DISSERTATION

Physical Mobility Modeling for TCAD Device Simulation

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften

eingereicht an der Technischen Universität Wien
Fakultät für Elektrotechnik und Informationstechnik

von

Zlatan Stanojević

Franz-Haas-Platz 6/1/15
1110 Wien, Österreich
Matr. Nr. 0325501
geboren am 26. März 1984 in Banja Luka

Wien, im Juni 2016

Abstract

Physical device modeling is one of the key technologies to continue semiconductor device scaling. Effects due to the quantum nature of electrons, the band structure of solids, as well as scattering and non-equilibrium transport dominate device performance. Dealing with these effects in device design demands simulation tools that account for them with sufficient accuracy, while being compatible with the Technology Computer Assisted Design (TCAD) concept. The novel contributions of this work to the state-of-the-art are divided into modeling approaches and computational methods.

The contributions in the field physical mobility modeling are the following: A new, *six-valley* effective-mass model for holes is derived from $\mathbf{k}\cdot\mathbf{p}$-theory allowing to qualitatively capture effects of confinement and strain on the valence band structure. A new approach to modeling carrier scattering by polar-optical phonons is devised, where the electrostatic Green's function is used to model the interaction between carriers and oscillating dipoles, thereby taking channel geometry and material variations into account. A novel model for surface and interface roughness scattering is developed; it represents a generalization of the roughness scattering model due to Prange and Nee to non-planar channels of arbitrary shape, allowing consistent modeling of roughness scattering in planar and non-planar structures, such as nanowires and FinFETs. Two new approaches for mobility modeling are discussed based on the solution of the *linearized Boltzmann transport equation* to consistently treat band anisotropy and anisotropy of the scattering processes: One approach computes the first-order response of the distribution function, while the other substitutes the response by a coupled microscopic relaxation time tensor.

The contributions in the field of computational methods include the following innovations: A finite-volume discretization method is devised which is element-based rather than edge-based allowing for discretization of anisotropic effective mass and $\mathbf{k}\cdot\mathbf{p}$ Hamiltonians while preserving their physical properties. A new numerical method is developed which performs an exhaustive search for eigenvalues within a given interval for large sparse systems at negligible added cost; this method allows to maintain a predefined error tolerance when calculating a channel's subbands and its properties. A method for discretizing the scattering operator of the linearized Boltzmann transport equation in $\mathbf{k}$-space is presented based on *symbolic contour integration*.

Finally, the developed methods are applied to calculate the mobilities and conductivities

of a number of existing devices as well as device concepts that are potential candidates for future technology nodes. NMOS and PMOS devices of the 22 nm node as published in 2012 are investigated for channel mobility under variation of crystal orientation and application of strain. It is found that varying the orientation of both the substrate and the fin can improve channel mobility significantly.

Regarding future technology nodes, two device are chosen for investigation: a MISFET with an InGaAs channel and a junction-less Si FET. The properties of the InGaAs device are modeled in detail and excellent agreement with measured results is achieved, highlighting the accuracy of the modeling framework presented in this work. The junction-less device is benchmarked against an inversion-mode device. It is shown that while surface roughness scattering is greatly suppressed in the junction-less device, the resulting mobility improvement is not sufficient to offset the mobility degradation due to increased impurity scattering.

Kurzfassung

Physikalische Bauelemente-Modellierung ist eine der Schlüsseltechnologien für das Fortsetzen der Skalierung von Halbleiter-Bauelementen. Effekte aufgrund der Quanten-Natur von Elektronen, der Bandstruktur von Festkörpern, sowie Streuung und Nicht-Gleichgewichts-Transport dominieren die Bauelement-Eigenschaften. Mit diesen Effekten im Bauelement-Design umzugehen verlangt Simulationswerkzeuge, die diese Effekte mit ausreichender Genauigkeit berücksichtigen. Die Beiträge dieser Arbeit, die den Stand der Technik erweitern, sind in Modellierungsansätze und rechnerische Methoden unterteilt.

Die Beiträge im Bereich der physikalischen Modellierung der Beweglichkeit sind folgende: Ein neuartiges *Sechs-Täler*-Effektive-Masse-Modell für Löcher wird aus der $\mathbf{k}\cdot\mathbf{p}$-Theorie hergeleitet, welches die Einflüsse von Confinement und mechanischer Verspannung auf die Valenzbandstruktur zu erfassen erlaubt. Ein neuer Zugang zur Modellierung von Ladungsträger-Streuung durch polar-optische Phononen wird entwickelt, in welchem die elektrostatische Green'sche Funktion verwendet wird, um die Wechselwirkung zwischen Ladungsträgern und oszillierenden Dipolen zu modellieren, wodurch die Geometrie des Kanals und die Variation von Materialeigenschaften berücksichtigt werden. Ein neues Modell für Streuung durch Oberflächen- und Grenzflächen-Rauigkeit wird entwickelt; es stellt die Verallgemeinerung des Modells für Oberflächen-Rauigkeit von Prange und Nee auf nicht-planare Kanäle beliebiger Form dar, wodurch eine Konsistente Modellierung von Streuung durch Rauigkeit in planaren und nicht-planaren Strukturen, wie Nanowires und FinFETs, möglich wird. Zwei neue Ansätze für Modellierung von Beweglichkeit basierend auf der *linearisierten Boltzmann-Transport-Gleichung* werden besprochen, mit dem Ziel die Anisotropie von Bändern und die Anisotropie von Streuprozessen konsistent zu behandeln. Ein Ansatz berechnet die Antwort erster Ordnung der Verteilungsfunktion, während im anderen die Antwort durch einen gekoppelten mikroskopischen Relaxationszeittensor substituiert wird.

Die Beiträge im Bereich der rechnerischen Methoden umfassen folgende Innovationen: Eine Finite-Volumen-Diskretisierungsmethode wird entwickelt, welche nicht kanten- sondern element-basiert ist, was die Diskretisierung von anisotropen Hamilton-Operatoren, auf Basis anisotroper effektiver Massen und $\mathbf{k}\cdot\mathbf{p}$-Theorie, unter Erhaltung deren physikalischer Eigenschaften ermöglicht. Eine neue numerische Methode wird entwickelt, welche eine erschöpfende Suche nach Eigenwerten in einem Intervall für große, schwach besetzte

Systeme bei vernachlässigbarem zusätzlichen Rechenaufwand durchführt; diese Methode erlaubt es eine vordefinierte Fehlertoleranz einzuhalten, wenn die Subbänder eines Kanals sowie deren Eigenschaften bestimmt werden. Eine Methode zur Diskretisierung des Streuoperators der linearisierten Boltzmann-Transportgleichung im **k**-Raum basierend auf *symbolischer Konturintegration* wird ebenso vorgestellt.

Schließlich werden die entwickelten Methoden angewendet, um die Beweglichkeiten und Leitfähigkeiten von existierenden Bauelementen sowie von Bauelement-Konzepten, die potenzielle Kandidaten für zukünftige Technologie-Knoten sind, zu berechnen. NMOS- und PMOS-Transistoren des 22 nm-Technologie-Knotens, welche 2012 veröffentlicht wurden, werden untersucht und deren Kanal-Beweglichkeit unter Variation der Kristallorientierung und Anwendung von mechanischer Verspannung bestimmt. Es stellt sich heraus, dass eine Änderung der Orientierungen sowohl des Subtrats als auch der Finne die Kanal-Beweglichkeit deutlich verbessern können.

Bezüglich zukünftiger Technologie-Knoten werden zwei Bauelemente zur näheren Untersuchung ausgewählt: ein MISFET mit einem InGaAs-Kanal sowie ein übergangsloser (*junction-less*) Si FET. Die Eigenschaften des InGaAs-Bauelements werden im Detail modelliert und eine exzellente Übereinstimmung mit Messergebnissen wird erreicht, was die Genauigkeit der in dieser Arbeit vorgestellten Modelle unterstreicht. Der übergangslose Transistor wird mit einem inversionsbasierten verglichen. Es wird gezeigt, dass während Streuung durch Rauigkeit im übergangslosen Bauelement stark unterdrückt wird, die daraus resultierende Verbesserung der Beweglichkeit nicht ausreicht, um deren gleichzeitige Verschlechterung durch verstärkte Störstellen-Streuung zu kompensieren.

