 2. With increasing the temperature to $T > E_{\mathrm{C}}/k_{\mathrm{B}}$, thermal fluctuations diminish the Coulomb blockade effect and the $I_{\mathrm{DS}}(V_{\mathrm{DS}})$ characteristic becomes more and more linear with increasing T. The deviation is still seen close to $V_{\mathrm{DS}} = 0$ which is better resolved by measuring the differential conductance $\mathrm{d}I_{\mathrm{DS}}/\mathrm{d}V_{\mathrm{DS}}$ as a function of V_{DS}: As shown in Fig. 18, a dip is visible around $V_{\mathrm{DS}} = 0$. Based on rate equations it can be shown [52,53] that the depth of the dip scales like $E_{\mathrm{C}}/3k_{\mathrm{B}}T$, whereas the full-width $V_{1/2}$ at half of the dip depth is described by

$$\frac{e\,V_{1/2}}{(M+1)\,k_{\mathrm{B}}T} = 5.439\cdots \,. \tag{30}$$

This allows to use such an array as a primary thermometer since $V_{1/2}$ does not depend on the device parameters except of the number M of islands. It has turned out that slight variations in the device parameters (island size and tunnel junction) do not significantly affect the validity of (30). Such thermometers are nowadays commercially available products (from Nanoway, Finland). The measurable temperature range depends on the single-electron charging energy E_{C} which can be designed by the junction and island size. Such single-electron devices might be able to replace established temperature standards used at low

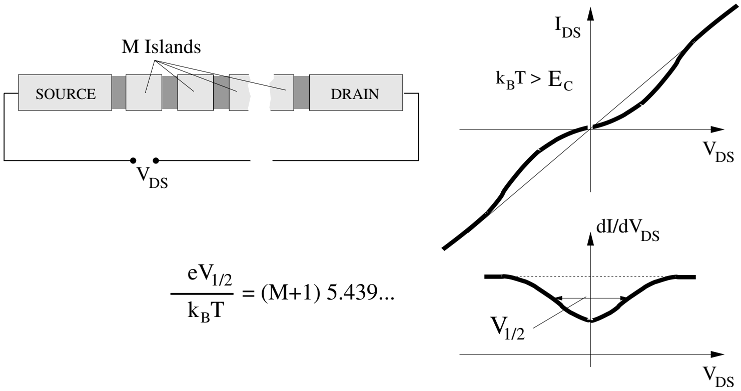


Fig. 18. Primary thermometer.

temperature, i.e., in the range of few milliKelvin to few tens of Kelvin. Two-dimensional arrays of small islands show similar behaviour [53].

13 Breakdown of the Single-Electron Tunneling Picture

In the limit of weak tunnel coupling and at low but finite temperature, the dynamics of single-electron transport is usually described by temperature-dependent rate equations [15,21,54,18] revealing the basic features of Coulomb blockade and single-electron tunneling. By this approach, only processes involving a tunneling event of an electron through one of the barriers are taken into account. This does not work in the case of strong tunnel coupling and – as pointed out in Sect. 14 – sometimes even not in the weak tunnel coupling regime.

Besides thermally induced fluctuations in the number of electrons on the island, quantum fluctuations occur and become stronger with increasing the tunnel coupling to the lead electrodes. Simple examples for this are so-called *co-tunneling* events (Fig. 19) [55]: An electron from one of the leads occupies the island while at the same time another electron leaves the island to one of the leads. Since the charge state on the island is not changed by this *correlated tunneling* event, no single-electron charging energy has to be paid. Even in the Coulomb blockade regime, this leads to a net current flow between source and drain for $|V_{\mathrm{DS}}| > 0$. The charge state of the island is only virtually changed. Under finite V_{DS} bias, the electron system confined in the quantum dot can even be excited by such correlated tunnel processes (*inelastic cotunneling*). Important

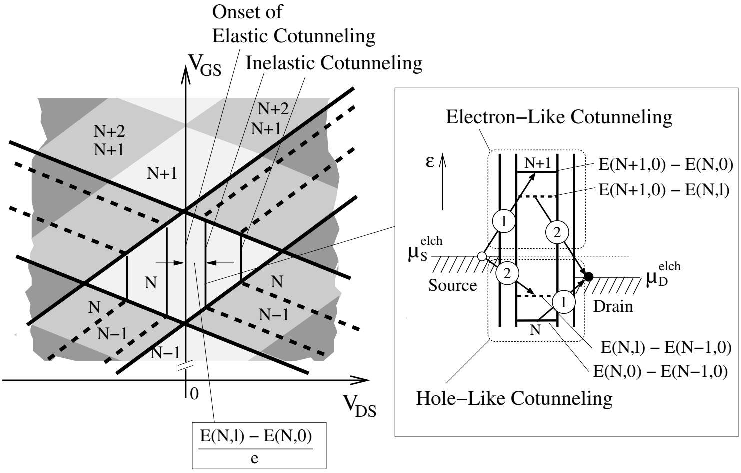


Fig. 19. Cotunneling as the simplest correlated tunneling event: Adding an electron while at the same time an electron leaves the island allows electron transport between source and drain even in the Coulomb blockade regime. Transport channels due to cotunneling open at positions in $|V_{\mathrm{DS}}| > 0$ (independent of V_{GS}) which are given by the energy difference leading to an excitation of electron system confined in the quantum dot. Such an excitation in the quantum dot can also be taken away by cotunneling.

to note, transport channels due to correlated tunneling are opened at certain threshold values of V_{DS}, independent of V_{GS} (see Fig. 19). This distinguishes them from transport channels opened for single-electron transport. Opening such a cotunneling channel leads to a *step-like* change in the differential conductance $\mathrm{d}I_{\mathrm{DS}}/\mathrm{d}V_{\mathrm{DS}}$ with increasing $|V_{\mathrm{DS}}|$. *Elastic cotunneling*, which uses the transition between the groundstates $|n,0\rangle$ and $|n+1,0\rangle$ as the intermediate transition, can already occur at $V_{\mathrm{DS}} = 0$.

This virtual occupation leads effectively to a broadening of the energy levels depicted in the energy schemes for the quantum dot. Usually these correlated tunneling processes can be treated as a small contribution. However, this is not always true as shown in the following.

14 Kondo Effect in Single Quantum Dot Systems

Figure 20b shows the differential conductance $\mathrm{d}I_{\mathrm{DS}}/\mathrm{d}V_{\mathrm{DS}}$ through a small quantum dot (Fig. 20a) as a function of V_{DS} and V_{GS}. For the case of weak tunnel coupling to both leads, the Coulomb blockade region is well resolved. With increasing the tunnel coupling while keeping the temperature, the Coulomb blockade region is no longer well defined, but the remarkable feature is the appearance of a peak in the differential conductance at $V_{\mathrm{DS}} = 0$ over the whole Coulomb blockade regime [56–59]. It becomes stronger with increasing the tunnel coupling, but disappears with increasing the temperature (Fig. 20c). It means that the quantum dot is highly conductive at low temperature and less conductive at high temperature. Important to note, the position of this zero-bias anomaly remains unaffected by V_{GS}, although the electronic states of the dot are shifted by V_{GS}, which indicates that the island is effectively not charged, i.e., that correlated electron tunneling is here of importance. It has been observed [60] that even the conductance $2e^2/h$ is reached for this zero-bias anomaly. Zero-bias anomalies are not observed for all Coulomb blockade regions, i.e., certain requirements have to be fulfilled.

Predicted in 1988 [61,62] and experimentally demonstrated in 1998 [56], the interpretation of this zero-bias anomaly is based in the simplest case on the so-called Anderson impurity model [63]. The model has been used to describe the Kondo effect observed at low temperature in the resistivity of metal slightly doped with magnetic impurities. The (extended) Anderson impurity model is depicted in Fig. 21: A spin-degenerate localized electron state is tunnel coupled to two electron reservoirs. Its energy lies below the Fermi level of the reservoirs, i.e., it is always occupied by an electron with spin-up or spin-down. Occupation of the localized state by two electrons at the same time is suppressed due to the electron-electron interaction $U = 2E_{\mathrm{C}}$ on the island. Solving this problem, it turns out that correlated electron tunneling of lowest order (cotunneling) is not enough to descibe the transport through such an island: The electronic state of the island hybridizes with the electronic states of the leads forming a spin-singlet state, although the energy level of this localized state is deep below the Fermi level of the reservoirs. At low temperature, even a small tunnel coupling to the

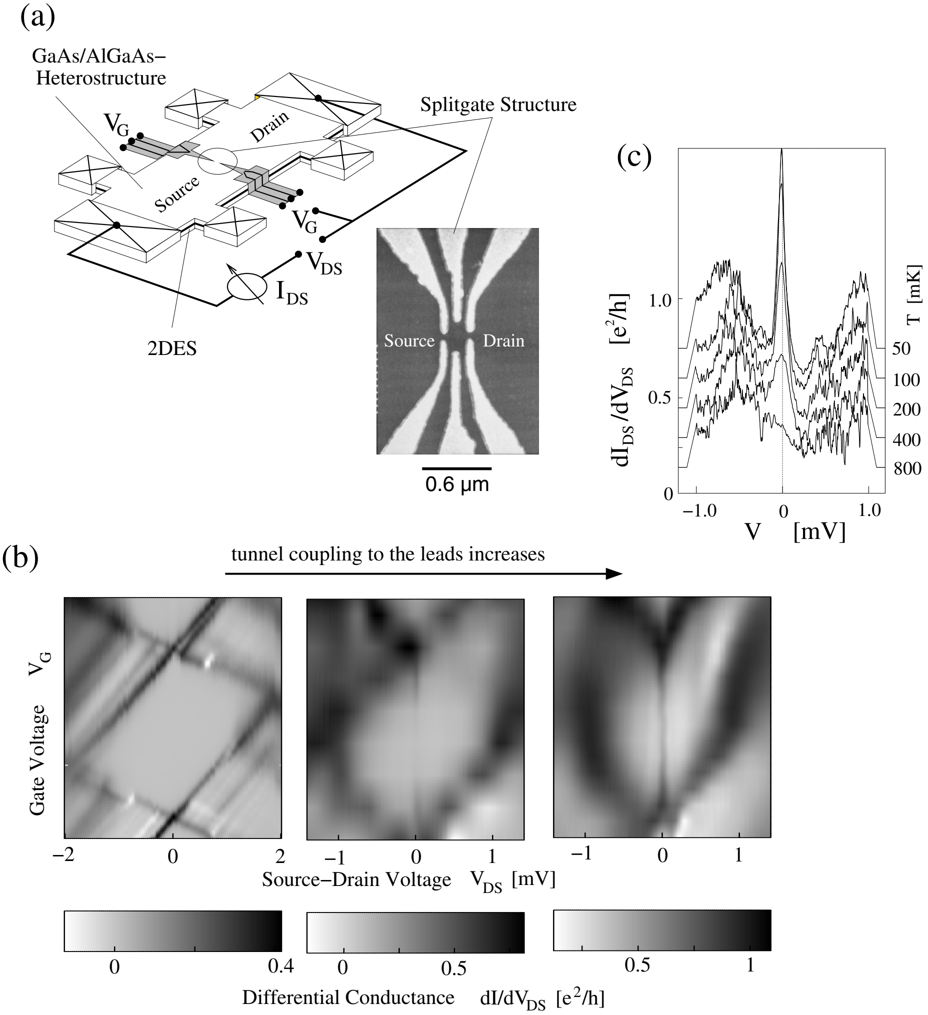


Fig. 20. (a) Sketch of the experimental arrangement of a single quantum dot defined in a two-dimensional electron system by electrostatic depletion. (b) Differential conductance as a function of the drain-source voltage and the gate voltage for different tunnel coupling to the leads. A zero-bias anomaly – identified as a Kondo peak – develops at $V_{\mathrm{DS}} = 0$ within the Coulomb blockade region. (c) Temperature dependence of the Kondo peak taken in the middle of a Coulomb blockade region (from another sample). (from J. Schmid et al., MPI-FKF)

leads causes correlated tunneling of electrons permanently flipping the spin state of the island. This leads to an effective density of state on the site of the impurity pinned to the Fermi level of the reservoirs (see Fig. 21). Electron transport is possible around $V_{\mathrm{DS}} = 0$. The weaker the tunnel coupling and the deeper the impurity level, the lower the temperature has to be to observe this Kondo effect.

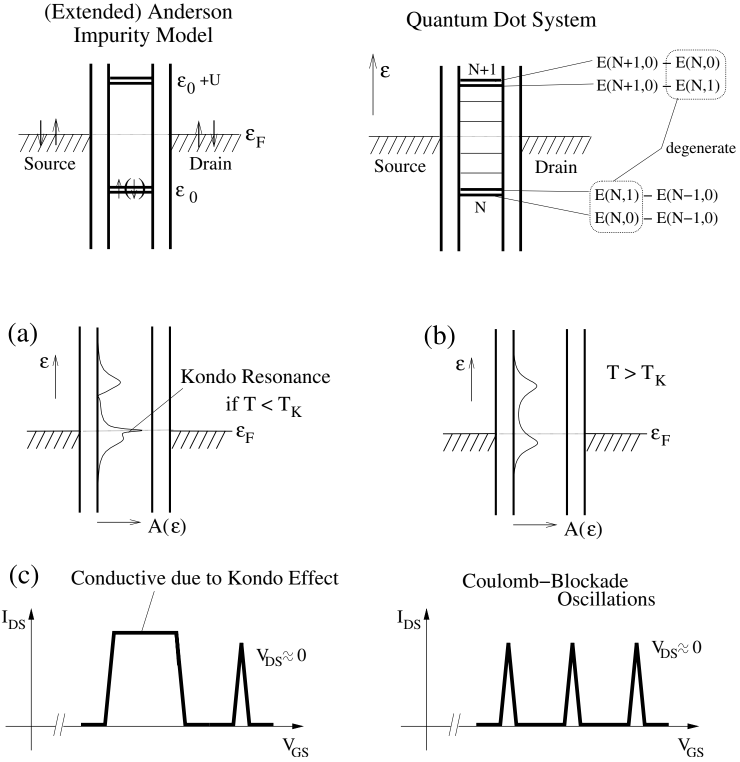


Fig. 21. The Anderson impurity model in comparison to the energy scheme of a quantum dot system with (spin-)degenerate groundstate. (a) Solving the model for low temperature, a resonance at the Fermi level is found which disappears at higher temperature (see (b)). (c) At low temperature, the Coulomb blockade disappears in the respective CBO valley but recovers at higher temperature.

The reference scale is given by the so-called *Kondo temperature* T_K

$$k_\mathrm{B} T_\mathrm{K} = \frac{\sqrt{\Gamma U}}{2} \exp\left[-\frac{\pi \left(\varepsilon_\mathrm{F} - \varepsilon_0\right)\left(U + \varepsilon_0 - \varepsilon_\mathrm{F}\right)}{\Gamma U} \right] \tag{31}$$

where the energy Γ describes the broading of the energy level due to the tunnel coupling of the impurity (quantum dot) state to the leads, and $\varepsilon_\mathrm{F} - \varepsilon_0$ the energetical distance of the level on the impurity site to the Fermi level of the reservoirs. A large U – basically given by the electron-electron interaction – and large Γ enlarges the Kondo temperature, i.e., the Kondo effect is observed at higher temperature.

Magnetic field dependent measurements reveal that spin-degeneracy usually is responsible for the Kondo effect in quantum dot systems. Suggested by the Constant Interaction Model, at the beginning the Kondo effect has been expected only for an odd number of electron on the quantum dot (*odd-even parity effect*).

However, it can also be observed for even electron numbers [59,64,65]. The electronic structure of a quantum dot is more complex than assumed by the CIM.

15 Two Electrostatically Coupled Single-Electron Transistors: More than the Sum of Two

The Anderson impurity model describes two separate electron systems labeled by an index which is usually identified with the spin quantum number (see Fig. 22a). The only interaction between both 'spin' electron systems happens on the impurity (quantum dot) site: Occupation by two electrons at the same time is suppressed due to the Coulomb interaction on this site. Interpreting the 'spin' index of the Anderson impurity model as the index distinguishing between two spatially separated electron systems, another realization of the Anderson impurity model becomes feasible [66]: a system consisting of electrostatically coupled quantum dots with separate leads to each quantum dot (see Fig. 22a). The mapping works [66] if (1) an energetical degeneracy is present in occupying either the upper or the lower quantum dot, (2) the groundstate of each quantum dot is not degenerate, excited states are energetically well separated.

An experimental setup to implement this arrangement is shown in Fig. 22b: By etching the pattern shown as an SEM image into a GaAs-AlGaAs heterostructure containing two 2DESs separated by a insulating 40 nm thick AlGaAs barrier, two strongly electrostatically coupled quantum dots are formed. By alloying metal contacts and by using top and back gates for locally depleting the upper or lower 2DES, the quantum dots are separately contacted.

In Fig. 22c, the conductance through the upper quantum dot is shown as a function of the gate voltages $V_{1,2}$ and V_{G} (see Fig. 22b). A honeycomb-like structure is visible which reflects strong electrostatic interaction between both quantum dots. Along the lines marked by 'a', single-electron tunneling occurs through the upper quantum dot. Along the lines marked by 'c', single-electron fluctuations are possible for the lower quantum dot, but not visible in the current through the upper quantum dot. Along the lines marked by 'b', current through the upper quantum dot is detected – although not expected within the single-electron tunneling picture for electrostatically coupled quantum dots. Along such lines, an energy degeneracy of having an additional electron either on the upper or lower quantum dot exists – one prerequisite of the Anderson model. Due to the predictions for the Anderson model, we expect to see a peak in the differential conductance versus drain-source voltage at the positions along the lines marked by 'b'. Such a trace taken in the middle of a line 'b' is shown in Fig. 22d. The observed peak indicates [67] that a simple co-tunneling process – adding an electron in the upper quantum dot while at the same time taking off an electron from the lower quantum dot (and vice versa) – is not enough to explain the electron transport. Correlated tunneling processes of higher order have to be taken into account – Kondo physics is present.

In conclusion, closely packed single-electron transistors with atom-like islands might show not only electrostatic interaction but might form also a correlated

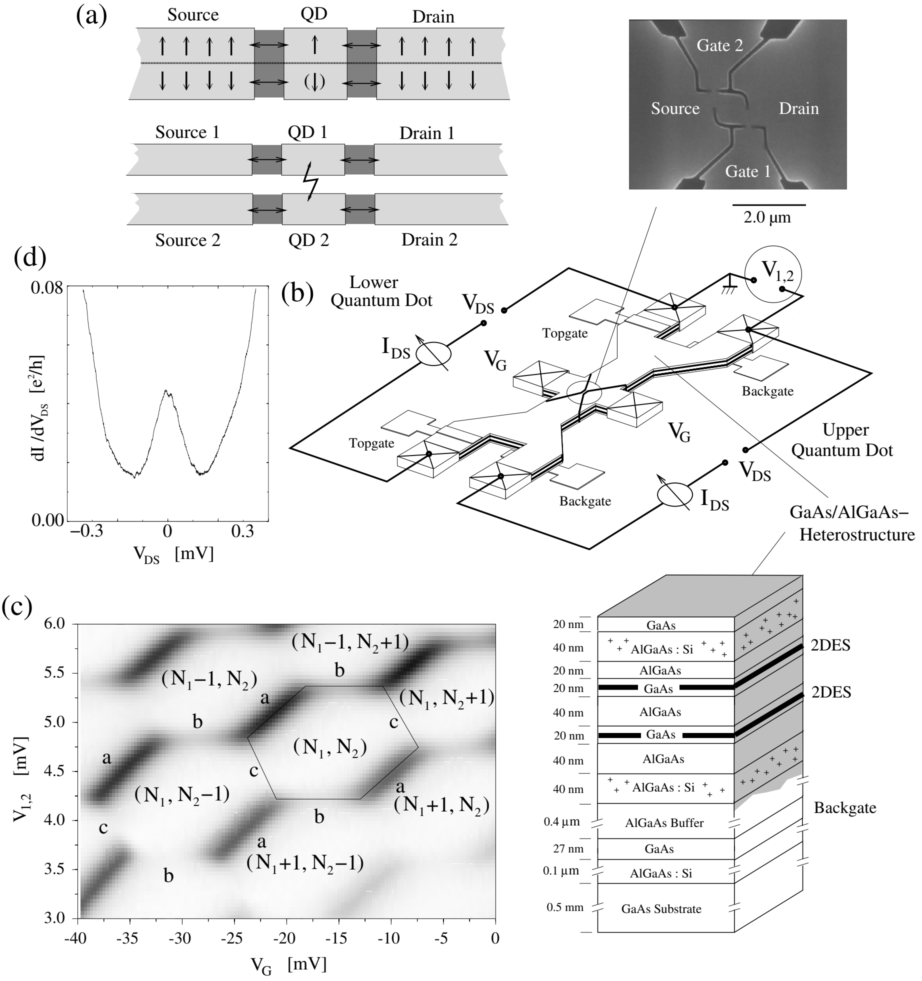


Fig. 22. (a) Scheme of the (extended) Anderson impurity model - top: two systems of different spin orientation, bottom: two spatially separated systems. In both cases, the systems interact only electrostatically on the QD site(s). (b) Sketch of the experimental setup of two quantum dots with separate leads. At top, a scanning electron microscope image of the etched pattern defining two quantum dots on top of each other. (c) Conductance through the upper quantum dot versus two gate voltages. (d) Differential conductance versus drain-source voltage taken in the middle of a line marked by ‘b’ in (c). As expected from the analogy to the Anderson impurity model, a zero-bias anomaly is observed. (from U. Wilhelm et al., MPI-FKF)

quantum mechanical state making them highly conductive in the regime where at higher temperature (beyond the Kondo temperature of the arrangement) Coulomb blockade is observed.

References

1. *Single Charge Tunneling*, volume B 294 of *NATO ASI Series*, ed. by H. Grabert, M.H. Devoret (Plenum Press, New York 1992)
2. K.K Likharev: 'Single-electron devices and their applications'. Proceedings of the IEEE **87**, 606 (1999)
3. U. Meirav, E.B. Foxman: 'Single-electron phenomena in semiconductors'. Semicond. Sci. Technol. **10**, 255 (1995)
4. L.P. Kouwenhoven, Ch.M. Marcus, P.L. McEuen, S. Tarucha, R.M Westerwelt, N.S. Wingreen: 'Electron transport in quantum dots'. In: *Mesoscopic Electron Transport*, ed. by L.L. Sohn et al. (Kluwer Academic Publishers, Dordrecht 1997)
5. L.P. Kouwenhoven, D.G. Austing, S. Tarucha: 'Few-electron quantum dots'. Rep. Prog. Phys. **64**, 701 (2001)
6. T. Chakraborty: *Quantum Dots – A survey of the properties of artificial atoms* (North-Holland, Amsterdam 1999)
7. G. Schön: 'Single-electron tunneling'. In: *Quantum Transport and Dissipation*, ed. by T. Dittrich, P. Hänggi, G. Ingold, G. Kramer, B. Schön, W. Zwerger (VCH, Weinheim 1997) chapter 3
8. H. Schoeller: 'Transport theory of interacting quantum dots'. In: *Mesoscopic Electron Transport*, ed. by L.L. Sohn et al.(Kluwer Academic Publishers, Dordrecht 1997)
9. T.M. Eiles, J.M. Martinis, M.H. Devoret: 'Even-odd asymmetry of a superconductor revealed by the Coulomb blockade of Andreev reflection'. Phys. Rev. Lett. **70**, 1862 (1993)
10. M. Tinkham: *Introduction to Superconductivity* (McGraw-Hill, New York 1996)
11. E.V. Sukhorukov, D. Loss: 'Spintronics and spin-based qubits in quantum dots'. phys. stat. sol. **224**, 855 (2001)
12. Y. Makhlin, G. Schön, A. Shnirman: 'Quantum-state engineering with josephson-junction devices'. Rev. Mod. Phys. **73**, 357 (2001)
13. D.V. Averin, K.K. Likharev: 'Coulomb blockade of single-electron tunneling, and coherent oscillations in small tunnel junctions'. J. Low Temp. Phys. **62**, 345 (1986)
14. K.K. Likharev: 'Single-electron transistors: Electrostatic analogs of the DC SQUIDS'. IEEE Transactions on Magnetics **23**, 1142 (1987)
15. D.A. Averin, A.N. Korotkov, K.K. Likharev: 'Theory of single-electron charging of quantum wells and dots'. Phys. Rev. B **44**, 6199 (1991)
16. T.A. Fulton, G.D. Dolan: 'Observation of single-electron charging effects in small tunnel junctions'. Phys. Rev. Lett. **59**, 109 (1987)
17. L.D. Hallam, J. Weis, P.A. Maksym: 'Screening of the electron-electron interaction by gate electrodes in semiconductor quantum dots'. Phys. Rev. B **53**, 1452 (1996)
18. D. Pfannkuche, S.E. Ulloa: 'Selection rules for spectroscopy of quantum dots'. Advances in Solid State Physics **35**, 65 (1996)
19. M. Kastner: 'Artificial atoms'. Phys. Today **46**, 24 (1993)
20. R.C. Ashoori: 'Electrons in artifical atoms'. Nature **379**, 413 (1996)
21. C.W.J. Beenakker: 'Theory of Coulomb-blockade oscillations in the conductance of a quantum dot'. Phys. Rev. B **44**, 1646 (1991)
22. S.M. Reimann, M. Manninen: 'Electronic structure of quantum dots'. Rev. Mod. Phys. **74**, 1283 (2002)
23. J. Weis, R.J. Haug, K. v. Klitzing, K. Ploog: 'Transport spectroscopy of a single quantum dot'. Semicond. Sci. Technol. **9**, 1890 (1994)
24. J. Weis, R.J. Haug, K. v. Klitzing, K. Ploog: 'Competing channels in single-electron tunneling through a quantum dot'. Phys. Rev. Lett. **71**, 4019 (1993)

25. A.T. Johnson, L.P Kouwenhoven, W. de Jong, N.C. van der Vaart, C.J.P.M. Harmans, C.T. Foxon: ‘Zero-dimensional states and single electron charging in quantum dots’. Phys. Rev. Lett. **69**, 1592 (1992)
26. E.B. Foxman, P.L. McEuen, N.S. Wingreen, Y. Meir, P.A. Belk, N.R. Belk, M.A. Kastner: ‘Effects of quantum levels on transport through a Coulomb island’. Phys. Rev. B **47**, 10020 (1993).
27. J.M. Kinaret, Y. Meir, N.S. Wingreen, P. Lee, X.-G. Wen: ‘Conductance through a quantum dot in the fractional quantum Hall regime’. Phys. Rev. B **45**, 9489 (1992)
28. D. Weinmann, W. Häusler, B. Kramer: ‘Spin blockades in linear and nonlinear transport through quantum dots’. Phys. Rev. Lett. **74**, 984 (1995)
29. J.J. Palacios, L. Martin-Moreno, C. Tejedor: ‘Magnetotunneling through quantum boxes in a strong-correlation regime’. Europhys. Letters **23**, 495 (1993)
30. D. Pfannkuche, S.E. Ulloa: ‘Selection rules for transport excitation spectroscopy of few-electron quantum dots’. Phys. Rev. Lett. **74**, 1194 (1995)
31. K. Jauregui, W. Häusler, D. Weinmann, B. Kramer: ‘Signatures of electron correlations in the transport properties of quantum dots’. Phys. Rev. B **53**, 1713 (1996)
32. G. Zimmerli, R.L. Kautz, J.M. Martinis: ‘Voltage gain in the single-electron transistor’. Appl. Phys. Lett. **61**, 2616 (1992)
33. P. Lafarge, H. Pothier, E.R. Williams, D. Esteve, C. Urbina, M.H. Devoret: ‘Direct observation of macroscopic charge quantization’. Z. Phys. B **85**, 327 (1991)
34. V.A. Krupenin, D.E. Presnov, A.B. Zorin, M.N. Niemeyer: ‘Aluminum single electron transistors with islands isolated from the substrate’. J. Low Temp. Phys. **118**, 287 (2000)
35. R.J. Schoelkopf, P. Wahlgren, A.A. Kozhevnikov, P. Delsing, D.E. Prober: ‘The radio-frequency single-electron transistor (rf-SET): A fast and ultrasensitive electrometer’. Science **280**, 1238 (1998)
36. Y.Y. Wei, J. Weis, K. von Klitzing, K. Eberl: ‘Single-electron transistor as an electrometer measuring chemical potential variations’. Appl. Phys. Lett. **71**, 2514 (1997)
37. M.J. Yoo, T.A. Fulton, H.F. Hess, R.L. Willett, L.N. Dunkelberger, R.J. Chichester, L.N. Pfeiffer, K.W. West: ‘Scanning single-electron transistor microscopy: Imaging individual charges’. Science **276**, 579 (1997)
38. M. Nonnenmacher, M.P. O’Boyle, H.K. Wickramasinghe: ‘Kelvin probe force microscopy’. Appl. Phys. Lett. **58**, 2921 (1991)
39. J. Weis, R.J. Haug, K. von Klitzing, K. Ploog: ‘Single-electron tunneling transistor as a current rectifier with potential-controlled current polarity’. Semicond. Sci. Technol. **10**, 877 (1995)
40. A.N. Korotkov, R.H. Chen, K.K. Likharev: ‘Possible performance of capacitively coupled single-electron transistors in digital circuits’. J. Appl. Phys. **78**, 2520 (1995)
41. J. Weis: Electrical Transport Through Quantum Dot Systems. Habilitationsschrift, Universität Stuttgart, Stuttgart, Germany 2002
42. A.W. Lo: ‘Some thoughts on digital components and circuit techniques’. IRE Trans. on Electronic Computers **10**, 416 (1961)
43. J.R. Tucker: ‘Complementary digital logic based on the ”Coulomb blockade”’. J. Appl. Phys. **72**, 4399 (1992)
44. L. Kouwenhoven: ‘Coupled Quantum Dots as Artifical Molecules’. Science **268**, 1440 (1995)
45. R.H. Blick, D. Pfannkuche, R.J. Haug, K. von Klitzing, K. Eberl: ‘Formation of a Coherent Mode in a Double Quantum Dot’. Phys. Rev. Lett. **80**, 4032 (1998)

46. L.P. Kouwenhoven, A.T. Johnson, N.C. van der Vaart, A. van den Enden, C.J.P.M. Harmans, C.T. Foxon.: 'Quantized current in a quantum dot turnstile'. Z. Phys. B **85**, 381 (1991)
47. H. Pothier, P. Lafarge, C. Urbina, D. Esteve, M.H. Devoret: 'Single-Electron Pump Based on Charging Effects'. Europhysics Letters **17**, 249 (1992)
48. R.L. Kautz, M.W. Keller, J.M. Martinis: 'Leakage and counting errors in a seven-junction electron pump'. Phys. Rev. B **60**, 8199 (1999)
49. V.I. Talyanskii, J.M. Shilton, M. Pepper, C.J.B. Ford, E.H. Linfield, D.A. Ritchie, G.A.C. Jones: 'Single-Electron transport in a one-dimensional channel by Radio Frequencies'. Phys. Rev. B **56**, 15180 (1997)
50. J. Ebbecke, G. Bastian, M. Blöcker, K. Pierz, F.J. Ahlers: 'Enhanced quantized current driven by surface acoustic waves'. Appl. Phys. Lett. **77**, 2601 (2000)
51. A. Erbe, C. Weiss, W. Zwerger, R.H. Blick: 'Nanomechanical Resonator Shuttling Single Electrons at Radio Frequencies'. Phys. Rev. Lett. **87**, 096106 (2001)
52. J.P. Pekola, K.P. Hirvi, J.P. Kauppinen, M.A. Paalanen: 'Thermometry by arrays of tunnel-junctions'. Phys. Rev. Lett. **73**, 2903 (1994)
53. J.P. Pekola, L.J. Taskinen, Sh. Farhangfar: 'One- and two-dimensional tunnel junction arrays in weak Coulomb blockade regime: Absolute accuracy in thermometry'. Appl. Phys. Lett. **76**, 3747 (2000)
54. D. Weinmann, W. Häusler, B. Kramer: 'Transport properties of quantum dots'. Ann. Physik **5**, 652 (1996).
55. D.V. Averin, Y.V. Nazarov: 'Macroscopic quantum tunneling of charge and co-tunneling'. In: *Single Charge Tunneling*, volume B 294 of *NATO ASI Series*, ed. by H. Grabert, M.H. Devoret (Plenum Press, New York, 1992) pp. 217–247
56. D. Goldhaber-Gordon, H. Shtrikman, D. Mahalu, D. Abusch-Magder, U. Meirav, M.A. Kastner: 'Kondo effect in a single-electron transistor'. Nature **391**, 156 (1998)
57. S.M. Cronenwett, T.H. Oosterkamp, L.P. Kouwenhoven: 'A tunable Kondo effect in quantum dots'. Science **281**, 540 (1998)
58. J. Schmid, J. Weis, K. Eberl, K. von Klitzing: 'A quantum dot in the limit of strong coupling to reservoirs'. Physica B **256**, 182 (1998)
59. J. Schmid, J. Weis, K. Eberl, K. von Klitzing: 'Absence of odd-even parity behaviour for Kondo resonances in quantum dots'. Phys. Rev. Lett. **84**, 5824 (2000)
60. W.G. van der Wiel, S. De Franceschi, T. Fujisawa, J.M. Elzerman, S. Tarucha, L.P. Kouwenhoven: 'The Kondo effect in the unitary limit'. Science **289**, 210 (2000)
61. L.I. Glazman, M.É. Raĭkh: 'Resonant Kondo transparency of a barrier with quasilocal impurity states'. JETP Lett. **47**, 453 (1988)
62. T.K. Ng, P.A. Lee: 'On-site Coulomb repulsion and resonant tunneling'. Phys. Rev. Lett. **61**, 1768 (1988)
63. P.W. Anderson: 'Localized magnetic states in metals'. Phys. Rev. **124**, 41 (1961)
64. S. Sasaki, S. De Franceschi, J.M. Elzerman, W.G. van der Wiel, M. Eto, S. Tarucha, L.P. Kouwenhoven: 'Kondo effect in an integer-spin quantum dot'. Nature **405**, 764 (2000)
65. M. Keller, U. Wilhelm, J. Schmid, J. Weis, K. von Klitzing, K. Eberl: 'Quantum dot in high magnetic fields: Correlated tunneling of electrons probes the spin configuration at the edge of the dot'. Phys. Rev. B **64**, 033302 (2001)
66. U. Wilhelm, J. Schmid, J. Weis, K. von Klitzing: 'Two electrostatically coupled quantum dots as a realization of the Anderson impurity model'. Physica E **9**, 625 (2001)
67. U. Wilhelm, J. Schmid, J. Weis, K. von Klitzing: 'Experimental evidence for spinless Kondo effect in two electrostatically coupled quantum dot systems'. Physica E **14**, 385 (2002)

Full Counting Statistics in Quantum Contacts

Wolfgang Belzig

Department of Physics and Astronomy, University of Basel, Klingelbergstr. 82, 4056 Basel, Switzerland

Abstract. Full counting statistics is a fundamentally new concept in quantum transport. After a review of basic statistics theory, we introduce the powerful Green's function approach to full counting statistics. To illustrate the concept we consider a number of examples. For generic two-terminal contacts we show how counting statistics elucidates the common (and different) features of transport between normal and superconducting contacts. Finally, we demonstrate how correlations in multi-terminal structures are naturally included in the formalism.

1 Introduction

The probabilistic interpretation is a fundamental ingredient of quantum mechanics. While the wave function determines the full quantum state of a system and its evolution in time, observable quantities are related to hermitian operators. Expectation values of these operators determine the average value of a large number of identical measurements. However, an individual measurement yields in general a different result. Applying this idea to a current measurement in a quantum conductor, leads directly to the concept of *full counting statistics* (FCS): during a given time interval t_0 a certain number of charges will pass the conductor. To predict the statistical properties of the number of transferred charges we need a probability distribution. The theoretical goal is to find this distribution.

1.1 Overview

In this article we give an introduction to the field of *full counting statistics in mesoscopic electron transport*. We will concentrate on the powerful technique – using Keldysh-Green's functions – which at the same time is also based on microscopic theory. To accomplish this goal we will first review concepts of basic statistics, which are relevant for counting statistics. In the next section we address the microscopic derivation of FCS using Keldysh-Green's functions. In the rest of the article we demonstrate the use of counting statistics in a number of examples, like two-terminal contacts with normal and superconducting leads, diffusive metals and, finally, multi-terminal structures. But first we review briefly the development of the field.

W. Belzig, Full Counting Statistics in Quantum Contacts, Lect. Notes Phys. **658**, 123–143 (2005)
http://www.springerlink.com/

1.2 History

Full counting statistics has its roots in quantum optics [1], where the number statistics of photons is used, e. g., to characterize coherence properties of photon sources. The major step to adopt the concept to mesoscopic electron transport has been undertaken by Levitov and Lesovik [2]. Since then the theory of FCS of charge transport in mesoscopic conductors has advanced substantially, see [3,4]. In [2] it was shown that scattering between uncorrelated Fermi leads with probability T is described by a binomial statistics $P(N) = \binom{M}{N} T^N (1-T)^{M-N}$. Here, $P(N)$ is the probability, that out of $M = 2et_0V/h$ independent attempts N charges are transferred. Furthermore, Levitov and coworkers studied the counting statistics of diffusive conductors [5], time-dependent problems [6] and of a tunnel junction [7]. A theory of full counting statistics based on the powerful Keldysh-Green's function method was initiated by Nazarov [8]. This formulation allows a straightforward generalization to systems containing superconductors [9,10] and multi-terminal structures [11,12]. Classical approaches to FCS were recently put forward for Coulomb blockade systems [13,14], and, for chaotic cavities based on a stochastic path-integral approach [15]. The field of counting statistics in the quantum regime is closely related to the fundamental measuring problem of quantum mechanics, which has been addressed in a number of works [6,16–21]. Expressing the FCS of charge transport by the counting statistics of photons emitted from the conductor provides an interesting alternative to classical counting of electrons [22]. Counting statistics has been addressed by now for many different phenomena

- Andreev contacts [23]
- generic quantum conductors [13,24–26]
- adiabatic quantum pumping [27–30]
- qubit-readout [17,31–33]
- superconducting contacts in equilibrium [9]
- proximity effect structures [10,34–37]
- cross-correlations with normal [38] or superconducting contacts [12,39]
- entangled electron pairs [40,41]
- phonon counting [42]
- relation between photon counting and electron counting [43]
- current biased conductors [44]
- interaction effects: weak and strong Coulomb blockade [14,45,46]
- multiple Andreev reflections in superconducting contacts [47,48].

Very recently, an important experimental step forward was achieved. Reulet, Senzier, and Prober measured for the first time the third cumulant of current fluctuations produced by a tunnel junction [49]. Surprisingly the measured voltage dependence deviated from the expected voltage-independent third cumulant of a simple tunnel contact [2,25]. A subsequent theoretical explanation is that the third cumulant is in fact susceptible to environmental effects [50]. This experiment has already triggered some theoretical activity [26,51,52].

2 Full Counting Statistics

The fundamental quantity of interest in quantum transport is the probability distribution

$$P_{t_0}(N_1, N_2, \ldots, N_M) \equiv P(\boldsymbol{N})\,, \tag{1}$$

which denotes for a M-terminal conductor the probability that during a certain period of time t_0 N_1 charges enter through terminal 1, N_2 charges enter through terminal 2, ..., and N_M charges enter through terminal M (negative N_i correspond to charges leaving the respective terminal). The same information is contained in the cumulant generating function (CGF), defined by

$$S(\chi) = \ln\left[\sum\nolimits_{\boldsymbol{N}} e^{i\boldsymbol{N}\boldsymbol{\chi}} P(\boldsymbol{N})\right]\,, \tag{2}$$

where we introduced the vector of counting fields $\boldsymbol{\chi} = (\chi_1, \chi_2, \ldots, \chi_N)$. The normalization condition requires $\sum_{\boldsymbol{N}} P(\boldsymbol{N}) = 1 \leftrightarrow S(\boldsymbol{\chi} = \boldsymbol{0}) = 0$.

2.1 Charge Conservation

We are interested in the long-time limit of the charge counting statistics, which means that no extra charges remain inside the conductor after the counting interval. If we count only the total number of transferred charges, we simply have to consider $P(N) = \sum_{\boldsymbol{N}} \delta_{\sum N_\alpha, N} P(\boldsymbol{\chi})$, or, equivalently, to put all counting fields equal $S(\chi_1 = \chi, \chi_2 = \chi, \ldots, \chi_N = \chi)$. Charge conservation now means that $S(\chi_1 = \chi, \chi_2 = \chi, \ldots, \chi_N = \chi) = 0$. As a consequence the CGF depends only on differences between counting fields. This has the direct interpretation, that a difference $\chi_\alpha - \chi_\beta$ is related to a charge transfer between terminal α and β. In general, this means that we need only $M-1$ counting fields to describe a M-terminal structure. If one of the counting fields, e. g. χ_M, has been eliminated, the charge transfer into terminal M can be restored from the CGF, in which all other χ_α are equal $\chi_\alpha - \chi_M$. In the special case of a two-terminal device, the CGF depends only on $\chi \equiv \chi_1 - \chi_2$. We denote this below with $S(\chi)$. Later we will see that the CGF's are in general *periodic* functions of χ, i. e. $S(\chi + 2\pi) = S(\chi)$. This ensures that the total charge transfered is an integer multiple of the electron charge e, which makes sense, since we are talking about electron transport and want to neglect transient effects.

However, the interesting question what the charge of an elementary event is, can be answered by FCS. Suppose the a CGF has the property $S(\chi + 2\pi/n) = S(\chi)$. Direct calculation shows that

$$P(Q) = \int \frac{d\chi}{2\pi} e^{-iN\chi + S(\chi)} = \begin{cases} P_n(Q/n) & , (Q \bmod n) = 0 \\ 0 & , (Q \bmod n) \neq 0 \end{cases}, \tag{3}$$

where $P_n(N)$ is the distribution $S_n(\chi) = S(\chi/n)$. The probability distribution vanishes for all N which are not multiples of n, thus the elementary charge transfer is in units of ne, where e is the electron charge. This has interesting consequences in the context of superconductivity, in which multiple charge transfers can occur [23,47,48], or for fractional charge transfer [25].

2.2 Correlations

One commonly addressed question is, if two different events (say the charges transfered into terminals α and β) are independent or not. For independent events the probability distributions are separable and we find that $\langle N_\alpha^k N_\beta^l \rangle = \langle N_\alpha^k \rangle \langle N_\beta^l \rangle$. In terms of the CGF this means that the CGF is the sum of two terms: one which depends only on χ_α and a second one, which depends only on χ_β. On the contrary, if the CGF can not be written as such a sum, the charge transfers in terminals α and β are correlated.

2.3 Special Distributions (Two Terminals)

If the elementary events are uncorrelated, the probability distribution is *Poissonian.* With the average number of events is $\bar{N}$ we have

$$P_{\text{Poisson}} = \frac{\bar{N}}{N!} e^{-\bar{N}} \leftrightarrow S(\chi) = \bar{N}\left(e^{i\chi} - 1\right) . \tag{4}$$

In the context of electron transport we encounter this distribution mostly for *tunnel junctions* with an almost negligible transmission probability at low temperatures. Here $\bar{N} = G_T V t_0 / e$ is simply related to the voltage bias and the tunnel conductance.

As second example we consider the binomial (or Bernoulli) distribution. This is obtained if an event occurs with a probability T and the number of tries is fixed to N_0:

$$P_{\text{binomal}} = \binom{N_0}{N} T^N (1-T)^{N_0 - N} \leftrightarrow S(\chi) = N_0 \ln\left[1 + T\left(e^{i\chi} - 1\right)\right] . \tag{5}$$

In some sense this is the most fundamental distribution in quantum transport: it gives the statistics of a voltage biased single channel quantum conductor if we identify $N_0 = eVt_0/h$.

2.4 Special Distributions (Many Terminals)

For uncorrelated processes the CGF takes the simple form

$$S(\boldsymbol{\chi}) = \sum_{\alpha,\beta} \bar{N}_{\alpha,\beta}\left(e^{i(\chi_\alpha - \chi_\beta)} - 1\right) . \tag{6}$$

The resulting distribution is just the product of Poisson distributions, taking into account total charge conservation. An important example is a multinomial distribution for N_0 independent attempts, which can have different outcomes with probabilities T_α. It has the form

$$S(\boldsymbol{\chi}) = N_0 \ln\left[1 + \sum_\alpha T_\alpha \left(e^{i\chi_\alpha} - 1\right)\right] . \tag{7}$$

3 Theoretical Approach to Full Counting Statistics

3.1 General Theory

We will follow here the approach to FCS using the Green's function technique [8]. Quantum-mechanically we define the cumulant generating function by [8–10,25]

$$e^{S(\chi)} = \langle \mathcal{T}_K e^{-i\frac{1}{2e}\int_{C_K} dt\chi(t)I(t)} \rangle . \tag{8}$$

Here, $\mathcal{T}_K$ denotes time ordering along the Keldysh-contour C_K, depicted in Fig. 1. The time-dependent field $\chi(t)$ is defined as $\pm\chi$ for $t \in C_{1(2)}$, i.e. $\chi(t)$ changes sign between the upper and the lower branch of C_K. $\hat{I}(t)$ is the usual operator of the current through a certain cross section. Expansion in the *counting field* yields the cumulants. In the second order we find the 2^{nd} cumulant as

$$C_2(t_0) = \int_0^{t_0} dt \int_0^{t_0} dt' \left\langle \delta\hat{I}(t)\delta\hat{I}(t') \right\rangle . \tag{9}$$

Higher cumulants yield more complicated expressions.

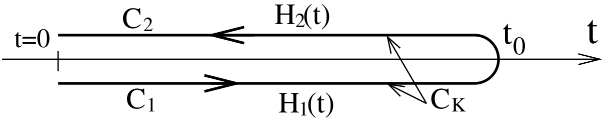


Fig. 1. Keldysh time-ordering contour

3.2 Current Correlation Functions

The cumulants $C_n(t_0)$ are directly related to experimentally accessible quantities like current noise or the third cumulant of the current fluctuations. Let us demonstrate the relation for the low-frequency current noise, defined by

$$S_I = 2\Delta f \int_{-\infty}^{\infty} d\tau \left\langle \delta\hat{I}(\tau)\delta\hat{I}(0) \right\rangle , \tag{10}$$

where $\delta\hat{I}(\tau) = \hat{I}(\tau) - \langle\hat{I}\rangle$ and $\Delta f = f_{\max} - f_{\min}$ is the frequency band width, in which the noise is measured. The factor of 2 enters here to conform to the review article [3]. We now transform in (9) the integration variables from t, t' to $T = (t + t')/2, \tau = t - t'$. In the limit $t_0 \equiv (\Delta f)^{-1}$ much larger than the correlation time of current-fluctuations, the integral over T can be evaluated and we obtain from (9) the desired result $S_I/2$. Similar arguments hold for higher cumulants, for which the expression corresponding to (9) are less trivial, however. In [49] it was noted that C_3 depends in an quite unusual way on the frequency band measured, i.e. it is proportional to $2f_{\max} - f_{\min}$, which made it possible to prove experimentally that the third cumulant is actually measured.

3.3 Keldysh-Green's Functions

So far we have formally defined the CGF quantum mechanically. The relation to standard quantum-field theory methods is made in the following way. We introduce the standard Green's function [53] in the presence of a time-dependent Hamiltonian

$$H_c(t) = H_0 + \frac{1}{2e}\chi(t)\hat{I}\,, \tag{11}$$

where the time-dependence is only in the 'counting' field $\chi(t)$. The counting field couples to the operator $\hat{I}$ of the current through a cross section, which intersects the conductor entirely. The single-particle operators corresponding to H_0 and I are denoted by h_0 and j.

Using the matrix notation for the Keldysh-Green's functions, we arrive at the equation of motion

$$\left[i\frac{\partial}{\partial t} - \hat{h}_0 - \frac{\chi}{2e}\bar{\tau}_3\hat{j}_c\right]\check{G}(t,t';\chi) = \delta(t-t')\,. \tag{12}$$

Here, $\bar{\tau}_3$ denotes the third Pauli matrix in the Keldysh space and is a result of the unusual time-dependence of the counting field. The relation of the Green's function (12) to the CGF (8) is obtained from a diagrammatic expansion in χ (the calculation is formally equivalent to the calculation of the thermodynamic potential in an external field, see e. g. [54]). One obtains the relation [8]

$$\frac{\partial S(\chi)}{\partial \chi} = \frac{it_0}{e}\mathrm{Tr}\left[\bar{\tau}_3\hat{j}\check{G}(t,t;\chi)\right] \equiv \frac{it_0}{e}I(\chi)\,, \tag{13}$$

where we have restricted us to a static situation, for which $\check{G}(t,t)$ is independent of time. Note, that the *counting current* $I(\chi)$ should not be confused with the standard electrical current, which is actually given by $I_{el} = I(0)$. Rather, $I(\chi)$ contains (via an expansion in χ) *all current-correlators* at once. It nevertheless resembles a current in the usual sense. E. g., it follows from (12) that the counting current is conserved.

