Simulation of single-electron devices

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Ph. D.

presented by
ANDREAS SCHOLZE
Dipl. Phys., University of Jena
born June 14, 1969
citizen of Germany

accepted on the recommendation of
Prof. Wolfgang Fichtner, examiner
Prof. Klaus Ensslin, co-examiner

2000
Seite Leer / Blank leaf
Acknowledgements

I wish to thank Prof. Wolfgang Fichtner for his guidance and assistance during my years at IIS, and Prof. Klaus Ensslin for reading and co-examining this thesis. Special thanks goes to PD Andreas Schenk for supervising the scientific aspects of this work.

Thanks to all members of the simulation group: Andreas Wettstein, Bernhard Schmithüsen, Bernd Witzigman, Andreas Witzig, Thomas Körner, Götz Leonhard, Christian Schuster, Fabian Bufler and Frank Geelhaar. I wish to express my gratitude for the benefits that I have gained from conversations with Michael Stopa (TU Munich), Massimo Macucci (University of Pisa) and Klaus Gärtner (WIAS Berlin). Furthermore, I wish to thank Thomas Heinzel (ETH Zürich) for his help with the experimental data and Diederick Fokkema (ISE Inc.) for his support with numerical software.

Also, I would like to thank the technical staff of IIS for providing excellent working conditions and all other people at the institute for their contribution to the excellent atmosphere that I enjoyed very much.

I would like to express my appreciation to Prof. Karl Hess who gave me the opportunity to spend a total of five months at the Beckman Institute of Advanced Science and Technology at the University of Illinois in Urbana-Champaign.

This work has been supported by the Swiss Priority Program in Micro and Nano System Technology (MINAST project 7.05 SIMNAD).
# Contents

## Preface

Preface

Preface (German)

## 1 Introduction

1.1 Evolution of single-electron devices

1.2 Modern semiconductor device simulation

1.3 Single-electron device simulation

1.4 Objectives

1.5 Layout of this work and general remarks

## 2 Introduction to single-electron tunneling and Coulomb blockade

2.1 Introduction

2.2 Orthodox theory of Coulomb blockade

2.2.1 The rules of the game

2.2.2 Single-electron box

2.2.3 Single-electron transistor

2.3 Constant interaction model

2.4 Example structures: vertical quantum dots
3 Theory of single-electron tunneling in semiconductors

3.1 Introduction .............................................. 29
3.2 Model definition ......................................... 30
3.3 Master equation approach ............................... 32
  3.3.1 Master equation .................................. 32
  3.3.2 Probability distributions and transition rates .... 33
  3.3.3 Detailed-balance equations ....................... 35
  3.3.4 Stationary current .................................. 36
3.4 Linear-response regime ................................. 37
  3.4.1 Linear-response conductance ...................... 37
  3.4.2 Limiting cases ..................................... 42
  3.4.3 Amplitude and lineshape ......................... 44

4 Transport coefficients ................................. 47

4.1 Introduction .............................................. 47
4.2 Equilibrium carrier densities .......................... 48
  4.2.1 Energy bandstructure and applied forces .......... 48
  4.2.2 Quasi-equilibrium approximation .................. 51
  4.2.3 Boundary conditions and Fermi level ............. 54
  4.2.4 Low dimensional electron gases ................... 55
4.3 Total free energies ..................................... 60
  4.3.1 Equilibrium thermodynamics ...................... 60
  4.3.2 Helmholtz free energy and external forces ....... 61
  4.3.3 Helmholtz free energy and charge quantization ... 63
4.3.4 Free energy minimization

4.4 Tunneling rates
   4.4.1 Bardeen's transfer Hamiltonian method
   4.4.2 Adiabatic theory for quantum dot systems

5 Numerics
   5.1 Introduction
   5.2 Space discretization
      5.2.1 Finite differences
      5.2.2 Discretization of $\nabla \cdot (\gamma \nabla)$
   5.3 Linear equation solving and eigenvalue problems
      5.3.1 Linear Poisson equation
      5.3.2 Schrödinger equation
   5.4 Nonlinear equation solving
      5.4.1 Introduction
      5.4.2 Newton-Raphson method
      5.4.3 Broyden's method
      5.4.4 Globally converging Newton schemes
   5.5 Self-consistent algorithms
      5.5.1 Predictor-corrector approach
      5.5.2 Algorithm implementation in SIMNAD
      5.5.3 Examples

6 Simulation examples
   6.1 GaAs/AlGaAs heterostructure
6.2 Device simulation ................................................. 98
   6.2.1 Geometry modeling and grid .................................. 98
   6.2.2 Surface and donors .......................................... 100
   6.2.3 Simulation results ........................................... 102
6.3 Comparison with measurements ................................. 115

7 Achievements, failures and prospects .......................... 119

A Thermodynamic properties of localized systems ............. 123

B Transition rates .................................................. 129
   B.1 Time-dependent perturbation theory .......................... 129
   B.2 Effective transition rates .................................... 132

C Density-functional theory ....................................... 135
   C.1 Hohenberg-Kohn theorem .................................... 136
   C.2 Kohn-Sham equation .......................................... 137
   C.3 Local-density approximation ................................. 139

Bibliography .......................................................... 141

Curriculum Vitae .................................................... 149
Preface

Devices utilizing the repulsion of additional electrons from tiny metallic or semiconducting islands (Coulomb blockade) in conjunction with correlated tunneling of single electrons from externally controlled reservoirs are called single-electron devices. In contrast to conventional field-effect transistors where switching is accomplished by electrostatically opening a channel and allowing for a macroscopic current to flow, the switching in single-electron transistors (SETs) is based on adding or removing a single electron from a quantum dot, which is the analogue to the channel. As such, single-electron devices represent the ultimate physical limit of miniaturization in electronics which is thought to be reached within the next decades.

This work is concerned with the numerical modeling of single-electron devices. The main focus is on the simulation of the local quantities such as the carrier densities and the electrostatic potential and the terminal characteristics of single-electron transistors.

After giving an introduction into the physics of single-electron devices by means of the so-called orthodox theory, the physical models used for the description of a SET driven in linear response are derived. The distribution functions and the transition rates for a quantum dot in equilibrium with the drain and the source reservoir and the linearization of the corresponding master equation is discussed in detail. The evaluation of the resulting equation for the linear-response conductance requires the calculation of the self-consistent many-body groundstate of a system of quantized electrons within the framework of density-functional theory as well as the calculation of the tunneling rates which are evaluated on the basis of Bardeen’s transfer-Hamiltonian formalism. Finally, the numerical methods used for the implementation of a simulation software, especially devised for the simulation of single-electron transistors, are discussed. Nonlinear solvers are of key importance for the proposed simulation approach, and therefore, much space is devoted to the quasi-Newton methods employed to solve the nonlinear Schrödinger/Poisson equation which results from carrier confinement in semiconductor quantum
dots, -wires, and -wells. An introduction into methods for linear-equation solving and eigenvalue problems at finite-difference grids is given.

Besides the focus of this work on basic models and methods for single-electron device simulation, the simulator was employed in a comprehensive study of a GaAs/AlGaAs heterostructure SET. The device was modeled using the realistic layer structure and the exact lateral gate geometry assuming various levels of quantization in the leads, the tunneling barriers and the quantum dot. Despite some discrepancies with the experimental data which are assumed to be due to an oversimplified model for narrow split-gate structures, the main features in the Coulomb-blockade characteristics could be explained on the basis of the self-consistent calculations. In addition to the conductance results, a discussion of the capacitances of the system is given on the basis of preprocessed data from the simulation output. Especially, the problem of a meaningful definition of a quantum dot self-capacitance was studied in more detail.

This work shows that the proposed method provides a physically sound and numerically stable framework for further investigations of single-electron transistors. The software implementation, for which the preliminary name SIMNAD (Simulation and Modeling of Nanodevices) is used, allows for the simulation of a variety of structures including GaAs/AlGaAs heterostructures and silicon SETs. However, for increased predictive capabilities more investigations into various problems including the treatment of narrow split-gate structures on active surfaces or the role of discrete doping centers, for instance, is certainly mandatory.
Einleitung


Diese Arbeit zeigt, dass die vorgeschlagene Methode einen physikalisch korrekten und numerisch stabilen Rahmen für zukünftige Untersuchungen an Einzelelektronen-Transistoren bietet. Die Implementierung, für die der vorläufige Name SIMNAD (Simulation and Modeling of Nanodevices) verwendet wird, erlaubt die Simulation einer Vielzahl von Strukturen, einschließlich von GaAs/AlGaAs Heterostrukturen und Silizium Einzelelektronen-Transistoren. Um eine bessere Übereinstimmung mit experimentellen Daten zu erreichen, müssen allerdings weitere Untersuchungen zu verschiedenen Problemen, wie zum Beispiel der Behandlung eng beieinander liegender Kontakte auf aktiven Oberflächen oder der Rolle diskreter Dotierzentren, vorgenommen werden.
Chapter 1

E stupor m'eran le cose non conte.
My mind was stunned by what it did not know.

Dante, »Purgatorio, XV«

Introduction

1.1 Evolution of single-electron devices

Since the invention of the transistor by John Bardeen and William Shockley in 1948 constant miniaturization has lead to devices with dimensions well below the 100 nm limit. Originally designed to simply emulate the vacuum tube, the transistor soon proved to provide a lot more potential for speed, power consumption and reliability. In its most widespread configuration as metal-oxide-semiconductor field-effect transistor (MOSFET) these potentials were fully exploited and initialized what we now call the information age.

The relentless quest for device miniaturization is something that has been with semiconductor technology since its early evolution starting in the 1950s. This search is driven by a few facts which became obvious at the latest when the first steps towards integrated circuits where made in the late 1950s. First, the reduced volume of a miniaturized transistor allows for lower power dissipation. Second, the consequently lower capacitances and shorter interconnects lead to faster systems. Finally, relative costs drop with more devices on ever-increasing chip sizes. According to the latest "road-map" [1], chips containing 100 million transistors will start to reach the markets within the next three years.

The end of the millennium sees the silicon-based semiconductor industry
well at the height of its evolution with even further transistor miniaturization ahead. But nevertheless, the scaling limits for the classical MOSFET structure are already beginning to show. Transistors with channel lengths of a few tens of nanometers are predicted to be governed by the quantized nature of the carriers. Besides the efforts to overcome the limitations intrinsic to the classical MOSFET structure, there is an intense search for new device principles that not only allow for the scaling limits to be pushed even further but make use of the quantized nature of the charge carriers evident in small structures.

Since the 1980s, developments in both semiconductor technology and theory have lead to a completely new field of research focusing on devices whose operation is based on the discrete nature of electrons tunneling through thin potential barriers. These devices which exhibit charging effects including Coulomb blockade are referred to as single-electron devices. The potential of the single-electron concept is rather breathtaking. In principle, it should be possible to operate in the limit of one transferred carrier per bit, the ultimate limit of semiconductor based logic. However, enormous obstacles have to be overcome. Semiconductor technology for single-electron devices requires the reproducible manufacturing of very small conducting islands and their accurate positioning with respect to external electrodes. It was not before the mid 1980s that especially lithography technology reached the deep sub-micron range making structuring in the 100 nm range possible. In the mean time, single-electronics has matured and a variety of concepts and device architectures is available. A typical device structure is shown in Fig. 1.1. As are most structures in the field, this device is based on the III/V-compound material

![GaAs/AlGaAs heterostructure quantum dot. The metal gates are structured using electron beam lithography (Stanford University).](image)
system GaAs/AlGaAs which has the advantage of providing a very clean inter-
face, therefore being especially suitable for heterostructure devices. For a long time, semiconductor heterostructures have been used to achieve two-
dimensional electron gases (2DEG) in HEMTs (high-electron-mobility tran-
sistor). Heterostructure single-electron transistors are related to HEMTs, how-
ever, with the difference that the 2DEG is patterned by electrostatic contacts
(Schottky contacts) at the top of the device in order to create a small charged
island, the so called quantum dot which is separated from the source and drain
regions by tunnel barriers. The electron gas in the quantum dot is essentially
zero-dimensional (0DEG) exhibiting properties similar to atoms. This led to
the term artificial atoms for this kind of structure [2]. GaAs/AlGaAs single-
electron transistor (SET) prototypes were build as early as 1987. These de-
vices had islands several hundreds of nanometers across because of techno-
logical limitations. Therefore, they had to be cooled to temperatures below
1 K to make observation of single-electron charging effects possible. Devices,
however, whose functionality can only be guaranteed at cryostatic temper-
atures are somewhat impractical for real-world applications. Consequently,
one of the major driving forces behind the development of single-electron de-
vices was the quest for room temperature operation which requires the island
sizes to be in the sub-100 nm range.

Just recently, the first single-electron transistors were realized in silicon. Silicon is the basic material of modern semiconductor technology, and there-

![Figure 1.2: Single-electron transistor in SOI technology. Left: Schematic of the device structure. Right: SEM micro-graph of the channel after oxidation (Princeton University).](image)
fore, provides one of the most promising paths towards industrial application of the single-electron concept for memory- and switching devices. Figure 1.2 shows one of the most prominent structures processed in silicon-on-insulator (SOI) technology. The central island is formed by a 10 nm silicon grain which is reduced to its final size by oxidation. The size of the island allows the observation of the Coulomb blockade of conductance at in general higher temperatures as in GaAs/AlGaAs structures [3].

A similar technology was used for the fabrication of a Si single-electron transistor memory device [4] with a self-aligned quantum-dot floating gate (Fig. 1.3 left). Due to the self-alignment, the floating gate is accurately positioned above the channel wire. If additional charges tunnel from the channel onto the floating gate, the threshold voltage shifts and a memory effect similar to classical floating gate transistors is observed. This device concept was even able to exhibit room temperature operation which is a prerequisite for further industrial application [4],[5]. Figure 1.3 (right) shows a side view of a similar structure [6]. Other concepts pursued comprise multiple nano-crystals coupled to the channel of a conventional MOSFET [7] which is especially interesting, since it could minimize offset charge effects.

The future of the single-electron concept is by no means clear. Several problems are not solved yet. The major flaws of single-electron technology are

**Figure 1.3:** SEM micro-graphs of quantum dot flash memory devices. *Left:* Self-aligned floating gate on a narrow channel FET (Fujitsu). *Right:* View of a quantum dot flash memory device along the channel (IBM).
1.2 Modern semiconductor device simulation

From the very beginning of semiconductor technology it was thought that numerical, physics-based analysis of devices could help a great deal in their understanding. Nowadays, simulation and modeling of semiconductor devices both at the process and the device level has become one of the most important development methodologies in industry and eventually has developed into an ever-growing industry itself. Complete TCAD (technology-computer-aided-design) suites, that include process-simulation- and device-simulation tools connected to advanced meshing- and optimization software are readily avail-

the sensitivity to random background charges, the necessity for lower junction capacitance to raise operation temperatures, and the sensitivity to fabrication parameters.

How exactly SET devices can be used as building blocks for further logic circuits is yet another unsolved problem. Even though several logic schemes have been proposed, none is presently threatening to oust the semiconductor standard CMOS (complementary metal-oxide-semiconductor) technology. Some of the schemes, however, are similar to CMOS technology where bits are represented by voltage levels. Nevertheless, it remains unclear whether these devices have significant gain. Other concepts are based on a bit representation by charges. An intriguing, but as-yet-unrealized proposal in this direction is the quantum-cellular automata (QCA) concept [8]. Mesoscopic cells of quantum dots can be polarized, and since the polarization of a cell can effect that of a nearby cell, a polarization signal can be propagated along a chain of cells. Together with principle problems regarding the read-in or read-out of polarization states and the stability of the states it is not clear that there will be any speed or density advantage and low temperatures seem to be mandatory.

It is clear that silicon MOSFETs have their limits. When these limits will be reached is not so obvious. The demise of silicon technology has been predicted (and postponed) many times in the last decades. At the same time, semiconductor companies have grown increasingly reluctant to work on new technologies whose potential might be exploitable in a far future. In any case, miniaturization will remain an important aspect of VLSI circuits and single-electron related effects will almost certainly plague, and perhaps, eventually revolutionize the semiconductor industry in the next decades.
All of the device simulation tools available today are essentially partial-differential equation (PDE) solvers, that are specialized to solve a particular system of PDEs relevant to semiconductor modeling. These so-called semiconductor (device-) equations describe the static and dynamic behavior of carriers in semiconductors under the influence of external fields that cause deviations from the thermal equilibrium [9]. The system of equations known as drift-diffusion equations consist of the Poisson equation

$$\nabla \cdot (\epsilon \nabla \phi) = -q(p - n + C),$$

the continuity equations for electrons and holes

$$\nabla \mathbf{J}_n - q \frac{\partial n}{\partial t} = -q R_n$$
$$\nabla \mathbf{J}_p + q \frac{\partial p}{\partial t} = q R_p$$

and the current-density equations for electrons and holes

$$\mathbf{J}_n = -q \mu_n n \nabla \phi + q D_n \nabla n$$
$$\mathbf{J}_p = -q \mu_p p \nabla \phi - q D_p \nabla p,$$

where $\phi$ is the electrostatic potential, $n$ and $p$ are the electron- and hole densities, respectively, $C$ is the total density of ionized donors, $\mathbf{J}_n$ and $\mathbf{J}_p$ are the current densities and $R_n$ and $R_p$ can be understood as functions describing generation and recombination of carriers. The material dependent parameters are the permittivity $\epsilon$, the mobilities $\mu_n$ and $\mu_p$, and the carrier diffusion constants $D_n$ and $D_p$. Numerical solutions of this system of equations, maybe extended by the energy balance equations with the electron and hole temperature as variables (hydrodynamic model), is the main goal of classical device simulation.

The drift-diffusion equations are macroscopic equations which describe the flow of charged particles using appropriate models for their macroscopic physical behavior. The most important physical models are for the densities of states of electrons and holes which are based on a bandstructure model, the carrier mobilities $\mu_n$ and $\mu_p$, and the generation/recombination rates $R_n$ and $R_p$. The semi-classical models for the densities of states of electrons and holes are based on the assumption that the carriers are freely moving particles in parabolic bands with an effective mass $m^*$ obeying Fermi-Dirac statistics.
This simple assumption, however, is not sufficient for carriers in narrow MOSFET channels for instance. The local carrier densities are significantly lower than the classically predicted ones. Needed is a quantum treatment of the carriers, i.e. the solution of a (usually one-dimensional) Schrödinger equation [10]. Inclusion of quantum effects is still a conceptually and numerically difficult problem due to the nonlocal relationship between the potential and the quantum mechanically calculated charge density [11].

The carrier mobilities $\mu_n$ and $\mu_p$ and the generation/recombination rates $R_n$ and $R_p$ should be calculated on the basis of physical models. In most cases this is a difficult issue and the agreement of the microscopic model with the experimental data is sometimes rather poor. Common is the application of simple analytical functions containing various adjustable parameters which guarantee the required fit and cause no numerical problems. However, the drawback is obvious. Since a multitude of parameters with no apparent physical meaning is available for tuning the functions, almost everything can be calculated and almost no certainty about the validity of the results can be given. Therefore, the development of advanced physical models for the parameters in the device equations is a major quest for nowadays device simulation [12].

A last point should be made. Much of the success of device simulation is due to the underlying process simulation. Process simulation is the first step in modern device modeling. The particular steps in the fabrication of a semiconductor device (the semiconductor process) such as deposition, etching and implantation of dopands are simulated. The result is a representation of the geometry and a doping (-density) profile which is subsequently used to automatically generate a simulation mesh by means of a mesh-generation software. Process simulation is in particularly demanding since problems as moving boundaries are difficult to handle, and instabilities of the simulation due to the mesh quality are still a rather common problem. Alternatively, the mesh has to be generated manually using a mesh-editor. The disadvantage of this procedure is that complicated three-dimensional meshes are hard to generate by hand and a control over the mesh quality is difficult to achieve. Also, manual mesh generation makes complex optimization of device geometries within the scope of TCAD experiments (DoE, design-of-experiments) almost impossible.
1.3 Single-electron device simulation

The drift-diffusion equations described in the previous section provide no appropriate mathematical model for single-electron devices. Here, the transport is dominated by the quantized nature of the tunneling electrons. If the transport description for an electronic device requires the inclusion of quantum-mechanical properties of the carriers, the choice is between two models (Fig. 1.4). Dynamic models on one hand are based on equations of motion for Green's functions or for quantum-phase-space distributions such as Wigner functions. Kinetic models on the other hand use time-independent representations of the quantum-mechanical states solving kinetic- or rate-equations.

Transport in arbitrary single-electron circuits consisting of tunnel junctions, capacitors, and voltage sources is described by a rate equation for the stochastic probability distribution \( P(M) \)

\[
\frac{d}{dt} P(K) = \sum_{M \neq K} \left[ P(M)W(M, K) - P(K)W(K, M) \right],
\]

where \( \sum_K P(K) = 1 \) and \( W(M, K) \) is the rate for the transition from state \( M \) to state \( K \) [20]. The state variable is the electron number. For a given state of the system the transition rates have to be calculated according to system parameters. An iterative method based on solution of Eq. (1.2) for circuits containing up to a few tens of islands was introduced by Fonseca et al. [14]. An alternative approach is based on a Monte-Carlo method [13]. A commercially available program that combines the master equation approach with the Monte-Carlo approach is SIMON from TU Vienna [15].

All these methods use the results of the phenomenological orthodox theory which treats single-electron transitions between purely capacitively coupled islands (metallic grains) or its extension towards confined levels, the constant-interaction model (an account of these models is given in Chapter 2). The advantage of these methods is the possibility to evaluate complex circuits containing a multitude of islands and tunneling junctions. Even cotunneling can be included in the calculations. The method is only slowed by the effort that is invested in calculating the transition rates and the probability distributions. Most of the time, analytical capacitance results are used which retain the simplicity needed for the numerical evaluation. The accuracy strongly depends on the physical models which are used to calculate the capacitances and the tunneling rates.
1.3. Single-electron device simulation

Figure 1.4: Classification of methods for numerical simulations of quantum transport in electronic devices.

For semiconductor structures the above mentioned methods are not accurate enough, especially if a discrete energy spectrum is present at a small island (then often referred to as quantum dot). A self-consistent description of the confinement potential is needed and methods based on density-functional theory (DFT) are used. The screening of the carriers is treated by solving an nonlinear (Schrödinger-) Poisson equation for a realistic geometry. This method allows for a detailed analysis of single-electron charging and the calculation of the exact shapes and peak heights of the Coulomb-blockade peaks[41].

A drawback of the self-consistent models is the numerical effort. Such calculations are extremely costly, especially if employed for 3D geometries. As such they are mostly limited to a single solid state device, i. e. to a single-electron transistor for which an analytical solution of the transport equation, Eq. (1.2), in the linear-response regime with respect to the junction voltage exists [26]. However, the hope is that device simulations can help to extract parameters such as the capacitances for the analysis of more complex circuits using stochastic methods. This is the classical path in modern semiconductor
device simulation. The first step is the actual device simulation using a self-consistent solid-state simulator. Then, in a second step, a parameterized model (SPICE-model for instance) is extracted and used in a circuit simulation for a more complex circuit environment.

1.4 Objectives

This work is concerned with the theoretical outlay and the numerical implementation of a simulation software, designed to simulate single-electron transistors. The developed physical models and numerical algorithms, however, are in principle applicable in the much wider context of nanodevice modeling and simulation.

Two main objectives had to be reached. First, concepts from conventional TCAD had to be advanced towards the nano-scale range, including solutions for the full quantum mechanical equations that describe the properties of electrons in quantum dots and quantum wires. Second, a sound description of electron transport in nanostructures, including the tunneling of particles through arbitrarily shaped potential barriers, separating the quantum dot from the macroscopic regions of the device had to be developed.

The approach chosen is based on a self-consistent analysis of the charge densities and the potential landscape in (almost) arbitrarily shaped semiconductor structures. Zero-dimensional confinement in semiconductor quantum dots can be simulated in conjunction with one- and two-dimensional confinement in quantum wells and quantum wires. This framework is applied to single-electron transistors by simulating the quantum dot using various models for the charge density of zero-dimensional electron gases. The charge density models for quantum wires- and quantum wells are applied to the source and the drain reservoir. Using the results from the self-consistent structure calculation, the tunneling rates through electrostatically formed barriers are computed. With the Helmholtz free energy which is calculated from the potential and the charge distribution in the device, the linear-response conductance is computed.

All these concepts had to be implemented in a both advanced and numerically stable simulation software. State-of-the-art algorithms for nonlinear equation solving had to be pushed even further to include the quantum mechanical equations necessary to describe a single-electron device. The valid-
ity of the simulation approach is proved by simulating example devices. A GaAs/AlGaAs SET has been modeled and first calculations of the electronic structure and the conductance characteristic have been performed. The results are compared with experimental findings.

Finally, the software was embedded in the TCAD suite of ISE AG Zürich. This enclosed the use of the ISE data structures in order to utilize the meshing and visualization tools part of the ISE TCAD suite.

1.5 Layout of this work and general remarks

This work has been organized in the following manner: In Chapter 2 an introduction to single-electron tunneling based on the orthodox theory for metallic grains with extensions towards finite temperatures and confined states in semiconductor structures is given.

Chapter 3 concentrates on the theory of single-electron tunneling in semiconductors. Based on a linearized master equation approach, a formula for the linear-response conductance of single-electron transistors is derived and limiting cases are discussed.

Chapter 4 is devoted to the discussion of the two main transport coefficients that enter the transport equation derived in the antecedent chapter. The first coefficient is the Helmholtz free energy which is used to evaluate the Gibbs distribution for the quantum dot occupations. The second is the tunneling rate for single-barrier tunneling, which is calculated using Bardeen's transfer-Hamiltonian formalism.

Chapter 5 gives an account of the numerical methods used. Both linear- and nonlinear solvers are needed for the solution of a nonlinear Schrödinger- Poisson-type equation in a discretized form. The main focus lies on globally converging Newton-Raphson methods that are used to solve the nonlinear systems of equations involved.

Chapter 6 discusses the functioning of the self-consistent simulation approach at a particular example, a GaAs/AlGaAs heterostructure SET.

In Appendix A, the Gibbs distribution for localized systems under the influence of an external reservoir is derived. The equations for the transition rates in single-electron tunneling employing time-dependent perturbation the-
ory are derived in Appendix B. Appendix C gives an overview over the foundations of density-functional theory and the local-density approximation.

Finally, some remarks have to be made: The SI (MKSA) system of units is used throughout the work. However, one exception is made. Densities are given in cm$^{-3}$ in order not to deviate too much from well established common usage and nm is the distance unit. For all constants such as the Boltzmann constant, Planck constant or any other constant all units are converted to MKSA units, i. e. cm, nm $\rightarrow$ m.
Chapter 2

Qui si convien lasciare ogni sospetto;
ogni viltà convien che qui sia morta.
Now here you must leave all distrust behind
let all your cowardice die on this spot.

DANTE, »Inferno, III«

Introduction to
single-electron tunneling and
Coulomb blockade

2.1 Introduction

The evolution of a consistent theory for single-electron tunneling goes back to the early 1950s. An unusual dependence of the resistance of thin metal films was explained by Gorter as due to the charging of small metallic grains [16]. Transferring electrons through the barriers between the grains would result in charging of the neighboring grains, and therefore, in an increase of the electrostatic energy by

\[ E_C = \frac{e^2}{2C}, \]

where \( C \) is the effective capacitance of the island. \( E_C \) became later known as the Coulomb charging energy. This energy suppresses electron transfer unless
it is overcome by either thermal excitations at a temperature $T$, i. e.

$$T \sim T_0 = \frac{E_C}{k_B}$$  \hspace{1cm} (2.2)

or by an externally applied voltage $V$, i. e.

$$V \sim V_t = \frac{E_C}{e} = \frac{e}{2C}.$$  \hspace{1cm} (2.3)

For this suppression of electron transport a new term was coined: *Coulomb blockade* (CB) of tunneling. Independently from Gorter, this effect was observed and explained by Neugebauer and Webb [17]. Yet another effect, *Coulomb exclusion* became known in the late 1960s through the work of Giaever and Zeller [18] and Lambe and Jaklevic [19]. If the voltage $V$ is less then the threshold voltage $V_t$, the system is in the Coulomb blockade state. With the voltage $V$ exceeding $V_t$, Eq. (2.3), an electron tunnels through the barrier into the grain. However, because of the Coulomb repulsion it can not be followed by another electron unless the externally applied voltage $V$ is increased even further by $\Delta V = e/C$. This can readily be seen by looking at the electrostatic energy of such a grain

$$U = \frac{Q^2}{2C} - QV,$$  \hspace{1cm} (2.4)

with the discrete charge $Q = -Ne$ and the externally applied voltage $V$. At $T < T_0$ the electron number is given at the minimum electrostatic energy and $N(V)$ is a staircase-like function commonly known as the *Coulomb staircase*. At temperatures of order of $T_0$ the staircase which otherwise exhibits sharp corners is washed out, totally vanishing above $T_0$.

The first quantitative theory for two barrier systems was given by Kulik and Shekhter in 1975 [20]. They used an equation similar to Eq. (2.4), treating tunneling as a perturbation to the equilibrium Hamiltonian. A master equation for the time evolution of probabilities $P(N)$ which contains all relevant information of the system behavior was derived. This approach is the basis of the so called *orthodox theory* of single-electron tunneling. The theory of correlated tunneling as is used now was eventually developed by Averin and Likharev [21] in the mid-1980s and almost immediately supported by experiments on single-electron transistors [22]. In the early 1990s the orthodox theory was generalized towards the inclusion of effects of both charge and energy conservation in semiconductor structures. Based on the pioneering papers of Meir *et al.* [23], Averin *et al.* [24] and Beenakker [26] a consistent theory that explains the principle features in the characteristics of single-electron tunneling
2.2 Orthodox theory of Coulomb blockade

2.2.1 The rules of the game

A theory of single-electron tunneling has as its subject the charging of a small island (grain, quantum dot etc.). The orthodox theory focuses on the charging of metallic islands with diameters $d \geq 100$ nm. The capacitance of such an island is given by the equation for a conducting sphere $C = \frac{2\pi\varepsilon_0 d}{e}$ and can be estimated to be $\geq 5$ aF. The resistance of the tunneling barrier $R_t$ has to be above the quasi-classical limit

$$R_t \gg R_Q = \hbar\pi/2e^2 \simeq 6.4 \text{ k}\Omega,$$

where $R_Q$ is the resistance quantum. This can be understood considering the tunneling rate $\Gamma = V/eR_t$ at some bias voltage $V$. The tunneling time between two subsequent tunneling events is $\tau = \Gamma^{-1} = eR_t/V$. The term tunneling time refers to the characteristic time of junction relaxation, not the duration of the actual tunneling event itself, which is of order of $\tau_t = \Gamma_t^{-1} = 10^{-15}$ s. The uncertainty principle states the duration of the tunneling event according to $\tau_t eV \geq \hbar$. Hence, Eq. (2.5) means that $\tau_t \leq \tau$, e. i. tunneling events are well separated in time and one speaks of sequential tunneling of single electrons. Essentially, this condition ensures that the wavefunction of an island electron is well localized there. In systems where the tunneling resistance $R_t$ is small compared to the scale provided by $R_Q$ suppression of single-electron charging occurs and delocalized states are available for charge transport.¹

¹Eq. (2.5) ensures localization of electrons at the conducting island. Barriers with high transparency effectively suppress the localization and make single-electron manipulation possible.
Figure 2.1: Plot of the parameters for single-electron tunneling parameterized by the tunneling frequencies $f = \tau^{-1} = (R_tC)^{-1}$. Room temperature operation is only possible in the lower right-hand side field between the dashed lines.

As already stated, the temperature must be small enough that the charging energy exceeds the energy of thermal excitations. From Eq. (2.1) and Eq. (2.2) it can readily be seen that the Coulomb blockade regime is present for

$$k_B T \ll \frac{e^2}{2C}.$$ (2.6)

This condition can be fulfilled either by lowering the temperature or by decreasing the capacitance which means to reduce the island size. Usually, experiments are performed at temperatures of a few mK and for structures with island sizes of a few hundred nanometers. Figure 2.1 shows a plot of the critical parameters for single-electron tunneling. For room temperature operation capacitances in the aF-range and island sizes in the sub 100 nm range (silicon: $d \leq 5$ nm) are required. These limitations hindered experiments and detailed evaluation of the orthodox theory especially in semiconductors for a long time. Nowadays, with advanced nanofabrication techniques at hand it is possible to structure islands of known geometry which are separated by

This is in contrast to the field of "quantum-electronic devices" which operate above the limit given by the resistance quantum. Of course, single-electron transistors use quantum properties of matter, but so do conventional bipolar or field-effect transistors (Likharev [25]).
well defined barriers and can be acted upon by leads connecting the island to voltage controlled reservoirs.