Contents

1 Introduction **1**
- 1.1 The "Evolution" of the Field-Effect Transistor . . . 1
- 1.2 Principles of Transistor Operation . . . 3
- 1.3 Modeling and Simulation of Novel Devices . . . 5
- 1.4 Outline . . . 7

2 Physics of Transport Modeling **8**
- 2.1 Electronic Structure . . . 9
  - 2.1.1 The $\mathbf{k}\cdot\mathbf{p}$ Model . . . 11
  - 2.1.2 The Effective Mass Model . . . 20
  - 2.1.3 Electronic Structure of Confined Systems . . . 23
- 2.2 Carrier Density and Electrostatics . . . 26
- 2.3 Scattering Processes . . . 28
  - 2.3.1 Non-Polar Lattice Scattering . . . 31
  - 2.3.2 Coulomb Scattering . . . 35
  - 2.3.3 Polar-Optical Phonon Scattering . . . 38
  - 2.3.4 Surface and Interface Roughness Scattering . . . 42
  - 2.3.5 Alloy Disorder Scattering . . . 46
- 2.4 Transport and Mobility . . . 47
  - 2.4.1 Linearizing the Boltzmann Transport Equation . . . 47
  - 2.4.2 Conductivity and Mobility Extraction . . . 49

3 Computational Foundation **54**
- 3.1 Model Concept . . . 55
- 3.2 Data Level . . . 56
  - 3.2.1 Geometry and Topology . . . 56
  - 3.2.2 Data Storage . . . 57
  - 3.2.3 Discretization . . . 57
  - 3.2.4 I/O and Configuration . . . 61
- 3.3 Modeling Level . . . 61
  - 3.3.1 Expressions . . . 61

3.3.2 Problem Specification and Assembly . . . 63
3.3.3 Contour Integration . . . 66
3.4 Algebraic Level . . . 67
3.4.1 Abstraction of Linear Operations . . . 67
3.4.2 Working with Expressions . . . 67
3.4.3 Solvers . . . 68
3.4.4 Fast Fourier Transform . . . 70
3.5 Extension Through Modules . . . 70
3.5.1 Module Loading . . . 70
3.5.2 Software Development Kit . . . 71
3.5.3 Literate Modeling . . . 71

4 Model Implementation **73**
4.1 Basic Models . . . 74
4.1.1 Poisson . . . 74
4.1.2 Self-Consistent Loop . . . 75
4.1.3 Strain . . . 77
4.2 Model Chains . . . 77
4.3 Carrier Models . . . 78
4.3.1 Classic 3D Carrier Gas with Parabolic Band Structure . . . 78
4.3.2 Confined Carrier Gas with Parabolic Band Structure . . . 79
4.3.3 Confined Carrier Gas with Non-Parabolic ($\mathbf{k}\cdot\mathbf{p}$) Band Structure . . . 81
4.4 Mobility Models . . . 84
4.4.1 Mobility Calculation for Parabolic Bands . . . 84
4.4.2 Mobility Calculation for Non-Parabolic Bands . . . 85
4.5 Scattering Models . . . 88
4.5.1 Scattering Model Interface . . . 88
4.5.2 Coulomb Scattering Template . . . 89
4.5.3 Non-Polar Phonon Scattering . . . 90
4.5.4 Alloy Disorder Scattering . . . 90
4.5.5 Ionized Impurity Scattering . . . 90
4.5.6 Polar-Optical-Phonon Scattering . . . 91
4.5.7 Surface and Interface Roughness Scattering . . . 91

5 Results **93**
5.1 Inversion-mode Channels . . . 93
5.1.1 22 nm n-Type Silicon FinFET . . . 94
5.1.2 22 nm p-Type Silicon FinFET . . . 99
5.1.3 InGaAs-Based Devices . . . 101
5.2 Junction-Less Channels . . . 107

6 Conclusion and Outlook **110**
6.1 Conclusion . . . 110

6.2 Outlook 110
6.2.1 Boltzmann Transport Equation in Phase Space 111
6.2.2 Quantum Transport 111

Bibliography **113**

CHAPTER 1 Introduction

Contents

1.1 The "Evolution" of the Field-Effect Transistor 1
1.2 Principles of Transistor Operation 3
1.3 Modeling and Simulation of Novel Devices 5
1.4 Outline 7

1.1 The "Evolution" of the Field-Effect Transistor

Since the invention of the field-effect transistor (FET), solid-state electronics are steadily getting cheaper, faster, and more ubiquitous. There is no industry with a growth quite like that observed in the semiconductor industry, making electronics devices cheaper and at the same time more functional every year. This trend was observed by Gordon E. Moore who formulated an economic law named after him: The law states that the number of transistors per chip is likely to double roughly every two years [1]. Throughout the second half of the 20^{th} century, this increase was achieved by *scaling*: The mere reduction of the device dimensions – from hundreds of micrometers to hundreds of nanometers – allowed to increase the density of transistors bringing a reduction of fabrication cost. As a welcome side-effect, scaling also improved transistor performance, making the devices faster and less power-consuming.

At the beginning of the 21^{st} century it became apparent that scaling alone would not suffice to keep Moore's Law on track. As the transistor channels became shorter the control the transistor's gate could exert over the channel slipped away, as illustrated in Fig. 1.1. This implied that the ratio between on-current and off-current required for practical circuits could not be maintained for channel lengths below 100 nm. What happened since then appears as some sort of evolutionary process when viewed from the outside, with a series of innovations introduced step-by-step every two technology nodes as shown in Fig. 1.2.

The introduction of strained silicon marked the first innovation step which aimed at the on-current. Strain increases the carrier mobility in silicon resulting in higher

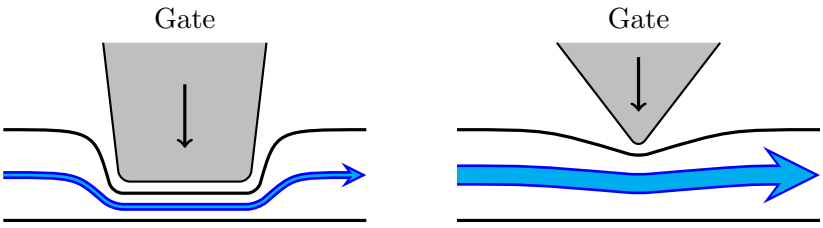


Figure 1.1: "The Garden Hose Analogy:" As scaling makes the gate smaller, it gets less effective in controlling the channel. The device becomes harder to turn off.

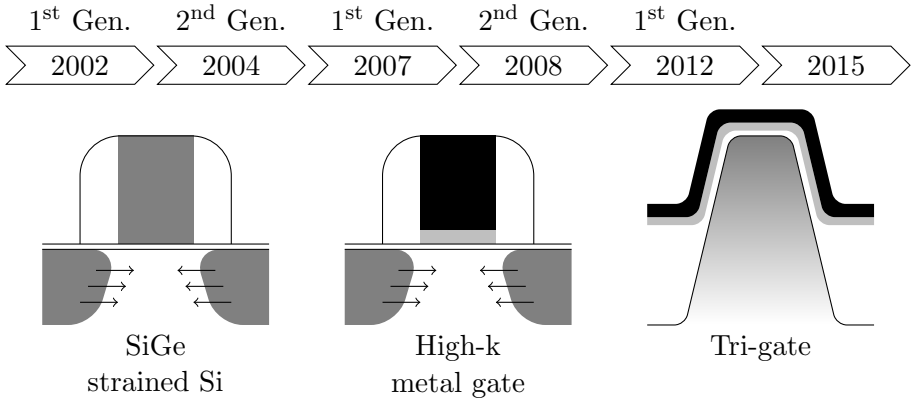
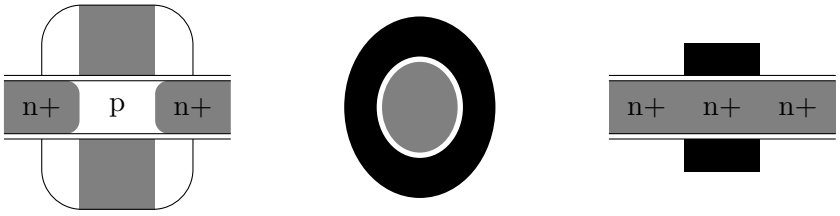


Figure 1.2: The last ten years of CMOS evolution featured the introduction of strain engineering in the MOSFET channel [2, 3], the usage of high-k dielectrics and metal gates [4, 5], and eventually the advent of the tri-gate transistor [6].

on-current and, hence, in a higher on/off-current ratio. The next step was to replace the SiO_2-dielectric by a material with higher permittivity combined with a metal gate. This innovation restored the gate's capability of controlling the channel, thus allowing further scaling, but only for two more technology nodes. With the possibilities of strain and material engineering seemingly exhausted, the next step to go was to alter the geometry of the transistor. Rather than having current flow in a sheet beneath the gate, it is directed through a fin surrounded by the gate on three sides: the tri-gate FET or FinFET was introduced.

Today, the FinFET is the state-of-the art in semiconductor device fabrication. The future of the FET evolution is unclear, although on thing is certain: To continue the scaling, device engineers will have to use every trick they have at their disposal. These are not limited to but will likely include dopant engineering, altering the channel geometry, applying different crystal orientations, strain engineering, and adding new, previously unused materials to the process.