3.4 A Simplification

In a typical mesoscopic transport problem we can access the full counting statistics based on the separation into terminals (or reservoirs) and a scattering region. Terminals provide boundary conditions to Green's function far away from the scattering region. These are usually determined by external current or voltage sources and include material properties like superconductivity. Let us now take the following parameterization of the current operator in (12)

$$\hat{j}(\boldsymbol{x}) = (\nabla F(\boldsymbol{x}))\lim_{\boldsymbol{x}\to\boldsymbol{x}'}\frac{ie}{2m}\left(\nabla_{\boldsymbol{x}} - \nabla_{\boldsymbol{x}'}\right)\hat{\sigma}_3\,. \tag{14}$$

$F(\boldsymbol{x})$ is chosen such that it changes from 0 to 1 across a cross section C, which intersects the terminal, but is of arbitrary shape. Here we have introduced a matrix $\hat{\sigma}_3$ in the current operator, occurring e. g. in the context of superconductivity. We assume that the change from 0 to 1 should occur on a length scale Λ, for which we assume $\lambda_F \ll \Lambda \ll l_{\mathrm{imp}}, \xi_0$ (Fermi wave length λ_F, impurity mean free path l_{imp}, and coherence length $\xi_0 = v_F/2\Delta$). With this assumption we can reduce (12) *inside the terminal* to its quasiclassical version (see [53])

$$\boldsymbol{v}_F \nabla \check{g}(\boldsymbol{x}, \boldsymbol{v}_F, t, t', \chi) = \left[-i\frac{\chi}{2}(\nabla F(\boldsymbol{x}))\boldsymbol{v}_F \check{\tau}_K \, , \, \check{g}(\boldsymbol{x}, \boldsymbol{v}_F, t, t', \chi) \right] . \qquad (15)$$

Here $\check{\tau}_K = \bar{\tau}_3 \hat{\sigma}_3$ is the matrix of the current operator and $\check{g}$ obeys the normalization condition $\check{g}^2 = 1$. Other terms can be neglected due to the assumptions we have made for Λ. The counting field can then be eliminated by the gauge-like transformation

$$\check{g}(\boldsymbol{x}, \boldsymbol{v}_F, t, t', \chi) = e^{-i\chi F(\boldsymbol{x})\check{\tau}_K/2} \check{g}(\boldsymbol{x}, \boldsymbol{v}_F, t, t', 0) e^{i\chi F(\boldsymbol{x})\check{\tau}_K/2} . \qquad (16)$$

We assume now that the terminal is a diffusive metal of negligible resistance. Then the Green's functions are constant in space (except in the vicinity of the cross section C) and isotropic in momentum space. Applying the diffusive approximation [53] in the terminal leads to a transformed terminal Green's function

$$\check{G}(\chi) = e^{-i\chi\check{\tau}_K/2} \check{G}(0) e^{i\chi\check{\tau}_K/2} , \qquad (17)$$

on the right of the cross section C (where $F(\boldsymbol{x}) = 1$) with respect to the case without counting field. Consequently, the counting field is entirely incorporated into a *modified boundary condition* imposed by the terminal onto the mesoscopic system.

3.5 Summary of the Theoretical Approach

This concludes the theoretical approach to counting statistics of mesoscopic transport. Let us briefly summarize the scheme to follow. The FCS can be obtained by a slight extension of the usual Keldysh-Green's function approach, which is widely employed to treat quantum transport problems. Making use of the separation of the mesoscopic structure into *terminals* and a *scattering* region, the formalism boils down to a very powerful, but nevertheless simple rule: we have to apply the *counting rotation* (17) to a terminal, thus providing new boundary conditions (now depending on the *counting field* χ) to the scattering problem. We then proceed 'as usual' and calculate the current in the terminal, which again depends on χ. Finally the counting statistics is obtained from (13).

4 Two-Terminal Contacts

4.1 Tunnel Contact

To illustrate the theoretical method we first calculate the counting statistics of a tunnel junction. As usual the system is described by a tunnel Hamiltonian

$H = H_1 + H_2 + H_T$, where $H_{1(2)}$ describe the left(right) terminal and H_T describes the tunneling. The current is calculated in second order in the tunneling amplitudes and we obtain $I(\chi) = \frac{G_T}{8e} \int dE \mathrm{Tr}\left(\check{\tau}_K \left[\check{G}_1(\chi), \check{G}_2\right]\right)$, where G_T is the conductance of the tunnel junction and we have included the counting field in $\check{G}_1$. The CGF is (using $(\partial/\partial\chi)G_1(\chi) = (i/2)\left[\check{\tau}_K, \check{G}_1(\chi)\right]$)

$$S(\chi) = i\frac{t_0}{e}\int_0^\chi d\chi' I(\chi') = \frac{G_T t_0}{4e^2}\int dE \mathrm{Tr}\left\{\check{G}_1(\chi), \check{G}_2\right\}\,, \tag{18}$$

which is the general expression for the FCS of a tunnel junction. We use the pseudo-unitarity $\check{\tau}_K^2 = \check{1}$ to write

$$S(\chi) = N_{12}(e^{i\chi} - 1) + N_{21}(e^{-i\chi} - 1)\,, \tag{19}$$

where $N_{ij} = (t_0 G_T/16e^2)\int dE \mathrm{Tr}\left[(1 + \check{\tau}_K)\check{G}_i(1 - \check{\tau}_K)\check{G}_j\right]$ denotes the average number of charges tunneling from i to j. The statistics is therefore a bidirectional Poisson distribution [25]. It is easy to see that the cumulants are $C_n = N_{12} + (-1)^n N_{21}$. If either $N_{21} = 0$ or $N_{12} = 0$ we obtain the Schottky limit. Furthermore, in equilibrium $N_{12} = N_{21}$ and the FCS is $(2G_T k_B T t_0/e^2)(\cos(\chi) - 1)$, which is non-Gaussian, remarkably.

4.2 General CGF for Quantum Contacts

Using the method presented in the previous section, we can find the counting statistics for all conductors, which are characterized by a set of transmission coefficients $\{T_n\}$. Nazarov has shown that the transport properties of such a contact are described by a *matrix current* [55]

$$\check{I}_{12} = -\frac{e^2}{\pi}\sum_n \frac{2T_n\left[\check{G}_1, \check{G}_2\right]}{4 + T_n\left(\{\check{G}_1, \check{G}_2\} - 2\right)}\,. \tag{20}$$

Here, $\check{G}_{1(2)}$ denote the matrix Green's functions on the left and the right of the contact. We should emphasize that the matrix form of (20) is crucial to obtain the FCS, since it is valid for any matrix structure of the Green's functions. The *scalar current* is obtained from the matrix current by

$$I_{12} = \frac{1}{4e}\int dE \mathrm{Tr}\check{\tau}_K \check{I}_{12}\,. \tag{21}$$

To find the FCS, we apply the counting rotation (17) to terminal 1, i. e. $\check{G}_1$ becomes χ-dependent. It turns out that the CGF can then be found generally from the relations (13), (20), and (21). To integrate (13) with respect to χ, we need the relations $i(\partial/\partial\chi)\check{G}_1(\chi) = [\check{\tau}_K, \check{G}_1(\chi)]$ and $\mathrm{tr}[\check{G}_1(\chi), \{\check{G}_1(\chi), \check{G}_2\}^n] = 0$. We find [9]

$$S(\chi) = \frac{t_0}{2\pi}\sum_n \int dE \mathrm{Tr}\ln\left[1 + \frac{T_n}{4}\left(\{\check{G}_1(\chi), \check{G}_2\} - 2\right)\right]\,. \tag{22}$$

Table 1. Characteristic functions of some generic conductors. The transmission eigenvalue densities are normalized to G/G_Q, where $G_Q = 2e^2/h$ is the quantum conductance. The third column displays the CGF-density, which determines the CGF via $S(\chi) = (t_0 G/4eh) \int dE \mathrm{tr}\check{s}(\{\check{G}_1(\chi), \check{G}_2\}/2)$

	$\rho(T)[G/G_Q]$	$\check{s}(\mathit{\Lambda})$
Single channel	$\delta(T - T_1)$	$\ln(1 - T_1(\mathit{\Lambda} - 1)/2)$
Diffusive connector	$\frac{1}{2}\frac{1}{T\sqrt{1-T}}$	$\frac{1}{4}\mathrm{arcosh}^2(\mathit{\Lambda})$
Dirty interface	$\frac{1}{\pi}\frac{1}{T^{3/2}\sqrt{1-T}}$	$\sqrt{2(1+\mathit{\Lambda})}$
Chaotic cavity	$\frac{2}{\pi}\frac{1}{\sqrt{T}\sqrt{1-T}}$	$4\ln\left(2 + \sqrt{2(1+\mathit{\Lambda})}\right)$

This is a very important result. It shows that the counting statistics of a large class of constrictions can be cast in a common form, independent of the contact types.

Note, that (22) is just the sum over CGF's of all eigenchannels. Thus, we can obtain the CGF's of all constrictions from a known transmission eigenvalue density. These are known for a number of generic contacts (see e.g. [56] and Table 1), can be determined numerically, or can be taken from experiment. Below we will discuss several illustrative examples for single channel contacts.

4.3 Normal Contacts

Consider first a single channel with transmission T between two normal reservoirs. They are characterized by occupation factors $f_{1(2)} = [\exp((E - \mu_{1(2)})/k_B T_e) + 1]^{-1}$ (T_e is the temperature). We obtain the result [2,6] (see Appendix)

$$S(\chi) = \frac{2t_0}{h}\int dE \ln\left[1 + T_{12}(E)\left(e^{i\chi} - 1\right) + T_{21}(E)\left(e^{-i\chi} - 1\right)\right] . \qquad (23)$$

Here, we introduced the probabilities $T_{12} = T f_1(E)\,(1 - f_2(E))$ for a tunneling event from 1 to 2 and $T_{21}(E)$ for the reverse process. We see that the FCS (for each energy) is a trinomial of an electron going from left to right, from right to left, or no scattering at all. The *counting factors* $e^{\pm i\chi} - 1$ thus correspond to single charge transfers from 1 to 2 (2 to 1).

At zero temperature and $\mu_1 - \mu_2 = eV \geq 0$ the argument of the energy integral is constant in the interval $\mu_1 < E < \mu_2$ and we obtain the *binomial form* $S(\chi) = \frac{2et_0|V|}{h}\ln\left[1 + T\left(e^{i\chi} - 1\right)\right]$. Note that for reverse bias $\mu_2 > \mu_1$ the CGF has the same form, but with a counting factor $e^{-i\chi} - 1$. The prefactor denotes the *number of attempts* $M = et_0 V/h$ to transfer an electron [1]. If the transmission probability is unity the FCS is non-zero only for $N = M$, which

[1] The non-integer values of $M(t_0)$ occur due to the quasiclassical approximation [6]. A more careful treatment reveals that M itself is described by a probability distribution. For large M the difference is negligible.

therefore constitutes the maximal number of electrons occupying an energy strip eV that can be sent through one (spin-degenerate) channel in a time interval t_0. In equilibrium it follows from (23) that the counting statistics is [57]

$$S(\chi) = -\frac{2t_0 k_B T_{el}}{h} \arcsin^2\left(\sqrt{T}\sin\frac{\chi}{2}\right) . \tag{24}$$

The fluctuations are non-Gaussian, except for $T = 1$, when $S(\chi) = -\frac{t_0 k_B T_{el}}{h}\chi^2$.

4.4 SN-Contact

The FCS of a contact between a superconductor and a normal metal also follows from the general expression (22). Using the Green's functions given in the Appendix we find the result [23]

$$S(\chi) = \frac{t_0}{2\pi}\sum_n \int dE \ln\left[1 + \sum_{q=-2}^{2} A_{nq}(E)\left(e^{iq\chi} - 1\right)\right] . \tag{25}$$

The coefficients $A_{nq}(E)$ are related to a charge transfer of $q \times e$. For example, a term $\exp(2i\chi) - 1$ corresponds just to an Andreev reflection process, in which two charges are transfered simultaneously. Explicit expressions for the various coefficients are given in [23,58]. The most interesting regime is that of pure Andreev reflection: $eV, k_BT \ll \Delta$. Here, we obtain

$$\begin{aligned} S(\chi) = \frac{t_0}{h}\int dE \ln [1 &+ R_A f_+ f_- \left(e^{i2\chi} - 1\right) \\ &+ R_A(1 - f_+)(1 - f_-))\left(e^{-i2\chi} - 1\right)] , \end{aligned} \tag{26}$$

where $R_A = T^2/(2 - T)^2$ is just the Andreev reflection probability and $f_\pm = f(\pm E)$ denotes the occupation with electrons above(below) the chemical potential of the superconductor. For low temperatures $k_BT_e \ll eV \ll \Delta$, the CGF becomes

$$S(\chi) = \frac{2et_0|V|}{h} \ln\left[1 + R_A\left(e^{i2\chi} - 1\right)\right] . \tag{27}$$

The CGF is now π-periodic, which according to Sect. 2 reflects that the charge transfer of an elementary event is now $2e$, a consequence of Andreev reflection. Quite remarkably, the statistics is again a simple binomial distribution. In equilibrium, we can adapt the result from 24 to find

$$S(\chi) = -\frac{2t_0 k_B T_{el}}{h} \arcsin^2\left(\sqrt{R_A}\sin\chi\right) \quad (\text{for } \chi \in [-\pi/2, \pi/2]) . \tag{28}$$

The counting statistics is also non-Gaussian, except for $R_A = 1$.

4.5 Superconducting Contact

Now we turn to a slightly more involved problem: a contact between two superconductors biased at a finite voltage V. For $eV < 2\mathit{\Delta}$ the transport is dominated by multiple Andreev reflections (MAR). The microscopic analysis of the average current and the shot noise calculations suggest that the current at subgap energies proceeds in "giant" shots, with an effective charge $q \sim e(1 + 2\mathit{\Delta}/|eV|)$. However, the question of size of the charge transfered in an elementary event can only be rigorously resolved by the FCS. The answer was given by Cuevas and the author [47] based on a microscopic Green's function approach. Independently, Johansson, Samuelsson and Ingerman [48] arrived at the same conclusion using a different method.

Now, what would we like to have? In Sect. 2 we have discussed that one can speak of multiple charge transfers if the CGF allows an interpretation in terms of elementary events, which are described by counting factors $e^{in\chi} - 1$, where n denotes the charge transfered in the process. How can we ever hope to obtain this from the general formula (22)? We have to calculate the determinant of a 4×4-matrix, which can give only factors of the type $e^{i2\chi}$ or even smaller charges. The answer to this puzzle is that we have to re-interpret the matrix structure in (22), since the Green's functions of superconductors at a finite bias voltage are essentially non-local in energy. The general result for the CGF can be written as $S(\chi) = \mathrm{Tr} \ln \check{Q}$, where Tr$= \int_0^{t_0} dt \mathrm{tr}$ and $\check{Q}(t) = 1 + (T/4)(\{\check{G}_1 \overset{\otimes}{,} \check{G}_2\} {-} 2)(t,t)$. Here $\check{G}_1 \otimes \check{G}_2(t,t') = \int dt'' \check{G}_1(t,t'' \check{G}_2(t'',t')$. Let us set the chemical potential of the right electrode to zero and represent the Green's functions by $\check{G}_1(t,t') = e^{ieVt\bar{\tau}_3} \check{G}_S(t-t') e^{-ieVt'\bar{\tau}_3}$ and $\check{G}_2(t,t') = \check{G}_S(t-t')$. Here, we have not included the dc part of the phase, since it can be shown that it drops from the expression of the dc FCS at finite bias. The Fourier transform leads to a representation of the form $\check{G}(E,E') = \sum_n \check{G}_{0,n}(E) \delta(E - E' + neV)$, where $n = 0, \pm 2$. Restricting the fundamental energy interval to $E - E' \in [0, eV]$ we can represent the convolution as *matrix product*, i.e. $(G_1 \otimes G_2)(E,E') \to (\check{G}_1 \check{G}_2)_{n,m}(E,E') = \sum_k (G_1)_{n,k}(E,E')(G_2)_{k,m}(E,E')$. The trace in this new representation is written as $\int_0^{eV} dE \sum_n \mathrm{Tr} \ln \left(\check{Q}\right)_{nn}$. In this way, the functional convolution is reduced to matrix algebra for the infinite-dimensional matrix $\check{Q}$. From these arguments it is clear that the statistics is a *multinomial* distribution of *multiple* charge transfers:

$$S(\chi) = \frac{t_0}{h} \int_0^{eV} dE \ln \left[1 + \sum_{n=-\infty}^{\infty} P_n(E,V) \left(e^{in\chi} - 1 \right) \right] . \tag{29}$$

General expressions for the probabilities $P(E,V)$ have been derived in [47].

Here, we will pursue a different path and study a toy model. Let us neglect all set $f^{R,A}(|E| < \mathit{\Delta}) = 1$, $g^{R(A)}(|E| > \mathit{\Delta}) = \pm 1$, and equal to zero otherwise. Physically, this means that we neglect Andreev reflections above the gap and replace the quasi-particle density of states by a constant $|E| > \mathit{\Delta}$. This simplifies the calculation a lot, since the matrix trace now becomes finite. Let us for example consider a voltage $eV = 2\Delta/4$. In that case, we consider the determinant of the matrix

$$\det\left[1-\frac{\sqrt{T}}{2}\begin{pmatrix} \hat{Q}_-(\chi) & 1 & & & \\ 1 & 0 & e^{-i\chi\hat{\tau}_3} & & \\ & e^{i\chi\hat{\tau}_3} & 0 & 1 & \\ & & 1 & 0 & e^{-i\chi\hat{\tau}_3} \\ & & & e^{i\chi\hat{\tau}_3} & \hat{Q}_+ \end{pmatrix}\right], \tag{30}$$

where $Q_\pm(\chi)$ describe quasi-particle emission (injection) and off-diagonal pairs $e^{\pm\chi}$ are associated with Andreev reflection. Evaluating the determinant we find $S(\chi) = \frac{\Delta t_0}{2h}\ln\left[1+P_5\left(e^{in\chi}-1\right)\right]$, where $P_5 = T^5/(16-20T+5T^2)^2$. This expression describes binomial transfers of 5 charges with probability P_5. For general subharmonic voltages $2\Delta/(n-1)$ we find

$$S(\chi) = \frac{2\Delta t_0}{(n-1)h}\ln\left[1+P_n\left(e^{in\chi}-1\right)\right], \tag{31}$$

where the probabilities are given by

$$\begin{aligned} &P_2 = \tfrac{T^2}{(2-T)^2}\,,\ P_3 = \tfrac{T^3}{(4-3T)^2}\,,\ P_4 = \tfrac{T^4}{(8-8T+T^2)^2}\,,\ P_5 = \tfrac{T^5}{(16-20T+5T^2)^2} \\ &P_6 = \tfrac{T^6}{(2-T)^2(16-16T+T^2)^2}\,,P_7 = \tfrac{T^7}{(64-112T+56T^2-7T^3)^2}\,. \end{aligned} \tag{32}$$

Note the limiting cases of these probabilities $P_n \sim T^n/4^{n-1}$ for $T \ll 1$ and $P_n = 1$ for $T \to 1$. We conclude this section by saying that the general results for the CGF [47] allow for a fast and efficient calculation of all dc-transport properties of contacts between superconductors (which may contain magnetic impurities, phonon broadening or other imperfections).

5 Quantum Noise in Diffusive SN-Structures

In this section, we illustrate a further advantage of the Keldysh-Green's functions approach to counting statistics. We consider a normal metallic diffusive wire connected on one end to a normal metal reservoir and on the other side to a superconductor. The wire is supposed to have a mean free path $l \gg \lambda_F$, a corresponding diffusion coefficient $D = v_F l/3$, and a length L. For $eV, k_BT \ll \Delta$ the transport occurs through Andreev reflection at the interface to the superconductor. This system shows a quite remarkable property, which is the so-called reentrance effect of the conductance. The energy difference $2E$ of electron-hole pairs leads to a dephasing on a length scale $\xi_E = \sqrt{D/2E}$. This has the consequence that the (otherwise) normal wire becomes partially superconducting and the conductance increases with decreasing energy. However, once the coherence length ξ_E reaches L the conductance *decreases* again. Finally for $E = 0$ the conductance is *exactly* equal to the conductance in the normal state. This is the reentrance effect occurring at an energy of the order of the Thouless energy $E_c = \hbar D/L^2$. In Fig. 2 (left panel, dotted curve) the resulting differential conductance at zero temperature is plotted.

The transport in this system is described by a matrix diffusion equation for the Keldysh Green's functions, the so-called Usadel equation

$$-\frac{D}{\sigma}\nabla\check{I} = \left[-iE\hat{\tau}_3, \check{G}\right] \ , \ \check{I} = -\sigma\check{G}\nabla\check{G} \, . \tag{33}$$

In these equations $\sigma = 2e^2N_0D$ is the conductivity. The boundary conditions for this equation are that the Green's functions in the terminal approach the bulk solution $\check{G}_N$ or $\check{G}_S$, respectively. This equation is in general difficult to solve, even if one is interested in the average current only. However, we can calculate the noise and the counting statistics using the recipe outlined in Sect. 3 and obtain the noise in the full parameter range of (33).

Before considering (33) in its full generality, we consider the limiting cases of low and high energies (compared to E_c). For $E = 0$ the r.h.s. is absent and the system is completely analogous to a diffusive connector as discussed in Sect. 4. From Table 1 and using the eigenvalues (52) we find

$$\begin{aligned} S(\chi) = \frac{t_0G}{16e^2}\int dE\mathrm{arcosh}^2 \Big[2\big(f_+f_-(e^{2i\chi}-1) \\ +(1-f_+)(1-f_-)(e^{-2i\chi}-1)\big)-1\Big] \, . \end{aligned} \tag{34}$$

This result shows, once again, that the charges are transfered in pairs. It is interesting to compare with the CGF for a diffusive wire between two normal metals, for which we obtain [5,8]

$$\begin{aligned} S(\chi) = \frac{t_0G}{4e^2}\int dE\mathrm{arcosh}^2 \Big[2\big(f_1(1-f_2)(e^{i\chi}-1) \\ +f_2(1-f_1)(e^{-i\chi}-1)\big)-1\Big] \, . \end{aligned} \tag{35}$$

We see that the only difference in the CGF between the SN- and the NN-case is in the counting factors, and a prefactor 1/4. Note, that this coincidence only occurs for the diffusive connector, but is by no means a general rule. At zero temperature the results simplify and we find

$$S^{\mathrm{SN}}(\chi) = \frac{1}{2}S^{\mathrm{NN}}(2\chi) \ , \ S^{\mathrm{NN}}(\chi) = \frac{t_0GV}{4e}\mathrm{arcosh}^2\left(2e^{i\chi}-1\right) \ , \tag{36}$$

a surprising simple relation between the CGF for the Andreev wire and the normal diffusive wire. It is easy to see that the cumulants obey the general relation $C_n^{\mathrm{SN}} = 2^{n-1}C_n^{\mathrm{NN}}$. We observe that we can read off the effective charge from the ratio $C_n^{\mathrm{SN}}/C_n^{\mathrm{NN}} = (q_{\mathrm{eff}}/e)^{n-1}$ and, indeed, find $q_{\mathrm{eff}} = 2e$. This result for the effective charge is a special property of the *diffusive connector*.

At energies large compared to E_c it is also possible to find the CGF for the Andreev wire in general. Then the proximity effect in the wire is absent and it turns out [36] that the wire can be effectively mapped on a normal circuit, consisting of two identical wires in series to which twice the voltage is applied and twice the counting field. Thus, for $E \gg E_c$ we obtain $S^{SN}(\chi)$ from $S^{NN}(\chi)$

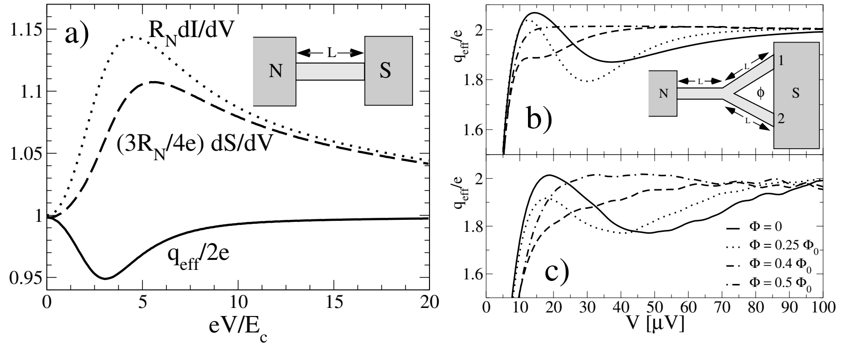


Fig. 2. Noise in diffusive SN-systems. Left panel a): the differential conductance and the noise show a reentrant behavior. The effective charge, defined as $q_{\text{eff}}(E) = (3/2)dS/dI$ reveals that the correlated Andreev pair transport suppresses the noise below the uncorrelated Boltzmann-Langevin result $2e$. Right panels b) and c): Effective charge of the Andreev interferometer shown in the inset (realized experimentally in [35]). The upper panel b) shows the theoretical predictions and the lower panel c) the experimental results. The theoretical results contain no fitting parameter (the Thouless energy $E_c = 30\,\mu$eV was extracted from the sample geometry and the experimental temperature of 43 mK was included in the calculation). Therefore, it is reasonable that the deviations between experimental and theoretical results come from possible heating effects in the experiment, which are not accounted for in the theoretical calculation

by the replacement $\chi \to 2\chi$ and $G \to G/2$, which exactly brings us to (refeq:cgf-diffusive-andreev) and shows that the counting statistics is again the same in the incoherent limit.

The full quantum-mechanical calculation of the energy-dependent shot noise can be performed on the basis of the approach of Sect. 3 [10]. We expand up to linear order in χ, i.e. $\check{G}(\chi) = \check{G}_0 - i(\chi/2)\check{G}_1$ and $\check{I}(\chi) = \check{I}_0 - i(\chi/2)\check{I}_1$. Substituting in (33) we find

$$\frac{D}{\sigma}\frac{\partial}{\partial x}\check{I}_1 = \left[-iE\bar{\tau}_3\,,\,\check{G}_1\right]\,,\ \check{I}_1 = -\sigma\left(\check{G}_0\frac{\partial}{\partial x}\check{G}_1 + \check{G}_1\frac{\partial}{\partial x}\check{G}_0\right)\,. \qquad (37)$$

The boundary conditions at the reservoirs read $\check{G}_1(0) = \left[\check{\tau}_{\text{K}}, \check{G}_{\text{L}}\right]$ at the left end and $\check{G}_1(L) = 0$ at the right end. Finally the noise is $S_{\text{I}} = -e\int dE\text{Tr}\check{\tau}_{\text{K}}\check{I}_1(x)$. By taking the trace of (37) multiplied with $\check{\tau}_{\text{K}}$ it follows that it does not matter, where the noise is evaluated, as it should be. From these equations the generalization of the Boltzmann-Langevin equation to superconductors can be derived [59], which allows for a faster numerical solution. The results for the energy dependent noise is shown in the left panel of Fig. 2. A direct comparison of the differential shot noise and the differential conductance (for zero temperature) shows the difference in the energy dependence. The effective charge defined as $q_{\text{eff}} = (3/2)dS/dI$ displays the clear deviation of the quantum noise from the Boltzmann-Langevin result of $2e$. At energies below the Thouless energy E_c the effective charge is suppressed below $2e$. This shows that the correlated Andreev pair transport suppresses the noise below the uncorrelated Boltzmann-Langevin result.

To probe the pair correlations in diffusive superconductor-normal metal-heterostructures experimentally it is most convenient to use an Andreev interferometer. An example is shown in the left part of Fig. 2. A diffusive wire connected to a normal terminal is split into two parts, which are connected to two different points of a superconducting terminal. By passing a magnetic flux through the loop one can effectively vary the phase difference between the two connections to the superconductor. Such a structure has been experimentally realized by the Yale group [35]. In Fig. 2 we present a direct comparison between our theoretical predictions and the experimentally obtained effective charge. Note, that we have included the experimental temperature in the theoretical modeling. The finite temperature explains the strong decrease of the effective charge in the regime $|eV| \leq k_B T$, where the noise is fixed by the fluctuation-dissipation theorem. The disagreement between theory and experiment in this regime stems solely from differences in the measured temperature-dependent conductance from the theoretical prediction. We attribute this to heating effects. The qualitative agreement in the shot-noise regime $|eV| \geq k_B T$ is satisfactory, if one takes into account, that we have no free parameters for the theoretical calculation. Both, experiment and theory show a suppression of the effective charge for some finite energy, which is of the order of the Thouless energy and depends on flux in a qualitative similar manner. Remarkably for half-integer flux the effective charge is completely flat, in contrast to what one would expect from circuit arguments based on the conductance distribution in the fork geometry. Currently we have no explanation for this behavior, and therefore more work is needed in this direction.

6 Multi-terminal Circuits

In circuits with more than two terminals it is of particular interest to study non-local correlations of currents in different terminals. For that purpose we need a slight extension of the theoretical approach of Sect. 3, suitable for multi-terminal circuits. We will now introduce this method.

6.1 Circuit Theory

To study transport in general mesoscopic multi-terminal structures the so-called circuit theory for quantum transport was developed by Nazarov [55,60]. Its main idea, borrowed from Kirchhoff's classical circuit theory, is to represent a mesoscopic device by discrete elements, which resemble the known elements of electrical transport. We briefly repeat the essentials of the circuit theory. Topologically, one distinguishes three elements: terminals, nodes and connectors. Terminals are the connections to the external voltage or current sources and provide boundary conditions, specifying externally applied voltages, currents or phase differences in the case of superconductors. The actual circuit is represented by a network of nodes and connectors, the first determining the approximate layout and the second describing the connections between different nodes, respectively.

To describe quantum effects it is necessary to represent the variables describing a node by *matrix Green's function* $\check{G}$, which can be either Nambu or Keldysh matrices, or a combination thereof. Consequently, we describe the current through a connector by a *matrix current* $\check{I}$, which relates the fluxes of all elements of $\check{G}$ on neighboring nodes. The current has been derived by Nazarov [55] and is given by (20) for a connector, characterized by a set of transmission coefficients $\{T_n\}$. Note that the *electrical current* is obtained from $I_{12} = \frac{1}{4e}\int dE \mathrm{Tr}\check{\tau}_K \check{I}_{12}$. The boundary conditions are given in terms of fixed matrix Green's functions $\check{G}_i$, which are determined by the applied potential, the temperature, the type of lead, and a counting field χ_i.

Once the network is determined and all connectors are specified, the transport properties can be found by means of the following circuit rules. We associate an (unknown) Green's function $\check{G}_j$ to each node j. The two rules are

1. $\check{G}_j^2 = \check{1}$ for the Green's functions of all internal nodes j.
2. The total matrix current in a node is conserved: $\sum_i \check{I}_{ij} = 0$, where the sum goes over all nodes or terminals connected to node j and each matrix current is given by (20).

Finally, the observable currents into the terminals are given by $I_i = \sum_j I_{ij}$, where the sum runs over all nodes connected to the terminal i. To obtain the counting statistics, we finally integrate all currents $I_i(\boldsymbol{\chi}) = (\partial/\partial\chi_i)S(\boldsymbol{\chi})$ to find the CGF $S(\boldsymbol{\chi})$.

6.2 Multi-tunnel Junction Structure

A general expression of $S(\boldsymbol{\chi})$ can be obtained for a system of an arbitrary number of terminals connected to one common node by tunnel contacts, see Fig. 3 [38,12]. At the same time it nicely demonstrates the application of the circuit theory rules, presented above. Let us denote the unknown Green's function of the central

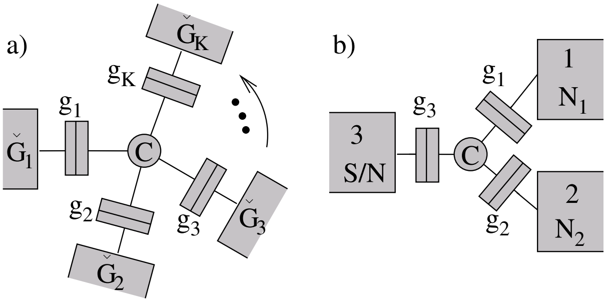


Fig. 3. Multi-tunnel junction structure: a) general setup with K terminals connected to a common node. b) beam splitter setup in which terminal 3 is either a normal metal or a superconductor

node by $\check{G}_c(\boldsymbol{\chi})$. The matrix current from a terminal α ($\alpha = 1, \dots, K$) into the central node is given by the relation

$$\check{I}_\alpha(\boldsymbol{\chi}) = \frac{g_\alpha}{2}\left[\check{G}_c(\boldsymbol{\chi}), \check{G}_\alpha(\chi_\alpha)\right] , \tag{38}$$

where $g_\alpha = G_Q \sum_n T_n$ is the conductance of the respective tunnel junction, for which we have assumed that all $T_n \ll 1$ and $g_\alpha \gg G_Q$ to avoid Coulomb blockade. The Green's function of the central node is determined by matrix current conservation, reading $\sum_{\alpha=1}^{K} \check{I}_\alpha = [\sum_{\alpha=1}^{K} g_\alpha \check{G}_\alpha, \check{G}_c]/2 = 0$. Employing the normalization condition $\check{G}_c^2 = 1$, the solution is

$$\check{G}_c(\boldsymbol{\chi}) = \frac{\sum_{\alpha=1}^{K} g_\alpha \check{G}_\alpha(\chi_\alpha)}{\sqrt{\sum_{\alpha,\beta=1}^{K} g_\alpha g_\beta \left\{\check{G}_\alpha(\chi_\alpha), \check{G}_\beta(\chi_\beta)\right\}}} . \tag{39}$$

To find the cumulant-generating function (CGF) $S(\boldsymbol{\chi})$ we integrate the equations $\partial S(\boldsymbol{\chi})/\partial \chi_\alpha = (-it_0/4e^2) \int dE \mathrm{Tr} \check{\tau}_K \check{I}_\alpha(\boldsymbol{\chi})$ [11]. We obtain

$$S(\boldsymbol{\chi}) = \frac{t_0}{2e^2} \int dE \mathrm{Tr} \sqrt{\sum_{\alpha,\beta=1}^{M} g_\alpha g_\beta \left\{\check{G}_\alpha(\chi_\alpha), \check{G}_\beta(\chi_\beta)\right\}} . \tag{40}$$

This is the general result for an M-terminal geometry in which all terminals are tunnel-coupled to a common node. It is valid for arbitrary combinations of normal metal and superconductor, fully accounting for the proximity effect. Note, that we have dropped the normalization of $S(\boldsymbol{\chi})$ to write the expression more compact.

6.3 Normal Metals

If all terminals are normal metals, the matrices in (40) are all diagonal and trace is trivial. We obtain

$$S(\boldsymbol{\chi}) = \frac{t_0}{2e^2} \int dE \sqrt{g_\Sigma^2 + \sum_{\alpha \neq \beta} g_\alpha g_\beta f_\alpha(E)(1 - f_\beta(E)) \left(e^{i(\chi_\alpha - \chi_\beta)} - 1\right)} , \tag{41}$$

where f_α is the occupation function of terminal α. Here, we introduced the abbreviation $g_\Sigma = \sum_{\alpha=1}^{N} g_\alpha$ for the sum of all conductances. We note, that the statistics is essentially non-Poissonian, despite the fact that we are considering tunnel junctions.

We now restrict us to two terminals (in which case we have to consider only one counting field $\chi = \chi_1 - \chi_2$). For zero temperature and voltage bias V the CGF reads then

$$S(\chi) = \frac{t_0 V}{2e} \sqrt{g_\Sigma^2 + 4 g_1 g_2 (e^{i\chi} - 1)} , \tag{42}$$

the result for a double tunnel junction first obtained by de Jong [13] using a master equation approach. We obtain as limiting cases for an asymmetric junction (either $g_1 \ll g_2$ or $g_1 \gg g_2$) Poisson statistics $S(\chi) = (t_0 V g_1 g_2/(g_1 + g_2))(\exp(i\chi) - 1)$.

Next we consider a three terminal structure, which is voltage biased such that the mean current $\bar{I}_3$ in lead 3 vanishes (voltage probe) and a transport current $\bar{I} = g_1 g_2/(g_1 + g_2)V$ flows between terminals 1 and 2. The CGF is [61]

$$S(\boldsymbol{\chi}) = \frac{t_0|V|}{2e} \left(g_2 \sqrt{g_\Sigma^2 + 4g_3 g_1 (e^{-i\chi_1} - 1) + 4g_1 g_2 (e^{i\chi_2 - i\chi_1} - 1)} \right.$$
$$\left. + g_1 \sqrt{g_\Sigma^2 + 4g_3 g_2 (e^{i\chi_2} - 1) + 4g_1 g_2 (e^{i\chi_2 - i\chi_1} - 1)} \right) . \quad (43)$$

It is interesting to note that the presence of the voltage probe makes the CGF asymmetric under the transformation $g_1 \leftrightarrow g_2$, whereas the current is symmetric. In certain limits in which the square roots in (43) can be expanded one is able to find the counting statistics. E. g. in the strong-coupling limit $g_3 \gg (g_1 + g_2)$ we find

$$S(\boldsymbol{\chi}) = \bar{N} \left[e^{-i\chi_1} + e^{i\chi_2} - 2 \right] . \quad (44)$$

The CGF is simply the sum of two Poisson distributions, demonstrating drastically the effect of the voltage probe. It completely suppresses the correlation between electrons entering and leaving the central node.

Another interesting geometry is a beam splitter configuration, in which a voltage bias is applied between one terminal and the other two. We find

$$S^N(\chi_1, \chi_2) = \frac{t_0|V|}{2e} \sqrt{g_\Sigma^2 + g_1 g_3 \left(e^{i\chi_1} - 1\right) + g_3 g_2 \left(e^{i\chi_2} - 1\right)} . \quad (45)$$

In the limit that $g_1 + g_2$ and g_3 are very different, we can expand the CGF and find for the CGF $S(\chi) = N_1 e^{i\chi_1} + N_2 e^{i\chi}$, i. e. the tunneling processes into the two terminals are uncorrelated. The corresponding probability distribution is simply the product of two Poisson distributions.

6.4 SN-Contact

We now consider the case of a double tunnel junction, in which one of the terminals is superconducting. From the general result (40) and (52) we find after some algebra

$$S(\chi) = \frac{t_0|V|}{e\sqrt{2}} \sqrt{g_1^2 + g_2^2 + \sqrt{\left(g_1^2 + g_2^2\right)^2 + 4g_1^2 g_2^2 (e^{i2\chi} - 1)}} . \quad (46)$$

Remarkably, the statistics is fundamentally different from the corresponding normal case (42). Still, the elementary events are transfers of pairs of electrons, which, however, are correlated in a more complicated way than normal electrons. If the junction is very asymmetric, the FCS reduces to Poissonian transfer of electron pairs. This is similar to the effect of decoherence between electrons and holes for energies of the order of the Thouless energy [34].

For the beam splitter configuration we are also able to find the FCS analytically. The CGF is [12]

$$S(\chi_1, \chi_2) = \frac{Vt_0}{\sqrt{2}e} \times \tag{47}$$

$$\sqrt{g_S^2 + \sqrt{g_S^4 + 4g_3^2 g_1^2 (e^{i2\chi_1} - 1) + 4g_3^2 g_2^2 (e^{i2\chi_2} - 1) + 8g_3^2 g_1 g_2 (e^{i(\chi_1+\chi_2)} - 1)}},$$

where we abbreviated $g_S^2 = g_3^2 + (g_1 + g_2)^2$. From this result we see that the elementary processes are now double charge transfers to either terminal of a splitting of a Cooper pair among the two terminals. It is interesting to note, that, if we assume that $g_1 + g_2$ and g_3 are very different (but $g_1 \approx g_2$), we obtain non-separable statistics

$$S(\chi) = N_{11} e^{i2\chi_1} + N_{22} e^{i2\chi_2} + N_{12} e^{i(\chi_1+\chi_2)} . \tag{48}$$

This expression can not be written as a sum of two independent terms. Furthermore, the last term is positive, which implies that current cross-correlation $S_{12} = -(2e^2/t_0)(\partial^2/\partial\chi_1\partial\chi_2)S(\chi_1, \chi_2)|_{\chi_1,\chi_2 \to 0}$ are *positive*. Equation (48) provides a simple explanation for this surprising effect: it is a consequence of independence of the different events, contributing to the current. This result, in fact, holds for a large class of superconducting beam splitters [36,39,62,63].

7 Conclusion

We have tried to give a pedagogical introduction to the field of counting statistics. Many technical details have been left out, but we have tried to cover the essence of the derivation and concentrated on looking at concrete examples. For a more thorough study we recommend the recent book *Quantum Noise in Mesoscopic Physics* [4] or the original literature. While a number of aspects have already been explored, many open questions remain, e. g. experimental strategies to measure FCS, strongly interacting systems, or spin-dependent problems. For the future, we expect even more activity in the field and, consequently, even more interesting results will emerge.

Acknowledgement

The ideas presented here are the results of numerous discussions with many people. In particular I would like to mention D. Bagrets, C. Bruder, J. C. Cuevas, Yu.V. Nazarov, and P. Samuelsson. This work was supported by the Swiss NSF and the NCCR Nanoscience.

Appendix

We summarize here the matrix-Green's function for superconducting and normal contact, as they were used in the text. The time-dependent Green's functions are expressed by their Fourier transforms $\check{G}_0(t-t') = \int (dE/2\pi)\ e^{-iE(t-t')}\check{G}_0(E)$. The energy-dependent Green's functions in the Keldysh×Nambu-space have the form

$$\check{G}(E) = \begin{pmatrix} (\bar{A}-\bar{R})\bar{f}+\bar{R} & (\bar{A}-\bar{R})\bar{f} \\ (\bar{A}-\bar{R})(1-\bar{f}) & (\bar{R}-\bar{A})\bar{f}+\bar{A} \end{pmatrix}, \tag{49}$$

where the advanced, retarded and occupation Nambu matrices are

$$\bar{A}(\bar{R}) = \begin{pmatrix} g_{A(R)} & f_{A(R)} \\ f_{A(R)} & -g_{A(R)} \end{pmatrix} \quad , \quad \bar{f}(E) = \begin{pmatrix} f(E) & 0 \\ 0 & f(-E) \end{pmatrix}. \tag{50}$$

The phase φ of the superconducting order parameter as well as the electrical potential eV enter via the gauge transformation $\check{G}(t,t') = \check{U}(t)\check{G}_0(t-t')\check{U}^\dagger(t')$. Here $\check{U}(t) = \exp\left[i\phi(t)\bar{\tau}_3/2\right]$, where $\phi(t) = \varphi + eVt$.

In the calculation of the FCS of contacts between normal metals and superconductors we frequently need the eigenvalues of anti-commutators of two Green's functions. For two normal metals $\{\check{G}_{N1}(\chi), \check{G}_{N2}\}/2$ is diagonal and the eigenvalue is

$$\left[1 + 2f_1(E)\left(1-f_2(E)\right)\left(e^{i\chi}-1\right) + 2f_2(E)\left(1-f_1(E)\right)\left(e^{-i\chi}-1\right)\right], \tag{51}$$

for the electron block and the same expression with $E \to -E$ for the 'hole'-block in Nambu space.

In the case of Andreev reflection, i. e. for $eV, k_BT_{el} \ll \Delta$, we find for $\{\check{G}_N(\chi), \check{G}_S\}/2$ the two eigenvalues

$$\pm\sqrt{f_N(E)f_N(-E)\left(1-e^{i2\chi}\right) + (1-f_N(E))(1-f_N(-E))\left(1-e^{-i2\chi}\right)}. \tag{52}$$

References

1. L. Mandel, E. Wolf: *Optical Coherence and Quantum Optics* (Cambridge University Press, Cambridge 1995).
2. L.S. Levitov, G.B. Lesovik: Pis'ma Zh. Eksp. Teor. Fiz. **58**, 225 (1993).
3. Ya.M. Blanter, M.Büttiker: Phys. Rep. **336**, 1 (2000).
4. *Quantum Noise in Mesoscopic Physics*, ed. by Yu.V. Nazarov (Kluwer, Dordrecht 2003).
5. H. Lee, L.S. Levitov, A.Yu. Yakovets: Phys. Rev. B **51**, 4079 (1996).
6. L.S. Levitov, H.W. Lee, G.B. Lesovik: J. Math. Phys. **37**, 4845 (1996).
7. H. Lee, L.S. Levitov: Phys. Rev. B **53**, 7383 (1996).
8. Yu.V. Nazarov: Ann. Phys. (Leipzig) **8**, SI-193 (1999).
9. W. Belzig, Yu.V. Nazarov: Phys. Rev. Lett. **87**, 197006 (2001).
10. W. Belzig, Yu.V. Nazarov: Phys. Rev. Lett. **87**, 067006 (2001).
11. Yu.V. Nazarov, D. Bagrets: Phys. Rev. Lett. **88**, 196801 (2002).
12. J. Börlin, W. Belzig, C. Bruder: Phys. Rev. Lett. **88**, 197001 (2002).
13. M.J.M. de Jong: Phys. Rev. B **54**, 8144 (1996).
14. D.A. Bagrets, Yu.V. Nazarov: Phys. Rev. B **67**, 085316 (2003).
15. S. Pilgram *et al.*: Phys. Rev. Lett. **90**, 206801 (2003).
16. L.S. Levitov, G.B. Lesovik: Pis´ma Zh. Eksp. Teor. Fiz. **55**, 534 (1992).
17. Yu. Makhlin, G. Schön, A. Shnirman: Phys. Rev. Lett. **85**, 4578 (2000).

18. Yu.V. Nazarov, M. Kindermann: cond-mat/0107133.
19. A. Shelankov, J. Rammer: Europhys. Lett. **63**, 485 (2003).
20. I. Klich, in [4]
21. G.B. Lesovik, N.M. Chtchelkatchev: Pis´ma Zh. Eksp. Teor. Fiz. **77**, 464 (2003).
22. C.W.J. Beenakker, H. Schomerus: Phys. Rev. Lett. **86**, 700 (2001).
23. B.A. Muzykantskii, D.E. Khmelnitzkii: Phys. Rev. B **50**, 3982 (1994).
24. Ya.M. Blanter, H. Schomerus, C.W.J. Beenakker: Physica E **11**, 1 (2001).
25. L.S. Levitov, M. Reznikov: cond-mat/0111057 (unpublished).
26. D. Gutman, Y. Gefen, A. Mirlin: cond-mat/0210076.
27. A. Andreev, A. Kamenev: Phys. Rev. Lett. **85**, 1294 (2000).
28. L.S. Levitov: cond-mat/0103617.
29. Yu. Makhlin, A. Mirlin: Phys. Rev. Lett. **87**, 276803 (2001).
30. B.A. Muzykantskii, Y. Adamov: Phys. Rev. B **68**, 155304 (2003).
31. M.-S. Choi, F. Plastina, R. Fazio: Phys. Rev. Lett. **87**, 116601 (2001); Phys. Rev. B **67**, 045105 (2003).
32. H.-A. Engel, D. Loss: Phys. Rev. B **65**, 195321 (2002).
33. A.A. Clerk: cond-mat/0301277.
34. P. Samuelsson: Phys. Rev. B **67**, 054508 (2003).
35. B. Reulet *et al.*: Phys. Rev. Lett. **90**, 066601 (2003).
36. W. Belzig, P. Samuelsson: Europhys. Lett. **64**, 253 (2003).
37. E.V. Bezuglyi, E.N. Bratus', V.S. Shumeiko, V. Vinokur: cond-mat/0311176.
38. Yu.V. Nazarov, D. Bagrets: Phys. Rev. Lett. **88**, 196801 (2002).
39. P. Samuelsson, M. Büttiker: Phys. Rev. B **66**, 201306(R) (2002).
40. F. Taddei, R. Fazio: Phys. Rev. B **65**, 075317 (2002).
41. L. Faoro, F. Taddei, R. Fazio: cond-mat/0306733.
42. M. Kindermann, C.J.W. Beenakker: Phys. Rev. B **66**, 224106 (2002).
43. M. Kindermann, Yu.V. Nazarov, C.W.J. Beenakker: Phys. Rev. Lett. **88**, 063601 (2002).
44. M. Kindermann, Yu.V. Nazarov, C.W.J. Beenakker: Phys. Rev. Lett. **90**, 246805 (2003).
45. D.A. Bagrets, Yu.V. Nazarov: cond-mat/0304339.
46. M. Kindermann, Yu.V. Nazarov: Phys. Rev. Lett. **91**, 136802 (2003).
47. J.C. Cuevas, W. Belzig: Phys. Rev. Lett. **91**, 187001 (2003).
48. G. Johansson, P. Samuelsson, A. Ingerman: Phys. Rev. Lett. **91**, 187002 (2003).
49. B. Reulet, J. Senzier, D.E. Prober: Phys. Rev. Lett. **91**, 196601 (2003).
50. C.W.J. Beenakker, M. Kindermann, Yu.V. Nazarov: Phys. Rev. Lett. **90**, 176802 (2003).
51. K.E. Nagaev, S. Pilgram, M. Büttiker: cond-mat/0306465.
52. A.V. Galaktionov, D.S. Golubev, A.D. Zaikin: cond-mat/0308133.
53. J. Rammer, H. Smith: Rev. Mod. Phys. **58**, 323 (1986).
54. A.A. Abrikosov, L.P. Gorkov, I.E. Dzyaloshinski: *Methods of Quantum Field Theory in Statistical Physics*, (Dover, New York 1963).
55. Yu.V. Nazarov: Superlattices Microst. **25**, 1221 (1999).
56. C.W.J. Beenakker: Rev. Mod. Phys. **69**, 731 (1997).
57. L.S. Levitov in [4].
58. W. Belzig, in [4].
59. M. Houzet, F. Pistolesi: cond-mat/0310418.
60. Yu.V. Nazarov: Phys. Rev. Lett. **73**, 134 (1994).
61. J. Börlin: Diploma Thesis, University of Basel (2002).
62. P. Samuelsson, M. Büttiker: Phys. Rev. Lett. **89**, 046601 (2002).
63. F. Taddei, R. Fazio: Phys. Rev. B **65**, 134522 (2002).

Quantum Dots Attached to Ferromagnetic Leads: Exchange Field, Spin Precession, and Kondo Effect

Jürgen König[1], Jan Martinek[2,3,4], Józef Barnaś[4,5], and Gerd Schön[2]

[1] Institut für Theoretische Physik III, Ruhr-Universität Bochum, 44780 Bochum, Germany

[2] Institut für Theoretische Festkörperphysik, Universität Karlsruhe, 76128 Karlsruhe, Germany

[3] Institute for Materials Research, Tohoku University, Sendai 980-8577, Japan

[4] Institute of Molecular Physics, Polish Academy of Sciences, 60-179 Poznań, Poland

[5] Department of Physics, Adam Mickiewicz University, 61-614 Poznań, Poland

Abstract. Spintronics devices rely on spin-dependent transport behavior evoked by the presence of spin-polarized electrons. Transport through nanostructures, on the other hand, is dominated by strong Coulomb interaction. We study a model system in the intersection of both fields, a quantum dot attached to ferromagnetic leads. The combination of spin-polarization in the leads and strong Coulomb interaction in the quantum dot gives rise to an exchange field acting on electron spins in the dot. Depending on the parameter regime, this exchange field is visible in the transport either via a precession of an accumulated dot spin or via an induced level splitting. We review the situation for various transport regimes, and discuss two of them in more detail.

1 Introduction

The study of spin-dependent tunneling through quantum dots resides in the intersection of two active and attractive fields of physics, namely spintronics [1–3] and transport through nanostructures [4–6]. Both the investigation of spin-dependent electron transport on the one hand and the study of strong Coulomb interaction effects in transport through nanostructures on the other hand define by now well-established research areas. The combination of both concepts within one system is, however, a very new field which is still in its early stages. Its attractiveness originates from the rich physics expected from the combination of two different paradigms. A suitable model system for a basic study of the interplay of spin-dependent transport due to spin polarization in ferromagnetic electrodes and Coulomb charging effects in nanostructures is provided by a quantum dot attached to ferromagnetic leads.