The small diameters of single-electron transistors make them candidate-devices for further ultra-dense logic. However, care has to be taken with respect to power consumption/dissipation. The power $p$ of switching in single-electron devices is given by the charging energy, Eq. (2.1), divided by the switching time $\tau = R_tC$

$$p = \frac{E_C}{\tau} = \frac{e^2}{2R_tC^2}. \tag{2.7}$$

Assuming that silicon-based technology will still be the technology of future electronic devices, one can take the "road-map" [1] estimate for the year 2012 of $10^{12}$ transistors/chip realized with the same number of single-electron transistors. According to Eq. (2.7) a room temperature operating SET would have a power consumption of $p \sim 10^{-8}$ W ($R_t = 25R_Q = 160$ kΩ and $C = 2$ aF). This would mean a total of $10^4$ W/chip has to be dissipated if all transistors are switched at the same time. With lower tunnel resistances $R_t$ and therefore higher operational frequencies (Fig. 2.1) the power consumption is increasing linearly. A decrease of the capacitance even leads to a quadratic increase of the power consumption as can be seen from Eq. (2.7).

A last very important point is to be made. At temperatures close to zero Kelvin the charge at the island is an integer multiple of $e$. In such systems, tunneling events which occur via two tunneling junctions are not independent anymore. The tunneling events are correlated by the need to keep the island charge an integer multiple of $e$. However, each of the charges at the two junctions can be fractional which makes it possible to tune the junction by external gating.

### 2.2.2 Single-electron box

Charging and de-charging of a single island is related to changes in the Helmholtz free energy $F$ of the system. Within the orthodox model the free energy of a single-electron box (Fig. 2.2) under the influence of an external gate is written in terms of macroscopic charges

$$F(N) = \int_0^{-Ne} dQ \phi(Q), \tag{2.8}$$
where \( \phi(Q) \) is the potential at the single-electron node and \( Q \) the discrete charge. A system of capacitors is described by a capacitance matrix, which relates potentials and charges at the circuit. For the single-electron box as in Fig. 2.2 this is

\[
\begin{pmatrix}
Q \\
\phi
\end{pmatrix} = C \begin{pmatrix}
V_g \\
\phi
\end{pmatrix}, \quad \text{width} \quad C = \begin{pmatrix}
C_\Sigma & -C \\
-C_g & C_\Sigma
\end{pmatrix}.
\tag{2.9}
\]

Rewriting of Eq. (2.9) in terms of the unknown quantities \((Q_g, \phi)\) leads to

\[
\begin{pmatrix}
Q_g \\
\phi
\end{pmatrix} = \begin{pmatrix}
C_\Sigma - CC_g C_\Sigma^{-1} & -CC_\Sigma^{-1} \\
C_g C_\Sigma^{-1} & C_\Sigma^{-1}
\end{pmatrix} \begin{pmatrix}
V_g \\
Q
\end{pmatrix}, \tag{2.10}
\]

which determines the potential at the single-electron node \( \phi \). The gate capacitance is \( C_g \), and \( V_g \) is the gate voltage. The total capacitance \( C_\Sigma \) which is the sum of all capacitances \( C_\Sigma = C + C_g \) is assumed to be independent of \( N \). With Eq. (2.8) the free energy takes the form

\[
F(N) = \frac{(Ne)^2}{2C_\Sigma} - Ne \frac{C_g}{C_\Sigma} \frac{V_g}{2}. \tag{2.11}
\]

The first term on the right-hand side represents the electrostatic or charging energy \( U(N) \) of \( N \) electrons corresponding to Eq. (2.1). The second term gives the work (potential energy) done by the voltage source. It is assumed that only one gate has significant capacitive coupling to the island. An external charge \( Q_{ext} = C_g \frac{V_g}{2} \) is introduced which can be understood as an externally induced polarization charge which is bound to the gate field and can be varied continuously by means of a gate voltage in contrast to \( Q \) which is restricted to integer multiples of \( e \). In terms of \( Q_{ext} \) the free energy can be written as

\[
F(N) = \frac{(Ne - Q_{ext})^2}{2C_\Sigma} - \frac{Q_{ext}^2}{2C_\Sigma}. \tag{2.12}
\]

The last term in Eq. (2.12) is again independent of \( N \). Figure 2.3 displays the free energy variations with changing gate voltage. The probability of finding \( N \) electrons at the island is given by

\[
P(N) = Z^{-1} \exp(-\beta[F(N) - EFN]), \tag{2.13}
\]
2.2. Orthodox theory of Coulomb blockade

where $Z$ is the $N$-independent partition function, $E_F$ is the Fermi energy of the reservoir which can be related to the ground contact of the circuit. Having a net current flowing through the device, the allowed numbers of electrons at the island must be fluctuating between $N$ and $N - 1$. The net number of electrons is then $N - 1/2$ (Fig. 2.3). This means that there are non-zero probabilities for two different numbers of electrons, i.e. $P(N) = P(N - 1)$. Midway between these degeneracy points adding (or removing) an electron moves the free energy away from the minimum. This is equivalent to the condition $F(N) - F(N - 1) \equiv \mu(N) = E_F$ (energy-conservation law) for the entire system. Substitution of Eq. (2.12) into this condition gives

$$E_F^* = \left( N - \frac{1}{2} \right) \frac{e^2}{C} + e \frac{Q_{\text{ext}}}{C},$$  

(2.14)

where $E_F^*$ is defined as the renormalized Fermi energy of the island. If one neglects changes in the Fermi energy, the constant spacing of the island charge increments by one is simply $\Delta V_g = e/C_g$.

\[ \text{Figure 2.3: Free energies (top): as the gate voltage is increased the charge } Q \text{ for which the energy is minimized changes from } -(N - 1)e \text{ to } -(N - 1/4)e. \]

\[ \text{Fermi energies (bottom): when } Q = -(N - 1/2)e \text{ the gap in the Fermi energies vanishes and a current can flow (after [2]).} \]
2.2.3 Single-electron transistor

Now consider a simple tunnel junction circuit containing one island (Fig. 2.4). The island is located between two tunnel junctions with capacities $C_1$ and $C_2$. It can electrostatically be controlled by the gate capacitance $C_g$ (Fig. 2.4(b)). The total capacitance of the central island is $C_\Sigma = C_1 + C_2 + C_g$. Such a setup is called a single-electron transistor. The device is reminiscent of a MOSFET, but with a small conducting island embedded between two tunnel junctions, instead of the usual inversion channel.

First, a reduced circuit that contains only the basic parts of the SET, the island and the two tunnel junctions in series (Fig. 2.4(a)) is considered. Simple circuit analysis reveals that the voltage drops across the junctions are

$$V_1 = \frac{C_2 V + Ne - Q_0}{C_\Sigma} \quad \text{and} \quad V_2 = \frac{C_1 V - Ne + Q_0}{C_\Sigma}, \quad (2.15)$$

where $V = V_1 + V_2$ and the island charge is $Q = Q_2 - Q_1 + Q_0 = -Ne + Q_0$. The charges at the two capacitors $Q_1$ and $Q_2$ are given as $Q_1 = C_1 V_1$ and $Q_2 = C_2 V_2$, and $Q_0$ is some background charge which is due to polarization of the island related to charged defects in the vicinity (not contained in the circuit Fig. 2.4). The electrostatic energy is given as

$$U(N) = \frac{C_1 C_2 V^2 + (Ne - Q_0)^2}{2C_\Sigma}, \quad (2.16)$$

and the work done by the voltage source is

![Figure 2.4: Equivalence circuit for a SET structure: (a) basic circuit consisting of two tunnel junctions, (b) capacitively coupled SET. The dashed lines mark the island.](image-url)
2.2. Orthodox theory of Coulomb blockade

\[ W(N_1, N_2) = \frac{(N_1C_2 + N_2C_1)eV}{C_\Sigma}, \] (2.17)

where \( N_1 \) and \( N_2 \) are the electron numbers at the junctions with \( N = N_1 - N_2 \). Additional control of the island charge is obtained by introducing a gate capacitance \( C_g \) (Fig. 2.4(b)). The free energy is calculated using above equations with the substitutions \( C_1 \to C_1 + \alpha C_g, C_2 \to C_2 + (1 - \alpha)C_g \) and \( Q_0 \to Q_0 + C_g(V_g - \alpha V) \equiv Q_{\text{ext}}. \alpha \) is a factor depending on the circuit. For the circuit in question the choice is \( \alpha = 1 \). Consequently, the free energy of the capacitively gated single-electron transistor is

\[ F(N_1, N_2) = \frac{(C_1 + C_g)C_2V^2 + (Ne - Q_{\text{ext}})^2}{2C_\Sigma} - \frac{[N_1C_2 + N_2(C_1 + C_g)]eV}{C_\Sigma}. \] (2.18)

The condition for the transition from a state with higher energy to a state with lower energy can be evaluated. The change in free energy for an electron tunneling through either junction is

\[ \Delta F(N_1 \pm 1, N_2) = \frac{e}{C_\Sigma} \left[ \frac{e}{2} \pm (Ne - C_gV_g - Q_0 + C_2V) \right] \]

\[ \Delta F(N_1, N_2 \pm 1) = \frac{e}{C_\Sigma} \left[ \frac{e}{2} \pm (-Ne + C_gV_g + Q_0 + (C_1 + C_g)V) \right]. \]

At zero temperature only transitions with \( \Delta F < 0 \) are allowed. Evaluation of

\[ \text{Slope} = \frac{C_g}{C_1 + C_g} \quad \text{Slope} = \frac{e}{C_\Sigma} \quad \text{Slope} = -\frac{C_g}{C_2} \]

\[ \text{Slope} = \frac{e}{C_\Sigma} \]

**Figure 2.5:** Stability diagram for the single-electron transistor. Stable operation is only possible in the shaded regions. The system is assumed to be free of background charges, i.e. \( Q_0 = 0 \).
this condition leads to the stability diagram given in Fig. 2.5. Stable operation is only possible for values of $V$ and $V_g$ within the shaded regions.

Figure 2.6(a) displays the $IV$-characteristics for a symmetric junction circuit (Fig. 2.4(a)), i.e. with $C_1 = C_2$ and $R_1 = R_2$. For $|V| < e/C_\Sigma$ the current is zero. The suppression of tunneling for low biases known as Coulomb blockade. Due to the charging energy, a gap has opened in the available chemical potentials. Only if the junction voltage $V$ is above the threshold given by the charging energy, the blockade can be overcome and a current flows. Then the junction behaves like a resistor. If an electron enters via junction one another leaves via junction two. This is known as (spatially-) correlated tunneling of electrons.

Figure 2.6(b) shows the $IV$-characteristics for a circuit with highly asymmetric junction resistances, for example $R_1 \ll R_2$. Carriers that enter through one junction are kept from escaping through the second junction by the high resistance. Eventually, the electron will leave through junction two, however, for most of the time the island is charged with an excess carrier. With increasing bias more electrons have the chance to populate the island. The $IV$-curve exhibits a staircase-like characteristics, commonly referred to as Coulomb staircase. If the asymmetry is reversed, i.e. $R_1 \gg R_2$, the island is more and more depleted and a descending instead of an ascending Coulomb staircase appears.

Figure 2.7 shows the $IV_g$-characteristics of a capacitively coupled SET (Fig. 2.4(b)) at a bias voltage of $V \to 0$. This plot is equivalent to a cut along the $x$-axis through the stability plot (Fig. 2.5). The current oscillates with the period $e/C_g$. In the instable regions, where correlated tunneling occurs, the current peaks to the value $V/2(R_1 + R_2)$. These so called Coulomb blockade peaks are sharp peaks at $V = 0$ (and $T = 0$ K of course), however, with increasing bias voltage the line width increases as well. At finite tempera-
2.3 Constant interaction model

The orthodox theory of Coulomb blockade was developed to explain single-electron charging phenomena observed in small metallic grains. In these grains the level discreteness gains importance only for structures with diameters below 1 nm. Above this limit, the level separation is in general much smaller than the Coulomb charging energy $e^2/C$ and the free energy is equal to a capacitive energy. These basic assumptions are made in the simple picture for the free energy used in the previous section.

In a semiconductor structure the island is called quantum dot. It exhibits a discrete level spectrum for much larger island sizes than in metallic systems, and the level separation increases indirectly proportional to the square of the dot diameter $d$. Assuming a symmetric square box of a given semiconductor material characterized by an effective mass $m^*$ with potential walls of infinite

\[
\frac{V}{2(R_1+R_2)}
\]

\[
\frac{e}{C_g}
\]

\[
\frac{3e}{C_g}
\]

**Figure 2.7:** $IV_g$-characteristics of a single-electron transistor for vanishing junction bias, $V \to 0$, and finite temperature.

tures some electrons can overcome the Coulomb charging energy for values of $V_g$ between the peaks. The result is a non-vanishing conductance in the blockade regions. Then, the conductance peaks acquire a finite width. In fact, Fig. 2.7 already shows the characteristics of an SET at finite temperature, since some level broadening is visible. The Coulomb blockade is almost completely washed out at sufficiently high temperatures, $k_B T \sim e^2/C_\Sigma$. 

height, the level spectrum is calculated according to

\[ \varepsilon_p = \frac{1}{2m^*} \left( \frac{\hbar \pi p}{d} \right)^2, \tag{2.19} \]

where the lower effective mass in semiconductors (GaAs: \( m^* = 0.067 \)) in comparison to the mass of free electrons used in metals (\( m^* = 1 \)) is the reason for the significance of the level separation. If the level spacing \( \Delta \varepsilon = \varepsilon_{p+1} - \varepsilon_p \) is of the order of the Coulomb charging energy, \( \Delta \varepsilon \sim e^2/C_\Sigma \), the Coulomb blockade is affected by it. A quick estimate can be obtained for a GaAs (\( \epsilon_r = 13 \)) quantum dot with a 2DEG of \( d = 400 \) nm. The charging energy can be estimated to be 1 meV (using the equation for a metallic sphere \( e^2/C_\Sigma = e^2/2\pi\epsilon_0\epsilon_r d \)). According to Eq. 2.19 the level spacing is \( \Delta \varepsilon = 0.03 \) meV. For a diameter of \( d = 50 \) nm the level spacing would be around \( \Delta \varepsilon = 2 \) meV and the charging energy is estimated to be 8 meV. Island sizes of 50 nm are well within the reach of modern lithography tools.

Even though the orthodox theory can still serve as a useful model for single-electron charging effects, a complete picture including the interplay of the discrete level spectrum and the capacitive charging, especially of small semiconductor quantum dots is only possible if the electron gas is treated quantum-mechanically. The electrons in a QD comprise a many-body system of strongly interacting Fermions in a complex potential. The quantum mechanical states of such a system are the solutions of a complex system of equations taking all possible interactions of the system and of the system with its environment into account. The issue of finding a meaningful expression for the quantum states is omitted here and a discrete set of eigenlevels with energies \( \varepsilon_p \) (energies relative to the Fermi level) is introduced in an "ad-hoc" manner. The free energy \( F \) of the QD is the sum of the capacitive energy according to the orthodox model and of the energies of all occupied single-particle levels. If one assumes that one electron occupies one level, which can be spin-degenerate of course, the following form of the free energy is obtained

\[ F(N) = \frac{(Ne)^2}{2C_\Sigma} - Ne \frac{C_g}{C_\Sigma} V_g + \sum_{p=1}^{N} \varepsilon_p. \tag{2.20} \]

The actual values of \( \varepsilon_p \) need to be calculated numerically. Within the orthodox theory the spacing of conductance peaks is calculated using Eq. 2.20 and \( F(N + 1) - F(N) = E_F \), yielding the following relation for the spacing of
the conductance peaks

\[
e \frac{C_g}{C_{\Sigma}} \Delta V_{g}^{i} = \begin{cases} 
\frac{e^2}{C_{\Sigma}} + (\varepsilon_{i+1} - \varepsilon_i) & (i \text{ even}) \\
\frac{e^2}{C_{\Sigma}} & (i \text{ odd})
\end{cases}
\]  

(2.21)

where $\Delta V_{g}^{i} = V_{g}^{i+1} - V_{g}^{i}$ and $V_{g}^{i}$ marks the center position of the $i$th peak. The dependence of the level spacing on whether $i$ is even or odd reflects the spin-degenerated nature of the levels. The separation of the change of the free energies of a QD into a capacitive part which is constant and a fluctuating part related to the discrete spacing of the single-particle levels according to Eq. (2.20) is known as constant interaction (CI) model.

### 2.4 Example structures: vertical quantum dots

In this section an illustration of the main effects related to single-electron tunneling in semiconductor nanostructures is given by shortly discussing vertical quantum dot structures in GaAs. Vertical quantum dots have thin barriers which allow a very precise control of just a few electrons ($N \leq 20$). Therefore, these structures are especially suitable for studying effects related to the discrete energy spectrum of quantum dots.

Figure 2.8(a) shows a schematic diagram of a vertical two terminal structure [32]. The InGaAs quantum dot is located between two AlGaAs barriers. The two barriers have different widths. The inclusion of Indium reduces the

![Schematic diagram of vertical quantum dot](image)

**Figure 2.8:** (a): Schematic diagram of vertical quantum dot. (b): The Coulomb blockade oscillations for a 500 nm dot [32].
bandgap so that the conduction band is below the Fermi level. This means that electrons are accumulated even when the junction voltage $V$ is zero. The dot is surrounded by a circular Schottky gate placed at the side of the mesa close to the quantum dot. With a gate voltage $V_g$ applied, the quantum dot can be squeezed electrostatically.

Figure 2.8(b) shows the Coulomb blockade oscillations measured at small $V$. Due to the small number of electrons ($N = 0 \ldots 20$) the peak spectrum is significantly influenced by the strong Coulomb interaction and quantum-mechanical effects in the few-electron regime. Since the electron gas in the dot has the shape of a circular disc and the confining potential is harmonic, the system exhibits a high degree of symmetry with degenerate single-particle states.

With the arguments of the constant interaction model, different level spacings should lead to different spacings in the conductance peaks (Eq. (2.21)). Moreover, the shell structure of the levels is also reflected in the height of the individual peaks. Strong modulations of both the peak height and the spacing of the conductance peaks, especially for the electron numbers $N = 0 \ldots 12$ can be seen in Fig. 2.8(b). Figure 2.9 displays the stability plot, i.e. the differential conductance $\partial I/\partial V$ of the structure in Fig. 2.8. The electron number $N$ is fixed in each of the white diamond-shaped areas, where Coulomb blockade is present ($\partial I/\partial V = 0$). The irregular shape of these areas is also an effect related to the level spacing and the

Figure 2.9: Stability plot for the quantum dot structure in Fig. 2.8 [32].

Figure 2.10: Coulomb staircase for a vertical dot [33].
2.4. Example structures: vertical quantum dots

shell-filling. One can clearly see that the numbers 2, 6 and 12 are especially pronounced. These numbers are the *magic numbers* of the two-dimensional harmonic oscillator potential, i.e. those electron numbers $N$ at which an additional shell in the level spectrum is occupied.

A measurement of the Coulomb staircase for a similar structure [33] is shown in Fig. 2.10. At negative voltages $V$ electrons are accumulated in the dot and a step occurs whenever an electron is trapped. The few-electron regime is again reflected by the low number of electrons which leads to modulations in the width of the plateaus of the staircase due to the strong Coulomb interaction. The width of the staircase steps is larger for low electron numbers which hints on a shift in the total capacitance $C_N$ towards higher values for more electrons.
Seite Leer / Blank leaf
Chapter 3

Ora incomincian le dolenti note.
And now the notes of anguish start to play.

DANTE, »Inferno, V«

Theory of single-electron tunneling in semiconductors

3.1 Introduction

In the last chapter an introduction to single electron tunneling was given. The basis for the theoretical predictions was the so-called orthodox theory and the constant interaction model. The shortcomings of the orthodox theory are evident. Equation (2.21) is sufficient to describe the spacing of the conductance peaks. However, no information about their amplitude and lineshape can be obtained. This requires the solution of a kinetic equation. Furthermore, the assumption that the capacitance of the island is constant and especially that $C_S$ is independent of the electron number is not valid in small semiconductor quantum dots.

In this chapter a linear response theory is discussed, which leads to an equation for the conductance of a SET in the limit of vanishingly small source-drain voltages. The effects of finite temperatures and discrete level spacings are taken into account as well as the shape of the tunneling barriers. This analysis was first performed by Kulik and Shekhter [20] for metallic islands within the orthodox theory and extended by Averin et al. [24] and Beenakker [26] to semiconductor quantum dots.
3.2 Model definition

The model for a single-electron transistor consists of a confinement region (island, quantum dot) with a discrete level spectrum $\varepsilon_i$ ($i = 0, 1, \ldots$). Each level carries one or zero electrons. The occupation of the $i$th level is therefore described by an integer number, $n_i = 0, 1$, and the occupation of the level system, i.e. a specific occupation number configuration is given by the set $\{n_i\}$ where the number of electrons is integer and $N = \sum_i n_i$. The confined region is weakly coupled to two electron reservoirs via tunneling barriers (junctions). The reservoirs which are named source and drain are taken to be in thermal

![Figure 3.1: (a) Schematic drawing of a SET. (b) Potential cut through the structure. Electrons on the island occupy discrete levels (long dashed lines). The chemical potentials $\mu(N)$ according to discrete numbers of electrons are indicated by straight lines. Electrons in the reservoirs continuously occupy all energies up to the Fermi levels.](image-url)
equilibrium at temperature $T$ and Fermi energy $E_F$, i.e. the voltage $V$ which is applied between the reservoirs is small.

The device is considered to be driven in linear-response, i.e. the conductance $G$ is defined as $G \equiv I/V$ in the limit $V \to 0$. In the absence of charging effects a net current flows through the quantum dot if the Fermi energy $E_F$ in the leads is aligned with one of the discrete single-particle levels. In a more general case, at $T = 0$ K current flows if the Fermi energy in the reservoirs aligns with one of the chemical potentials $\mu(N)$ of the quantum dot. The set of chemical potentials $\{\mu(N)\}$ is given as the change in the free energies of the quantum dot when adding or removing an electron, and therefore, the condition for the lift of the Coulomb blockade is

$$E_F = \mu(N) \equiv F(N) - F(N - 1).$$

The tunneling through the two barriers is described by sets of tunneling rates for all states $k$ in the quantum dot to the left (drain-) and the right (source-) reservoir, $\{\Gamma^d_k\}$ and $\{\Gamma^s_k\}$. The characteristic energies of the system are given by the charging energy $e^2/C_S$, the single-particle level separation $\Delta \varepsilon$, the thermal energy $k_B T$, and $\hbar \Gamma_t$. The last term denotes the intrinsic width of the transmission resonance $\hbar \Gamma_t = \hbar (\Gamma^s + \Gamma^d)$. It is assumed that for all levels in the dot $k_B T \gg \hbar \Gamma_t$, i.e. the finite width of the transmission resonance through the quantum dot can be disregarded. This allows to characterize each dot state by an occupation number. Furthermore, it is assumed that the junction resistances are above the quasi-classical limit (Eq. (2.5) in Section 2.2), i.e. $R_Q \ll R_t = R_1 R_2 / (R_1 + R_2)$ and the tunneling time $\tau_t = \Gamma_t^{-1}$ is smaller than the junction relaxation time $\tau^{-1}$. This ensures that the tunneling events are well separated in time and the current flow does not disturb the equilibrium. It is therefore possible to use a master equation to describe the transport. In the following section it is assumed that inelastic scattering takes place exclusively in the reservoirs and not in the dot. The effect of inelastic scattering in the quantum dot is discussed in [26].

The transition of an electron from the $k$th state in the dot to a final state in either the drain or the source reservoir, $\varepsilon^{d,s}_k$, requires energy conservation. The initial configuration of electrons in the quantum dot is characterized by the Helmholtz free energy

$$F(\{n_i\}, n_k = 1, N) \equiv F(n_1, \ldots, n_{k-1}, 1, n_{k+1}, \ldots; N),$$

and the final configuration by

$$F(\{n_i\}, n_k = 0, N - 1) \equiv F(n_1, \ldots, n_{k-1}, 0, n_{k+1}, \ldots; N - 1).$$
For a given drain-source voltage $V$ the fraction $\eta$ of the voltage that drops over the source barrier is $\eta = R_d/(R_d + R_s)$. According to Eq. (3.1) the quantum dot chemical potential $\mu(N)$ is given as

$$\mu(N) \equiv F(\{n_i\}, n_k = 1, N) - F(\{n_i\}, n_k = 0, N - 1). \quad (3.2)$$

The energy conservation condition for a transition into a final energy state in the source reservoir is

$$\epsilon_i^s(N) = \epsilon_i^s(N) + \eta eV. \quad (3.3)$$

Similarly, the energy conservation condition for a transition from an initial reservoir state in the source, $\epsilon_i^d(N)$, to the $k$th dot state reads

$$\epsilon_i^d(N) = \epsilon_i^d(N + 1) + \eta eV. \quad (3.4)$$

$N$ is the number of electrons before the tunneling event takes place. For the transitions between the dot and the drain reservoir the following condition applies for the final state $\epsilon_i^d(N)$

$$\epsilon_i^d(N) = \mu(N) - (1 - \eta)eV \quad (3.5)$$

and the initial state $\epsilon_i^s(N)$

$$\epsilon_i^s(N) = \mu(N + 1) - (1 - \eta)eV. \quad (3.6)$$

The parameter $\eta$ drops out of the final equation for the linear-response conductance. The derivation of the conductance expression follows the line-of-arguments given by Beenakker [26].

### 3.3 Master equation approach

#### 3.3.1 Master equation

In general, the probability distribution $P(N, t)$ is a time-dependent property that describes the probability for the stochastic variable $N$ of the system to have the value $N$. Transitions from a state of the system described by the variable $N$ to a state described by $M$ will produce changes in the corresponding occupation factors $P(N, t)$ and $P(M, t)$,

$$dP(N, t) = -dP(M, t) = P(N, t)W(N, M)dt.$$
For discrete stochastic variables the summation over all possible transition processes results in the following *master equation*

\[ \frac{\partial}{\partial t} P(N, t) = \sum_{M \neq N} [P(M, t)W(M, N) - P(N, t)W(N, M)], \quad (3.7) \]

where \( W(N, M) \) is the transition probability distribution or transition rate (usually assumed to be time independent) for the transition of the system from state \( \mathcal{N} \) to state \( \mathcal{M} \). The probability distribution is normalized according to

\[ \sum_{N} P(N, t) = 1. \]

In a quantum dot the stochastic variable \( \mathcal{N} \) is the electron number. Single-electron transitions between quantum dot states occur if the number of electrons in the dot is increased or decreased by one. The stationary solution to the master equation can therefore be given as the solution to the following rate equation

\[ 0 = \sum_{M=\{\frac{N-1}{N+1}\}} [P(M)W(M, N) - P(N)W(N, M)], \quad (3.8) \]

where \( W(N, M) \) is the total transition rate from the \( N \)-electron to the \( M \)-electron system. The set of discrete non-equilibrium probability distributions \( \{P(N)\} \) is the solution of the master equation.

### 3.3.2 Probability distributions and transition rates

The stationary equilibrium probability-distribution function \( P_{eq}(\{n_i\}) \) for particular occupation configurations \( \{n_i\} \) is given by the Gibbs distribution of the grand canonical ensemble theory

\[ P_{eq}(\{n_i\}) = Z^{-1} \exp \left\{ -\frac{1}{k_B T} [F(\{n_i\}, N) - E_F N] \right\}, \quad (3.9) \]

where \( \{n_i\} \) is a specific set of occupation numbers, \( F(\{n_i\}, N) \) is the free energy of this configuration and \( Z \) is the *grand partition function*

\[ Z = \sum_{\{n_i\}} \exp \left\{ -\frac{1}{k_B T} [F(\{n_i\}, N) - E_F N] \right\}. \quad (3.10) \]
The equilibrium probability $P_{eq}(N)$ for the dot to contain $N$ electrons is

$$P_{eq}(N) = \sum_{\{n_i\}} P_{eq}(\{n_i\}) \delta_{N, \sum_i n_i}. \quad (3.11)$$

The single-particle occupation factor for the dot levels is given by the \textit{conditional-probability distribution function} $g(\varepsilon_k | N)$, i.e. the probability that the $k$th single-particle level in the quantum dot is occupied \textit{if} the dot contains $N$ electrons

$$g(\varepsilon_k | N) = \frac{1}{P_{eq}(N)} \sum_{\{n_i\}} P_{eq}(\{n_i\}) \delta_{n_k, 1} \delta_{N, \sum_i n_i} \quad (3.12)$$

(from now on the symbol $g_N(\varepsilon_k)$ is used instead). The sum over the occupation configurations $\{n_i\}$ on the right hand side of Eq. (3.12) is the \textit{joint-probability distribution} to have $N$ electrons in the dot \textit{and} the $k$th single-particle level occupied. The Gibbs distribution for localized systems under the influence of external reservoirs is discussed in Appendix A.

The transition rates between the many-particle ground states of the isolated quantum dot can be derived by treating the tunneling Hamiltonian as a perturbation to the time-dependent Schrödinger equation. An account of this approach is given in Appendix B. The total transition rate $W$ is the sum over the partial transition rates of the reservoirs, i.e. the summation is over the reservoirs $r \in \{s, r\}$

$$W(N, M) = \sum_r W^r(N, M), \quad (3.13)$$

which are given as sums over all partial tunneling rates to or from specific energy levels $\varepsilon_k$. Transitions to dot states are characterized by

$$W^r(N, N + 1) = \sum_k \Gamma_k f(\varepsilon_f^r(N) - E_F^r)[1 - g_N(\varepsilon_k)]$$

$$W^r(N - 1, N) = \sum_k \Gamma_k f(\varepsilon_f^r(N - 1) - E_F^r)[1 - g_{N-1}(\varepsilon_k)]. \quad (3.14)$$

This leads to an increase of the electron number in the dot. The next equations refer to transitions from filled dot states to empty reservoir states decreasing the electron number in the dot

$$W^r(N, N - 1) = \sum_k \Gamma_k [1 - f(\varepsilon_f^r(N) - E_F^r)] g_N(\varepsilon_k)$$

$$W^r(N + 1, N) = \sum_k \Gamma_k [1 - f(\varepsilon_f^r(N + 1) - E_F^r)] g_{N+1}(\varepsilon_k), \quad (3.15)$$
3.3. Master equation approach

where $c_{i,f}^\dagger$ are the initial and final states in the reservoirs, $f$ is the Fermi-Dirac distribution for continuous states

$$f(\epsilon - E_F) = \left[1 + \exp\left(\frac{\epsilon - E_F}{k_B T}\right)\right]^{-1},$$

and $g$ is the occupation probability for the discrete energy levels in the quantum dot which is taken to be the Gibbs distribution for discrete single-particle levels (Eq. (3.12)) in the most general case. For large electron numbers the Fermi-Dirac distribution for continuous states can be used instead [31].