It is important to mention here, that the FinFET is far from being the only option for the continuation of device scaling. Over the years a number of transistor concepts was



Double gate UTB | Gate-all-around (GAA) FET | Junction-Less FET

Figure 1.3: Over the years a number of alternative technologies have been proposed to replace the planar MOSFET. The double-gate ultra-thin-body (DG UTB) [7–11] and the Gate-all-around (GAA) [12] attempt to improve electrostatic control of the gate by reducing the transistor body to a thin film or a nanowire, respectively. The junction-less transistor [13] takes a radical step further: a highly conductive silicon body is actively pinched off by the surrounding gates, thus making it a normally-on device, as opposed to traditional normally-off FETs.

proposed and implemented. Silicon-on-insulator (SOI) based technologies are considered a potential alternative to the mainstream bulk technology. Technologies like SOI and SON (silicon-on-nothing) focus on improving the electrostatic control by the gate by thinning the silicon body, a principle also employed in the FinFET. Some of the promising SOI technologies are shown in Fig. 1.3

1.2 Principles of Transistor Operation

Every transistor works based on the same basic principle - electron (or hole) transport over an energy barrier. The *height* of the barrier controls the current that can pass through the device and the barrier itself can be controlled either by voltage or current. The idea for such a device is much older than microelectronics. Patents for a device which one nowadays would call a transistor were granted in the 1930's to Julius Edgar Lilienfeld where a "Device for controlling electric current" is described [14, 15]. At that time semiconductors were poorly understood. However, this understanding was necessary to build a working transistor. Semiconductors have an energy gap or forbidden band in their electronic structure, which is what the the energy barrier consists of in every transistor.

Figure 1.4 shows the working principle of a transistor - in particular a field-effect transistor (FET). Other transistor types exist but the main difference between them is the way the barrier is controlled. For a *metal-oxide-semiconductor* FET (MOSFET) it is controlled through a capacitor, for *junction* FET (JFET) or *metal-semiconductor* FET (MESFET) through a reverse-biased (Schottky) diode, and for a bipolar transistor through a forward-biased diode.

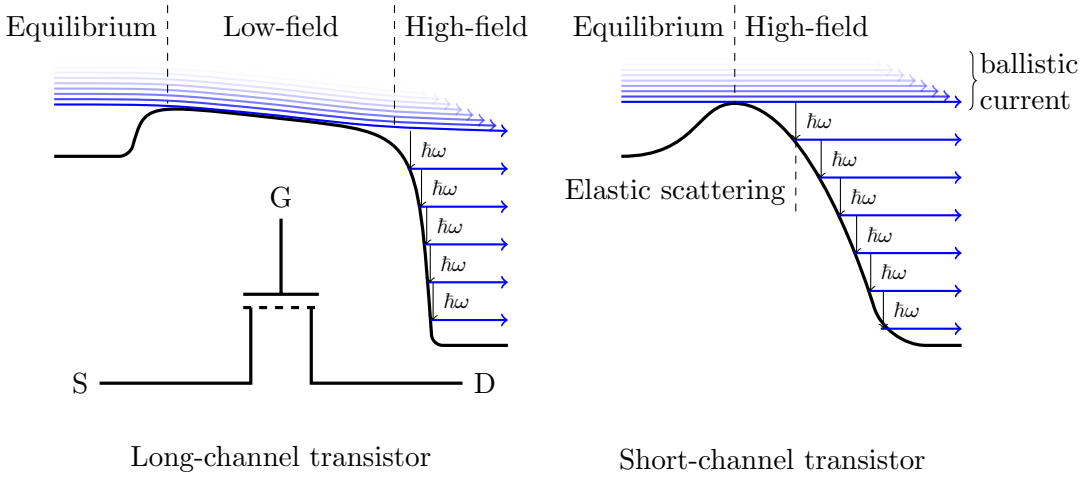


Figure 1.4: The working principle of a transistor: To travel from source (S) to drain (D) the carriers must surmount a barrier, the height of which is controlled electrostatically by the gate terminal (G). Only carriers with an energy above the top of the barrier [16] can pass the barrier. In a long-channel transistor carriers move through an extended low-field region before reaching the drain. In the low-field region scattering is mostly elastic, only small amounts of energy are lost to acoustic lattice vibrations. In the high field region before the drain, the carriers are accelerated to high kinetic energies which are then relaxed by emitting quantized lattice vibrations of energy $\hbar\omega$ into the drain.

Controlling the current by voltage (or current) is the necessary basis for electronic circuit elements such as amplifiers, current sources and current or voltage-controlled switches. For digital applications it is crucial to have a high-quality switch. An ideal switch would be infinitely fast, consume no power, and have zero resistance in the on-state. The MOSFET emerged as the dominant device for digital applications. Apart from its low fabrication cost, the MOSFET has the lowest gate leakage, allowing switches that consume little power. The maximum current that passes over the barrier is diminished by scattering. Scattering causes electrons and holes to dissipate both momentum and energy and thus contributes to the electrical resistance in the on-state.

The amount of scattering carriers undergo in a device is characterized by a quantity called mobility. Although in modern, nanometer-sized transistors lumping all transport-related phenomena into a single quantity does not accurately reflect physics of transport, the mobility is still an important figure of merit in device engineering.

1.3 Modeling and Simulation of Novel Devices

Modeling and simulation is instrumental in the development of semiconductor technology. There are two main areas at which simulation is used in semiconductor technology, (i) *process simulation*, which deals with the simulation of fabrication processes, such as ion implantation, layer deposition, oxidation, thermal annealing, etc., and (ii) *device simulation* which aims to predict the characteristics of a device, before it is manufactured. Usually, process and device simulation are used in combination: A process is simulated to obtain a device structure, which is then fed into the device simulator. Both areas comprise what is known as *technology computer-assisted design* (TCAD).

For several decades, traditional device simulation was based on the drift-diffusion equations for modeling carrier transport. The crucial parameter therein is the aforementioned carrier mobility. The entire physics of carrier transport is condensed in this one quantity which is a function of temperature, doping concentration, the electric field, strain, and geometry. In a top-down approach the mobility function is fitted to reproduce measured device characteristics. One ends up with an analytical expression for mobility with an ample number of parameters, which need to be adjusted to fit the measured results.

As devices get smaller, empirical mobility models lose validity. What happens, is that one set of parameters, fitted to one device design will fail to reproduce the properties of a slightly altered design. In other words: At small scales, empirical mobility models lose their universality. Ultimately, this strongly diminishes the value purely empirical models for ultra-scaled devices and novel architectures such as the FinFET.

In a bottom-up approach the mobility function is derived from the underlying physics of carrier transport. This adds significantly to the complexity of the model, since additional equations need to be solved (numerically) such as the Schrödinger equation and the Boltzmann transport equation. However, the parameters of these equations are more directly related to transport physics and the materials of which the device is made. These are a handful of well-known, measurable material property constants and more universal than empirical model parameters. Physically-grounded mobility models are still valid at

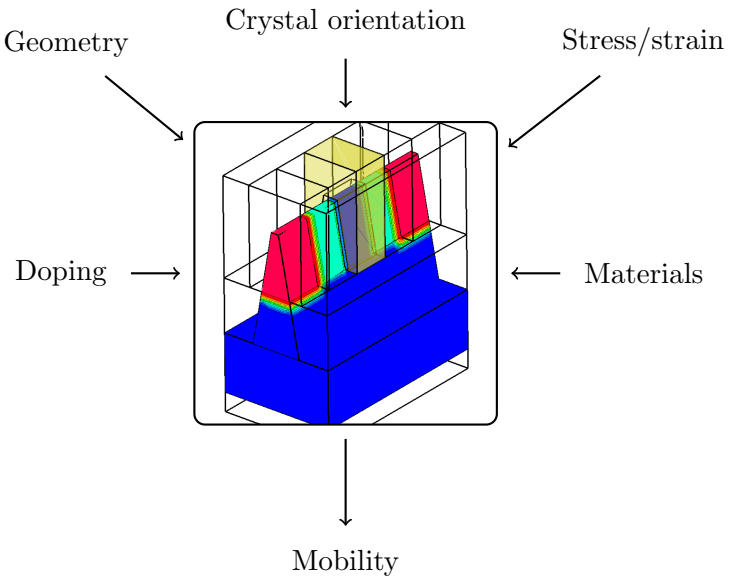


Figure 1.5: The goal of the work is to provide a computational framework capable of computing the electronic structure and mobility of nano-scaled transistor channels. The framework should take into account the effects of geometry, material composition, crystal orientation, doping, and mechanical stress. All these properties can be varied in the process of designing a semiconductor device.

the nano-scale and their parameters are portable between similar technologies.

The aim of this work is to develop a unified computational framework for the purpose of evaluating mobility in a wide range of transistor channel types. Device geometry, material composition, doping profile, crystal orientation, and stress distribution need to be taken into account in order to achieve the desired unification shown in Fig. 1.5. A central requirement is that the methods developed are TCAD-compatible, i.e. robust and efficient enough to be used in device design and engineering.