1.1 Some Concepts of Spintronics

The field of spin- or magnetoelectronics [1–3] has attracted much interest, for both its beautiful fundamental physics and its potential applications. A famous example, which has already proven technological relevance, is the spin valve based on either the giant magnetoresistance effect (GMR) in magnetic multilayers

J. König et al., Quantum Dots Attached to Ferromagnetic Leads: Exchange Field, Spin Precession, and Kondo Effect, Lect. Notes Phys. **658**, 145–164 (2005)
http://www.springerlink.com/

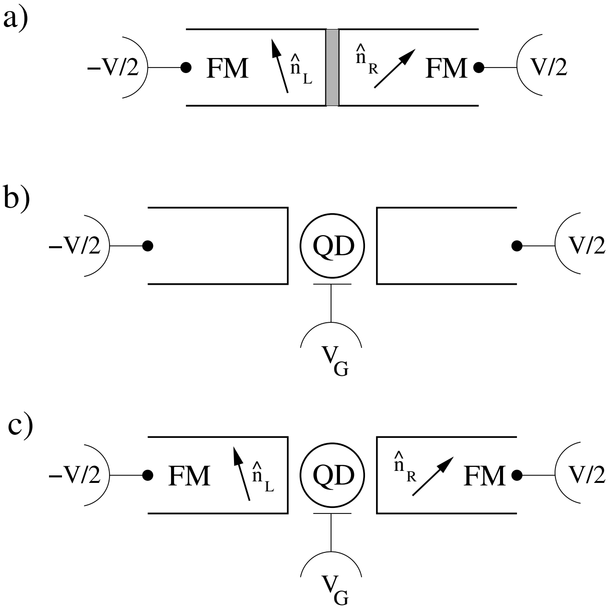


Fig. 1. a) Spin valve: a single tunnel junction between two ferromagnets (FM) with magnetization orientations $\hat{\mathbf{n}}_{\mathrm{L}}$ and $\hat{\mathbf{n}}_{\mathrm{R}}$, respectively. b) Quantum dot. c) Quantum-dot spin valve: a quantum dot is connected to two ferromagnetic leads (FM)

or the tunnel magnetoresistance (TMR) in magnetic tunnel junctions. In both cases, the transport properties depend on the relative magnetization orientation of the magnetic layers or leads involved, an information conveyed by the spin polarization of the transported electrons. In the case of a single magnetic tunnel junction, the tunneling current is maximal for parallel alignment of the leads' magnetization orientations, while it is minimal for antiparallel alignment. This can be easily understood within a non-interacting-electron picture, as proposed by Jullière [7]: the tunnel current of electrons with given spin direction is proportional to the product of the corresponding spin-dependent densities of states in the source and drain electrode, which leads to a reduction of transport in the case of antiparallel alignment.

This concept has been extended [8] to describe also noncollinear arrangements, as depicted in Fig. 1a, where the magnetization directions of the leads enclose an arbitrary angle ϕ. In this situation, the ϕ-dependent part of the tunneling current is proportional to the overlap of the spinor part of the majority-spin wave functions in the source and drain electrode, i.e. proportional to $\cos\phi$, as it has been experimentally confirmed recently [9].

In heterostructures that consist of a nonmagnetic metal sandwiched by ferromagnetic electrodes, the concept of spin accumulation becomes important. Once the spin diffusion length is larger than the size of the nonmagnetic region, the information about the relative orientation of the leads' magnetization is mediated through the middle part. In the antiparallel configuration an applied bias voltage leads to a pile-up of spin in the nonmagnetic metal, since electrons with one type of spin (say spin up) are preferentially injected from the source electrode, while electrons with the other type of spin (spin down) are pulled out from the drain electrode. This piling up of spin splits the chemical potentials for spin-up and spin-down electrons in the normal metal such that electrical transport through the whole device is reduced.

As spin is a vector quantity, transport through a ferromagnetic-nonmagnetic-ferromagnetic heterostructure can be tuned by manipulating the direction of the spins in the middle part. The prototype for such a concept is the spin field-effect transistor proposed by Datta and Das [10]. Spin-polarized electrons are injected from a ferromagnetic metal into a ballistic conducting channel provided by a two-dimensional electron gas in a semiconductor heterostructure. Due to the Rashba effect, the electrons in the semiconductor experience a spin-orbit coupling, whose strength can be tuned by a gate voltage. This spin-orbit coupling leads to a rotation of the spins in the conducting channel as they move along towards the drain electrode. The total transmission through the device, then, depends on the relative orientation of the rotated spins and the magnetization direction of the drain electrode.

1.2 Transport Through Nanostructures

Tunneling transport through nanostructures, such as semiconductor quantum dots (Fig. 1b) or small metallic islands, is strongly affected by Coulomb interaction, and a non-interacting electron picture is no longer applicable [4–6]. Coulomb-blockade phenomena arise at low temperature, such that the corresponding energy scale is smaller than the charging energy, the energy scale for adding or removing one electron from the dot or island. Small quantum dots with a size of the order of the Fermi wavelength have a discrete level spectrum. If the level spacing is large enough, transport through single levels is possible. This situation defines a simplest but very generic model, the Anderson-impurity model, for studying Coulomb interaction in nanostructures.

When the level is occupied with one electron since double occupancy is prohibited by charging energy, the dot possesses a local spin. At low temperature and large dot-lead tunnel-coupling strength, a ground state with complex manybody correlations forms, which manifests itself in the so-called Kondo effect [11]. The local spin is screened by the spins of the conduction electrons in the leads, and accompanied with this, electrical transport through the quantum dot is strongly enhanced.

1.3 Quantum-Dot Spin Valves

The scheme of a quantum-dot spin valve, a quantum dot attached to ferromagnetic leads, is illustrated in Fig. 1c. Successful fabrication of either quantum-dot systems or magnetic heterostructures has been achieved by a large number of experimental groups. To attach ferromagnetic electrodes to quantum dots, though, is quite a challenging task, and only very recently first results have been reported.

Let us start with metallic single-electron devices. Both spin-dependent tunneling and Coulomb blockade have been found in magnetic tunnel junction with embedded Co clusters [12]. All-ferromagnetic metallic single-electron transistors have been manufactured, using either single-island [13,14], or multi-island structures [15,16]. Magnetoresistance of single-electron transistors with a normal metallic island in a cobalt-aluminum-cobalt structure has been measured [17]. In all these examples, the level spectrum on the island is continuous, and many levels are involved in transport.

Our focus, however, is on single-level quantum dots. The difficulty lies in the incompatibility of the usual materials showing ferromagnetism (metals) and those usually forming quantum dots (semiconductors). There are different strategies to overcome this problem. One possibility is the use of ferromagnetic semiconductors (Ga,Mn)As as lead electrodes coupled to, e.g., self-assembled InAs quantum dots [18]. A very promising approach is to contact an ultrasmall aluminum nanoparticle, which serves as a quantum dot, to ferromagnetic metallic electrodes. In this way, quantum dots with one magnetic (nickel or cobalt) and one nonmagnetic (aluminum) electrodes have been fabricated [19]. Another important system is a magnetic impurity inside the tunneling barrier of ferromagnetic tunnel junction [20]. An alternative route is to use carbon nanotubes as quantum dots and to place them on metallic contacts. Coulomb-blockade phenomena and even the Kondo effect has been observed in such systems [21,22]. Spin-dependent transport through carbon nanotubes attached to ferromagnetic electrodes has been investigated in [23,24]. A more challenging scheme is a ferromagnetic single-molecule transistor [25], where a single molecule is attached to ferromagnetic electrodes. To some extent, there is also a relation between the quantum-dot spin valve and a single magnetic-atom spin on a scanning tunneling microscope tip. For the latter, precession of the single spin in an external magnetic field has been detected in the power spectrum of the tunneling current [26].

This progress on the experimental side has stimulated a number of theoretical activities [27–40] on spin-dependent transport through either metallic single-electron transistors or quantum dots.

The motivation for studying quantum-dot spin valves can be formulated from two different perspectives, depending on from which side one starts to approach the problem. Coming from the spintronics side, one may ask how the concepts introduced there, such as spin accumulation and spin manipulation, manifest themselves in quantum dots, and how the presence of strong Coulomb interaction gives rise to qualitatively new behavior as compared to non-interacting electron systems. On the other hand, when starting from Coulomb-interaction effects in quantum dots, one may ask how the spin-polarization of the leads changes the picture. As mentioned above, the screening of a local spin on the quantum dot by the lead-electron spins is crucial for the Kondo effect to develop. This screening behavior is affected by spin asymmetry introduced due to a finite spin polarization of the lead. In this case, it is a priori not clear whether a Kondo-correlated state can still form or not.

To comprise all this in a single question, we ask whether the combination of strong Coulomb interaction and finite spin-polarization gives rise to qualitatively new phenomena that are absent for either non-interacting or unpolarized electrons. The answer is: yes, it does. We predict that single electrons on the quantum dot experience an exchange field, which effectively acts like a local magnetic field. The main goal of this paper is to illustrate the origin of this exchange field, its properties, and its implications on transport. Of course, the latter depends on the considered transport regime, and the observable consequences can be quite different. In the present paper, we concentrate on two particular regimes, the case of weak dot-lead coupling but noncollinear magnetization directions and the case of very strong coupling but collinear configuration. Other limits will only be commented on shortly, as for these cases work is still in progress and will be presented elsewhere.

2 The Model

We consider a small quantum dot with one energy level ϵ participating in transport. The dot is coupled to ferromagnetic leads, see Fig. 1c. The left and right lead are magnetized along $\hat{\boldsymbol{n}}_{\mathrm{L}}$ and $\hat{\boldsymbol{n}}_{\mathrm{R}}$, respectively. The total Hamiltonian is

$$H = H_{\mathrm{dot}} + H_{\mathrm{L}} + H_{\mathrm{R}} + H_{\mathrm{T,L}} + H_{\mathrm{T,R}}\,. \tag{1}$$

The first part, $H_{\mathrm{dot}} = \epsilon \sum_\sigma c_\sigma^\dagger c_\sigma + U n_\uparrow n_\downarrow$, describes the dot energy level plus the charging energy U for double occupation. In the presence of an external magnetic field, the energy level experiences a Zeeman splitting, i.e., becomes spin-dependent. The leads are modeled by $H_r = \sum_{k\sigma} \epsilon_{k\sigma} a_{rk\sigma}^\dagger a_{rk\sigma}$ with $r = \mathrm{L}, \mathrm{R}$. In the spirit of a Stoner model of ferromagnetism [41], there is a strong spin asymmetry in the density of states $\rho_{r\sigma}(\omega)$ for majority ($\sigma = +$) and minority ($\sigma = -$) spins. Throughout all of our calculations presented here, we approximate the density of states to be energy independent, $\rho_{r\sigma}(\omega) = \rho_{r\sigma}$. Real ferromagnets will have a structured density of states [42]. This fact, however, will only modify details of the results and not the main physical picture. The ratio $p = (\rho_{r+} - \rho_{r-})/(\rho_{r+} + \rho_{r-})$ characterizes the degree of spin polarization in the leads. For simplicity, we assume here $\rho_{\mathrm{L}+} = \rho_{\mathrm{R}+} \equiv \rho_+$ and $\rho_{\mathrm{L}-} = \rho_{\mathrm{R}-} \equiv \rho_-$. Nonmagnetic leads are described by $p = 0$, and $p = 1$ represents half metallic leads, which accommodate majority spins only. We emphasize that the magnetization directions of leads can differ from each other, enclosing an angle ϕ.

Tunneling between leads and dot is described by the standard tunneling Hamiltonian. For the left tunnel barrier we get

$$H_{\mathrm{T,L}} = t \sum_{k\sigma=\pm} \left(a_{\mathrm{L}k\sigma}^\dagger c_\sigma + h.c. \right) , \tag{2}$$

where $c_\pm$ are the Fermi operators for an electron on the quantum dot with spin along $\pm\hat{\boldsymbol{n}}_r$. For the right barrier, an analogous expression holds. As $\hat{\boldsymbol{n}}_{\mathrm{L}}$ may differ from $\hat{\boldsymbol{n}}_{\mathrm{R}}$, an ambiguity arises in the definition of $c_\pm$. This is no problem for collinear, i.e., parallel or antiparallel, configuration of the leads. In this case, $\hat{\boldsymbol{n}}_{\mathrm{L}} = \pm\hat{\boldsymbol{n}}_{\mathrm{R}}$ provides a natural quantization axis for the dot spin.

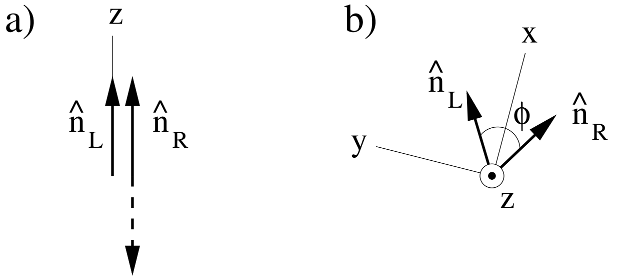


Fig. 2. Choice of the used coordinate system: a) For collinear configuration of the leads' magnetization, i.e., parallel (solid arrow for $\hat{\boldsymbol{n}}_\mathrm{R}$) or antiparallel (dashed arrow), the z-axis is along $\hat{\boldsymbol{n}}_\mathrm{L}$. In this case, we use the tunneling Hamiltonian in the form of (2). b) For noncollinear arrangements, the z-axis is perpendicular to both $\hat{\boldsymbol{n}}_\mathrm{L}$ and $\hat{\boldsymbol{n}}_\mathrm{R}$. Here, the tunneling Hamiltonian in the form of (3) is used. The quantum-dot spin is always quantized along the z-axis

For noncollinear leads, however, the form (2) of the tunnel Hamiltonian is no longer useful. To describe the scenario properly, we find it convenient to quantize the dot spin neither along $\hat{\boldsymbol{n}}_\mathrm{L}$ nor $\hat{\boldsymbol{n}}_\mathrm{R}$, but along the axis perpendicular to both $\hat{\boldsymbol{n}}_\mathrm{L}$ and $\hat{\boldsymbol{n}}_\mathrm{R}$. To be explicit, we choose the coordinate system defined by $\hat{\boldsymbol{e}}_x = (\hat{\boldsymbol{n}}_\mathrm{L} + \hat{\boldsymbol{n}}_\mathrm{R})/|\hat{\boldsymbol{n}}_\mathrm{L} + \hat{\boldsymbol{n}}_\mathrm{R}|$, $\hat{\boldsymbol{e}}_y = (\hat{\boldsymbol{n}}_\mathrm{L} - \hat{\boldsymbol{n}}_\mathrm{R})/|\hat{\boldsymbol{n}}_\mathrm{L} - \hat{\boldsymbol{n}}_\mathrm{R}|$, and $\hat{\boldsymbol{e}}_z = (\hat{\boldsymbol{n}}_\mathrm{R} \times \hat{\boldsymbol{n}}_\mathrm{L})/|\hat{\boldsymbol{n}}_\mathrm{R} \times \hat{\boldsymbol{n}}_\mathrm{L}|$, and quantize the dot spin along the z-direction, see Fig. 2. The tunnel Hamiltonian, then, becomes

$$H_{\mathrm{T,L}} = \frac{t}{\sqrt{2}} \sum_k (a^\dagger_{\mathrm{L}k+}, a^\dagger_{\mathrm{L}k-}) \begin{pmatrix} e^{i\phi/4} & e^{-i\phi/4} \\ e^{i\phi/4} & -e^{-i\phi/4} \end{pmatrix} \begin{pmatrix} c_\uparrow \\ c_\downarrow \end{pmatrix} + h.c. \,, \tag{3}$$

and $H_{\mathrm{T,R}}$ is the same but with L → R and $\phi \to -\phi$. The special choice of the coordinate system implies that both up and down spins of the dot are equally-strongly coupled to the majority and minority spins of the leads. There, are, however, phase factors $e^{\pm i\phi/4}$ are involved, similar to multiply-connected quantum-dot systems dubbed Aharonov-Bohm interferometers [43]. The two spin directions ↑ and ↓ in the dot correspond to the quantum dots placed in the two arms of the Aharonov-Bohm interferometer, and the angle ϕ plays the role of the Aharonov-Bohm phase, which measures the total magnetic flux enclosed by the arms of the interferometer in units of the flux quantum. We note, however, that our model translates to a very special kind of Aharonov-Bohm interferometer: the dot in each interferometer arm accommodates only a single level instead of a doubly-degenerate one, and Coulomb interaction occurs between the two dots, instead of within each of them.

The two different choices we use for the collinear and noncollinear configuration, in which we use either (2) or (3), respectively, are illustrated in Fig. 2. In both cases, the tunnel coupling leads to a finite width of the dot level. Its energy scale is given by $\Gamma = \sum_r \Gamma_r$ with $\Gamma_r = \pi |t|^2 \sum_{\sigma=\pm} \rho_\sigma$ [44].

3 Exchange Field

As pointed out in the introduction, the qualitative new physics introduced by the combination of spin-polarized leads and strong Coulomb interaction in the dot, is the existence of an exchange field acting on electron spins in the dot. This exchange field is intrinsically present in the model described by the Hamiltonian (1) together with the spin-dependent density of states. It is, therefore, automatically contained in any consistent treatment of the model for a given transport regime, as we will see in the subsequent sections. Nothing has to be added by hand. Nevertheless, we find it instructive to derive an explicit analytic expression by making use of the following heuristic procedure.

Each of the two leads will contribute to the exchange field separately. To keep the discussion transparent, we consider the effect of one lead only. The total exchange field is, then, just the sum over both leads. The first step is to derive an effective Hamiltonian for the subspace of the total Hilbert space in which the quantum dot is singly occupied. This is the regime of interest, as far as the exchange field in concerned, since both an empty and a doubly-occupied dot have zero total spin, and an exchange field would be noneffective. By taking into account virtual excitations to an empty or doubly-occupied dot within lowest-order perturbation theory in the tunnel coupling, in analogy to the Schrieffer-Wolff transformation [11] employed in the context of Kondo physics for magnetic impurities in nonmagnetic metals, we arrive at an effective spin model for the dot spin operators $S^{\pm}$ and S^z (quantized along the magnetization direction of the considered lead),

$$\begin{aligned} H_{\text{spin}} = {} & S^{+}|t|^2 \sum_{kq} \left(\frac{1}{U+\epsilon-\epsilon_q} + \frac{1}{\epsilon_k - \epsilon} \right) a^{\dagger}_{rk\downarrow} a_{rq\uparrow} \\ & + S^{-}|t|^2 \sum_{kq} \left(\frac{1}{U+\epsilon-\epsilon_k} + \frac{1}{\epsilon_q - \epsilon} \right) a^{\dagger}_{rq\uparrow} a_{rk\downarrow} \\ & + S^{z}|t|^2 \left(\sum_{qq'} \frac{1}{U+\epsilon-\epsilon_{q'}} a^{\dagger}_{rq\uparrow} a_{rq'\uparrow} - \sum_{kk'} \frac{1}{U+\epsilon-\epsilon_{k'}} a^{\dagger}_{rk\downarrow} a_{rk'\downarrow} \right) \\ & - S^{z}|t|^2 \left(\sum_{qq'} \frac{1}{\epsilon_q - \epsilon} a_{rq'\uparrow} a^{\dagger}_{rq\uparrow} - \sum_{kk'} \frac{1}{\epsilon_k - \epsilon} a_{rk'\downarrow} a^{\dagger}_{rk\downarrow} \right) . \end{aligned} \tag{4}$$

Note that the information about the different densities of states for up- and down-spins is included in the summation over q, q' (used for spin-up electrons) and k, k' (used for spin down), respectively. In addition, there is a term describing potential scattering, but this does not contribute to the exchange field we are aiming at.

In a second step we employ in (4) a mean-field approximation for the lead-electron states, making use of $\langle a^{\dagger}_{rk\sigma} a_{rk'\sigma'} \rangle = f_r(\epsilon_{k\sigma}) \delta_{kk'} \delta_{\sigma\sigma'}$ and $\langle a_{rk\sigma} a^{\dagger}_{rk'\sigma'} \rangle = [1 - f_r(\epsilon_{k\sigma})] \delta_{kk'} \delta_{\sigma\sigma'}$, where $f_r(\omega)$ is the Fermi function of lead r. The terms proportional to $S^{\pm}$ drop out. The resulting effective Hamiltonian, then, reads

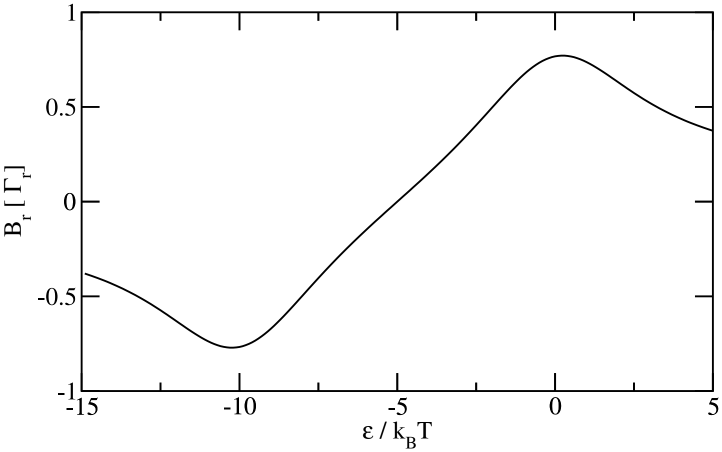


Fig. 3. The exchange field as a function of the level position ϵ for $U/k_\mathrm{B}T = 10$ and $p = 1$

$H_\mathrm{eff} = -S^z B_r$ with the exchange field (for simplicity we include the gyromagnetic factor in the definition)

$$B_r = \int' d\omega (\rho_+ - \rho_-)|t|^2 \left(\frac{1 - f_r(\omega)}{\omega - \epsilon} + \frac{f_r(\omega)}{\omega - \epsilon - U} \right) \tag{5}$$

$$= -\frac{p\Gamma_r}{\pi} \mathrm{Re} \left[\Psi \left(\frac{1}{2} + i \frac{\beta(\epsilon - \mu_r)}{2\pi} \right) - \Psi \left(\frac{1}{2} + i \frac{\beta(\epsilon + U - \mu_r)}{2\pi} \right) \right] , \tag{6}$$

where $\Psi(x)$ denotes the digamma function, μ_r is the electrochemical potential of lead r, and the prime at the integral sign in (5) symbolizes Cauchy's principal value. For illustration, we plot the exchange field as a function of the level position in Fig. 3.

From the explicit form (6) of the exchange field we derive the following properties:

(i) It vanishes in the case of a non-interacting quantum dot, $U = 0$.
(ii) The exchange field is proportional to the degree of spin-polarization p in the lead. This means that both strong Coulomb interaction and finite spin-polarization are required to generate the exchange field.
(iii) It depends on the tunnel coupling strength Γ. In the treatment lined out above, Γ enters linearly as a global prefactor.
(iv) The magnitude and even the sign of the exchange field depends on the level position ϵ. In particular, there is a value of ϵ at which the exchange field vanishes (in our model with flat density of states this happens at $\epsilon - \mu_r = U/2$, i.e., when the total system is particle-hole symmetric).

Furthermore, we notice from (5) that not only electronic states around the Fermi energy of the lead are involved. Instead, it is rather the full band that matters. This means that a precise simulation of realistic materials requires a knowledge of the detailed density of states, to be inserted in the integral in (5).

This will modify the details of the exchange field such as its precise dependence on the level position ϵ.

4 Transport Regimes

After introducing the notion of the exchange field, the question of how it affects the transport behavior arises immediately. The answer to this question depends on the transport regime under consideration. In particular, we will identify two mechanisms by which the exchange field enters. One scenario is the generation of a level splitting between up and down spins in the quantum dot, with the level splitting given by the exchange field (6). But this is not the only possibility. Even in situations where the generated level splitting is negligible, the exchange field can affect the dot state and, thus, the transport behavior by rotating an accumulated spin on the dot, which can pile up there in non-equilibrium due to an applied bias voltage. A complete picture of the various different transport regimes goes beyond the scope of the present paper. Instead, we will concentrate on two specific limits, namely weak dot-lead coupling but noncollinear magnetization in linear response, and strong coupling but collinear configuration of the leads. For some other regimes, that are currently under investigation, we will only give some short comments and refer the reader to forthcoming publications.

In the limit of weak dot-lead coupling, $\Gamma \ll k_{\mathrm{B}}T$, referred to as sequential-tunneling regime, transport is dominated by processes of first order in Γ (unless both ϵ and $\epsilon + U$ are shifted into the Coulomb-blockade region). First-order transport probes the state of the quantum dot to zeroth order (since the tunneling between dot and leads necessary for transport already trivially involves a factor Γ). Therefore, the level splitting generated by the exchange field cannot be probed by first-order transport. Nevertheless, the exchange field plays a role via the second of the above mentioned mechanisms. Once a finite spin is accumulated on the quantum dot, with a direction noncollinear to the exchange field, the latter will induce a precession of the accumulated spin. For this to happen, a noncollinear configuration of the leads' magnetic moments is required, as otherwise accumulated spin, if any, and exchange field are pointing in the same direction.

In the Coulomb-blockade regime, sequential tunneling is exponentially suppressed, and transport is dominated by cotunneling, which are second-order processes. But also on resonance, second-order corrections become important for intermediate coupling strengths, $\Gamma \sim k_{\mathrm{B}}T$. Second-order transport is affected by the generated level splitting, and the exchange field plays a role even for a collinear arrangement of the leads' magnetization.

A very dramatic signature of the level splitting generated by the exchange field is predicted for the limit of low temperature and large coupling strength, $k_{\mathrm{B}}T \le k_{\mathrm{B}}T_{\mathrm{K}} \ll \Gamma$, for which the Kondo effect can appear (T_{K} is the Kondo temperature). Since a finite level splitting, e.g., due to a Zeeman term induced by an external magnetic field, quickly destroys the Kondo effect, the exchange field has quite an important, at first glance destructive, consequence. As we will see below in more detail, however, by applying an appropriately-tuned external magnetic field one can compensate for the induces level splitting and, thus, recover the Kondo effect. For this discussion, we restrict ourselves to collinear configurations.

4.1 First-Order Transport in Linear Response

Here, we only present the major steps and main results. Details of the calculations can be found in [30,31]. The first step is to relate the linear conductance $G^{\text{lin}} = (\partial I/\partial V)\big|_{V=0}$ to the Green's functions of the dot. For first-order transport, we obtain

$$G^{\text{lin}} = \frac{e^2}{h}\Gamma \int d\omega \; \Big\{ \text{Im}\, G^{\text{ret}}_{\downarrow\downarrow}(\omega) f'(\omega) \\ + p \sin\frac{\phi}{2} \left[f(\omega)\,\frac{\partial G^{>}_{\downarrow\uparrow}(\omega)}{\partial(eV)} + [1-f(\omega)]\,\frac{\partial G^{<}_{\downarrow\uparrow}(\omega)}{\partial(eV)} \right] \Big\} \; . \quad (7)$$

Here, $f(\omega)$ is the Fermi function, $G_{\sigma\sigma'}(\omega)$ are the Fourier transforms of the usual retarded, greater and lesser Green's functions. Contributions involving the Green's functions $G_{\uparrow\uparrow}(\omega)$ and $G_{\uparrow\downarrow}(\omega)$ are accounted for in a prefactor 2. Since Γ already appears explicitly in front of the integral, all Green's functions are to be taken to zeroth order in Γ. In this limit, we find $-(1/\pi)\text{Im}\, G^{\text{ret}}_{\downarrow\downarrow}(\omega) = (P^0_0 + P^{\downarrow}_{\downarrow})\delta(\omega - \epsilon) + (P^{\uparrow}_{\uparrow} + P^{d}_{d})\delta(\omega - \epsilon - U)$, $G^{>}_{\downarrow\uparrow}(\omega) = 2\pi i P^{\downarrow}_{\uparrow}\delta(\omega - \epsilon - U)$, and $G^{<}_{\downarrow\uparrow}(\omega) = 2\pi i P^{\downarrow}_{\uparrow}\delta(\omega - \epsilon)$, where $P^{\chi}_{\chi'} = \langle |\chi'\rangle\langle\chi| \rangle$ are elements of the stationary density matrix (to zeroth order in Γ) of the quantum-dot subsystem, with $\chi, \chi' = 0$ (empty dot), $\uparrow, \downarrow$ (singly-occupied dot), and d (doubly-occupied dot).

The main task is now to determine the density-matrix elements to zeroth order in Γ. They contain the information about the average occupation and spin on the quantum dot. The diagonal matrix elements, P^{χ}_{χ}, are nothing but the probabilities to find the quantum dot in state χ, i.e., the dot is empty with probability $P_0 \equiv P^0_0$, singly occupied with $P_1 \equiv P^{\uparrow}_{\uparrow} + P^{\downarrow}_{\downarrow}$, and doubly occupied with $P_d \equiv P^d_d$. A finite spin can only emerge for single occupancy. The average spin $\hbar\boldsymbol{S}$ with $\boldsymbol{S} = (S_x, S_y, S_z)$ is related to the matrix elements $P^{\chi}_{\chi'}$ via $S_x = \text{Re}\, P^{\downarrow}_{\uparrow}$, $S_y = \text{Im}\, P^{\downarrow}_{\uparrow}$, and $S_z = (1/2)(P^{\uparrow}_{\uparrow} - P^{\downarrow}_{\downarrow})$. To obtain the density-matrix elements by using the real-time transport theory developed in [45], we solve a kinetic equation formulated in Liouville space. The details are found in [30,31].

It is remarkable that on the r.h.s of (7), derivatives of Green's function with respect to bias voltage V appear. As a consequence, the linear conductance is not only determined by equilibrium properties of the quantum dot, but linear corrections in V are involved as well. This is consistent with the observation that, in equilibrium, the density matrix is diagonal with the matrix elements determined by the Boltzmann factors, i.e., the average spin on the quantum dot vanishes at $V = 0$ [46]. With applied bias voltage, though, a finite spin can accumulate. Therefore, to be sensitive to the relative magnetization direction of the leads, the linear conductance has to be connected to the differential spin accumulation $(d\boldsymbol{S}/dV)\big|_{V=0}$.

The results we find can be summarized as follows. At finite bias voltage, spin is accumulated on the dot. Here, we only need its contribution linear in V and find

$$\left.\frac{\partial|\boldsymbol{S}|}{\partial(eV)}\right|_{V=0} = \frac{pP_1}{4k_BT}\cos\alpha(\phi)\sin\frac{\phi}{2}\,, \tag{8}$$

where P_1 is the equilibrium probability for a singly occupied dot. The spin is lying in the y-z-plane enclosing an angle α with the y-axis, where

$$\tan\alpha(\phi) = -\frac{B}{\Gamma[1-f(\epsilon)+f(\epsilon+U)]}\cos\frac{\phi}{2}\,. \tag{9}$$

In the absence of an exchange field, the accumulated spin is oriented along $\hat{\boldsymbol{n}}_L - \hat{\boldsymbol{n}}_R$, i.e., it has a y-component only, $\alpha = 0$. The exchange field B, though, leads to a precession of the spin about the x-axis. The factor $1/\Gamma[1-f(\epsilon)+f(\epsilon+U)]$ in (9) can be identified as the life time of the dot spin, limited by tunneling out of the dot electron or by tunneling in of a second electron with opposite spin. Since both this life time and the exchange field are of first order in Γ, the angle α acquires a finite value.

The differential spin accumulation $dS/d(eV)$ in units of k_BT is illustrated in the middle panel of Fig. 4. It is clear that single occupation of the dot is required for spin accumulation, i.e., the plotted signal is high in the valley between the two conductance peaks. The lower panel of Fig. 4 shows the evolution of the rotation angle α as a function of the level energy ϵ. This angle is large in the valley between the conductance peaks, getting close to $\pm\pi/2$. A special point is $\epsilon = -U/2$, at which, due to particle-hole symmetry, the exchange interaction vanishes. As a consequence, α shows a sharp transition from positive to negative values, accompanied with a peak in the accumulated spin.

The linear conductance is given by

$$G^{\text{lin}} = G^{\text{lin,max}}\left(1 - p^2\cos^2\alpha(\phi)\sin^2\frac{\phi}{2}\right)\,. \tag{10}$$

The conductance is maximal for parallel magnetization, $\phi = 0$. Its value is $G^{\text{lin,max}} = (\pi e^2/h)(\Gamma/k_BT)[1-f(\epsilon+U)]f(\epsilon)[1-f(\epsilon)+f(\epsilon+U)]/[f(\epsilon)+1-f(\epsilon+U)]$. The upper panel of Fig. 4 depicts the linear conductance for five different values of the angle ϕ. For parallel magnetization, $\phi = 0$, there are two conductance peaks located near $\epsilon = 0$ and $\epsilon = -U$, respectively. With increasing angle ϕ, transport is more and more suppressed due to the spin-valve effect. However, this suppression is not uniform, as would be in the absence of the exchange field. In contrast, the spin-valve effect is less pronounced in the valley between the two peaks, where the rotation angle α is large. A large angle α reduces both the magnitude of the accumulated spin, as discussed above, and the relative angle to the magnetization of the drain electrode. Both enhance transport as compared to the situation without the exchange field. As a consequence, the two conductance peaks move towards each other with increasing ϕ.

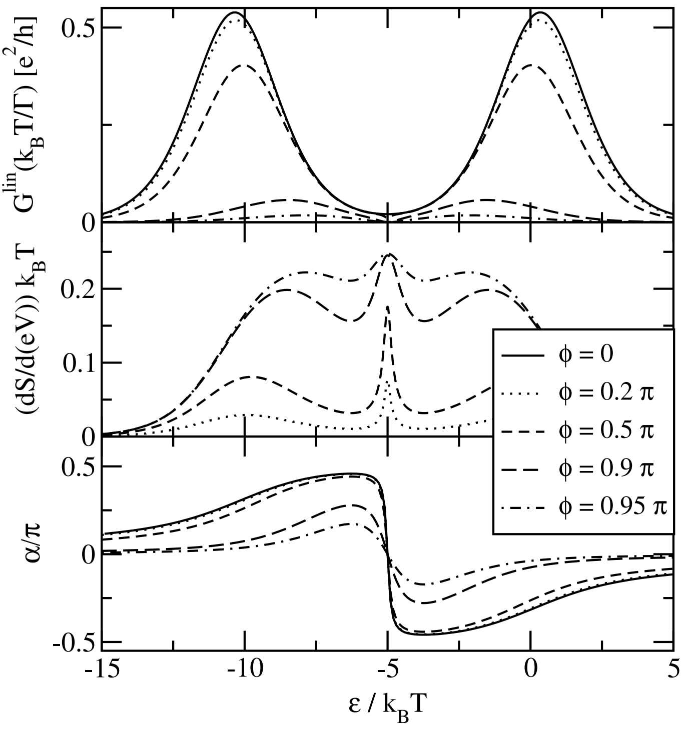


Fig. 4. Upper panel: Linear conductance (normalized by Γ/k_BT and plotted in units of e^2/h) as a function of level position ϵ for five different angles ϕ. Middle panel: Derivative of accumulated spin S with respect to bias voltage V normalized by k_BT. Lower panel: angle α between the quantum-dot spin and the y-axis. In all panels we have chosen the charging energy $U/k_BT = 10$ and $p = 1$

Another way to illustrate the influence of the exchange field is to plot the ϕ-dependence of the linear conductance, see Fig. 5. For values of the level position ϵ at which the rotation angle α is small, $\epsilon/k_BT = 3$ and 1, the ϕ-dependence of the conductance is almost harmonic, as it is for single magnetic tunnel junction. For $\epsilon/k_BT = -1$ and -3, however, the spin-valve effect is strongly reduced, and conductance is enhanced, except in the regime close to antiparallel magnetization, $\phi = \pi$. The conductance, then, stays almost flat over a broad range, and then establishes the spin-valve effect only in a small region around $\phi = \pi$.

4.2 First-Order Transport in Nonlinear Response

A rather complete analysis of first-order transport through quantum-dot spin valves, which covers both the linear- and nonlinear-response regime is presented in [31]. There, we derive generalized rate equations for the dot's occupation and accumulated spin, which provide the basis of quite an intuitive understanding of the behavior of the quantum-dot state.

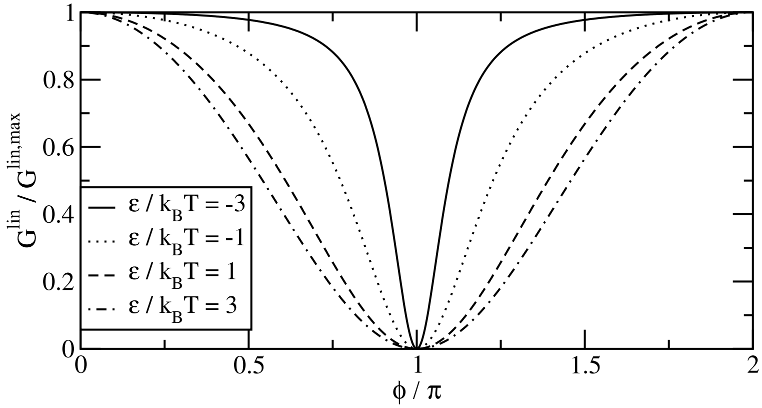


Fig. 5. Normalized linear conductance as a function of ϕ for $U/k_BT = 10$, $p = 1$, and four different values of the level position

In the non-linear-response regime, the physics of spin accumulation is more involved as for linear response. The accumulated spin tends to align anti-parallel to the drain electrode, leading to a spin blockade, i.e., a stronger spin-valve effect. This contrasts with the exchange field which, by rotation of the accumulated spin, tends to weaken the spin-valve effect. By the interplay of these two countersteering mechanisms, a very pronounced negative differential conductance is predicted.

4.3 Second-Order Transport

While first-order transport does not probe the spin splitting generated by the exchange field, second-order transport does. Therefore, in second-order transport, the exchange splitting plays a role even for collinear configuration of the leads' magnetizations. For parallel alignment, the exchange field gives rise to a gate-voltage dependent, finite spin polarization of the dot, $n_\uparrow \neq n_\downarrow$, even at zero bias. This polarization vanishes (at zero bias) for antiparallel orientation and symmetric coupling, since, in this case, the total exchange field adds up to zero. A detailed analysis of this transport regime will be presented in [32], which includes, among other things, the prediction and explanation of a peculiar zero-bias behavior for some circumstances.

4.4 The Kondo Effect

A very sensitive probe to the exchange field is provided by the Kondo effect, which occurs in singly-occupied quantum dots below a characteristic temperature, $k_\mathrm{B}T \leq k_\mathrm{B}T_\mathrm{K} \ll \Gamma$. The singly-occupied dot defines a local spin with two degenerate states, spin up and down. The local spin can be flipped by higher-order tunneling processes, in which the electron tunnels out of the dot, and another one with opposite spin enters from one of the leads. By these processes, the dot- and the lead-electron spins are coupled to each other. At low temperature, a highly-correlated state is formed, in which the local spin is totally screened. This Kondo-correlated state is accompanied with an increased transmission through the dot, and gives rise to a sharp zero-bias anomaly in the current-voltage characteristics.

How does a finite spin polarization in the leads modify this picture? As it turns out, there are two mechanisms influencing the Kondo effect. First, the exchange field lifts the spin degeneracy on the quantum dot. This is analogous to the situation of a Kondo dot in the presence of an external magnetic field. For the latter it is well known, that the zero-bias anomaly splits by twice the Zeeman energy. Due to the same reason, the exchange-field induces a splitting of the zero-bias anomaly for our model system, but now in the absence of an external magnetic field. In the presence of an external magnetic field both exchange- and magnetic-field induced splittings contribute. In particular, for a properly-tuned magnetic field the level splitting is compensated, and a single zero-bias anomaly is recovered.

The second mechanism by which the finite spin-polarization influences the Kondo effect is the screening of the quantum-dot spin. Naturally, both up- and down-spin electrons in the leads are crucial for the screening. An imbalance of majority and minority spins in the leads, therefore, weakens the screening capability. As we will see below, this leads to a reduced Kondo temperature $T_\mathrm{K}(p)$, which even vanishes for $p = 1$.

Recently, the possibility of the Kondo effect in a quantum dot attached to ferromagnetic electrodes was discussed in a number of publications [33–39], and it was shown, that the Kondo resonance is split and suppressed in the presence of ferromagnetic leads [37–39]. It was shown that this splitting can be compensated by an appropriately tuned external magnetic field to restore the Kondo effect [37,38], as we discuss in detail below.

In the following, we mainly concentrate on the case of parallel alignment of the leads' magnetization. For antiparallel alignment and symmetric coupling to the left and right lead, the exchange field vanishes (at zero bias voltage), and the usual Kondo resonance as for nonmagnetic electrodes forms.

Perturbative-Scaling Approach. An analytical access to the problem, which provides an intuitive picture of the involved physics, is the perturbative-scaling approach. For detail of the following calculations we refer to [37]. We make use of the poor man's scaling technique [47], performed in two stages [48]. In the first stage, when high-energy degrees of freedom are integrated out, charge fluctuations are the dominant. Afterwards, in the second stage, we map the resulting model to a Kondo Hamiltonian, and integrate out the degrees of freedom involving spin fluctuations. As we will see, each of the two stages will account for one of the two above mentioned different mechanisms by which the spin-polarized leads influence the Kondo effect, respectively.

The scaling procedure starts at an upper cutoff D_0, given by the onsite repulsion U. Charge fluctuations lead to a renormalization of the level position

ϵ_σ according to the scaling equations

$$\frac{d\epsilon_\sigma}{d\ln(D_0/D)} = |t|^2 \rho_{\bar{\sigma}} \,, \tag{11}$$

where $\bar{\sigma}$ is opposite to σ. Since the renormalization is spin dependent, a spin splitting is generated. In the presence of a magnetic field, this generated spin splitting simply adds to the initial Zeeman splitting $\varDelta\epsilon$. We obtain the solution $\varDelta\widetilde{\epsilon} = \widetilde{\epsilon}_\uparrow - \widetilde{\epsilon}_\downarrow = -(1/\pi)p\varGamma \ln(D_0/D) + \varDelta\epsilon$. The scaling of (11) is terminated [48] at $\widetilde{D} \sim -\widetilde{\epsilon}$. When plugging in $D_0 = U$ and $D = \epsilon$, we recover that the generated level splitting exactly reflects the zero-temperature limit of the exchange field (6).

To reach the strong-coupling limit, we tune the external magnetic field B_ext such that the total effective Zeeman splitting vanishes, $\varDelta\widetilde{\epsilon} = 0$. In the second stage of Haldane's procedure [48], spin fluctuations are integrated out. To accomplish this, we perform a Schrieffer-Wolff transformation [11] to map the Anderson model (with renormalized parameters $\widetilde{D}$ and $\widetilde{\epsilon}$) to a Kondo Hamiltonian, see (4). Since we are interested in low-energy excitations only, we neglect the energy dependence of the coupling constants and arrive at

$$\begin{aligned} H_\text{Kondo} = {} & J_+ S^+ \sum_{rr'kq} a^\dagger_{rk\downarrow} a_{r'q\uparrow} + J_- S^- \sum_{rr'kq} a^\dagger_{rq\uparrow} a_{r'k\downarrow} \\ & + S^z \left(J_{z\uparrow} \sum_{rr'qq'} a^\dagger_{rq\uparrow} a_{r'q'\uparrow} - J_{z\downarrow} \sum_{rr'kk'} a^\dagger_{rk\downarrow} a_{r'k'\downarrow} \right) , \end{aligned} \tag{12}$$

plus terms independent of either dot spin or lead electron operators, with $J_+ = J_- = J_{z\uparrow} = J_{z\downarrow} = |t|^2/|\widetilde{\epsilon}| \equiv J_0$ in the large-U limit. Although initially identical, the three coupling constants $J_+ = J_- \equiv J_\pm$, $J_{z\uparrow}$, and $J_{z\downarrow}$ are renormalized differently during the second stage of scaling. The scaling equations are

$$\frac{d(\rho_\pm J_\pm)}{d\ln(\widetilde{D}/D)} = \rho_\pm J_\pm (\rho_\uparrow J_{z\uparrow} + \rho_\downarrow J_{z\downarrow}) \tag{13}$$

$$\frac{d(\rho_\sigma J_{z\sigma})}{d\ln(\widetilde{D}/D)} = 2(\rho_\pm J_\pm)^2 \tag{14}$$

with $\rho_\pm = \sqrt{\rho_\uparrow \rho_\downarrow}$, $\rho_\sigma \equiv \sum_r \rho_{r\sigma}$. To solve these equations we observe that $(\rho_\pm J_\pm)^2 - (\rho_\uparrow J_{z\uparrow})(\rho_\downarrow J_{z\downarrow}) = 0$ and $\rho_\uparrow J_{z\uparrow} - \rho_\downarrow J_{z\downarrow} = J_0 p(\rho_\uparrow + \rho_\downarrow)$ is constant as well. I.e., there is only one independent scaling equation. All coupling constants reach the stable strong-coupling fixed point $J_\pm = J_{z\uparrow} = J_{z\downarrow} = \infty$ at the Kondo energy scale, $D \sim k_B T_K$. For the parallel configuration, the Kondo temperature in leading order,

$$T_\text{K}(P) \approx \widetilde{D} \exp\left\{ -\frac{1}{(\rho_\uparrow + \rho_\downarrow) J_0} \, \frac{\text{artanh}(p)}{p} \right\} , \tag{15}$$

depends on the polarization p in the leads. It is maximal for nonmagnetic leads, $p = 0$, and vanishes for $p \to 1$.

The unitary limit for the P configuration can be achieved by tuning the magnetic field appropriately, as discussed above. In this case, the maximum conductance through the quantum dot is $G^P_{\mathrm{max},\sigma} = e^2/h$ per spin, i.e., the same as for nonmagnetic leads.

Numerical Renormalization Group. Although perturbative scaling provides an instructive insight in the relevant physical mechanisms, it is a approximate method, and its reliability is, a priori, not clear. The numerical renormalization-group (NRG) technique [11], on the other hand, is one of the most accurate methods available to study strongly-correlated systems in the Kondo regime. Recently, it was adapted to the case of a quantum dot coupled to ferromagnetic leads [38,39].

The NRG study [38,39] confirms the predictions of the perturbative scaling analysis. The Kondo resonance is split, as a consequence of the exchange field. By appropriately tuning an external magnetic field, this splitting can be fully compensated and the Kondo effect can be restored [38]. Precisely at this field, the occupancy of the local level is the same for spin up and down, $\langle n_\uparrow \rangle = \langle n_\downarrow \rangle$, a fact that follows from the Friedel sum rule. Moreover, the Kondo effect has unusual properties such as a strong spin polarization of the Kondo resonance and for the density of states. Nevertheless, the quantum dot conductance is found to be the same for each spin channel, $G_\uparrow = G_\downarrow$. Furthermore, by analyzing the spin spectral function, the Kondo temperature can be determined, and the functional dependence on p as given by (15) has been confirmed.

More recently, the NRG scheme has been extended to account for structured densities of states [40]. The generated spin splitting found in this case is found to coincide with the exchange field defined in (6), when the energy-dependent density of states is included in the integral.

Nonequilibrium Transport Properties. To get a qualitative understanding of how the exchange field appears in nonlinear transport, we employ an equations-of-motions scheme with the usual decoupling scheme [49], but generalized by a self-consistent determination of the level energy to account for the exchange field in a correct way. We skip all technical details here (they are given in [37]), and go directly to the discussion of the results.

In Fig. 6 we show the differential conductance as a function of the transport voltage. For nonmagnetic leads, there is a pronounced zero-bias maximum (Fig. 6a), which splits in the presence of a magnetic field (Fig. 6b). For magnetic leads and parallel alignment, we find a splitting of the peak in the absence of a magnetic field (Fig. 6c), which can be tuned away by an appropriate external magnetic field (Fig. 6d). In the antiparallel configuration, the opposite happens, no splitting at $B_{\mathrm{ext}} = 0$ (Fig. 6e) but finite splitting at $B_{\mathrm{ext}} > 0$ (Fig. 6f) with an additional asymmetry in the peak amplitudes as a function of the bias voltage.

We conclude by mentioning that very recent experimental results [24,25] indicate confirmation of our theoretical predictions.

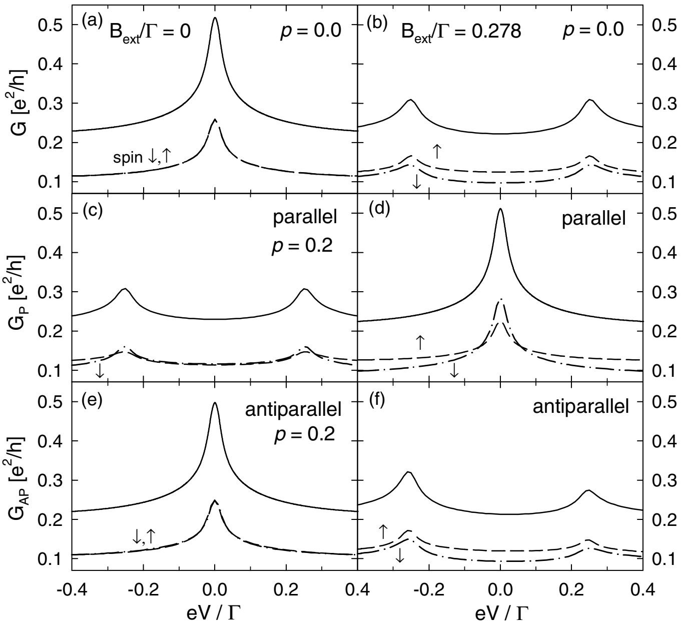


Fig. 6. Total differential conductance (solid lines) as well as the contributions from the spin up (dashed) and the spin down (dotted-dashed) channel vs. applied bias voltage V at zero magnetic field $B_{\mathrm{ext}} = 0$ (a,c,e) and at finite magnetic field (b,d,f) for normal (a,b) and ferromagnetic leads with parallel (c,d) and antiparallel (e,f) alignment of the lead magnetizations. The degree of spin polarization of the leads is $p = 0.2$ and the other parameters are: $k_{\mathrm{B}}T/\Gamma = 0.005$ and $\epsilon/\Gamma = -2$

5 Summary

The interplay of charge and spin degrees of freedom in quantum dots coupled to ferromagnetic leads is investigated theoretically. The simultaneous presence of both spin polarization in the leads and strong Coulomb interaction in the quantum dot generates an exchange field that acts on the quantum-dot electrons. We analyze its influence on the dot state and the conductance for different transport regimes. Two mechanisms, which can be important, are identified. The exchange field can precess an accumulated quantum-dot spin, and it generates a level splitting. In the limit of weak dot-lead coupling, the spin precession leads to a nontrivial dependence of the linear conductance on the angle between the leads' magnetization. For strong dot-lead coupling, the exchange field is detectable in a splitting of the Kondo resonance, which can be tuned away by additionally applying an external magnetic field.