### 3.3.3 Detailed-balance equations

With Eq. (3.14) and Eq. (3.15) the stationary master equation, Eq. (3.8), for the stationary non-equilibrium distribution function $P(N)$ is

$$0 = P(N - 1) \sum_k [1 - g_{N-1}(\varepsilon_k)] \sum_r \Gamma^r_k f(\varepsilon^r_k (N - 1) - E^r_F) + P(N + 1) \sum_k g_{N+1}(\varepsilon_k) \sum_r \Gamma^r_k [1 - f(\varepsilon^r_k (N + 1) - E^r_F)]

- P(N) \sum_k [1 - g_N(\varepsilon_k)] \sum_r \Gamma^r_k f(\varepsilon^r_k (N) - E^r_F)

- P(N) \sum_k g_N(\varepsilon_k) \sum_r \Gamma^r_k [1 - f(\varepsilon^r_k (N) - E^r_F)].$$

(3.17)

This equation can be recast using the conditional-probability distribution Eq. (3.12) which leads to a set of equations for the non-equilibrium probability-distribution functions $P(\{n_i\})$

$$0 = \sum_k P(n_1, \ldots, n_{k-1}, 0, n_{k+1}, \ldots) \delta_{n_{k-1}} \sum_r \Gamma^r_k f(\varepsilon^r_k (N - 1) - E^r_F)$$

$$+ \sum_k P(n_1, \ldots, n_{k-1}, 1, n_{k+1}, \ldots) \delta_{n_{k-1}} \sum_r \Gamma^r_k [1 - f(\varepsilon^r_k (N + 1) - E^r_F)]$$

$$- \sum_k P(\{n_i\}) \delta_{n_{k-1}} \sum_r \Gamma^r_k f(\varepsilon^r_k (N) - E^r_F)$$

$$- \sum_k P(\{n_i\}) \delta_{n_{k-1}} \sum_r \Gamma^r_k [1 - f(\varepsilon^r_k (N) - E^r_F)].$$

(3.18)
The first two terms refer to the system evolving into the state with the occupation \( \{n_i\} \) either by adding one electron to the dot or by removing one electron from the dot to the drain or source reservoir. The last two terms refer to the system evolving out of the state with the occupation \( \{n_i\} \) due to an electron filling an unoccupied level in the dot and an electron leaving the dot. Equation (3.18) is equivalent to a set of \textit{detailed-balance equations}, one for each dot level \( k \).

\[
P(n_1, \ldots, n_{k-1}, 1, n_{k+1}, \ldots) \sum_r \Gamma^r_k \left[ 1 - f(\epsilon^r_k(\tilde{N} + 1) - E^r_F) \right] \\
= P(n_1, \ldots, n_{k-1}, 0, n_{k+1}, \ldots) \sum_r \Gamma^r_k f(\epsilon^r_k(\tilde{N}) - E^r_F),
\]

using the notation \( \tilde{N} \equiv \sum_{i \neq k} n_i \), i.e. \( N = \tilde{N} + 1 \) if the \( k \)th state is definitely occupied, \( \delta_{nk,1} \), and \( N = \tilde{N} \) if the \( k \)th state is unoccupied, \( \delta_{nk,0} \).

### 3.3.4 Stationary current

According to Kichhoff’s law the steady-state current through the source barrier equals that through the drain barrier, \( I = I^s = I^d \). It is given by

\[
I = -e \sum_N P(N) [W^s(N, N + 1) - W^s(N, N - 1)].
\]

Applying the results of the previous section this equation can be written as

\[
I = -e \sum_N \sum_k P(N) \{ \Gamma^s_k [f(\epsilon^s_k(N) - E^s_F)] [1 - g_N(\varepsilon_k)] \\
- \Gamma^s_k [1 - f(\epsilon^s_k(N) - E^s_F)] g_N(\varepsilon_k) \}.
\]

Again, the first term in the sum refers to an electron moving into the previously empty state \( k \) in the quantum dot. The second term refers to an electron moving from state \( k \) in the dot to the final state \( \epsilon^s_i \) in the source reservoir.
3.4 Linear-response regime

3.4.1 Linear-response conductance

Since it is impossible to solve above equations in a general form, a restriction to the case of small source-drain voltages $V$ is made. The conductance in the linear-response regime can be given analytically. As mentioned before, the *linear-response conductance* is defined as

$$G = \lim_{V \to 0} \frac{I}{V}. \quad (3.22)$$

Equilibrium conditions are assumed for the source and the drain reservoir, i.e. $E_F^s = E_F^d = E_F$. In this limit the non-equilibrium probability distribution $P\left(\{n_i\}\right)$ differs very little from the Gibbs distribution at equilibrium $P_{eq}\left(\{n_i\}\right)$. Therefore, the stationary non-equilibrium probability distribution is linearly expanded in terms of $V$

$$P\left(\{n_i\}\right) \equiv P_{eq}\left(\{n_i\}\right) \left(1 + \frac{eV}{k_BT} \Psi\left(\{n_i\}\right)\right). \quad (3.23)$$

Substitution of this expression into the balance equations, Eq. (3.19), yields

$$P_{eq}(n_1, \ldots, n_{k-1}, 1, n_{k+1}, \ldots) \left[1 + \frac{eV}{k_BT} \Psi(n_1, \ldots, n_{k-1}, 1, n_{k+1}, \ldots)\right]$$

$$\times \{\Gamma_k^s[1 - f(\epsilon + \eta eV)] + \Gamma_k^d[1 - f(\epsilon - (1 - \eta)eV)]\}$$

$$= P_{eq}(n_1, \ldots, n_{k-1}, 0, n_{k+1}, \ldots) \left[1 + \frac{eV}{k_BT} \Psi(n_1, \ldots, n_{k-1}, 0, n_{k+1}, \ldots)\right]$$

$$\times \{\Gamma_k^s f(\epsilon + \eta eV) + \Gamma_k^d f(\epsilon - (1 - \eta)eV)\}, \quad (3.24)$$

with $\epsilon \equiv F(\{n_i\}, n_k = 1, \tilde{N} + 1) - F(\{n_i\}, n_k = 0, \tilde{N}) - E_F$. Since the balance equations are linearized with respect to the junction voltage $V$, a Taylor expansion of the Fermi-Dirac factor including only the first-order term can be used in Eq. (3.24), i.e.

$$f(\epsilon + \eta eV) = f(\epsilon) + \eta eV f'(\epsilon)$$

$$= f(\epsilon) \left[1 - \eta \frac{f(\epsilon)}{1 + \exp\left(-\frac{\epsilon}{k_BT}\right)} \frac{eV}{k_BT}\right]. \quad (3.25)$$
and

\[ f(\epsilon - (1 - \eta)eV) = f(\epsilon) - (1 - \eta)eV f'(\epsilon) \]

\[ = f(\epsilon) \left[ 1 + (1 - \eta) \frac{f(\epsilon)}{1 + \exp \left( -\frac{\epsilon}{k_B T} \right)} \frac{eV}{k_B T} \right]. \quad (3.26) \]

The following identities are substituted

\[ 1 - f(\epsilon) = f(\epsilon) \exp \left( \frac{\epsilon}{k_B T} \right), \quad (3.27) \]

\[ k_B T f'(\epsilon) \left[ 1 + \exp \left( -\frac{\epsilon}{k_B T} \right) \right] = -f(\epsilon), \quad (3.28) \]

and

\[ P_{eq}(n_1, \ldots, n_k, 1, n_{k+1}, \ldots) \]

\[ = P_{eq}(n_1, \ldots, n_k, 0, n_{k+1}, \ldots) \exp \left(-\frac{\epsilon}{k_B T} \right). \quad (3.29) \]

Latter describes the so-called detailed-balance symmetry which means that if the level system is in occupation distribution \( \{n_i\} \), the probability for the \( k \)th level to be occupied compared to the probability that the level is empty differs by a Boltzmann factor. Another consequence is that at zero temperature no transitions are possible which would lead to the occupation of the \( k \)th level if previously empty, whereas at high temperatures both probabilities are equal.

Application of these equations leads to a new form of the balance equations, now for the first-order non-equilibrium correction \( \Psi(\{n_i\}) \) to the equilibrium probability distribution \( P_{eq}(\{n_i\}) \)

\[ \Psi(n_1, \ldots, n_k, 1, n_{k+1}, \ldots) \times \left\{ (\Gamma_k^s + \Gamma_k^d) + \frac{eV}{k_B T} [\Gamma_k^s \eta - \Gamma_k^d (1 - \eta)] \frac{\exp \left(-\frac{\epsilon}{k_B T} \right)}{1 + \exp \left(-\frac{\epsilon}{k_B T} \right)} \right\} \]

\[ \times \left\{ (\Gamma_k^s + \Gamma_k^d) + \frac{eV}{k_B T} [\Gamma_k^s \eta - \Gamma_k^d (1 - \eta)] \frac{1}{1 + \exp \left(-\frac{\epsilon}{k_B T} \right)} \right\}. \]
3.4. Linear-response regime

All factors $P_{eq}$ and $f$ cancel. Further, all terms proportional to $eV/k_BT$, i.e. all second order contributions are ignored and a final equation for $\Psi(\{n_i\})$ is obtained

$$\Psi(n_1, \ldots, n_{k-1}, 1, n_{k+1}, \ldots) = \Psi(n_1, \ldots, n_{k-1}, 0, n_{k+1}, \ldots) + \frac{\Gamma^d_k}{\Gamma^s_k + \Gamma^d_k} - \eta,$$  \hspace{1cm} (3.30)

for which the solution is

$$\Psi(\{n_i\}) = \text{const} + \sum_i n_i \left( \frac{\Gamma^d_i}{\Gamma^s_i + \Gamma^d_i} - \eta \right).$$  \hspace{1cm} (3.31)

The first-order non-equilibrium correction $\Psi(\{n_i\})$ to $P_{eq}(\{n_i\})$ is zero if $\eta = \Gamma^d_i/(\Gamma^s_i + \Gamma^d_i)$ for all $i$. This is the case if $\eta = 1/2$ and $\Gamma^s_i = \Gamma^d_i$, i.e. if the tunneling barriers are identical.

With this result the linear-response current through the quantum dot in first order in $V$ can be calculated. Linearization of the current equation, Eq. (3.21), after substitution of Eq. (3.23) and the subsequent use of the Taylor expansion of the Fermi-Dirac factor, Eq. (3.25), yields

$$I = -e \sum_k \sum_{\{n_i\}} \Gamma^s_k P_{eq}(\{n_i\}) \left( 1 + \frac{eV}{k_BT} \Psi(\{n_i\}) \right) \times \left\{ \delta_{n_k,0} f(\epsilon + \eta eV) - \delta_{n_k,1}[1 - f(\epsilon + \eta eV)] \right\}$$

and expanding the Fermi-Dirac factor

$$I = -e \sum_k \sum_{\{n_i\}} \Gamma^s_k P_{eq}(\{n_i\}) \times \left\{ \delta_{n_k,0} f(\epsilon) - \delta_{n_k,1}[1 - f(\epsilon)] + \delta_{n_k,0} \eta eV f'(\epsilon) + \delta_{n_k,1} \eta eV f'(\epsilon) \right\}$$

$$+ \frac{eV}{k_BT} \Psi(\{n_i\}) \delta_{n_k,0} f(\epsilon) - \frac{eV}{k_BT} \Psi(\{n_i\}) \delta_{n_k,1}[1 - f(\epsilon)]$$

$$+ \frac{eV}{k_BT} \Psi(\{n_i\}) \delta_{n_k,0} \eta eV f'(\epsilon) + \frac{eV}{k_BT} \Psi(\{n_i\}) \delta_{n_k,1} \eta eV f'(\epsilon).$$
Neglecting terms of the order $V^2$ (the last two terms) one arrives at

$$I = -e \frac{eV}{k_B T} \sum_k \sum_{\{n_i\}} \Gamma_k^s \rho_{eq}(\{n_i\})$$

$$\times \left\{ \frac{k_B T}{eV} \delta_{n_k,0} f(\epsilon) - \frac{k_B T}{eV} \delta_{n_k,1} [1 - f(\epsilon)] + \delta_{n_k,0} \eta k_B T f'(\epsilon) + \delta_{n_k,1} \eta k_B T f'(\epsilon) + \Psi(\{n_i\}) \delta_{n_k,0} f(\epsilon) - \Psi(\{n_i\}) \delta_{n_k,1} [1 - f(\epsilon)] \right\}. \tag{3.32}$$

The first two terms cancel because of Eq. (3.27) and Eq. (3.29) (detailed balance symmetry). If the identities Eqs. (3.27)-(3.29) are now substituted into the current equation one obtains

$$I = e \frac{eV}{k_B T} \sum_k \sum_{\{n_i\}} \Gamma_k^s \rho_{eq}(\{n_i\}) \left\{ (\delta_{n_k,0} + \delta_{n_k,1}) \eta \frac{f(\epsilon)}{1 + \exp \left( \frac{\epsilon}{k_B T} \right)} - \Psi(\{n_i\}) \delta_{n_k,0} f(\epsilon) + \Psi(\{n_i\}) \delta_{n_k,1} f(\epsilon) \exp \left( \frac{\epsilon}{k_B T} \right) \right\}, \tag{3.32}$$

and finally

$$I = \frac{e^2 V}{k_B T} \sum_k \sum_{\{n_i\}} \Gamma_k^s \rho_{eq}(\{n_i\}) \delta_{n_k,0} f(\epsilon) \times [\eta + \Psi(n_1, \ldots, n_{k-1}, 1, n_{k+1}, \ldots) - \Psi(n_1, \ldots, n_{k-1}, 0, n_{k+1}, \ldots)], \tag{3.33}$$

where the first two terms in Eq. (3.32) combine to give

$$\rho_{eq}(\{n_i\})(\delta_{n_k,0} + \delta_{n_k,1}) f(\epsilon) \frac{\eta}{1 + \exp \left( \frac{\epsilon}{k_B T} \right)} = \rho_{eq}(\{n_i\}) \delta_{n_k,0} \eta f(\epsilon),$$

and the last two terms give

$$\rho_{eq}(\{n_i\}) \Psi(\{n_i\}) f(\epsilon) \left[ \delta_{n_k,1} \exp \left( \frac{\epsilon}{k_B T} \right) - \delta_{n_k,0} \right] = \rho_{eq}(\{n_i\}) \delta_{n_k,0} f(\epsilon) \times [\Psi(n_1, \ldots, n_{k-1}, 1, n_{k+1}, \ldots) - \Psi(n_1, \ldots, n_{k-1}, 0, n_{k+1}, \ldots)].$$
It is now possible to write the current equation again in terms of $N$ since for $\delta_{n_k,0}$ the replacement $\tilde{N} = N$ can always be made, i.e.

$$\epsilon = F(\{n_i\}, n_k = 1, N + 1) - F(\{n_i\}, n_k = 0, N) - E_F. \quad (3.34)$$

Substituting Eq. (3.30) and dividing by $V$ leads to the final expression for the linear-response conductance

$$G = \frac{e^2}{k_B T} \sum_k \sum_{\{n_i\}} \frac{\Gamma^s_k \Gamma^d_k}{\Gamma^s_k + \Gamma^d_k} P_{eq}(\{n_i\}) \delta_{n_k,0} f(\epsilon). \quad (3.35)$$

The particular product between the Gibbs distribution for the quantum dot $P(\{n_i\})$ and the Fermi-Dirac distribution $f(\epsilon)$ for the reservoirs reflects the fact that tunneling requires an occupied initial state in the reservoir and an empty state in the quantum dot. Eq. (3.35) is the central result of the paper of Beenakker [26]. The same equation has also been obtained by Meir et al. [23] by solving an Anderson model in the limit $k_B T \gg \hbar \Gamma_t$. 

Within the constant-interaction model for semiconductor quantum dots already introduced in the previous chapter, the difference in the free energies using Eq. (2.20) is

$$F(n_k = 1, N + 1) - F(n_k = 0, N) = \varepsilon_k + U(N + 1) - U(N), \quad (3.36)$$

where $U(N)$ denotes the capacitive contributions to the free energy of a system with $N$ electrons. Inserting the Gibbs distribution $P(N)$, Eq. (3.11), and the single-particle occupation numbers $g_N(\varepsilon_k)$, Eq. (3.12), the equation for the linear-response conductance, Eq. (3.35), reduces to

$$G = \frac{e^2}{k_B T} \sum_k \sum_{N=0}^{\infty} \frac{\Gamma^s_k \Gamma^d_k}{\Gamma^s_k + \Gamma^d_k} P_{eq}(N) \left[ 1 - g_N(\varepsilon_k) \right] f(\varepsilon_k + U(N + 1) - U(N) - E_F) \quad (3.37)$$

or the equivalent equation

$$G = \frac{e^2}{k_B T} \sum_k \sum_{N=1}^{\infty} \frac{\Gamma^s_k \Gamma^d_k}{\Gamma^s_k + \Gamma^d_k} P_{eq}(N) g_N(\varepsilon_k)$$

$$\times \left[ 1 - f(\varepsilon_k + U(N) - U(N - 1) - E_F) \right]. \quad (3.38)$$

A first assumption that limits the applicability of these equations has already been made in the model definition: one assumes that the quantum mechanical broadening of the quantum dot levels is smaller than the thermal energy, $\hbar \Gamma_t \ll k_B T$. In the following section limiting cases for Eq. (3.37) and Eq. (3.38) related to the other significant energy scales are discussed.
3.4.2 Limiting cases

In the limit $k_B T \gg \Delta \varepsilon$, the discrete energy spectrum may be treated as a continuum of states. The Gibbs distribution is approximated by the Fermi-Dirac distribution, $g_N(\varepsilon_k) = f(\varepsilon_k - \bar{\mu}(N))$, with a given chemical potential $\bar{\mu}(N)$ determined by

$$
\sum_k f(\varepsilon_k - \bar{\mu}(N)) = N.
$$

The Gibbs distribution $P_{eq}(N)$ is taking its classical form

$$
P_{\text{class}}(N) = \frac{\exp\left\{ -\frac{U(N) + N(\bar{\mu} - E_F)}{k_B T} \right\}}{\sum_N \exp\left\{ -\frac{U(N) + N(\bar{\mu} - E_F)}{k_B T} \right\}},
$$

where $\bar{\mu}$ is the chemical potential of the quantum dot in equilibrium. The summation over $k$ in Eq. (3.38) is replaced by an integration over $\varepsilon$, multiplied by the density of states $\rho(\varepsilon)$. If $k_B T \ll \bar{\mu}, E_F$, the density of states and the tunneling rates are evaluated at $\bar{\mu}$ using that $\mu(N) \approx \text{const} = \bar{\mu}$ for all $N$ which contribute to the classical Gibbs distribution. The conductance becomes

$$
G = \frac{e^2 \rho}{k_B T} \Gamma^s \Gamma^d \sum_{N=1}^{\infty} P_{\text{class}}(N) g(U(N) - U(N - 1) + \bar{\mu} - E_F),
$$

if $\Delta \varepsilon \ll k_B T \ll \bar{\mu}, E_F$, (3.39)

with

$$
\int_{-\infty}^{\infty} dy f(y)[1 - f(y + x)] = x \left[ 1 - \exp\left( -\frac{x}{k_B T} \right) \right]^{-1} \equiv g(x).
$$

If furthermore $k_B T \gg e^2/C_\Sigma$, the charging energy can be ignored, and Eq. (3.39) with $g = k_B T$ states that the conductance is simply that of the two tunnel barriers in series, i.e.

$$
G = \frac{G^s G^d}{G^s + G^d}, \quad \text{if} \quad e^2/C_\Sigma, \Delta \varepsilon \ll k_B T \ll \bar{\mu}, E_F. \quad (3.40)
$$

The conductances of the source and drain barrier, $G^s$ and $G^d$, are given by the thermally averaged Landauer formula

$$
G^{s,d} = -\frac{e^2}{h} \int_0^{\infty} d\varepsilon T^{s,d}(\varepsilon) \frac{df}{d\varepsilon}, \quad (3.41)
$$
where the transmission probability of a barrier \( T(\varepsilon) \) equals the tunneling rate \( \Gamma(\varepsilon) \) divided by the attempt frequency \( \nu(\varepsilon) = 1/\hbar\rho(\varepsilon) \),

\[
T^{s,d}(\varepsilon) = \frac{\hbar \Gamma^{s,d}(\varepsilon)}{\nu(\varepsilon)} = \hbar \frac{\Gamma^{s,d}(\varepsilon)}{\Gamma_{s,d}(\varepsilon)} \rho(\varepsilon). \tag{3.42}
\]

If the tunneling rates at high barriers are assumed to be independent of energy the conductance of the barriers for each side becomes \( G^{s,d} = e^2 \Gamma^{s,d} \rho \), evaluated at \( E_F \) this leads to the conductance of the quantum dot

\[
G = e^2 \rho \frac{\Gamma^{s} \Gamma^{d}}{\Gamma^{s} + \Gamma^{d}} = \frac{e^2}{h} \frac{T^{s} T^{d}}{T^{s} + T^{d}} \equiv G_\infty,
\]

if \( e^2/C_\Sigma \Delta \varepsilon \ll k_B T \ll \bar{\mu}, E_F \). \tag{3.43}

The conductance \( G_\infty \) depends solely on the barrier height and width which determine the transmission probabilities \( T^{s,d} \), however, it is independent of the size of the dot. The assumption that the quantum mechanical broadening of the levels should be small, \( \hbar \Gamma_t \ll k_B T \), i.e. the levels should be sharply defined together with \( \rho \sim 1/\Delta \varepsilon \) in Eq. (3.42) requires that \( T^{s,d} \ll 1 \). Hence, the general criterion for the tunneling rates of the barriers implies that the conductance of the quantum dot is \( G \ll e^2/h \).

Coulomb blockade oscillations become visible at low temperatures. In the regime where \( k_B T \ll \Delta \varepsilon \), the term with \( k = N = N_{min} \) in Eq. (3.37) gives the dominant contribution to the conductance. All \( P_{eq}(N) \) are negligible small for \( N \) not equal to \( N_{min} \) or \( N_{min} - 1 \), so that \( N_{min} \) minimizes the value \( \Delta(N) = \Omega(N) - \Omega(N - 1) \) with \( \Omega(N) = F(N) - N E_F \), i.e.

\[
\Delta(N) = \varepsilon_N + U(N) - U(N - 1) - E_F,
\]

and the probability \( P_{eq}(N) \) is

\[
P_{eq}(N_{min}) = \frac{\exp\left[-\Omega(N_{min})/k_B T\right]}{\exp\left[-\Omega(N_{min})/k_B T\right] + \exp\left[-\Omega(N_{min} - 1)/k_B T\right]} = f(\Delta_{min}).
\]

Using the equivalence

\[
f(x)[1 - f(x)] = -k_B T f'(x), \tag{3.44}
\]

the conductance in the low temperature limit is therefore given as

\[
G = -e^2 \frac{\Gamma^{s}_{N_{min}} \Gamma^{d}_{N_{min}}}{\Gamma^{s}_{N_{min}} + \Gamma^{d}_{N_{min}}} f'(\Delta_{min}), \text{ if } k_B T \ll \Delta \varepsilon. \tag{3.45}
\]
This formula is equivalent to the one that describes thermally broadened resonant tunneling. However, Eq. (3.45) includes the effects of charging as well as the discrete level spectrum.

### 3.4.3 Amplitude and lineshape

With the results from the previous section it is possible to give simple approximations for the amplitude and the lineshape of the conductance oscillations.

The classical regime, $\Delta \varepsilon \ll k_B T$, was first studied by Kulik and Shekhter [20] within the orthodox theory. A continuum of states is assumed to participate in the conductance. If $\Delta \varepsilon \ll k_B T \ll e^2/C_\Sigma$, only one electron number, $N_{\text{min}}$, contributes to the sum in Eq. (3.39) minimizing the difference $\Delta(N) = U(N) - U(N - 1) + \bar{\mu} - E_F$. If defining $\Delta_{\text{min}} = \Delta(N_{\text{min}})$, Eq. (3.39) reduces to

$$
G = \frac{e^2 \rho}{k_B T} \frac{\Gamma^d}{\Gamma^s + \Gamma^d} P_{\text{class}}(N_{\text{min}}) g(\Delta_{\text{min}})
$$

The final result for the shape and the height of the peak is therefore

$$
\frac{G}{G_{\text{max}}} = \frac{\Delta_{\text{min}}/k_B T}{\sinh(\Delta_{\text{min}}/k_B T)} \approx \cosh^{-2} \left( \frac{\Delta_{\text{min}}}{2.5 k_B T} \right),
$$

$$
G_{\text{max}} = \frac{e^2}{2 \Delta \varepsilon} \frac{\Gamma^d}{\Gamma^s + \Gamma^d}, \quad \text{if} \quad \Delta \varepsilon \ll k_B T \ll e^2/C_\Sigma.
$$

In the classical regime the maximum conductance is temperature independent and half the high-temperature value of Eq. (3.43) with negligible charging energy. The reason is the correlation of subsequent tunneling events by the charging energy. One electron can only tunnel onto the quantum dot if another one is tunneling off. Therefore, the probability for one tunnel event decreases to one half. The width of the peaks is linear in temperature.

The situation is distinctly different in the quantum- or resonant-tunneling regime, $k_B T \ll \Delta \varepsilon < e^2/C_\Sigma$. Tunneling occurs through a single level labeled by $N_{\text{min}}$ in Eq. (3.45). The conductance oscillations and the peak height
3.4. Linear-response regime

\( G_{\text{max}} \) are given by

\[
\frac{G}{G_{\text{max}}} = -4k_B T f'(\Delta_{\text{min}}) = \cosh^{-2}\left( \frac{\Delta_{\text{min}}}{2k_B T} \right)
\]

\[
G_{\text{max}} = \frac{e^2}{4k_B T} \frac{\Gamma_{N_{\text{min}}}^{s} \Gamma_{N_{\text{min}}}^{d}}{\Gamma_{N_{\text{min}}}^{s} + \Gamma_{N_{\text{min}}}^{d}}, \quad \text{if} \quad k_B T \ll \Delta \ll e^2/C_S.
\] (3.48)

The lineshape in the quantum regime is almost the same as in the classical regime, except for the different effective temperatures \( T^* \), which is \( 2.5T \) in the classical and \( 2T \) in the quantum regime. The peak maximum \( G_{\text{max}} = G_\infty (\Delta \epsilon / 4k_B T) \) decreases linearly with increasing temperature in the quantum regime while remaining constant in the classical regime. At the same time, the peaks broaden and the total area under the conductance peaks remains constant.

It should be noted that above theory is not suitable for the case, where the natural linewidth \( \hbar \Gamma_t \) exceeds the thermal energy, \( \hbar \Gamma_t \gg k_B T \). As a compromise the Breit-Wigner formula can be used

\[
G = g \frac{e^2}{\hbar} \frac{\Gamma_{N_{\text{min}}}^{s} \Gamma_{N_{\text{min}}}^{d}}{\Gamma_{N_{\text{min}}}^{s} + \Gamma_{N_{\text{min}}}^{d}} \Gamma (\Delta/2)^2 + (\Gamma/2)^2,
\] (3.49)

where \( g \) is the degeneracy of the resonant level and \( \Delta \) is the energetic separation of the resonant level from the Fermi level [34].

Neglecting the influence of the energy or the temperature on the tunneling rates and assuming the same rates \( \Gamma_{k}^{s,d} \) for all levels \( k \) as well as equidistant energy levels \( \Delta \epsilon = \text{const} \), the conductance can be evaluated using Eq. (3.37) in order to demonstrate the distinct regimes. Figure 3.2 shows the temperature dependence of the conductance peak maxima. Also shown are the minima which merge with the maxima as \( k_B T \) approaches \( e^2/C_S \) (classical regime). In the resonant tunneling regime

![Figure 3.2: Temperature dependence of the maxima and minima of the Coulomb blockade peaks for \( \Delta \epsilon = 0.01 e^2/C_S \) and \( \Gamma_{s} = \Gamma_{d} \). (From Beenakker [26].)](image-url)
$k_B T \ll \Delta \varepsilon$ the peak height increases with reduced temperatures due to the vanishing thermal broadening of the resonance.

Figure 3.3(a) displays the temperature dependence of the peaks in the classical regime. The width increases with $T$. For low temperatures the peak heights are independent of temperature. The crossover from the classical to the resonant-tunneling regime is shown in Fig. 3.3(b).

![Figure 3.3: Temperature dependence of the Coulomb blockade peaks for parameters $\Delta \varepsilon = 0.01e^2/C_{\Sigma}$ in the classical regime (a) and in the quantum regime (b). In (a) the temperature increases from $k_B T/(e^2/C_{\Sigma}) = 0.075$ in [a] to 2 in [f]. In (b) the temperature increases from $k_B T/\Delta \varepsilon = 0.5$ in [a] to 15 in [d]. (From Beenakker [26].)](image)
Chapter 4

Transport coefficients

4.1 Introduction

In the previous chapter a formalism for the calculation of the linear-response transport through a single-electron transistor was derived. The main results are condensed in an equation for the linear-response conductance, Eq. (3.35), for a system composed of a quantum dot weakly connected to a drain and a source reservoir via tunneling barriers. The linear-response conductance is given as a function of the Helmholtz free energies of the system, $F(\{n_i\}, N)$, and sets of tunneling rates for the two barriers, $\{\Gamma^k_i\}$ and $\{\Gamma^d_i\}$. The free energies and the tunneling rates are termed transport coefficients since they contain all physical relevant information about the many-particle ground state of the system and the shape of the tunneling barriers which are necessary for the evaluation of the transport equation (Eq. (3.35)).

In Chapter 2 the free energy was obtained within the orthodox theory and the constant interaction model at zero temperature, i.e.

$$F(N) = \frac{(Ne)^2}{2C_{\Sigma}} - Ne \frac{C_g}{C_{\Sigma}} V_g + \sum_{k=1}^{N} \varepsilon_k.$$  

However, while this simplified model was sufficient to provide a basic understanding of the key effects associated with Coulomb blockade in semiconductor quantum structures, it will fail to give precise results for small structures
where the level spectrum is strongly influenced by the screened electrostatic potential. Especially, the parameterization of the free energy in a circuit-like approach using capacitances and voltage sources is not a sufficiently accurate model for structures with small quantum dots. More appropriate is the calculation of the free energies using a solid state model for the carrier densities and the electrostatic potential. Furthermore, the resulting self-consistent potential can also be used to calculate the tunneling rates for realistic electrostatic barrier shapes.

In the following chapter the self-consistent theory for the calculation of the free energies on one side and the transfer Hamiltonian formalism for the calculation of the tunneling rates on the other side is discussed.

### 4.2 Equilibrium carrier densities

#### 4.2.1 Energy bandstructure and applied forces

The electronic structure of a bulk semiconductor is best described by its bandstructure. The bandstructure reflects the properties of the crystal potential in which the electrons are more or less free to move. The periodicity of the crystal potential in real space allows the calculation of the energy bands in the reciprocal space of wavevectors \( \mathbf{k} \)-space. This assumes energy bands which are spatially invariant.

An externally applied force field modifies the bandstructure and leads to spatially varying energy bands. Consider an undisturbed bulk bandstructure \( E_{n0}(\mathbf{k}) \), associated with a particular semiconducting material. Force fields applied to the bandstructure can act in two ways. Either, they can affect the crystal potential of the unit cell, and therefore change the bandstructure \( E_{n0}(\mathbf{k}) \) such that

\[
E_n(\mathbf{k}, \mathbf{r}) = E_{n0}(\mathbf{k}) + \delta E_n(\mathbf{r}),
\]

where \( \delta E_n(\mathbf{r}) \) is the perturbation and \( n \) the band index, or they can produce a potential energy \( \Psi(\mathbf{r}) \) which is superimposed on the bandstructure such that

\[
\mathcal{E}_n(\mathbf{k}, \mathbf{r}) = E_n(\mathbf{k}, \mathbf{r}) + \Psi(\mathbf{r}).
\]

\( \mathcal{E}_n(\mathbf{k}, \mathbf{r}) \) is the total energy of the \( n \)th band under the influence of an external
4.2. Equilibrium carrier densities

potential. The exact Schrödinger equation for the crystal in a force field is

\[ \mathcal{H}\Phi(k) = \mathcal{E}\Phi(k), \]  

(4.3)

with a Hamilton operator of the form

\[ \mathcal{H} = -\frac{\hbar^2}{2m} \nabla^2 + \mathcal{U}(r) + \Psi(r), \]  

(4.4)

where \( \mathcal{U} \) is the crystal potential operator. If the externally applied forces are slowly varying over the dimension of the crystal unit cell, Eq. (4.3) transforms into the equivalent Schrödinger equation

\[ \mathcal{H}_n A(r) = \mathcal{E}_n A(r), \]  

(4.5)

where \( A(r) \) are Wannier functions, and \( \mathcal{H}_n \) is the effective Hamiltonian for a band with the index \( n \), i.e.

\[ \mathcal{H}_n = -\frac{\hbar^2}{2} \nabla \left( \frac{1}{m^*_n(r)} \nabla \right) + \Psi(r). \]  

(4.6)

Equation (4.6) has the form \( \mathcal{H}_n = E_n(-i\nabla, r) + \Psi(r) \). Application of the correspondence principle \((-i\nabla \rightarrow \mathbf{k})\) yields

\[ \mathcal{H}_n(k, r) = E_n(k, r) + \Psi(r) = \mathcal{E}_n(k, r). \]  

(4.7)

The Hamiltonian in Eq. (4.7) contains no explicit potential term for the periodic crystal background potential. Electrons (or holes) are described as particles with a spatially varying effective mass \( m^*_n \) which is different from the mass \( m \) of the free electron. The effective mass depends on the material and the particular band \( n \) which is assumed to be parabolic around the minimum of \( E_n \) at \( k = k_0 \).