1.4 Outline

This thesis is structured as follows: **Chapter 2** deals with the physics of carrier transport in nanoscale devices and lays out the models that are necessary for physical mobility modeling, such as electronic structure, carrier density and electrostatics, scattering processes, transport, and methods for calculating mobility itself. **Chapter 3** is a general overview of computational methods used to tackle the problems and equations from the physics chapter. **Chapter 4** discusses the particularities of implementing the models from the physics chapter using the methods presented in the computational methods chapter. Finally, the implemented models are put to use and some case studies are presented along with results in **Chapter 5**.

CHAPTER 2 Physics of Transport Modeling

Contents

2.1 Electronic Structure 9
2.1.1 The **k**·**p** Model 11
2.1.2 The Effective Mass Model 20
2.1.3 Electronic Structure of Confined Systems 23
2.2 Carrier Density and Electrostatics 26
2.3 Scattering Processes 28
2.3.1 Non-Polar Lattice Scattering 31
2.3.2 Coulomb Scattering 35
2.3.3 Polar-Optical Phonon Scattering 38
2.3.4 Surface and Interface Roughness Scattering 42
2.3.5 Alloy Disorder Scattering 46
2.4 Transport and Mobility 47
2.4.1 Linearizing the Boltzmann Transport Equation 47
2.4.2 Conductivity and Mobility Extraction 49

As stated in the introduction, the goal is to model mobility in general and low-field mobility in particular taking into account the following physical properties of the transistor channel: geometry, material composition, doping, crystal orientation, and mechanical stress. Each of these can be viewed as a degree of freedom when thinking up a novel device design. Therefore, it is desirable to have a model framework that is able to predict the qualities of a transistor channel based on the input of these properties. Being a widely used property in engineering, the channel low-field mobility serves as a metric to rate different device realizations. Even though its particular value can only be observed in long-channel devices, it serves as a figure of merit even for short-channel devices.

However, none of the above properties relates to transport or low-field mobility *directly*. Instead the influence is mediated through electronic structure and scattering processes,

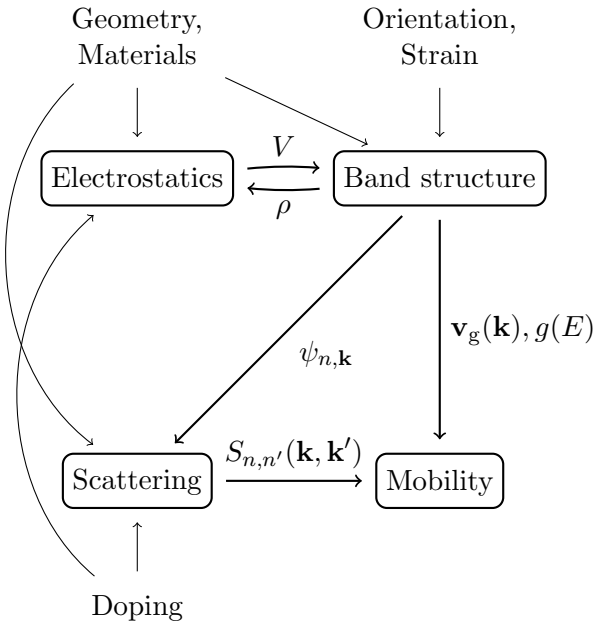


Figure 2.1: The device parameters influence the channel mobility only indirectly via the band structure and the carrier scattering. The interdependencies between the physical models are shown: Electrostatics and electronic structure affect each other via potential V and space charge density ρ, electronic structure affects mobility directly through group velocity $\mathbf{v}_\mathrm{g}(\mathbf{k})$ and density of states $g(E)$, and indirectly through the transition rates $S_{n,n'}(\mathbf{k},\mathbf{k}')$ which themselves depend on the electronic wave functions.

as illustrated in Fig. 2.1. For instance, crystal orientation or strain affect primarily the electronic structure which determines quantities such as density of states or group velocity. These derived quantities are those which affect transport and, eventually, the mobility.

In order to obtain a reliable, predictive low-field mobility figure, each of the models must deal with the information available to it in a versatile and accurate fashion. The network of interdependencies between the models will serve as a guide through the remainder of this chapter: Starting with the electronic structure in Section 2.1 and its coupling with electrostatics in Section 2.2, we will continue to scattering processes in Section 2.3 before combining both with semi-classical transport in Section 2.4, where low-field mobility will be derived.

2.1 Electronic Structure

The electronic structure or band structure is the dispersion relation of electrons in a crystal. Its significance comes from the fact that due to the periodic structure of a

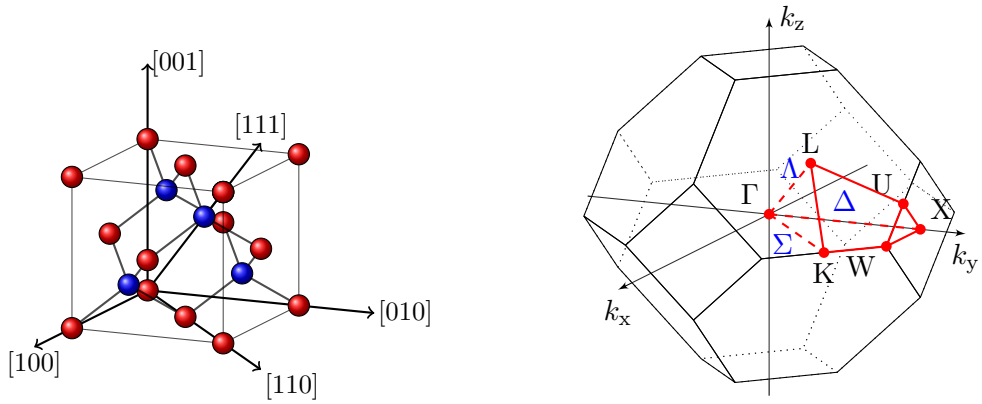


Figure 2.2: Left: the unit cell of a diamond and zinc-blende lattice; right: the corresponding Brillouin zone in reciprocal space. The point of highest symmetry is Γ in the center of the zone ($\mathbf{k} = \mathbf{0}$), the edges Δ, Σ, and Λ correspond to the crystal axes ⟨100⟩, ⟨110⟩, and ⟨111⟩, respectively.

crystal, the potential of the crystal affecting electrons is also periodic. Thus, rather than having electrons scatter off the ionized atomic cores of the crystal, the periodic potential alters the kinetic behavior of electrons. Electrons traverse the crystal in energy bands, unperturbed. Scattering within a band occurs, if the crystal potential deviates from the perfectly periodic form, e.g. by displacement of atoms due to lattice vibrations or by crystal defects that break the periodicity.

Semiconductors are known to be materials the energy bands of which are interrupted by a band gap, i.e. an energy range where no propagating states exist. The energy bands below the gap are referred to as *valence bands* and the bands above as *conduction bands*, where the boundaries between each of the bands and the bandgap are called the *band edges*.

The electrical properties of a semiconductor are related to the band structure. More specifically it was found, that most properties of semiconductors can be derived from the effects occurring near the band edges [17]. Current, for instance, flows only in states near the band edges because statistically the number of electrons available for transport decreases roughly exponentially with energy above the conduction band edge. Similarly, current in the valence band is due to moving electron vacancies or holes the number of which also decreases exponentially for low energies. Consequently, for not too high electric fields, current flow in semiconductors can be explained and analyzed by looking at the band structure near the band edges.

Fig. 2.2 shows the crystal lattice of diamond or zinc-blende type along with its first Brillouin zone (BZ), also known as Wigner-Seitz cell, which reflects the lattice symmetries in Fourier space or *reciprocal* space. Since in quantum mechanics, position and momentum are linked by the Fourier transform, an electronic state can be equivalently represented in real space and reciprocal space.

The diamond/zinc-blende crystal is not only periodic, but also has certain symmetries.

The lattice is said to be *invariant* with respect to a number of geometrical transformations. One such transformation would be mirroring the crystal along the [100]-axis while rotating it around the same axis by 90 degrees. Performing such a transformation will leave us with the same lattice structure as before the transformation. The set of such transformations the lattice is invariant to is called a *point group*, or a *space group*, if translations are also allowed. The symmetry groups of diamond and zinc-blende are denoted as O_h^7 and T_d^2, respectively [18].

The crystal symmetry is reflected in reciprocal space as well. The Brillouin zone is thus also invariant under a number of transformations, such as mirroring, rotation, and translation. The consequence of this is that only a small section of the Brillouin zone, called the *irreducible wedge*, contains the entire band structure information. The irreducible wedge is highlighted red in Fig. 2.2. The band structure in the remaining zone can be obtained by rotating or mirroring the irreducible wedge. The edges and corners of the irreducible wedge are each denoted by specific letters: Γ, X, L, W, K, and U for the points and Δ, Λ, and Σ for the edges. Due to symmetry, band extrema, i.e. the aforementioned band edges, are always found along these edges, mostly in one of the points.