Acknowledgments

The presented work is based on joint publications with L. Borda, M. Braun, R. Bulla, J. von Delft, H. Imamura, S. Maekawa, M. Sindel, Y. Utsumi, and I. Weymann, all of whom we thank for fruitful collaboration.

We thank G. Bauer, A. Brataas, P. Bruno, T. Costi, A. Fert, L. Glazman, W. Hofstetter, B. Jones, C. Marcus, J. Nygård, A. Pasupathy, D. Ralph, A. Rosch, S. Takahashi, D. Urban, and M. Vojta for discussions. This work was supported by the Deutsche Forschungsgemeinschaft under the Center for Functional Nanostructures and the Emmy-Noether program, and by the European Community under the 'Spintronics' RT Network of the EC RTN2-2001-00440, Project PBZ/KBN/044/P03/2001, and the Centre of Excellence for Magnetic and Molecular Materials for Future Electronics within the EC Contract G5MA-CT-2002-04049.

References

1. S.A. Wolf, D.D. Awschalom, R.A. Buhrman, J.M. Daughton, S. von Molnar, M.L. Roukes, A.Y. Chtchelkanova, D.M. Treger: Science **294**, 1488 (2001)
2. *Semiconductor Spintronics and Quantum Computation*, ed. by D.D. Awschalom, D. Loss, and N. Samarth (Springer, Berlin 2002)
3. S. Maekawa, T. Shinjo: *Spin Dependent Transport in Magnetic Nanostructures* (Taylor & Francis 2002)
4. D.V. Averin, K.K. Likharev: in *Mesoscopic Phenomenon in Solids*, ed. by B.L. Altshuler, P.A. Lee, R.A. Webb (Amsterdam: North-Holland 1991)
5. *Single Charge Tunneling: Coulomb Blockade Phenomena in Nanostructures*, NATO ASI Series B: Physics 294, ed. by H. Grabert, M.H. Devoret (Plenum Press, New York 1992)
6. *Mesoscopic Electron Transport*, ed. by L.L. Sohn, L.P. Kouwenhoven, G. Schön (Kluwer, Dordrecht 1997)
7. M. Jullière: Phys. Lett. A **54**, 225 (1975)
8. J.C. Slonczewski: Phys. Rev. B **39**, 6995 (1989)
9. J.S. Moodera, L.R. Kinder: J. Appl. Phys. **79**, 4724 (1996); H. Jaffrès, D. Lacour, F. Nguyen Van Dau, J. Briatico, F. Petroff, A. Vaurès: Phys. Rev. B **64**, 064427 (2001)
10. S. Datta, B. Das: Appl. Phys. Lett. **56**, 665 (1990)
11. A.C. Hewson: *The Kondo Problem to Heavy Fermions* (Cambridge Univ. Press 1993)
12. L.F. Schelp, A. Fert, F. Fettar, P. Holody, S.F. Lee, J.L. Maurice, F. Petroff, A. Vaurès: Phys. Rev. B **56**, 5747 (1997)
13. H. Brückl, G. Reiss, H. Vinzelberg, M. Bertram, I. Mönch, J. Schumann: Phys. Rev. B **58**, 8893 (1998)
14. K. Ono, H. Shimada, S. Kobayashi, Y. Ootuka: J. Phys. Soc. Jpn. **65**, 3449 (1996); K. Ono, H. Shimada, Y. Ootuka, J. Phys. Soc. Jpn. **66**, 1261 (1997)
15. S. Mitani, S. Takahashi, K. Takanashi, K. Yakushiji, S. Maekawa, H. Fujimori: Phys. Rev. Lett. **81**, 2799 (1998); H. Imamura, J. Chiba, S. Mitani, K. Takanashi, S. Takahashi, S. Maekawa, H. Fujimori: Phys. Rev. B **61**, 46 (2000); K. Yakushiji, S. Mitani, K. Takanashi, S. Takahashi, S. Maekawa, H. Imamura, H. Fujimori: Appl. Phys. Lett. **78**, 515 (2001)

16. K. Yakushiji, S. Mitani, K. Takanashi, H. Fujimori: J. Appl. Phys. **91**, 7038 (2002)
17. C.D. Chen, Y.D. Yao, S.F. Lee, J.H. Shyu: J. Appl. Phys. **91**, 7469 (2002)
18. Y. Chye, M.E. White, E. Johnston-Halperin, B.D. Gerardot, D.D. Awschalom, P.M. Petroff: Phys. Rev. B **66**, 201301(R) (2002)
19. M.M. Deshmukh, D.C. Ralph: Phys. Rev. Lett. **89**, 266803 (2002)
20. R. Jansen, J.S. Moodera: Appl. Phys. Lett. **75**, 400 (1999); S. Tanoue, A. Yamasaki: J. Appl. Phys. **88**, 4764 (2000)
21. J. Nygård, D.H. Cobden, P.E. Lindelof: Nature **408**, 342 (2000)
22. M.R. Buitelaar, T. Nussbaumer, C. Schönenberger: Phys. Rev. Lett. **89**, 256801 (2002)
23. A. Jensen, J. Nygård, J. Borggreen: in *Proceedings of the International Symposium on Mesoscopic Superconductivity and Spintronics*, ed. by H. Takayanagi, J. Nitta (World Scientific 2003) pp. 33-37; B. Zhao, I. Mönch, H. Vinzelberg, T. Mühl, C.M. Schneider: Appl. Phys. Lett. **80**, 3144 (2002); J. Appl. Phys. **91**, 7026 (2002); K. Tsukagoshi, B.W. Alphenaar, H. Ago: Nature **401**, 572 (1999)
24. J. Nygård, C.C. Markus, private communication.
25. D.C. Ralph, A. Pasupathy, private communnication.
26. Y. Manassen, R.J. Hamers, J.E. Demuth, A.J. Castellano: Phys. Rev. Lett. **62**, 2531 (1989); C. Durkan, M.E. Welland: Appl. Phys. Lett. **80**, 458 (2002)
27. J. Barnaś, A. Fert: Phys. Rev. Lett. **80**, 1058 (1998); S. Takahashi, S. Maekawa: Phys. Rev. Lett. **80**, 1758 (1998); A. Brataas, Yu.V. Nazarov, J. Inoue, G.E.W. Bauer: Phys. Rev. B **59**, 93 (1999); M. Pirmann, J. von Delft, G. Schön: J. Mag. Mag. Mat. **219**, 104 (2000); J. Barnaś, J. Martinek, G. Michalek, B.R. Bulka, A. Fert: Phys. Rev. B **62**, 12363 (2000); J. Martinek, J. Barnaś, S. Maekawa, H. Schoeller, G. Schön: Phys. Rev. B **66**, 014402 (2002)
28. W. Rudziński, J. Barnaś: Phys. Rev. B **64**, 085318 (2001); G. Usaj, H.U. Baranger: Phys. Rev. B **63**, 184418 (2001); A. Cottet, W. Belzig, C. Bruder: cond-mat/0308564
29. J. Martinek, J. Barnaś, A. Fert, S. Maekawa, G. Schön: J. Appl. Phys. **93**, 8265 (2003)
30. J. König, J. Martinek: Phys. Rev. Lett. **90**, 166602 (2003)
31. M. Braun, J. König, J. Martinek: submitted to Phys. Rev. B.
32. I. Weymann, J. Martinek, J. König, J. Barnaś, G. Schön: to be published.
33. N. Sergueev, Q.F. Sun, H. Guo, B.G. Wang, J. Wang: Phys. Rev. B **65**, 165303 (2002)
34. P. Zhang, Q.K. Xue, Y. Wang, X.C. Xie: Phys. Rev. Lett. **89**, 286803 (2002)
35. B.R. Bułka, S. Lipinski: Phys. Rev. B **67**, 024404 (2003)
36. R. Lopez, D. Sanchez: Phys. Rev. Lett. **90**, 116602 (2003)
37. J. Martinek, Y. Utsumi, H. Imamura, J. Barnaś, S. Maekawa, J. König, G. Schön: Phys. Rev. Lett. **91**, 127203 (2003)
38. J. Martinek, M. Sindel, L. Borda, J. Barnaś, J. König, G. Schön, J. von Delft: Phys. Rev. Lett. **91**, 247202 (2003)
39. M.S. Choi, D. Sanchez, R. Lopez: Phys. Rev. Lett. **92**, 056601 (2004)
40. J. Martinek, M. Sindel, L. Borda, J. Barnaś, R. Bulla, J. König, G. Schön, S. Maekawa, J. von Delft, to be published
41. K. Yosida: *Theory of Magnetism* (Springer, Berlin 1996).
42. *Handbook of the Band Structure of Elemental Solids*, ed. by D.A. Papaconstantopoulos (Plenum Press 1986)
43. J. König, Y. Gefen: Phys. Rev. Lett. **86**, 3855 (2001); Phys. Rev. B **65**, 045316 (2002); B. Kubala, J. König, Phys. Rev. B **65**, 245301 (2002)

44. We note that, for all the effects discussed in this paper, the (spin-dependent) density of states ρ_σ and the (spin-independent) tunneling amplitudes t enter via the combination $\rho_\sigma |t|^2$. As a consequence, all the conclusions drawn in this paper are valid also for a system with spin-dependent tunneling amplitudes but nonmagnetic leads, or a mixture of both.
45. J. König, H. Schoeller, G. Schön: Phys. Rev. Lett. **76**, 1715 (1996); J. König, J. Schmid, H. Schoeller, G. Schön: Phys. Rev. B **54**, 16820 (1996); H. Schoeller, in Ref. [6]; J. König: *Quantum Fluctuations in the Single-Electron Transistor* (Shaker, Aachen 1999)
46. In equilibrium and zeroth order in Γ, the average spin on the dot vanishes: majority-spin electrons enter the quantum dot at a higher rate than minority spins, but this is exactly compensated by a higher rate for the majority spins to leave the dot.
47. P.W. Anderson: J. Phys. C **3**, 2439 (1970)
48. F.D.M. Haldane: Phys. Rev. Lett. **40**, 416 (1978)
49. Y. Meir, N.S. Wingreen, P.A. Lee: Phys. Rev. Lett. **70**, 2601 (1993); N.S. Wingreen, Y. Meir: Phys. Rev. B **49**, 11040 (1994)

Transport of Interacting Electrons Through a Quantum Dot in Nanowires

Igor V. Gornyi[1,2] and Dmitri G. Polyakov[1,2]

[1] Forschungszentrum Karlsruhe, Institut für Nanotechnologie (INT), Postfach 3640, 76021 Karlsruhe, Germany
[2] A.F. Ioffe Physico-Technical Institute, 194021 St.Petersburg, Russia

Abstract. We generalize the fermionic renormalization group method to analytically describe transport through a double barrier structure in a one-dimensional system. Focusing on the case of weakly interacting electrons, we investigate thoroughly the dependence of the conductance on the strength and the shape of the double barrier for arbitrary temperature T, down to zero T. We systematically analyze the contributions to renormalized scattering amplitudes from characteristic scales absent in the case of a single impurity, without restricting the consideration to the model of a single resonant level. Both a sequential resonant tunneling for high T and a resonant transmission for T smaller than the resonance width are studied within the unified treatment of transport through strong barriers. For weak barriers, we show that two different regimes are possible. Moderately weak impurities get strong due to the renormalization, so that transport is described in terms of theory for initially strong barriers. The renormalization of very weak impurities does not yield any peak in the transmission probability; however, remarkably, the interaction gives rise to a sharp peak in the conductance.

1 Introduction

Effects related to the Coulomb interaction between electrons become increasingly prominent in systems of lower spatial dimensionality as their size is made smaller. Recent experimental progress in controlled preparation of nanoscale devices has led to a revival of interest in the transport properties of one-dimensional (1D) quantum wires. Owing to the particular geometry of the Fermi surface, systems of dimensionality one are unique in that the Coulomb correlations in 1D change a noninteracting picture completely and thus play a pivotal role in low-temperature physics. A remarkable example of a correlated 1D electron phase is the Luttinger-liquid model [1,2]. In this model, arbitrarily weak interactions ruin the conventional Fermi liquid phenomenology by essentially modifying low-energy excitations across the Fermi surface. As a result, the tunneling density of states develops power-law singularities on the Fermi surface. Interactions between electrons moving in opposite directions lead to striking transport properties of a Luttinger liquid in the presence of impurities. In particular, even a single impurity yields a complete pinning of a Luttinger liquid with repulsive interactions [3,4].

Evidence has recently emerged pointing towards the existence of the Luttinger liquid in metallic single-wall carbon nanotubes [5,6]. The Luttinger liquid

I.V. Gornyi and D.G. Polyakov, Transport of Interacting Electrons Through a Quantum Dot in Nanowires, Lect. Notes Phys. **658**, 165–184 (2005)
http://www.springerlink.com/

behavior was observed via the power-law temperature and bias-voltage dependence of the current through tunneling contacts attached to the nanotubes. Further technological advances have made possible the fabrication of low-resistance contacts between nanotubes and metallic leads (see, e.g., [7–10] and references therein). These recent developments have paved the way for systematic transport measurements in Luttinger liquids with impurities.

Here, we study transport through a double barrier in a 1D liquid. In the 1D geometry, two impurities in effect create a quantum dot inside the system. Resonant tunneling through the two impurities is a particularly attractive setup to investigate the correlated transport in an inhomogeneous Luttinger liquid: due to the resonant behavior of the current, the interplay of Luttinger-liquid correlations and impurity-induced backscattering is more easily accessible to transport measurements. Two striking experimental observations have been reported recently. In [11], a resonant structure of the conductance of a semiconductor single-mode quantum wire was attributed to the formation (with reduction of electron density by changing gate voltage) of a single disorder-induced quantum dot. In [12], two barriers were created inside a carbon nanotube in a controlled way with an atomic force microscope. In both cases, the amplitude of a conductance peak G_p as a function of temperature T showed power-law behavior $G_p \propto T^{-\gamma}$ with the exponent γ noticeably different from $\gamma = 1$. The latter is the value of γ expected in the absence of interactions provided T lies in the range $\varGamma \ll T \ll \varDelta$, where $\varGamma$ is the width of a resonance in the transmission coefficient and $\varDelta$ is the single-particle level spacing. The width of a conductance peak w followed a linear T dependence $w \propto T$ in both experiments.

On the theoretical side, resonant tunneling in a Luttinger liquid was studied in a number of papers [3,13–18]. In particular, the width $\varGamma \propto T^{\alpha_e}$ was shown [13,16] to shrink with decreasing temperature. The exponent α_e depends on the strength of interaction and describes tunneling into the end of a semi-infinite liquid. The dimensionless peak conductance (in units of e^2/h) obeys $G_p \sim \varGamma/T$ in the above range of T, which indeed leads to a smaller value of $\gamma = 1 - \alpha_e$. The reduced exponent γ reported in [11] was positive (and different for different conductance peaks, in the range $\gamma \sim 0.5 - 0.8$), whereas in [12] the reported value of $\gamma \simeq -0.7$ was negative. More specifically, in [12], the conductance as a function of the gate voltage showed certain traces of periodicity characteristic to the Coulomb blockade regime. Surprisingly, both the amplitude G_p and the width w were reported to vanish with decreasing T, in sharp contrast to the noninteracting case. While such behavior is known to be possible for very strong repulsive interaction [13,16], the required strength of interaction would then be much larger than expected and indeed reported (see [5,6,19] and references therein) in carbon nanotubes. Roughly a doubling (or even a larger factor) of the expected [19,5,6] exponent α_e, which is $\alpha_e \sim 0.6 - 1.0$, would be necessary to fit the experimental data.

It is thus desirable to examine the resonant tunneling in a Luttinger liquid in a broad range of temperature down to $T = 0$ and for various parameters of the barriers. There are a variety of techniques to construct the low-energy transport theory [1,2,20]. The method we develop here is valid for weak interaction and is based on the renormalization group (RG) approach of [21], which was applied earlier in a variety of contexts [22–25]. One of the appeals of this kind of theory is that it allows one to treat weak and strong scatterers on an equal footing, which is technically significantly less straightforward in the bosonization method [2]. The RG approach enables us to investigate in detail the resonant transport of weakly interacting spinless electrons. Within the fermionic RG approach, we confirm earlier results [3,13,16] obtained within bosonic field theories. We examine the conductance through a double barrier for arbitrary strength and an arbitrary shape of the barrier, not restricting ourselves to the model of a single resonant level. In particular, we demonstrate the existence of narrow conductance peaks for two weak impurities, which is in sharp contrast to the noninteracting case. We do not find any trace of the correlated tunneling mechanism proposed in [12,26]. Part of this work was presented in [27].

2 Fermionic Renormalization Group

2.1 Single Impurity: Basic Results

We begin with a brief description of transport through a single structureless impurity in the spirit of the RG approach [21]. Without interaction, the impurity is characterized by a transmission coefficient t_0 and reflection coefficients r_{L0} and r_{R0}, from the left and from the right respectively (we put the impurity at the center of coordinates, $x = 0$). Suppose that the energy dependence of the bare scattering matrix can be neglected far from the boundaries of an energy band $(-D_0, D_0)$ around the Fermi level. The energy scale D_0 serves as the ultraviolet cutoff of RG transformations and, physically, is of the order of v_F/d (throughout the paper we put $\hbar = 1$) or the Fermi energy ϵ_F, whichever is smaller. Here d is the radius of interaction and v_F is the Fermi velocity. Deep inside the band $(-D_0, D_0)$, we linearize the energy spectrum around the Fermi level. The differential RG equations [21] read

$$\partial t/\partial\mathcal{L} = -\alpha t\,\mathrm{R}\ , \quad \partial r_{L,R}/\partial\mathcal{L} = \alpha r_{L,R}\mathrm{T}\ , \tag{1}$$

where $\mathcal{L} = \ln(D_0/|\epsilon|)$, the energy ϵ is measured from the Fermi level and the transmission probability $\mathrm{T} = 1 - \mathrm{R} = |t|^2$. The boundary conditions at $\mathcal{L} = 0$ set the scattering amplitudes at their noninteracting values t_0, $r_{L0,R0}$. Throughout the paper we consider spinless electrons, for which the interaction constant is

$$\alpha = (V_f - V_b)/2\pi v_F\ , \tag{2}$$

where V_f and V_b are the Fourier transforms of a pairwise interaction potential yielding forward (V_f) and backward (V_b) scattering. The forward scattering does not lead to transitions between two branches of right- and left-movers, whereas the backscattering does. We assume that $\alpha > 0$.

Note that, for spinless electrons, the interaction-induced backward scattering and forward scattering relate to each other as direct and exchange processes, so that the backscattering only appears in the combination $V_f - V_b$ and thus merely redefines parameters of the Luttinger model (formulated [1] in terms of forward-scattering amplitudes only). In particular, the backscattering does not lead to any RG flow for α. For spinful electrons this is valid only to one-loop order [1]. It is also worth mentioning that for a point interaction $V_f = V_b$, so that $\alpha = 0$, hence for spinless electrons one has to start with a finite-range interaction. However, the RG flow for the scattering matrix (1) occurs for $|\epsilon| < v_F/d$ and is governed solely by the constant α. It follows that on low-energy scales one can effectively consider the interaction as local, $V_{\mathrm{eff}}(x - x') = 2\pi\alpha v_F \delta(x - x')$, and formally deal exclusively with forward scattering. A non-zero range of interaction for $k_F d \gg 1$ manifests itself only in the boundary conditions to (1) at $|\epsilon| \sim D_0 = v_F/d$ and therefore does not affect the singular behavior of the renormalized scattering matrix at $\epsilon \to 0$. We assume that the Coulomb interaction between electrons is screened by external charges (e.g., by metallic gates, in which case d is given by the distance to the gates) and that a resulting $\alpha \ll 1$.

Integration of (1) gives [21]

$$\frac{\mathrm{R}}{\mathrm{T}} = \frac{\mathrm{R}_0}{\mathrm{T}_0}\left(\frac{D_0}{|\epsilon|}\right)^{2\alpha} . \tag{3}$$

The phases of the scattering amplitudes are not affected by the renormalization. Equations (1) are equivalent to a one-loop renormalization, so that (3) is valid to first order in interaction $\sim O(\alpha)$ in the exponent of the power-law scaling. As follows from (3), whatever the initial values of T_0, at $\alpha > 0$ they all flow to the fixed point of (1) at zero transmission [3], $\mathrm{T} = 0$ at $\epsilon = 0$. In the limits of a weak impurity (both $\mathrm{R}_0 \ll 1$ and $\mathrm{R} \ll 1$) and a strong tunneling barrier ($\mathrm{T}_0 \ll 1$), (3) coincides with the RG results obtained by bosonization [3], provided $\alpha \ll 1$. Equation (3) gives the transmission probability for electrons with energy ϵ at $T = 0$. For finite T, the renormalization stops at $|\epsilon| \sim T$.

Beyond the microscopic scale D_0, it is instructive to introduce two more energetic scales that characterize the renormalization of the transmission coefficient by weak interaction, D_p and D_r:

$$\ln(D_p/D_0) = -1/\alpha \ , \quad D_r/D_0 = \mathrm{R}_0^{1/2\alpha} \ . \tag{4}$$

The energy D_p defines the scale on which a perturbation theory in interaction breaks down. If $|\epsilon| < D_p$, the interaction requires a non-perturbative treatment. The scale D_p does not depend on R_0 and is much smaller than D_0 for $\alpha \ll 1$. The energy D_r defines the scale on which a perturbation theory in the impurity strength breaks down. If $|\epsilon| < D_r$, a weak impurity with $\mathrm{R}_0 \ll 1$ yields strong reflection, $\mathrm{R} \sim 1$. Provided $\mathrm{R}_0 \ll 1$, the scales D_p and D_r are parametrically different and, for any α, $D_r \ll D_p$. We will see in Sect. 2.3 that the scale D_r is of central importance in RG theory for a double-barrier structure.

2.2 Renormalization Group for a Double Barrier

Consider two potential barriers located at $x = 0$ and $x = x_0$ and let the distance x_0 be much larger than the width of each of them. The spatial structure itself yields an energy dependence of the total (describing scattering on both impurities) transmission and reflection amplitudes, $t(\epsilon)$ and $r_{L,R}(\epsilon)$. Specifically, without interaction the energy $\Delta = \pi v_F / x_0$ gives a period of oscillations in the total scattering amplitudes with changing ϵ_F. An RG description of a double barrier requires a generalization of the RG [21] to the case when the bare amplitudes are energy dependent. A question, however, arises if the total amplitudes generated by RG transformations are expressed in terms of themselves only. The answer depends on the parameter Δ / D_0, where $D_0 = \min\{\epsilon_F\,,\, v_F/d\}$. If $\Delta \ll D_0$, the RG transformations generate more terms than are encoded in the total S-matrix, namely the amplitudes to stay inside the dot $A_{\mu,-\mu}(\epsilon)$ and those to escape from the dot to the left(right) $d_\mu^\pm(\epsilon)$ for right(left)-movers ($\mu = \pm$).

We now derive non-perturbative amplitudes for a double barrier using an appropriate RG scheme. To account for the ϵ dependence of the bare amplitudes, the derivation of the RG from the perturbative results necessitates introduction of two energies, ϵ and D. The latter is a flow parameter in RG transformations, i.e., an ultraviolet cutoff rescaled after tracing over states with energies ϵ' in the interval $|\epsilon'| \in (D, D_0)$. The renormalization stops at $D = \max\{|\epsilon|, T\}$. The system of one-loop RG equations for a double barrier reads [27]

$$\begin{aligned}\partial t(\epsilon, D)/\partial \mathcal{L}_D \;&= \hat{I}_{\epsilon'}(\epsilon, D)\Big\{ L_+(\epsilon, \epsilon'; D) + \theta(-\epsilon') t(\epsilon, D) \\ &\quad \times \left[\, r_R(\epsilon, D) r_R^*(\epsilon', D)\, \chi_{\epsilon-\epsilon'} + r_L(\epsilon, D) r_L^*(\epsilon', D) \,\right]\Big\}\;, \qquad (5)\\ \partial r_L(\epsilon, D)/\partial \mathcal{L}_D &= \hat{I}_{\epsilon'}(\epsilon, D)\; \Big\{ L_-(\epsilon, \epsilon'; D) + \theta(\epsilon') r_L(\epsilon', D) \\ &\quad + \theta(-\epsilon')[\, t^2(\epsilon, D) r_R^*(\epsilon', D)\, \chi_{\epsilon-\epsilon'} + r_L^2(\epsilon, D) r_L^*(\epsilon', D) \,]\Big\}\;, \quad (6)\end{aligned}$$

and similar equations for other amplitudes. Here $\mathcal{L}_D = \ln(D_0/D)$ (we introduced D dependent amplitudes),

$$\begin{aligned} L_\mu(\epsilon, \epsilon') &= d_-^-(\epsilon) A_{+-}(\epsilon') d_-^\mu(\epsilon)(\chi_{\epsilon-\epsilon'} - 1) + d_+^-(\epsilon) A_{-+}(\epsilon') d_+^\mu(\epsilon)(1 - \chi_{\epsilon'-\epsilon})\;, (7)\\ \hat{I}_{\epsilon'}(\epsilon, D)\{...\} &= -\frac{\alpha}{2\ln \Lambda}\left[\int_D^{\Lambda D} + \int_{-\Lambda D}^{-D}\right] \frac{d\epsilon'}{\epsilon - \epsilon'}\,\{\ldots\}\;, \qquad (8)\end{aligned}$$

$\Lambda \gg 1$ is restricted by the condition $\alpha \ln \Lambda \ll 1$, and $\chi_\epsilon = \exp(2\pi i \epsilon / \Delta)$.

The essence of the RG procedure is a perturbative treatment of contributions to the renormalized amplitudes at energy ϵ from all states with energies ϵ' in the interval $|\epsilon'| \in (D, \Lambda D)$, starting from $D = D_0/\Lambda$. The RG equations thus differ from the Hartree-Fock equations in that all the amplitudes depend on D and, moreover, the Hartree-Fock type integration over projected states with energies ϵ' only goes over the interval $|\epsilon'| \in (D, \Lambda D)$ instead of $(0, D_0)$. In effect, each step of the RG transformations accounts for the scattering off the Friedel oscillations in a *finite* spatial region, $|x|, |x - x_0| \in (v_F/\Lambda D, v_F/D)$. Moreover, the Friedel oscillations are only partly modified, through the (already performed) renormalization of the reflection amplitudes at energies larger than D. At the same time, the scattering matrix at energies smaller than D is taken at its bare value. This should be contrasted with the Hartree-Fock approach, where the scattering amplitudes are determined by interaction processes on all energy scales on every step of the Hartree-Fock iterations.

At finite temperature T, one should substitute the Fermi distribution function $n_F(\epsilon)$ for the step functions in (5) and (6) according to $\theta(\pm\epsilon) \to n_F(\mp\epsilon)$. The factor $(\epsilon - \epsilon')^{-1}$ in (8) effectively stops the renormalization at $D \sim |\epsilon|$, while the factors $n_F(\pm\epsilon')$ do so at $D \sim T$, otherwise the renormalization can be carried out down to $D = 0$. The infrared cutoff at $D \sim T$ establishes a characteristic spatial scale of $L_T = v_F/T$. Due to the thermal smearing, the Friedel oscillations decay exponentially on a scale of L_T. The RG equations (5),(6) should be solved with proper boundary conditions at $D = D_0$: $t_0(\epsilon) = t_1 t_2 S^{-1}(\epsilon)$, $r_{L0}(\epsilon) = r_1 + r_2 t_1^2\, \chi_\epsilon S^{-1}(\epsilon)$, and similarly for other amplitudes. Here $S(\epsilon) = 1 - r_2 r_1' \chi_\epsilon$, and $r_{R0}(\epsilon) = -r_{L0}^*(\epsilon) t_0(\epsilon)/t_0^*(\epsilon)$ by unitarity.

We are now in a position to solve the system of RG equations (5),(6) by integrating out all states with energies $|\epsilon'| > \max\{|\epsilon|, T\}$. We begin with the case $D_0 \gg \varDelta$, which is a typical case unless interaction is very long ranged. We proceed in two steps. Let us first integrate over $D \gg \varDelta$. This can be done for arbitrary ϵ. Specifically, if $|\epsilon| > \varDelta$, this will already solve the problem by providing us with fully renormalized amplitudes. In the more interesting case of $|\epsilon| < \varDelta$, we will only sum up contributions to the renormalized amplitudes from states with $|\epsilon'| > \varDelta$ and, as a second step, will have to proceed with renormalization for $D < \varDelta$.

2.3 Separate Renormalization of Two Impurities: $D \gg \varDelta$

Since the renormalization for $D \gg \varDelta$ involves many resonant levels, the amplitudes contain slowly varying parts and parts oscillating rapidly with changing ϵ' on a scale of $\varDelta$. Integration over ϵ' in (5),(6) allows us to separate the slow and fast variables: as a result, the dependence of the amplitudes on D will be slow on the scale of $\varDelta$. To construct the solution to the RG equations, note that an important parameter $D/D_{r_{\min}}$ is available, where $D_{r_{\min}} = \min\{D_{r_1}, D_{r_2}\}$ and $D_{r_{1,2}}$ are defined for each of two barriers by (4). If both barriers are initially (i.e., at $D = D_0$) strong ($|t_{1,2}| \ll 1$), then this parameter is small for all $D < D_0$. However, if one or both of the barriers are initially weak, there is a range of $D \in (D_{r_{\min}}, D_0)$ where at least one barrier still remains weak.

It is useful first to examine some general properties of integrals over ϵ' that appear in the course of renormalization. We see that the averaging over ϵ' involves two types of integrals

$$\mathcal{I}_1 = \int_{|\epsilon|}^{D_0} \frac{d\epsilon'}{\epsilon'} \frac{1}{S(\epsilon')}\ , \quad \mathcal{I}_2 = \int_{|\epsilon|}^{D_0} \frac{d\epsilon'}{\epsilon'} \frac{\chi_{\epsilon'}}{S(\epsilon')}\ , \tag{9}$$

where $\mathcal{I}_{1,2}$ are related by $\mathcal{I}_2 = (\mathcal{I}_1 - \mathcal{L})/r_1' r_2$. The integrals (9) are evaluated in different ways depending on whether at least one of the barriers is weak ($|r_1' r_2| \ll 1$) or both barriers are strong ($|r_1' r_2| \simeq 1$). In the former case the integrand of $\mathcal{I}_1$ is only slightly modulated, so that one can expand the factor $S^{-1}(\epsilon')$ and average over harmonics $\chi^n(\epsilon')$. Then only zero harmonics yield singular (logarithmic) corrections and in the leading-log approximation we have

$$\mathcal{I}_1 = \mathcal{L} \ , \quad \mathcal{I}_2 = 0 \ . \tag{10}$$

In the opposite case of strong barriers, sharp resonances appear that are described by a Breit-Wigner formula for $S^{-1}(\epsilon')$ and give $|\mathcal{I}_1| \simeq |\mathcal{I}_2|$:

$$\mathcal{I}_1 = \mathcal{L}/2 \ , \quad \mathcal{I}_2 = -\mathcal{L}/2r_1'r_2 \ . \tag{11}$$

The situation repeats itself in the RG equations (5),(6). The difference in the factor of 1/2 between the values of $\mathcal{I}_1$ in (10),(11) implies that the renormalization should be carried out differently in the regions $D \gg D_{r_{\min}}$ and $\varDelta \ll D \ll D_{r_{\min}}$.

For $D \gg D_{r_{\min}}$, similarly to (10), after the averaging over ϵ' only zero harmonics contribute to the renormalization. The solution at $D \gg D_{r_{\min}}$ has the form of Fabry-Perot equations with the reflection and transmission amplitudes of each of two barriers renormalized separately, according to the RG (1),(3) for a single impurity:

$$\begin{aligned}\partial t_{1,2}(D)/\partial \mathcal{L}_D &= -\alpha t_{1,2}(D)|r_{1,2}(D)|^2 \ , &(12)\\ \partial r_{1,2}(D)/\partial \mathcal{L}_D &= \ \alpha r_{1,2}(D)|t_{1,2}(D)|^2 \ . &(13)\end{aligned}$$

Accordingly, the amplitudes describing a Fabry-Perot resonance with the replacement

$$\begin{aligned} t_{1,2} \qquad &\to \frac{(D/D_0)^\alpha t_{1,2}}{[\,|r_{1,2}|^2 + (D/D_0)^{2\alpha}|t_{1,2}|^2\,]^{1/2}} \ , &(14)\\ r_{1,2}(r_{1,2}') &\to \frac{r_{1,2}(r_{1,2}')}{[\,|r_{1,2}|^2 + (D/D_0)^{2\alpha}|t_{1,2}|^2\,]^{1/2}} &(15)\end{aligned}$$

solve (5),(6) averaged over harmonics for $D \gg \max\{D_{r_{\min}}, \varDelta\}$.

On the other hand, if $D_{r_{\min}} \gg \varDelta$, there is an interval of $D \in (\varDelta, D_{r_{\min}})$ in which each of the impurities is strongly reflecting, so that the averaged equations are again simplified by summing over resonance poles, similarly to (11). We get an independent renormalization of $t_{1,2}(D)$ according to

$$t_{1,2}(D)/t_{1,2}(D_{r_{\min}}) = (D/D_{r_{\min}})^{\alpha/2} \ , \tag{16}$$

and the scaling exponent is now half that for $D \gg D_{r_{\min}}$. We thus have two solutions given by (14),(15) and (16), respectively, that match onto each other at $D \sim D_{r_{\min}}$. Due to the slow power-law dependence, for $\alpha \ll 1$ the matching is exact.

We conclude that the key difference between the renormalization for D larger and smaller than $D_{r_{\min}}$ is that for $D \gg D_{r_{\min}}$ the transmission amplitudes for each barrier are renormalized with the exponent α, whereas for $D \ll D_{r_{\min}}$ with the exponent $\alpha/2$. In both limiting cases, two barriers are renormalized separately for $D \gg \varDelta$. It is worth stressing that generally the independent renormalizations of two barriers cannot be derived from RG equations written in terms of only $t(\epsilon)$ and $r_{L,R}(\epsilon)$, i.e., the terms $L_\mu(\epsilon,\epsilon';D)$ are of crucial importance in the derivation of (12)–(15). However, the renormalization of resonant tunneling amplitudes for energies near resonances allows for another formulation which involves $t(\epsilon)$ and $r_{L,R}(\epsilon)$ only, we will return to this issue in Sect. 2.4.

If ϵ is close to one of resonant energies, (5),(6) can be further simplified by expanding χ_ϵ near the resonance: the renormalized amplitudes for strong barriers take then the form of Breit-Wigner amplitudes with D dependent widths $\varGamma_{1,2}(D) = (\varDelta/2\pi)|t_{1,2}(D)|^2 \propto D^\alpha$. Specifically, for initially strong barriers:

$$\varGamma_{1,2}(D) = \frac{\varDelta}{2\pi}\,|t_{1,2}|^2\left(\frac{D}{D_0}\right)^\alpha , \tag{17}$$

where $|t_{1,2}|^2 \ll 1$ are the bare transmission probabilities at $D = D_0$ and resonant peaks are sharp [i.e., $\varGamma_{1,2}(D) \ll \varDelta$] for all $D < D_0$. If at least one barrier is initially weak, the resonant structure develops only at $D \ll D_{r_{\mathrm{min}}}$. Provided one barrier is initially weak (assume this is the right barrier and $D_{r_2} \gg \varDelta$), whereas the other is strong, then

$$\varGamma_1(D) = \frac{\varDelta}{2\pi}\,|t_1|^2\left(\frac{DD_{r_2}}{D_0^2}\right)^\alpha , \quad \varGamma_2(D) = \frac{\varDelta}{2\pi}\left(\frac{D}{D_{r_2}}\right)^\alpha . \tag{18}$$

If $\varDelta \ll D_{r_2} < D_{r_1} \ll D_0$, i.e., both barriers are initially weak, then

$$\varGamma_1(D) = \frac{\varDelta}{2\pi}\left(\frac{DD_{r_2}}{D_{r_1}^2}\right)^\alpha \tag{19}$$

and $\varGamma_2(D)$ is given again by (18).

To summarize this section, we have found the fully renormalized scattering amplitudes for $|\epsilon| > \varDelta$ if $|\epsilon| \gg T$. Also, if $T \gg \varDelta$, substituting $D \to T$ solves the problem for arbitrary ϵ. However, when both $|\epsilon|, T \ll \varDelta$, we should proceed with the renormalization in the range $D \ll \varDelta$.

2.4 Single Resonance: $D \ll \varDelta$

Let us now consider (5),(6) for $|\epsilon|, |\epsilon'| \ll \varDelta$. In this limit, the terms (7) containing the amplitudes $A_{\mu,-\mu}(\epsilon')$ to stay inside the dot become irrelevant in the RG sense: the phase factors $\chi_{\epsilon-\epsilon'}$ in (7) can be expanded about $\epsilon, \epsilon' = 0$, which leads to the cancellation of the singular factor $(\epsilon-\epsilon')^{-1}$ in (8). As a result, the terms $L_\mu(\epsilon,\epsilon';D)$ do not contribute to the renormalization at $D \ll \varDelta$. The factors $\chi_{\epsilon-\epsilon'}$ should also be omitted in the terms of (5),(6) that are proportional to $r_R^*(\epsilon')$. Thus we are led to a coupled set of RG equations that describe also a single impurity with energy dependent scattering amplitudes: the spatial structure of the double barrier system is of no importance for the renormalization at $D \ll \varDelta$. However, the boundary conditions in the double-barrier case should be written at $D \sim \varDelta$, instead of $D \sim D_0$. We obtain for $|\epsilon|, |\epsilon'|, D \ll \varDelta$:

$$\partial t(\epsilon,D)/\partial\mathcal{L}_D \;= -(\alpha/2)\;t(\epsilon,D)\,[\,r_R(\epsilon,D)\;\overline{r_R^*}(D) + r_L(\epsilon,D)\;\overline{r_L^*}(D)\,]\;, \tag{20}$$

$$\partial r_L(\epsilon,D)/\partial\mathcal{L}_D = (\alpha/2)\,[\,\overline{r_L}(D) - t^2(\epsilon,D)\;\overline{r_R^*}(D) - r_L^2(\epsilon,D)\;\overline{r_L^*}(D)\,]\;, \tag{21}$$

and similarly for $r_R(\epsilon, D)$, where the bar over the reflection amplitudes denotes the averaging (8) over ϵ'.

We integrate now (20),(21) assuming that each of two barriers is characterized by $D_{r_{1,2}} \gg \Delta$. This condition means that either the barriers are strong initially at $D = D_0$ or get strong in the course of renormalization (12),(13) before D equals Δ. We will analyze the case of both or one of $D_{r_{1,2}}$ being smaller than Δ in Sect. 2.6.

Consider first the case of a resonance energy ϵ_0 lying exactly on the Fermi level, $\epsilon_0 = 0$. Let $\Gamma(D)$ be a renormalized width of the resonance peak at the Fermi energy (to be found below). If $D \gg \Gamma(D)$, then $|\overline{r_{L,R}}(D)| \simeq 1$, which allows for a significant simplification of (20),(21). It is convenient to introduce phase-shifted amplitudes $\tilde{r}_L = r_L e^{-i\varphi_{r_1}}$, $\tilde{r}_R = r_R e^{-i\varphi_{r'_2}+2\pi i(\epsilon_F+\epsilon_0)/\Delta}$, $\tilde{t} = t e^{-i(\varphi_{t_1}+\varphi_{t_2})}$, where φ_{r_1} is the phase of r_1, etc., in obvious notation. Then we get, by putting the averaged amplitudes far from the resonance $\overline{\tilde{r}_{L,R}}(D) = 1$:

$$\partial\tilde{t}(\epsilon, D)/\partial\mathcal{L}_D = -(\alpha/2)\ \tilde{t}(\epsilon, D)\,[\,\tilde{r}_L(\epsilon, D) + \tilde{r}_R(\epsilon, D)\,]\ , \tag{22}$$

$$\partial\tilde{r}_{L,R}(\epsilon, D)/\partial\mathcal{L}_D = (\alpha/2)\,[\,1 - \tilde{r}^2_{L,R}(\epsilon, D) - \tilde{t}^2(\epsilon, D)\,]\ , \tag{23}$$

with the following solutions

$$\tilde{t}(\epsilon, D) = \frac{[u_+^2(D) - u_-^2(D)]^{1/2}}{u_+(D) + 2i\epsilon}\ , \tag{24}$$

$$\tilde{r}_L(\epsilon, D) = \frac{u_-(D) + 2i\epsilon}{u_+(D) + 2i\epsilon}\ , \quad \tilde{r}_R(\epsilon, D) = \frac{-u_-(D) + 2i\epsilon}{u_+(D) + 2i\epsilon}\ , \tag{25}$$

where $u_\pm(D) = \Gamma_\pm(\Delta)(D/\Delta)^\alpha$, and $\Gamma_\pm(\Delta) = \Gamma_1(\Delta) \pm \Gamma_2(\Delta)$ should be found by matching onto (17)–(19). The width of a resonant tunneling peak $\Gamma(D)$ is thus given by $u_+(D)$.

Note that the only condition we have assumed in the above derivation is $D \gg \Gamma(D)$ with $D = \max\{|\epsilon|, T\}$, otherwise ϵ in (24)–(25) may be arbitrary. Thus, (24)–(25) give the shape of the ϵ dependence of fully renormalized amplitudes for the case of temperature $T \gg \Gamma(T)$ (with T substituted for D). In particular, the width of the resonance behaves as T^α:

$$\Gamma(T) = \Gamma_+(\Delta)(T/\Delta)^\alpha. \tag{26}$$

As follows from (24), while the resonance becomes sharper with decreasing T, the peak value of the transmission amplitude is not renormalized, since the D dependent factors cancel in (24) at $\epsilon = 0$. The absence of renormalization stems from the vanishing of the sum $\tilde{r}_L(0, D) + \tilde{r}_R(0, D)$ in (22).

We recognize (24)–(25) as Breit-Wigner solutions that take into account renormalization at $D < \Delta$. On the other hand, we have already obtained Breit-Wigner formulas in Sect. 2.3, where the renormalization has been carried out for $D \gg \Delta$, for ϵ close to a resonance energy. In particular, the results of Sect. 2.3 apply for $|\epsilon| \ll \Delta$ if $\epsilon_0 = 0$. The matching of the two solutions at $D \sim \Delta$ implies that (20),(21) are in fact valid in a broader range of D, namely for $D \ll D_{r_{\min}}$, provided only that one averages $r_R(\epsilon',D)$ over ϵ' together with the phase factor $\chi_{\epsilon'}$. It follows that in the case of strong barriers close to resonances the RG equations can be cast in the form (22),(23) containing $t(\epsilon,D)$ and $r_{L,R}(\epsilon,D)$ only. Note also that for $D_0 < \varDelta$ the boundary conditions to (22),(23) are fixed at $D = D_0$, which leads to the change $\varDelta \to D_0$ in $u_\pm(D)$.

At this point, one might be concerned about a possible contribution to $t(\epsilon = 0, D)$ from other resonances. Indeed, in the derivation of (24), which gives no renormalization of $t(\epsilon = 0, D)$, we approximated $|r(\epsilon',D)|$ by unity for large $|\epsilon'|, D \gg \varGamma(D)$. Corrections coming from other resonances are clearly small in the parameter $\varGamma_{1,2}(D)/\varDelta$ but one should check if they might contribute to the renormalization of $t(\epsilon = 0, D)$. Relaxing the above approximation by allowing for resonant "percolation" of electrons through the barriers at $D > \varDelta$ does give a perturbative correction to the RG (22):

$$\frac{\partial[\delta\tilde{t}(\epsilon,D)]}{\partial\mathcal{L}_D} = -\alpha\tilde{t}(\epsilon,D)\frac{\pi u_-(D)}{4\varDelta}\left[\,\tilde{r}_L(\epsilon,D) - \tilde{r}_R(\epsilon,D)\,\right] , \tag{27}$$

which, in contrast to (22), does not vanish at $\epsilon = 0$ (unless the double barrier is symmetric: the correction is then always zero). However, (27) tells us that the correction is irrelevant since $u_-(D)$ itself scales to zero as D^α. We thus conclude that the "single-peak approximation" of (22),(23) correctly describes the renormalization of the resonant amplitudes for all $D \gg \varGamma(D)$.

2.5 Inside a Peak: $D \ll D_s$

Now that we have integrated out all $D \gg u_+(D)$, let us continue with the renormalization for D inside a resonant tunneling peak. The point at which D and $u_+(D)$ become equal to each other yields a new characteristic scale D_s:

$$D_s = \varGamma_+(\varDelta)(D_s/\varDelta)^\alpha = \varGamma_+(\varDelta)[\,\varGamma_+(\varDelta)/\varDelta\,]^{\alpha'}, \tag{28}$$

where $\varGamma_+(\varDelta)$ is obtained from (17)–(19) depending on the ratio of $D_{r_{1,2}}$ and D_0. To leading order in α the exponent $\alpha' = \alpha/(1-\alpha) \to \alpha$. As will be seen below, the significance of D_s is that the width of the tunneling resonance saturates with decreasing D on the scale of D_s.

For $D \ll D_s$, (20),(21) can be simplified since the scattering amplitudes now depend on a single variable, which is $D = \max\{|\epsilon|, T\}$. The averaged reflection amplitudes $\overline{r_{L,R}}(D)$ coincide then with $r_{L,R}(D)$, and the RG equations can be written in precisely the same form as for a single impurity:

$$\partial t(D)/\partial\mathcal{L}_D = -\alpha t(D)\mathrm{R}(D) \ , \quad \partial r_{L,R}(D)/\partial\mathcal{L}_D = \ \alpha r_{L,R}(D)\mathrm{T}(D) \ , \tag{29}$$

with matching conditions at $\mathcal{L}_D = \ln(D_0/D_s)$. The difference between the single structureless impurity and the resonance peak is that in the latter case the ultraviolet cutoff is D_s.

Consider first the symmetric case of identical barriers. Assuming, as in Sect. 2.4, that the resonance energy $\epsilon_0 = 0$, we get

$$\tilde{t}(D) = \frac{u_+(D)}{u_+(D) + 2i\epsilon} \ , \tag{30}$$

which turns out to be valid down to $D = 0$. A remarkable consequence of (30) is that the resonance in the symmetric case is perfect, $\mathrm{T} = 1$ at $\epsilon = 0$. While this is trivial for noninteracting electrons, weak interaction in the Luttinger liquid is seen to preserve the perfect transmission, in agreement with the result obtained by a bosonic RG [3]. So long as inelastic scattering is not taken into account, the perfect transmission at $\epsilon = 0$ is not affected by finite T, either, as is seen from (30) if one puts $D = T$. However, the width of the resonance does depend on T. At $T = 0$, the width is finite and given by D_s, which follows from (30) for $D = |\epsilon|$. For $T \gg D_s$, the width obeys (26).

The shape of the perfect resonance depends on the parameter T/D_s. If $T \ll D_s$, the reflection probability as a function of $|\epsilon|$ behaves near $\epsilon = 0$ first as ϵ^2 for $|\epsilon| \ll T$ and then as $|\epsilon|^{2(1-\alpha)}$ for $T \ll |\epsilon| \ll D_s$. For larger energies, the transmission probability falls off with increasing $|\epsilon|$ as

$$\mathrm{T}(\epsilon) = (D_s/2|\epsilon|)^{2(1-\alpha)} \ . \tag{31}$$

This lineshape should be contrasted with the Lorentzian which describes the transmission peak for $T \gg D_s$ up to $|\epsilon| \sim T$, at which point a crossover to (31) occurs.

Let us now turn to the asymmetric case. Inspecting (29), we see that a new characteristic scale D_- emerges:

$$D_- = D_s[|\mathit{\Gamma}_-(\mathit{\Delta})|/\mathit{\Gamma}_+(\mathit{\Delta})]^{1/\alpha}, \tag{32}$$

which coincides with D_s for strongly asymmetric barriers but vanishes for symmetric ones. For $T > D_-$ we get the same results for $\mathrm{T}(\epsilon)$ as in the symmetric case, only with an overall factor of

$$\lambda = \frac{\mathit{\Gamma}_+^2(\mathit{\Delta}) - \mathit{\Gamma}_-^2(\mathit{\Delta})}{\mathit{\Gamma}_+^2(\mathit{\Delta})} = \frac{4\mathit{\Gamma}_1(\mathit{\Delta})\mathit{\Gamma}_2(\mathit{\Delta})}{[\,\mathit{\Gamma}_1(\mathit{\Delta}) + \mathit{\Gamma}_2(\mathit{\Delta})\,]^2} \ . \tag{33}$$

However, for $T \ll D_-$ a new feature in the behavior of $\mathrm{T}(\epsilon)$ shows up, namely a power-law fall off with decreasing $|\epsilon|$. The function $\mathrm{T}(D)$, as obtained from (29), for $D \ll D_s$ reads

$$\mathrm{T}(D) = \frac{\lambda(D/D_s)^{2\alpha}}{1 - \lambda\,[\,1 - (D/D_s)^{2\alpha}\,]} \ . \tag{34}$$

One sees that $\mathrm{T}(\epsilon)$ behaves as (Fig. 1a)

$$\mathrm{T}(\epsilon) = \lambda(|\epsilon|/D_-)^{2\alpha} \tag{35}$$

in the interval $T \ll |\epsilon| \ll D_-$ and saturates at smaller energies: $\mathrm{T}(\epsilon = 0) = \lambda(T/D_-)^{2\alpha}$. Thus, in the limit $T \ll D_-$, the resonant transmission probability as a function of ϵ exhibits a double-peak structure, see Fig. 1a. If, however, the barriers are only slightly asymmetric, the gap near $\epsilon = 0$ develops in a range of ϵ which is much narrower than the width of the resonance peak. Specifically, $\mathrm{T}(\epsilon)$ first arises with increasing $|\epsilon|$ for $T \ll |\epsilon| \ll D_-$, then there is a plateau with an energy independent transmission for $D_- \ll |\epsilon| \ll D_s$, and $\mathrm{T}(\epsilon)$ starts to fall off as $|\epsilon|$ is further increased.