Modifications of the bandstructure due to nonuniform lattice spacing according to locally varying materials is described by the band offset \( \delta E_n(r) = \Delta E_n(r) \). Usually, the band offset is given for the conduction band (conduction-band offset). The valence-band offset can be determined with the bandgap \( \mathcal{E}_g \). The potential energy of the externally applied forces in semiconductor devices is the potential energy of the electrostatic field, i.e.

\[ \Psi(r) = -q\phi(r), \]  

(4.8)

where \( \phi \) is the electrostatic- or Hartree potential. Now, the effective Hamiltonian in Eq. (4.7) is written

\[ \mathcal{E}_n(k, r) = E_{n0}(k) + \Delta E_n(r) - q\phi(r). \]  

(4.9)
For some fixed point in k-space, usually a minimum (valley), Eq. (4.9) can be written as

$$\mathcal{E}_n(r) = E_n(k_n, r) - q\phi(r), \quad n = c, v; \quad \nu = 1, \ldots, \quad (4.10)$$

where c refers to the lowest conduction band, v to the highest valence band and \(\nu\) is the valley index.

Figure 4.1 shows the energy band diagram for two materials with different bandstructure. The reference for the band diagram is the infinite- or force-free vacuum level \(E_0\). It is defined as the energy an electron would have if it is completely free from the influence of the material or the applied external forces. The local vacuum level \(\mathcal{E}_1\) is defined as the vacuum level in the presence of forces, i.e., in the presence of an electrostatic field

$$\mathcal{E}_1(r) = E_0 - q\phi(r). \quad (4.11)$$

Therefore, in the absence of forces \(E_0 = \mathcal{E}_1 = E_1\). The energy needed to excite an electron from the conduction band to the local vacuum level is the electron affinity \(\chi\). It accounts for the work against image forces and surface dipoles. The work function \(\Phi\) of the semiconductor is defined as the change in free energy of the material if an electron is added or taken away from the system. The work function changes with the doping. Therefore, it is more practicable to characterize the material using the electron affinity. The conduction-band offset \(\Delta E_c\) used in the band energy, Eq. (4.9), is calculated

---

**Figure 4.1:** Energy band diagram for nonuniform material before (left) and after (right) contact. The different chemical potentials \(\mu_1\) and \(\mu_2\) align to a constant Fermi energy \(E_F\) in equilibrium. The potential \(\phi\) is the built-in potential if no external potentials are applied.
4.2. Equilibrium carrier densities

as

$$\Delta E_c(r) = \chi_1(r) - \chi_2(r), \quad (4.12)$$

(electron-affinity rule) where $\chi_1$ and $\chi_2$ are the electron affinities of two materials building a heterojunction at $r$.

Before contact, each of the materials is characterized by a chemical potential $\mu$. The effect of the electrostatic potential on the energy levels due to $\Psi(r) = -q\phi(r)$ is described by a locally varying chemical potential, the electrochemical potential $\zeta$ which is defined as follows

$$\zeta(r) = \mu(r) - q\phi(r). \quad (4.13)$$

Before contact, the materials are characterized by different chemical potentials, $\mu$, electron affinities, $\chi$, and energy gaps $\mathcal{E}_g = \mathcal{E}_c - \mathcal{E}_v$ (Fig. 4.1). After contact, the difference in the chemical potentials leads to electron transfer from right-hand to the left-hand side minimizing the free energy and establishing thermodynamic equilibrium characterized by an uniform electrochemical potential at a Fermi energy $E_F = \zeta = \text{const}$. In $n$-doped materials the electron transfer ionizes donors on the right-hand side and creates excess electrons on the left-hand side. An internal electrostatic potential is induced across the junction which is flat in regions far from the junction where charge neutrality holds. The potential difference between the two materials can be understood as a potential drop (voltage) which is built-in into the junction because of its internal composition (material composition, doping). It is especially not an externally applied voltage.

### 4.2.2 Quasi-equilibrium approximation

The electrostatic potential $\phi(r)$ in a semiconductor structure depends on the charge $\rho(r)$. In general, $\rho$ also depends on the current flow. The self-consistent evaluation of potential, carrier densities and current densities is the fundamental problem of device theory. However, in many cases an adequate estimate for the densities and the potential can be calculated by neglecting the current. This is called quasi-equilibrium approximation. It is assumed that the device can be subdivided into regions, each of which is locally in thermal equilibrium with a Fermi level set by the voltage of a contact to which the region is connected. This assumption holds exactly if no external voltages are applied to the contacts and the structure is naturally in thermal equilibrium. If different
voltages are applied to the contacts, the regions with different Fermi levels have to be separated by insulating regions where the carrier densities are zero and no current flow is possible.

In the quasi-equilibrium approximation the system of equations for the potential, the carrier densities, and the current densities given in Section 1.2 reduces to one equation for the electrostatic potential and the carrier densities, the *non-linear Poisson equation*

\[-\nabla \cdot (\varepsilon \nabla \phi) = \rho[\phi],\]  

(4.14)

which is a global equation, i.e. which is solved for the whole device domain subject to appropriate boundary conditions. The total charge density is

\[\rho[\phi] = q(p[\phi] - n[\phi] + N^+_d[\phi] - N^-_a[\phi]),\]  

(4.15)

where \(p\) is the density of holes, \(n\) the density of the electrons, and \(N^+_d, N^-_a\) are the densities of the ionized impurities. The equilibrium carrier densities for electrons and holes in bulk regions are given by the following state equations for carriers in parabolic bands

\[n(r) = N_c \mathcal{F}_{1/2}(\eta_c(r)), \quad \eta_c(r) = \frac{E_F - E_c(r)}{k_B T},\]  

(4.16a)

\[p(r) = N_v \mathcal{F}_{1/2}(\eta_v(r)), \quad \eta_v(r) = \frac{E_v(r) - E_F}{k_B T},\]  

(4.16b)

where \(E_c\) is the conduction band edge, which was previously defined as

\[E_c(r) = E_{c0}(k_\nu) + \Delta E_c(r) - q\phi(r).\]  

(4.17)

A particular minimum \(\nu\) in the conduction band is characterized by the spatially invariant term \(E_{c0}(k_\nu)\). The valence-band edge is given by \(E_v = E_g - E_c\) and \(\mathcal{F}_{1/2}\) is a statistical distribution function which in this particular case is the generic Fermi-Dirac integral of order 1/2. The general form of the Fermi-Dirac integral of order \(\alpha\) is

\[\mathcal{F}_\alpha(\eta) = \begin{cases} \frac{1}{\Gamma(\alpha + 1)} \int_0^\infty \frac{t^\alpha \exp(-t\eta)}{\exp(t - \eta) + 1} \, dt, & \alpha > -1 \\ \frac{1}{1 + \exp(-\eta)}, & \alpha = -1 \end{cases}\]  

(4.18)

\[\mathcal{F}'_\alpha(\eta) = \mathcal{F}_{\alpha-1}(\eta), \quad \alpha \geq 0.\]
4.2. Equilibrium carrier densities

The factor $N_c$ is the *effective conduction-band density of states*:

$$N_c = 2 \left( \frac{2\pi m^*_n k_B T}{\hbar^2} \right)^{3/2} M_c, \quad (4.19)$$

with $M_c$ the number of energetically degenerate conduction band valleys and $N_v$ is the *effective valence-band density of states*:

$$N_v = 2 \left( \frac{2\pi m^*_p k_B T}{\hbar^2} \right)^{3/2} M_v, \quad (4.20)$$

where $M_v$ is the number of energetically degenerate valence band valleys. It is not useful to introduce an intrinsic density $n_i$ or an intrinsic Fermi level $E_i$ since these quantities are not constant through a heterostructure. Furthermore, the mass action law $pn = n_i^2$ is not valid in a degenerate semiconductor.

The total charge density, Eq. (4.15), also contains contributions from the ionized impurity densities of donors $N_{d^+}$ and of acceptors $N_{a^-}$. The state equations for these quantities are given as

$$\frac{N_{d^+}}{N_d} = \left[ 1 + g_d \exp \left( \frac{E_F - \mathcal{E}_d(r)}{k_B T} \right) \right]^{-1}, \quad (4.21a)$$

$$\frac{N_{a^-}}{N_a} = \left[ 1 + g_a \exp \left( \frac{\mathcal{E}_a(r) - E_F}{k_B T} \right) \right]^{-1}, \quad (4.21b)$$

where $g_d$ and $g_a$ are the degeneracy factors for donors and acceptors. The energies of the donor levels $\mathcal{E}_d$ and $\mathcal{E}_a$ are defined with respect to the same energy scale as the conduction band edge $\mathcal{E}_c$ and the valence band edge $\mathcal{E}_v$.

The screening equation for the structure is obtained by combining all equations in this section into the non-linear Poisson equation, Eq. (4.14), which is solved for the electrostatic potential $\phi(r)$. Material parameters such as $\Delta E_c$ are determined by the design of the structure and the Fermi energy is fixed by the external circuitry. The solution $\phi(r)$ has local screening properties: increasing the potential decreases the charge density and vice versa.
4.2.3 Boundary conditions and Fermi level

The Poisson equation is solved within a bounded domain subject to the appropriate boundary conditions. In principle, two different kinds of boundaries are distinguished. First, real physical boundaries to contacts and interfaces are represented by Dirichlet boundary conditions. Another boundary type is due to the limited size of the simulation domain. This second category of artificial boundaries \((\partial \Omega_a)\) represents no boundary in the physical sense and is only introduced to enable the simulation. Usually, Neumann boundary conditions, i.e. the condition of zero gradient of the electrostatic potential normal to the boundary are used, i.e.

\[
\mathbf{n} \nabla \phi(\mathbf{r}) = 0, \quad \mathbf{r} \in \partial \Omega_a, \tag{4.22}
\]

where \(\mathbf{n}\) is the unit normal vector of the boundary. Therefore, it is guaranteed that no currents are flowing in and out the device except contact currents.

The physical boundary conditions are either for Ohmic contacts \((\partial \Omega_o)\) or for Schottky contacts \((\partial \Omega_s)\). For the \(i\)th voltage controlled Ohmic contact the boundary condition is

\[
\phi^{(i)}(\mathbf{r}) - \phi_{bi}^{(i)}(\mathbf{r}) - \phi^{(i)}_D = 0, \quad \mathbf{r} \in \partial \Omega_o. \tag{4.23}
\]

where \(\phi^{(i)}_D\) is the externally applied bias, and \(\phi_{bi}^{(i)}\) is the built-in potential which is defined through a local charge neutrality condition using Eq. (4.16) and Eq. (4.21), i.e.

\[
0 = p(\mathbf{r}) - n(\mathbf{r}) + N^+_d(\mathbf{r}) - N^-_a(\mathbf{r}), \quad \mathbf{r} \in \partial \Omega_o. \tag{4.24}
\]

The terminals, together with the circuit nodes act as charge reservoirs with given Fermi levels which are determined by the applied voltages. Terminals which are connected to the substrate via Ohmic contacts are assumed to be in equilibrium with the substrate, i.e. the Fermi level in the substrate region connected to the Ohmic contact is pinned through the applied voltage such that

\[
E_F = -q\phi^{(i)}_D. \tag{4.25}
\]

In general, different regions connected to different terminals can have different Fermi levels. However, since a quasi-equilibrium approximation has been made, these regions then have to be separated by insulating regions. Otherwise, a current would flow and the quasi-equilibrium assumption would be violated.
4.2. Equilibrium carrier densities

For Schottky barriers the following Dirichlet boundary condition for the electrostatic potential is used

\[ \phi^{(i)}(r) - \phi^{(i)}_{\text{bi}}(r) + \phi^{(i)}_s - \phi^{(i)}_D = 0, \quad r \in \partial \Omega_s, \quad (4.26) \]

where \( \phi^{(i)}_s \) is the Schottky barrier of the \( i \)th contact given as the difference between the work function of the metal and the electron affinity of the material the contact is connected to (Schottky model),

\[ \phi^{(i)}_s = \Phi^{(i)}_m - \chi. \quad (4.27) \]

The height of the Schottky barrier is a characteristic quantity for the semiconductor/metal contact. Biased Schottky contacts are not generally in equilibrium with the substrate. The Fermi level in the terminal node, which is given by the applied voltage \( \phi^{(i)}_D \), is in general different from the substrate Fermi level, which is given by some ground electrode (Ohmic contact) far away from the Schottky contact. Additionally to the Poisson equation, the current equations have to be solved to get the bending of the quasi-Fermi potential at the Schottky contact and the contact current right.

4.2.4 Low dimensional electron gases

Kohn-Sham system

The electron charge in bulk regions is given by Eq. (4.16a) with the three dimensional semi-classical effective conduction band density of states \( N_c \). This approximation is only valid for electrons in parabolic energy band extrema, i.e. in bands with a quadratic dispersion relationship. The charge density for localized levels in the bandgap due to carrier confinement in one or more space dimensions has to be calculated using a different representation of the density of states. Electrons in systems with confinement are described by an equation of the form

\[
\left\{-\frac{\hbar^2}{2m^*} \nabla \cdot \left( \frac{1}{m^*} \nabla \right) + \mathcal{V}_{\text{eff}}(r) - E_i\right\} \psi_i(r) = 0, \quad \text{in } \Omega, \quad (4.28)
\]

where the effective potential \( \mathcal{V}_{\text{eff}}(r) \) for electrons is given as

\[ \mathcal{V}_{\text{eff}}(r) = \Delta E_c(r) - q\phi(r) + V_{xc}(r). \quad (4.29) \]
Figure 4.2 schematically shows a possible subdivision of the simulation domain. For the interface between the semi-classical region $\hat{\Omega}$ and the confinement region $\Omega$ proper boundary conditions for Eq. (4.28), which is only solved in region $\Omega$, have to be supplied. Homogeneous Dirichlet boundary conditions for the eigenfunctions $\psi_i$ are suitable for confined carriers. Consequently, the carrier densities vanish at the boundary of $\Omega$ and a depletion zone is surrounding the confined region. Naturally, the eigenfunctions should not feel the boundary conditions too much.

Although Eq. (4.28) looks like a Schrödinger equation, it is what is called a single-band effective-mass equation (Ben-Daniel-Duke form [35]). The crystal potential is periodic on the atomic scale and does not appear explicitly. It is incorporated through the position dependent effective mass for the conduction band $m_c^*$ and the position dependent conduction band offset $\Delta E_c$. Equation (4.28) is an effective single-particle equation for the $i$th single-particle level characterized by an eigenvalue $\mathcal{E}_i = \mathcal{E}_i(V_{\text{eff}})$ and the corresponding envelope wavefunction $\psi_i = \psi_i(V_{\text{eff}})$.

The electron screening is treated within a mean-field approximation, i.e. each electron moves in the effective potential produced by all other electrons. The form of the effective potential in Eq. (4.29), however, goes beyond the popular Hartree approximation by incorporating not only the classical interaction due to the electrostatic potential energy, $-q\phi$, but an additional term $V_{\text{xc}}$ which accounts for the quantum mechanical interactions due to exchange- and correlation. The Schrödinger type Eq. (4.28) with an appropriate formulation for the exchange- and correlation potential is known within the framework of density-functional theory (DFT) as the Kohn-Sham equation [51]. A short account of DFT is given in Appendix C.

Density-functional theory states that the two potential terms $-q\phi[n](r)$ and $V_{\text{xc}}[n](r)$ are unique functionals of the electron charge density [49]. The charge density is compiled from the single-particle wavefunctions according
4.2. Equilibrium carrier densities

to the quantum mechanical expression

$$n(V_{eff})(r) = \sum_{i=0}^{\infty} N_i(V_{eff})|\psi_i(V_{eff})(r)|^2, \quad r \in \Omega. \quad (4.30)$$

The numbers $N_i$ are the occupation factors according to

$$N_i(V_{eff}) = f(\mathcal{E}_i(V_{eff}) - \mu), \quad (4.31)$$

where $\mu$ denotes the chemical potential, and $f$ is the thermodynamic equilibrium distribution function for electrons. The distribution function takes different forms depending on the reduced dimension of the electron gas. With $d$, the dimensionality of the electron gas, this is

$$f(s) = c\mathcal{F}_\alpha \left( -\frac{s}{k_B T} \right), \quad \alpha = \begin{cases} -1 & \text{if} \quad d = 3 \\ -\frac{1}{2} & \text{if} \quad d = 2 \\ 0 & \text{if} \quad d = 1 \end{cases}$$

where $c$ is some positive constant, $T$ the temperature of the electron gas and $s$ is defined as the difference of the eigenvalue $\mathcal{E}_i$ corresponding to the wavefunction $\psi_i$ used in Eq. (4.30) and the chemical potential, i.e. $s = \mathcal{E}_i(V_{eff}) - \mu$. The chemical potential for a given electron number, $\mu(N)$, is defined by the charge conservation law, i.e.

$$N = \int_{\Omega} d\mathbf{r} \ n(V_{eff})(r) = \sum_{i=0}^{\infty} f(\mathcal{E}_i(V_{eff}) - \mu(N)). \quad (4.32)$$

Electron densities

Confinement is possible in one space dimension (quantum wells: 2DEG), in two dimensions (quantum wires: 1DEG) or in all three space dimensions (quantum dots: 0DEG). For zero-dimensional electron gases in quantum dots, i.e. $\Omega \subset \mathbb{R}^3$, the distribution function is given by the Fermi-Dirac distribution

$$f(s) = 2g\mathcal{F}_{-1} \left( -\frac{s}{k_B T} \right) = \frac{2g}{1 + \exp \left( \frac{s}{k_B T} \right)}, \quad (4.33)$$

with $\mathcal{E}_i = \varepsilon_i$, the single-particle energy of the $i$th quantum dot level. For one-dimensional electron gases in quantum wires, i.e. $\Omega \subset \mathbb{R}^2$, the distribution
function is

\[ f(s) = 2g \left( \frac{m^*_{\perp} k_B T}{2\pi h^2} \right)^{\frac{1}{2}} \mathcal{F}_{-\frac{1}{2}} \left( \frac{-s}{k_B T} \right) \]

\[ = \frac{g}{\pi h} \left( 2m^*_{\perp} k_B T \right)^{\frac{1}{2}} \int_0^\infty t^{-\frac{1}{2}} \frac{dt}{1 + \exp \left( t + \frac{s}{k_B T} \right)} , \quad (4.34) \]

with \( E_i = \epsilon_i \), the position dependent subband energy of the \( i \)th one-dimensional subband. The Schrödinger equation, Eq. (4.28), is solved in slices alongside the 1D-channel and \( m^*_{\perp} = m^*_{\perp,n} \) is the conduction band effective mass component perpendicular to \( \Omega \). For a two-dimensional electron gas in quantum wells, i.e. \( \Omega \subset \mathbb{R}^1 \), the distribution function is given as

\[ f(s) = 2g \frac{m^*_{\perp} k_B T}{2\pi h^2} \mathcal{F}_0 \left( \frac{-s}{k_B T} \right) = g \frac{m^*_{\perp} k_B T}{\pi h^2} \ln \left( 1 + \exp \left( \frac{-s}{k_B T} \right) \right) , \quad (4.35) \]

with \( E_i = \epsilon_i \), the position dependent subband energy of the \( i \)th two-dimensional subband. The Schrödinger equation, Eq. (4.28), is solved one-dimensionally and \( m^*_{\perp} = m^*_{\perp,n} \) is the component of the conduction band effective mass perpendicular to \( \Omega \). The factor 2 arises from the spin degeneracy of the eigenstates. \( g \) is the valley degeneracy of the particular material. Two-dimensional systems with one-dimensional confinement are intensively studied for 2DEGs in semiconductor inversion and accumulation layers (see for instance [30]).

Quantum dots are regions with confinement in all three space dimensions. In some structures, the electrons are confined in tiny areas of typical sizes below a few hundred nanometers (typical for GaAs systems are \( \sim 100 \) nm). The energy level spectrum is discrete and the spacing of the levels can be larger than there thermal broadening. The Fermi-Dirac distribution, which was used in the previous section as a distribution function for electrons amongst the levels, is not a good approximation any more. The electrons in the quantum dot are isolated from the reservoirs, and therefore, the definition of a Fermi energy for the dot electrons is somewhat meaningless. The electron gas in a quantum dot is described by \textit{canonic ensemble} theory. The electron number is fixed at an integer value \( N \) and the occupation factors \( N_i \) for the electrons in an effective potential \( V_{\text{eff}} \) are determined by the Gibbs distribution for localized
4.2. Equilibrium carrier densities

states \(g(\varepsilon_i | N)\) as already introduced in Chapter 3 (Eq. (3.12)), i.e.

\[
N_i = \frac{2g}{Z(N)} \sum_{\{n_k\}} \exp \left( -\frac{1}{k_B T} \sum_k n_k \varepsilon_k \right) \delta_{n_i,1} \delta_{N,\sum n_k},
\]

(4.36)

where \(Z(N)\) is the \(N\)-dependent partition function in the canonic ensemble theory

\[
Z(N) = \sum_{\{n_k\}} \exp \left( -\frac{1}{k_B T} \sum_k n_k \varepsilon_k \right) \delta_{N,\sum n_k},
\]

(4.37)

and \(\mathcal{F}(N) = -k_B T \ln(Z(N))\) is defined as the free energy of the internal degrees of freedom. Again, the factor 2 arises from the spin degenerate nature of the eigenstates \(\varepsilon_i\), \(g\) is the valley degeneracy of the particular material and \(T\) the temperature of the electron gas. In practical calculations, the Fermi-Dirac distribution, Eq. (4.33), can be used instead since for larger electron numbers in highly populated quantum dots the deviations of the Fermi-Dirac distribution from the more general Gibbs distribution is small [31]. However, the difference can be significant for small numbers of \(N\) (see Appendix A).

In some cases, the electron gas in a quantum dot is strongly confined in one dimension, and therefore, its quasi-2D nature allows a shortcut to the very expensive solution of the 3D-Schrödinger equation. The wavefunction \(\psi(r)\) is separable and an one-dimensional equation is solved in the direction of the strong confinement (which is taken to be the \(z\)-direction for instance) at every point in the \(xy\)-plane, i.e.

\[
\left[ -\frac{\hbar^2}{2} \frac{\partial}{\partial z} \left( \frac{1}{m^*_z} \frac{\partial}{\partial z} \right) + V_{\text{eff}}(x, y, z) \right] \xi_n^{x,y}(z) = \epsilon_n(x, y) \xi_n^{x,y}(z).
\]

(4.38)

Assuming only a single subband \(\epsilon_n(x, y)\), the index \(n\) can be dropped and the charge distribution in the \(xy\)-plane is calculated with the effective potential \(\epsilon(x, y) = \epsilon_0(x, y)\), i.e.

\[
\left[ -\frac{\hbar^2}{2} \sum_{i \in \{x,y\}} \frac{\partial}{\partial i} \left( \frac{1}{m^*_i} \frac{\partial}{\partial i} \right) + \epsilon(x, y) \right] \phi_m(x, y) = \varepsilon_m \phi_m(x, y).
\]

(4.39)

The \(m\)th single-particle wavefunction within the single-subband approximation is \(\psi_m(r) \rightarrow \psi_m^{0}(x, y, z) = \phi_m(x, y) \xi_0^{x,y}(z)\) and \(\varepsilon_m\) is the energy of the corresponding single-particle level.
4.3 Total free energies

4.3.1 Equilibrium thermodynamics

The first and second law of thermodynamics are condensed in an equation for the total internal energy \( E \) of a system (Euler's equation) such that

\[
E = TS - PV + \sum_i \mu_i N_i, \tag{4.40}
\]

where the intensive variables (not depending on the particle concentration) temperature \( T \), pressure \( P \), and chemical potential \( \mu_i \) are related to the extensive variables (depending on the particle concentration) entropy \( S \), volume \( V \), and particle number \( N_i \). The index of the chemical potential \( \mu_i \) and of the particle number \( N_i \) refers to a particular species of particles in the system. With Eq. (4.40) the *Helmholtz free energy* can be defined as

\[
F = E - TS, \tag{4.41}
\]

and the *Gibbs free energy* as

\[
G = E - TS + PV, \tag{4.42}
\]

both of which have minimal properties at thermal equilibrium if \( T \) and \( N_i \) are constant. \( F \) is at minimum for constant volume and \( G \) for constant pressure. In an isothermal transformation the change of the Helmholtz free energy is completely transformed into work. Using Eq. (4.40) a system consisting of a single species of particles with particle number \( N \) and the total charge \( Q = -qN \) has the free energy

\[
F = -PV + (\mu - q\phi)N. \tag{4.43}
\]

The Fermi energy \( E_F \) is defined as the change in free energy of the crystal when an electron is added or taken away

\[
\left( \frac{\partial F}{\partial N} \right)_{T,V} = E_F = \mu - q\phi, \tag{4.44}
\]

where \( \phi \) is the internal electrostatic potential energy (built-in potential) of the system. The built-in potential is zero in equilibrium with no excess carriers present. Only in this case the Fermi energy equals the chemical potential. The
Fermi energy is constant in thermal equilibrium. In a non-equilibrium situation (externally applied forces) the (non-equilibrium) Fermi energy is often referred to as quasi-Fermi potential \( \zeta = \mu - q\phi \), where \( \phi \) is the sum of the internal- and the external potential. The property \( \Omega = F - E_F N \) is the grand canonical potential.

### 4.3.2 Helmholtz free energy and external forces

To calculate the total (Helmholtz) free energy of a charged system consisting of \( i = 1 \ldots M \) distinct elements with the total equilibrium charges \( Q_i \) and the voltages \( V_i \) the following semi-classical expression is used:

\[
F = \frac{1}{2} \sum_{i=1}^{M} Q_i V_i - \sum_{i=1}^{M} \int_0^t dt' \, I_i(t') V_i(t'),
\]

where the \( I_i \) are the currents provided by the external circuitry (voltage sources). The first term in Eq. (4.45) is the capacitive (potential-) energy \( U \) stored in the device. Voltages and charges are related by the elements of the symmetric capacitance matrix \( C_{ij} \), i.e. the charge at the \( i \)th element is related to the voltages \( V_j \) at all elements such that

\[
Q_i = \sum_{j=1}^{M} C_{ij} V_j.
\]

The diagonal elements of the capacitance matrix are the positive self-capacitances. The negative off-diagonal elements are mutual capacitances or coefficients of induction which account for the amount of change in the \( i \)th charge due to the change in the \( j \)th voltage if all other voltages are kept constant. \( \{C_{ij}\} \) is the capacitance matrix with \( M(M - 1)/2 \) independent elements.

The device is separated into elements which contain a continuous space charge \( \rho \) (dot, leads, donor layer) and metal plates which are equipotential regions (gates, contacts). The electrostatic energy \( U \) is written as

\[
U = \frac{1}{2} \int_{\Omega} d\mathbf{r} \, \rho(\mathbf{r}) \phi(\mathbf{r}) + \frac{1}{2} \sum_{i \in \text{gates}} Q_i V_i.
\]

The potential \( \phi \) is the electrostatic potential, i.e. the solution of the Poisson equation with the charge density \( \rho \), subject to Dirichlet boundary conditions at
gates and contacts. The electrostatic potential can also be expressed in terms of the Greens function of the Poisson equation, \( G(\mathbf{r}, \mathbf{r}') = 1/\left(\varepsilon(\mathbf{r}')|\mathbf{r} - \mathbf{r}'|\right) \), i.e.

\[
\phi(\mathbf{r}) = \frac{1}{2} \int_{\Omega} d\mathbf{r}' G(\mathbf{r}, \mathbf{r}') \rho(\mathbf{r}').
\]  
(4.48)

The integral over the space charges replaces the summation over the \( M \) charged elements in Eq. (4.45). The energy contribution from the gates which are kept at constant voltages \( V_i \) is calculated with the surface charges \( Q_i \) which are obtained using Gauss law, i.e.

\[
Q_i = \int_{\partial \Omega_i} d\mathbf{r} \sigma(\mathbf{r}) \quad \text{with} \quad \sigma(\mathbf{r}) = -\varepsilon(\mathbf{r}) \frac{\partial \phi(\mathbf{r})}{\partial \mathbf{n}_f},
\]  
(4.49)

where \( \sigma(\mathbf{r}) \) is the surface charge density and \( \mathbf{n}_f \) is the unit surface normal vector of the \( i \)th gate at \( \partial \Omega_i \).

The second term in Eq. (4.45) is the work \( W \) which is needed to charge all elements which are connected to external circuitry (such as the gates and the source and drain region) when \( N \) electrons are transferred to the quantum dot region. It is assumed that the charging is fast compared with the internal relaxation times, i.e. the tunneling time \( \tau^1 \). Voltage sources are considered to be ideal with zero internal resistances (low-impedance environment). Therefore, it is reasonable to assume that the voltages are kept at constant level before and after the charging \( V_i(t) = V_i(0) \). Using \( I = dQ/dt \) the current integral in Eq. (4.45), i.e. the work \( W_i \) for the charging of the \( i \)th element is written

\[
W_i = V_i \int_0^t dt' I_i(t') = V_i [Q_i(t) - Q_i(0)] = V_i Q_i.
\]  
(4.50)

The charge at the \( i \)th gate at \( t = 0 \) can be arbitrarily set to \( Q_i(0) = 0 \). Combining these results, the final form of the semi-classical Helmholtz free energy is given as

\[
F = \frac{1}{2} \sum_{i \in \text{leads}} \int_{\Omega_i} d\mathbf{r} \rho(\mathbf{r}) \phi(\mathbf{r}) - \frac{1}{2} \sum_{i \in \text{gates}} \int_{\Omega_i} d\mathbf{r} \rho(\mathbf{r}) \phi(\mathbf{r}) - \frac{1}{2} \sum_{i \in \text{gates}} Q_i V_i.
\]  
(4.51)

\(^1\)This again is not the tunneling duration \( \tau_t \), moreover it is the system relaxation time, i.e. the time between the tunneling events and \( \tau_t \ll \tau \) (see Chapter 2).
4.3. Total free energies

The work needed to recharge the $i$th element during a single tunneling event in transition of the dot from the $N$-particle to the $(N + 1)$-particle system can be calculated in a similar manner, i.e.

$$\Delta W_i = \int_0^T dt' I_i(t') V_i(t') = \Delta Q_i V_i,$$  \hspace{1cm} (4.52)

where $\Delta Q_i = Q_i(N + 1) - Q_i(N)$. The charges $Q_i$ are surface charges calculated using Eq. (4.49) if the $i$th element is a metal gate.