2.1.1 The k·p Model

A number of models for the band structure of semiconductors exists. The three most widely used models are, the (semi-empirical) tight-binding model (TB), the empirical pseudopotential model (EPM), and the **k·p** model. The tight-binding model on the one hand arrives at the band structure by assuming electrons to be quasi-bound by the potentials of the atoms in the crystal. It uses atomic orbitals as basis-functions to expand the coupling between the bound states at adjacent atomic sites. The empirical pseudopotential method on the other hand assumes electrons to be quasi-free as in a vacuum. It expands the potential of the ionized atoms and core electrons in terms of plane-wave basis-functions. Ab-initio methods, such as the density functional theory (DFT), can be used for band structure calculation, and would constitute a fourth class of electronic structure models. However, DFT is a much broader theoretical framework widely used in quantum chemistry.

The **k·p** model differs from EPM and TB in the sense that the basis-functions are chosen based on a different kind of assumption. The assumption is that the Bloch-functions are known at a certain point in **k**-space, commonly a point of high symmetry. Let's assume that the Schrödinger equation in the crystal consists of a kinetic energy operator and a periodic crystal potential,

$$H\left|\psi\right\rangle = \left[\frac{p^2}{2m_e} + V\right]\left|\psi\right\rangle = E\left|\psi\right\rangle, \tag{2.1}$$

where $\mathbf{p}$ signifies the momentum operator $-\mathrm{i}\hbar\boldsymbol{\nabla}$ and m_e the electron rest mass. According to the Bloch-theorem, the wave-function can be separated into a product of plane wave and the *Bloch-function*,

$$\left|\psi\right\rangle = e^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\left|n,\mathbf{k}\right\rangle, \tag{2.2}$$

where n and $\mathbf{k}$ represent band index and wave vector, respectively. The Bloch-function $u_{n,\mathbf{k}}(\mathbf{r}) = \langle \mathbf{r}|n,\mathbf{k}\rangle$ has the same periodicity as the crystal lattice. Inserting Eq. (2.2) into Eq. (2.1), results in an effective Schrödinger equation for the Bloch-function,

$$\left[\frac{p^2}{2m_e} + V + \frac{\hbar\mathbf{k}\cdot\mathbf{p}}{m_e} + \frac{\hbar^2k^2}{2m_e}\right]|n,\mathbf{k}\rangle = E_{n,\mathbf{k}}\,|n,\mathbf{k}\rangle\,. \tag{2.3}$$

The Bloch-function at $\mathbf{k} = \mathbf{0}$ would then satisfy

$$\left[\frac{p^2}{2m_e} + V\right]|n,\mathbf{0}\rangle = E_{n,\mathbf{0}}\,|n,\mathbf{0}\rangle\,. \tag{2.4}$$

Since the Bloch-functions at $\mathbf{k} = \mathbf{0}$ form a complete orthogonal basis set, the Bloch-functions away from $\mathbf{k} = \mathbf{0}$ can be expanded in terms of $|n,\mathbf{0}\rangle$,

$$|n,\mathbf{k}\rangle = \sum_m c_{n,\mathbf{k}}\,|n,\mathbf{0}\rangle\,. \tag{2.5}$$

Inserting the expansion into Eq. (2.3) and multiplying it from left by $\langle n,\mathbf{0}|$, produces an eigenvalue problem of the coefficients $c_{n,\mathbf{k}}$,

$$E_{n,\mathbf{k}}c_{n,\mathbf{k}} = \left[E_{n,\mathbf{0}} + \frac{\hbar^2k^2}{2m_e}\right]c_{n,\mathbf{k}} + \frac{\hbar}{m_e}\mathbf{k}\cdot\sum_m \langle n,\mathbf{0}|\,\mathbf{p}\,|m,\mathbf{0}\rangle\,c_{n,\mathbf{k}}. \tag{2.6}$$

This system of equations represents an eigenvalue problem for the coefficients $c_{n,\mathbf{k}}$. which are also called *envelope functions*. Taking all the energy bands up to vacuum level into account would in theory give a very precise band structure description. However, a high number of matrix elements $\langle n,\mathbf{0}|\,\mathbf{p}\,|m,\mathbf{0}\rangle$ would need to be determined in that case. Instead, the equation system in Eq. (2.6) is reduced to contain only a few bands, i.e. the ones closest to a band edge. The remaining *remote* bands are treated as a perturbation.

To get an estimate on the influence of such remote bands, we can look a the case of a single band. Here, only one band is retained in Eq. (2.6) and all other bands are considered remote. It that case, perturbation theory would give

$$E_{n,\mathbf{k}} = E_{n,\mathbf{0}} + \frac{\hbar^2k^2}{2m_e} + \frac{\hbar}{m_e}\mathbf{k}\cdot\langle n,\mathbf{0}|\,\mathbf{p}\,|n,\mathbf{0}\rangle + \frac{\hbar^2}{m_e^2}\sum_{m\neq n}\frac{|\mathbf{k}\cdot\langle n,\mathbf{0}|\,\mathbf{p}\,|m,\mathbf{0}\rangle\,|^2}{E_{n,\mathbf{0}} - E_{m,\mathbf{0}}}. \tag{2.7}$$

The denominator $E_{n,\mathbf{0}} - E_{m,\mathbf{0}}$ causes the influence of the remote bands to diminish the larger the energy spacing is between them and the bands of interest.

The Dresselhaus-Kip-Kittel Model for Holes

In diamond and zinc-blende crystals, the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian for the valence band is due to Dresselhaus, Kip, and Kittel [19]. It forms a second order expansion of the band structure at the zone center, i.e. the Γ-point in $\mathbf{k}$-space. Semiconductors with diamond lattice, such as Si and Ge, have two atoms per unit cell, each with four electrons in their outer shells; thus each unit cell contributes eight valence electrons resulting in four two-fold

spin-degenerate bands. The first s-type band is far below the valence band edge and only included as perturbation. The remaining three bands are superposition of p-type states and form the valence band edge at $\mathbf{k} = 0$. They are called the heavy hole and light hole and split-off band. When spin-orbit coupling is neglected, two heavy hole bands and a light hole band, each two-fold spin-degenerate, meet at the Γ point forming a six-fold degeneracy. Omitting the spin-degeneracy, the valence band can be described by a thee-band effective Hamiltonian, written as a 3×3 matrix,

$$\mathbf{H}_{3\times3} = \frac{\hbar^2}{2m_e}\begin{bmatrix} Lk_x^2 + M(k_y^2+k_z^2) & Nk_xk_y & Nk_xk_z \\ Nk_xk_y & Lk_y^2 + M(k_x^2+k_z^2) & Nk_yk_z \\ Nk_xk_z & Nk_yk_z & Lk_z^2 + M(k_x^2+k_y^2) \end{bmatrix} + \begin{bmatrix} l\varepsilon_{xx} + m(\varepsilon_{yy}+\varepsilon_{zz}) & n\varepsilon_{xy} & n\varepsilon_{xz} \\ n\varepsilon_{xy} & l\varepsilon_{yy} + m(\varepsilon_{xx}+\varepsilon_{zz}) & n\varepsilon_{yz} \\ n\varepsilon_{xz} & n\varepsilon_{yz} & l\varepsilon_{zz} + m(\varepsilon_{xx}+\varepsilon_{yy}) \end{bmatrix}. \quad (2.8)$$

L, M, and N are the band curvature parameters and l, m, and n the deformation potentials for the various strain components $\varepsilon_{\xi\eta}$. Other notations for the L, M, and N parameters are common, such as the A, B, C parameters [19],

$$\begin{aligned} L &= A + 2B & A &= \tfrac{1}{3}(L+2M) \\ M &= A - B & B &= \tfrac{1}{3}(L-M) \\ N &= -\sqrt{3C^2+9B^2} & C^2 &= \tfrac{1}{3}[N^2-(L-M)^2], \end{aligned} \quad (2.9)$$

or the Luttinger Parameters γ_1, γ_2, and γ_3 [20],

$$\begin{aligned} L &= -(\gamma_1+4\gamma_2) & \gamma_1 &= -\tfrac{1}{3}(L+2M) \\ M &= -(\gamma_1-2\gamma_2) & \gamma_2 &= -\tfrac{1}{6}(L-M) \\ N &= -6\gamma_3 & \gamma_3 &= -\tfrac{1}{6}N. \end{aligned} \quad (2.10)$$

The deformation potentials can be equivalently expressed in terms of strain parameters a_v, b, and d [17],

$$\begin{aligned} l &= -(a_v+2b) & a_\mathrm{v} &= -\tfrac{1}{3}(l+2m) \\ m &= -(a_v-b) & b &= -\tfrac{1}{3}(l-m) \\ n &= -\sqrt{3}d & d &= -\tfrac{1}{\sqrt{3}}n. \end{aligned} \quad (2.11)$$

Spin-orbit coupling lifts the degeneracies of the valence bands. At the Γ-point it splits the the six-fold degeneracy resulting in a four-fold degenerate band edge and a two-fold degenerate split-off band. To describe spin-orbit coupling, the three-band Hamiltonian need to be extended to a six-dimensional basis, using the 3×3 Hamiltonian for each spin

state and adding a Hamiltonian describing the spin-orbit interaction,

$$\mathbf{H}_{6\times6} = \begin{bmatrix} \mathbf{H}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{H}_{3\times3} \end{bmatrix} - \frac{E_{\mathrm{so}}}{3} \begin{bmatrix} 0 & \mathrm{i} & 0 & 0 & 0 & -1 \\ -\mathrm{i} & 0 & 0 & 0 & 0 & \mathrm{i} \\ 0 & 0 & 0 & 1 & -\mathrm{i} & 0 \\ 0 & 0 & 1 & 0 & -\mathrm{i} & 0 \\ 0 & 0 & \mathrm{i} & \mathrm{i} & 0 & 0 \\ -\mathrm{i} & -\mathrm{i} & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2.12}$$

where E_{so} is the energy difference between the valence band edge and the split-off band at Γ [21].