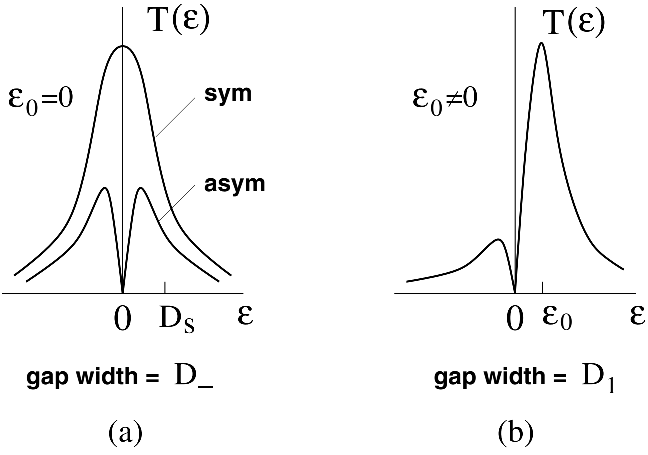


Fig. 1. Transmission peak structure for zero temperature and strong barriers: (a) $\epsilon_0 = 0$, symmetric vs asymmetric barriers; (b) $0 < |\epsilon_0| < D_s$, symmetric barriers

Above, we analyzed the behavior of $\mathrm{T}(\epsilon)$ for the resonance energy $\epsilon_0 = 0$, i.e., when it coincides with the Fermi level. Let us now examine the case $\epsilon_0 \neq 0$. Again, let the barriers first be symmetric. Then, in (30), the resonant denominator changes to $u_+(D) + 2i(\epsilon - \epsilon_0)$ and D is, as before, $\max\{|\epsilon|, T\}$. The innocent looking shift $\epsilon \to \epsilon - \epsilon_0$ leads at $T = 0$ to dramatic consequences for transmission at the Fermi energy, $\epsilon = 0$. Namely, $\mathrm{T}(\epsilon)$ is now seen to vanish at $\epsilon = 0$ and zero T, whatever ϵ_0 unless it is exactly zero. A new characteristic scale D_1 becomes relevant at $\epsilon_0 \neq 0$: it is defined by $u_+(D_1) = 2|\epsilon_0|$, which is rewritten as

$$D_1 = D_s(2|\epsilon_0|/D_s)^{1/\alpha} \ . \tag{36}$$

The significance of the energy D_1 is that the width of the gap in $\mathrm{T}(\epsilon)$ around $\epsilon = 0$ at $T = 0$ is given by D_1 for $|\epsilon_0| < D_s$. Note that $D_1 \ll |\epsilon_0|$ for $|\epsilon_0| \ll D_s$.

The shape of the resonant peak as a function of ϵ changes in an essential way for $|\epsilon_0| < D_s$. Specifically, if $T \gg D_1$, the changes are weak; however, for $T \ll D_1$ a range of ϵ arises, $T \ll |\epsilon| \ll D_1$, within which $\mathrm{T}(\epsilon)$ behaves as (Fig. 1b)

$$\mathrm{T}(\epsilon) = (|\epsilon|/D_1)^{2\alpha}. \tag{37}$$

The power-law fall off (37) saturates at $\mathrm{T}(\epsilon = 0) = (T/D_1)^{2\alpha}$.

We thus see that the width of the resonance in the transmission through a symmetric barrier exactly at the Fermi energy $\mathrm{T}(\epsilon = 0)$ as a function of ϵ_0 vanishes as $T \to 0$. On the other hand, the width of the resonance in the transmission at $\epsilon_0 = 0$ as a function of ϵ is finite even at $T = 0$ and is given by D_s. This peculiar feature is in sharp contrast to the resonant tunneling of noninteracting electrons, for which the two widths are the same.

For asymmetric barriers, $\mathrm{T}(\epsilon)$ does not change substantially with increasing $|\epsilon_0|$ as long as $D_1 \ll D_-$ and is given by the formulas for symmetric barriers with an overall reduction of $\mathrm{T}(\epsilon)$ by a factor of λ (33) otherwise.

2.6 Weak Barriers

Let us now examine the resonant transmission in the case of at least one barrier being initially so weak that the renormalization does not make it strong at $D \sim \varDelta$. We begin with the case of both barriers characterized by $D_{r_{1,2}} \ll \varDelta$. The total reflection coefficient $r_L(\epsilon, D)$ as obtained from (5),(6) in the limit $|r_L| \ll 1$ is given (with$D = \max\{|\epsilon|, T\}$) by

$$r_L(\epsilon, D) = (r_1 - r_2\chi_\epsilon)(D_0/D)^\alpha \ . \tag{38}$$

Suppose first that the barrier is symmetric and $\epsilon_0 = 0$. Then (38) simplifies to

$$\mathrm{R}(\epsilon, D) = 2[\,1 - \cos(2\pi\epsilon/\varDelta)\,]\,(D_r/D)^{2\alpha}. \tag{39}$$

One sees that reflection is enhanced by interaction, but the reflection coefficient is always small, $\mathrm{R} \ll 1$ for any ϵ, if $D_r \ll \varDelta$. No sharp features in the ϵ dependence of the scattering amplitudes emerge around the Fermi energy.

It is now instructive to introduce a weak asymmetry $\mathrm{R}_- = |\mathrm{R}_2 - \mathrm{R}_1|$, such that $\mathrm{R}_- \ll \mathrm{R} \simeq \mathrm{R}_{1,2}$. Given that the asymmetry is weak, it can manifests itself only at small energies. Expanding (38) about $\epsilon = 0$, we get for $|\epsilon| \ll \varDelta$:

$$\mathrm{R}(\epsilon, D) = \left[(\mathrm{R}_-/2\mathrm{R})^2 + (2\pi\epsilon/\varDelta)^2\right](D_r/D)^{2\alpha}\,. \tag{40}$$

As can be seen from (40), asymmetry sets two new characteristic scales of energy: $(\mathrm{R}_-/\mathrm{R})\varDelta$ and a smaller scale

$$\delta_- = D_r(\mathrm{R}_-/2\mathrm{R})^{1/\alpha}. \tag{41}$$

Provided that temperature $T \ll (\mathrm{R}_-/\mathrm{R})\varDelta$, the reflection coefficient starts to grow with approaching the Fermi level at $|\epsilon| \sim (\mathrm{R}_-/\mathrm{R})\varDelta$. The enhancement of reflection is cut off by temperature before R becomes of order unity if T is not too low, specifically if $\delta_- \ll T$. However, if $T \ll \delta_-$, then reflection gets strong at $|\epsilon| \sim \delta_-$. To describe the scattering probabilities at $|\epsilon| < \delta_-$, one should solve (29), derived in the same way it was done in Sect. 2.5, now with matching onto the perturbative (in $\mathrm{R}_{1,2}$) solution (38) anywhere in the region $\delta_- \ll |\epsilon| \ll \varDelta$. For $D \ll (\mathrm{R}_-/\mathrm{R})\varDelta$ the solution reads:

$$\mathrm{T}(D) = 1/\left[\,1 + (\delta_-/D)^{2\alpha}\right]\,. \tag{42}$$

Thus, the weak double barrier remains slightly reflecting after the renormalization provided that $T \gg \delta_-$. However, for both $T, |\epsilon| \ll \delta_-$ the transmission probability is small: within the range $T \ll |\epsilon| \ll \delta_-$, $\mathrm{T}(\epsilon)$ behaves as (Fig. 2a)

$$\mathrm{T}(\epsilon) = (|\epsilon|/\delta_-)^{2\alpha}\ , \tag{43}$$

and saturates for smaller $|\epsilon|$ at $\mathrm{T}(\epsilon = 0) = (T/\delta_-)^{2\alpha}$. Comparing (43),(35) with each other, we see that the energy δ_- is a counterpart of D_- for the case of a weak barrier.

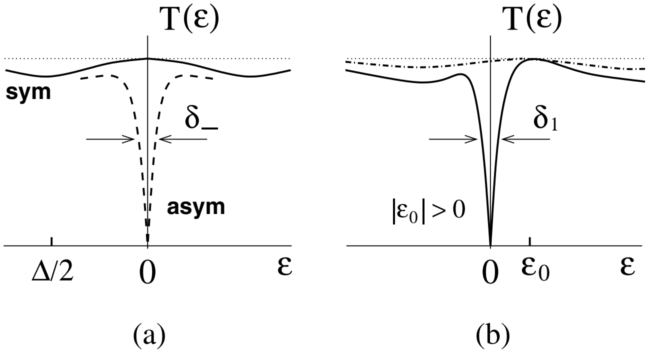


Fig. 2. Transmission coefficient for zero temperature and weak barriers: (a) $\epsilon_0 = 0$, symmetric (solid) and asymmetric (dashed) barriers; (b) $\epsilon_0 \neq 0$, symmetric barriers, the bare transmission (dash-dotted) is strongly renormalized (solid)

Generalizing to $\epsilon_0 \neq 0$, we have for $|\epsilon_0| \ll \varDelta$ a shift $\epsilon \to \epsilon - \epsilon_0$ in (40), while $D = \max\{|\epsilon|, T\}$. A new energy scale $\delta_1 \ll \epsilon_0$ appears, at which the reflection coefficient becomes of order unity in the symmetric case:

$$\delta_1 = D_r \, (2\pi|\epsilon_0|/\varDelta)^{1/\alpha}, \tag{44}$$

analogous to D_1 in (36) for tunneling barriers. The energy δ_1 gives the width of the gap in the transmission probability at the Fermi level at $T = 0$. For $T \ll |\epsilon| \ll \delta_1$, we get a power-law vanishing of $\mathrm{T}(\epsilon) = (|\epsilon|/\delta_1)^{2\alpha}$ with decreasing $|\epsilon|$ (see Fig. 2b) and a saturation for smaller $|\epsilon|$ at $(T/\delta_1)^{2\alpha}$. A general expression for the scattering probabilities, valid for arbitrary $D_{r_{1,2}} \ll \varDelta$ and $|\epsilon|, D \ll \varDelta$, can be obtained from (29):

$$\frac{\mathrm{R}(\epsilon, D)}{\mathrm{T}(\epsilon, D)} = \left[\left(\mathrm{R}_1^{1/2} - \mathrm{R}_2^{1/2} \right)^2 + (\mathrm{R}_1 \mathrm{R}_2)^{1/2} \left(\frac{2\pi}{\varDelta} \right)^2 (\epsilon - \epsilon_0)^2 \right] \left(\frac{D_0}{D} \right)^{2\alpha} . \tag{45}$$

Equation (45) reproduces (40)–(44) in the corresponding limits.

We conclude that if the barriers are symmetric but ϵ_0 is nonzero, or if the barriers are asymmetric, the transmission probability vanishes (Fig. 2) at the Fermi level in the limit $T \to 0$. We will see in Sect. 3 that these features lead to the emergence of a sharp peak in the low-T conductance as a function of ϵ_0 even for two weak impurities, provided only that they are slightly asymmetric.

Finally, when two strongly asymmetric barriers are located nearby, so that at $D \sim \varDelta$ one barrier is strongly reflecting whereas the other is still weak, the effect of the latter on the transmission probability remains small for any D. Let us take the example $D_{r_1} \gg \varDelta$ and $D_{r_2} \ll \varDelta$. Then we get for $D_{r_1} \gg D \gg \varDelta$

$$\mathrm{T}(\epsilon, D) = (D/D_{r_1})^{2\alpha} \left[1 + 2(D_{r_2}/D)^{\alpha} \cos\theta \right] , \tag{46}$$

where $\theta = 2\pi(\epsilon - \epsilon_0)/\varDelta$. One sees that the presence of the weak impurity only leads to a weak modulation with changing ϵ. For $D < \varDelta$, the independent renormalization of the weaker impurity is suppressed by reflection from the strong barrier and $\mathrm{T}(D)$ behaves as $(D/D_{r_1})^{2\alpha}$ down to $D = 0$.

3 Conductance Peak

The solution to the problem of transmission through a double barrier given in the preceding sections allows us to examine the linear conductance of the system $G(\epsilon_0, T)$ as a function of temperature T and the energy distance between the Fermi level and a resonance level ϵ_0. Recall that we have studied the elastic transmission of interacting electrons, i.e., the energy ϵ of an incident electron before and after the transmission is the same. At finite T, there are also inelastic processes, characterized by the inelastic scattering length L_{in}. Neglecting the inelastic scattering is legitimate if $L_{\text{in}} \gg L_T = v_F/T$, which is satisfied in the present problem for weak interaction $\alpha \ll 1$. Under these conditions (within the one-loop approximation, i.e., keeping only first order terms in the exponents), one can use the Landauer-Büttiker formalism relating the conductance and the transmission probability. The conductance $G(\epsilon_0, T)$ in units of e^2/h reads

$$G(\epsilon_0, T) = \int d\epsilon \; \mathrm{T}(\epsilon)\,(-\partial n_F/\partial \epsilon) \; . \tag{47}$$

We are interested in the low-temperature regime with $T \ll \varDelta$, otherwise we intend to keep T arbitrary, i.e., T may be as small as zero.

3.1 Strong Barriers

Consider first the case of strong barriers (more precisely, the bare transmission through the barriers may be high, $\mathrm{T}_{1,2} \simeq 1$, but we assume that the barriers get strong before the RG flow parameter D equals the single-particle energy spacing inside the dot, $\varDelta$), i.e., $D_{r_{\min}} = D_0(\min\{\mathrm{R}_1, \mathrm{R}_2\})^{1/2\alpha} \gg \varDelta$. Then we have a sharp peak of the transmission probability centered at $\epsilon = \epsilon_0$ whose width is $\max\{D_s, \varGamma(T)\} \ll \varDelta$, where D_s and $\varGamma(\mathrm{T})$ are defined in (26),(28). That is, the width of the peak in $\mathrm{T}(\epsilon, T)$ is $\varGamma(T) = D_s(T/D_s)^\alpha$ for $T \gg D_s$, whereas for smaller $T \ll D_s$ the width is of order D_s and does not depend on T.

***$T \gg D_s$*, Sequential Tunneling.** For $T \gg D_s$, the shape of the conductance peak is given by (Fig. 3)

$$G(\epsilon_0, T) = \zeta\, G_p \cosh^{-2}(\epsilon_0/2T) \; , \tag{48}$$

where the peak value of the conductance

$$G_p = \pi \lambda \varGamma(T)/8T \; , \tag{49}$$

with λ defined in (33), and $\zeta = (\max\{|\epsilon_0|, T\}/T)^\alpha$. The width of the conductance peak w is of order T, as for noninteracting electrons; however, the power-law behavior of $G_p(T)$ is seen to be modified by interaction, in accordance with the results derived in [13,16]. Note that the scaling of $G_p \propto T^{\alpha-1}$ is governed by the single-particle density of states $\rho_e(T)$ for tunneling into the end of a Luttinger liquid, namely $G_p \propto \rho_e(T)/T$ [20].

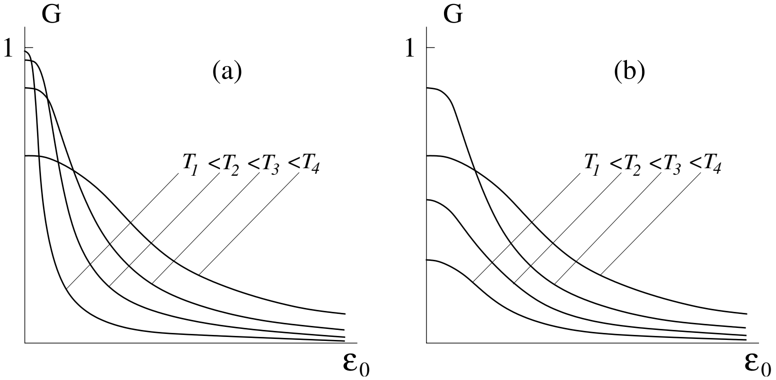


Fig. 3. Strong barriers: Conductance G (in units of e^2/h) as a function of ϵ_0 for symmetric (a) and asymmetric (b) barriers

We recognize (48),(49) as the conventional sequential tunneling formulas, but with a T dependent resonance width $\Gamma(T)$. Far in the wings of the resonance the exponential fall off (48), $G(\epsilon_0, T) \sim \lambda T^{-1}\Gamma(|\epsilon_0|)\exp(-|\epsilon_0|/T)$, crosses over onto the co-tunneling (determined by the processes of fourth order in tunneling amplitudes) power law $G(\epsilon_0, T) = \lambda\Gamma^2(T)/4\epsilon_0^2$, as usual. The crossover between the sequential tunneling and co-tunneling regimes occurs at $|\epsilon_0| \simeq T\ln[\,T/\Gamma(T)\,]$. In fact, this formula is valid [16] for an arbitrary strength of interaction with $\Gamma(T) \propto \rho_e(T) \propto T^{\alpha_e}$, where α_e (equal to α for a weak interaction) is the end-tunneling exponent. It is worthwhile to note that for strong enough interaction (namely for $\alpha_e > 1$) the sequential mechanism of tunneling is effective for the resonance peak for all T, down to $T = 0$ [16]. Moreover, for $\alpha_e > 1$ the crossover to the co-tunneling regime shifts towards larger $|\epsilon_0|$ with increasing strength of interaction.

***T* ≪ *D_s*, Symmetric Barriers.** Let us now turn to low temperatures $T \ll D_s$, where processes of all orders in the tunneling amplitudes are important. Consider first the symmetric case (Fig. 3a). The main contribution to the integral over ϵ in (47) comes from $|\epsilon| \sim T \ll D_s$, so that the shape of the conductance peak is a Lorentzian:

$$G(\epsilon_0, T) = \Gamma^2(T)/\,[\,\Gamma^2(T) + 4\epsilon_0^2\,]\ . \tag{50}$$

We see that the height of the peak $G_p = 1$ and the width

$$w = D_s(T/D_s)^\alpha \tag{51}$$

exhibits a power-law temperature dependence with an exponent depending on the strength of interaction. The vanishing of w as $T \to 0$ should be contrasted with the behavior of the peak in $\mathrm{T}(\epsilon, T)$ (Fig. 1), whose width is D_s for low T. In the limit $T \to 0$, the conductance peak becomes infinitely narrow but the resonance at $\epsilon_0 = 0$ persists down to $T = 0$, in accordance with [3]. We thus confirm the persistence [3] of the perfect resonance at $T = 0$ by means of the fermionic RG. For finite $T \ll D_s$, there is a small correction $1 - G_p \sim (T/D_s)^{2(1-\alpha)}$, which comes from the non-perfect transmission for finite ϵ at $\epsilon_0 = 0$ after the thermal averaging (47).

T $\boldsymbol{\ll}$ ***D*$_s$, Asymmetric Barriers.** We recall that the renormalization in the asymmetric case is governed by the scale D_- [defined in (32)] which describes the degree of asymmetry. The double-peak structure of the transmission coefficient as a function of ϵ_0 for $T \ll D_-$ translates into a complete vanishing of the conductance peak at $T \to 0$. Specifically, for $T \gg D_-$ the conductance $G(\epsilon_0, T)$ is given by (50) for symmetric barriers with an overall factor of λ. However, for $T \ll D_-$ the transmission at the Fermi level falls off with decreasing T (Fig. 1) and so does the conductance peak (Fig. 3b):

$$G(\epsilon_0, T) = \frac{\lambda(T/D_s)^{2\alpha}}{(D_-/D_s)^{2\alpha} + (2\epsilon_0/D_s)^2} \ , \tag{52}$$

which gives

$$G_p = \lambda(/D_-)^{2\alpha} \ , \quad w = D_s|\Gamma_-(\Delta)|/\Gamma_+(\Delta) \ . \tag{53}$$

Thus, at small $T \ll D_-$, the height of the conductance peak goes down as T decreases, whereas the width of the peak does not depend on T any longer. This kind of behavior of the resonance peak was predicted in [3]: $G_p(T)$ scales as $\rho_e^2(T)$, as for a single impurity. The role of asymmetry was emphasized and expressions similar to (53) were obtained in [24]. However, the authors of [24] do not distinguish between the scales D_- and D_s.

3.2 Weak Barriers

Consider now the case of weak barriers, i.e., $D_{r_{1,2}} \ll \Delta$. Naively one could think that scattering on weak barriers cannot possibly yield a sharp peak of conductance. Indeed, the transmission probability as a function of ϵ (Fig. 2) does not have any peak at $\epsilon = \epsilon_0$, in contrast to the case of resonant tunneling. At high $T \gg D_r$, $G(\epsilon_0, T)$ is a weakly oscillating (with a period Δ) function of ϵ_0. The only difference with the non-interacting case is an enhanced amplitude of the oscillations.

Let us show that in fact the interaction-induced vanishing of $\mathrm{T}(\epsilon)$ at the Fermi energy $\epsilon = 0$ for $T = 0$ does lead to a narrow Lorentzian peak of $G(\epsilon_0, T)$ (see Fig. 4), provided that T is low enough and the barriers are not too asymmetric.

Symmetric Barriers. Integration (47) of the transmission probability (45) for symmetric barriers yields

$$G(\epsilon_0, T) = \left[1 + (2\pi\epsilon_0/\Delta)^2 \, (D_r/T)^{2\alpha}\right]^{-1} \ , \tag{54}$$

which indeed describes a Lorentzian peak (Fig. 4a) with the height $G_p = 1$ and the width

$$w = (\Delta/\pi)(T/D_r)^{\alpha}. \tag{55}$$

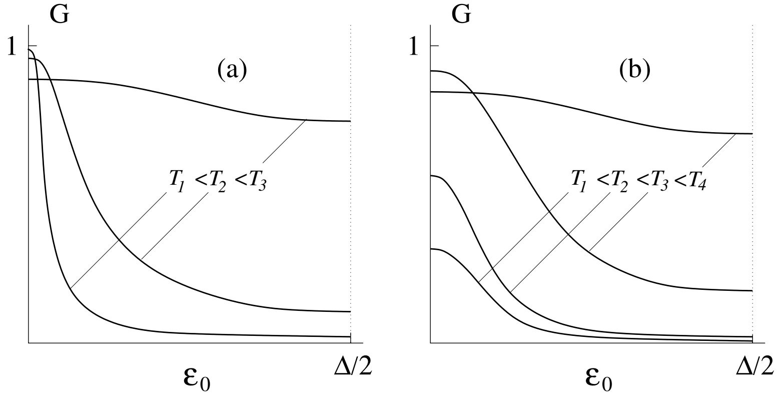


Fig. 4. Weak barriers: Conductance G (in units of e^2/h) as a function of ϵ_0 for symmetric (a) and asymmetric (b) barriers

It follows that the peak is narrow, $w \ll \varDelta$, provided that $T \ll D_r$. In the limit $T \to 0$, the width of the peak is infinitesimally small. Similarly to the case of resonant tunneling, the resonance at $\epsilon_0 = 0$ remains perfect in the presence of interaction at $T = 0$. At finite T, there is a small correction

$$1 - G_p \sim (T/\varDelta)^2 (D_r/T)^{2\alpha} \ . \tag{56}$$

Equation (56) describes also the high-T behavior of G_p in the case of slightly asymmetric barriers (see below).

Asymmetric Barriers. Introducing a weak asymmetry $\mathrm{R}_- = |\mathrm{R}_1 - \mathrm{R}_2| \ll \mathrm{R} \simeq \mathrm{R}_{1,2}$, we get for $T \ll \delta_-$, where δ_- is defined in (41):

$$G(\epsilon_0, T) = \frac{\mathrm{R}^2 (T/D_r)^{2\alpha}}{\mathrm{R}_-^2/4 + \mathrm{R}^2 (2\pi\epsilon_0/\varDelta)^2} \ . \tag{57}$$

The height and the width of the peak are

$$G_p = (T/\delta_-)^{2\alpha} \ , \quad w = (\varDelta/2\pi)\mathrm{R}_-/\mathrm{R} \ . \tag{58}$$

Thus, the asymmetry leads (Fig. 4b) to vanishing G_p at $T \to 0$ and the width is seen to saturate with decreasing T, similarly to (53). It is worth noting that the dependence of G_p on T is non-monotonic: $G_p \propto T^{2\alpha}$ grows with increasing T for $T \ll \delta_-$, continues to grow in the range $\delta_- \ll T \ll w$ according to $1 - G_p \propto T^{-2\alpha}$, but goes down for $w \ll T \ll \varDelta$, where the correction behaves similarly to the case of symmetric barriers, $1 - G_p \propto T^{2(1-\alpha)}$. The conductance peak is narrow provided the asymmetry is weak, $\mathrm{R}_- \ll \mathrm{R}$. If the asymmetry is strong, $\mathrm{R}_- \simeq \mathrm{R}_1 + \mathrm{R}_2$, the peak is completely destroyed.

4 Conclusions

In conclusion, we have studied transport of spinless electrons through a double barrier. We have described a rich variety of regimes depending on the strength of the barrier, its shape, and temperature. We have developed a fermionic RG approach to the double barrier problem, which has enabled us to treat on an equal footing both the resonant tunneling and resonant transmission through weak impurities. In the latter case, we have demonstrated how the interaction-induced renormalization in effect creates a quantum dot with tunneling barriers with a pronounced resonance peak structure. Moreover, we have shown that even very weak impurities, for which the renormalized transmission coefficient does not exhibit any peak, may give a sharp peak in the conductance as a function of gate voltage, provided that the double barrier is only slightly asymmetric. In contrast, the resonant structure is shown to be completely destroyed for a strongly asymmetric barrier. All the regimes we have studied may be characterized by three different types of behavior of the conductance peak height G_p and the peak width w on temperature T: (i) for high temperature $T \gg \Gamma(T)$, $G_p \propto T^{\alpha-1}$ and $w \propto T$; (ii) for lower T, depending on the shape of the barrier (whether it is symmetric or asymmetric), either G_p does not depend on T and $w \propto T^{\alpha}$ or (iii) $G_p \propto T^{2\alpha}$ and w is constant. None of the regimes (i–iii) supports $G_p \propto T^{2\alpha-1}$ and $w \propto T$, as proposed in [12]. Further experiments would be useful to resolve the puzzle.

Acknowledgments

We thank V.V. Cheianov, F. Evers, L.I. Glazman, D.L. Maslov, A.D. Mirlin, Yu.V. Nazarov, and P. Wölfle for interesting discussions. This work was supported by SFB 195 and SPP "Quanten-Hall-Systeme" of DFG, and by RFBR.

References

1. J. Sólyom: Adv. Phys. **28**, 201 (1979); J. Voit: Rep. Prog. Phys. **58**, 977 (1994)
2. A.O. Gogolin, A.A. Nersesyan, A.M. Tsvelik: *Bosonization and Strongly Correlated Systems* (Cambridge University, Cambridge 1998)
3. C.L. Kane, M.P.A. Fisher: Phys. Rev. B **46**, 15233 (1992)
4. A. Furusaki, N. Nagaosa: Phys. Rev. B **47**, 4631 (1993)
5. M. Bockrath, D.H. Cobden, J. Lu, A.G. Rinzler, R.E. Smalley, L. Balents, P.L. McEuen: Nature **397**, 598 (1999)
6. Z. Yao, H. Postma, L. Balents, C. Dekker: Nature **402**, 273 (1999)
7. Z. Yao, C.L. Kane, C. Dekker: Phys. Rev. Lett. **84**, 2941 (2000)
8. J. Nygård, D.H. Cobden, P.E. Lindelof: Nature **408**, 342 (2000)
9. M. Bockrath, W. Liang, D. Bozovic, J.H. Hafner, C.M. Lieber, M. Tinkham, H. Park: Science **291**, 283 (2001)
10. R. Krupke, F. Hennrich, H.B. Weber, D. Beckmann, O. Hampe, S. Malik, M.M. Kappes, H. v. Löhneysen: Appl. Phys. A **76**, 397 (2003)
11. O.M. Auslaender, A. Yacoby, R. de Picciotto, K.W. Baldwin, L.N. Pfeiffer, K.W. West: Phys. Rev. Lett. **84**, 1764 (2000)
12. H.W.Ch. Postma, T. Teepen, Z. Yao, M. Grifoni, C. Dekker: Science **293**, 76 (2001)
13. A. Furusaki, N. Nagaosa: Phys. Rev. B **47**, 3827 (1993)
14. M. Sassetti, F. Napoli, U. Weiss: Phys. Rev. B **52**, 11213 (1995)

15. H. Maurey, T. Giamarchi: Europhys. Lett. **38**, 681 (1997)
16. A. Furusaki: Phys. Rev. B **57**, 7141 (1998)
17. A. Braggio, M. Grifoni, M. Sassetti, F. Napoli: Europhys. Lett. **50**, 236 (2000)
18. A. Komnik, A.O. Gogolin: Phys. Rev. Lett. **90**, 246403 (2003)
19. R. Egger, A.O. Gogolin: Phys. Rev. Lett. **79**, 5082 (1997); Eur. Phys. J. B **3**, 281 (1998); C.L. Kane, L. Balents, M.P.A. Fisher: Phys. Rev. Lett. **79**, 5086 (1997).
20. M.P.A. Fisher, L.I. Glazman: in *Mesoscopic Electron Transport*, ed. by L.L. Sohn, L.P. Kouwenhoven, G. Schön (Kluwer, Dordrecht 1997)
21. K.A. Matveev, D. Yue, L.I. Glazman: Phys. Rev. Lett. **71**, 3351 (1993); D. Yue, L.I. Glazman, K.A. Matveev: Phys. Rev. B **49**, 1966 (1994)
22. C.L. Kane, K.A. Matveev, L.I. Glazman: Phys. Rev. B **49**, 2253 (1994)
23. S.-W. Tsai, D.L. Maslov, L.I. Glazman: Phys. Rev. B **65**, 241102 (2002)
24. Yu.V. Nazarov, L.I. Glazman: Phys. Rev. Lett. **91**, 126804 (2003)
25. A perturbative-in-α fermionic RG approach similar to that of Ref. [21] has been used as a basis for numerical simulations in V. Meden, W. Metzner, U. Schollwöck, O. Schneider, T. Stauber, K. Schönhammer: Eur. Phys. J. B **16**, 631 (2000); V. Meden, W. Metzner, U. Schollwöck, K. Schönhammer: J. Low Temp. Phys. **126**, 1147 (2002)
26. M. Thorwart, M. Grifoni, G. Cuniberti, H.W.Ch. Postma, C. Dekker: Phys. Rev. Lett. **89**, 196402 (2002); M. Thorwart, M. Grifoni: Chem. Phys. **281**, 477 (2002)
27. D. G. Polyakov, I. V. Gornyi: Phys. Rev. B **68**, 035421 (2003)

Interference and Interaction in Metallic Nanostructures

Heiko B. Weber

Forschungszentrum Karlsruhe, Institut für Nanotechnologie (INT), Postfach 3640,
76021 Karlsruhe, Germany

Abstract. The electron transport across metallic nanobridges at low temperatures and in good contact with electric leads is reviewed. The smallness of the samples induces two phenomena which are different from macroscopic metals. First, interference effects like universal conductance fluctuations and Aharonov-Bohm effect are observable as a result of quantum-mechanical phase coherence over the entire sample. Second, interaction effects are considered, which are induced by a reduced screening of the carrier charges. Examples for these phenomena are discussed both at zero- and finite bias voltages.

1 Introduction

The conductance of a macroscopic conductor is generally described by Ohm's law and is related to microscopic processes in the framework of the Boltzmann equation. The conductance G can be written as a product of the geometry factor and the material-specific conductivity

$$G = \frac{A}{L} \cdot \sigma \tag{1}$$

with A the cross section of the conductor, L its length and σ the conductivity, which is an average local quantity. In the presence of scattering processes, the electron movement is diffusive. Then, the conductivity can be related to the mean free time τ between scattering events that affect the momentum state of the carrier. Drude's result is

$$\sigma = \frac{ne^2\tau_{\mathrm{Drude}}}{m} \tag{2}$$

with e the elementary charge, n the charge carrier density and m the effective mass of the conduction electrons [1].

A different description has to be chosen when the sample is nanometer-sized. To give a drastic example, the conductance of a single-atom contact of gold (see Fig. 1) is close to the conductance quantum $G_0 = 2e^2/h$, which is a fundamental constant. One might expect that a longer chain of gold atoms with several gold atoms in a row will have a reduced conductance according to (1). Experiments and theory have demonstrated that the conductance value is again very close to G_0 and depends only weakly on the exact geometry [2]. This surprising result indicates that the concepts to describe electron transport on the nanoscale

H.B. Weber, Interference and Interaction in Metallic Nanostructures, Lect. Notes Phys. **658**, 185–203 (2005)
http://www.springerlink.com/

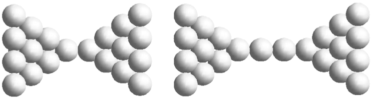


Fig. 1. Sketch of gold contacts with a single-atom cross section. The conductance of both contacts is $\approx 2e^2/h$, independent of the length. This provides a drastic example how electron transport at the nanoscale is completely different from macroscopic physics (Ohm's law)

are different from the macroscopic ones. Indeed, the conductance is frequently described by the Landauer formula

$$G = \frac{2e^2}{h} \sum_i \tau_i \tag{3}$$

which is a summation over i available modes with τ_i being the corresponding transmission probabilities. This description essentially treats the electron transport as a wave-like propagation, where electrons may access some propagation modes and are either transmitted or reflected. In strong contrast to (1), the geometry, the mean free path and the electron density are not explicitly included, although they determine implicitly the conductance in a nontrivial way.

It turns out that the Landauer description is not only suitable for the single-atom scale, but is also very useful on the *meso* scale in between macroscopic and atomistic dimensions (greek: *mesos* = middle). When the sample is entirely quantum mechanically phase coherent, the concept of ensemble averaging breaks down, which is the base of the Boltzmann description. Discoveries made since the 1980s showed that this happens at surprisingly large length scales, which are typically around 1 µm in a metallic sample at low temperatures ($T \approx 1$ K).

The present article describes interference and interaction effects in mesoscopic metallic conductors. First, the theoretical framework will be sketched, then the focus will be on experimental phenomena. Whereas many groups world wide have contributed to the research in this field, the phenomena will be explained using experimental results we obtained over the last decade in Karlsruhe.

2 Conductance is Transmission

Consider a perfect conductor with length L (no impurities) which has a cross section sufficiently small that only one current carrying mode is occupied. In order to determine the conductance, it is connected to two reservoir-like leads. This means that the electron system in each lead is in thermodynamical equilibrium. Hence, the temperature and the electrochemical potential is perfectly defined. The current across the conductor is basically the product of the number of charge carriers, their velocity v, the elementary charge e. In a more accurate quantum mechanical description this can be written as

$$I = \frac{e}{L} \sum_{k,\sigma} v_k (f_L(\epsilon_k) - f_R(\epsilon_k)) \tag{4}$$

or for a long conductor, in which the k states are dense (the factor 2 appears due to spin degeneracy):

$$I = \frac{2e}{\pi} \int dk \cdot v_k (f_L(\epsilon_k) - f_R(\epsilon_k)) \tag{5}$$

where f_L and f_R denote the Fermi-Dirac distribution function

$$f_{L,R} = \frac{1}{1 + \exp\left(\frac{E - \mu_{L,R}}{k_B T}\right)}$$

of the left and the right lead, respectively. The current is non-zero when a finite voltage V is applied, which shifts the electrochemical potential μ by eV with respect to the opposite side. When the transition to energy coordinates is desired, the one-dimensional density of states comes into play, which is

$$\rho = \frac{1}{v_k \hbar} \tag{6}$$

and thus the velocities exactly cancel and drop out of the calculation for the current:

$$I = \frac{2e}{h} \int d\epsilon (f_L(\epsilon_k) - f_R(\epsilon_k)), \tag{7}$$

which results in the conductance

$$G = \frac{2e^2}{h} \int d\epsilon \left(- \frac{\partial f}{\partial \epsilon} \right) . \tag{8}$$

Hence, the conductance of a perfect single-mode one-dimensional wire connected to two reservoirs has the above value which simplifies at zero temperature to $G_0 = 2e^2/h$, which is the conductance quantum. It corresponds to the quantum resistance of $R_0 = \frac{h}{2e^2} = 12.9$ kΩ. It is remarkable that this result is independent of the size (within the above-mentioned assumptions) and of the material. Note that in the 3D case in the absence of scatterers zero resistance would be expected.

For the metallic nanosamples that are described in this chapter, the assumption of the single-mode conductor is not appropriate. For this purpose, the concept can be extended to the case of a many-mode conductor[3]. This is frequently called the Landauer formalism: The conductor is considered as a scattering region, which is placed in between two reservoir-like leads. In the scattering region, there are several input channels and several output channels (cf Fig. 2). Waves which come from the left side are either backscattered in one of the left-hand leads, or transmitted in one of the right-hand leads. Technically, the scattering center is given by a transmission matrix (T matrix) for each energy. Its matrix elements t_{nm} relate a wave incoming from the left in channel n with probability 1 to the outgoing wave leaving the scattering region along channel m with the probability t_{nm}. The incoming wave is thus distributed over all outgoing channels or is reflected in one of the input channels such that current conservation is respected. This matrix can in principle be calculated starting from a

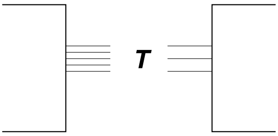


Fig. 2. The transmission matrix T distributes waves coming from the left reservoir via input channels to outgoing waves to the right reservoir. The number of channels may be different for both sides

Schrödinger equation, a procedure, which will not be discussed in this article (see [4]).

Once the T matrix is known for each energy, the calculation of the conductance is straightforward. The total transmission is given by a sum over all transmitted amplitudes $T = \sum_i \sum_j |t_{i,j}|^2 = tr(t^\dagger t)$. Hence, the conductance reads

$$G = \frac{2e^2}{h} \int d\epsilon \left(-\frac{\partial f}{\partial \epsilon} \right) tr(t^\dagger t) \, . \tag{9}$$

At zero temperature, this simplifies to

$$G = \frac{2e^2}{h} tr(t^\dagger t) \, . \tag{10}$$

The eigenvectors of $(t^\dagger t)$ are called transmission eigenchannels. In their basis, the conductance formula simplifies to

$$G = \frac{2e^2}{h} \sum_i \tau_i \tag{11}$$

with real eigenvalues τ_i which range between 0 and 1, representing the transmission probability of this mode.

We find again formulas similar to (8), but generalized to many modes. Within this formalism, the conductance is the sum of transmission values, multiplied by the conductance quantum. It is valid if the interaction between charge carriers can be neglected. Because the incoming waves are simply redistributed to the output waves and phase relations and the superposition principle are respected, the formalism is particularly useful for describing interference effects in phase coherent samples. For a more thorough treatment, [4] is recommended. In Sect. 5 we will discuss interaction effects, which are not tractable with the Landauer formalism.

3 Quantum Interference Experiments

3.1 The Mesoscopic Aharonov–Bohm Effect

If we consider a coherent metallic bridge with mesoscopic dimensions (for example a copper bridge, which is 50 nm wide and thick, 1000 nm long with a

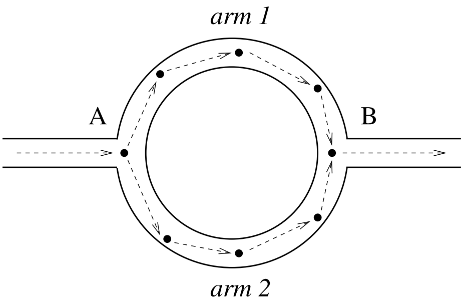


Fig. 3. A mesoscopic Aharonov Bohm ring: When an electron comes from the left hand side, it splits at A into two partial waves, which propagate along the two arms. In each arm, many scattering processes change the phase in a deterministic, but random way. At B, both waves interfere according to their relative phase

resistance of $R = 20\ \Omega$), its resistance will roughly be determined by the classical formulas ((1) and (2)) and the quantum mechanical corrections will be rather small. In order to access the quantum effects experimentally, a further tuning parameter is required. In a semiconductor device this can be given by electrostatically induced changes of the geometry (see also the contribution of Jürgen Weis in this book)). In the case of a metallic sample, such changes may be induced by the magnetic field. The following experiment will elucidate the basic mechanism. Consider a metallic conductor, which splits into two arms (Fig. 3). Charge carriers which come from the left may choose one of the arms 1 and 2 with probability T_1 or T_2, respectively. In a quantum mechanical description electrons are considered as waves and split at point A into two partial waves which propagate along the two arms. At point B these partial waves rejoin and are subject to interference. Following the last section, the interference will affect the transmission and consequently the conductance. The amplitude of the reconvened beam will depend on the relative phase ϕ according to

$$T_{A\to B} = |T_1 + T_2|^2 = |T_1|^2 + |T_2|^2 + 2\,|T_1|\,|T_2| \cdot \cos\phi \tag{12}$$

where ϕ is unknown, because it depends on many microscopic details, which add random phases in each arm. However, we know from quantum mechanics that the interference between trajectories can be altered by the presence of electromagnetic fluxes. Aharonov and Bohm have demonstrated theoretically that for any pair of trajectories that interferes with a transmission probability T in the absence of an electromagnetic field, the probability changes to

$$T' = T \exp\left(\frac{ie}{\hbar}\int(\mathit{\Phi}_{\mathrm{el}}dt - \boldsymbol{A}d\boldsymbol{r})\right), \tag{13}$$

where $\mathit{\Phi}_{\mathrm{el}}$ and $\boldsymbol{A}$ are the potentials of the electric field $\boldsymbol{E} = \boldsymbol{\nabla}\mathit{\Phi}_{\mathrm{el}}$ and magnetic field $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$, respectively [1]. As a consequence, the magnetic flux $\mathit{\Phi}_{\mathrm{magn}} =$

[1] This can be deduced from the Schrödinger equation when the usual substitution for the momentum $\boldsymbol{p} \to \boldsymbol{p} - e\boldsymbol{A}$ is carried out [5].

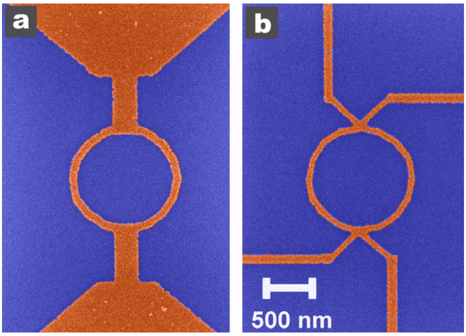


Fig. 4. SEM pictures of metallic Aharonov-Bohm rings. a: two-terminal setup, b: four-terminal setup, which will be discussed in Sect. 4.1.

$\int \boldsymbol{B} d\boldsymbol{s}$ in between the trajectories affects immediately the quantum mechanical phase ϕ which is ruling the interference and can continuously be tuned by the magnetic field $\boldsymbol{B}$

$$\Delta\phi(B) = \frac{e}{\hbar}\oint \boldsymbol{A} d\boldsymbol{l} = \frac{e}{\hbar}\int \boldsymbol{B} d\boldsymbol{s} = 2\pi\frac{\Phi_{\mathrm{magn}}}{\Phi_0} \tag{14}$$

where Stokes' theorem was applied. For simplicity, $\boldsymbol{B}$ is usually chosen perpendicular to the ring plane and then $\Phi_{\mathrm{magn}} = \pi r^2 B$. The fundamental constant $\Phi_0 = h/e = 4.15 \cdot 10^{-15}\ Tm^2$ is called the flux quantum. Remarkably, the effect is not caused by the field in which the electron is moving, but by the magnetic flux which is enclosed by the two paths. An equivalent experiment with free electrons has been proposed by Aharonov and Bohm in 1959 [6] and being performed by Chambers 1960 [7]: An electron beam was used in a double slit experiment with a coil in between the two slits, which provided a magnetic flux (in the space between the two partial beams). By constantly increasing the current in the coil, the interference pattern behind the double slit could be continously shifted.

Here, the same physics is recovered in a solid state device with conduction electrons. We have already seen that the transmission probability is intimately linked to the conductance and therefore we would expect that also the conductance depends sensitively on the interference term and thus on the applied flux. An experiment that displays these features (first carried out in 1984 [8]) must provide the following prerequisites: (i) A ring-shaped sample has to be patterned, which is micrometer-sized in order to allow for a full phase-coherence at low temperatures ($T < 1$ K). Here, a ring diameter of 1 µm has been chosen. (ii) The arms of the ring should be rather narrow to define the cross-section of the ring sufficiently well (the results of a finite aspect ratio are discussed below). For our experiment, $w = 80$ nm were reliably achievable. (iii) The ring has to be contacted by two leads within the phase coherent area (Fig. 4a); the case of a four-terminal measurement displayed in Fig. 4b is described in Sect. 4.1. Such an Aharonov-Bohm (AB) ring can be produced by e-beam lithography and the lift-off technique. Figure 4a shows such a copper sample on a scanning electron micrograph. The resistance of this particular sample is $R = 55.6\ \Omega$,

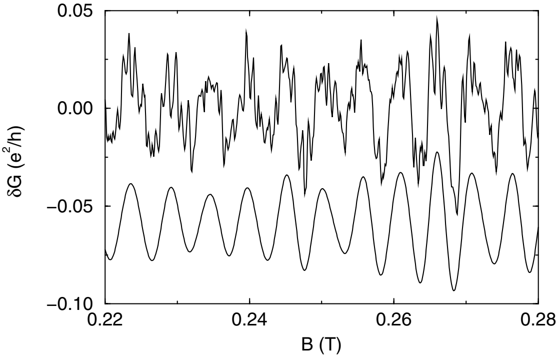


Fig. 5. The conductance oscillations $\delta G = G(B) - \langle G \rangle$ in an Aharonov-Bohm ring induced by the magnetic field B. The upper data set are raw data with both aperiodic and periodic fluctuations, the lower data set after narrow-band filtering

corresponding to a conductance of $G = 464\ \frac{e^2}{h}$. Figure 5 (upper data set) shows the fluctuations of the conductance obtained with this sample at $T = 20$ mK as a function of the magnetic field. The signal is fluctuating. Only a minor part of these fluctuations is electronic noise, the dominant part is reproducible when the measurement is repeated (of course, some electronic noise is also present). Both periodic and aperiodic fluctuations (so called universal conductance fluctuations, see below) appear. When an appropriate narrow-band filter procedure picks out the periodic signal (lower data set in Fig. 5), a long-range periodicity can be found, with some modulation on top. According to (14), the signal is periodic in the flux and therefore the basic "frequency" of $G(B)$ depends on the area $S = \pi r^2$ of the ring:

$$\Phi_0 = B_C \cdot S\,. \tag{15}$$

In order to enclose one flux quantum $\Phi_0 = B{\cdot}S$ in the ring, a field of $B_C = 62$ mT is needed in this case. Figure 5 shows also some modulation or beatings, which indicate that the frequency is not sharply assigned. This is a consequence of the finite width of the ring: The area in between two paths in the upper and lower ring is not accurately defined. For example, an electron which is travelling along the outermost possible path and one which is travelling the innermost path are differently affected by the magnetic field. An important finding is that the amplitude of the oscillation in a metallic AB-ring with many open paths is on the order of $G_0 = 2e^2/h$, no matter of the value of conductance. Whereas in a single-mode conductor with a conductance $G \approx G_0$ this may be a large effect, in a metallic nano-ring with many open channels, the effect is small, but not negligible.

3.2 Universal Conductance Fluctuations

Now we consider a metallic nanobridge as schematically shown in Fig. 6. Again, we assume phase coherence over the entire bridge, which is connected to two ideal

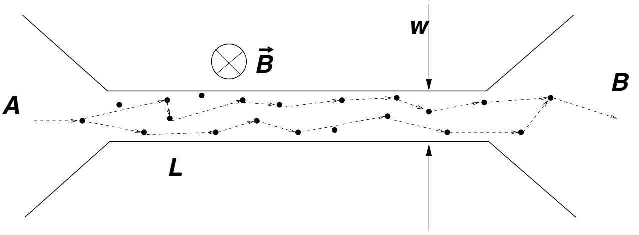


Fig. 6. A mesoscopic bridge in between reservoir-like leads. Electrons coming from point A may propagate along different diffusive trajectories. When they rejoin in point B, the quantum mechanical phases of the two trajectories have to be respected

leads. Conduction electrons travelling along the bridge can again take different paths according to the diffusive motion. In Fig. 6, only two of such paths are drawn for simplicity. An electron starting at point A splits into two partial waves which travel along different paths and rejoin in point B, where again interference is important. In analogy to the AB ring experiment, the relative phase can be tuned by an external magnetic field. However, the geometry is not well defined, many paths are affected simultaneously and for the conductance pattern no periodicity can be expected.

Figure 7 displays conductance data obtained with several of such bridges [2]. Similarly to Fig. 5, there are reproducible fluctuations observable as a function of the magnetic field. In this case, however, only aperiodic fluctuations can be detected. In order to characterize such curves, two statistical values can be extracted from $G(B)$: the amplitude of the fluctuations $\delta G = \langle G(B) - \langle G \rangle \rangle$ and a correlation field B_C, defined as the width (HWHM) of the autocorrelation function

$$F(\varDelta B) = \frac{1}{2B_0} \int_{-B_0}^{B_0} \delta G(B') \delta G(B' + \varDelta B) dB' \,. \tag{16}$$

B_C characterizes the magnetic field scale on which the conductance changes significantly. It corresponds in this respect qualitatively to the period of the AB-fluctuations.

In analogy to the AB oscillations, the amplitude is $\delta G \approx 0.25\ e^2/h$, irrespective of the conductance of the sample (this is why they are called *universal* conductance fluctuations). For the correlation length, a simple argument yields a qualitative value: the two paths shown in Fig. 6 include an area which is on the order of $l \cdot w$. Imagine the bridge is twice as wide, two such typical paths would include twice as much area and consequently twice as much flux. So, within a factor C of the order of one,

$$B_c = C \frac{\varPhi_0}{wl} \tag{17}$$

gives a reasonable estimate of the field scale on which the correlations are completely destroyed and new correlations are randomly established. Figure 7 shows

[2] The first experimental observation of this effect in metals was reported in 1984 [9]

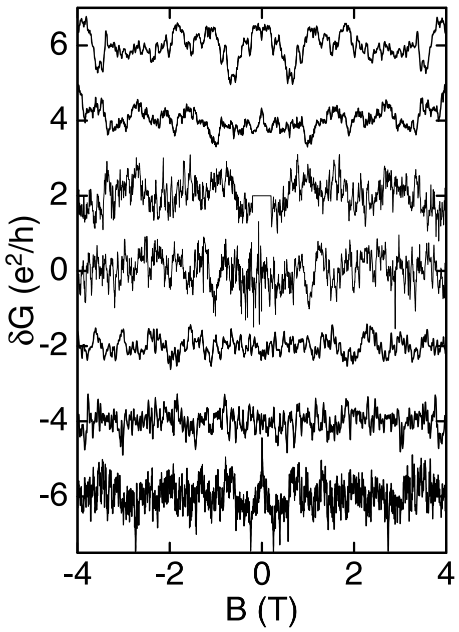


Fig. 7. Universal conductance fluctuations recorded with metallic nanowires of different widths (from top to bottom $w = 45, 80, 100, 120, 160, 240, 320$ nm). One can clearly see that the fluctuation width is reduced for larger w, whereas the fluctuation amplitude δG is unaltered. Note also the symmetry $G(B) = G(-B)$

a systematic investigation of bridges with different width [10]. One can see immediately that the average fluctuation amplitude is very similar, whereas the fluctuation width on the field scale gets broader for narrower bridges. In Fig. 8, the correlation fields B_C for bridges with different widths w are determined statistically and qualitatively confirm equation (17).