4.3.3 Helmholtz free energy and charge quantization

In the case of bound states in a zero-dimensional electron gas the free energy has to be modified including the influence of the confined states on the total energy of the quantum dot [42], [43], [45]. The Helmholtz free energy, Eq. (4.45), is extended by two terms, i.e.

$$F(N) = \sum_{\{n_k\}} n_k \varepsilon_k^0 + E_{xc} + \frac{1}{2} \sum_{i=1}^M Q_i V_i - \sum_{\{n_k\}} \int_0^t dt' I_i(t') V_i(t').$$  \hspace{1cm} (4.53)

The first sum over the bare (non-interacting) dot levels $\varepsilon_k^0$ with a particular occupation configuration $\{n_k\}$ is the interaction-free kinetic energy of the quantum dot $T_s$. The exchange-correlation energy $E_{xc}$ of the electrons in the local-density approximation (LDA) corresponds to a (local) exchange-correlation potential $V_{xc}$ according to Eq. (C.11) (see Appendix C for a discussion of the LDA). The last two terms are the same as in Eq. (4.45) in the previous section. The total energy contribution of the quantum dot electrons to the Helmholtz free energy is

$$E_{tot}(\{n_k\}, Q_{dot}, V_{dot}) = \sum_{k} n_k \varepsilon_k^0 + \frac{1}{2} Q_{dot} V_{dot} + E_{xc}. $$  \hspace{1cm} (4.54)

From the Kohn-Sham equation, Eq. (4.28), with the Kohn-Sham effective potential, Eq. (4.29), the self-consistent energies are written as

$$\varepsilon_k = \langle \psi_k | - \frac{\hbar^2}{2m^*} \nabla \cdot \left( \frac{1}{m^*} \nabla \right) + \Delta E_c(r) - q\phi(r) + V_{xc}(r) | \psi_k \rangle. $$  \hspace{1cm} (4.55)

With the definition of $\varepsilon_k^0$ this results in

$$\varepsilon_k = \varepsilon_k^0 + \langle \psi_k | - q\phi(r) | \psi_k \rangle + \langle \psi_k | V_{xc}(r) | \psi_k \rangle. $$  \hspace{1cm} (4.56)
Subsequent summation over $k$ after weighting with the occupation numbers $n_k$ leads to the following expression for the interaction-free kinetic energy

\[ T_s({n_k}) = \sum_k n_k \varepsilon_k + \sum_k n_k \left( q \int_{\Omega} d\mathbf{r} |\psi_k|^2 \phi(\mathbf{r}) - \int_{\Omega} d\mathbf{r} |\psi_k|^2 V_{xc}(\mathbf{r}) \right). \]

If the occupation configuration dependence of the last two terms is ignored and the discrete occupation numbers $n_k$ are replaced by the non-integer occupation numbers according to the Gibbs- or Fermi-Dirac distribution, one obtains

\[ T_s({n_k}) = \sum_k n_k \varepsilon_k + q \int_{\Omega} d\mathbf{r} n(\mathbf{r}) \phi(\mathbf{r}) - \int_{\Omega} d\mathbf{r} n(\mathbf{r}) V_{xc}(\mathbf{r}). \quad (4.57) \]

It is not straightforward to find a vindication for this approximation. However, small thermal variations in the level occupancies have a negligible effect on the self-consistent results ($\varepsilon_k, \phi, \rho,$ and $Q_i$) which are implicit functions of the electron number and the applied voltages [41], [43]. This is especially true for low temperatures, where the level spacing is hardly affected by the temperature. The above result is now inserted into Eq. (4.54) where the electrostatic energy of the quantum dot is replaced according to

\[ \frac{1}{2} Q_{dot} V_{dot} \rightarrow -\frac{1}{2} q \int_{\Omega} d\mathbf{r} n(\mathbf{r}) \phi(\mathbf{r}). \quad (4.58) \]

The final expression for $E_{tot}$ becomes

\[ E_{tot}({n_k, \rho, \phi}) = \sum_k n_k \varepsilon_k + \frac{1}{2} q \int_{\Omega} d\mathbf{r} n(\mathbf{r}) \phi(\mathbf{r}) \]

\[ + \int_{\Omega} d\mathbf{r} n(\mathbf{r}) V_{xc}(\mathbf{r}). \quad (4.59) \]

The integration is over the quantum dot area $\Omega$ only and the electron density $n$ is positive. The second term in Eq. (4.59) is the negative electrostatic energy of the electrons. This is because the electrostatic energy is counted twice if the sum over the self-consistent particle levels from a mean field theory as the Kohn-Sham theory is included in the total energy calculation. The last sum provides a similar correction for the exchange- and correlation energy.

The total energy of the quantum dot, $E_{tot}$, replaces the corresponding terms in the Helmholtz free energy given by Eq. (4.53). This leads to the
4.3. Total free energies

The final form of the Helmholtz free energy, \( F(\{n_k\}, N) \), which is used to compute the Gibbs distribution, Eq. (3.9), for the quantum dot occupation and the two-terminal linear-response conductance, Eq. (3.37). In the low temperature limit, \( n_k = 1 \) for all \( k \leq N \) and \( n_k = 0 \) otherwise, i.e. the electrons occupy only the lowest states in the dot and the free energy can be denoted \( F_0(N) \) with \( N \) being an integer number and \( \sum_k n_k = N \).

4.3.4 Free energy minimization

In the following, the method used to minimize the free energy to determine the equilibrium number of electrons in the dot and the conductance is discussed. The chemical potential \( \mu \) of the quantum dot is ramped and the self-consistent ground state and the free energy is calculated using the equations from the previous section for maybe five values of \( \mu \) above the reservoir Fermi level. The number of electrons in the quantum dot is generally non-integer (at finite temperatures) and the free energy values give a parabola if plotted as a function of \( \mu \). Now, the free energies at integer numbers of electrons, \( F(\{n_k\}, N) \), are determined by spline interpolation. These free energy values are used in the Gibbs distribution Eq. (3.11) to calculate the equilibrium number of electrons in the dot and the conductance according to Eq. (3.35) (note that the Fermi energy used in these equations is the reservoir Fermi energy).

![Figure 4.3](image)

Figure 4.3: Left: Coulomb blockade peak and discrete charging (dashed line) of a GaAs quantum dot at 1 K. Right: free energy curves at the conductance peak (a) and between two peaks (b). The zero of the free energy is set arbitrarily.
Figure 4.3 shows an example of a conductance peak and the discrete charging of a GaAs quantum dot. The electron number in the dot is increased from 62 to 63 by ramping the gate voltage. Also displayed are the two free energy parabola for the top of the conductance peak (a) and a point half-way to the next peak (b). It can be seen that the free energies for electron number 62 and 63 are degenerate at the top of the peak and consequently the equilibrium number of electrons is 62.5. Between the peaks, at point (b), the electron number is an integer value and the minimum of the parabola is to be found at \( N = 63 \).

4.4 Tunneling rates

4.4.1 Bardeen’s transfer Hamiltonian method

The transfer Hamiltonian formalism was introduced by Bardeen [36] in 1961. It has proven to be very successful in the description of a number of phenomena including tunneling between normal metals and superconductors and was even able to predict the Josephson effect [38]. The transfer Hamiltonian formalism is based on the separation of the system into subsystems. Consider a two reservoir/barrier system described by the following model Hamiltonian

\[ \mathcal{H} = \begin{cases} \mathcal{H}_L & \text{for } r \in \Omega_L, \\ \mathcal{H}_R & \text{for } r \in \Omega_R. \end{cases} \]  

(4.60)

\( \mathcal{H}_L \) and \( \mathcal{H}_R \) are the Hamiltonians for the subsystems. Each subsystem contains a reservoir connected to an electrode and a semi-infinite barrier Fig. (4.4). \( \Omega_L \) and \( \Omega_L \) denote the half-spaces for which the Hamiltonians \( \mathcal{H}_L \) and \( \mathcal{H}_R \) are adequate descriptions. The eigenstates and eigenfunctions of the two subsystems are given as the solutions to two Schrödinger equations

\[ \mathcal{H}_L \psi_0 = E_0 \psi_0, \]
\[ \mathcal{H}_R \psi_k = E_k \psi_k. \]  

(4.61)

In other words, it is assumed that the electron tunnels from a given state \( E_0 \) on the left hand side to a state \( E_k \) on the right hand side which is part of a given set of eigenstates of \( \mathcal{H}_R \). Now, the solution of the time-dependent Schrödinger equation with the model Hamiltonian

\[ i\hbar \frac{\partial}{\partial t} \psi(t) = \mathcal{H} \psi(t) \]  

(4.62)
Figure 4.4: (a): Schematic view of the tunnel junction. The electron tunnels from a given reservoir state \( E_0 \) to a discrete quantum dot state \( E_k \) (b), (c): Left and right subsystem characterized by the Hamiltonians \( \mathcal{H}_L \) and \( \mathcal{H}_R \), respectively.

is written as a linear combination of \( \psi_0 \) and the various \( \psi_k \) which are independent of time

\[
\psi(t) = a(t)\psi_0 e^{-\frac{i}{\hbar}E_0 t} + \sum_k b_k(t)\psi_k e^{-\frac{i}{\hbar}E_k t}.
\]  

(4.63)

At \( t = 0 \) one assumes \( \psi = \psi_0 \) and therefore, the expansion coefficients have to vanish, \( b(0) = 0 \). Furthermore, the expansion is restricted to first order in \( a \) which leads to \( a = 1 \) and \( \dot{a} = 0 \). Inserting the eigenfunction \( \psi(t) \) into the time-dependent Schrödinger equation yields the following equation for the expansion coefficient

\[
i\hbar \dot{b}_k = e^{-\frac{i}{\hbar}(E_k - E_0) t} M_k,
\]  

(4.64)
with the matrix element for the transition $M_{0\to k} = M_k$

$$M_k = \int_{\Omega} \mathrm{d}r \, \psi_k^*(H - E_0) \psi_0.$$  \hspace{1cm} (4.65)

The transition rate $\Gamma_{0\to k} = \Gamma_k$ follows from

$$\Gamma_k = \lim_{t\to \infty} |b_k(t)|^2 t^{-1}, \quad \text{i.e.} \quad \Gamma_k = \frac{2\pi}{\hbar} |M_k|^2 \delta(E_k - E_0). \hspace{1cm} (4.66)$$

The remaining task is to evaluate the matrix element, Eq. (4.65), which can be done if $\psi_0$ and the $\psi_k$ are known. Bardeen suggested a more useful form of the matrix element which is derived now. Using the model Hamiltonian, Eq. (4.60), in Eq. (4.65)

$$M_k \approx \int_{\Omega_R} \mathrm{d}r \, \psi_k^*(H_R - E_0) \psi_0, \hspace{1cm} (4.67)$$

where the integration is over the right half-space only since $\langle \psi_k | H - E_0 | \psi_0 \rangle$ vanishes on the left-hand side. Since $\langle \psi_k | H - E_k | \psi_0 \rangle$ vanishes on the right-hand side, the matrix element $M_k$ can be approximated by

$$M_k \approx \int_{\Omega_R} \mathrm{d}r \, [\psi_k^*(H - E_0) \psi_0 - \psi_0(H - E_k) \psi_k^*]$$

$$\approx \int_{\Omega_R} \mathrm{d}r \, [\psi_k^*H \psi_0 - \psi_0 H \psi_k^* - \psi_k^*E_0 \psi_0 + \psi_0 E_k \psi_k^*]$$

All transitions are considered to be elastic tunneling events, i.e. $\psi_0 E_k \psi_k^* = \psi_k^* E_0 \psi_0$ and energy is always conserved. If a model Hamiltonian of the form

$$\mathcal{H} = -\frac{\hbar^2}{2m^*} \nabla^2 + \mathcal{V}_{\text{eff}}(r) \hspace{1cm} (4.68)$$

(i.e. Kohn-Sham Hamiltonian$^2$) is assumed with the effective potential $\mathcal{V}_{\text{eff}}$ as a function of $r$ then $\psi_0 \mathcal{V}_{\text{eff}} \psi_k^* = \psi_k^* \mathcal{V}_{\text{eff}} \psi_0$ and the matrix element becomes

$$M_k \approx -\frac{\hbar^2}{2m^*} \int_{\Omega_R} \mathrm{d}r \, [\psi_k^* \nabla^2 \psi_0 - \psi_0 \nabla^2 \psi_k^*]. \hspace{1cm} (4.69)$$

$^2$Note that the form of the Hamiltonian Eq. (4.68) implies that the bandstructure is spatially constant, i.e. the effective mass $m^*$ is not spatially dependent. For an extension of the theory towards tunnel junctions with varying bandstructure the reader is referred to [37].
4.4. Tunneling rates

Using Green’s theorem the integral Eq. (4.69) can be evaluated, and the final form of the matrix element is the following surface integral

\[
M_k \approx -\frac{\hbar^2}{2m^*} \int_{\partial \Omega_R} df \left[ \psi_k \frac{\partial \psi_0}{\partial n} - \psi_0 \frac{\partial \psi_k}{\partial n} \right]
\]  
(4.70)

over the surface of the right-hand side half-space \( \Omega_R \) with the vector \( n \) being the unit normal vector of the surface, \( n \perp \partial \Omega_R \).

4.4.2 Adiabatic theory for quantum dot systems

The transfer Hamiltonian formalism [36] is applied in a straightforward manner to a reservoir/quantum dot system. The quantum dot is separated from the reservoir by a narrow quasi-1D channel. By applying a voltage to an electrostatic gate above the channel a narrow constriction, Fig. (4.5), is created. Since the constriction is formed electrostatically, its boundaries are smooth and electron scattering can be neglected. An ideal quantum point contact (QPC) between the reservoir region and the quantum dot is formed.

The wavefunction inside the constriction is one-dimensional [40]. The width of the constriction can not be considered constant, however. The simplest theory for this case is based on the adiabatic approximation which assumes that the cross section of the channel changes so slowly that only negligible scattering between the subbands occurs. Transport is possible only in the \( x \)-direction and confinement of carriers occurs in the \( yz \)-plane. At each value of \( x \) along the channel a two-dimensional Schrödinger equation with the self-consistent effective Kohn-Sham potential \( V_{\text{eff}}(x, y, z) \) for the transverse eigenfunctions \( \phi_n(x; y, z) \) and the subband energies \( \epsilon_n(x) \) is solved, i.e.

\[
\left[ -\frac{\hbar^2}{2m^*_\perp} \left( \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + V_{\text{eff}}(x, y, z) - \epsilon_n(x) \right] \phi_n(x; y, z) = 0,
\]  
(4.71)

where \( m^*_\perp \) is the electron effective mass perpendicular to the transport direction. The effective mass is assumed to be constant throughout the \( yz \)-plane, i.e. the channel cross-section. The full wave function \( \psi \) satisfying the three-dimensional Schrödinger equation factorizes as

\[
\psi \rightarrow \psi_{nk}(x, y, z) = \phi_n(x; y, z) \xi_k(x),
\]  
(4.72)

where the \( \xi_k(x) \) are the solutions of the coupled-mode equation, i.e. the
Figure 4.5: (a) Schematic view of the constriction. (b) The barrier potential, the energies of the lowest subband and the quantum dot spectrum. \( \alpha \) is the classical turning point in the reservoir.

One-dimensional Schrödinger equation with the \( n \)th subband energy \( \epsilon_n(x) \) as effective potential

\[
\left[-\frac{\hbar^2}{2m^*} \frac{\partial^2}{\partial x^2} + \epsilon_n(x)\right] \xi_k(x) = E_{nk} \xi_k(x).
\]  

The total energy is

\[
E \rightarrow E_{nk} = \epsilon_n(x) + \frac{\hbar^2 k^2}{2m^*}.
\]  

Now, the case of an electron moving from a reservoir state labeled \( E_{nk} \) to a 0D-quantum dot state \( \epsilon_p \) is considered. Energy conservation in the tunneling process requires that \( E_{nk} = \epsilon_p \). It is convenient to assume that \( V_{\text{eff}} \) is constant outside a larger range to the left hand side of the barrier \( x < x_L \), i.e.
4.4. Tunneling rates

\( \epsilon_n(x) = \mathcal{V}_{\text{eff}, \text{L}} = \text{const for } x < x_L \leq a, \text{ Fig. (4.5)}. \) Since the constriction is formed electrostatically, its boundaries are smooth and scattered components \( \exp(-i k x) \) of the wavefunction can be neglected. The wave function components in the \( x \)-direction in the outer regions are taken to be plane waves \( \xi_k(x) = \exp(i k x) \) where \( k = k_L \) and \( k(x) \) is the positive root

\[
\frac{\hbar^2}{2 m^*} k^2(x) = E_{n_k} - \epsilon_n(x). \tag{4.75}
\]

The classical turning point \( a \) in the reservoir is that particular point in \( x \) where \( E_{n_k} = \epsilon_p \), i.e. where the energies of the electron before and after the tunneling process are matched and the classically forbidden region starts. Classically, the wave is reflected at this point. Quantum mechanically, it can tunnel through the barrier. The WKB approximation is used for the \( x \)-component of the wavefunction in \( x_L < x \leq a, \text{ Fig. (4.5)}, \) i.e.

\[
\xi_k(x) = \exp[w(x)] \quad \text{with} \quad w(x) = i \int_{x}^{x_L} \, dx' \, k(x') \tag{4.76}
\]

as long as \( k(x) \) varies slowly enough (in particular, if \( |dk/da| < k^2 \)). The WKB approximation is also used in the forbidden region \( x > a \). The wave function component in \( x \) is

\[
\xi_\kappa(x) = \exp[w(x)] \quad \text{with} \quad w(x) = \int_{a}^{x} \, dx' \, \kappa(x') \tag{4.77}
\]

and \( \kappa(x) \) is the positive root

\[
\frac{\hbar^2}{2 m^*} \kappa^2(x) = \epsilon_n(x) - E_{n_k} \tag{4.78}
\]

again with the assumption of \( |d\kappa/da| \ll \kappa^2 \). The WKB wave functions in the different regions would have to be joined correctly as \( x \) increases through the classical turning point \( a \).

Using Eq. (4.70), the matrix element for the transition from the \( n \)th transverse state in the constriction to the \( p \)th localized quantum dot state can be approximated. The constriction is assumed to be narrow enough that only a single transverse state is below the Fermi level. As the electron moves away from the constriction, the channel becomes wider and the number of transverse states grows. However, within the constriction a single mode is present and the matrix element is calculated only for the transition from the ground state
$(n = 0)$. Both the channel region and the quantum dot region overlap and the matrix element is calculated by integrating over a surface $\partial \Omega$ in the $yz$-plane at some point $x_b$ (usually taken as the mid-point of the barrier, Fig. (4.5)). This leads to

$$M_p \approx -\frac{\hbar^2}{2m^*_\perp} \xi_{\kappa}(x_b) \int \int_{\partial \Omega(y,z)} dy \, dz \, \phi_0(x_b; y, z) \left[ \kappa(x_b) \psi_p(x_b; y, z) - \frac{\partial}{\partial x} \psi_p(x_b; y, z) \right], \quad (4.79)$$

where $\kappa(x_b)$ is defined by Eq. (4.78) with $E_{0k} = \varepsilon_p$. Inserting the matrix element in Eq. (4.66) the tunneling rate $\Gamma_p$ is obtained. Equation (4.79) is the central result of this section.
Chapter 5

Provando e riprovando.
By proof and refutation.
DANTE, »Paradiso, III«

Numerics

5.1 Introduction

Subject of this chapter are the numerical methods employed to solve the equations for the linear-response model which is used to simulate single-electron transistors. The basic equation is the nonlinear Poisson equation

\[ \nabla \cdot (\varepsilon \nabla \phi) + \rho[\phi] = 0, \]

(5.1)

where the charge density \( \rho \) is given as the sum of all charges in the device \( \rho = q(p - n + N_d^+ - N_a^-) \). A considerable portion of the charge is calculated by solving Schrödinger-type equations for confined electrons whose interaction is not sufficiently described by the classical Coulomb repulsion alone but requires the inclusion of complex quantum-mechanical many-body interactions. In general, it is not possible to find analytical solutions for this equations. Therefore, the development of efficient numerical methods is of paramount importance for single-electron device simulation.

The numerical solution of Eq. (4.14) proceeds in three steps: first the device is subdivided into small elements with constant electrostatic potential and carrier densities. Then the equations in these domains are approximated by algebraic equations using a finite-difference discretization scheme. This leads to a fairly large system of nonlinear equations for the electrostatic potential \( \phi \) and the charge density \( \rho \). In a third step this system of equations is solved.
5.2 Space discretization

5.2.1 Finite differences

A discretization approach in which the differential operators are approximated by differences is known as the method of *finite differences* (FD). The device domain is partitioned in subdomains separated by grid lines which run parallel to the (Cartesian) coordinate axes. Such \(d\)-dimensional grids can be represented by products of \(d\) one-dimensional grids. Consequently, they are called *tensor-product grids*.

The nomenclature for a two-dimensional nonuniform grid is shown in Fig. 5.1. In three space dimensions \((d = 3)\) a point on the grid is numbered by a discrete triple \(p = (i, j, k)\). The coordinates of this point are given by \(r_p = (x_i, y_j, z_k)\) where the elements of this triple are the positions of the \(i\)th, \(j\)th and \(k\)th grid line parallel to the \(x\)-, \(y\)- and \(z\)-axis, respectively. The separation of the \(i\)th from the \((i + 1)\)th grid line in the \(x\)-direction is given as

\[
h_i^x = x_{i+1} - x_i, \quad i = 0, \ldots, n_x, \tag{5.2}
\]

where \(n_x\) is the highest grid line number in the direction of the \(x\)-axis and a total of \(n_x + 1\) lines is present in this direction. A scalar quantity \(u(x, y, z)\) at the grid is represented by a finite number of sampling points \(u_{i,j,k}\) with

\[
\begin{align*}
  u_{i,j,k} & = u(x_i, y_j, z_k), \\
  u_{i+\frac{1}{2},j,k} & = u\left(\frac{x_i + x_{i+1}}{2}, y_j, z_k\right), \quad i = 0, \ldots, n_x - 1 \\
  u_{i-\frac{1}{2},j,k} & = u\left(\frac{x_{i-1} + x_i}{2}, y_j, z_k\right), \quad i = 0, \ldots, n_x - 1 \tag{5.3}
\end{align*}
\]

and \(i = 0, \ldots, n_x, j = 0, \ldots, n_y\) and \(k = 0, \ldots, n_z\) if not stated otherwise.

Assuming that \(u\) is three times continuously differentiable, a Taylor expansion can be used to find an approximation for the first order derivative \(u'(x)\) for \(x \in (x_{i+1} - x_i)\). The Taylor expansion of \(u(x_i)\) is given as

\[
u(x_i) = u(x) + (x_i - x) \frac{du}{dx}(x) + \frac{(x_i - x)^2}{2} \frac{d^2u}{dx^2}(x) + O((x_i - x)^3).
\tag{5.4}
\]

Subtracting an equivalent equation for \(u(x_{i+1})\) from Eq. (5.4) and using the definition of \(u\) in the middle between \(x_i\) and \(x_{i+1}\) (Eq. (5.3)) the derivative at
5.2. Space discretization

Figure 5.1: Elements and vertices in a finite difference discretization. The circles mark the nodes and the dashed area is the box volume associated with node \( i, j \). The shaded areas are the different elements surrounding the node \( i, j \).

\( x_{i+\frac{1}{2}} \) is given as

\[
\frac{\partial}{\partial x} u_{i+\frac{1}{2}, j, k} = \frac{u_{i+1, j, k} - u_{i, j, k}}{h_i^x} + \mathcal{O}(h_i^x)^3. \tag{5.5}
\]

This leads to an approximation for the first order derivative which has second order accuracy, i.e. the local truncation error \( \mathcal{O} \) is of third order. The above definition of the first order derivative is known as central-difference approximation. It should be mentioned that without the central difference approximation the local truncation error would be of second order, \( \mathcal{O}((h_i^x)^2) \).

Tensor-product grids have two major drawbacks. The first is the poor control one has over the point density of the grid. Refining the grid in regions where the gradients of the sampled properties are high always creates an abundance of unneeded points in other regions. The number of grid points determines both memory and CPU time requirements which can seriously limit the utility of finite differences. The second is related to the rectangular properties of the grid which makes it difficult to model non-rectangular device features.
The main advantage, however, is the regularity of the sparse linear systems emerging from the discretization. The simple banded structure of the matrices allows for the use of very simple data structures and algorithms. Finite differences are therefore very appealing if a discretization scheme is needed which can be handled with little effort.

5.2.2 Discretization of $\nabla \cdot (\hat{\gamma} \nabla)$

The partial differential equations, such as the non-linear Poisson equation, Eq. (4.14), and the eigenvalue problem associated with the Kohn-Sham equation, Eq. (4.28), are characterized by the same form of the linear operator $\nabla \cdot (\hat{\gamma} \nabla)$. The material dependent tensor $\hat{\gamma} = (\gamma_{\nu\mu})_{\nu,\mu \in \{x,y,z\}}$ is the dielectric tensor in the Poisson equation with the components $\gamma_{\nu\mu} = \epsilon_{\nu\mu}$ and the tensor of the reciprocal effective mass in the Kohn-Sham equation, $\gamma_{\nu\mu} = 1/m^*_{\nu\mu}$. In both cases the tensor is symmetric, i.e. $\gamma_{\nu\mu} = -\gamma_{\mu\nu}$. For symmetric tensors there always exists a coordinate system in which the tensor is diagonal

$$\hat{\gamma} = (\gamma_{\nu\mu} \delta_{\mu,\nu})_{\nu,\mu \in \{x,y,z\}} = \begin{pmatrix} \gamma^x & 0 & 0 \\ 0 & \gamma^y & 0 \\ 0 & 0 & \gamma^z \end{pmatrix}.$$  \hspace{1cm} (5.6)

Furthermore, in isotropic media, $\gamma^x = \gamma^y = \gamma^z$. This will be assumed to be the case in the next steps.

The finite difference approximation to the differential operator $\nabla \cdot (\hat{\gamma} \nabla)$ can be derived using the box-integration method. An algebraic equation for an operator at a grid point $p = (i, j, k)$ is derived by integration over the box volume

$$\Omega_{i,j,k} = \{x_{i-1} \leq x \leq x_{i+\frac{1}{2}}; y_{j-\frac{1}{2}} \leq y \leq y_{j+\frac{1}{2}}; z_{k-\frac{1}{2}} \leq z \leq z_{k+\frac{1}{2}}\}.$$  \hspace{1cm} (5.7)

Application of Gauss law yields

$$\int_{\Omega_{i,j,k}} \mathbf{d} \mathbf{r} \nabla \cdot (\hat{\gamma}(\mathbf{r}) \nabla u(\mathbf{r})) = \int_{\partial \Omega_{i,j,k}} \mathbf{d} f \hat{\gamma}(\mathbf{r}) \nabla u(\mathbf{r}),$$  \hspace{1cm} (5.8)

which leads to a surface integral over the surface of the box defined by the six planes bisecting the line connecting a vertex to its next neighbor (for vertices
on the boundary, only the volume inside is included). The result is a real, symmetric matrix $\hat{A}$ which operates on the vector $u$ as

$$(\hat{A}u)_{i,j,k} = \sum_{\nu = \{x,y,z\}} \left[ -\frac{2}{h_i^\nu (h_{i-1}^\nu + h_i^\nu)} \gamma_{i+\frac{1}{2},j,k} u_{i,j,k} 
-\frac{2}{h_{i-1}^\nu (h_{i-1}^\nu + h_i^\nu)} \gamma_{i-\frac{1}{2},j,k} u_{i,j,k} 
+\frac{2}{h_i^\nu (h_{i-1}^\nu + h_i^\nu)} \gamma_{i+\frac{1}{2},j,k} u_{i+1,j,k} 
+\frac{2}{h_{i-1}^\nu (h_{i-1}^\nu + h_i^\nu)} \gamma_{i-\frac{1}{2},j,k} u_{i-1,j,k} \right],$$

where $(\hat{A}u)_{i,j,k}$ denotes the row (column) of the symmetric matrix associated with the box $(i, j, k)$ multiplied by $u$, i.e. the left-hand side of the algebraic equation for $u$ in the box $(i, j, k)$. The order of the matrix is $n = n_x \times n_y \times n_z$.

Because of the number of off-diagonal entries the matrix is tridiagonal for $\Omega_i, j, k \in \mathbb{R}^1$, a five-point matrix for $\Omega_i, j, k \in \mathbb{R}^2$, and a seven-point matrix for $\Omega_i, j, k \in \mathbb{R}^3$, i.e. the summation in Eq. (5.2.2) is restricted for problems with dimensions $d = 1$ or $d = 2$.

Neumann boundary conditions, especially homogeneous boundary conditions (used in the Poisson equation for the electrostatic potential at artificial boundaries) are given for vertices in the surface region $B$ as

$$n \nabla u(r) = 0, \quad r \in \partial \Omega_B. \quad (5.9)$$

They reduce to $u_{i,j,k+1} = u_{i,j,k-1}$ ($n = (0, 0, 1)$) in the $z$-direction. Implementation of Dirichlet boundary conditions in the surface region $D$ is trivially performed by setting $u_{i,j,k} = u_0$ at $(x_i, y_j, z_k) \in D$ where $u_0$ is a physically meaningful value for either Ohmic- or Schottky contacts.

### 5.3 Linear equation solving and eigenvalue problems

#### 5.3.1 Linear Poisson equation

The operator of the linear Poisson equation is the Laplace operator $\mathcal{L} = \nabla \cdot (\hat{\varepsilon} \nabla)$ which acts on the electrostatic Potential $\phi$. The dielectric tensor $\hat{\varepsilon}$ is taken for inhomogeneous but isotropic materials, which makes all entries of
the matrix equal to \( \epsilon \). Application of Eq. (5.2.2) and integration of the right hand side of the linear equation results in a real symmetric matrix problem for the matrix \( \hat{L} \)

\[
\hat{L}\phi = -\Omega \rho,
\]

(5.10)

where \( \phi \) and \( \rho \) are vectors of the vertex representations of the potential and the total charge density, and \( \Omega \) is the diagonal matrix of the box volumes. The linear system in Eq. (5.10) is equilibrated according to

\[
(\hat{D}^{-1/2}\hat{L}\hat{D}^{-1/2})(\hat{D}^{1/2}\phi) = -\hat{D}^{-1/2}\Omega \rho, \quad \text{with} \quad \hat{D} = \text{diag}(\hat{L}), \quad (5.11)
\]

in order to restrict the eigenvalue spectrum of the operator to \( 0 \leq \lambda_i \leq 2 \), and therefore improve the conditioning of the matrix [60].

The solution of the linear Poisson equation requires the inversion of the equilibrated matrix \( \hat{L} = \hat{D}^{-1/2}\hat{L}\hat{D}^{-1/2} \). Due to the fact that the box discretization couples only between different grid points which are neighbors separated by the box faces, the linear system Eq. (5.10) is sparse, i.e. only a small portion of the matrix entries is nonzero. Furthermore, since the FD method is used, the nonzero entries always fill the same positions in the matrix. To limit time and memory requirements for the inversion of linear matrices it is mandatory to employ data structures and algorithms that make use of sparsity of the problem. A common data structure for sparse matrices is the YSMP\(^1\) format: all nonzero entries of the matrix are stored in a dense vector. In a vector of the same length the corresponding column indices are listed, and in a third vector pointers to the start of each row in the other vectors are stored. The row start vector has length \( n + 1 \), the last entry being a sentinel. This format is also known as condensed row format or \((ia, ja, a)\)-format because of an often used naming of the three vectors.

The linear system can be solved by sparse direct methods, i.e. Gaussian elimination, or by sparse iterative methods, usually derivatives of the conjugate gradient method (CG) [54]. The matrices resulting from three-dimensional grids are often rather large, and only iterative methods can be applied. CG based methods as BiCG [55], CGS [56] and CGSTAB [57], all combined with ILU preconditioning [58] are nowadays methods of choice in large scale device simulations.

\(^1\)Yale Sparse Matrix Package
5.3.2 Schrödinger equation

The Schrödinger equation (which is called Kohn-Sham equation in the case of the presence of a Kohn-Sham potential) constitutes a matrix eigenvalue problem for the Hamilton operator matrix \( \hat{H} \)

\[
\hat{H}\psi_n = E_n\Omega\psi_n, \quad (5.12)
\]

where the operator \( \mathcal{H} = \mathcal{T} + \mathcal{V} \), the Hamilton operator with the kinetic part \( \mathcal{T} = -\frac{1}{2} \nabla \cdot (1/m^* \nabla) \), is discretized using Eq. (5.2.2). The inhomogeneous reciprocal effective mass tensor \( 1/n_{\text{eff}} \) is assumed to be diagonal. The potential operator \( \mathcal{V} \) represents the effective (Kohn-Sham-) potential and is given as the product of the unity matrix times the vector of the vertex potential \( V(r) \rightarrow (IV)_{ik} \), therefore being a contribution to the diagonal of \( \hat{H} \). \( E_n \) and \( \psi_n \) are the eigenvectors and eigenvalues of the operator.