Valence Band Parameters for Si, Ge, and SiGe

The Dresselhaus-Kip-Kittel (DKK) model can be used to model the valence band structure of the most widely used semiconductors, such as Ge, Ge, and SiGe alloys, which have a diamond crystal lattice. The description is also valid for semiconductors with zinc-blende type lattices, such as GaAs. However, since these semiconductors have a direct band gap the DKK model is extended to also include the conduction band valley at Γ. Such models will be discussed in Section 2.1.1.

To use the DKK for modeling the valence band structure of a particular semiconductor, actual figures are needed for the band parameters L, M, N, spin-orbit splitting E_{so}, and strain parameters l, m, n. One way to determine them is through measurement: cyclotron resonance measurements, as done by the authors of the DKK model, can provide values for the band parameters L, M, and N. Alternatively, the values can be obtained from theoretical calculations based on other models, such as non-local EPM [22], or first principles calculations. Table 2.1 gives a list of parameters for the DKK model for Si and Ge at room temperature.

To obtain the parameters for a SiGe alloy as functions of Ge content x, some kind of interpolation needs to be applied. It has to reproduce the values of pure Si and Ge for $x = 0$ and $x = 1$, respectively. A common way to interpolate alloy properties is using a quadratic function of x. For instance,

$$a_0(x) = a_0(\mathrm{Si})(1-x) + a_0(\mathrm{Ge})x + \beta x(1-x) = 5.431\,\text{Å} + x0.2\,\text{Å} + x^2 0.027\,\text{Å} \tag{2.13}$$

gives the average lattice constant for SiGe, where β is called a bowing parameter [23]. For $\beta = 0$ a linear interpolation is obtained. This works well if the values of both materials are similar. For SiGe, however, the L and N parameters in Ge are several times larger than their counterparts in Si and a linear or quadratic interpolation will not suffice.

A method for the interpolation of SiGe band parameters was proposed in [24]. The method is based on the approach by Lawaetz [25], which relates the matrix elements used to derive the band parameters to lattice properties, such as lattice constant or ionicity. In [19] the parameters L, M, and N are expressed in terms of F, G, H_1, and H_2, which

are the k^2-perturbations of each remote band,

$$\begin{aligned} L &= F + 2G \\ M &= H_1 + H_2 \\ N &= F - G + H_1 - H_2. \end{aligned} \tag{2.14}$$

The strongest perturbations are due to the conduction bands states directly above the valence band edge,

$$\begin{aligned} F &= -\frac{E_\mathrm{p}}{E_0} \\ H_1 &= -\frac{E_{\mathrm{p}'}}{E_{0'}}, \end{aligned} \tag{2.15}$$

where E_p and $E_{\mathrm{p}'}$ are the $\mathbf{k}\cdot\mathbf{p}$ matrix elements of the hole states and each of the two conduction bands, respectively. As Fig. 2.3 shows, the ordering of the two conduction bands is opposite for Si and Ge, and while $E_{0'}$ has a similar value in both materials, E_0 is lower in Ge by about 3 eV. Since F is inversely proportional to E_0, F is much larger in Ge than in Si and so are L and N. Experiments [26, 27] have shown that the dependence of the band energies at the Γ point follows a linear rule, so E_0 can be expressed as

$$E_0(x) = E_0(\mathrm{Si})(1-x) + E_0(\mathrm{Ge})x. \tag{2.16}$$

One problem, however, is that E_0(Si) is notoriously difficult to measure [28], and the values are only available at very low temperatures.

The matrix elements E_p and $E_{\mathrm{p}'}$, as well as those used in G and H_2 are related to lattice properties by the semi-empirical model of Lawaetz [25], which multiplies E_p, $E_{\mathrm{p}'}$, G, and H_2 by a scaling factor,

$$\delta(x) = [1 + \alpha(D(x) - 1)] \left(\frac{a_0(\mathrm{Si})}{a_0(x)} \right)^2, \tag{2.17}$$

where D is an enhancement factor accounting for d-orbital core states in Ge. In Si its value is 1.0 since there are no d-orbital states, in Ge the value was found to be 1.25 [29]. For SiGe the value is obtained by linear interpolation [24]. The computed relation between valence band parameters and alloy composition is shown in Fig. 2.4.

Kane Model

In direct-band semiconductors, such as GaAs, GaSb, InP, InAs, or InSb, but also indirect semiconductors with a low-lying Γ-valley, a model is required that couples conduction and valence bands. This coupling is responsible for the non-parabolicity of the conduction band in the Γ point. The four-band $\mathbf{k}\cdot\mathbf{p}$ model due to Kane [37] is structurally similar to the DKK model. The effective Hamiltonian reads

$$\mathbf{H}_{4\times4} = \frac{\hbar^2}{2m_e} \begin{bmatrix} A'k^2 + E_0 & Bk_yk_z + \mathrm{i}k_xP & Bk_xk_z + \mathrm{i}k_yP & Bk_xk_y + \mathrm{i}k_zP \\ Bk_yk_z - \mathrm{i}k_xP & L'k_x^2 + M(k_y^2 + k_z^2) & N'k_xk_y & N'k_xk_z \\ Bk_xk_z - \mathrm{i}k_yP & N'k_xk_y & L'k_y^2 + M(k_x^2 + k_z^2) & N'k_yk_z \\ Bk_xk_y - \mathrm{i}k_zP & N'k_xk_z & N'k_yk_z & L'k_z^2 + M(k_x^2 + k_y^2) \end{bmatrix}. \tag{2.18}$$

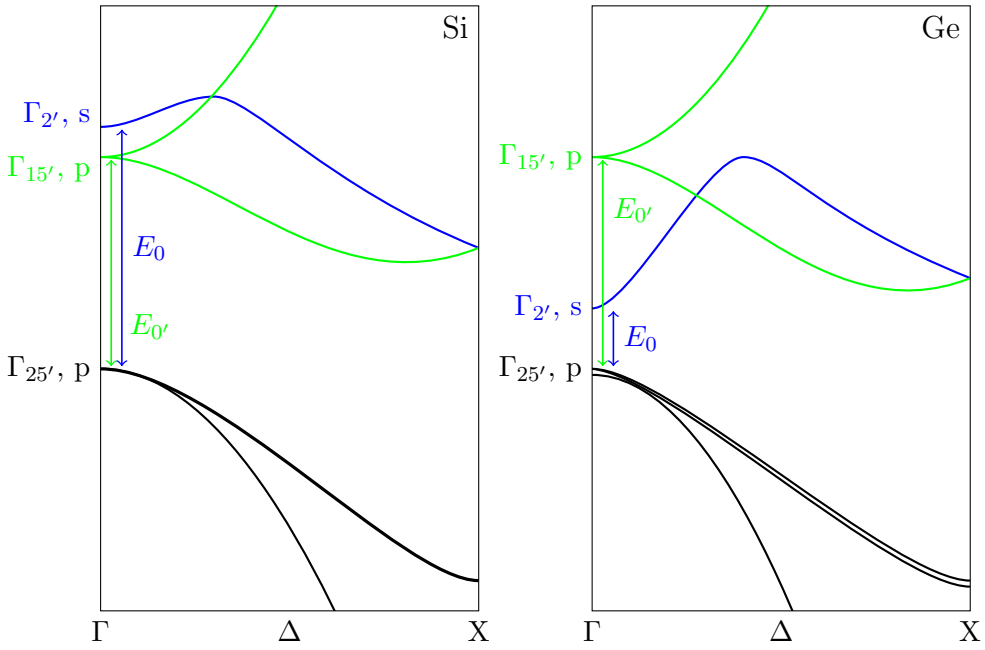


Figure 2.3: Qualitative plot of the band structure in Si and Ge along the Δ axis; the states at the valence band edge are similar to p-type atomic orbitals. The bands closest to the valence bands edge are the conduction bands at E_0, which correspond to an s-type orbital and $E_{0'}$ which is again of p-type. Both are considered *remote* bands in the **k·p** description of the valence band and are thus treated as second-order perturbation. The magnitude of the perturbation is inversely proportional to the transition energy. In Ge, E_0 is lowered significantly (by $\sim 3\,\mathrm{eV}$) as compared to Si, which is responsible for the non-linear dependence of the band parameters on the Ge mole fraction.