For a given sample, the conductance fluctuations generate a pattern as a function of the magnetic field B, which is specific for this sample. But a nominally equivalent sample with the same geometry will have a different pattern. This difference arises due to the importance of microscopic details, in particular the position of the scattering centers which determine the diffusion paths. Allowing for some microscopic reorientation of defects in a given sample (for example by warming it up and recooling it thereafter) is often sufficient to alter the conductance pattern substantially. If one would measure the zero-field conductance of a large ensemble of bridges with identical geometry, the variation of the conductance is expected to show the same statistical fluctuations $\delta G \approx 0.25\ e^2/h$. Variation of the magnetic field for a given sample or alternatively investigating many samples at a fixed field value provide two different ensembles which both explore the phase space randomly. According to the ergodic hypothesis the same statistical variance is then expected [11,12]. Another parameter which randomly

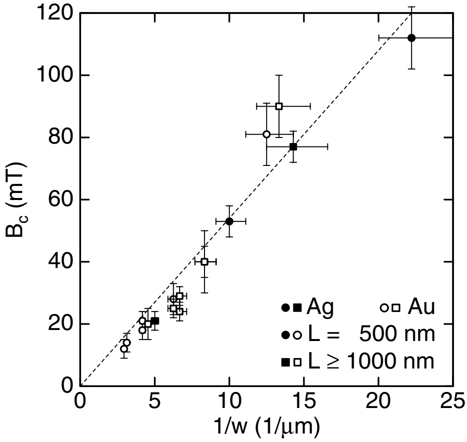


Fig. 8. A statistical evaluation of the data displayed in Fig. 7 yields a correlation field B_C for bridges of different width. The straight line confirms that $B_C \propto 1/w$

probes the phase space for a given sample is the Fermi velocity of the conduction electrons. Related experiments are described in the next section.

4 Interference out of Equilibrium

The conductance measurements described so far were carried out at small voltages: the excitation energies provided by the voltage was comparable to the thermal broadening $eV \approx k_B T$. This is the equilibrium regime, in which the voltage does not play an important role. When the energy exceeds $k_B T$, the voltage can be a relevant energy scale. Some examples may elucidate the physics involved.

4.1 Conductance Fluctuations at Finite Voltages

Applying a voltage to a nanobridge means shifting the electrochemical potential of the two leads by eV with respect to each other (cf (5)). Instead of characterizing the conductance I/V, it is more convenient to look at the differential conductance dI/dV. The meaning of this physical observable can be seen in Fig. 9: As a consequence of the imbalance of the electrochemical potentials, current can flow from occupied states in the left reservoir to unoccupied states at the right reservoir at all energy levels between μ_L and μ_R. When the applied voltage V is further increased by δV, the current response is increased by δI. The ratio between these values $\delta I/\delta V$ is the differential conductance: In the simple case, this corresponds to the transparency of the additionally opened energy level. But in some cases the current channels which are already active are altered by the increase of the voltage. Then, a slight increase of V may affect the transparency of all levels.

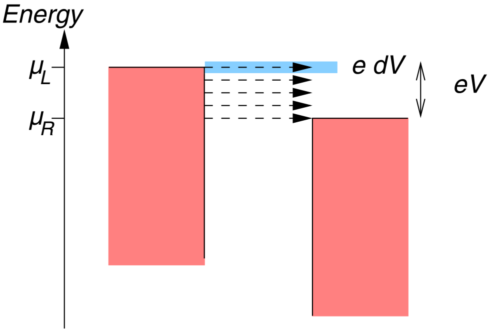


Fig. 9. When a voltage is applied to a nanostructure (not shown) in between two reservoir-like electrodes, the two electrochemical potentials μ_L and μ_R of the leads are shifted with respect to each other by eV. Electrons from occupied states (shaded areas) in the left lead can flow into unoccupied states in the right lead. When the voltage is slightly modified by dV, the current correspondingly varies by dI. The differential conductance at a given voltage V is resulting as $dI/dV(V)$

Experimentally, the differential conductance is often measured by a modulation technique. A small excitation voltage V_{exc}(comparable to k_BT/e) is added to the bias voltage V_0. The current response has an AC component which is proportional to the differential conductance.

$$V(t) = V_0 + V_{\text{exc}} \sin(\omega t) \tag{18a}$$

$$I(t) = GV_0 + dI/dV V_{\text{exc}} \sin(\omega t + \delta) \,. \tag{18b}$$

The AC current response can be measured with good resolution using lock-in techniques.

Beyond which voltage does the physical picture change? The key issue for the interference experiments is that the momenta and the corresponding wavelengths for the various energy levels in Fig. 9 are different. When we look at waves with different wavelengths with a defined phase relationship, they will run out of phase after a given time τ_{ph}. If the time τ_C needed to traverse the sample is shorter than τ_{ph}, the interference pattern will only slightly be altered. If it is larger, then one part of the electrons will be incoherent to another part on their diffusion along the nanostructure. Expressed in terms of energies, the length of the sample implies an energy scale $E_C = h/\tau_C = hD/L^2$, which is also called the Thouless energy. As long as eV is smaller than this energy scale, all conduction electrons are phase coherent. If, however eV exceeds the Thouless energy E_C, the electrons are partially incoherent to each other. If the applied voltage is, for example $V = N \cdot E_C$, then N incoherent channels at energy levels will cause fluctuations. The incoherent superposition will then scale with the number $N^{1/2}$ and consequently with the voltage $U^{1/2}$. This has been predicted by Larkin and Khmel'nitskii 1986 [13].

The consequences can be seen in an interference experiment at finite voltage, for example in an AB-ring. In our experiment, we vary the current and record

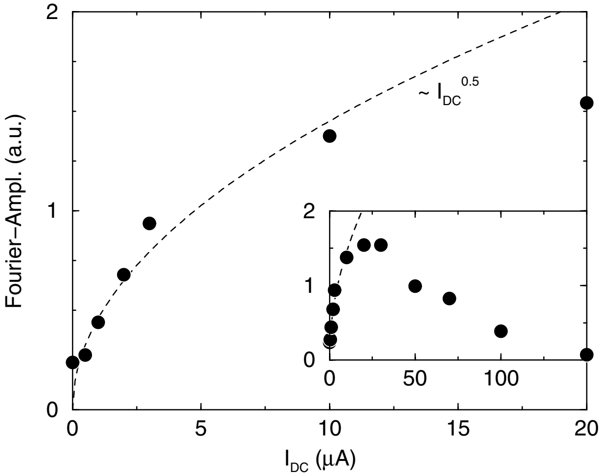


Fig. 10. The amplitude of Aharonov-Bohm oscillations in the differential resistance in a mesoscopic ring as a function of the applied bias current. One can see the square-root increase up to a level where incoherent processes set in and destroy the phase coherence (see inset)

the conductance oscillations[3] [14]. The Fourier amplitude that describes the amplitude of the oscillatory part increases with the square root of I_{DC} (Fig. 10). Tuning the currents higher, an additional trivial effect comes into play: heating destroys the phase coherence (see inset). When we limit ourselves to the regime where heating can be neglected, quantum interference becomes visible as a function of finite bias voltages.

These curves are gained by evaluating many data sets $G(B)$, each recorded at a given bias current for a magnetic field sweep. If we focus our attention onto one of such data sets, an interesting fundamental question arises, which is based on the combination of two arguments: (i) We have seen that the conductance is proportional to the cosine of the flux and consequently $G \propto \cos(B/B_C + \alpha)$, with a zero-field phase of alpha which naively can be assumed to be a random value due to the interference of two partial waves (cf Fig. 3). (ii) A very general symmetry argument, which is based on Onsager's relation [15], tells us that for a two-terminal measurement (i.e. within the phase coherent region there are two reservoir-like leads), the conductance should not change when the time is reversed [16]. One can easily rationalize that a time reversal would change the magnetic field on the one hand, but the current and consequently the voltage along the sample on the other hand, therefore the conductance $G = I/V$ should not be affected. Hence, one expects

$$G(B) = G(-B) . \tag{19}$$

[3] Due to technical limitations, the leads were rather far apart from the ring and thus the voltage applied is not accurately defined. Our results compare to the Larkin-Khmel'nitskii results only qualitatively.

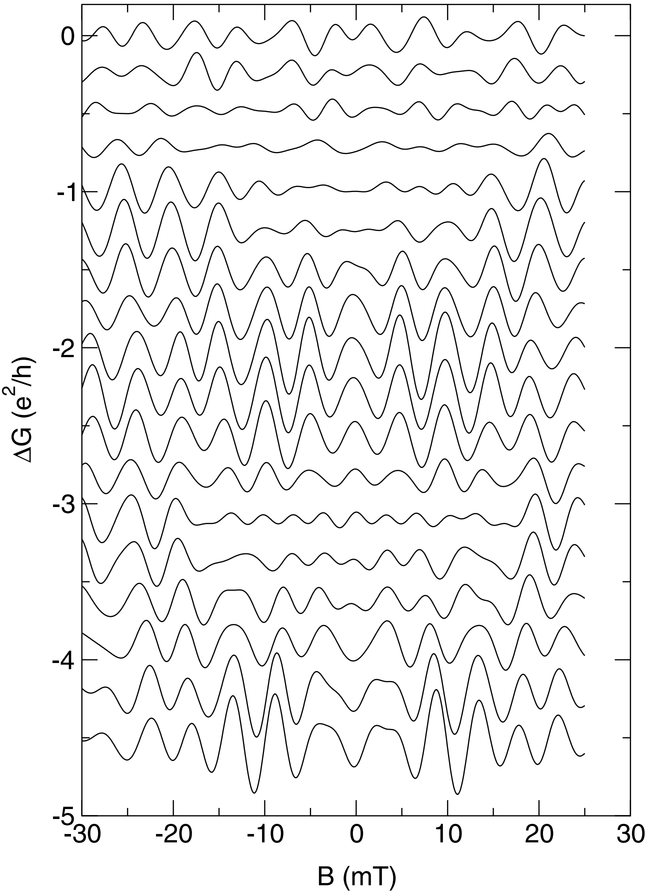


Fig. 11. Filtered Aharonov-Bohm oscillations in a two-terminal setup for different applied bias voltages. At zero field, either a maximum or a minimum is observable, with some flat transition regions. This symmetry is a consequence of Onsager's relation

Combining (i) and (ii), it can easily be deduced that $G \propto \cos(B/B_C)$ or $G \propto -\cos(B/B_C)$, any other phase α is not in agreement with time reversal symmetry.

Our approach to address this phenomenon experimentally is to vary the current in an AB ring in order to induce a random change in the interference pattern and to study the $dI/dV(V,B) = G_V(B)$ traces with respect to their zero-point symmetry [17]. Figure 11 shows such data with a mesoscopic AB ring. The plot shows a number of $dI/dV(B)$ curves recorded at various bias current values, which are in this context only a parameter to vary the interference patterns randomly. The data are filtered such that only the periodic part is visible.

It can be seen that for all bias currents the differential conductance $G_V(B)$ shows a symmetry with respect to $B = 0$. Whereas in the uppermost set there is a minimum at $B = 0$, this minimum is broadened at higher currents and is converted to a maximum at $B = 0$ for $I = 0.36$ μA and then again to a minimum at $I = 6.8$ μA. Only slight asymmetries can be observed [4]. For comparison, a ring was manufactured for which two current leads and two voltage leads are present within the phase coherent area (see Fig. 4b). In a classical system this setup would allow for the accurate determination of the conductance of the ring only (and eliminates effects in the leads). In a mesoscopic device, a four-terminal measurement remains sensitive to details in the leads which are within the phase coherent area, because they affect the wave pattern in the sample. In our example, the above-mentioned symmetry is not valid and replaced by a more complicated relationship[5]. Data collected with the four-terminal AB ring indeed show that the symmetry from (19) is absent and the phase shift α has arbitrary values.

The same symmetry rules hold for the universal conductance fluctuations. For the two-terminal measurements in Fig. 7 it can be seen that the fluctuations of all data sets are symmetric with respect to magnetic field inversion. Again, in a four-terminal measurement they are not symmetric.

5 Interaction Effects in Nanobridges

For the interference effects described so far, electrons are considered as waves and the electrostatic interaction between the charge carriers can be neglected. In a (macroscopic) metal, the electrons are indeed only weakly interacting due to screening by the surrounding charge carriers. This very efficient suppression of the long range Coulomb interaction leads to a physical picture of a metal at low temperatures in which the electron system is consisting of quasi independent, only weakly interacting *quasiparticles*, which are excitations of the Fermi liquid ground state [1,19]. For a perfect crystalline material, the eigenfunctions are k states (plane waves), which rearrange dynamically to screen any charge perturbation. This can, for example, be calculated on the textbook level by the Lindhard approximation, which yields a time- and space dependent dielectric function $\epsilon(\boldsymbol{k}, \omega)$.

When impurities and scatterers are present in the metal and the electronic motion is diffusive, the $\boldsymbol{k}$ states are no longer eigenstates of the system. It turns out that many properties of a metal remain unaffected by this change, but at very low temperatures a rather drastic change is induced by the diffusive nature of electron motion. The reason is that diffusion slows down the electrons on long distances considerably. This is formally described by a propagator which is called

[4] These asymmetries presumably arise from experimental imperfections

[5] For a four terminal device, the symmetry relation $G_{mn,kl}(B) = G_{kl,mn}(-B)$ is valid, where the permutation of indices reflects a permutation of the current leads m, n and voltage leads k, l[18,16]. Of course, the two-terminal relationship (19) is just a special case of this more general rule. The simplification in the two-terminal case is due to the fact that the tensorial character of the conductance has no importance, whereas it matters in the four-terminal case.

the Diffuson

$$\mathcal{D}(q,\omega) = \frac{1}{(-i\omega + Dq^2)\tau} \tag{20}$$

with q, ω being the difference in the momenta and frequencies of two interacting particles, respectively. $\mathcal{D}(q,\omega)$ diverges for $q,\omega \to 0$ and enters in a perturbative calculation of the density of states (DOS) and the conductance, even for macroscopic samples, causing temperature- and voltage dependent corrections to these observables.

As a consequence of the diffusive motion, two particles with similar momentum will follow very similar trajectories and will consequently stay much longer nearby than plane waves, resulting in an enhancement of interaction. A second consequence is the reduced efficiency of screening. This can easily be understood: Imagine a charge suddenly appearing in the metal. The surrounding electrons have to rearrange until the probe charge is screened. For crystalline systems, the screening by the conduction electrons is fast. For a diffusive system, the electrons are hindered by the slower diffusive motion and can rearrange only with a retardation around the probe charge. Hence, the dynamical screening is less efficient. As a consequence, the interaction is increased compared to the case of a perfect crystal.

These two mechanisms increase the interaction in a diffusive metal. Let us consider an elementary excitation from the usually assumed metallic ground state: the generation of an electron hole pair. Whereas in a perfect metal, due to screening, such an exciton experiences a very weak attractive interaction, in a diffusive conductor it is stabilized by the Coulomb interaction between the negative charge of the electron-like and the positive of the hole-like quasiparticle. Hence, this excited state may be even lower in energy than what we assumed to be the ground state. As a consequence the resulting ground state of the interacting electron system is a many-body state which is at least partially depleted around the Fermi edge (these states have slightly moved away from the Fermi energy in both directions because the "excited" states are more favorable).

The expression for the DOS correction has to be calculated according to the dimensionality of the sample, which is not of interest at this point. A detailed derivation is given in [20]. Here we want to focus on effects which appear in the conductance of metallic nanobridges. If they are sufficiently small[6], they are effectively zero dimensional. The influence of the diffusion correction can be observed both in the temperature dependence and in the voltage dependence of the conductance.

For the experiment, a small bridge is embedded in between two large and thick reservoir-like leads [21]. The leads are thick and bulky in order to guarantee reservoir-like properties albeit the heat generated by the current. Figure 12 shows an SEM picture of such a sample.

A first manifestation of interaction effects can be detected in the temperature dependence of the conductance. Figure 13 shows conductance data $G(V=0,T)$

[6] The sample is zero-dimensional with respect to the discussed effects, when all scales are smaller than the thermal diffusion length $L_T = \sqrt{\frac{hD}{k_B T}}$.

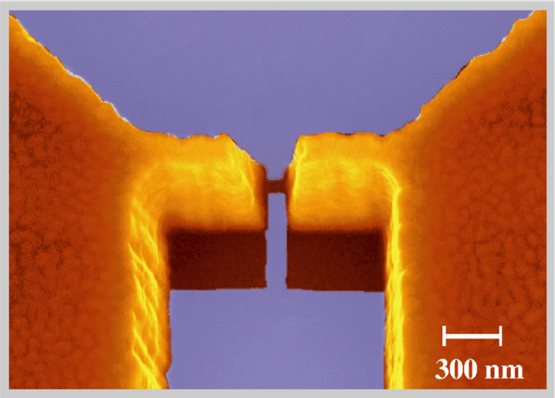


Fig. 12. SEM image of a nanobridge designed for non-equilibrium measurements. The bulky electrodes (thickness 700 nm) serve as reservoir-like leads and conduct the arising heat away from the nanobridge. The bridge itself (width and length 80 nm, thickness 10 nm) is in good contact with the leads

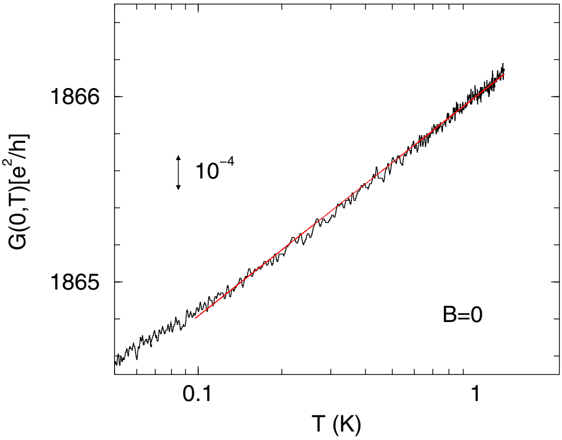


Fig. 13. Zero-bias conductance of a metallic nanobridge, showing a logarithmic dependence of G on T (cf (21))

as a function of temperature on a semi-logarithmic scale. The straight line indicates a logarithmic temperature dependence.

$$G(0,T) = G_{\text{background}} + b \cdot \ln(T/1\text{K}\ ) \tag{21}$$

remarkably, $b \approx 0.3\ e^2/h$ for many samples with different properties. In metallic films, there are several effects which are known to cause logarithmic temperature dependencies: weak localization [22], electron-electron interaction [20] and the Kondo effect [23] when magnetic impurities are present. In this case, weak localization is absent due to the small lateral size of the bridge and Kondo impurities

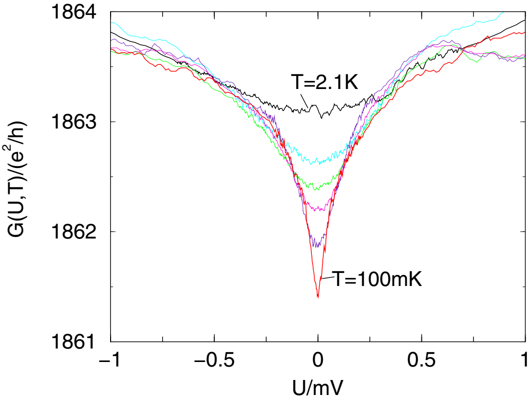


Fig. 14. Raw data of the differential conductance as a function of voltage for different temperatures. For lower temperatures, a narrow anomaly occurs, which is due to interaction effects

can be excluded because the behavior described by (21) is unaffected in strong magnetic fields.

Further information is gained when the differential conductance is measured. Figure 14 is displaying conductance data of the same sample as a function of the applied voltage, recorded at various temperatures. The anomaly is rather weakly affecting the conductance, except at rather small voltages, where a dip in $G(V,T = \text{const}) = dI/dV(V)$ appears. This reminds us of the dip in the DOS due to electron-electron interaction. It can be seen in Fig. 14 that this dip gets more pronounced (both deeper and narrower) when low temperatures are reached. Qualitatively, the voltage dependence and the temperature dependence are similar for many interaction effects, because both finite voltage and finite temperature provide excitations around the Fermi edge. As a natural choice we will look at the data as a function of eV/k_BT, because this combination plays an important role in any statistical description of the electron system.

For this purpose, we subtract the zero-bias conductance from each curve, the result is normalized by the factor b deduced from (21), and the resulting data is plotted as a function of eV/k_BT, where T is the temperature for each data set. All curves collapse onto a single curve, which is displayed in a semi-logarithmic plot in Fig. 15. Following this procedure, we can write a down a scaling law

$$\frac{G(V,T) - G(0,T)}{b} = H(eV/k_BT) \tag{22}$$

with a function H, which is universal in the sense that it appears in many different samples and does not depend on the material, the conductance or the exact geometry as long as it is zero dimensional. We can see that $H \approx 0$ for $eV < k_BT$, which reflects the fact that the voltage is not relevant when it is within the thermal smearing range. For $eV > k_BT$, H follows a straight line, which corresponds to a logarithmic voltage dependence. For the data sets at

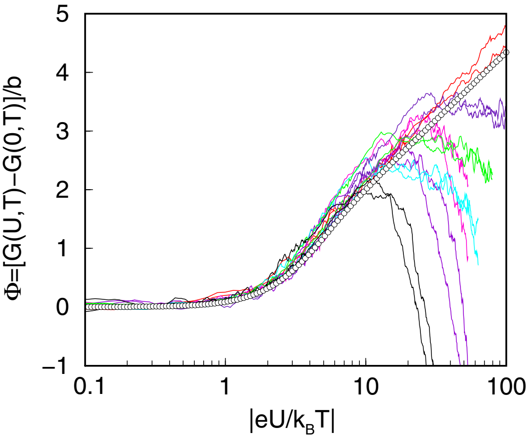


Fig. 15. Scaling plot of the data shown in Fig. 14. All data collapse onto a single curve. The circles indicate a theory which includes electron-electron interaction effects

lowest temperature, this logarithmic shape persists up to $eV/k_T \approx 100$, whereas for higher temperatures, the curves start to deviate earlier from the common scaling curve due to additional effects at higher voltages. The open circles in Fig. 15 are theoretical calculations for the diffusion correction to the conductance based on a non-equilibrium Green's functions approach [21,24] . The theory results in the same scaling law (22), a very similar value for b and the scaling curve fits excellently and without any free parameter to the experimental data, as can be seen in Fig. 15. The dip in the conductance corresponds to the diffusion correction which suppresses the DOS at the Fermi level, but also affects the propagation of charge carriers.

This behavior was also observed in metallic islands which are connected to the electric leads by high-resistance metal strips [25] and, similarly, but with a different law also in carbon-nanotube contacts [26], which have conductances much smaller than e^2/h. The common origin of these phenomena is the interaction between the charge carriers, which have to propagate across the nanojunctions while they interact with the other electrons which already occupy the nanobridge. This zero-bias anomaly is hence a precursor of Coulomb blockade, occuring at high transparency. Coulomb blockade, which appears in islands which are connected to the leads by tunnel barriers with $G \ll e^2/h$ may suppress the conductance entirely below a certain threshold voltage (see also the contribution of Jürgen Weis in this book). The interaction effects described in this section provide a correction which is small, but still present, when $G \gg e^2/h$.

Acknowledgments

It is a pleasure to thank my colleagues who contributed to this work. The Aharonov-Bohm work was mainly carried out by Ralf Häussler, the UCF work by Elke Scheer, all experiments were done in the group of Hilbert von Löhneysen and were supported by the DFG through SFB 195. I further like to thank Detlef Beckmann and Peter Pfundstein. For many theoretical discussions thanks to Ferdinand Evers and Hans Kroha.

References

1. N.W. Ashcroft, N.D. Mermin: *Solid State Physics* (Saunders College, Philadelphia 1976)
2. N. Agrait, A. Levy-Yeyati, J M. van Ruitenbeek: Phys. Rep. **377**, 86 (2003)
3. D.S. Fischer, P.A. Lee: Phys. Rev. B **23**, 6851 (1981)
4. S. Datta: *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, Cambridge 1995)
5. J. J. Sakurai: *Modern Quantum Mechanics* (Addison Wesley, 1994)
6. Y. Aharonov, D. Bohm: Phys. Rev. **115**, 485 (1959)
7. R.G. Chambers: Phys. Rev. Lett. **5**, 3 (1960)
8. R.A. Webb, S. Washburn, C.P. Umbach, R.B. Laibowitz: Phys. Rev. Lett. **54**, 2696 (1085)
9. C.P. Umbach, S. Washburn, R.B. Laibowitz, R.A. Webb: Phys. Rev. B **30**, 4048 (1984)
10. E. Scheer, H. v. Löhneysen, H. Hein: Physica B **218**, 85 (1996)
11. B.L. Altshuler, V.E. Kravtsov, I.V. Lerner: Sov. Phys. JETP Lett. **43**, 441 (1986)
12. P.A. Lee, A.D. Stone: Phys. Rev. Lett. **55**, 1622 (1985)
13. A.I. Larkin, D.E. Khmel'nitskii: Sov. Phys. JETP **64**, 1075 (1986)
14. R. Häussler, H.B. Weber, H. v. Löhneysen: J. Low. Temp. Phys. **118**, 467 (2000)
15. L. Onsager: Phys. Rev. **38**, 2265 (1931); H.B.G. Casimir: Rev. Mod. Phys. **17**, 343 (1945)
16. M. Büttiker: IBM J. Res. Develop. **32**, 3 (1988)
17. R. Häussler, E. Scheer. H.B. Weber, H. v. Löhneysen: Phys. Rev. B **64**, 085404 (2001)
18. A.D. Benoit, S. Washburn, C.P. Umbach, R.B. Laibowitz, R.A. Webb: Phys. Rev. Lett. **57**, 1765 (1986)
19. J.M. Ziman: *Principles of the theory of solids* (Cambridge University Press, London 1972)
20. B.L. Altshuler, A.G. Aronov: *Electron-Electron interactions in Disordered Systems* (North Holland, Amsterdam 1985)
21. H.B. Weber, R.Häussler, H. v. Löhneysen, J. Kroha: Phys. Rev. B. **63**, 165436 (2001)
22. G. Bergmann: Phys. Rep. **107**, 1 1984
23. L. Kouwenhoven, L. Glazman: Physics today, January 2001
24. P. Schwab, R. Raimondi: Eur. Phys. J. B **30**, 5 (2002)
25. V.A. Krupenin, A.B. Zorin, D.E. Presnov, M.M. Savvateev, J. Niemeyer: J. Appl. Phys. **90**, 2411 (2001)
26. A. Bachtold, M. De Jonge, K. Grove-Rasmussen, P.L. McEuen, M. Buitelaar, C. Schönenberger, Phys. Rev. Lett. **87**, 166801 (2001)

Ab Initio Calculations of Clusters

Florian Weigend

Forschungszentrum Karlsruhe, Institut für Nanotechnologie (INT), Postfach 3640,
76021 Karlsruhe, Germany

1 Introduction

Both *ab initio* quantum chemistry and the investigation of atomic or molecular clusters are fields of active research nowadays. This contribution presents on one hand some advances in making programs efficient for calculations of large systems, on the other it shows some fruitful combinations of calculated and experimental results. Respecting the context of the CFN summer school, it is addressed rather to non-quantum chemists, to give an insight into contemporary quantum chemical possibilities in calculations of clusters as well as to show the strengths and weaknesses of several methods. In the following chapter the very basics of quantum chemistry are discussed (as far as necessary for what follows), next some details of implementation to increase efficiency of HF, DFT and MP2 algorithms are shown. In the last section three different applications of quantum chemical methods for atomic and for molecular clusters are presented. Firstly, as an example for the HF+MP2-method, the determination of the structure and bonding situation in negatively charged clusters of water molecules is discussed, secondly a similar study of anionic gold clusters, but this time by means of density functional theory (DFT) is presented. In both cases, by combining theory and experiment results are obtained which would not have been found by calculations alone or only by experiments. The last application presented here is a DFT study on clusters of magnesium (up to Mg_{309}), focussing on the validity of the simple models of electronic and atomic shells leading to 'magic' atom and electron numbers.

2 Quantum Chemical Methods

2.1 Schrödinger Equation, Born-Oppenheimer Approximation

Quantum chemistry simply means solving the stationary Schrödinger equation

$$\hat{H}\Psi = E\Psi \tag{1}$$

for a system of many (interacting) electrons and nuclei. By using the Born-Oppenheimer approximation, i.e. by keeping the nuclei at fixed positions, this means solving (1) for many (interacting) electrons in the potential of the nuclei. As a consequence, the energy of a system shows a parametric dependence on the positions of the nuclei, defining the so-called potential hypersurface. Points

F. Weigend, Ab Initio Calculations of Clusters, Lect. Notes Phys. **658**, 205–219 (2005)
http://www.springerlink.com/

of interest usually are local minima of this surface, i.e. points, where the first derivative (with respect to all nuclear displacements) of the energy is zero (and the second derivative, the curvature, is positive). A convenient way to find local minima is to calculate energy and gradients of the energy (i.e. forces) repeatedly and to relax the positions of the nuclei due to these gradients, until a local minimum is reached.

Let us next discuss how to solve the Schrödinger equation for a system of interacting particles in presence of the external potential of the nuclei, i.e. for the electronic Hamiltonian

$$H = \underbrace{-\sum_{A}^{N}\sum_{\alpha}^{n}\frac{Z_\alpha}{|\boldsymbol{R}_A - \boldsymbol{r}_\alpha|} - \frac{1}{2}\sum_{\alpha}^{n}\nabla_\alpha^2}_{h_\alpha} + \sum_{\alpha}^{n}\sum_{\beta>\alpha}^{n}\frac{1}{|\boldsymbol{r}_\beta - \boldsymbol{r}_\alpha|}\,. \tag{2}$$

N is the number of nuclei A (charge Z_A), n the number of electrons α, $\boldsymbol{R}$ and $\boldsymbol{r}$ are the respective distances (in atomic units). The first term is the potential energy of the electrons, the second one the kinetic, and the last one describes the electron-electron interaction, which prohibits a straightforward exact solution of (1). Thus one has to look for approximate solutions.

2.2 Hartree–Fock Theory

In absence of the third term in (2) the exact solution of (1) would be a Slater determinant. Nevertheless, also in presence of the third term one can use a Slater determinant as a trial function for the wave function and get the best possible energy by the variation principle. In this way one obtains the Hartree-Fock equations

$$\underbrace{\left[h(\alpha) + \sum_i (J_i(\alpha) - K_i(\alpha))\right]}_{F(\alpha)}\varphi_i(\alpha) = \varepsilon_i(\alpha)\varphi_i(\alpha)\,. \tag{3}$$

The two-electron operator $(1/\boldsymbol{r}_{\alpha\beta})$ is replaced by the Coulomb operator J, describing the interaction of electron a with the averaged field of the $n-1$ other electrons, and by the exchange operator K, as a consequence of the Pauli principle. Together with the one-electron part h they build the Fock operator F. As F depends on the molecular orbitals φ, (3) has to be solved iteratively. The molecular orbitals $\varphi_i = |i\rangle$ are expanded in terms of atom-centred (usually Gaussian type) basis functions $|m\rangle$

$$|p\rangle = \sum_\mu c_{p\mu}|\mu\rangle\,, \tag{4}$$

which leads to the Hartree-Fock matrix equations, that can be efficiently solved by computers yielding the coefficients $\boldsymbol{c}$ and thus the Hartree-Fock wave function

Ψ_{HF}. The expectation value of the exact Hamiltonian (not that of the Fock operator) is the Hartree-Fock energy:

$$\begin{aligned} E_{HF} &= \langle \Psi_{\mathrm{HF}} | H | \Psi_{\mathrm{HF}} \rangle \\ &= \underbrace{D_{\nu\mu} h_{\nu\mu}}_{E_1} + \underbrace{\tfrac{1}{2}(\nu\mu|\kappa\lambda) D_{\nu\mu} D_{\kappa\lambda}}_{E_J} \underbrace{- \tfrac{1}{2}(\nu\kappa|\mu\lambda) D_{\nu\mu} D_{\kappa\lambda}}_{E_K} \,. \end{aligned} \tag{5}$$

The integrals $(\nu\mu|\kappa\lambda)$ describe the electron-electron interaction

$$(\nu\mu|\kappa\lambda) = \int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \nu(\boldsymbol{r}_1) \mu(\boldsymbol{r}_1) \frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \kappa(\boldsymbol{r}_2) \lambda(\boldsymbol{r}_2) \tag{6}$$

and $\boldsymbol{D}$ is the density matrix, i.e. the matrix representation of the density in the basis $|\mu\rangle$

$$D_{\nu\mu} = \sum_i n_i c_{\nu i} c_{\mu i} \,. \tag{7}$$

n_i denotes the occupation number of orbital $|i\rangle$. The Hartree-Fock method is usually applicable to main group compounds. Errors in equilibrium distances here amount to some pm, calculated IR frequencies are ca. 10% too large, which is in line with overestimated binding energies. It is not suited for use on metallic systems or for main group compounds containing extremely delocalised electrons. The computational costs for this procedure formally increase as N^4 (N is a measure for the size of the molecule), as in (6) one has four indices running over all (atom-centred) basis functions. For larger systems many integrals are insignificant, which can be used to achieve an asymptotic scaling behaviour of N^2.

2.3 Deficiencies of the HF Method, Electron Correlation

The Hartree-Fock equations were derived using the variational principle yielding a mean-field description for the electrons. The variational principle has the consequence that the HF energy is always higher than the exact energy of the system. This energy difference is called the correlation energy.

$$E_{\mathrm{corr}} = E_{\mathrm{exact}} - E_{\mathrm{HF}} \tag{8}$$

It has its origins in the mean-field character of the HF solution, as can be illustrated by the following example. Let us consider a very simple many-electron system, the He-atom containing a nucleus and two electrons. We keep one electron fixed (at a distance R to the nucleus) and look for the probability of finding the second one on a sphere of the same radius. In the Hartree-Fock picture this probability is constant all over the sphere, as the second electron “senses” only the mean field of the first electron. In fact, the probability of finding the second electron is reduced in the vicinity of the first electron by the coulomb repulsion of the two electrons, which means that the motion of the two electrons is not uncorrelated, but that they are able ‘to avoid each other’, leading to a lower total energy.

To account for this effect one can in principle, do two different things. On the one hand one can perform a HF calculation and correct the errors afterwards (e.g. by a perturbative correction) on the other one can directly modify the Hamiltonian in an appropriate way using density functional theory.

2.4 Møller–Plesset Perturbation Theory

By analogy to the correlation energy, (8), a perturbation operator V is defined as the difference between the exact Hamiltonian and the Fock operator

$$\hat{V} = \hat{H}_{\text{exact}} - \hat{H}_{\text{HF}} = \sum_{\alpha<\beta} \frac{1}{r_{\alpha\beta}} - \sum_{\alpha} \left\{ \hat{J}(\alpha) - \hat{K}(\alpha) \right\} . \qquad (9)$$

By insertion of $H = H_{\text{HF}} + V$ and $\Psi = \Psi_{\text{HF}} + \lambda\Psi_1 + ...$ in the Schrödinger equation one gets a set of equations and by expanding the first order perturbed wave function Ψ_1 in terms of excited Hartree-Fock wave functions one finally obtains the second order perturbed energy, E_{MP2} (the first order is already included in (5))

$$E_{\text{MP2}} = \sum_{iajb} t_{ij}^{ab}(ia|jb) \quad , \quad t_{ij}^{ab} = \sum_{iajb} \frac{2\,(ia|jb) - (ib|ja)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \qquad (10)$$

i and j denote molecular orbitals that are occupied in the HF wave function and a and b denote unoccupied ones. The above procedure yields excellent results (even accounting for dispersive interactions), when Hartree-Fock already provides a good description, but if HF fails, MP2 will fail, too. Note that the denominator of (10) contains energy differences of occupied and unoccupied orbitals; thus this procedure will fail for systems with a small energy difference between the highest occupied and the lowest unoccupied orbital (which is typical for metallic systems). Furthermore, for calculation of E_{MP2} the electron interaction integrals have to be transformed into the basis of the molecular orbitals, leading to an N^5 scaling behaviour.

2.5 Density Functional Theory

A computationally much more economic way to account for effects of electron correlation is given by DFT methods. To keep matters as simple as possible, just replace in (5) the Hartree-Fock exchange part, E_{K}, by an exchange-correlation energy, E_{XC}, which can be expressed as a functional of the electron density $\rho = |\Psi|^2$

$$E_{\text{XC}} = \int f(\rho(\boldsymbol{r}), \nabla\rho(\boldsymbol{r}), ...) d^3r = \underbrace{k \int \rho(\boldsymbol{r})^{\frac{4}{3}} d^3r}_{\text{Dirac}} + ... \qquad (11)$$

Unfortunately the exact form of E_{XC} is known only for the free electron gas (Dirac exchange). Much effort was and is spent on developing functionals suited for molecules; well established and widely used is that of Becke (for exchange) and Perdew (for correlation), BP86 [1,2].

DFT is an economic way to calculate electronic structures including correlation effects. As Coulomb integrals have to be evaluated as in HF, the scaling behavior is the same (as long as no 'tricks' are applied, see below). For main group compounds DFT is nearly as good as HF+MP2, for (large) transition metal compounds, and in particular for metallic clusters, there currently is no reasonable alternative to DFT.

3 Program Developments

The first calculations on metallic clusters were done in the late 80s, e.g. Li_{10} (30 electrons) or Na_9 (99 electrons). Nowadays one is able to treat systems with ca. 9000 electrons (Ga_{309}) on a single PC. If one takes into account that for large systems DFT scales as N^2, this means an increase of efficiency by a factor of ca. 10^4, which of course can partly be put down to the rapid increase in the power of computers, but also is a result of the hard work spent on making programs more efficient. As an example, the application of the RI-approximation to increase the efficiency of DFT, HF and MP2 is discussed with several additional methods mentioned at the end.

3.1 RI Approximation

The most demanding step in HF, DFT or MP2 calculations is the evaluation (and the transformation) of the electron-electron integrals (6). By using the following method the computational costs can be enormously reduced [3]. Let us expand a product of basis functions (Gaussian functions) in a series of so-called auxiliary basis functions $P(\boldsymbol{r})$(also Gaussian functions)

$$\rho_{\nu\mu}(\boldsymbol{r}) = \nu(\boldsymbol{r})\mu(\boldsymbol{r}) \approx \tilde{\rho}_{\nu\mu}(\boldsymbol{r}) = \sum_P c^P_{\nu\mu} P(\boldsymbol{r})\,. \tag{12}$$

To determine the expansion coefficients we require the Coulomb self-interaction of the error of the expansion to become minimal:

$$\int \{\tilde{\rho}_{\nu\mu}(\boldsymbol{r}_1) - \rho_{\nu\mu}(\boldsymbol{r}_1)\} \frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \{\tilde{\rho}_{\nu\mu}(\boldsymbol{r}_2) - \rho_{\nu\mu}(\boldsymbol{r}_2)\}\, d\boldsymbol{r}_1 d\boldsymbol{r}_2 \;=\; \min\;. \tag{13}$$

This leads to

$$c^P_{\nu\mu} = \sum_Q (\nu\mu|Q)(Q|P)^{-1} \tag{14}$$

and finally to

$$(\nu\mu|\kappa\lambda) \approx (\nu\mu|\kappa\lambda)_{\mathrm{RI}} = \sum_P c^P_{\nu\mu}(P|\kappa\lambda) = \sum_{QP} (\nu\mu|Q)(Q|P)^{-1}(P|\kappa\lambda)\,, \tag{15}$$

which is formally similar to inserting the resolution of the identity (RI).

This can be used for a more efficient calculation of the Coulomb part of the Fock matrix [4,5]. In a conventional algorithm the evaluation of the Coulomb matrix

$$J_{\nu\mu} = \sum_{\kappa\lambda} (\nu\mu|\kappa\lambda) D_{\kappa\lambda} \tag{16}$$

requires an N^4 step for the calculation of the integrals and one more N^4 step for the contraction with the density matrix. Using the approximated integrals instead of the exact ones, gives

$$J^{\mathrm{RI}}_{\nu\mu} = \sum_{\kappa\lambda} \sum_{P,Q} \underbrace{(\nu\mu|Q)\, \underbrace{(Q|P)^{-1}\, \underbrace{\underbrace{(P|\kappa\lambda)}_{N^2 N_x}\, D_{\kappa\lambda}}_{N^2 N_x}}_{N_x^2}}_{N^2 N_x} \; . \tag{17}$$

When doing the operations from the right to the left, as indicated by the brackets, obviously the most expensive step, the calculation of N^4 integrals, is replaced by steps that scale as N^3 or N^2. We note that accuracy and efficiency of RI methods clearly depend on the quality of the auxiliary basis sets. If one uses optimised auxiliary basis sets, errors are ca. 1 order of magnitude smaller than that of basis set changes and efficiency is increased by a factor of ca. 10.

In connection with MP2 the RI approximation can be used to reduce the computing costs of integral evaluations and transformations [6,7]:

$$(ia|jb)_{\mathrm{RI}} = \sum_{\substack{PQR \\ \nu\mu\kappa\lambda}} \underbrace{c_{i\nu} c_{a\mu} (\nu\mu|Q)(Q|P)^{-\frac{1}{2}}}_{B^P_{ia}} \underbrace{(P|R)^{-\frac{1}{2}} (R|\kappa\lambda) c_{\kappa j} c_{\lambda b}}_{B^P_{jb}} \; . \tag{18}$$

If one first calculates quantities $\boldsymbol{B}$, and then performs the step $\boldsymbol{B} \times \boldsymbol{B}$, the 4-index integrals are replaced by 3-index integrals and the exponent of the scaling behavior for both calculation and transformation of integrals is reduced by one. The remaining multiplication $\boldsymbol{B} \times \boldsymbol{B}$ still scales as N^5, but as it is a matrix multiplication step it can be done with great efficiency. We note, in passing, that the RI approximation can be applied to the HF exchange part in a very similar way to that for MP2 [8]. In both cases the efficiency increases with the size of the atomic basis set.

3.2 Other Developments Useful for Large Systems

In DFT calculations of large systems the dominating step remains the evaluation of the Coulomb part even when the RI approximation is used. Further reduction in computing costs for this part is achieved if interactions of distant electrons are described by multipole expansions. In this way the number of integrals to be calculated is drastically reduced for large systems [9].

A problem which often occurs in calculations of metal clusters is that the occupation of orbitals chosen at the beginning of a calculation is usually not correct at the end, i.e. does not fulfil the *aufbau* principle. A convenient way around this is to adjust the occupation during iterations. This is best done by calculating orbital occupations every iteration by a Gaussian error function from the orbital energies [10]. This leads to fractional occupation numbers in the vicinity of the HOMO-LUMO gap. The broadness of the energy region with fractional occupation numbers is determined by a parameter which can automatically be decreased during SCF procedure so that a non-fractional occupation that fulfils the *aufbau* principle results.

Finally mention should be made of some previous developments which are nevertheless important for efficient calculations of clusters: optimised grids for efficient numerical evaluation of E_{XC} [11]; exploitation of symmetry for all point groups [12]; direct methods for calculation of integrals (i.e. no storage of integrals on disc) together with integral screening (neglect of insignificant integral batches) [13].

4 Applications

4.1 The Water Hexamer Anion

As an example for the use of the HF+MP2 method as well as an example for successful combination of theory and experiment we present a study of anionic water clusters [14]. Small closed-shell molecules are usually not able to bind an additional electron unless they have a large dipole moment. Experiments indicate that a cluster of two water molecules is just able to bind an electron. In these experiments, water is heated in a chamber with a hole allowing steam to escape through this hole. Just at the moment when the molecules leave the chamber electrons are added to this jet. The water molecules in this jet form clusters that are analysed by mass spectrometry and by photoelectron detachment spectrometry yielding information about the number of molecules in a cluster and the vertical detachment energy (VDE), i.e. the binding energy of the additional electron. No information about the geometry of these clusters can be retrieved from experiments and, in particular, how the additional electron is bound, i.e. whether it is dipole-bound, or rather located at the surface or inside the cluster. Experiments show large intensities for a six-membered cluster with a VDE of ca. 0.45 eV. The motivation for these experiments was to obtain information about solvated charged particles in water ('solvated electrons'). Quantum chemical calculations aimed at finding the geometric structure and the binding mechanism were performed with the following strategy:

1. Consider many isomers of $(H_2O)_6^-$.
2. Optimise geometrical structure of all isomers.
3. Calculate the energy of the neutral species in the geometry of the anionic one to get the VDE as energy difference for each isomer.
4. Most promising candidates are those which fit VDE and lowest total energy.

MP2 is known to yield a good description of hydrogen bonds (compared to HF or DFT). Thus the HF+(RI-)MP2 method was chosen for this problem.

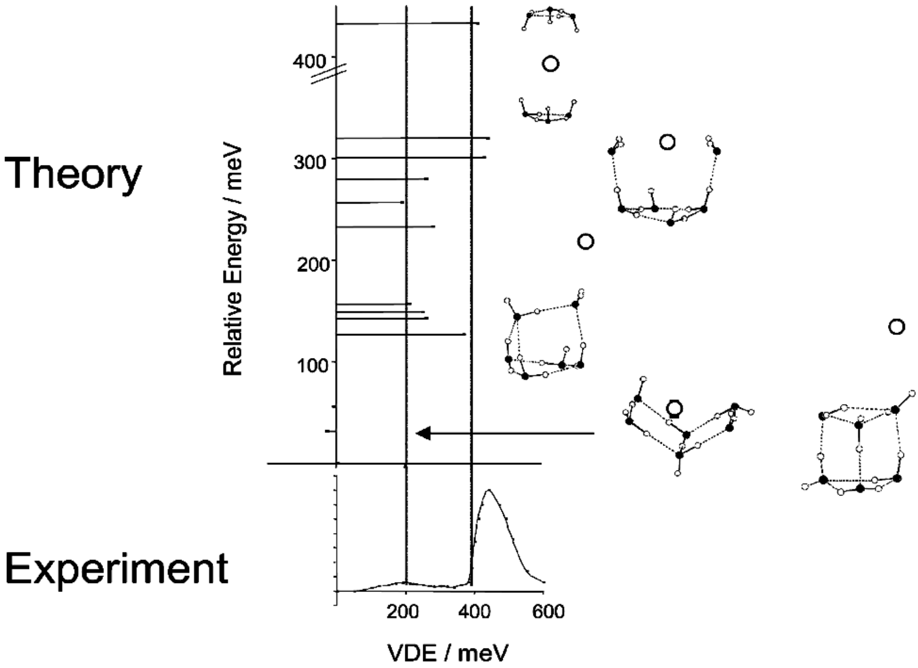


Fig. 1. The water hexamer anion. Lower panel: experimental VDE spectrum, upper panel: results of MP2 calculations. Hydrogen bonds are drawn as dashed lines, the position of the floating center is indicated by an open circle. The different types of isomers can be divided into four groups (electron inside the cluster, electron on the surface, electron outside, minima of neutral clusters). For each group at least one representative isomer is shown

As mentioned previously, molecular orbitals are expanded by atom-centered basis functions and this has serious consequences on the adequate description of dipole-bound states: the dipole bound electron is rather far away from any of the atomic centers. Thus, at the location of the electron no basis functions are provided for the description of electronic density leading to an unrealistic disfavoring of these kinds of states, compared to those where the additional electron is close to one of the atoms. Matters can be much improved if one provides basis functions that are not centered on one of the atoms but somewhere around the molecule ('floating bases'). In our calculations the position of the floating bases was optimised simultaneously with the position of the atoms; in this way one also obtains information about the location of the electron. After assessing the accuracy of the method at the water dimer anion (also in calculations we found the additional electron to be bound very weakly), the work of Lee et al. [15] was used as source for reasonable isomers (these authors mainly used DFT methods without floating bases and got less reliable results due to the above-mentioned reasons). The results of our calculations together with the experimental VDE spectrum are shown in Fig. 1. One observes different total energies, i.e. different stabilities for the different types of binding. Least stable is the cluster with the electron inside. Those binding the electron at the cluster surface are of higher stability, and even more stable are the ones with a dipole-bound electron. Of highest stability are two systems which are believed to be lowest minima for neutral $(H_2O)_6$ clusters. From the point of stability these would be the most promising candidates, but now the second criterion, the binding energy of the additional electron, has to be investigated. It turns out that these two isomers show negative VDEs, i.e. these clusters are not able to bind the additional electron. This indicates that the $(H_2O)_6^-$ cluster is only metastable and it explains why in the experiment one has to add the electrons before the clusters are formed from single molecules. As expected, the dipole-binding clusters, which form the next stable group of isomers, all have clearly positive VDEs. The VDE of the most stable isomer in this group, ∼0.4 eV, agrees with the experimental value quite well. Concerning the VDE, the surface-bound states also show matching VDEs, but they are energetically clearly disadvantaged. Thus we conclude that the additional electron in $(H_2O)_6^-$ is bound by the dipole moment of the cluster.