In order to transform this equation into a standard matrix eigenvalue problem both sides are premultiplied by \( \Omega^{-1/2} \). Substituting \( \mathbf{I} = \Omega^{-1/2}\Omega^{1/2} \) one arrives at

\[
(\Omega^{-1/2}\hat{H}\Omega^{-1/2})(\Omega^{1/2}\psi_n) = E_n(\Omega^{1/2}\psi_n) \quad (5.13)
\]

or \( \tilde{H}\tilde{\psi}_n = E_n\tilde{\psi}_n \) with the still Hermitian matrix \( \tilde{H} = \Omega^{-1/2}\hat{H}\Omega^{-1/2} \) and the eigenvectors \( \tilde{\psi}_n = \Omega^{1/2}\psi_n \) [60]. The matrix is real and symmetric without an applied magnetic field, which results in real eigenvalues and eigenvectors. Boundary conditions are such that the eigenfunctions \( \tilde{\psi}_n \) vanish at the boundary of the Schrödinger grid (homogeneous Dirichlet boundary condition) which comprises only a small portion of the total grid.

The same problem as with linear equations arises with eigenvalue problems. The matrix representations of the operators are too large for direct methods. Alternatively, one uses iterative methods which calculate the eigenvalues and vectors at the lower bound of the spectrum up to a desired number. This is a reasonable approach, since only the lowest eigenvalues are occupied, and therefore physically meaningful within a Kohn-Sham framework.

Especially, subspace (Krylow subspace) iteration methods have proven to be extremely powerful for large eigenvalue problems arising not only from space discretizations of the Schrödinger equation. A rather new method in this context is the Jacobian-Davidson style QR method (JDQR) [59] which is able to obtain a group of eigenvalues near a target \( \lambda_t \) without the use of
a *shift and invert* strategy. JDQR is trying to find the smallest eigenvalues of the matrix \((\hat{A} - \lambda_t I)\). A search subspace is generated onto which the given eigenproblem is projected. The much smaller projected eigenproblem is solved and this leads to approximations for the wanted eigenvectors and eigenvalues of the larger problem (Davidson part). Then, a correction equation for a selected eigenpair is considered. The solution of the correction equation defines an orthogonal correction for the current eigenvector approximation (Jacobi part). The correction is used in the expansion of the search subspace and the process is repeated again. JDQR aims to find \(k_{\text{max}}\) Schur pairs \((\mathbf{q}, \lambda)\) near \(\lambda_t\). It consists of an outer iteration, in which an approximate Schur pair is computed and an inner iteration in which a linear system is solved. The linear system is preconditioned by a matrix \(K\) which is an approximation of the matrix \((\hat{A} - \lambda_t I)\). Deflation is applied whenever a Schur vector has converged, i.e. the linear system is projected onto the subspace orthogonal to the found Schur vectors.

### 5.4 Nonlinear equation solving

#### 5.4.1 Introduction

Solving nonlinear equations like the nonlinear Poisson equation is a task that can only be performed numerically. Traditionally, all terms are moved to the left and one solves for

\[
F(x) = 0 \quad \text{and} \quad F : \mathbb{R}^N \to \mathbb{R}^N,
\]

where \(F\) is the \(N\)-dimensional vector function whose components \(F_i\) are the individual equations to be satisfied simultaneously. The functions \(F_i\) are functions of real variables. Furthermore, it is assumed that they are continuous and differentiable with respect to these variables. In many cases the vector functions are extremely complicated and hence laborious to compute. This has the immediate consequence that during any iterative solution process the bulk of the computing time will be spent in evaluating the vector function. Thus, the most efficient method will be that which requires the smallest number of function evaluations.

Other than for linear equations, root finding for nonlinear equations invariably proceeds by iteration. Starting from an initial guess vector for \(x\), the
solution is improved until some predetermined convergence criterion is satisfied. Even for smoothly varying functions, most of the convergence behavior of an algorithm depends on the quality of the initial guess. Providing good initial guesses requires at least some insight into the non-linear problem, i.e. some analysis concerning the properties of the vector function \( F \).

Most of the methods applied today in semiconductor device simulations are based on the familiar Newton-Raphson method. However, in most cases of actual interest, Newton methods can not be applied in a pure way since they are not in general globally converging. The combination with damping- or line-search schemes to ensure global convergence is a very common practice. Newton-Raphson methods are bad performers essentially in two cases. First, if the initial guess is only a poor approximation to the solution, and second, if the Jacobian matrix can not be given explicitly and a poor approximation to the algebraic expression for the derivatives or numerically calculated derivatives have to be used.

In single-electron and other quantum-mechanics-based simulations, the problems arise mainly with the quantized charge densities which can not be given as functions locally dependent on the electrostatic potential. Therefore, the Jacobian matrix has to be approximated in some way. In most cases the semi-classical expressions for the charge densities are used to approximate the Jacobian entries which are due to quantized charges. However, quadratic convergence can not be guaranteed with these approximations, and in many cases the Newton iterations will fail to converge at all.

The following section is mainly dealing with inexact and damped Newton-Raphson schemes as arise from low-temperature calculations including partial and complete confinement of carriers. The methods are especially designed to reduce the number of function evaluations required, and to work around the problems due to the inexact Jacobians.

### 5.4.2 Newton-Raphson method

The most important methods for nonlinear equation solving in one dimension are bisection and bracketing. The only elementary method that works in multidimensional problems is the Newton-Raphson method. The nonlinear equation is linearized in the vicinity of the solution using a truncated Taylor
expansion

$$F(x + \delta x) = F(x) + \tilde{J}\delta x + O(\delta x^2), \quad (5.15)$$

where $\tilde{J}$ is the Jacobian matrix whose elements are the first order derivatives of the elements of the vector function

$$J_{ij} = \frac{\partial F_i}{\partial x_j}. \quad (5.16)$$

By neglecting terms of order $\delta x^2$ and setting $F(x + \delta x) = 0$, one obtains a set of linear equations for the corrections $\delta x$,

$$-\tilde{J}^{-1}F(x) = \delta x, \quad (5.17)$$

that move the vector function closer to zero according to

$$x_{\text{new}} = x_{\text{old}} + \delta x. \quad (5.18)$$

In this sense, Newton's method replaces a system of nonlinear equations with a set of linear equations, but since the solutions of the two systems are not identical in general, the process must be repeated until a certain accuracy is achieved. The Newton method converges quadratically if applied close to the solution.

### 5.4.3 Broyden's method

**Secant updating methods**

The high cost per iteration of the Newton method and/or the difficulties for some problems to calculate the Jacobian matrix has lead to alternative methods, analogous the one-dimensional secant method. These methods gradually build up an approximation to the Jacobian based on successive evaluations of the function vector without explicitly calculating derivatives. Computational effort is saved by updating a factorization of the approximate Jacobian matrix at each iteration (employing techniques similar to the Sherman-Morrison formula). Because of these two features they are often referred to as *secant updating methods*. However, the price to be payed is that these methods generally have super-linear and not quadratic convergence.
5.4. Nonlinear equation solving

Broyden’s original method

One of the most effective schemes for secant updating is Broyden’s method [61], which begins with an approximate Jacobian matrix which is updated at each iteration. If the negative inverse of the approximate Jacobian is denoted \( \hat{G}^{(m)} = -J^{(m)}_{\text{approx}}^{-1} \) the \( m \)th quasi-Newton step \( \delta \mathbf{x}^{(m)} \) is the solution of

\[
\hat{G}^{(m)} \mathbf{F}^{(m)} - \delta \mathbf{x}^{(m)} = 0,
\]

(5.19)

where \( \delta \mathbf{x}^{(m)} = \mathbf{x}^{(m+1)} - \mathbf{x}^{(m)} \) (cf. Eq. (5.18)). In general, \( \mathbf{F}(\mathbf{x} + \delta \mathbf{x}) \neq 0 \) and Eq. (5.15) can be recast into a quasi-Newton secant condition which is such that \( \mathbf{F}(\mathbf{x} + \delta \mathbf{x}) \) satisfies

\[
\delta \mathbf{x}^{(m)} + \hat{G}^{(m)} \delta \mathbf{F}^{(m)} = 0,
\]

(5.20)

where \( \delta \mathbf{F}^{(m)} = \mathbf{F}^{(m+1)} - \mathbf{F}^{(m)} \). This is the generalization of the secant approximation to the derivative \( F' \) in one dimension. It is both convenient and numerically advantageous to scale the vectors and redefine \( \delta \mathbf{x} \) and \( \delta \mathbf{F} \) as

\[
\delta \mathbf{x}^{(m)} = \frac{\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}}{\| \mathbf{F}^{(m+1)} - \mathbf{F}^{(m)} \|_2}
\text{ and } \delta \mathbf{F}^{(m)} = \frac{\mathbf{F}^{(m+1)} - \mathbf{F}^{(m)}}{\| \mathbf{F}^{(m+1)} - \mathbf{F}^{(m)} \|_2}.
\]

(5.21)

Note, that now \( \delta \mathbf{F}^{(m)} \cdot \delta \mathbf{F}^{(m)} = 1 \). Since \( \hat{G}^{(m)} \) generally fails to satisfy Eq. (5.20) an updating formula for \( \hat{G}^{(m)} \) minimizing \( \| \hat{G}^{(m+1)} - \hat{G}^{(m)} \|_2 \) is used

\[
\hat{G}^{(m+1)} = \hat{G}^{(m)} - (\delta \mathbf{x}^{(m)} + \hat{G}^{(m)} \delta \mathbf{F}^{(m)}) \delta \mathbf{F}^{(m)\text{T}},
\]

(5.22)

where \( \hat{G}^{(m+1)} \) is consistent with Eq. (5.20). Equation (5.22) is the Broyden updating formula which is often applied using the Sherman-Morrison formula for an analytical inversion of the approximate (quasi-) Jacobian.

Modifications of Broyden’s method

Vanderbilt and Louie [63] introduced a modified version of the original method in which they retain the informations from all previous iteration steps during the updating procedure. Even though the overall performance of the method is improved over the original method, the storing and multiplications
of \( N \times N \) matrices remained a shortcoming especially in large-scale applications. A new updating formula was derived using least-squares minimization of an error function, which includes all previous vectors \( \delta x \) and \( \delta F \) weighted with a factor \( \omega_n \) for the \( n \)th out of \( m \) steps and a factor \( \omega_0 \) assigned to the error in the inverse Jacobian. With the error function defined as

\[
E = \omega_0 \| \hat{G}^{(m+1)} - \hat{G}^{(m)} \|_2 + \sum_{n=1}^{m} \omega_n^2 \| \delta x^{(n)} + \hat{G}^{(m+1)} \delta F^{(n)} \|_2,
\]

subject to a variational ansatz according to \( \partial E / \partial G^{(m+1)}_{ij} \), one obtains

\[
\hat{G}^{(m+1)} = \hat{A}^{(m+1)} (\hat{B}^{(m+1)})^{-1}
\]

with the auxiliary matrices \( \hat{A} \) and \( \hat{B} \) according to

\[
\hat{A}^{(m+1)} = \hat{G}^{(m)} - \hat{X}^{(m)} \hat{F}^{(m)T} \quad \text{and} \quad \hat{B}^{(m+1)} = I + \hat{F}^{(m)} \hat{F}^{(m)T},
\]

where, the following notation for the vectors \( \hat{F}^{(n)} = (\omega_n / \omega_0) F^{(n)} \) and \( \delta \hat{x}^{(n)} = (\omega_n / \omega_0) \delta x^{(n)} \) is introduced

\[
\hat{X}^{(n)} = (\delta \hat{x}^{(1)}, \ldots, \delta \hat{x}^{(n)}) \in \mathbb{R}^{(N \times n)}
\]

\[
\hat{F}^{(n)} = (\hat{F}^{(1)}, \ldots, \hat{F}^{(n)}) \in \mathbb{R}^{(N \times n)}
\]

Still, a matrix inversion which has to be performed to get \( (\hat{B}^{(m+1)})^{-1} \) is necessary. Johnson [64] suggested an expansion to infinite order in terms of \( \delta F \) in order to avoid explicit matrix inversions. By applying the Sherman-Morrison-Woodbury formula\(^2\), the following form of \( (\hat{B}^{(m+1)})^{-1} \) was derived

\[
(\hat{B}^{(m+1)})^{-1} = I + \hat{F}^{(m)} \hat{\beta}^{(m)} \hat{F}^{(m)T}
\]

with

\[
\hat{\beta}^{(m)} = (I + \hat{F}^{(m)} \hat{F}^{(m)})^{-1} \in \mathbb{R}^{(m \times m)},
\]

where \( \hat{\beta} \) is a \( m \times m \) matrix and not a large \( N \times N \) matrix. Using the result in Eq. (5.24), a new updating formula is obtained, i.e.

\[
\hat{G}^{(m+1)} = \hat{G}^{(m)} - (\hat{G}^{(m)} \hat{F}^{(m)} + \hat{X}^{(m)}) \hat{\beta}^{(m)} \hat{F}^{(m)T}.
\]

\(^2\)Sherman-Morrison-Woodbury formula: suppose \( \hat{A} \in \mathbb{R}^{N \times N}, \hat{U} \in \mathbb{R}^{N \times k}, \hat{V} \in \mathbb{R}^{N \times k} \) and both \( \hat{A} \) and \( (I + \hat{V}^T \hat{A}^{-1} \hat{U}) \) are nonsingular, then

\[
(\hat{A} + \hat{U} \hat{V}^T)^{-1} = \hat{A}^{-1} - \hat{A}^{-1} \hat{U} (I + \hat{V}^T \hat{A}^{-1} \hat{U})^{-1} \hat{V}^T \hat{A}.
\]

(5.25)
This method can be improved even further if \( \hat{G}^{(m+1)} \) is written in terms of \( \hat{G}^{(1)} \) [64], i.e.

\[
\hat{G}^{(m+1)} = \hat{G}^{(1)} - \hat{Z}^{(m)} \hat{F}^{(m)T}
\]

(5.27)

with the matrices \( \hat{U}^{(m)} \) and \( \hat{Z}^{(m)} \) defined as follows

\[
\hat{U}^{(m)} = \hat{G}^{(1)} \hat{F}^{(m)} + \hat{X}^{(m)}
\]

\[
\hat{Z}^{(m)} = (\hat{U}^{(m)} + [\hat{Z}^{(m-1)} | 0]) \hat{\beta}^{(m)}.
\]

The proof of this new updating formula can easily be accomplished by induction. The new method requires the storage of the current matrix \( \hat{\beta} \), and the vectors \( x \) and \( F \) as well as the vectors \( \delta x \) and \( \delta F \) from the previous iterations (storage of 2\((m+1)N\) vectors and a \( m \times m \) matrix).

The following approximation for the updating formula can be made if \( \omega_0 \ll \omega_n \)

\[
\hat{G}^{(m+1)} = \hat{G}^{(1)} - \hat{U}^{(m)} \hat{\beta}^{(m)} \hat{F}^{(m)T}.
\]

(5.28)

It only remains to find an appropriate form of \( \hat{G}^{(1)} \). It is straightforward to choose it as a diagonal constant matrix \( \hat{G}^{(1)} = \alpha \hat{I} \) and \( 0 < \alpha < 1 \). According to Powell [65], the appropriate choice of \( \alpha \) should be such that

\[
\delta F^{(2)} \cdot \delta F^{(2)} < \delta F^{(1)} \cdot \delta F^{(1)}.
\]

More elaborate choices for \( \hat{G}^{(1)} \) can certainly be envisaged. Another question is the choice of the weights \( \omega_m \). Johnson [64] suggested to use

\[
\omega_m = (\delta F^{(m)} \cdot \delta F^{(m)})^{-1/2},
\]

i.e. the inverse of the norm of the vector function.

5.4.4 Globally converging Newton schemes

Newton’s method may fail to converge when started far from the solution. Unfortunately, in \( N \)-dimensions there is no analogy to bisection that can provide a fail-safe hybrid method when combined with Newton iterations. Nevertheless, it is possible to safeguard the Newton’s method and open it to a much broader range of applications. Two methods will be discussed her, one is the damped Newton method especially in the form of the Bank-Rose damped Newton method. The other on is an extrapolation method which can be based on Broyden’s method as discussed in the previous section.

In the damped Newton method, the Newton step \( \delta x \) is calculated as usual at every calculation, but then the new iterate is taken to be

\[
x^{(m+1)} = x^{(m)} + \alpha^{(m)} \delta x^{(m)} \quad \text{with} \quad 0 < \alpha^{(m)} \leq 1,
\]

(5.29)
where $\alpha^{(m)}$ is a scalar parameter. The motivation is that far from the solution the Newton step is likely to be unreliable, in most of the cases to big. The factor $\alpha^{(m)}$ can be adjusted in such a way that $x^{(m+1)}$ is a better approximation to the solution if the step is damped. Sufficiently close to the solution $\alpha^{(m)}$ should approach unity and the convergence properties of the classical Newton method are retained.

One method to estimate the damping factor was introduced by Bank and Rose [66]. They showed that under certain conditions global and quadratic convergence is achieved if one uses

$$\alpha^{(m)} = \frac{1}{1 + \kappa_k \|F^{(m)}\|_2}$$

with $\kappa_k$ such that $\alpha^{(m)}$ satisfies a sufficient decrease condition

$$\alpha^{(m)} \delta \leq (1 - \|F^{(m+1)}\|_2/\|F^{(m)}\|_2)$$

where $\delta > 0$ is a fixed, small number which can be taken as the machine epsilon. It is readily noted that $\alpha^{(m)}$ approaches unity sufficiently fast as $\|F^{(m)}\|_2 \rightarrow \|F^{(m+1)}\|_2$.

The method after Bank and Rose is a popular method in the field of semiconductor device modeling. Other than the extrapolation methods discussed in the next section, it does not alter the direction of the Newton descent, i. e. only the length of the correction vector is adapted. Convergence properties of the method, therefore, remain predictable and are investigated thoroughly [66]. Practical experience with this method which shows its capabilities has been reported by Fichtner and Rose [67] for instance.

An alternative method to the Bank-Rose damping method is based on an extrapolation scheme used to determine the optimal correction vector, i. e.

$$x^{(m+1)} = x^{(m+1)}(x^{(m)}, \delta x^{(m)}, \ldots, x^{(m-n)}, \delta x^{(m-n)})$$

with $n$ being the extrapolation depth. Laux [60] used the Anderson extrapolation method [68] to extrapolate the correction vector in an Newton-Raphson iteration. This combination proved to be superior in convergence if applied to the nonlinear Poisson equation describing the charge density of confined electrons in an two dimensional channel.

Care should be taken in using extrapolation methods since in contrast to damping methods not only the length of the correction vector but also its direction is altered. This leads to in general poorer convergence properties compared to damped Newton methods. Especially, quadratic convergence is not
5.5. Self-consistent algorithms

achieved in most of the cases. However, extrapolation methods can be advantageous if the vector function is very expensive to evaluate. The Bank-Rose method requires at least two function evaluations, in many cases even more to find an appropriate choice for \( \alpha^m \). If, for instance, the evaluation of the function vector includes the solution of Schrödinger equations, this can consume a considerable amount of time.

5.5 Self-consistent algorithms

Two different nonlinear equation solvers for the solution of the nonlinear Poisson equation are implemented in the SIMNAD software. Both algorithms are based on Newton-Raphson iterations and a predictor-corrector type approach for the evaluation of the vector function \( \mathbf{F} \). They are mainly distinguished by the methods used to force and stabilize the convergence. The first algorithm uses the Bank-Rose damping method and the second Broyden extrapolation steps.

5.5.1 Predictor-corrector approach

Before the methods and their implementations are explained in more detail, the main idea of the predictor-corrector method in the solution of the nonlinear Poisson equation involving the Schrödinger equation for confined states is explained.

As already outlined in Section 4.2.4, the electrons in bound states or sub-bands in the case of partial confinement are described by the Kohn-Sham equation, Eq. (4.28), for the envelope wavefunctions. The charge densities in regions with zero-, one- or two-dimensional confinement are given by Eq. (4.30) with the (Fermi-Dirac-) distribution functions Eqs. (4.33), (4.34) and (4.35). Apparently, the exact dependence of the electron densities \( n \) on the electrostatic potential is not known. This results in an incorrect contribution to the Jacobian if a Newton-Raphson method is employed. Newton-Raphson methods can only be used without the Schrödinger equation coupled via the charge densities, i.e. with semi-classical charge densities as in Eq. (4.16). In many approaches the dependence of the electron charge density on the electrostatic potential is approximated assuming a semi-classical form for \( n \) to calculate \( \partial n / \partial \phi \) (see for instance [60]).
A more advanced solution is to use an approximate expression (predictor equation) for the quantum electron density \( \tilde{n}[\phi] \) in \( \Omega \) for which an explicit dependence on the electrostatic potential can be given. Consequently, the Poisson equation

\[
\nabla \cdot (\epsilon \nabla \phi) + \rho[\tilde{n}[\phi], \phi] = 0
\]

(5.33)

can be linearized using a Newton-Raphson approach (for all charges other than the quantized charges, the semi-classical expressions are used). The predicted result for \( n \) is then corrected in an outer iteration, i.e. the Schrödinger equation is solved and the correct expression for the charge density is evaluated. An expression for a predictor density which depends explicitly on \( \phi \) was derived by Trellakis et al. [69]. They applied first order perturbation theory to the Schrödinger equation in a quantum wire. The one-dimensional electron charge density is given according to Eq. (4.30) and Eq. (4.34), i.e.

\[
n[\phi] = N_{1D} \sum_n \psi_n^2[\phi] \mathcal{F}^{-\frac{1}{2}} \left( \frac{\mu - \epsilon_n[\phi]}{k_BT} \right),
\]

(5.34)

with the dimension-dependent factor

\[
N_{1D} = 2g \left( \frac{m^*_n k_BT}{2\pi \hbar^2} \right)^{\frac{3}{2}}.
\]

(5.35)

A similar expression (predictor) is derived

\[
\tilde{n}[\phi] = N_{1D} \sum_n (\psi_n^{(k)}[\phi])^2 \mathcal{F}^{-\frac{1}{2}} \left( \frac{\mu - \epsilon_n^{(k)} + q(\phi - \phi^{(k)})}{k_BT} \right)
\]

(5.36)

in which the subband energy \( \epsilon_n^{(k)} \) from the last outer step (corrector) is augmented by the change in the electrostatic potential, i.e.

\[
\epsilon_n[\phi] \rightarrow \epsilon_n^{(k)} - q(\phi - \phi^{(k)}).
\]

(5.37)

This form of the density is in particular suitable for Newton-Raphson methods since an expression for the contribution to the Jacobian can be given

\[
\frac{\partial \tilde{n}[\phi]}{\partial \phi} = \frac{q N_{1D}}{k_BT} \sum_n (\psi_n^{(k)}[\phi])^2 \mathcal{F}^{-\frac{3}{2}} \left( \frac{\mu - \epsilon_n^{(k)} + q(\phi - \phi^{(k)})}{k_BT} \right).
\]

(5.38)
In principle, this first order perturbation ansatz can be extended towards the electron densities in quantum wells (2DEG) and quantum dots (0DEG) resulting in similar expressions. If the electrostatic potential is sufficiently converged, it is used as \( \phi^{(k+1)} \) in the Schrödinger equation and \( n^{(k+1)} \) is used to calculate a new exchange-correlation potential \( V_{xc}[\tilde{n}^{(k+1)}] \).

The predictor-corrector method for the solution of the nonlinear Poisson equation can be summarized as follows. First, a predictor type equation \( \mathbf{F}_{\text{pred}} = 0 \) is solved using the approximate density \( \tilde{n}[\phi] \), Eq. (5.36), i.e.

\[
\mathbf{F}_{\text{pred}} = \nabla \cdot (e \nabla \phi^{(k+1)}) + \rho[\tilde{n}[\phi^{(k+1)}], \phi^{(k)}], \{\psi_{n^{(k)}}, \epsilon_{n^{(k)}}\}], \phi^{(k+1)}]. \tag{5.39}
\]

Then a corrector step is performed which includes the solution of the Schrödinger equation with the electrostatic potential solution of the predictor equation \( \phi^{(k+1)} \) and the exchange-correlation potential calculated with \( \tilde{n}^{(k+1)} \)

\[
\begin{align*}
-\frac{\hbar^2}{2} \nabla \cdot \left( \frac{1}{m^*} \nabla \right) + \Delta E_c - q\phi^{(k+1)} + V_{xc}[\tilde{n}^{(k+1)}] \psi_{n^{(k+1)}} \\
= \epsilon_{n^{(k+1)}} \psi_{n^{(k+1)}}. \tag{5.40}
\end{align*}
\]

With the exact (corrector-) electron density \( n[\phi] \), Eq. (5.34), the correct nonlinear function vector \( \mathbf{F}_{\text{corr}} \) is evaluated

\[
\mathbf{F}_{\text{corr}} = \nabla \cdot (e \nabla \phi^{(k+1)}) + \rho[n[\phi^{(k+1)}], \{\psi_{n^{(k+1)}}, \epsilon_{n^{(k+1)}}\}], \phi^{(k+1)}]. \tag{5.41}
\]

The iterations are repeated until \( ||\mathbf{F}_{\text{corr}}||_2 < \epsilon \).

### 5.5.2 Algorithm implementation in SIMNAD

Figure 5.2 shows a flow chart of a Newton-Raphson method with Bank-Rose damping for the above described nonlinear Poisson equation. The vector function \( \mathbf{F} \) is composed of the values of the residual of the Poisson equation at discrete grid points so that the order of the nonlinear problem is given by the size of the grid, i.e. by the number of nodes. The argument vector \( \mathbf{x} \) is the scalar electrostatic potential \( \phi \) at discrete grid points. For the predictor form of the vector function \( \mathbf{F}_{\text{pred}} \) the electron density \( \tilde{n}[\phi] \) for a 1DEG is chosen according to Eq. (5.36) or the corresponding expressions containing the explicit dependence on the electrostatic potential \( \phi \) for 2DEG and 0DEG. For the corrector form of the vector function the exact expression \( n[\phi] \) according to Eq. (5.34) is used. All other densities (electrons outside the confinement re-
Figure 5.2: Flow chart of a Bank-Rose damped Newton-Raphson algorithm for a predictor-corrector type nonlinear problem. The convergence criterion for the corrector equation $F_{\text{corr}}$ is $\varepsilon_1$ and for the predictor equation $F_{\text{pred}}$ the convergence criterion is $\varepsilon_2$. Usually, one chooses $\varepsilon_2 = 0.1 \varepsilon_1$, i.e. the predictor equation is solved to a higher accuracy than the corrector equation.
5.5. Self-consistent algorithms

gion, holes and ionized donors) are treated within the semi-classical equations given in Section 4.2.2.

The second algorithm is based on a Newton-Raphson method combined with a Broyden extrapolation scheme, Fig. 5.2. Using the multidimensional secant approach (Broyden's method) described in Section 5.4.3. The new update vector is extrapolated from the previous steps. This method has the advantage over the Bank-Rose method that the vector function has to be computed only once during one Newton step, whereas the Bank-Rose method requires multiple, but at least two evaluations to determine the correct damping factor $\alpha^{(m)}$. This can be extremely costly if the evaluations have to be performed for vectors of high order ($N \sim 50\,000 \ldots 100\,000$). On the other side

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{flowchart.pdf}
\caption{Flow chart of a Newton-Raphson algorithm combined with a Broyden extrapolation scheme for a predictor-corrector type nonlinear problem.}
\end{figure}
the use of the Broyden extrapolation method requires storage of at least \(2n\) vectors of order \(N\). The number \(n\) is the number of previous iteration steps used for the extrapolation.

The success and the speed of convergence heavily depend on the choice of the initial vector \(x_{\text{init}}\), i.e., the choice of the initial electrostatic potential. Usually, this is not a big issue in quasistationary calculations where a gate voltage is ramped with considerable smoothness and so the boundary conditions of the Poisson equation change not very much from one run of the solver to the next. In this case the potential from the previous iteration is always a good estimate. However, at the very beginning of the calculation, when no information from previous runs is available, the Laplace equation, subject to the appropriate boundary conditions is solved and the (homogeneous) potential is used as the initial choice.

5.5.3 Examples

In order to demonstrate the convergence behavior of the non-linear methods presented, calculations for a test structure are performed. A GaAs/AlGaAs structure similar to the one discussed in the next chapter is used. The 2DEG in the GaAs/AlGaAs heterojunction is modeled assuming three independent

![Figure 5.4: Convergence of the Bank-Rose algorithm (○) with a poor approximation of the initial guess potential. The other curve (△) shows the Bank-Rose damping factor \(\alpha\). The dotted lines mark the convergence threshold for the corrector equation (\(\varepsilon_1\)) and the predictor equation (\(\varepsilon_2\)), respectively.](image)
regions with one-dimensional confinement in the $z$-direction. Two regions are related to the drain and source region, respectively, and the third region is the quantum dot region.

Figure 5.4 shows the convergence for the residual of the non-linear Poisson equation $\|F^{(m)}\|_2$ and the damping factor $\alpha$ for the Bank-Rose algorithm. The solution of the homogeneous Poisson equation was used as initial estimate for the potential which is a rather poor approximation resulting in a slow convergence during the first steps. The residual exhibits the typical seesaw-like behavior of predictor-corrector equations since the residual always increases if the corrector equation is solved. A good performance of the Newton-Raphson method is achieved for the predictor equation because of the exact Jacobian involved. The calculation was performed at a temperature of 4.2 K.

The result of this calculation is used as initial guess for the next run where the 2DEG model previously used for the quantum dot is replaced by a 0DEG model solving for approximately 50 . . . 60 eigenvalues in the quantum dot. The performance of the Bank-Rose algorithm was compared to that of the Broyden-extrapolation algorithm in which the following parameters were used: $\omega_0 = 10^{-4}$ and $\mathbf{G}^{(1)} = 0.5 \mathbf{I}$. The calculations were now performed at 1 K. The Broyden-extrapolation algorithm is usually faster in calculations where a good initial guess potential can be provided, i.e. it is mostly used in quasistationary calculations where a gate voltage is ramped in small steps and the result from the previous calculation can be used as initial estimate. The

![Figure 5.5: Comparison of the Bank-Rose algorithm (□) with the Broyden-extrapolation algorithm (○). The dotted lines mark the convergence threshold for the corrector equation ($\varepsilon_1$) and the predictor equation ($\varepsilon_2$), respectively.](image-url)
Bank-Rose algorithm generally proved to be more stable and converges even if the initial estimate is rather poor. It is therefore the method of choice if a good guess for the potential is needed or no previously calculated results for a structure are available.
Chapter 6

And here Alice began to get rather sleepy, and went on saying to herself, in a dreamy sort of way, 'Do cats eat bats? Do cats eat bats?' and sometimes 'Do bats eat cats?' for, you see, as she couldn't answer either question, it didn't much matter which way she put it.

LEWIS CARROLL, »Alice's Adventures in Wonderland«

Simulation examples

6.1 GaAs/AlGaAs heterostructure

The most commonly used material system for quantum dot structures is gallium arsenide, GaAs, which combined with the ternary solution aluminum-gallium arsenide, Al$_x$Ga$_{1-x}$As, serving as barrier material, allows for the creation of very thin epitaxial layers as a result of the almost equal lattice constants. The strong electrostatic confinement at the interface between GaAs and AlGaAs can be used to constrain electrons to move in a plane, thus forming a two-dimensional electron gas (2DEG).

The 2DEG can be given any desired shape using lithographic methods. The shape of the gas is defined either by etching a pattern or by metallic gates which electrostatically deplete the 2DEG underneath. The latter method has the advantage that the depletion is not permanent and can be switched off. Local depletion of the electron gas is associated with a local increase of the electrostatic potential relative to the undepleted region. The resulting potential step in the 2DEG is smooth because of the large depletion length of $\approx$ 100 nm. This is the basis for the so-called split-gate technique used to define small channels or dots with variable size and smooth boundaries by means of several
Chapter 6. Simulation examples

Figure 6.1: Schematic of the horizontal structure of the GaAs/AlGaAs heterostructure. The dashed line marks the 2DEG.

gates deposited on the top semiconductor layer. The gates are automatically isolated from each other because of the Schottky barriers which are formed by the metal/semiconductor contact if operated in reverse bias, i.e. by applying negative voltages with respect to the semiconductor ground.