Table 2.1: List of band parameters the valence band of the pure Si and Ge, as well as SiGe alloys; listed are the band parameters $A/B/C$, the deformation potentials $a_\mathrm{v}/b/d$ [17], the spin-orbit splitting parameter E_so, and the direct transition energies at the Γ-point E_0 and $E_{0'}$. For pure Si and Ge, these parameters are taken from experiments and theoretical calculations referenced in the respective table header. The alternative notations $L/M/N$ and $l/m/n$ as well as the Luttinger parameters were computed using the formulas given in Eqs. (2.9) to (2.11). The $L/M/N$ parameters for SiGe alloys are interpolated as functions of the Ge content x. The Γ-point energies E_so, E_0, and $E_{0'}$ have an apparently linear dependence on x as demonstrated experimentally [26, 27]. The similarity of the deformation potential of Si and Ge indicate that a linear interpolation is adequate here as well. The band parameters, however, show large non-linear variations due to their inverse dependence on the direct transitions E_0 and $E_{0'}$.

	Unit	Si	Ge	SiGe
L	1	-5.53	-30.44	$-E_\mathrm{p}/E_0 + 2G$
M	1	-3.64	-4.73	$-E_{\mathrm{p}'}/E_{0'} + H_2$
N	1	-8.32	-33.93	$-E_\mathrm{p}/E_0 - G - E_{\mathrm{p}'}/E_{0'} - H_2$
A	1	-4.27	-13.3	$\frac{1}{3}[-E_\mathrm{p}/E_0 + 2G - 2E_{\mathrm{p}'}/E_{0'} + 2H_2]$
B	1	-0.63	-8.57	$\frac{1}{3}[-E_\mathrm{p}/E_0 + 2G + E_{\mathrm{p}'}/E_{0'} - H_2]$
$\lvert C\rvert$	1	4.93	12.78	$[\frac{1}{3}(2E_{\mathrm{p}'}/E_{0'} + 3G)(2E_\mathrm{p}/E_0 - G + 2H_2)]^{\frac{1}{2}}$
γ_1	1	4.27	13.3	$-\frac{1}{3}[-E_\mathrm{p}/E_0 + 2G - 2E_{\mathrm{p}'}/E_{0'} + 2H_2]$
γ_2	1	0.315	4.285	$-\frac{1}{6}[-E_\mathrm{p}/E_0 + 2G + E_{\mathrm{p}'}/E_{0'} - H_2]$
γ_3	1	1.387	5.655	$-\frac{1}{6}[-E_\mathrm{p}/E_0 - G - E_{\mathrm{p}'}/E_{0'} - H_2]$
l	eV	1.74	4.56	$l(\mathrm{Si})(1-x) + l(\mathrm{Ge})x$
m	eV	-4.56	-4.14	$m(\mathrm{Si})(1-x) + m(\mathrm{Ge})x$
n	eV	8.31	9.18	$n(\mathrm{Si})(1-x) + n(\mathrm{Ge})x$
a_v	eV	2.46	1.24	$a_\mathrm{v}(\mathrm{Si})(1-x) + a_\mathrm{v}(\mathrm{Ge})x$
b	eV	-2.1	-2.9	$b(\mathrm{Si})(1-x) + b(\mathrm{Ge})x$
d	eV	-4.8	-5.3	$d(\mathrm{Si})(1-x) + d(\mathrm{Ge})x$
E_so	eV	0.044	0.29	$E_\mathrm{so}(\mathrm{Si})(1-x) + E_\mathrm{so}(\mathrm{Ge})x$ [30]
E_0	eV	4.1	0.8	$E_0(\mathrm{Si})(1-x) + E_0(\mathrm{Ge})x$
$E_{0'}$	eV	3.4	3.14	$E_{0'}(\mathrm{Si})(a_0(\mathrm{Si})/a_0(x))^{1.92}$ [31]
References		[22, 32–34]	[33, 35, 36]	[24, 25]

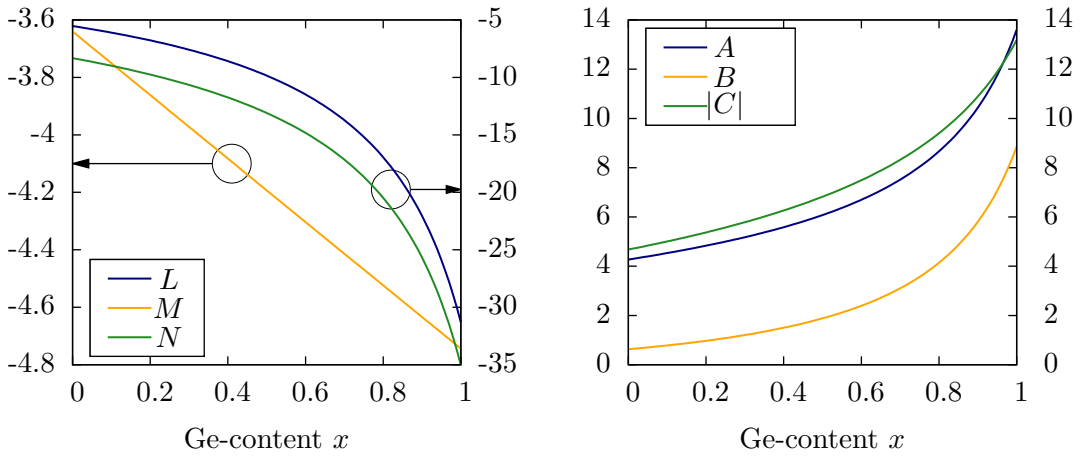


Figure 2.4: SiGe band parameters as functions of the Ge content x. L and N exhibit a clearly non-linear dependence on x due to their dependence on the perturbation parameter F, which is inversely proportional to the direct energy transition E_0.

Here, the lower right 3×3 block has the same structure as the DKK Hamiltonian, albeit with modified parameters L' and N'. The E_0 transition is now directly included in the models while it was modeled as a remote band in the DKK model. Thus, the parameters L and N, which depend on the perturbation parameter F, have different values than in the DKK model. The coupling parameter P is commonly expressed as

$$P = \frac{1}{\hbar}\sqrt{\frac{E_\mathrm{P}}{2m_e}}, \tag{2.19}$$

where E_P is called the Kane energy. Similar to the approach in the DKK model, the Kane model can be extended to eight bands and a spin-orbit Hamiltonian can be added to include spin-orbit coupling. Deformation potentials were omitted in Eq. (2.18), but apply in the same way as for the DKK Hamiltonian plus a hydrostatic deformation potential a_c for the conduction band.

An extensive list of parameters for III-V compound semiconductors is found in [38], including ternary and quaternary alloys.

k·p Hamiltonians for Electrons

In indirect semiconductors the conduction band minimum lies in some point different from the Γ-point. In Si, this is along the Δ-edge, whereas in Ge it is at the L-point. As in the case of holes, a **k**·**p** model can be formulated here as well. However, the expansion point $\mathbf{k} = \mathbf{0}$ is not placed a Γ but at or near the respective valley minima.

In Si, a two-band **k**·**p**-model due to Hensel, Hasegawa, and Nakayama [39] is used. The two bands in question are Δ_1 and $\Delta_{2'}$ and the point of expansion $\mathbf{k} = \mathbf{0}$ is placed at the X-point. Since the band structure is periodic in **k**-space, $\Delta_{2'}$ corresponds to Δ_1 in

Parameter	Value	Unit
m_l	0.916	m_e
m_t	0.196	m_e
M	0.235	m_e
k_0	0.15	$2\pi/a_0$
Ξ_u	9.2	eV
$\Xi_{\mathrm{u}'}$	7.0	eV

Table 2.2: Parameters for the two-band $\mathbf{k}\cdot\mathbf{p}$ model for Si.

the next Brillouin zone and vice versa. The bands are degenerate at the X-point. The remaining bands are treated as second-order perturbation. The effective Hamiltonian of the two-band $\mathbf{k}\cdot\mathbf{p}$ model reads [40]

$$\mathbf{H}_{2\times2} = \begin{bmatrix} \frac{\hbar^2k_x^2+\hbar^2k_y^2}{2m_\mathrm{t}} + \frac{\hbar^2(k_z-k_0)^2}{2m_\mathrm{l}} + \Xi_\mathrm{u}\varepsilon_{zz} & -\frac{\hbar^2k_xk_y}{M} + \Xi_{\mathrm{u}'}\varepsilon_{xy} \\ -\frac{\hbar^2k_xk_y}{M} + \Xi_{\mathrm{u}'}\varepsilon_{xy} & \frac{\hbar^2k_x^2+\hbar^2k_y^2}{2m_\mathrm{t}} + \frac{\hbar^2(k_z+k_0)^2}{2m_\mathrm{l}} + \Xi_\mathrm{u}\varepsilon_{zz} \end{bmatrix}, \tag{2.20}$$

where $k_0 = 0.15(2\pi/a_0)$ denotes the $\mathbf{k}$-space distance between the X-point and the Δ-valley minimum, m_l and m_t are longitudinal and transversal effective masses of the electrons, and Ξ_u and $\Xi_{\mathrm{u}'}$ are the uniaxial and shear deformation potentials. The two-band $\mathbf{k}\cdot\mathbf{p}$ parameters for Si are given in Table 2.2.