4.2 Small Gold Cluster Anions

Another example for combining theory and experiments is now discussed. Small gold clusters were investigated in a way which is in principle the same as that for the water clusters, but differs in some important details [16]. Firstly, we are now considering metallic systems, which usually cannot be treated accurately with HF+MP2 (neither with other HF + post-HF techniques), on one hand because correlation effects become too large (at least for larger systems), on the other simply because of high computing costs. Currently the only feasible way for treating (large) metal clusters is density functional theory. Secondly, the VDE is a less helpful criterion to distinguish different isomers, as for metallic systems additional charge is always located at the surface leading to similar VDEs for most of them. Besides energy, the cross section appeared to be an appropriate criterion to distinguish isomers: it depends on the cluster shape and it can be measured from drift cell experiments as well as calculated from quantum chemically optimised geometrical structures. In the experiment the cross section is measured by measuring the mobility of (mass-selected) clusters in a cell filled with helium; due to scattering at the He atoms, the time to pass through the cell increases linearly with the cross section of the clusters. The cross section can also be calculated from the quantum chemically determined geometric structure, e.g. by hard sphere scattering or more sophisticated methods. Experimentally and theoretically determined cross sections (i.e. cross sections divided by the value of a fit function originally obtained for cations) for Au_n^- clusters, $n < 14$ are shown in Fig. 2. Considering the experimental data, it is most striking that for $n > 12$ relative cross sections are smaller than for $n < 12$. This indicates a principal change of the cluster shape at $n = 12$ resulting in two different cross sections for $n = 12$, corresponding to two different isomers. The DFT calculations reveal the following picture: in all cases investigated (ca. four for each cluster size) the most stable isomer was planar. For $n < 12$ the few stable three-dimensional

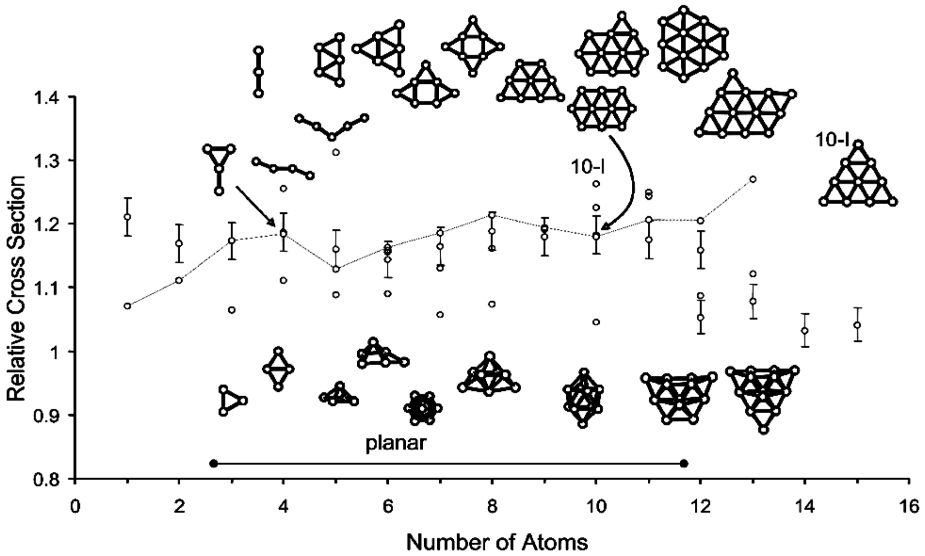


Fig. 2. Calculated and measured relative cross sections of gold cluster anions obtained by dividing both by the fit function $\Omega(n) = \frac{4}{3}\pi^{\frac{1}{3}}(r_{\mathrm{Au}} + r_{\mathrm{He}})^2$ with $r_{\mathrm{Au}} = 1.47\,\text{Å}$ and $r_{\mathrm{He}} = 1.15\,\text{Å}$. This plot projects out the increase of the cross section due to the increasing cluster size and thus illustrates the effect of different cluster shapes. The circles with error bars represent the experimental data, the open circles represent cross sections of candidates from DFT calculations. The line connects the 'best' candidates (based on stability and cross section). In most cases these are lowest in energy; exceptions are Au_4^-, Au_{10}^- and Au_{13}^- where only candidates that are slightly higher in energy are in line with the experiment

isomers are higher in energy by usually more than 1 eV. Also for $n = 12, 13$ the three-dimensional isomers are higher in energy, but only by 0.6 (0.25) eV. By combining the results, we may conclude, that up to $n = 12$ negatively charged gold clusters are planar, larger systems prefer three-dimensional packing. We further note that DFT slightly overestimates the stability of planar systems. The theoretically and experimentally observed planarity for comparably large systems is surprising, since one would expect, that for the usually non-directed metal-metal bonds space filling shapes would be preferred to achieve a high number of next neighbors, as is indeed usually observed for metals (see below). Obviously Au is an exception, probably due to relativistic effects, yielding a hybridisation of the 6s and the $5\mathrm{d}_{\mathrm{z}^2}$ orbital which leads to a preference for low coordination numbers.

4.3 A Theoretical Study on Clusters of Magnesium

Metal atoms usually tend to build space-filling systems as it is preferable in case of non-directed bonds to maximize the number of next neighbors and thus the number of bonds. The latter also implies the preference of closed polyhedra, like

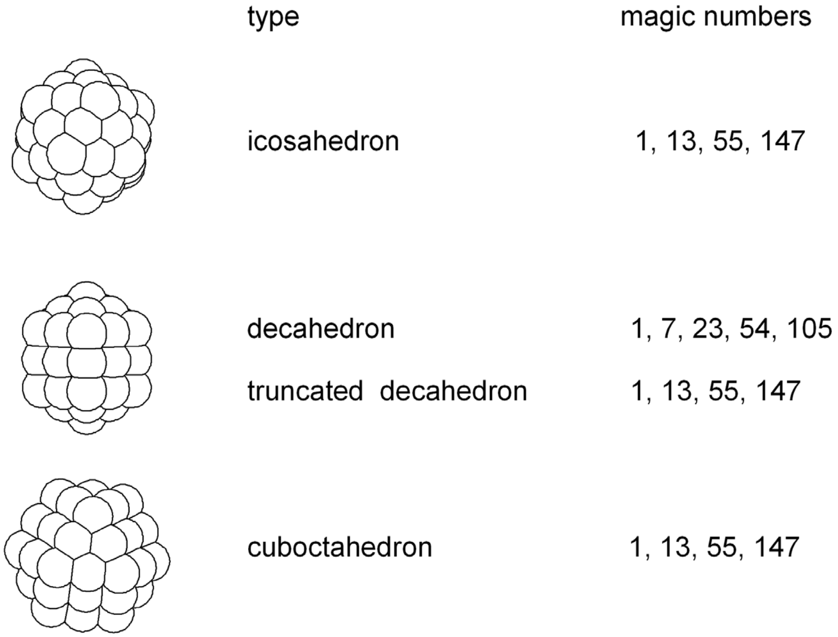


Fig. 3. Selected polyhedra and corresponding magic numbers

(cube-)octahedra, icosahedra, decahedra, etc., which leads to so-called 'magic' numbers of atoms; selected cases are shown in Fig. 3. One can also introduce "magic numbers" for the electrons as rationalised in the Jellium model. One assumes that the clusters are spherical and that effects of the positively charged cores can be replaced by a uniform positively charged background. This leads to a model of free electrons in the potential of a harmonic oscillator (in the simplest case). This results in a highly degenerate shell structure of spherical harmonics (1s)(1p)(2s1d)(2p1f)(3s2d1g)... , leading to magic electron numbers (whenever a shell is closed) of $2, 8, 20, 40, 70, 112, 168, 240, 330\ldots$. We investigated the stability of Mg clusters, in particular the occurrence of magic numbers and the validity of the electronic shell model [17].

To begin with one has to assess the accuracy of DFT for the problem. For this purpose we compared DFT and (the more accurate) CCSD(T) results for tetrahedral Mg_4. Indeed, the dissociation energy obtained with both methods was very similar (1.22 eV vs. 1.14 eV) as well as the equilibrium distance (309.4 pm vs. 310.3 pm), at least, if with the careful choice of an appropriate basis: usually basis sets for Mg are optimised for 'chemical' compounds, where Mg is bound rather ionically and thus partially oxidised. This leads to basis set exponents that are too steep for an accurate description for Mg-Mg bonds. Thus we extended the basis by a diffuse p and a diffuse d set and re-optimised all valence and polarisation functions at Mg_4, which led to the above-mentioned results.

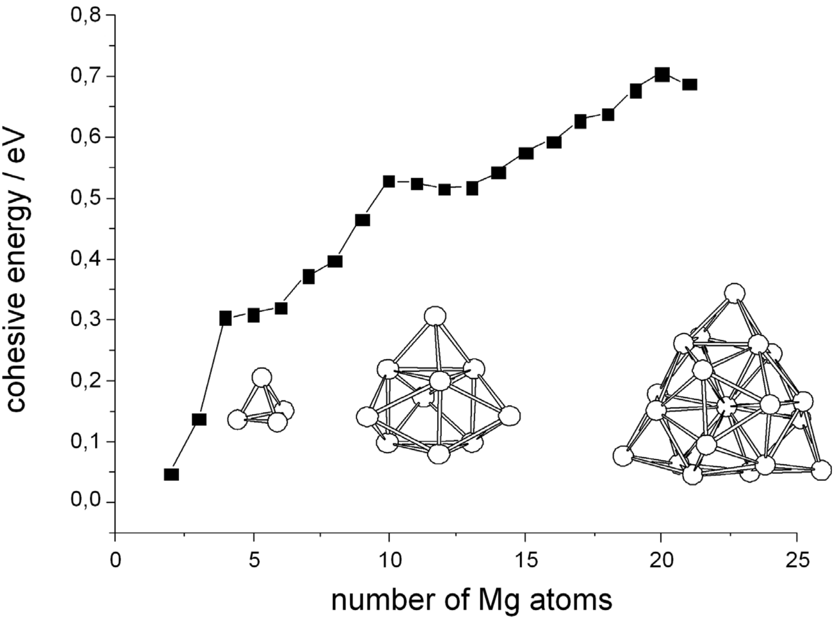


Fig. 4. Calculated cohesive energies for Mg_n ($n < 23$) and most stable isomers for $n = 4, 10, 20$

Firstly, we investigated the stability of small clusters (up to 22 atoms). For several clusters simulated annealing techniques were applied to get new structures. Higher and lower aggregates were created by removing or adding atoms. Especially around magic numbers (Mg_n, n=10,20) a variety of structures was investigated. Cohesive energies

$$\varepsilon_{\mathrm{coh}} = -\left[E(\mathrm{Mg}_n) - nE(\mathrm{Mg})\right]/n \tag{19}$$

are shown in Fig. 4. It is evident, that clusters of pronounced stability are those with 4, 10 and 20 Mg atoms corresponding to magic electron numbers of 8, 20 and 40 as well as to the magic atom numbers of tetrahedral systems (not shown in Fig. 3. Thus it is not too surprising that Mg_4 and Mg_{20} have tetrahedral topologies and Mg_{10} is at least related to a tetrahedron.

Next, the stability of larger clusters was investigated. The cohesive energy may be approximated as

$$\varepsilon_{\mathrm{coh}} \approx \varepsilon_{\mathrm{coh,bulk}} + a_{\mathrm{surface}} n^{-\frac{1}{3}} + a_{\mathrm{edge}} n^{-\frac{2}{3}} + a_{\mathrm{corner}} n^{-1} \,. \tag{20}$$

For such large systems a systematic search for local minima is currently not feasible for us. We restricted our investigations to selected symmetric closed polyhedral structures. A plot of the cohesive energy versus $n^{-\frac{1}{3}}$ for all calculated clusters is shown in Fig. 5. No clear picture arises from these data, but we can definitely state that up to 309 atoms the hcp packing of the solid state is disfavoured towards icosahedral packing, which is in line with experimental

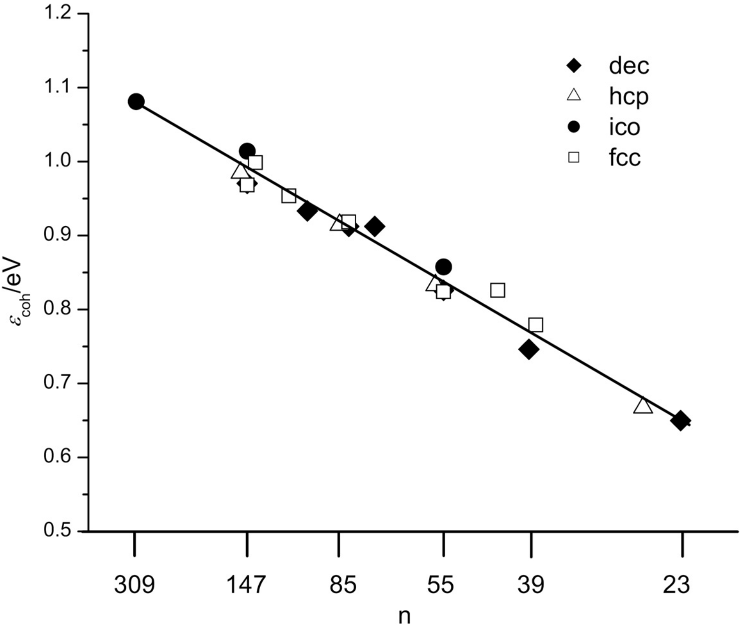


Fig. 5. Calculated cohesive energies for selected clusters versus $n^{-\frac{1}{3}}$. Structures derived from octahedral are denoted 'oct', those derived from decahedra 'dec'. The line corresponds to a linear regression including all structure types

findings [18]. The cohesive energy for the bulk evaluated from these data with the help of (19) leads to a value of 1.38 eV, reasonably close to the experimental one (1.51 eV).

Finally we comment on the validity of the shell model of electrons. According to the harmonic model one would expect an energetic structure of shells with $1, 3, 6, 10, 15...$ orbitals leading to magic electron numbers of $2, 8, 20, 40, 70...$. Investigating the density of states arising from the valence orbitals of ico55, cuboct55, trdec55 and hcp57, Fig. 6, one obtains the following picture. In all cases the first three shells are clearly separated from each other and also from the higher valence orbitals, yielding the magic numbers 2, 8 and 20. The next magic number, 40, is visible only for hcp57 and ico55, but all clusters show a zero density when a total number of 34 electrons is reached. Furthermore for ico55 one recognizes a zero density of states above 58 and 92 electrons. These numbers can be rationalized by closing of subshells that occur from distortions of the spherical shape of the clusters. In all cases the subshells with highest l quantum number are lowest in energy, which leads directly to the magic numbers 18 (1s+1p+1d), 34 (1s+1p+1d2s+1f), 58 (1s+2p+2s1d+1f2p+1g) and 92 (1s+2p+2s1d+1f2p+1g2d3s+1h). For the other clusters, where the deviation from a sphere is larger, only some of these numbers can be observed.

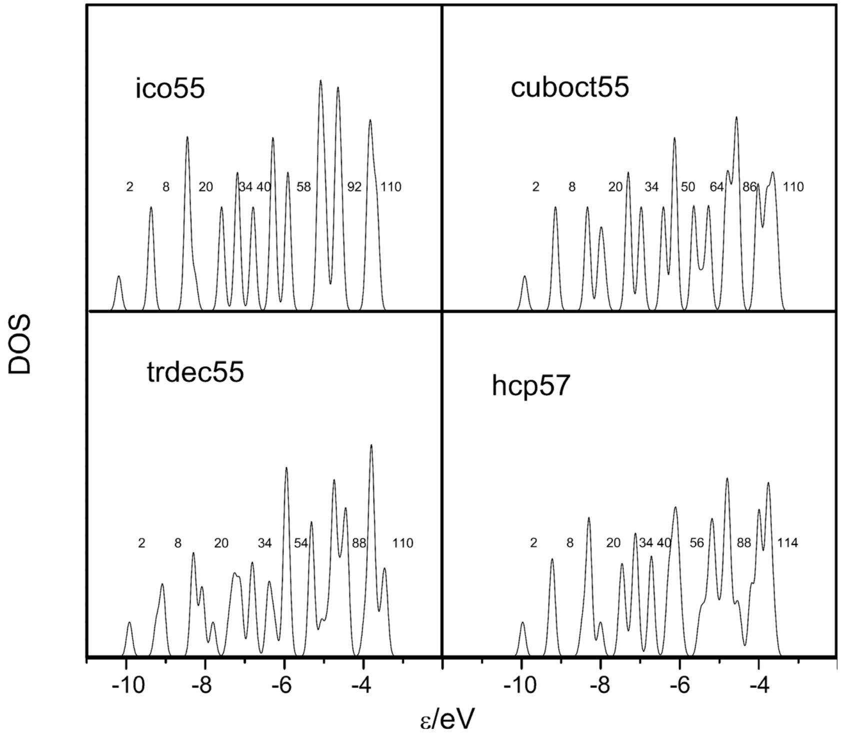


Fig. 6. Calculated valence density of states (arbitrary units) vs. orbital eigenvalues in eV for different cluster shapes of Mg_{55} (Mg_{57} for hcp). The calculated discrete levels were broadened by Gaussians of 0.1 eV FWHM to help the eye. The numbers denote the sum over valence electrons up to a given energy

5 Summary

In this contribution calculations on water clusters with HF+MP2 methods and DFT studies on metal clusters have been presented. The examples selected show fruitful combinations of theoretical and experimental data, as well as the test of simple models by quantum chemical calculations. The work on program developments necessary to make algorithms efficient for calculation of these large systems was discussed using the example of the RI approximation.

References

1. A.D. Becke: J. Chem. Phys. **98**, 5648 (1993)
2. J.P. Perdew: Phys. Rev. B **33**, 8822 (1986)
3. O. Vahtras, J. Almlöf, M.W. Ferereisen: Chem. Phys. Lett. **213**, 514 (1993)
4. K. Eichkorn, O. Treutler, H. Öhm, M. Häser, R. Ahlrichs: Chem. Phys. Lett. **240**, 283 (1995)
5. K. Eichkorn, F. Weigend, O. Treutler, R. Ahlrichs: Theor. Chem. Acc. **97**, 119 (1997)

6. F. Weigend, M. Häser: Theor. Chem. Act. **97**, 331 (1997)
7. F. Weigend, M. Häser, H. Patzelt, R. Ahlrichs: Chem. Phys. Lett. **297**, 143 (1998)
8. F. Weigend: Phys. Chem. Chem. Phys. **4**, 4285 (2002)
9. M. Sierka, A. Hogekamp, R. Ahlrichs: J. Chem. Phys. **118**, 9163 (2003)
10. P. Nava, M. Sierka, R. Ahlrichs: Phys. Chem. Chem. Phys. **5**, 4036 (2003)
11. O. Treutler, R. Ahlrichs: J. Chem. Phys. **102**, 346 (1995)
12. M. Häser: J. Chem. Phys. **95**, 8259 (1991)
13. M. Häser, R. Ahlrichs: J. Comput. Chem. **10**, 104 (1989)
14. F. Weigend, R. Ahlrichs: Phys. Chem. Chem. Phys. **1**, 4537 (1999)
15. S. Lee, J. Kim, S.J. Lee, K.S. Kim: Phys. Rev. Lett. **79**, 2308 (1997)
16. F. Furche, R. Ahlrichs, P. Weis, C. Jakob, S. Gilb, T. Birweiler, M. Kappes: J. Chem. Phys. **117**, 6982 (2002)
17. A. Köhn, F. Weigend, R. Ahlrichs: Phys. Chem. Chem. Phys. **3**, 711 (2001)
18. T.P. Martin, T. Bergmann, H. Göhlich, T. Lange: Chem. Phys. Lett. **176**, 343 (1991)

Nanostructured Materials: Reaction Kinetics and Stability

John H. Perepezko

University of Wisconsin-Madison, Department of Materials Science and Engineering, 1509 University Ave., Madison, WI 53706, USA, perepezk@engr.wisc.edu

Abstract. An important consequence of the expanding study of the nanocrystalline state is the recognition of new behavior that is exposed at the nanometer length scale, but this also requires the recognition of the scaling of conventional behavior. The synthesis pathways further emphasize the importance of reaction kinetics and especially nucleation processes where the nanometer length scale is central to the kinetics. Similarly, the observed phase selection during nanostructure synthesis is often different than that expected from the thermodynamics of bulk phase stability, but can be analyzed in terms of a scaling of the hierarchy of equilibrium and the influence of large characteristic driving free energies. At the same time, the reaction pathways that yield different phase states and microstructures can be described in terms of open or closed system conditions that reflect the manner in which the excess free energy is developed during synthesis. The principles that govern the genesis of nanostructured materials and the key issues concerning the reaction kinetics and stability are illustrated from the observed behavior in specific amorphous alloys, but the treatment also applies in general to materials systems.

1 Introduction

One of the highlights of the contemporary attention directed towards nanocrystalline materials is the major innovation in processing methodologies that have been developed to achieve nanostructured materials. The nanocrystalline state, where the microstructural size scales are in the 1 to 100 nm range, can be synthesized by a variety of processing routes starting with the vapor, liquid or solid state [1]. The earliest efforts were directed principally to attain laboratory scale quantities by a vapor condensation path. Subsequent efforts have yielded a large variety of methods based upon strategies involving vapor, liquid and solid state processing [2,3]. Within this large and growing menu of options it is possible to characterize the methods into two broad categories based upon open or closed systems that relate to the manner in which the driving free energy that motivates structure synthesis and change is developed during processing. An important consequence of the expanding study of the nanocrystalline state is the recognition of different materials behavior that can be exposed at the nanometer length scale [1].

A useful concept for the study of non-equilibrium processes that yield nanostructured materials is the distinction between closed and open system processes [4]. In a closed system, an energized state is achieved through a rapid temperature, pressure or composition change to create a certain level of undercooling

J.H. Perepezko, Nanostructured Materials: Reaction Kinetics and Stability, Lect. Notes Phys. **658**, 221–249 (2005)
http://www.springerlink.com/

or supersaturation (i.e. a metastable state), which then releases the excess free energy during relaxation towards equilibrium. With an open system, the energized state is often attained by a continuous incremental excess energy input to an initial state through the incorporation of excess lattice defects or solute on a localized spatial scale and time interval that is short compared to the relaxation time so that the relaxation process is influenced [5–7].

Although there can be synthesis and processing history related effects that influence the behavior and properties, the present discussion will focus on some of the key thermodynamic and kinetics issues that are relevant to nanostructured materials. The nanostructure can be expressed as isolated volumes within a bulk volume or as aggregates of nanoparticles. Regardless of the aggregation condition or synthesis method there are common themes in the thermodynamic stability and reaction kinetics. The fundamental issues include the phase selection during the initial synthesis and the phase stability after synthesis and during subsequent treatment.

In order to treat the basic issues in phase stability within the limited coverage that is available, some of the key points in the scaling of the free energy to the nanoscale size are examined by considering selected thermodynamic relations for macroscopic systems and their modification for high interfacial area nanoscale systems. For certain alloy solutions that exhibit phase separation, the scaling from macroscale to nanoscale has a form that is influenced by the limitation on the spatial extent of diffusional phase decomposition. Similarly, the basic kinetic reactions are examined from the point of view of crystallization, but these points are general in terms of application to many aspects of synthesis reactions as well as the understanding of the properties and behavior of nanostructures. As a means of illustrating the use of the relevant thermodynamic and kinetics concepts to the analysis of nanostructure development, an example is considered for nanocrystallization of amorphous alloys.

2 Phase Stability Thermodynamics

2.1 Pure Components

In the analysis of relative phase stability and the driving free energies for different phase transformation reactions the Gibbs free energy, G, is the main function of interest. By definition, $G = H - TS$ where H is the enthalpy, T is the temperature and S is the entropy [8,9]. For reactions between two phases, the free energy change is the difference, $\Delta G = \Delta H - T\Delta S$. For example, during solidification of a liquid, $G_s - G_l = (H_s - H_l) - T(S_s - S_l)$ or $\Delta G_f = \Delta H_f - T\Delta S_f$. At equilibrium, $\Delta G_f = 0$ so that $\Delta S_f = \Delta H_f / T_m$. If ΔS_f is taken as independent of T (i.e. neglecting the small heat capacity correction), then $\Delta G_f = \Delta H_f(1 - T/T_m) = \Delta H_f(T_m - T)/T_m = \Delta H_f \Delta T / T_m$ so that the driving free energy (i.e. driving force) for the reaction is proportional to the undercooling, ΔT. The dependence of the molar free energy on temperature is illustrated in Fig. 1 for a liquid, a stable α phase and a metastable β phase where it is evident that the formation

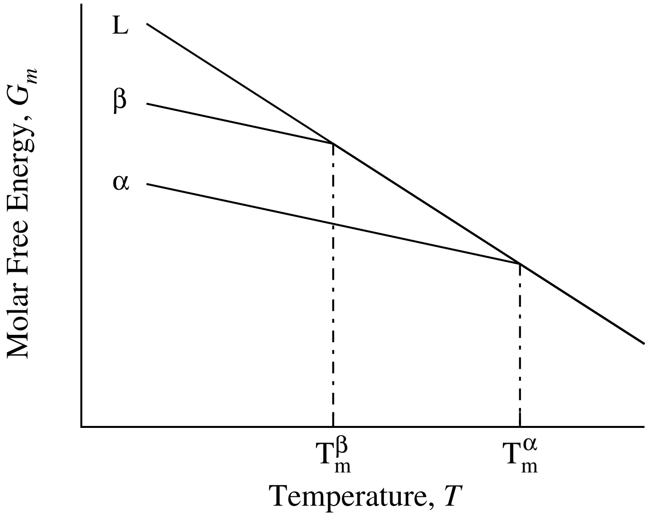


Fig. 1. Free Energy as a function of temperature for a single component. The melting point of the stable α phase T_m^α and the metastable β phase T_m^β is indicated.

of the metastable β phase requires a minimum liquid undercooling below the stable α phase melting point of $\Delta T = T_m^\alpha - T_m^\beta$.

From the second law of thermodynamics several differential relations can be developed (i.e. Maxwell relations)[8]. The relation for the Gibbs free energy for a single phase is

$$dG = VdP - SdT \tag{1}$$

where V is the volume and P is the pressure. For the liquid-solid case at equilibrium

$$V_s dP - S_s dT = V_l dP - S_s dT \tag{2}$$

or

$$\left(\frac{dP}{dT}\right)_{eq} = \frac{S_s - S_l}{V_s - V_l} = \frac{\Delta S_f}{\Delta V_f} = \frac{\Delta H_f}{\Delta V_f T_m} \tag{3}$$

which is the Clapeyron equation for the pressure dependence of the melting point. When (3) is applied to other two-phase coexistence such the solid-vapor and liquid-vapor equilibrium, the pressure -temperature phase diagram for a single component is developed as shown in Fig. 2. The solid lines represent the two phase coexistence conditions where there is one degree of freedom according to the Gibbs phase rule (i.e. $P + F = C + 2$ where P = the number of phases, F = the number of degrees of freedom and C = the number of components). The lines intersect at an invariant triple point where $F = 0$. It is useful to note that $(\partial G/\partial T)_P = -S$ and $(\partial G/\partial P)_T = V$.

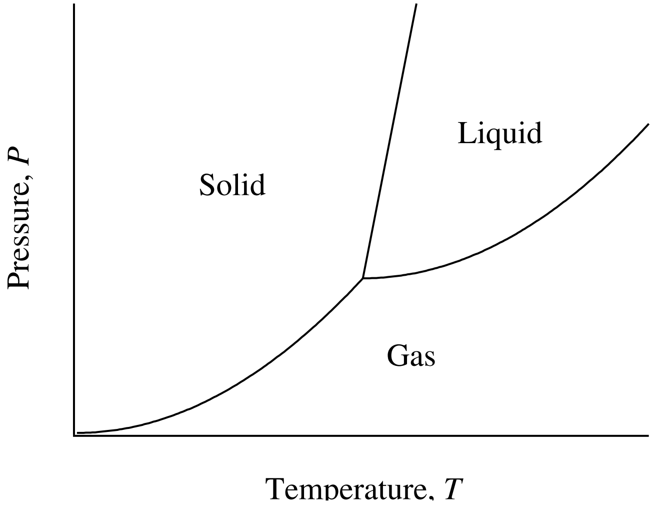


Fig. 2. A schematic pressure vs. temperature diagram for a single component.

2.2 Alloy Solutions

For alloy solutions, the chemical terms for the free energy are included by adding the partial molar free energy for each component, such as i or j for the binary case to (1) as

$$dG = VdP - SdT + dn_i \left(\frac{\partial G}{\partial n_i} \right)_{T,P,n_j} + dn_j \left(\frac{\partial G}{\partial n_j} \right)_{T,P,n_i} \tag{4}$$

The partial molar free energy represents the chemical potential, μ_i, that is defined as

$$\left(\frac{\partial G}{\partial n_i} \right)_{T,P,n_j} = \mu_i = \mu_i^0 + RT \ln a_i \tag{5}$$

where n is the number of moles, μ_i^0 is the standard state chemical potential, R is the gas constant and a_i is the activity that is the product of an activity coefficient, γ_i and the mole fraction, X_i (note that $X_i + X_j = 1$ and for an ideal solution $\gamma_i = 1$). For alloy formation (i.e. atomic mixing) at constant T and P for each phase with a given composition X_B in an A-B system, the molar free energy is

$$G = (1 - X_B)\mu_A + X_B \mu_B \tag{6}$$

For transformation between two solution phases such as liquid and solid

$$\Delta G_m = (1 - X_B)(\mu_A^s - \mu_A^l) + X_B(\mu_B^s - \mu_B^l) \tag{7}$$

At equilibrium, $\Delta G_m = 0$ so that $\mu_A^s = \mu_A^l$ and $\mu_B^s = \mu_B^l$. For the examination of interactions in solutions it is often useful to represent the free energy change for mixing, ΔG_m, in terms of an enthalpy of mixing, ΔH_m and a mixing

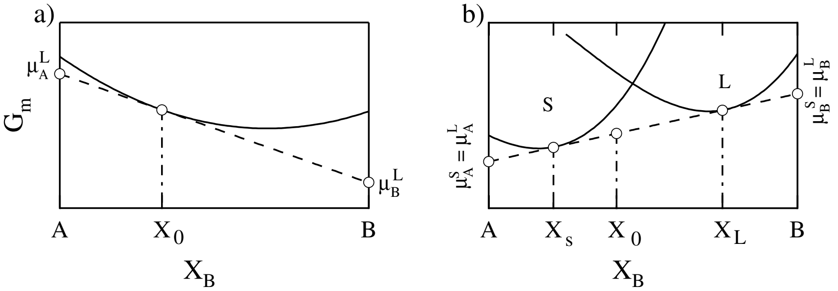


Fig. 3. Schematic free energy vs. composition diagrams for a binary alloy as (a) a single phase liquid solution with the component chemical potentials μ_α^L and μ_β^L determined by the pure component intersections with the tangent to the G_m vs. X_B curve at X_0 and as (b) a two-phase liquid-solid equilibrium at a temperature T_1, with common tangent points at X_S and X_L. The intersection of the G_L and G_S curves defines the T_0 point [12].

entropy,$\varDelta S_m = -R\left[(1-X_B)\ln(1-X_B)+X_B\ln(X_B)\right]$ that applies to ideal and regular solutions as [8]

$$\varDelta G_m = \varDelta H_m - T\varDelta S_m = \varDelta H_m + RT\left[(1-X_B)\ln(1-X_B)\right] \tag{8}$$

The composition dependence of $\varDelta G_m$ yields a concave curve when $\varDelta H_m < 0$ as illustrated in Fig. 3a. In addition, in Fig. 3b, the equilibrium condition of the equality of the chemical potential for each component in the coexisting phases is illustrated by the common tangent construction [8,9]. The various phase equilibria that are defined by a collection of free energy vs. compositions diagrams at different temperatures are the basis of the equilibrium phase diagram. As demonstrated in the schematic diagram in Fig. 4, the common tangent points on the free energy vs. composition diagram define the tie lines between coexisting phases at each temperature on the phase diagram [10].

An important application of the thermodynamic description of phase stability is the representation of the free energy change that accompanies a phase transformation (i.e. the driving free energy [9]). A change of phase requires a departure from equilibrium as the system develops a supersaturation or an undercooling that represents the driving free energy as indicated in Fig. 5. For a dilute solution where the activity is given by the mole fraction, the free energy change associated with a precipitation reaction where the α phase composition changes from X_0 to X_α as pure B precipitates at temperature T_1 is given by

$$\varDelta G = RT_1 \ln\left(\frac{X_0}{X_\alpha}\right) \tag{9}$$

Since the solvus phase boundary composition, X_s, may be represented as [8]

$$X_S = \exp\left(\frac{\varDelta S_S}{R}\right)\exp\left(-\frac{\varDelta H_S}{RT}\right) \tag{10}$$

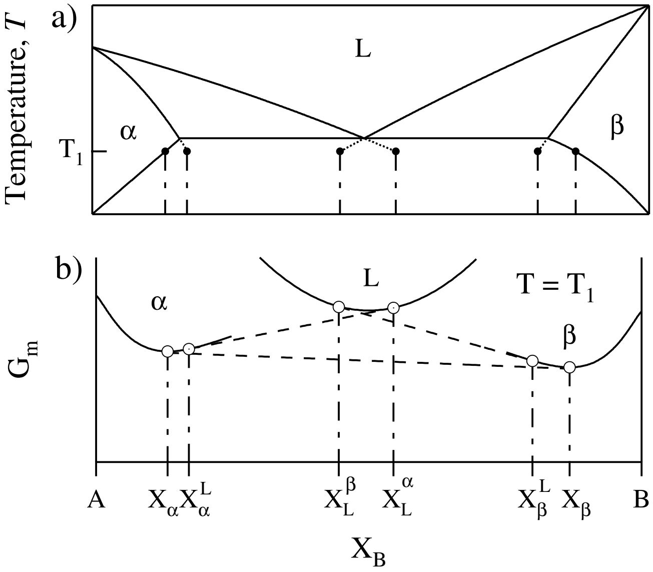


Fig. 4. Schematic phase diagram (a) and free energy vs. composition diagram (b) for a binary alloy that exhibits a eutectic reaction: $L \rightarrow \alpha + \beta$. The relationship between the phase boundaries on the phase diagram and the common tangents is illustrated for a temperature T_1 where both a stable $\alpha + \beta$ and the metastable $L + \alpha$ and $L + \beta$ two-phase equilibria could occur for different kinetic conditions.

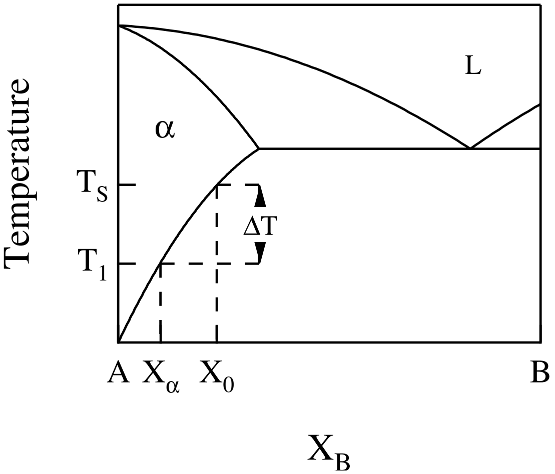


Fig. 5. Schematic illustration for the development of a supersaturation $[X_0/X_\alpha]$ as an α phase alloy of composition X_0 is undercooled by an amount ΔT below the α phase solvus boundary.

where ΔS_S is the entropy for solution and ΔH_S is the enthalpy of solution, the free energy change for precipitation can be expressed as

$$\Delta G = \frac{\Delta H_S \Delta T}{T_S} \tag{11}$$

where for X_0, $\Delta T = T_s - T_1$ and T_s is the solvus temperature. Again, this example demonstrates that a supersaturation can be expressed in terms of an equivalent undercooling.

2.3 Phase Stability Hierarchy

Nucleation-Controlled Reactions. Throughout the analysis of transformation behavior, it is commonly recognized that nucleation control is an important part of the initial stage of a reaction [11]. For example, the product phase density and phase selection are often considered as important signatures of nucleation control; especially when metastable phases and their associated undercooling or supersaturation are involved in the reaction [12–14]. Under nucleation control a strong temperature dependence of the product phase number density can develop and lead to a nanoscale microstructure, but this can be modified by the availability of heterogeneous nucleation sites. Nucleation limitations in diffusion reactions, especially those involved in reactive diffusion also represent a form of nucleation control that is important in nanostructure synthesis [15]. In terms of alloy metastability, nucleation control is important in allowing access to metastable states for measurement and analysis. By accentuating the kinetic factors limiting nucleation through the isolation or removal of active nucleation sites further excursion into higher levels of metastability become possible [16]. In effect, the observation of nucleation control is directly related to the factors that promote the development of large undercooling or supersaturation, enable the expression of kinetic transitions through competitive phase selection and result in the refinement of the product size to the nanoscale level.

Competitive Phase Selection Kinetics. The development of a transformation microstructure based upon stable equilibrium phases or metastable phases depends on the relative nucleation and growth kinetics of the competing structures that are illustrated schematically in Fig. 6 for a solidification reaction [12]. The thermodynamic relationships for the molar free energy G of a material as a liquid, stable phase α and metastable phase β are given in Fig. 6a. In Fig. 6b a function describing the nucleation barrier, ΔG^* is shown to illustrate the role of competitive nucleation. Similarly, in Fig. 6c the relative growth kinetics for stable and metastable phases are illustrated. It is clear that the thermodynamic undercooling to yield temperatures below the melting point of the metastable phase is only a minimum necessary condition for its development. In order to dominate the microstructure, the metastable phase must form at a larger undercooling than the minimum in order to allow it to have faster nucleation and growth kinetics than the competing stable phase.

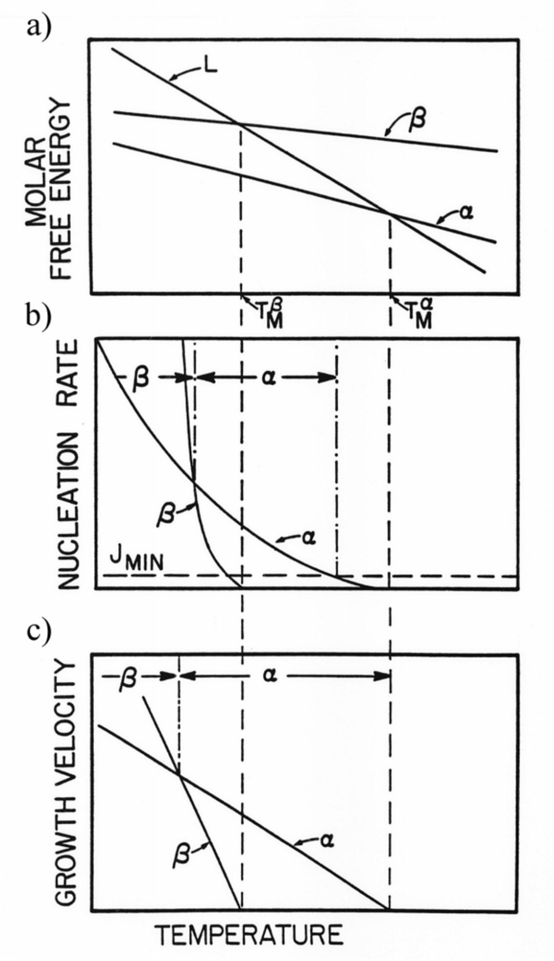


Fig. 6. Schematic representation of the operation of competitive phase selection kinetics which favors the formation of a metastable phase β from the liquid L at low temperature in spite of (a) the thermodynamic stability of α. (b) shows the temperature range for faster nucleation of β phase while (c) shows the temperature range for faster growth of the β phase.

A common theme in the development of new microstructural options by advanced non-equilibrium processing methods is the occurrence of metastable structural states. Often, the reactions that occur during the freezing of undercooled liquids or during other rapid transformations are viewed as non-equilibrium processes. However, it is also evident that some of the departures from full equilibrium can be considered in terms of different levels of metastability. In fact, a hierarchy of equilibrium can be identified based upon the severity of the kinetic constraints that affect the capability of a material to relax towards full equilibrium during processing [12]. As the rate of reaction becomes faster, kine-

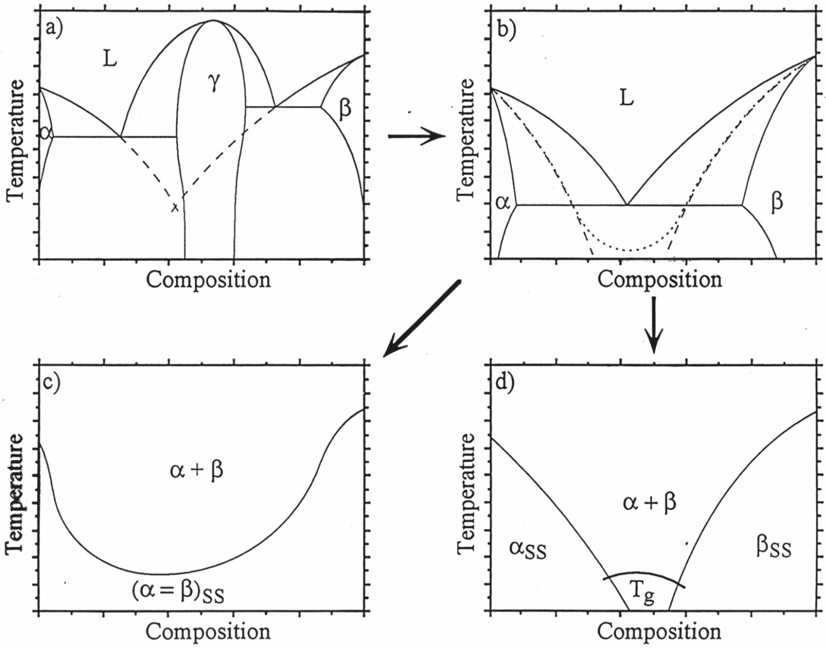


Fig. 7. Schematic illustration of some of the levels in the hierarchy of equilibrium. The equilibrium phase diagram of a system with an intermediate phase is given in (a) Included are the metastable extensions of the liquidus and solidus curves for the primary solution phases(dashed). Solidification under metastable equilibrium conditions can result in a bypassing of the intermediate phase to yield a metastable eutectic phase diagram between α and β. The T_0 curves for the primary solutions are included in (b). The extensions of these curves to temperatures below the metastable eutectic are indicated as dashed curves. If the primary phases have different crystal structures and low mutual solubility, then the T_0 curves might not intersect as in (d). Such a situation favors glass formation in the composition range where the T_g curve is greater than the T_0 curves in (c).

tic constraints that can arise from nucleation and growth limitations associated with an equilibrium product phase formation can develop and can expose alloy metastability that often coincides with nanostructure synthesis. For the suppression of the equilibrium phase or the formation of a kinetically favored metastable phase, it is still possible to analyze reactions in terms of a metastable equilibrium that is used locally at interfaces. The transition from stable to metastable equilibrium is illustrated in Fig. 7 where the kinetic suppression of an equilibrium γ phase (Fig. 7a) yields a metastable eutectic involving the α and β phases [17]. Moreover, it is expected that the application of the appropriate local equilibrium can be used when the processing involves nanoscale structures. Under extreme conditions, loss of interfacial equilibrium for either a stable or metastable phase can develop when even interfacial relaxation becomes too slow. With the loss of interfacial equilibrium, thermodynamics can still be used to restrict the possible range of compositions that can exist at an interface at various temperatures since the selection must yield a net reduction in the free energy of the system [13]. One way to represent the thermodynamic restrictions is based upon the application of T_0 curves which represent the locus of temperatures and compositions where the free energies of two phases are equal as illustrated in Fig. 3b for liquid and solid phases and thus define the limiting condition for partitionless transformation [13]. For example, as interfacial equilibrium is lost, the liquidus and solidus boundaries in Fig. 7b collapse to the T_0 curves. With isomorphous systems that exhibit complete solubility the T_0 curve is continuous with composition (Fig. 7c) while for alloys based upon components with different structures each crystal phase has a T_0 curve (Fig. 7d). At temperatures and compositions above the T_0 curves solute partitioning is required for solidification (Fig. 7d). Because of the diffusional constraint due to solute partitioning, a crystallization reaction can be inhibited by quenching to promote glass formation [18,19]. Within the overall hierarchy of stability the systematic examination of the different levels of kinetic constraints can provide useful insight into the thermodynamic analysis of alloy metastability and phase selection during nanostructure formation.

3 Nanostructure Considerations

3.1 Thermodynamic Modifications

The nanoscale is often reported as a linear dimension, but the important interfacial effects should be considered in terms of the interfacial area per unit volume A/V. For example, for a sphere: $A/V = 3/r = 3 \times 10^7$ m^{-1} for $r = 100$ nm. This is significant! Interfacial effects can be included in the free energy as [9]

$$dG = VdP - SdT + \Sigma\mu_i dn_i + \Sigma\sigma_i dA_i \tag{12}$$

where σ_i is the interfacial energy and A_i is the interfacial area. The increment in free energy due to interfaces is represented by the Gibbs-Thomson relation [9] in terms of the interface curvature κ, that is determined by the principal radii, r_i, as:

$$\Delta G = \kappa\sigma = \left(r_1^{-1} + r_2^{-1}\right)\sigma V_m \tag{13}$$

For a sphere, $\Delta G = 2\sigma V_m/r$. The effect of curvature on the melting point and the relative phase stability is indicated in Fig. 8 where the dependence of the melting point of a pure component on curvature $T_m(r)$, α can be expressed as

$$T_m(r) = T_m - \frac{2\sigma_{SL}V_m}{r\Delta S_f} \tag{14}$$

where T_m is the macroscopic size melting point. There are two key points that are illustrated in Fig. 8. First, the melting points for the α and β phases i.e. $T_m^\alpha(r)$ and $T_m^\beta(r)$ decrease with increasing curvature. Secondly, the β phase that is metastable for macroscopic sizes can become the stable phase with respect to the α phase at high curvature values. Furthermore, the phase stability relations

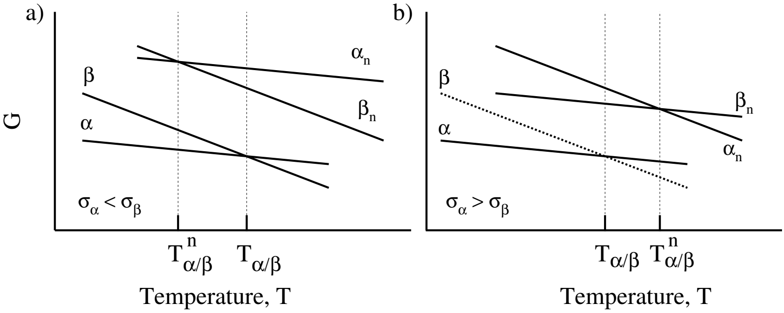


Fig. 8. Schematic Illustration of the modification of the stability of the α phase at the nanoscale. (a) when $\sigma_\alpha < \sigma_\beta$, there is an enhancement of a phase stability; (b) when $\sigma_\alpha > \sigma_\beta$, there is a reversal of phase stability between the α phase and the β phase(the superscript n refers to the nanoscale size).

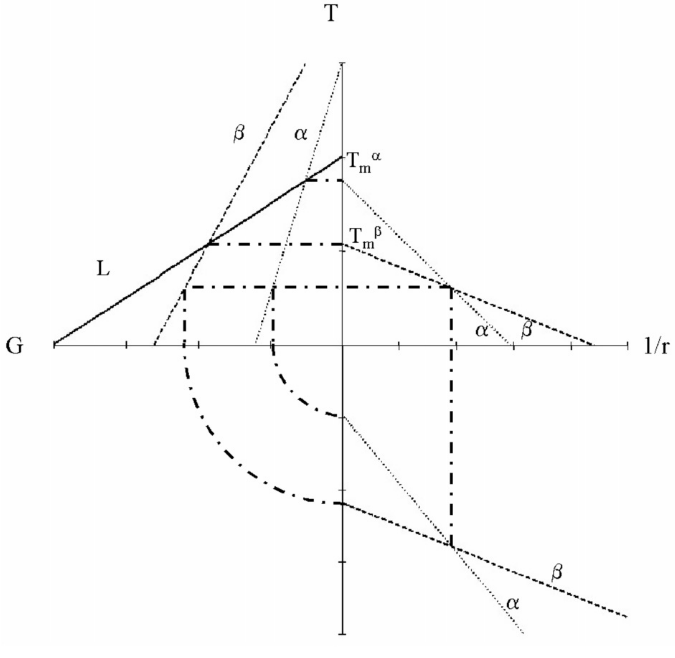


Fig. 9. Schematic illustration of the influence of curvature $(1/r)$ on the melting point of the α phase and the β phase and on the relative phase stability when $\sigma_{\alpha L} V_M / \Delta S_{\alpha f} > \sigma_{\beta L} V_M / \Delta S_{\beta f}$.

for a macroscopic scale system can be modified in different ways depending on the relative magnitude of σ for each phase for a nanostructured system. As illustrated in Fig. 9, when $\sigma_\alpha < \sigma_\beta$, the stability of α is enhanced at the nanoscale. However, when $\sigma_\alpha > \sigma_\beta$, there is a reversal of phase stability at the nanoscale. There are a number of examples of the reversal in relative phase stability when the size

Table 1. Representative driving free energies for various transformation reactions

Reaction Process	Free Energy	Typical Value (J/mol)$T = 1000$K	Remarks
Crystallization	$\Delta H_f \Delta T/T_m$	$3 \cdot 10^3$	
Mixing/ Interdiffusion	$RT(X_A \ln X_A$ $+X_B \ln X_B)$	$5 \cdot 10^3$	Ideal solution behavior
Oxidation	$\Delta G^0 = -RT \ln K$	$5 \cdot 10^4 - 5 \cdot 10^6$	ΔG^0 formation of oxide
Sublimation/ Deposition	$\Delta H_v \Delta T/T_s$	$10^4 - 10^5$	ΔH_v- sublimation enthalpy T_s- sublimation temperature
Grain growth	$2\sigma/r$	20 for $r = 1$ μm $2 \cdot 10^3$ for $r = 10$ nm	$\sigma = 1$ J/m^2
Precipitation	$RT \ln X_\alpha/X_0$	10^4	$X_\alpha/X_0 = 10$
Cold work (stored energy)	$\rho G b^2$	$10^2 - 10^3$	$G-$ shear modulus $b-$ Burgers vector $\rho-$ dislocation density

scale changes from macroscopic to nanostructured [20–24]. In fact, it is possible to consider that by suitable control over σ, such as through selective adsorption, the phase stability may be controlled in nanostructured systems [25–27].

The excess free energy due to curvature that acts to control the relative phase stability in nanoscale materials also acts to modify other thermodynamic behavior. For example with the same analysis that was applied to express the curvature dependence of the melting point, the dependence of the solubility may be represented in terms of particle size. This behavior is illustrated in Fig. 10 where it is evident that the increase in chemical potential with decreasing particle size will act to drive a diffusive flux between neighboring articles in a particle size distribution such as that produced by a precipitation reaction. An important consequence of the interparticle transport is the dissolution of the finest particles along with a concomitant increase in the average particle size in the distribution [10]. This coarsening or Ostwald ripening behavior is central to the kinetic stability of a nanoscale microstructure.