At low temperatures, the conducting electrons are free to move over large distances (several \( \mu \text{m} \)) without being scattered and therefore maintain phase coherence. For systems with size quantization which is connected with a change in the density of states, the Fermi wavelength \( \lambda_F \) is important. It defines the length scale on which the wave nature of a particle becomes obvious. The electron motion is confined and only standing wave solutions, i.e. bound states with discrete energies are allowed. The quantum mechanical wavefunction of a particle with the Fermi energy \( E_F \) has the typical wave length \( \lambda_F = 2\pi k_F^{-1} \) and \( k_F = \sqrt{2mE_F/h^2} \). Particles at the Fermi edge are essential for charge transport. In GaAs heterostructures the typical value for \( E_F \approx 10 \text{ meV} \) and \( \lambda_F \approx 50 \text{ nm} \).

Figure 6.1 shows an example for a GaAs/AlGaAs heterostructure \([70]\) with a patterned top gate. The 2DEG is situated around 45 nm below the gate at the interface between the GaAs substrate and a layer composed of Al\(_{0.32}\)Ga\(_{0.68}\)As. The conduction band edge and the electron charge density calculated at 1 K are shown in Fig. 6.2. The position of the conduction band at the exposed surface (\( z = 0 \)) is determined by Fermi level pinning. The conduction band edge is pinned 600 meV above the Fermi level which is taken to be zero. Figure 6.3 shows an atomic-force microscopy (AFM) picture of a split-gate structure \([71]\). The gates consist of a layered structure of titanium (Ti) and gold (Au). A 5 nm layer of Ti is first evaporated directly at the GaAs cap-layer. Then a 50-150 nm thick layer of gold is deposited. The Schottky
6.1. GaAs/AlGaAs heterostructure

barrier is determined mainly by the Ti/GaAs contact ($\phi_s \approx 800$ meV).

The gates are asterisked as follows: F is the finger gate which forms together with the gates Q1 and Q1 two quantum point contacts (QPC) by depleting electron charges in the 2DEG underneath the gate electrodes. Usually, negative voltages between $-200$ meV and $-300$ meV are sufficient to fully deplete the charges immediately underneath the gates. A further increase of the negative voltages leads to additional lateral depletion. The gate asterisked C is the control- or plunger gate which is used to vary the potential in the quantum dot by applying a control-gate voltage $V_g$. The gate geometry defines an island in the 2DEG which is separated from the drain- and source reservoir by two tunnel junctions, i.e. two quantum point contacts. The shape of the barriers can be con-

Figure 6.2: Calculated conduction-band edge (solid line) and electron-charge density (dashed line) of the GaAs/AlGaAs heterostructure in Fig. 6.1 in growth direction.

Figure 6.3: AFM picture of the split-gate structure of the GaAs/AlGaAs heterostructure SET in Fig. 6.1. The framed region marks the simulated domain. (Courtesy T. Heinzel [72].)
trolled with the gates \( F, Q1 \) and \( F, Q2 \), respectively. The island potential is controlled by the control gate \( C \).

Measurements are performed in a dilution refrigerator in liquid helium at temperatures below \( T = 1 \text{ K} \). The electron transport was studied by applying a small bias voltage (4.3 \( \mu \text{V AC} \)) between the drain reservoir to the left and the source reservoir to the right. The current was measured using standard lock-in techniques ([71] and references therein).

### 6.2 Device simulation

#### 6.2.1 Geometry modeling and grid

The device in Fig. 6.1 is mapped onto a three dimensional non-uniform tensor-product grid with 69 \( \times \) 36 \( \times \) 70 grid points along \( x \times y \times z \), comprising a total of 173,880 vertex nodes. The total area simulated is a brick-like region of 1000 nm \( \times \) 600 nm \( \times \) 290 nm. The substrate, which is extending 2.8 \( \mu \text{m} \) in the negative \( z \)-direction, is effectively cut-off at 290 nm imposing von Neumann boundary conditions at this surface as well as at the surfaces in the lateral \( xy \)-directions where the electric field vanishes at distances far enough from the active device region. The Fermi level in the substrate is set assuming an electrode (ohmic contact) outside the simulated region, i.e. a metal backgate or a highly doped layer connected to an ohmic contact. The boundary conditions at the Schottky gates (Fig. 6.4(a)) are imposed by use of Dirichlet boundary conditions at the relevant nodes (Fig. 6.4(b)) according to Eq. (4.26).

Critical is the decomposition of the 2DEG into regions where the calculation of the electron charge is performed using different models for the electron confinement. Figure 6.4(c) shows the domain decomposition in the 2DEG layer. The quantum dot region (0DEG) is situated in the central region of 585 nm \( \times \) 450 nm \( \times \) 39 nm comprising a total of 42,336 nodes. The wavefunctions and the electron density vanish at the boundaries of this domain. The negative voltages at the split-gates ensure that the area surrounding the 0DEG region in the \( xy \)-pane is depleted so that these boundary conditions are adequate. The two quantum wire regions (1DEGs) which are used to self-consistently model the potential and the 2D-wavefunctions at the quantum-point contacts (QPCs) are connected to the two quantum well regions (2DEGs) which give the charge densities in the drain and source regions (leads). In the
Figure 6.4: (a) Scheme of the split-gate structure. (b) Non-uniform tensor-product grid and Dirichlet boundary nodes at the gate layer. (c) Domain decomposition of the 2DEG layer.
1DEGs the confinement is in the $yz$-plane and the electron charge is calculated according to Eq. (4.35). Usually these regions are completely depleted in order to guarantee a quantum-point contact with a sufficiently high barrier. However, the calculation of the 2D-wavefunctions and of the 1D-subband energies is necessary since these quantities are needed in the evaluation of the tunneling rates (see Section 4.4, especially Eq. (4.79)). In general, the different regions can overlap. This is mandatory, especially between the 0DEG region (quantum dot) and the 1DEG regions (lead channels, QPCs), since both solutions for the Schrödinger equation are needed to evaluate the tunneling rates, i.e. to calculate the overlap integral in the matrix element Eq. (4.79) at some point in the barrier. However, no special boundary conditions for the charge densities are imposed since the QPC is always depleted at least in its central region. Alternatively, open-boundary conditions for the Schrödinger equation in the quantum dot would have to be formulated to account for the propagating components of the otherwise localized states.

### 6.2.2 Surface and donors

Part of the surface of the device is covered by metal gates for which Dirichlet boundary conditions for Schottky contacts are employed. The rest is exposed GaAs. Since the exposed surface covers a substantial portion of the surface, the choice of the appropriate boundary condition is crucial for the determination of the electrostatic potential in the whole device. The properties of exposed GaAs surfaces are dominated by a band of surface states near the middle of the bandgap. While the states underneath the metal gate can be assumed to be pinned to the metal Fermi level, those at the exposed surface are assumed to be pinned at the substrate Fermi level. *Fermi-level pinning* is modeled using Dirichlet boundary conditions for the electrostatic potential at the surface, i.e.

$$\phi(\mathbf{r}) = \phi_{\text{pin}}, \quad \mathbf{r} \in \partial \Omega_{\text{ex}}. \quad (6.1)$$

The conduction band is set to a characteristic value above the Fermi level which is assumed at $E_c - E_F = 0.6 \ldots 0.8$ eV for GaAs surfaces. The choice of an appropriate model for exposed surfaces is a somewhat controversial issue. The Fermi level pinning model leads to unrealistic discontinuous steps of the potential at the edges of the Schottky gates. Also, at low temperatures it is difficult to believe that charges from the 2DEG can move to the surface states in response to shifts in the gate voltages in order to maintain the exposed surface at equipotential and in equilibrium with the substrate. An alternative
6.2. Device simulation

choice is the *frozen surface* model. The surface is treated as a simple dielectric boundary with a fixed (frozen) surface charge density $Q_s$ which does not move in response to the gate voltage. The corresponding Neumann boundary condition reads

$$
\epsilon_s \frac{\partial \phi(r)}{\partial n} = Q_s, \quad r \in \partial \Omega_{\text{ex}},
$$

where $n$ is the unit normal vector of the boundary. The relatively high value of the dielectric constant of the semiconductor means that this boundary condition can be approximated by $\partial \phi/\partial n = 0$. The frozen surface model, however, is only valid if the exposed surface is large and smooth [73], [74]. If the separation of the gates is small as in split-gate structures, the above approximations are not valid and the dielectric has to be modeled as well for a more realistic description of the surface [75].

The structure was $\delta$-doped at the GaAs/AlGaAs interface by evaporating silicon during the growth process and at the same time stopping the GaAs deposition. The dopants are assumed to be positioned in an atomic monolayer. However, the Coulomb repulsion between the dopants and the high growth temperatures of 600 K lead to a spreading of the layer. The doping-sheet acquires a finite width which is assumed to be $1 \ldots 3$ nm depending on the doping density (see [76] and references therein). A typical value for the two-dimensional charge density in the doping layer is $5 \times 10^{12}$ cm$^{-2}$ [72]. If a doping sheet with a constant density and an effective width of 2 nm is assumed, the three-dimensional charge density in the sheet is $3 \times 10^{19}$ cm$^{-3}$. The discrete nature of the donors is neglected and the doping sheet is treated as jellium.

The doping density has to be reduced from that grown in the structure since deep levels are neglected. This is reasonable, since half of the electrons from the doping occupy deep levels. These levels are usually found well below the Fermi level and the associated charge can be assumed to be frozen out. This results in an effective doping density of $1 \ldots 1.5 \times 10^{19}$ cm$^{-3}$. The remaining uncertainty is used to adjust the number of electrons in the quantum dot to a particular value. Comparing the charge densities in the doping sheet and in the 2DEG (Fig. 6.2) one notices that the density of the ionized donors is much higher by two orders of magnitudes. Most of the electrons from the donors are obviously absorbed by the surface states.
6.2.3 Simulation results

The conductance characteristics of the heterostructure was calculated only for a small subspace of the available parameter space. The control gate voltage, i.e. the voltage $V_g$ at gate C was ramped from a negative value of $-800$ mV to the point at which the depletion of the 2DEG underneath the gate is breaking down. Usually, this happens at around $-200$ mV. All other gates were kept fixed at a value of $-700$ mV. The doping density in the $\delta$-doping sheet was used to adjust the charge density, i.e. the number of electrons in the quantum dot.

Figure 6.5 shows the conduction-band edge and the electron densities along the $z$-direction. The conduction-band offset between GaAs and AlGaAs is 360 meV. There is no substrate bias in this example and the substrate Fermi energy is set to zero. Displayed are the densities for a parabolic-band model (Bulk), a two-dimensional electron-gas model (2DEG), and a zero-dimensional electron-gas model (0DEG). Only the lowest subband is occupied.

Figure 6.5: Conduction-band edge and electron densities along the $z$-direction at 1 K. The circles at the conduction-band curve show the position of the grid planes in $xy$-direction. The inset displays the wavefunctions of the first two subbands ($\phi_0$, $\phi_1$) corresponding to the 2DEG model.
in the 2DEG model. The subband energies $\phi_1$ and $\phi_2$ are 31 meV and 55 meV above the conduction-band edge at the GaAs/AlGaAs heterojunction. This corresponds to the values of 24 meV and 46 meV calculated by Stern and Das Sarma [77] for a similar structure. The difference of the Fermi energy to the lowest subband is 6 meV.

The conduction-band edge and the electron-charge density in the lateral $xy$-plane is shown in Fig. 6.6. The electron density in the center (quantum dot) was calculated using a 0DEG model solving a 3D-Schrödinger equation. The number of electrons in the quantum dot was found to be $N \approx 50$. The electron gas shows a structure which is related to the shape of the 3D-wavefunctions.

**Figure 6.6:** Split-gate structure (top), electron density (middle) and conduction-band edge (bottom) in a lateral $xy$-plane. The density and band-edge cuts are taken 8 nm below the GaAs/AlGaAs heterojunction in the 0DEG/2DEG.
of the electrons which are used to calculate the quantum dot charge density. The form of the conduction band in the quantum dot region is far from being parabolic and the dot electron charge is distorted towards a non-circular shape. The shape of the confinement is determined by the gate geometry in the first place. This finding is in contrast to the calculations of Kumar et al. [78] which showed a nearly parabolic shape of the confinement potential for a similar layer structure. However, they calculated the self-consistent groundstate for only around 12 electrons in the quantum dot and for a top gate symmetric in the $xy$-plane. In order to compare with their results, the electron number in the quantum dot was adjusted to around 10 electrons using the donor density in the doping sheet as a tuning parameter. The charge density in this case was found to exhibit a similar shape to the one shown in Fig. 6.6 (middle). The diameter of the electron gas, however, was significantly reduced to around $\sim 100$ nm in the $x$-direction. Measurements in this low-filling regime would be difficult (at least with this device geometry) because of the increased barrier thickness. The electron charge density is concentrated more in the center of the quantum dot and the coupling of the single-particle eigenstates to the leads is very weak.

The single-particle level spectrum and a cut through the wavefunctions along the $x$-direction through the quantum-point contacts (QPC) are shown in Fig. 6.7. The level spectrum is not a harmonic oscillator spectrum associated with a parabolic confinement. The first two eigenstates are non-degenerate and a near degeneracy occurs only with the

Figure 6.7: Cut through the single-particle level spectrum and the conduction-band edge in the $x$-direction along the QPCs.
third and the fourth level which can also been seen in the 2D plots of the individual wavefunctions labeled 2 and 3 in the $xy$-plane (Fig. 6.8). The distorted shape of the quantum dot potential is also reflected in the shape of the individual wavefunctions shown in Fig. 6.8. One observes a tendency to form quasi-one-dimensional wavefunction-\textit{scars}. These scars are related to the classical trajectories of particles entering the dot, then bouncing within, and finally exiting. Two pronounced families of scars are visible. The first (S1) is related to the wavefunctions labeled 4, 6, 9, 12, 15, 19, 23, 28, 33 and the second (S2) to 5, 8, 11, 16, 20, 26, 31. The scars of one family have always the same shape and position in the quantum dot, however, the number of nodes of the wavefunctions increases with increasing energy. The classical trajectories for S1 and S2 are schematically shown in Fig. 6.9. The occurrence of scars is an indication that the quantum dot is already acquiring properties.
of a disordered ballistic\(^1\) structure in which conductance fluctuations can be understood as interference of phase-coherent electrons traversing the dot via a number of distinct classical paths. This is the dominant regime in large, highly populated \((N \geq 200)\) quantum dots.

Figure 6.10 shows the tunneling rates calculated using the wavefunctions in Fig. 6.8, versus the eigenenergies at a particular value of the gate voltage \((V_g = -400\) meV). The tunneling rates belonging to one particular scar family are linked by dotted lines. It can be seen that especially the rates related to S1 have almost the same value which can be explained by the shape of the wavefunctions which is similar, leading to similar values for the overlap integral between the lead and the

---

\(^{1}\)In ballistic structures the elastic mean free path \(l\) of the particles exceeds the structure size. This, however, is always the case in small quantum dots at low temperatures where the mean free part can be a few microns.
quantum dot wavefunction. The highest values, however, are related to the wavefunctions which are more uniformly distributed over the quantum dot area, showing some alignment along the $x$-axis (see for example states 18, 24, 29 and 35).

Figure 6.11 shows the electron charging of the quantum dot and the conductance versus the gate voltage. The envelope of the conductance peaks exhibits some modulation which is more pronounced at the lower temperature ($T = 0.7$ K). Over the given range the quantum dot is charged, successively increasing the number of electrons from 49 to 66. One of the main reasons for envelope modulation is that at finite temperature more channels are thermally accessible than at $T = 0$ K, i.e. when $N$ is in the middle of a shell, more channels are available for charge transport than if $N$ is just entering or leaving the shell. In Fig. 6.11 a pronounced step in the peak height occurs at $V_g \approx -400$ mV. On the other side, the conductance peaks show an overall tendency to linearly increase their peak height with an increasing number of electrons in the dot. This can be explained by the stronger coupling of the quantum dot states to the leads because of the extended size of the ODEG. Also visible is the thermal dependence of the conductance minimum between the peaks. In agreement with the predictions of Section 3.4.2 the minimum of the conductance peaks approaches the maximum if the temperature is increased.

At this point, one notes that the values of the conductance is very low even at the conductance peaks which have heights of a few $10^{-5} \frac{e^2}{h}$. Indeed, the

Figure 6.11: Electron charging $N$ (dotted line) and conductance $G$ (solid line) in units of the conductance quantum $e^2/h$ versus the gate voltage $V_g$ at two different temperatures.
Figure 6.12: Electron charging $N$ (dotted line) and conductance $G$ (solid line) in units of the conductance quantum $e^2/h$ versus the gate voltage $V_g$ at two different temperatures.

Noise level in conductance measurements experiments of this structure [72] is to be found at $\sim 10^{-3} e^2/h$ which means that above results are below the range which can be accessed in standard measurements. Therefore, the doping density in the doping sheet was changed and the number of electrons in the quantum dot was adjusted to values around $N = 85$ at a negative gate bias of $V_g = -800$ mV. Now, the conductance peaks exhibit values of the order of the conductance quantum which is $G_Q = 4e^2/h$. The reason for this is the already mentioned lateral extension of the 0DEG if the dot is getting filled with more electrons. The electron-wavefunction coupling to the leads becomes stronger resulting in higher tunneling rates, Fig. 6.13. The overall values of the tunneling rates are some four magnitudes larger then in the previous case. This corresponds to the experimental observations [72] which show the onset of the oscillations at $V_g = -1000$ mV with around $N = 80$ electrons in the quantum dot. The 2DEG below the gates starts to get depleted at negative gate voltages of around $-200$ mV. Above this value at around $-200 \ldots -100$ mV at the control gate, the dot starts to extend below the gate. Experimentally, one observes a significant change in the spacing of the Coulomb-blockade peaks. The spacing becomes smaller because of the much larger capacitance of the extended quantum dot. In the simulations, the region where the electron gas is modeled as a 0DEG has to be extended below the control gate if this happens. The calculations here presented explicitly exclude this case and the conductance calculation only includes the range where the gate regions are fully depleted. However, the beginning breakdown is always detected because of a sudden change in the conductance characteristics.
The modulations visible in both curves in Fig. 6.12 are due to coherent-resonant transport through the quantum dot. The quantum dot is acting as a resonator and the electrons are reflected by the quantum point contacts similar to a Fabry-Perot interferometer. The period of the oscillations is roughly 300 mV. The oscillations can not be attributed to shell filling effects for two reasons. First, shell-filling is mainly observable at low temperatures, usually dominating the conductance characteristics at a few tens of mK. Second, both the upper and the lower envelope of the peaks are modulated which is a significant feature also present in the measurements [72] and clearly identified as due to coherent-resonant transport. A quantum-mechanical interpretation can be given considering the tunneling rates for different wavefunction symmetries. It has already been ar-

**Figure 6.13:** Tunneling rates, i.e. elastic couplings of the dot wavefunctions to the leads versus the single-particle eigenenergies. The dotted lines link the rates which belong to one of the scar families (S1 and S2).

**Figure 6.14:** Schematic of the chemical potential as function of $V_g$. The Fermi level in the leads, $E_F$, is indicated.
gued that those states which are more situated in the center of the quantum dot, showing some alignment with the $x$-direction contribute the most to the transport by having the highest tunneling rates $\Gamma_p$. Strong coupling can be observed for example for the states labeled 24, 29, and 34 (Fig. 6.13). If one takes the two-fold spin-degeneracy of the states into account, a period for the envelope modulation of around 10 peaks should be visible in Fig. 6.12. In both curves, which are taken at different temperatures, modulations of roughly this period can be observed.

Figure 6.14 shows a schematic of the quantum-dot electron number $N$ and the quantum-dot chemical potential $\mu$ as a function of the gate voltage $V_g$ according to the orthodox theory at $T = 0$ K. The chemical potential approaches the Fermi level, $E_F$, in the leads at the step from electron number $N - 1$ to $N$. The occurring step in the chemical potential can serve as a measure for the self-capacitance of the quantum dot, i.e. $\Delta\mu = e^2/C_\Sigma$.

Figure 6.15 shows the gate-voltage dependence of the single-particle eigenvalue spectrum close to the quantum-dot chemical potential $\mu$. The sawtooth-like structure reflects the changes in the electrostatic and the chemical potential at the electron charging steps. The filling of a level pushes the energies of the already filled levels energetically lower. This requires addi-

**Figure 6.15:** Single-particle eigenvalue spectrum (dotted lines) and quantum dot chemical potential (solid line) versus gate voltage at $T = 0.8$ K. The step-like curve is the electron number in the dot. The Fermi energy in the leads is $E_F = 0$ eV. The occupation of a previously empty level occurs when the level crosses the chemical potential.
6.2. Device simulation

Tional energy and the charging energy is increased. Another effect is related to the degeneracies of the levels. Some of the levels are degenerate at some values of $V_g$ but the degeneracy lifts with increasing gate voltage (Fig. 6.16). Other levels are non-degenerate, becoming degenerate at some value of $V_g$ and eventually becoming non-degenerate again. This effect is related to changes in the shape of the quantum-dot potential due to subsequent filling of states with different symmetry (see Fig. 6.8). This leads to oscillations in the proportions of the spatial dimensions of the quantum-dot confinement, i. e. space-symmetries of the potential are established and broken again. Also visible are shells of various sizes which also change with the gate voltage.

Figure 6.16: Complete single-particle eigenvalue spectrum versus gate voltage over the whole voltage range at $T = 0.8 \text{ K}$. 
Gate capacitance $C_g$ from numerically differentiating the Coulomb staircase $N(V_g)$ at $T = 0.8 \, K$ in Fig. 6.12.

The gate capacitance $C_g$, i.e. the capacitive coupling of the control gate to the quantum dot may be defined as

$$C_g = \frac{dQ}{dV_g}, \quad (6.3)$$

where $Q$ is the quantum dot charge. Therefore, the gate capacitance is calculated by numerically differentiating the Coulomb-staircase, Fig. 6.17. Another method to evaluate the gate capacitance is via the spacing of the Coulomb-blockade peaks. The difference $\Delta V_g$ between the $N$th and the $(N+1)$th peak is used and the gate capacitance is calculated as

$$C_g(N) = \frac{e}{\Delta V_g} = \frac{e}{V_g(N+1) - V_g(N)}. \quad (6.4)$$

The capacitance is now a function of $N$. Both methods give similar results. The gate capacitance increases from 5 aF to a value of around 7 aF over the range of the conductance calculations. These values are too small if compared with the measured value which is $C_g = 32 \, aF \, [72]$. Consequently, the spacing of the conductance peaks is too wide compared with the measured results. A discussion of possible reasons for this discrepancy is given in the next section. If one now considers a factor $32 \, aF / 5 \, aF \sim 6$ by which the calculated spectra are stretched in comparison to the measured ones, these results correspond with each other.

The concept of capacitances is only meaningful for conductors, whose electrostatic potential can be specified by one number, the voltage. The concept of self-capacitances for a semiconductor quantum dot whose dimensions
are comparable with the screening length is therefore a somewhat delicate issue. One possible and still meaningful definition is based on differential capacitances [79], i.e., a self-capacitance $C_{dd}$ is defined as the amount of work, $\Delta V$, to bring fixed amount of charge, $\Delta Q$, to the quantum dot. For $\Delta V$ one uses

$$e\Delta V = \mu(N + \Delta N) - \mu(N), \quad (6.5)$$

where $\mu(N)$ is the chemical potential of the $N$-particle system. For single-electron charging the obvious choice for $\Delta N$ is $\Delta N = 1$ and $\Delta Q = e$. Therefore, the capacitance can be written as

$$\frac{e^2}{C_{dd}(N)} = \mu(N + 1) - \mu(N). \quad (6.6)$$

The quantum dot self-capacitance $C_{dd}(N)$ (Fig. 6.19) was calculated using the self-consistent structure results. The quantum dot chemical potential was extracted from the free energy minimization. The zig-zag shape of the capacitance curve is due to even-odd level filling of the spin-degenerate single-particle levels in the quantum dot (see also Eq. 2.21 in Chapter 2). This effect has previously been reported in self-consistent structure calculations of semiconductor quantum dots [28].

Another method to obtain a quantum dot self-capacitance should be mentioned here. A self-capacitance $C'_{dd}$ can be calculated by numerically differ-
differentiating the Coulomb-staircase with respect to the chemical potential, i.e.

\[
C'_{dd}(V_g) = \frac{dQ}{d\mu}. 
\]

The capacitance is now a function of the gate voltage \(V_g\) (Fig. 6.20). Striking features in the capacitance curve are the two singularities which result from the two extrema in the quantum-dot chemical potential left and right from the conductance peak. Also, the self-capacitance is negative between the two singularities in the range of the gate voltages where Coulomb blockade is lifted and the quantum dot charge is shifted from \(N\) to \(N + 1\). The negative capacitances reflect the highly unstable regime during the charging of the dot. The dot is free to oscillate between two groundstate energies which correspond to the two electron numbers. Its most stable configuration is acquired between the blockade peaks, where the electron number is integer at low temperatures. In this regime the chemical potential is almost linear and consequently the self-capacitance is almost constant. There, the value of the self-capacitance is around 100 ... 150 aF which corresponds to the classical electrostatic limit (for the GaAs quantum dot \(\epsilon = 13\), \(d = 200\) nm, \(C = 2\pi\epsilon_0\epsilon d \approx 150\) aF).

In the classical limit for large, highly populated quantum dots with metallic properties the self-capacitance is much higher then the values observed here. The consequence is that the steps in the chemical potential vanish since they are proportional to \(C_{\Sigma}^{-1}\) and the chemical potential becomes a constant. Consequently, the capacitance in the metallic regime is constant as well since the steps in the charging characteristics of the quantum dot vanish and the in-
6.3 Comparison with measurements

The measured linear-response conductance as a function of the control-gate voltage $V_C$ for the simulated structure is shown in Fig. 6.21 [72]. All other gates were kept at constant voltages defining the operating point of the device. The conductance shows dramatic oscillations of two orders of magnitude. The peak height is of the order of the conductance quantum $(e^2/\hbar)$. This measurement was performed at a very low temperature of 25 mK. The number of electrons in the quantum dot was estimated to be $N = 75 \ldots 250$. The simulated results shown in Fig. 6.12 in the previous section are obtained with a quantum dot with an electron number of $N = 87 \ldots 110$ and as such the sit-

Figure 6.20: Quantum dot self-capacitance $C_{dd}$ as a function of the gate voltage $V_g$ at $T = 0.8$ K.

crease of the island charge is linear. This resolves the problems of the negative capacitances and makes them a unique feature of single-electron charging and Coulomb blockade invisible in the classical (metallic) regime.
Figure 6.21: Coulomb-blockade oscillations measured as a function of the control-gate voltage $V_C$. All other voltages are kept constant. (Taken from [72]. Courtesy T. Heinzel.)

The simulation should closely resemble the measured one. The main difference to the experimental results is the much larger period of the conductance oscillations. This corresponds to the much lower simulated gate capacitance of $5 \ldots 7 \ \text{aF}$ compared to the measured one which is around $32 \ \text{aF}$. A possible reason for this discrepancy is the assumption of Fermi-level pinning at the exposed parts of the GaAs surface. More generally speaking, the exposed surface has a large influence on the general electrostatics of the device, also influencing the coupling of all other gates amongst each other and the coupling of the gates to the substrate. This is especially evident for narrow split-gate structures as the one used here [75].

Figure 6.22 shows the conductance oscillations over a much wider range of gate voltages and for several temperatures. One observes pronounced envelope modulations of the conductance peaks. The period of these modulations is around $50 \ldots 80 \ \text{mV}$. The simulated data (Fig. 6.12) show a similar modulation of the peak height, however, the period is again larger as with the measured data. Both the measured and the simulated modulations are due to coherent-resonant transport between the two QPCs, i.e. because of a Fabry-Perot like behavior of the quantum dot, acting as the cavity. Even though the simulations could not be performed for temperatures lower than $800 \ \text{mK}$, the
same trend in the height of the conductance peaks for different temperatures was observed. In agreement with the theoretical predictions the measurements show a smaller amplitude for the conductance oscillations and a general trend to higher conductance values between the peaks. This trend is confirmed by the simulations (Fig. 6.11 and Fig. 6.12). Also in agreement with the measurements, the envelope modulations become less pronounced.
Chapter 7

Just as a swimmer, still with panting breath, now safe upon the shore, out of the deep might turn for one last look at the dangerous waters, ...

DANTE, »Inferno, I«

Achievements, failures and prospects

This work describes the theoretical foundations and the numerical algorithms used in a simulation software for single-electron devices. A prototype simulator has been implemented and its capabilities are tested at a particular example, a GaAs/AlGaAs heterostructure SET. To some extent, the approach chosen is very limited because of its restriction to the quasi-equilibrium. Nevertheless, in the context of device simulation which relies on the drift-diffusion equations, inappropriate to simulate single-electron tunneling, this work means considerable progress.

Some aspects should be especially emphasized. First, the inclusion of size quantization and many-body effects in the solution of a non-linear Poisson equation and the implementation of stable numerical algorithms has been achieved. This allows for a self-consistent evaluation of single-electron tunneling. Second, a sound method for the calculation of the tunneling rates for soft electrostatic barriers which is based on the transfer-Hamiltonian method is included. Finally, these results are used in a free-energy minimization proce-
dure which allows for the computation of the discrete charging of the quantum dot and the linear-response characteristics of a SET at finite temperatures and for realistic geometries.

The software, for which the preliminary name SIMNAD (Simulation and Modeling of Nanodevices) is used, is embedded in the TCAD suite of ISE AG (Zürich). This allows for the utilization of the ISE grid editor TED-ISE which is especially suitable for tensor-product grids. Furthermore, the visualization of results can be done by graphical software such as INSPECT-ISE since all output files are compatible with the ISE datastructures.

Since the development of the SIMNAD simulation software was initialized in 1997, a lot of structures, some even with chances to find device applications, have been presented. It more and more seems to be clear that anything that could be of practical interest for logical circuits should be based on silicon. Simulation of silicon single-electron transistors is well within the scope of the SIMNAD software. The calculation of the quantized levels for a six-valley bandstructure is possible in all three confinement models (0DEG, 1DEG, and 2DEG) implemented in SIMNAD and for a (100)-substrate orientation. However, silicon SETs, most of them in SOI-technology, have rather complicated three-dimensional geometries that require huge grids and sometimes cause severe numerical problems. Also, the simulation of silicon quantum dots is often not possible using the single-subband approximation (Eq. (4.38) and Eq. (4.39)) that allows for a considerable speed-up of simulations of GaAs/AlGaAs heterostructure SETs.

Apart from silicon, GaAs/AlGaAs based heterostructures are still first choice when properties of confined electrons in zero-dimensional electron gases are investigated. The simulation of these structures can be carried out in a numerically stable and reasonably fast manner. However, there is still need for further investigation of the role of the exposed surface and the doping sheet including the doping levels in order to become really predictable. It should also be mentioned that the numerical algorithms implemented still fail to converge at very low temperatures below a few hundred mK. This regime, however, is experimentally of particular interest. At higher temperatures, effects related to the single-particle spectrum are often disguised by the thermal broadening of the levels.

Further work should start by dealing with the shortcomings of this work. Certainly, the direction of further work depends on what the simulation software will be used for. If a TCAD-oriented tool is planned, some work has to be
devoted to a more advanced discretization scheme at unstructured grids which goes beyond the simple tensor-product/finite-difference approach used in the present software. The use of automatic meshing, however, requires the formulation of adequate meshing criteria, especially suitable for the Schrödinger-Poisson system. At this point the author has no knowledge that a stable, self-consistent scheme for the Schrödinger-Poisson equation system employing a box discretization for instance on a general, unstructured grid has ever been implemented.

On the physics side, the inclusion of magnetic fields would be desirable since the application of magnetic fields reveals a lot about the electronic structure of the electronic gases in quantum dots. A whole literature of experimental data related to the effects of magnetic fields in quantum dots is available. Some work on the discretization of the vector potential at tensor-product grids using finite differences has been done in the course of this project [80]. A test implementation using the object-oriented program structure of SIMNAD is already available.

The SIMNAD code has the capability for providing the advanced application oriented research in academia as well as in industry with more insight into one of the most promising branches of modern device physics. The single-electron concept and single-electron devices are certainly beyond the ultimate scaling limits of the conventional CMOS, however, they may proof to be its ultimate replacement. Computer simulations and complex TCAD experiments will be of key importance with this technology. Goal of this work was certainly not to clarify the question whether or not the single-electron concept will have any applications. It is only reasonable to expect that the basic research undertaken today will eventually find use in further microelectronic- or nanoelectronic devices. Some of the effects discussed in this work may play a more significant role for conventional devices, intentionally or not. Others may only have the potential for niche applications. However, that single-electron devices will never have an impact on the industry is certainly a verdict to early to be proclaimed. The current status of nanodevices seems somewhat like the situation in the 1940s when there was no reason to believe that vacuum tubes could ever be displaced.
Seite Leer / Blank leaf
Appendix A

Thermodynamic properties of localized systems

_Grand canonical_ systems at a temperature $T$ are characterized by a Fermi energy $E_F$, a particle number $N$, a volume $V$, and a total energy $E$. In equilibrium the volume $V = V_0$, the particle number $N = N_0$, and the energy $E = E_0$ are strictly constant. The entropy of the system $S_0 = k_B T \ln \Gamma_0$ is maximal, therefore being constant as well. $\Gamma_0$ is the dimension of the classical phase space, i.e. gives the number of accessible states of the system.

The system (Fig. A.1) is subdivided into a reservoir $R$ with a set of thermodynamic variables $N', V', E', \Gamma', S'$ and a small subsystem $A$ characterized by $N, V, E, \Gamma, S$. The subsystem $A$ is embedded in the much larger reservoir $R$ and $V_0, N_0, E_0$ are properties of the system as a whole ($A + R$). The subsystem exchanges energy and particles with the reservoir. The particle number in the reservoir changes according to the _law of equilibrium changes_

\[ dE' = TdS' - PdV' + E_F dN'. \quad (A.1) \]

The important assumption is made that the subsystem $A$ is so small that $T$ and $E_F$ of the reservoir remain fixed even though the subsystem fluctuates among its possible states which are labeled by $\alpha$. It is furthermore assumed that $E_\alpha \ll E_0$ and $N_\alpha \ll N_0$. Even if some of the states $|\alpha\rangle$ may violate the above condition, it is only a small number which can safely be ignored.
Using Eq. (A.1), one obtains for the constant derivatives with respect to entropy $S'$ and the particle number $N'$

\[
\frac{\partial E'}{\partial S'} = T \quad \text{(A.2)}
\]

\[
\frac{\partial E'}{\partial N'} = E_F.
\]

Within this model the *Gibbs distribution* for the occupation numbers of the small subsystem in particle exchange with the reservoir can be derived. The number of possible states in $A$ is denoted $\Gamma$ and that in $R$ is denoted $\Gamma'$. Therefore, the total number of states in $A + R$ is $\Gamma \Gamma' = \Gamma_t$. In equilibrium the entropy is maximal, $\Gamma_t = \Gamma_0$ and $S_t = S + S' = S_0$. The probability for the system as a whole of being in a particular state is $P_{eq} = 1/\Gamma_0$. The number of choices for the reservoir states given the subsystem is in one particular state is $\Gamma' = |\alpha\rangle = |N_\alpha, E_\alpha\rangle$ is $\Gamma'_\alpha$. With $\Gamma_{t\alpha} = \Gamma_\alpha \Gamma'_\alpha$ and $\Gamma_\alpha = 1$ (since the system is in state $|\alpha\rangle$ by assumption), the probability of this state can be written as

\[
P_\alpha = \frac{\Gamma'_\alpha}{\Gamma_0}, \quad \text{(A.3)}
\]

i. e. out of $\Gamma_0$ possible states of the total system, there are $\Gamma'_\alpha$ states that find $A$ in state $|\alpha\rangle$.

The reservoir entropy $S'_\alpha = k_B T \ln \Gamma'_\alpha$ depends on the energy and the particle number in the subsystem

\[
S'_\alpha = S'(E_0 - E_\alpha, N_0 - N_\alpha), \quad \text{(A.4)}
\]

and with $S_0 = k_B T \ln \Gamma_0$ it follows that

\[
S_0 - S'_\alpha = -k_B \ln \frac{\Gamma'_\alpha}{\Gamma_0} = -k_B \ln P_\alpha. \quad \text{(A.5)}
\]

It has to be emphasized that $S_0 - S'_\alpha$ is not the entropy of the subsystem $A$. The entropy of the subsystem is zero, since $A$ is in a given state $|\alpha\rangle$. The
equilibrium entropy of $A + R$ is the average of $S'_{\alpha}$ over all possible states. The *thermodynamic average* of a quantity $f$ is

$$\overline{f} = \sum_{\alpha} P_{\alpha} f_{\alpha}, \quad \text{with} \quad \sum_{\alpha} P_{\alpha} = 1,$$

(A.6)

which for the entropy $S = \overline{S'_{\alpha}}$ leads to

$$S = -k_B \sum_{\alpha} P_{\alpha} \ln P_{\alpha}, \quad P_{\alpha} = \exp \left[ -k_B^{-1} (S_0 - S'_{\alpha}) \right].$$

(A.7)

With the law of equilibrium changes, Eq. (A.1), one writes

$$S'_{\alpha} = S' (E_0 - E_{\alpha}, N_0 - N_{\alpha}) = S' (E_0, N_0) - \left( \frac{\partial S'}{\partial E} \right)_{V',N'} E_{\alpha} - \left( \frac{\partial S'}{\partial N} \right)_{V',E'} N_{\alpha} = \text{const} - \frac{E_{\alpha}}{T} + E_F N_{\alpha},$$

(A.8)

which substituted into Eq (A.7) gives the Gibbs distribution

$$P_{\alpha} = Z^{-1} \exp \left[ -\frac{E_{\alpha} - N_{\alpha} E_F}{k_B T} \right],$$

$$Z = \sum_{\alpha} \exp \left[ -\frac{E_{\alpha} - N_{\alpha} E_F}{k_B T} \right],$$

(A.9)

using the normalization condition for the $P_{\alpha}$ and $Z$ is the partition function of the grand canonical ensemble, i.e. the *grand canonical partition function*. The thermodynamic averages of the energy $E = \overline{E}$ and the particle number $N = \overline{N}$ of the subsystem $A$ are calculated using the Gibbs distribution in Eq. (A.6) giving

$$E = \sum_{\alpha} E_{\alpha} P_{\alpha} \quad \text{and} \quad N = \sum_{\alpha} N_{\alpha} P_{\alpha}.$$

(A.10)

Inserting the Gibbs distribution into the equation for the entropy of the system $S$, Eq. (A.7), one obtains

$$k_B T \ln Z = E - TS - N E_F = F - N E_F = \Omega,$$

(A.11)

and consequently to an explicit form of the *grand canonical potential* $\Omega$ for the subsystem $A$, i.e.

$$\Omega = -k_B T \ln \sum_{\alpha} \exp \left( -\frac{E_{\alpha} - N_{\alpha} E_F}{k_B T} \right)$$

or

$$\Omega = -k_B T \ln Z.$$

(A.12)
All thermodynamic variables at fixed $V, T, E_F$ of the system can now be calculated by thermodynamic averaging of $E, S, N$ if the energies and the electron numbers $\{E_\alpha, N_\alpha\}$ for all states $|\alpha\rangle$ of the subsystem $A$ are known.

If the subsystem contains a fixed number of electrons and no particle exchange between the reservoir and the subsystem occurs, the system has to be treated as a canonical ensemble. Instead of keeping the chemical potential fixed, a constant electron number according to $N_\alpha = N$ for all $|\alpha\rangle$ is assumed and the thermodynamic variables of the system are determined by thermodynamic averaging at a particular $N, T, V$. The thermodynamic potential is the Helmholtz free energy $F(N)$

$$F(N) = -k_B T \ln \sum_\alpha \exp \left( - \frac{E_\alpha}{k_B T} \right)$$

with the canonical partition function $Z(N)$ and the Gibbs distribution

$$P_\alpha = Z^{-1}(N) \exp \left( - \frac{E_\alpha}{k_B T} \right)$$

$$Z(N) = \sum_\alpha \exp \left( - \frac{E_\alpha}{k_B T} \right).$$

To illustrate the principle difference between the Gibbs distribution and the Fermi-Dirac distribution for single-particle levels, the following example is used. Consider a two level system $\{E_1, E_2\}$ with $N = 1$ particles. According to Eq. (A.16) the occupation probability for the level $\alpha$ is given as

$$P_\alpha = P(E_\alpha | N = 1) = \frac{\exp \left( - \frac{E_\alpha}{k_B T} \right)}{\exp \left( - \frac{E_1}{k_B T} \right) + \exp \left( - \frac{E_2}{k_B T} \right)}.$$ 

If cast into a slightly different form, this results in

$$P_\alpha = \left[ 1 + \exp \left( \frac{E_\alpha - E_F}{k_B T^*} \right) \right]^{-1}$$

with $E_F = \frac{1}{2} (E_1 + E_2)$ and $T^* = \frac{1}{2} T$, which corresponds to a Fermi-Dirac distribution, however, at a fictitious temperature $T^*$, which is half the true temperature of the electronic system. The
Fermi energy $E_F$ can be defined as the average of the two energies $E_1$ and $E_2$. For higher electron numbers and the number of levels approaching infinity, the Gibbs distribution becomes equivalent to the Fermi-Dirac distribution. Numerical analysis confirming this fact for semiconductor quantum dots can be found in [31].
Appendix B

Transition rates

Subject of this appendix are the transition rates between the many-particle groundstates in the quantum dot. The calculation is based on time-dependent perturbation theory in lowest order in the tunneling terms.

B.1 Time-dependent perturbation theory

The evolution of the Hamiltonian operator $\mathcal{H}$ with time is described by the time-dependent Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle = \mathcal{H} |\Psi(t)\rangle,$$

(B.1)

where $|\Psi(t)\rangle$ is the time-dependent state vector of the system. The time-evolution of the state vector from an initial state $|\Psi(t_i)\rangle$ to a final state $|\Psi(t_f)\rangle$ can be written as

$$|\Psi(t_f)\rangle = \mathcal{U}(t_f, t_i)|\Psi(t_i)\rangle,$$

(B.2)

with a time propagator $\mathcal{U}$ of the form

$$\mathcal{U}(t_f, t_i) = \exp \left[ -\frac{i}{\hbar} \int_{t_i}^{t_f} dt \mathcal{H}(t) \right],$$

(B.3)
which reduces to

\[ U(t_f, t_i) = \exp \left[ -\frac{i}{\hbar} \mathcal{H}(t_f - t_i) \right] \quad (B.4) \]

if the Hamiltonian operator is not explicitly time-dependent. The propagator is usually evaluated via expansion of the exponential function in powers of the argument. The direct evaluation of the time propagator is in general a very difficult, if not impossible task. A very common procedure is to split the Hamiltonian into a time-independent part \( \mathcal{H}_0 \) for which the stationary Schrödinger equation \( \mathcal{H}_0 |\psi_i^{(0)}\rangle = E_i^{(0)} |\psi_i^{(0)}\rangle \) can be solved exactly and a time-dependent part \( \mathcal{H}_p(t) \)

\[ \mathcal{H} = \mathcal{H}_0 + \mathcal{H}_p(t). \quad (B.5) \]

The time-dependent part is treated as weak perturbation to the system. The time propagator is expanded in terms of \( \mathcal{H}_p \) in the basis of eigenstates of \( \mathcal{H}_0 \). The total propagator is given as a series

\[ U(t_f, t_i) = U_0(t_f, t_i) + \sum_{n=1}^{\infty} \left( \frac{1}{i \hbar} \right)^n U_n(t_f, t_i), \quad (B.6) \]

where \( U_0 \) is the propagator of the unperturbed system

\[ U_0(t_f, t_i) = \exp \left[ -\frac{i}{\hbar} \mathcal{H}_0(t_f - t_i) \right] \quad (B.7) \]

and \( U_n \) the \( n \)th order correction due to the presence of the perturbation

\[ U_n(t_f, t_i) = \int_{t_i}^{t_f} \mathrm{d}t_n \int_{t_i}^{t_n} \mathrm{d}t_{n-1} \cdots \int_{t_i}^{t_2} \mathrm{d}t_1 U_0(t_f, t_n) \mathcal{H}_p(t_n) \times \]

\[ \times U_0(t_n, t_{n-1}) \mathcal{H}_p(t_{n-1}) \cdots \mathcal{H}_p(t_1)U_0(t_1, t_i). \quad (B.8) \]

A small perturbation of the system due to \( \mathcal{H}_p \) can be assumed if the matrix elements of the operator are small. If, furthermore, the system is considered to evolve in 'short-time' periods then

\[ \langle \psi_i^{(0)} | \mathcal{H}_p | \psi_i^{(0)} \rangle (t_f - t_i) / \hbar \ll 1. \quad (B.9) \]

Consequently, the series converges very fast and only the lowest terms in \( n \) need to be considered. Since tunneling is always a small perturbation to the isolated quantum dot, i.e. the tunneling terms are always small compared to
the other energy scales in the system, only the first order term in the expansion of $\mathcal{U}$ in terms of the perturbation $\mathcal{H}_p$ needs to be considered. The perturbation does not explicitly depend on time, it rather depends on the time difference $t = t_f - t_i$. The probability amplitude, $A(t)$, for one tunneling event is defined as

$$|\psi_i^{(0)}\rangle \longrightarrow |\psi_f^{(0)}\rangle : \quad A_{f,i}(t) = \langle \psi_f^{(0)} | \mathcal{U}(t) | \psi_i^{(0)} \rangle. \quad (B.10)$$

The unperturbed operator (0th order perturbation) generates only the trivial time dependence but does not allow for transitions

$$|\psi_i^{(0)}(t)\rangle = \exp \left[ -\frac{i}{\hbar} E_i^{(0)} t \right] |\psi_i^{(0)}(0)\rangle. \quad (B.11)$$

However, one is interested in transitions between different many-particle eigenstates where $i \neq f$. In this case, the lowest perturbation gives the probability amplitude

$$A_{f,i}^{(1)}(t) = \langle \psi_f^{(0)} | \frac{1}{i\hbar} \int_0^t dt_1 \mathcal{U}_0(t - t_1) \mathcal{H}_p \mathcal{U}_0(t_1) | \psi_i^{(0)} \rangle, \quad (B.12)$$

which becomes after integrating over $t_1$

$$A_{f,i}^{(1)}(t) = M_{f,i} \frac{1}{E_f^{(0)} - E_i^{(0)}} \left[ \exp \left( -\frac{i}{\hbar} E_f^{(0)} t \right) - \exp \left( -\frac{i}{\hbar} E_i^{(0)} t \right) \right]$$

$$\text{with} \quad M_{f,i} = \langle \psi_f^{(0)} | \mathcal{H}_p | \psi_i^{(0)} \rangle, \quad (B.13)$$

where $M_{f,i}$ is the perturbation calculated as the matrix element of the eigenfunctions of the unperturbed Hamiltonian $\mathcal{H}_0$. The probability $W_{f,i}^{(1)}(t)$ that this perturbation has caused the transition is $|A_{f,i}^{(1)}(t)|^2$, the absolute square of the probability amplitude. This leads to the final form of the transition probability

$$W_{f,i}^{(1)}(t) = \frac{1}{t} \left| M_{f,i} \right|^2 \frac{(1/\hbar^2) \sin^2 \left[ (E_f^{(0)} - E_i^{(0)}) \frac{t}{2\hbar} \right]}{\left[ (E_f^{(0)} - E_i^{(0)}) \frac{t}{2\hbar} \right]^2}. \quad (B.14)$$

Assuming 'long enough' times reflected in the uncertainty relation between time and the energy difference $\hbar \ll (E_f^{(0)} - E_i^{(0)}) t$, the following limes leads to the transition rate $\gamma$ which is defined as

$$\gamma_{f,i} = \lim_{t \to \infty} W_{f,i}^{(1)}(t) t^{-1} \quad (B.15)$$
and therefore

\[ \gamma_{f,i} = \frac{2\pi}{\hbar} |M_{f,i}|^2 \delta(E_f^{(0)} - E_i^{(0)}). \]  

(B.16)

### B.2 Effective transition rates

The expression derived above is now used to calculate the transition rates for the many-particle dot states. The Hamiltonian of the unperturbed system consists of a contribution due to the quantum dot and one contribution for the source-side reservoir and the drain-side reservoir respectively.

\[ \mathcal{H}_0 = \mathcal{H}_{\text{dot}} + \mathcal{H}_s + \mathcal{H}_d. \]  

(B.17)

The eigenvalues \( E_{f,i}^{\text{dot}}, E_i^s, E_i^d \) and the eigenfunctions \( |\psi_f^{\text{dot}}\rangle, |\psi_i^s\rangle, |\psi_i^d\rangle \) for each operator are solutions to the time-independent Schrödinger equation for the particular subsystems. The total energy is \( E_i^{(0)} = E_i^{\text{dot}} + E_i^s + E_i^d \) and the total wavefunction factorizes as \( |\psi_i^{(0)}\rangle = |\psi_i^{\text{dot}}\rangle |\psi_i^s\rangle |\psi_i^d\rangle \). The tunneling Hamiltonian which acts as a small perturbation to the system is

\[ \mathcal{H}_p = \mathcal{H}_s^t + \mathcal{H}_d^t \quad \text{with} \quad \mathcal{H}_{s/d}^t = \mathcal{H}_{s/d}^{t,-} + \mathcal{H}_{s/d}^{t,+}, \]  

(B.18)

where one term \((-\rangle\) describes the tunneling out of the dot

\[ \mathcal{H}_{s/d}^{t,-} = \sum_{k,m} T_{k,m}^{s/d} c_{s/d,k}^\dagger c_m \]  

(B.19)

and the term \(\langle+\rangle\) describes the tunneling into the dot

\[ \mathcal{H}_{s/d}^{t,+} = \sum_{k,m} \left(T_{k,m}^{s/d}\right)^* c_m^\dagger c_{s/d,k} \]  

(B.20)

summing over all dot states and lead states. Therefore, it is assumed that transitions between the dot states occur solely if an electron is added or removed from the dot. Transitions which keep the number of electrons fixed and could be due to electron-phonon interactions are neglected here.

Transitions which reduce the number of electrons in the quantum dot are transitions that move one electron from the dot to the source or drain reservoir. Inserting the tunneling Hamiltonian from Eqs. (B.20) and Eq. (B.19)
into Eq. (B.13) leads to
\[ M^{s/d,-}_{f,i} = \sum_{k,m} T^{s/d}_{k,m} \langle \psi^i_f | (c^\dagger_{s/d,k} \psi^i_{f/d}) | \psi^i_{f/d} \rangle \]
\[ = \sum_{k,m} T^{s/d}_{k,m} D^{f,i}_{f,i,m} A^{s/d,-}_{f,i,m} B^{d/s,-}_{f,i,m} \]  \hspace{1cm} (B.21)

with the matrix elements
\[ D^{f,i}_{f,i,m} = \langle \psi^i_f | c_m | \psi^i_{f/d} \rangle \]
\[ A^{s/d,-}_{f,i,m} = \langle \psi^i_f | c^\dagger_{s/d,k} | \psi^i_{s/d} \rangle \]
\[ B^{d/s,-}_{f,i,m} = \langle \psi^i_f | \psi^i_{d/s} \rangle . \]

The expression for \( M^{s/d,-}_{f,i} \) contains a product of an operator acting on the left or right reservoir Hilbert space only (\( c^\dagger_{s/d,k} \)) and an operator acting exclusively on the dot Hilbert space (\( c_m \)). The matrix elements are inserted into Eq. (B.16) for the transition rate. The effective transition rate is calculated by summing over all possible final states in the reservoirs which can be filled leaving an initial dot state empty
\[ W^{D,-}_{f,i} = \sum_f \gamma^{s/d}_{f,i} . \]  \hspace{1cm} (B.22)

The following relations are used
\[ \langle \psi^i_{s/d} | \psi^i_{s/d} \rangle = \delta_{i,f} \quad \text{(completeness of } | \psi^i_{s/d} \rangle) \]
\[ \langle \psi^i_{s/d} | c_{s/d,k'} c^\dagger_{s/d,k} | \psi^s_{s/d} \rangle = [1 - f_{s/d}(E^s_{k} - E^s_{F})] \delta_{k,k'} \]
\[ \langle \psi^i_{f/d} | c^\dagger_{m'} c_m | \psi^i_{f/d} \rangle = f_d(E^d_{m'}) \delta_{m,m'} , \]

where \( f_{dot} \) is the occupation distribution function for the quantum dot and \( f_{s/d} \) is the Fermi-Dirac distribution function for the reservoirs which contains the reservoir chemical potential and the temperature. Since the initial and the final state in the dot are characterized by the electron number \( N \) the following notation is adopted for the transition rate \( W \), \( W^{D,-}_{f,i} = W^{s/d}(N, N - 1) \). Finally, one arrives at the expression
\[ W^{s/d}(N, N - 1) = \sum_m \Gamma^{s/d}_m(E^s_{k}) [1 - f_{s/d}(E^s_{k} - E^s_{F})] f_{dot}(E^d_{m}) \]
\[ \text{with } \Gamma^{s/d}_m(E^s_{k}) = \frac{2\pi}{\hbar} |T^{s/d}_{k,m}|^2 \delta(E^d_{m} - E^d_{s/d} + E^s_{k}) \]  \hspace{1cm} (B.23)
being the tunneling rate for the transition from the $m$th dot state to the $k$th state in the reservoir. The delta function selects the wavevector $k = k(E)$ in the reservoir and the energy of the tunneling electron is $E_{\text{f}}^{\text{dot}} - E_{\text{i}}^{\text{dot}}$.

The same procedure leads to a similar equation for transitions which increase the number of electrons in the quantum dot, i.e. which are transitions that move electrons from the source or drain reservoir into the dot

\[
W^{s/d}(N, N + 1) = \sum_{m} \Gamma^{s/d}_{m}(E_{k}^{s/d})f_{s/d}(E_{k}^{s/d} - E_{F}^{s/d})[1 - f_{\text{dot}}(E_{m}^{\text{dot}})]
\]

The results of Eqs. (B.23, B.24) are known in the literature as Fermi’s Golden rule.
Appendix C

Density-functional theory

The many-particle problem of interacting electrons in an external potential has been a key problem since the introduction of quantum mechanics in the late 1920s. A first solution in the form of a mean-field theory was provided by the Hartree theory in which the electrons are described by an effective single-particle Schrödinger equation and the effect of the other electrons, the screening, is included via an effective potential, the Hartree potential. While the Hartree theory still failed to include quantum mechanical interactions, the later introduced Hartree-Fock theory expanded the concept and included quantum mechanical exchange.

The Hartree-Fock theory proved to be useful for systems with small numbers of particles. However, the expansion of the many particle wavefunction in Slater-determinants, which even for small systems leads to huge numerical problems, made the application to solid-state systems prohibitive [46]. An early alternative to Hartree-Fock theory was the Thomas-Fermi theory [47, 48] which gave, on a heuristic basis, a description of the physical properties of an electron gas by use of a density-functional. However, it took until 1964 that an formally exact theory of the many-particle problem of a system of interacting electrons was formulated. In their pioneering paper Hohenberg and Kohn\(^1\) [49] proved that the ground state properties of an arbitrary system of interacting particles is uniquely represented by its charge density.

\(^1\)W. Kohn received the Nobel prize in Chemistry for his work on density-functional theory in 1998.
Appendix C. Density-functional theory

In the following section a short account of the theory developed by Hohenberg, Kohn and Sham, the density-functional theory (DFT) is given. A more detailed presentation of density-functional concepts can be found in [50].

C.1 Hohenberg-Kohn theorem

Consider a system of interacting particles described by the many-particle Schrödinger equation with the Hamiltonian operator

\[ \mathcal{H} = \mathcal{T} + \mathcal{V} + \mathcal{W} \quad \text{and} \quad \mathcal{H}\left|\Psi\right\rangle = E\left|\Psi\right\rangle \]  

(C.1)

where \( \mathcal{T} \) is the kinetic operator, \( \mathcal{V} \) the operator of the external potential, \( \mathcal{W} \) the operator of the electron-electron interaction which, at least for semiconductor systems, can assumed to be Coulombic, \( \Psi \) is the many particle wave-function, and \( E \) the total energy of the system. The Hohenberg-Kohn theorem states in its first part that the external potential \( V = \langle \mathcal{V} \rangle = \langle \Psi|\mathcal{V}|\Psi \rangle \) is always a unique functional of the true ground state charge density

\[ V = V[n] + \text{const.} \]  

(C.2)

In a second statement a (total) energy functional \( E_V[n] \) is postulated which is a unique functional of the charge density as well and which, assuming charge conservation, can be minimized giving the total energy of the system, \( E_V = E_V[n] \) and \( E_V[n] < E_V[n'] \) if \( n' \) is not the ground state density. Therefore, the ground state density determines the external potential and the many-particle Hamilton operator and consequently all ground state properties of the electron gas in a unique way. The restriction to the ground state properties, however, is one of the mayor weaknesses of density-functional theory. The properties of excited states are beyond the scope of the basic theory.

The second part of the Hohenberg-Kohn theorem enables one to explicitly calculate the total energy of the ground state. Introducing the following form of the total energy functional

\[ E_V[n] = \int \text{d}r \, n(r)V(r) + F[n] \]  

(C.3)

\(^2\)Originally density-functional theory was only introduced for electron gases, however, it is applicable to hole gases as well.
where $F[n]$ is a universal functional independent of the external potential. It contains the kinetic contribution to the total energy and the electron-electron interaction, and therefore is the same for all kinds of electronic systems

$$F[n] = T_s[n] + \frac{1}{2} \int \int \, \text{d}r \, dr' \, \frac{n(r)n(r')}{|r-r'|} + E_{xc}, \quad (C.4)$$

where $T_s$ is the functional of the kinetic energy for non-interacting electrons with the charge density $n$, the second term is the Hartree energy $E_H$, and the last term is the so-called exchange- and correlation energy which is defined as

$$E_{xc} = \langle \mathcal{T} \rangle - T_s[n] + \langle \mathcal{W} \rangle - E_H. \quad (C.5)$$

The exchange- and correlation energy contains the difference between the expectation value of the kinetic operator in the many-particle Hamiltonian, $\langle \mathcal{T} \rangle$, and the energy $T_s$ of the non-interacting system and the difference between the expectation value of the operator of the electron-electron interaction $\langle \mathcal{W} \rangle$ and the classical Hartree energy $E_H$. This definition of $E_{xc}$ ensures the decomposition of $F[n]$ according to Eq. (C.4). The two functionals $T_s$ and $E_{xc}$ are not known from the beginning. $E_{xc}$ is given within appropriate approximations as the local-density approximation (LDA).

### C.2 Kohn-Sham equation

The Hohenberg-Kohn theorem provides the basis for the calculation of the total energy using a variational principle with respect to the charge density. One now tries to find an appropriate representation of this charge density by assuming a system of non-interacting particles with the density $n_s$. According to the Hohenberg-Kohn theorem the total energy functional is given as

$$E_{V_s}[n_s] = T_s[n_s] + \int \, \text{d}r \, n_s(r)V_s(r) \quad (C.6)$$

where $n_s$ can be represented by a sum over squared single-particle wavefunctions $\psi_i(r)$ weighted by an occupation factor $N_i$

$$n_s = \sum_{i=0}^{\infty} N_i |\psi_i(r)|^2. \quad (C.7)$$

It can be shown (Kohn-Sham theorem) that for every system of interacting particles there exists a single particle potential $V_s$ so that the ground state
charge density \( n \) of the interacting system equals that of the non-interacting system \( n_s \). For the effective single-particle potential \( V_s \) an effective single-particle Schrödinger equation, the Kohn-Sham equation\(^3\) can be formulated

\[
\left[ -\frac{\hbar^2}{2m} \nabla^2 + V_s(r) \right] \psi_i(r) = \epsilon_i \psi_i(r). \tag{C.8}
\]

The system of interacting particles in an external potential \( V \) is now projected onto a system of non-interacting particles in an effective single-particle potential \( V_s \). Equation (C.8) can be solved exactly if the effective potential is known. The kinetic energy functional of non-interacting particles can be written as

\[
T_s = -\frac{\hbar^2}{2m} \sum_i \int dr' \psi_i^*(r) \nabla^2 \psi_i(r). \tag{C.9}
\]

Variation of the functional of the total energy Eq. (C.3) with respect to the charge density and consideration of the Kohn-Sham equation leads to an explicit formulation for the effective potential

\[
V_s(r) = V(r) + \int dr' \frac{n(r')}{|r - r'|} + V_{xc}(r). \tag{C.10}
\]

The second term is the Hartree potential \( V_H \), which is related to the electrostatic potential \( \phi \) as \( V_H = -q\phi \). The exchange-correlation potential is formally given by

\[
V_{xc}(r) = \frac{\delta E_{xc}[n]}{\delta n(r)}. \tag{C.11}
\]

Equations (C.7), (C.8) and (C.10) describe in an unique manner the system of non-interacting particles which is related to the system of interacting particles via the identical groundstate charge density. These equations have to be solved self-consistently. Usually, one begins with an assumed \( n(r) \), constructs \( V_s(r) \) from Eq. (C.10) and Eq. (C.11), and finds a new \( n(r) \). The Kohn-Sham total energy (ground-state energy) is given by

\[
E[n] = \sum_{i=0}^{\infty} N_i \epsilon_i - \frac{1}{2} \int dr n(r)V_H(r) + E_{xc} - \int dr n(r)V_{xc}(r). \tag{C.12}
\]

The interpretation of the single-particle eigenvalues \( \epsilon_i \) and the single-particle wavefunctions \( \psi_i \) which are the solutions of the Kohn-Sham equation remains

\(^3\)We restrict to the vacuum problem with \( m \) being the mass of the free electron.
C.3 Local-density approximation

They describe a system of non-interacting (quasi-) particles, but no direct physical meaning for the system of interacting particles can be attached to them\(^4\). Nevertheless, they are interpreted as single-particle (orbital-) energies which sufficiently accurate describe the bandstructure in the crystalline solid-state.

C.3 Local-density approximation

The relatively straightforward formulation of the Kohn-Sham formalism of DFT is only possible because all many-particle properties of the electron gas which are difficult to describe, especially the exchange- and correlation interaction between identical particles are relocated in the functional \(E_{xc}\). This functional is not explicitly known and, as already mentioned, one depends on appropriate approximations.

The most widely used approximation is the local-density approximation (LDA)

\[
E_{xc}^{\text{LDA}} = \int dr \epsilon_{xc}(n) \bigg|_{n(r)}, \tag{C.13}
\]

where \(\epsilon_{xc}\) is a function of the density (not a functional) which is defined as the exchange- and correlation energy per electron for a homogeneous electron gas. It locally approximates the density of the inhomogeneous electron gas. This approximation is certainly good in systems where the charge density is spatially nearly constant. However, LDA proved to be an astonishingly good approximation even in systems where the charge density is strongly inhomogeneous such as molecules (and of course quantum dots). The exchange- and correlation contribution to the effective potential is given according to Eq. (C.11)

\[
V_{xc}(r) = \left(n \frac{d}{dn} + 1\right) \epsilon_{xc}(n) \bigg|_{n(r)} \tag{C.14}
\]

The function \(\epsilon_{xc}\) has been determined for an interacting homogeneous electron gas by Ceperly and Alder [52] using a Quantum-Monte-Carlo method. In this work a parameterized form for the three-dimensional problem is used, which was given by Perdew and Zunger [53].

\(^{\text{4}}\)In fact they are the Lagrange-parameters of the variational scheme used to derive the Kohn-Sham effective potential Eq. (C.10).
Bibliography


[70] This heterostructure has been fabricated by W. Wegscheider at the Walter Schottky Institut of the Technische Universität München.


Curriculum Vitae

Andreas Scholze was born in Arnstadt (Germany), on June 14, 1969. After finishing secondary school in 1989, he studied Physics at the Universities of Jena (Germany) and Edinburgh (Scotland). He specialized in solid-state theory submitting a thesis on density-functional based analysis of diamond surface reconstructions. After graduating from the University of Jena, he joined the Integrated Systems Laboratory (IIS) of the Swiss Federal Institute of Technology in Zürich (ETHZ) working towards the Doctoral degree in Electrical Engineering. During the summer of 1997 he stayed as a visiting research associate at the Beckman Institute for Advanced Science and Technology at the University of Illinois in Urbana-Champaign (UIUC) with the group of Prof. K. Hess. He is currently working within the framework of a national project (MINAST) in the field of nanodevice simulation. The main focus of this work is on the establishment of numerical methods and algorithms for the simulation of single-electron devices at the solid-state level.