A useful perspective on the different diffusional reactions and phase formation processes that have been discussed can be developed by considering the relative magnitudes of the driving free energies associated with each of the reactions. A partial summary of the representative magnitudes for the driving free energies for a selection of reactions is presented in Table 1. The listings in Table 1 indicate that for typical processes involving chemical changes or phase transitions the driving free energies range from a few to 100 KJ/mole. While the excess free energy associated with a nanoscale microstructure is only a few KJ/mole, this level is sufficient to alter the relative phase stability and to modify the path of kinetic reactions.

3.2 Nanoscale Reaction Kinetics

Initial Stage of Interface Reaction. Besides thermodynamic comparisons, there are also kinetic features that are negligible in bulk sizes, but can become

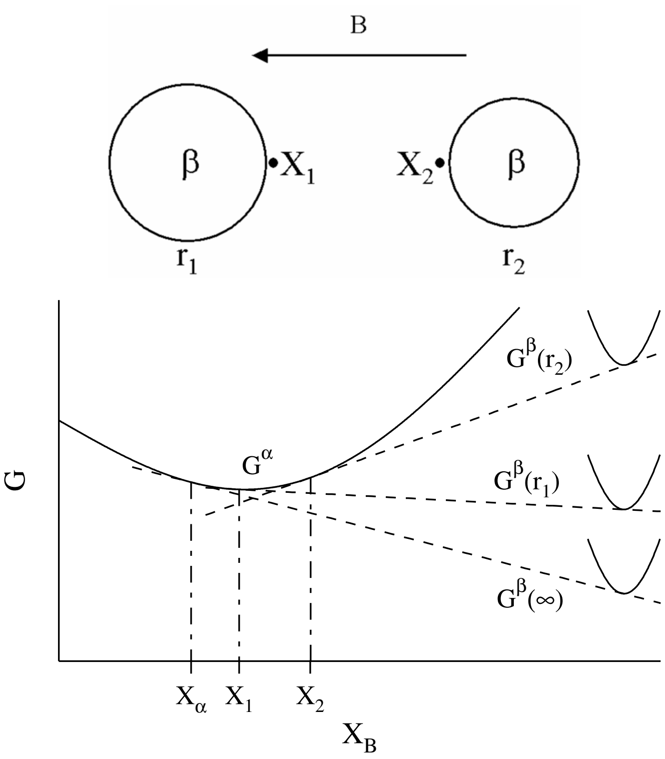


Fig. 10. Schematic free energy vs. composition diagram demonstrating the influence of interface curvature of a β phase particle on the solubility.

a dominant influence at the nanometer size scale. As the size scale of a microstructure reaches the nanometer level, the diminished length scale of diffusional processes has an important impact on the kinetics of solute redistribution and phase reactions. For example, with a limited spatial extent for diffusion it is necessary to consider the influence of large concentration (i.e. chemical potential) gradients on reaction kinetics. During the initial stages of reaction, the steep gradient acts to inhibit reaction kinetics and to allow access to metastable regions through a kinetic stabilization [16]. However, this is only a sufficient condition since Desré [28] has shown that a steep concentration gradient can also act to reduce the effective driving free energy for phase nucleation.

During the initial stage of an interface reaction there are several specific features that can have a significant impact on the reaction kinetics [16]. This condition is illustrated in Fig. 11, where the development of an intermediate β phase during the aging of a supersaturated α solid solution (Fig. 11b) is compared to β phase formation in a diffusion couple between pure A and B (Fig. 11c). In the conventional aging treatment, the formation of β occurs by a fluctuational nucleation process in a compositionally homogeneous, supersaturated matrix α phase of composition X_0. For example, it is clear that some diffusional mixing

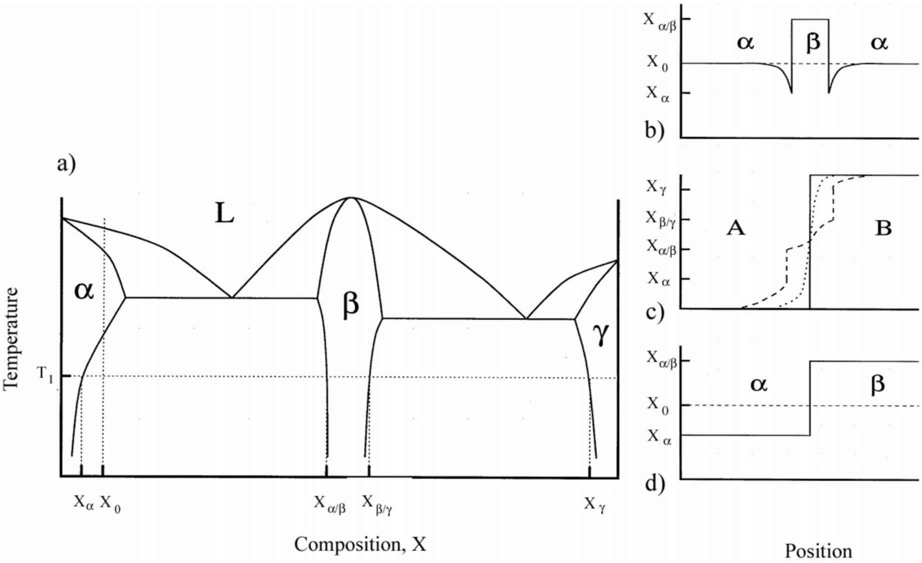


Fig. 11. Composition of phase development in an alloy of composition X_0 during (b) precipitation and (c) interdiffusion. In both cases, the reaction yields a common final state (d).

to at least the levels X_α and X_γ at T_1 must occur; otherwise a β nucleus could not develop since it would be in contact with unsaturated phases and dissolve. Moreover, if the α and γ phases are isostructural and do not undergo phase separation, the interdiffusion profile can extend smoothly across the original interface as shown in Fig. 11c (dotted curve). In this case all compositions from X_α to X_γ are available for nucleation of β phase. Once the β phase develops, the concentration profile is modified as shown in Fig. 11c (broken curve) for the growth of a β phase layer at the expense of γ phase. Of course, the common end state of the diffusion process for an overall composition X_0 is also given by Fig. 11d. The comparison in Fig. 11 clearly demonstrates that for an overall composition X_0 the end state of an intermediate phase formation by an aging reaction and an interdiffusion reaction are the same, but the pathways to this end state are quite different. The fundamental difference in behavior is closely related to the influence of a large concentration gradient, ∇X_B, on intermediate phase formation.

The role of the concentration gradient in delaying intermediate phase nucleation was first recognized by Desré and Yavari [28,29] and Gusak [30,31] who applied it to explain the observation of a critical thickness of an amorphous phase layer developed during solid state amorphization by diffusion. With a critical nucleation size, r^* and a concentration range, ΔX_B over which the intermediate phase can form, nucleation will be inhibited unless $\Delta X_B/r^* \geq \nabla X_B$. When the effect of the concentration gradient on the thermodynamics of phase formation is also considered, ∇X_B values of the order of 10^6 m^{-1} can have a significant

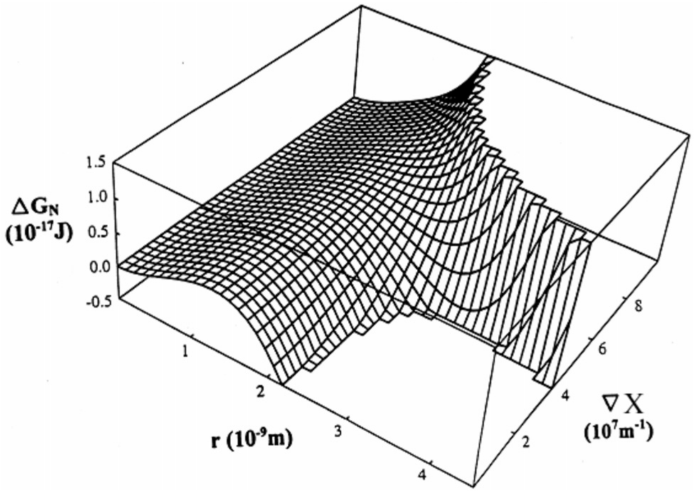


Fig. 12. Work of nucleus formation ΔG_N as a function of nucleus size r and imposed concentration gradient ∇X.

influence on the initial phase formation and can inhibit nucleation to allow metastable supersaturated regions to develop over a distance well in excess of the 100 nm scale usually considered to mark the start of nanocrystalline materials.

In order to illustrate the concentration gradient effect more completely, it is useful to consider the result of the Desré analysis [28] of the nucleation barrier, ΔG_N, for a cubical nucleus of length $2r$ under an imposed ∇X_B. For the volume nucleation of an intermediate phase of composition X_B^* from a solid solution of composition X_0 the work is given by

$$\Delta G_N \left(X_0, X^*\right) = 24\sigma_{sc} r^2 + 8\rho \Delta G_{sc} r^3 + \frac{4}{3}\rho\alpha \left(\nabla^2 X_B\right) r^5 \tag{15}$$

where σ_{sc} is the interfacial energy between the solution and the intermediate phase, ρ is the number of moles of atoms per unit volume, ΔG_{sc} is the free energy change for polymorphous transformation at composition X^*, and $\alpha = (\partial G_s^2/\partial X_B^2)$ for the solution. The first two terms in (15) represent the usual contribution of surface and volume effects to nucleus formation. The main consequence of the concentration gradient appears in the $\nabla^2 X_B r^5$ term of (15). Clearly, this term will be most important during the earliest period of interdiffusion. In fact, as shown in Fig. 12 which refers to nucleation of $Ni_{10}Zr_7$ in Ni-Zr amorphous layers [28,29] at the largest ∇X_B values nucleation is prohibited until a critical gradient, ∇X_B^*, given by

$$\nabla X_B^* = \frac{\rho}{9\sigma_{sc}} \frac{(2\Delta G_{sc})^{3/2}}{\alpha^{1/2}} \tag{16}$$

is reached, but even at ∇X_B^* the nucleation barrier is larger than that for a uniform solid solution. A similar result of impeded intermediate phase nucleation

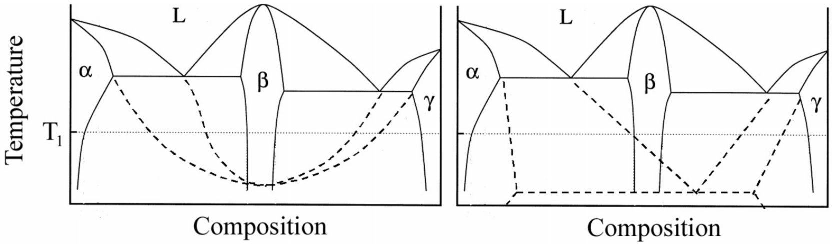


Fig. 13. Two possible forms of alloy metastability during interdiffusion.

in a large concentration gradient has been reported by Hoyt and Brush [32] who also examined the optimal nucleus shape. In order to minimize the effect of the gradient, it appears more favorable for the nucleus to spread out along the initial interface. The importance of a diffusion process before phase nucleation in thin films has been noted as well by Coffey and Barmak [33] and Philibert [34]. There are also transient effects during low temperature interdiffusion in multilayer samples that can influence phase formation [35]. Some of the time dependent effects can be traced to an asymmetric ∇X_B profile that often develops due to large differences in the component interdiffusion coefficients. These developments represent new kinetic features of interface reactions which can provide for a rational accounting of phase development [16,32,34,36].

While the influence of an imposed steep concentration gradient was originally applied in an attempt to understand solid state amorphization reactions, recent experience indicates that the thermodynamic consequences of ∇X_B on phase formation are general and may be required as part of a general model of initial phase formation during interdiffusion. Furthermore, the delay period introduced in order for the ∇X_B level to be reduced below the critical level for nucleation provides a kinetic constraint which exposes the interdiffusion zone to metastable equilibria. Two general forms of the alloy phase metastability that can develop are illustrated in Fig. 13. For the conditions depicted in Fig. 11 for isostructural terminal phases a metastable isomorphous equilibria develops before intermediate phase nucleation (Fig. 13a). Alternatively, for phases with different crystal structures complete solubility is interrupted by a two-phase composition step or by the development of a metastable extension of a high temperature phase (Fig. 13b). In solid state amorphization this phase is a liquid which develops below the glass transition. In effect, the amorphous layer provides for a metastable solubility and allows for ∇X_B to be reduced below the critical value for nucleation.

Nanoscale Phase Separation. Even for reactions that are not subject to a nucleation limitation such as spinodal decomposition, the kinetics can be influenced when the reaction occurs in nanometer size scale volumes. Spinodal decomposition occurs at the limit of metastability of a single phase solution to diffusional unmixing or phase separation. The analysis of spinodal decomposition requires the solution of the diffusion equation in an inhomogeneous system.

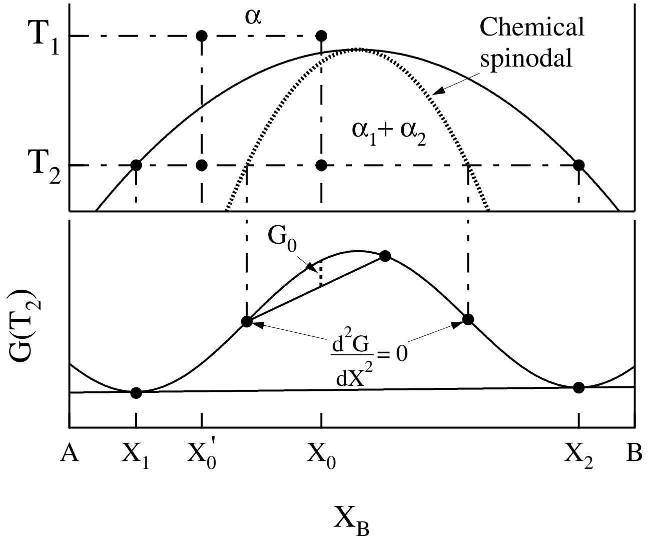


Fig. 14. Alloys between the spinodal points are unstable and can decompose into two coherent phases α_1 and α_2 without overcoming an activation energy barrier. Alloys between the coherent miscibility gaps and the spinodal are metastable and can decompose only after nucleation of the other phase.

The interdiffusion flux, $\tilde{J}$ is represented by [37,38]

$$\tilde{J} = -M\nabla(\mu_A - \mu_B) \tag{17}$$

where M is the mobility and is positive. For a homogeneous system,

$$\mu_A - \mu_B = \frac{\partial G}{\partial X_A} \tag{18}$$

so that

$$\tilde{J} = -M\frac{\partial^2 G}{\partial X_A^2}\nabla X_A \tag{19}$$

Since the interdiffusion coefficient $\tilde{D}$ can be defined as

$$\tilde{D} = M\frac{\partial^2 G}{\partial X_A^2} \tag{20}$$

It is evident that $\tilde{D} < 0$ within the spinodal composition range that is indicated in Fig. 14 where $\partial^2 G/\partial X_A^2 < 0$. In effect, this characteristic signifies the uphill diffusion process that yields the phase separation. For an inhomogeneous system it is necessary to include the local interactions that arise from ∇X_A and modify the free energy. The simplest form for the free energy that allows for the incorporation of the local interactions is expressed by [38]

$$\mu_A - \mu_B = \frac{\partial G}{\partial X_A} - 2K\nabla^2 C_A \tag{21}$$

where $K = N_v k_B T_c \psi^2$ with N_v the number of atoms per unit volume, k_B the Boltzmann constant, T_c the consolute temperature in Fig. 14 and ψ is the interaction parameter, which is of the order of the interatomic spacing a_0 [37]. For the inhomogeneous system the flux becomes

$$\tilde{J} = -M \frac{\partial^2 G}{\partial X_A^2} \nabla X_A - 2MK \nabla^2 X_A \tag{22}$$

from the divergence the diffusion equation is expressed as

$$\frac{\partial X_A}{\partial t} = M \frac{\partial^2 G}{\partial X_A^2} \nabla^2 X_A - 2MK \nabla^4 X_A \tag{23}$$

The solution of the diffusion equation gives

$$X_A - X_0 = \exp\left[R(\bar{\beta})t\right] \cos\left(\bar{\beta} r\right) \tag{24}$$

where $\beta = 2\pi/\lambda$ and λ is the wavelength of the composition profile. Since $R(\beta)$ has a sharp maximum with β only wavelengths near the maximum λ_m are observed.

During spinodal decomposition the onset of the reaction is set by the scale of diffusional unmixing that is expressed by the wavelength of the composition modulation, λ_c that is given by [37]:

$$\lambda_c = \sqrt{\frac{-8\pi^2 K}{\partial^2 G / \partial X_B^2}} \tag{25}$$

where K is a gradient energy factor and $(\partial^2 G/\partial X_B^2)$ is the curvature of the molar free energy as a function of composition, X. For a regular solution, $\partial^2 G/\partial X_B^2 = -N_v k_B (T_s - T)/(X_B - X_B^2)$, where $T_s - T$ is the undercooling below the spinodal onset temperature, T_s [38]. With a regular solution then the critical wavelength for an equiatomic alloys can be represented as

$$\lambda_c = \sqrt{\frac{32\pi^2 a_0^2 T_c}{(T_s - T)}} \tag{26}$$

so that λ_c is inversely proportional to the amount of undercooling below the spinodal boundary. Significant decomposition by spinodal unmixing can not develop until the value of λ_c is less than the size of the sample. For nanometer samples such as thin films or particles, the kinetic inhibition can allow for the presence of a single phase solid solution at an undercooling of several hundred degrees below the bulk chemical spinodal [38] as indicated in the plot given in Fig. 15 that shows the undercooling for the onset decomposition for the scaled λ/ψ values. While the persistence of an unstable solid solution below the spinodal boundary in nanometer samples may appear to be a nonequilibrium effect due to the processing method used to create nanocrystalline sample volumes, it is actually a natural consequence of the mechanism of the spinodal diffusion process. Lastly, in nanometer size scales there can be a strong and even dominant

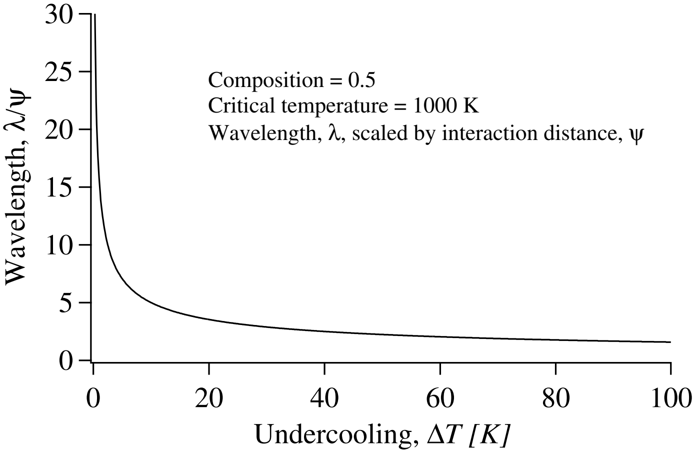


Fig. 15. The expected wavelength vs. undercooling below the spinodal for the hypothetical alloy with $T_c = 1000\text{K}$ and $X_B = 0.5$ [40].

influence of elastic strain energy on the relative phase stability and competitive kinetics. A full treatment of these effects is beyond the scope of the present discussion, but important developments can be found in the work of Johnson and Voorhees [39].

4 Nanocrystallization

The generation of nanocrystalline structures from the liquid or vapor requires the attainment of a high crystal nucleation rate which in turn is promoted by a large undercooling before the onset of crystallization. Actually, there are two pathways that may be followed to achieve the high crystal nucleation density. If a sample is rapidly quenched at a rate that happens to coincide with the conditions for a high nucleation rate a nanocrystalline structure is possible by direct quenching. However, under most conditions of rapid quenching it is difficult to control the processing and reproducibility. Instead, a direct cooling to an amorphous state and a subsequent low temperature crystallization treatment is usually preferred as a method of achieving reproducible nanostructure synthesis including the fabrication of nanostructures in bulk sample volumes [40–44].

The metallic glasses that provide the most effective routes to nanocrystallization are closely related to two important aspects of solidification that involve kinetic competition: (1) avoidance of crystallization upon cooling of the liquid and (2) the control of crystallization upon heating of the glass. Although there are connections between these aspects, including the common underlying important role of melt undercooling as a measure of liquid metastability, in each case the controlling reactions occur under regimes of different kinetic constraints [45]. In addition to the closed system methods involving liquid or vapor quenching, it is recognized that open systems involving continuous deformation or irradiation can drive a material towards nanocrystallinity. In this case, the stored energy due to defects, grain refinement and solute supersaturation is a measure of the level of metastability that is crucial to consider in the analysis of amorphization and the development of nanostructured microstructures [4,7,41].

4.1 Devitrification Reactions

The crystallization (or devitrification) behavior of amorphous materials is of central importance in the synthesis of nanostructured materials [46]. The reaction pathways that are operative during crystallization must be identified and controlled in order to develop successful strategies for the consolidation of amorphous powders or ribbons that can be processed into bulk nanostructured solids. Moreover, the control of the reaction path during crystallization provides for the option to develop nanoscale structures with different phase selection.

The different reaction paths and product selection options are identified in Fig. 1 which illustrates schematically the free energy relationships between an initial amorphous phase that is considered as an undercooled liquid solution and several crystalline product phases that include stable α and β phases and a metastable γ phase. Within the alloy composition ranges that are usually favored for glass formation there are several types of crystallization reactions that can be employed to develop nanocrystalline structures during controlled heating or isothermal reaction [42]. One of the simplest reactions is the direct transformation from the glass to a single phase crystal without composition change as illustrated in Fig. 16a and b by pathways (1) and (2) for either stable or metastable initial product phases. The composition invariant or polymorphic reaction can yield metastable structures such as supersaturated solid solution phases or metastable intermediate phases that can undergo further transformation that is indicated by pathways (1′) and (2′) in Fig. 16a and b.

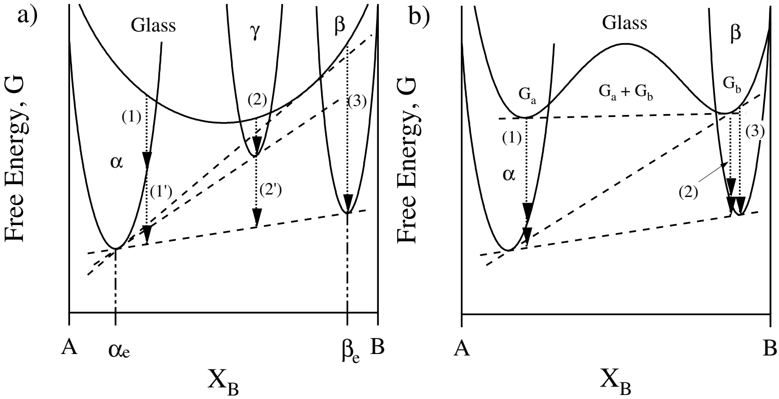


Fig. 16. Schematic Free Energy versus composition diagrams illustrating some of the possible nanocrystallization reactions of an amorphous phase. (a) reaction pathways for an alloy with a negative heat of mixing and a metastable γ phase. (b) reaction pathways for an alloy with a positive heat of mixing.

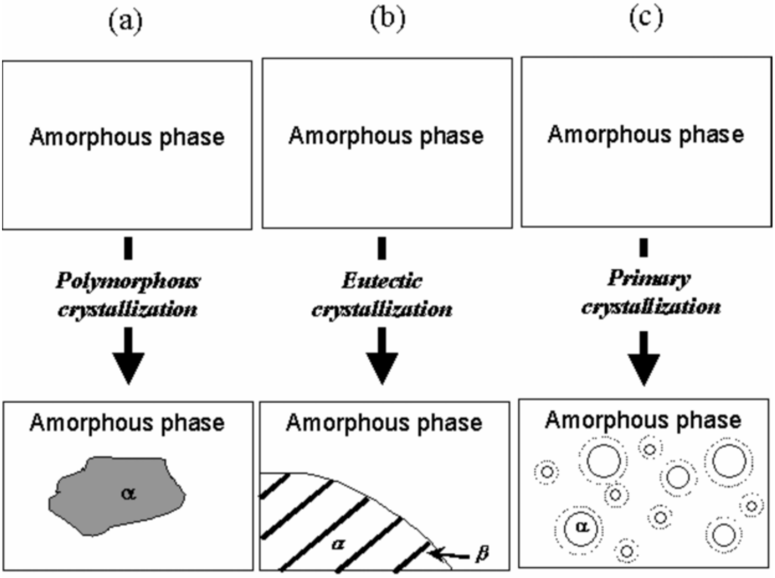


Fig. 17. Schematic illustration of the characteristic microstructural morphologies that develop during nanocrystallization by (a) polymorphic, (b) eutectic, and (c) primary phase reactions. In (c) the dotted curve around primary phase nanocrystals denotes the extent of the solute diffusion field.

With primary crystallization, a single phase is the initial product, but the reaction proceeds with a partitioning of solute to yield a solute lean primary crystal and a residual amorphous phase matrix that is enriched in solute. The kinetics of primary crystallization are evidently related to the rate of solute diffusion in the amorphous matrix that is necessary to dissipate the solute that is rejected during primary crystal growth. It is also apparent that primary crystallization does not result in a stable equilibrium product structure that is indicated by the compositions α_e and β_e in Fig. 16a and b. In order to complete the primary crystallization a subsequent multiphase crystallization develops either from the nucleation site provided by the primary crystal or directly from the amorphous phase. For example, with eutectic crystallization that is indicated by pathway (3) in Fig. 16a, the product phases (i.e. α and β) often develop by a coupled growth and appear with a lamellar or rod type of regular morphology in a spherulitic pattern. In this case the synthesis of a nanoscale microstructure requires a high density of α and β colonies with an ultrafine lamellar spacing. A schematic illustration of the characteristic microstructural morphologies associated with each of the nanocrystallization reactions is provided in Fig. 17.

Often, under high undercooling conditions metastable phase reactions can develop as a precursor to the formation of stable crystallization products. For example, as indicated in Fig. 16b the undercooled liquid or amorphous phase can undergo a phase separation reaction leading to the formation of two liquids with different compositions that are indicated by G_a and G_b in Fig. 16b. At low temperature or high undercooling, limited atomic mobility will result in a fine scale of phase separation that can extend into the nanoscale regime. Moreover, it has been established that in some cases the interfaces between the different liquid regions can serve as heterogeneous nucleation sites for subsequent crystallization reactions and establish high nucleation product number densities [47]. In addition, there is evidence that in some systems minor impurity levels can promote the development of phase separation reactions [14]. Another example of a precursor reaction is the formation of an intermediate phase as a metastable product as illustrated in Fig. 16a for the γ phase.

4.2 Kinetics of Nanocrystallization

One of the key requirements that must be satisfied for the development of a nanoscale microstructure by a crystallization reaction is the attainment of a very high nucleation product number density. A key concept in nucleation is the development of an activation barrier to the creation of a new phase due to the interface between the produce and parent structures. The origin of the nucleation barrier is illustrated in Fig. 18 for the capillarity or sharp interface analysis [48–52]. The work to form a nucleus of size r, $\Delta G(r)$, is given by a term to create the interface and volume free energy reduction from transformation as

$$\Delta G(r) = -\frac{12\pi r^2 \sigma_{LS} + 4\pi r^3 \Delta G^*}{3k_B T} \tag{27}$$

The magnitude of the barrier is $b\sigma_L^3 S/(3\Delta G_v^2)$ and occurs at the critical size $r^* = 2\sigma_{LS}/\Delta G_v$ where σ_{LS} is the liquid-solid interfacial energy, ΔG_v is the driving free energy for nucleation of a unit volume of product phase and $b = 16\pi/3$

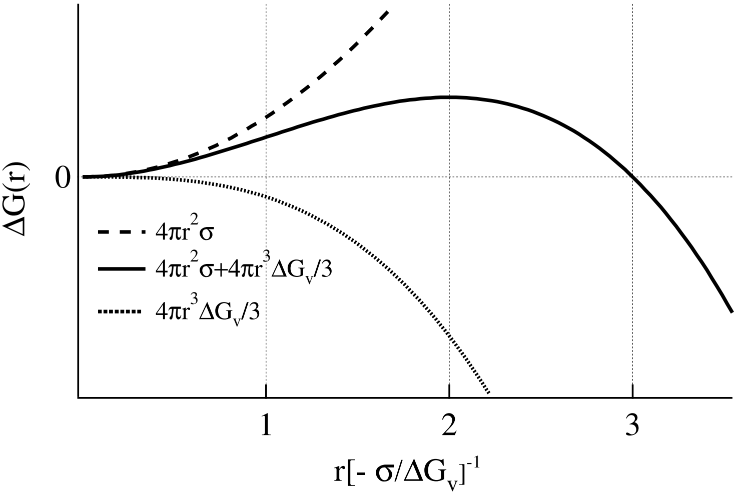


Fig. 18. The free energy change associated with homogeneous nucleation of a sphere of radius r.

for spherical nuclei. Nucleation is a fluctuational growth process in cluster size space so that the current of clusters or the nucleation rate is the result of both single atom addition and removal from the evolving cluster. At high driving free energy the nucleation rate is represented well by the forward flux as the product of three terms: the number of cluster surface sites, the jump frequency and a cluster concentration given by [11,16]

$$J(T) = \frac{4\pi r^{*2}}{a_0^2}\frac{D_l}{a_0^2}C(1)\exp\left[-\frac{\Delta G^*}{k_B T}\right] \tag{28}$$

where a_0 is the jump distance, D_l is the diffusivity, $C(1)\exp[-\Delta G^*/k_B T]$ is the concentration of critical clusters, $C(n^*)$, in a system with a monomer concentration of $C(1)$. The main features of the nucleation rate kinetics can be described by

$$J_i^s = \Omega_i \exp\left[-\frac{\Delta G^* f(\theta)}{k_B T}\right] \tag{29}$$

where J_i^s is the steady state nucleation rate on a volume ($i = v$) or surface basis ($i = a$). Respective values for the prefactor, Ω_i, activation barrier, ΔG^*, and the contact angle function, $f(\theta)$, are used in (28) and $k_B T$ is the thermal energy. With planar catalytic sites and spherical nuclei $f(\theta) = [2 - 3\cos(\theta) + \cos(3\theta)]/4$. The expressions for Ω_i involve a product of a nucleation site density on a catalytic surface or volume basis, the number of atoms on a nucleus surface and a jump frequency. For most cases, $\Omega_v = 10^{30}/\eta$ cm^{-3}s^{-1} and $\Omega_a = \phi 10^{22}/\eta$ cm^{-2}s^{-1} with η, the liquid shear viscosity (in poise) given by [12]

$$\eta = 10^{-3.3}\exp\left[\frac{3.34 T_L}{T - T_g}\right] \tag{30}$$

in terms of the liquidus temperature, T_L and the glass transition, T_g and ϕ, the fraction of active catalytic sites. Following the establishment of a supersaturation or undercooling, there is an initial time period during which the nucleation cluster population evolves towards the steady state distribution. During this transient period the time dependent nucleation rate, $J_i(t)$, is given by [51,53]

$$J_i(t) = J_i^s\left[1 + 2\Sigma(-1)^n \exp\left(-\frac{n^2 t}{\tau}\right)\right] \tag{31}$$

where τ is the time lag or delay time that is estimated by $r^{*2}/\pi^2 D_l$. In order to achieve a nanocrystalline microstructure (i.e. with a size scale $\leq$ 100 nm) in a fully crystallized volume, the nucleation number density should be at least of the order of 10^{21} m^{-3}. Of course, nanocrystallization can be achieved only if there are also restrictions on the kinetics of nanocrystal growth following nucleation.

The kinetic analysis of growth follows different forms that depend on the nature of the solute partitioning associated with phase growth. For example, during polymorphous transformation without solute redistribution, the growth rate, V, is controlled by interface attachment limited kinetics as represented by [10,52]

$$V = V_0 \exp\left[-\frac{Q_D}{RT}\right]\left(1 - \exp\left[\frac{\Delta G_v}{RT}\right]\right) \tag{32}$$

where V_0 is a prefactor of the order of 5×10^3 m/s and Q_D is the activation energy for interface jumps. At low temperature where $\Delta G_v \gg RT$ growth is diffusion controlled as expressed by [42]

$$V = V_0 \exp\left[-\frac{Q_D}{RT}\right] \tag{33}$$

For the case of eutectic reaction where the solute redistribution is limited to the reaction interface [42]

$$V \cong 4\tilde{D}_I \frac{\delta}{\lambda^2} \tag{34}$$

where $\tilde{D}_I$ is the interface diffusivity, δ is the thickness of the reaction front and λ is the lamellar spacing. With these kinetic modes, the reaction is relatively rapid and a metastable microstructure based upon nanocrystals and an amorphous phase with the original composition is possible if the kinetics of subsequent decomposition reactions to a more stable phase constitution is sluggish.

When growth requires a redistribution of solute as in primary crystallization, the kinetics are limited by the rate of diffusion of the rejected solute into the amorphous matrix. For evolving nanocrystals that are isolated from each other the growth rate has the following form [10]

$$V = \frac{\alpha}{2}\sqrt{\frac{D}{t}} \tag{35}$$

where α is a dimensionless factor that is evaluated from the compositions at the particle/matrix interface and the average composition and D will be controlled by the slowest diffusing solute in a multicomponent alloy. However, at high nucleation densities the isolation can be lost as the diffusion fields from neighboring nanocrystals begin to overlap (i.e. soft impingement). Under this condition there is a kinetic inhibition to further growth. Concurrent with the growth of nanocrystals, the highly refined sizes indicate that capillarity effects such as Ostwald ripening due to curvature driven transport (i.e. Gibbs Thomson effect) can be important.

5 Nanocrystallization of Amorphous Alloys

Following amorphization by melt quenching a number of metallic glasses do exhibit a clear glass transition signal, T_g, upon reheating. It is useful to note that the glass transition is not a phase transformation in a thermodynamic sense, but it is a kinetic manifestation of the slowing down of atomic transport in the liquid with cooling [55]. In fact, the calorimetric glass transition signal is due to the

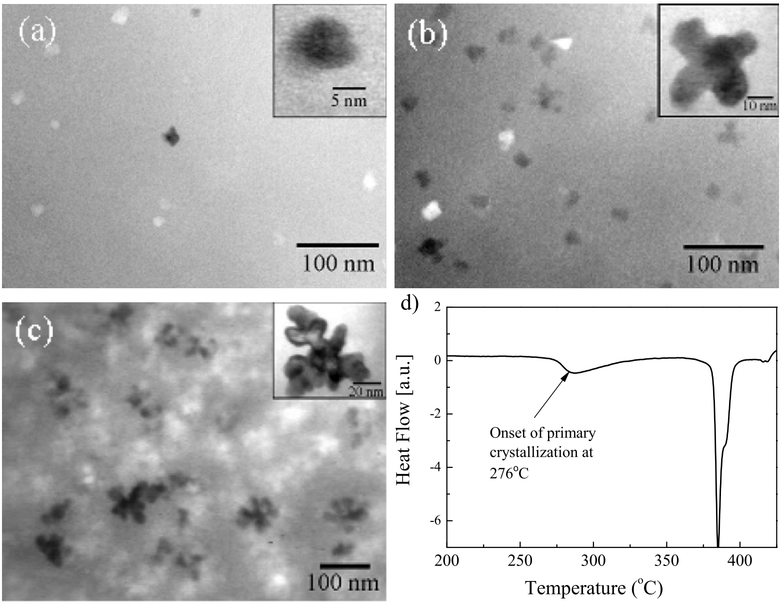


Fig. 19. TEM bright-field images from an $Al_{88}Y_7Fe_5$ melt-spun ribbon that was isothermally annealed at 245°C for (a) ten minutes; (b) 30 minutes; (c) 100 minutes and (d) continuous heating DSC trace at 40 K/min showing a primary crystallization onset at 276°C.

large change in heat capacity that occurs when a liquid becomes configurationally frozen. The slowing down of atomic transport is also reflected by an increase in liquid viscosity. The time for the liquid structure to relax during cooling is related to the viscosity and for typical laboratory measurement conditions T_g corresponds to a viscosity in the range of $10^{12} - 10^{-13}$ poise ($10^{11} - 10^{12}$ Pa-s).

Other amorphous alloys such as the marginal glass forming alloys do not show a clear T_g signal [56]. Instead, initial exothermic maxima are observed to develop that indicate a multiple stage crystallization [54,57] as shown in Fig. 19 for an amorphous $Al_{88}Y_7Fe_5$ ribbon after melt spinning and after initial crystallization. The microstructural analysis has established that for many Al-base alloys the initial crystallization corresponds to primary phase formation (i.e. Al) yielding a sample that contains a high density of nanocrystals within an amorphous matrix [58]. This behavior is of importance in understanding the kinetic control of glass formation. The two basic strategies to synthesize amorphous alloys are illustrated schematically in Fig. 20. With nucleation control, the undercooling that is achieved during cooling bypasses the nucleation reaction and the nucleation size distribution [45], $C(n)$ that may be retained by the cooling does not overlap with the critical nucleation size, n^* at the crystallization temperature, T_x. As a result, there is no precursor reaction to influence the evolution of crystalline clusters during subsequent thermal treatment. In this way, a clear separation in temperature between the T_g and T_x signals can be observed during reheating of

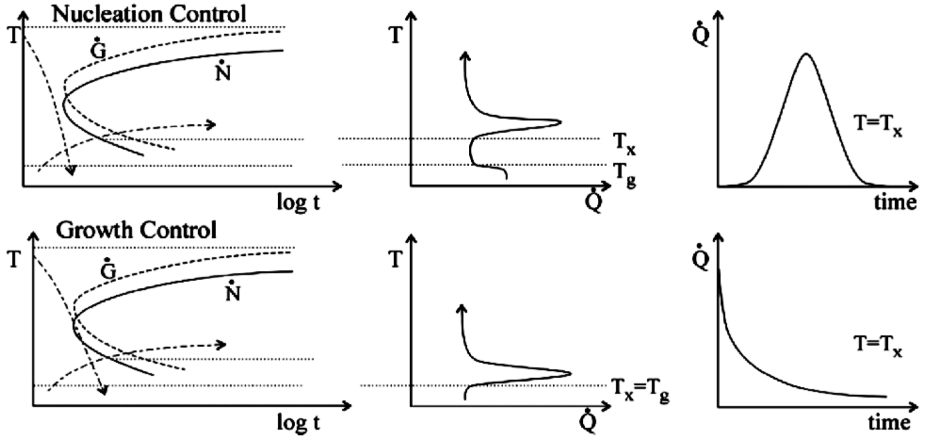


Fig. 20. The principal forms of kinetic control for metallic glass formation.

a glass. These kinetic conditions are the basis for bulk glass formation during slow cooling [41]. During isothermal annealing at T_x, the heat evolution rate exhibits a clear delay before the onset of the nucleation reaction and a peak maximum associated with the completion of nucleation and continued growth. On the other hand, under growth control conditions the cooling rate is insufficient to bypass the nucleation onset completely so that some small fraction of crystallites may form initially, but the rapidly rising viscosity and falling growth rate with continued cooling near T_g prevents rapid cluster growth. In addition, the cluster size distribution that is retained overlaps in size with the critical nucleation size at T_x. In this case as indicated in Fig. 20, upon reheating a sample with pre-existing crystallites (i.e. quenched-in nuclei), rapid crystallization due to the development of quenched-in clusters as well as additional nucleation ensues at T_x which will essentially coincide with T_g.

While many of the early metallic glass alloys were synthesized under growth controlled conditions (i.e. marginal glass formers) [59] the primary crystallization particle densities in these alloys are of the order of 10^{18} m^{-3}. For the class of amorphous Al and Fe base glasses, the primary crystallization number densities range from 10^{21} up to almost 10^{23} m^{-3}. Both of the basic mechanisms for glass formation that are outlined in Fig. 20 can yield a high number density of nanocrystals upon devitrification. With nucleation control, nanostructure development can be achieved by controlled reheating, since the maximum in the growth rate typically occurs at a higher temperature than the maximum in the nucleation rate. The different synthesis routes that are shown in Fig. 20 originate from the relative time scale for the onset of nucleation and melt cooling. The transition from growth control to nucleation control can be achieved either by and increase in the cooling rate or by lengthening the time for onset of nucleation, t_n. Since t_n is related to atomic transport in the liquid, it is evident that liquids with high viscosity (i.e. strong liquids [55]) are favored for bulk glass formation. It is also apparent that t_n can be lengthened by removing active nucleation sites from the melt [60]. In fact, this is the basis for the effectiveness of melt fluxing, which has been shown to promote bulk glass formation. The actual mechanism for the development of the ultra high number densities of nanocrystals is under active study and proposals based upon homogeneous [61] and heterogeneous [57] nucleation and precursor phase separation reaction [62] are under examination.

The attainment of nanocrystal dispersions of essentially pure Al with ultrahigh number densities is a critical component of the attractive structural performance, but an equally important characteristic is the high thermal stability. One indication of this stability is the wide temperature range between the primary crystallization and final crystallization of between $75 - 100°\mathrm{C}$ in Fig. 19. Within this range, there is a metastable two-phase coexistence involving the Al nanocrystals and the surrounding amorphous matrix with limited coarsening of the microstructure. The sluggish kinetics is related at least in part to the large differences in component atom sizes and diffusivities as well as the onset of impingement of the diffusion fields from neighboring nanocrystals [54]. Indeed, even at a particle density of 10^{21} m^{-3} the average nanocrystal separation is only about 100 nm. It is also evident that in order for the Al nanocrystals to grow, there is a rejection of solute (i.e. TM and RE) as is typical for primary crystallization reactions. The low solute diffusivities, especially for the large RE atom, act to limit the growth and the transport is limited further by the reduced concentration gradient due to impingement as indicated by the asymmetric primary crystallization exotherm in Fig. 19. This kinetic restriction inhibits further nanocrystal growth and accounts for the asymmetric crystallization peak and the remarkable thermal stability. In fact, since the amorphous matrix composition will also be enriched in TM and RE components, it is possible to use the solute redistribution during primary crystallization to enhance the stability of the amorphous matrix (i.e. raise T_g)[63].

6 Summary

An important theme in many of the contemporary modeling strategies of materials behavior is the development of scaling relations. For example, specific property scales allow for density compensated comparisons of properties. Similarly, appropriate ratios of characteristic parameters are often useful in formulating indices or metrics (dimensionless groups) to gage different regimes of operating conditions or performance for particular design conditions. For nanostructured materials the key scaling is in terms of the specific interface area (A/V) that is often represented by an interface curvature. Within the scope of available coverage a number of cases or examples have been developed to illustrate the value of this scaling. The thermodynamics underlying this scaling is clear, but as the single digit nanometer scale is reached the macroscopic representation of properties that are implicit in the scaling, such as the interfacial energy, can no longer be expected to apply without some modification. The study of nanostructured materials is a rich area where there is probably much more to learn than what is known at present. This is also an area where computational approaches can have significant and immediate impact. The small number of atoms allows for an extended duration of analysis within the existing (but always growing) computer power. At the same time the critical role of direct experimental studies that has been responsible for many of the new discoveries of nanostructured materials behavior and has provided unique insight to understanding novel behavior, will continue to be at the forefront.

Acknowledgements

It is a pleasure to acknowledge the conference chairmen and in particular Jörg Weissmüller for organizing an inspiring meeting and for the invitation to present our work. The support of the US Army Research Office for my work in this area is most gratefully acknowledged.

References

1. H. Gleiter: Prog. Mat. Sci. **33**, 223 (1989)
2. C. C. Koch: Nanostruc. Mater. **2**, 109 (1993)
3. K. Lu: Mat. Sci. & Engr. Rep. **R16**, 161 (1996)
4. G. Martin, P. Bellon: Sol. State Physics **50**, 189 (1996)
5. G. Martin: Phys. Rev. **B30**, 1424 (1984)
6. G. Martin, P. Bellon, P Sisson: Defect Diffusion Forum **143-147**, 385 (1997)
7. W. L. Johnson: Progress in Materials Science **30**, 86 (1981)
8. D. R. Gaskell: *Introduction to the Thermodynamics of Materials*, 3rd edn (Taylor & Francis, London 1995)
9. M. Hillert: *Phase Equilibria, Phase Diagrams and Phase Transformations: Their Thermodynamic Basis*, (Cambridge University Press, Cambridge UK 1998)
10. D. A. Porter, K. E. Easterling: *Phase Transformations in Metals and Alloys*, (Chapman and Hall, New York 1992)
11. J. H. Perepezko, M. J. Uttörmark: Metall. Mater. Tran. **27A**, 533 (1996)
12. W. J. Boettinger, J. H. Perepezko: Fundamentals of Solidification at High Rates. In: *Rapidly Solidified Alloys: Processes, Structures, Properties, Applications*, ed by H. H. Lieberman (Marcel Decker Inc., New York 1993) pp 17–78
13. J. C. Baker, J. W. Cahn: *Solidification*, (ASM, Metals Park OH 1971) pp 23
14. J. H. Perepezko, W. J. Boettinger: In *Alloy Phase Diagrams*, ed by T. B. Massalski and B. C. Giessen (Elsevier, New York 1983) pp 223
15. J. H. Perepezko: Mater. Sci. & Engr. **A226-228**, 374 (1997)
16. J. H. Perepezko, M. H. da Silva Bassani, J. H. Park, A. S. Edelstein, R. K. Everett: Mat. Sci. & Engr. **A195**, 1 (1995)
17. J. H. Perepezko, G. J. Wilde: J. Non-Cryst. Solids **274**, 271 (2000)
18. W. J. Boettinger: In *Rapidly Solidified Amorphous and Crystalline Alloys*, ed by B. H. Kear, B. C. Giessen and M. Cohen (North-Holland, Amsterdam 1982) pp 15
19. T. B. Massalski: In *Proc. 4th Int. Conf. In Rapidly Quenched Metals*, ed by T. Masumoto and K. Suzuki (The Japan Institute of Metals, Sendai 1982) pp 203
20. J. M. McHale, J. M. Anroux, A. J. Perrotte, A. Navrotsky: Science **277**, 778 (1997)
21. R. C. Garrie: J. Phys. Chem. **69**, 1238 (1965)
22. N. L. Wu, T. F. Wu: J. Mater. Res. **16**, 666 (2001)

23. H. Zhang, J. F. Banfield: J. Mater. Chem. **8**, 2073 (1998)
24. J. K. Dewhurst, J. E. Lowther: Phys. Rev. **B57**, 741 (1998)
25. J. M. McHale, A. Navrotsky, A. J. Perrotte: J. Phys. Chem **B101**, 603 (1977)
26. J. Tersoff: Appl. Phys. Lett. **83**, 353 (2003)
27. R. S. Williams, G. Medeiros-Ribeiro, T. I. Kamins, D. A. A. Ohlberg: Ann. Rev. Phys. Chem. **51**, 527 (2000)
28. P. J. Desré: Acta Metall. Mater. **39**, 2309 (1991)
29. P. J. Desré, R. Yavari: Phys. Rev. Lett. **64**, 13 (1990)
30. A. M. Gusak: Ukr. Phys. **35**, 725 (1990)
31. A. M. Gusak, A. V. Nasarov: J. Phys. Condens. Matter **4**, 4753 (1992)
32. J. J. Hoyt, L. N. Brush: J. Appl. Phys. **78**, 1559 (1995)
33. K. R. Coffey, K. Barmak: Acta Metall. Mater. **42**, 2905 (1994)
34. J. Philiber: Def. and Diff. Forum **95-98**, 493 (1993)
35. H. J. Highmore, A. L. Greer, J. A. Leake, J. E. Evetts: Mater. Lett. **6**, 401 (1998)
36. C. V. Thompson: J. Mater. Res. **7**, 367 (1992)
37. J. W. Cahn: Acta Metall. **9**, 795 (1961)
38. J. W. Cahn: AIME **242**, 166 (1968)
39. W. C. Johnson, P. W. Voorhees: J. Stat. Phys. **95**, 1281 (1999)
40. A. Inoue: Prog. Mat. Sci. **43**, 365 (1998)
41. D. Turnbull: Metall. Trans. **12A**, 695 (1981)
42. U. Köster, U. Schünemann: In: *Rapidly Solidified Alloys: Processes, Structures, Properties, Applications*, ed by H. H. Liebermann (Marcel Decker Inc., New york 1993) pp 303–337
43. A. L. Greer: Science **267**, 1947 (1995)
44. A. L. Greer: Metall. Mater. Trans. **27A**, 549 (1996)
45. J. H. Perepezko, R. J. Hebert: JOM **54**, 34 (2002)
46. K. Hono: Prog. Mat. Sci. **47**, 621 (2002)
47. W. D. Kingery, H. K. Bowen, D. R. Uhlmann: *Introduction to Ceramics*, 2nd Ed. (J. Wiley & Sons, New York 1976)
48. K. C. Russell: Advances in Colloid and Interface Sciences **13**, 205 (1980)
49. F. Spaepen: Sol. State. Phys. **47**, 1 (1994)
50. R. K. Trivedi: Theory of Capillarity. In *Lectures on the Theory of Phase Transformations*, ed by H. I. Aaronson (TMS, Warrendale PA 1999) pp 135–165
51. K. Kelton: Sol. State Phys. **45**, 75 (1975)
52. J. W. Christian: *The Theory of Transformations in Metals and Alloys*, 2nd edn (Pergamon Press, Oxford UK 1995)
53. D. Kaschiev: Surf. Science **14**, 109 (1969)
54. D. R. Allen, J. C. Foley and J. H. Perepezko: Acta Mater **46**, 431 (1998)
55. C. A. Angell: Science **267**, 1924 (1995)
56. J. C. Foley, D. R. Allen, J. H. Perepezko: Scripta Mat. **35**, 655 (1996)
57. J. H. Perepezko, R. J. Hebert, W. S. Tong: Intermetallics **10**, 1079 (2002)
58. J. H. Perepezko, R. J. Hebert, R. I. Wu, G. Wilde: J. Non-Cryst. Sol. **317**, 52 (2003)
59. A. L. Greer: Acta Metall. **30**, 171 (1982)
60. H. W. Kui, D. Turnbull: Appl. Phys. Lett. **47**, 796 (1985)
61. S. Omata, T. Tanaka, T. Ispida, A. Sato, A. Inoue: Phil. Mag. **A76**, 387 (1997)
62. A. K. Gangopadhyay, T. K. Croat, K. Kelton: Acta Mater. **48**, 4035 (2000)
63. J. C. Foley, D. R. Allen, J. H. Perepezko: Mat. Sci. & Engr. **A226-228**, 569 (1997)