In Ge, the conduction band minima lie at the L-points. The conduction band is non-degenerate (apart from spin) and a single-band effective mass description can approximate the conduction band structure near the L-point. A more accurate band description can be obtained by including the heavy hole bands 2.2 eV below the conduction band edge; reference [41] gives the expression

$$E = \frac{\hbar^2k_\perp^2}{2m_e} + \frac{\hbar^2k_\perp^2}{2}\left(\frac{1}{m_\mathrm{t}} - \frac{1}{m_e}\right)\frac{E_\mathrm{g,L}}{E_\mathrm{g,L}+E} + \frac{\hbar^2k_\parallel^2}{2m_\mathrm{l}}, \tag{2.21}$$

citing [42]. Here, m_l and m_t are again the longitudinal and transversal effective masses, and $E_\mathrm{g,L} = 2.2\,\mathrm{eV}$ is the energy difference between conduction and valence band in the L-point. An effective Hamiltonian matrix can be constructed, which the eigenvalue in Eq. (2.21),

$$\mathbf{H} = \begin{bmatrix} \frac{\hbar^2k_x^2+\hbar^2k_y^2}{2m_e} + \frac{\hbar^2k_z^2}{2m_\mathrm{l}} & \sqrt{\frac{E_\mathrm{g,L}}{2M}}\hbar(k_x+\mathrm{i}k_y) \\ \sqrt{\frac{E_\mathrm{g,L}}{2M}}\hbar(k_x-\mathrm{i}k_y) & -E_\mathrm{g,L} \end{bmatrix}, \tag{2.22}$$

where $k_z = k_\parallel$ and $1/M = 1/m_\mathrm{t} - 1/m_e$. Recently, a ten-band model for L-valley electrons in Ge was proposed in [43], which includes two heavy hole bands below and two conduction bands above the conduction band edge, as well as spin-orbit interaction.

In SiGe, the conduction band minimum depends on the Ge content x. Since the Si has Δ-valleys and Ge L-valleys, a cross-over between Δ and L valleys occurs at the mole

fraction $x = 0.85$. Depending on whether the Ge content of a particular alloy is below or above 0.85, the two-band $\mathbf{k}\cdot\mathbf{p}$ model or the L-valley $\mathbf{k}\cdot\mathbf{p}$ model can be used. For alloys with x close to 0.85 both Δ and L need to be modeled.

2.1.2 The Effective Mass Model

The effective mass model can be viewed as the special case of a $\mathbf{k}\cdot\mathbf{p}$ model with a single band. All the interactions with other bands are lumped together in one parameter, i.e. the effective mass. The dispersion relation is parabolic and reads

$$E(\mathbf{k}) = E_0 + \frac{\hbar^2}{2}\mathbf{k}\cdot\mathbf{m}^{-1}\cdot\mathbf{k}, \tag{2.23}$$

where $\mathbf{m}$ is the tensor-valued effective mass parameter. A single-band description is accurate close to the band edge, i.e. for very small values of kinetic energy.

Electrons

The effective mass model is mostly used to describe bands that are non-degenerate at the valley minimum. This is the case for the conduction bands of the most important semiconductors, such as the Δ-valleys in Si, the L-valleys in Ge, or the Γ-valley in GaAs.

Due to symmetry, the effective mass can be described by one or two effective mass parameters. For the Γ-valley, which is isotropic, the effective mass tensor becomes a scalar m^*, such that

$$E(\mathbf{k}) = E_\mathrm{c} + \frac{\hbar^2 k^2}{2m^*}. \tag{2.24}$$

For the X/Δ and L-valleys two parameters are sufficient, i.e. the longitudinal and transversal effective masses,

$$E(\mathbf{k}) = E_\mathrm{c} + \frac{\hbar^2 k_\mathrm{l}^2}{2m_\mathrm{l}} + \frac{\hbar^2 k_\mathrm{t}^2}{2m_\mathrm{t}}, \tag{2.25}$$

where "l" and "t" denote the longitudinal and transversal direction, respectively. In the case of X or Δ-valleys, the corresponding basis vectors for the first out of six valleys are

$$\mathbf{e}_\mathrm{l} = \begin{bmatrix}1\\0\\0\end{bmatrix}, \quad \mathbf{e}_\mathrm{t}^{(1)} = \begin{bmatrix}0\\1\\0\end{bmatrix}, \quad \mathbf{e}_\mathrm{t}^{(2)} = \begin{bmatrix}0\\0\\1\end{bmatrix}. \tag{2.26}$$

The basis vectors of the other two valleys are obtained by permuting the vector elements. In the case of L-valleys, the basis vectors read

$$\mathbf{e}_\mathrm{l} = \frac{1}{\sqrt{3}}\begin{bmatrix}1\\1\\1\end{bmatrix}, \quad \mathbf{e}_\mathrm{t}^{(1)} = \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\\0\end{bmatrix}, \quad \mathbf{e}_\mathrm{t}^{(2)} = \frac{1}{\sqrt{6}}\begin{bmatrix}1\\1\\-2\end{bmatrix}. \tag{2.27}$$

The basis of the other three valleys can be obtained by changing the sign of one row and permuting the elements ($x \to y$, $y \to z$, $z \to x$).

Holes

The valence band structure of the technologically most important semiconductors has a three-fold degeneracy at the highest-symmetry point Γ (if SO-coupling is neglected). The curvature of the dispersion relation at the Γ-point is ill-defined: Different curvatures are obtained when looking in the equivalent ⟨100⟩, ⟨110⟩, and ⟨111⟩ directions. These differences cannot be represented by a single effective mass tensor. This makes an approximation of the valence band structure by parabolic effective mass Hamiltonians less straightforward than it was the case for electrons. The complications involving the valence band become clear when looking at the energy contour plot of the valence dispersion relation shown in Fig. 2.5.

It is apparent that providing the valence dispersion relation along the Cartesian k_x, k_y, and k_z axes is insufficient to describe the dispersion relation as a whole. In fact the density of states is dominated by protrusions of the contour or "fingers" pointing at each of the equivalent ⟨110⟩ directions.

This raises the question whether the contribution of these fingers can be modeled in an effective way. In fact, two opposing fingers, e.g. [110] and $[\bar{1}\bar{1}0]$, can be approximated by an ellipsoid as shown by the overlay in Fig. 2.5. All the twelve fingers can thus be approximated as six intersecting ellipsoids. Each ellipsoid can be treated as as *valley*, with its own effective mass Hamiltonian. The effective mass of each of the six valleys is described by a tensor, with three masses, m_l, m_t, and m_z, where for a [110]-pointing ellipsoid, l corresponds to direction [110], t to $[1\bar{1}0]$, and z to [001].

Each ellipsoid corresponds to one of the following six envelope-function states in the three-band $\mathbf{k}\cdot\mathbf{p}$ model [44],

$$\begin{aligned} \psi_1 &= \frac{1}{\sqrt{2}}\left|\begin{matrix}1\\1\\0\end{matrix}\right\rangle, \quad \psi_2 = \frac{1}{\sqrt{2}}\left|\begin{matrix}1\\-1\\0\end{matrix}\right\rangle, \quad \psi_3 = \frac{1}{\sqrt{2}}\left|\begin{matrix}1\\0\\1\end{matrix}\right\rangle, \\ \psi_4 &= \frac{1}{\sqrt{2}}\left|\begin{matrix}1\\0\\-1\end{matrix}\right\rangle, \quad \psi_5 = \frac{1}{\sqrt{2}}\left|\begin{matrix}0\\1\\1\end{matrix}\right\rangle, \quad \psi_6 = \frac{1}{\sqrt{2}}\left|\begin{matrix}0\\1\\-1\end{matrix}\right\rangle. \end{aligned} \tag{2.28}$$

Inserting the states into the Hamiltonian from Eq. (2.8), one obtains effective mass parameters for each of the six valleys represented by the ellipsoids [44],

$$m_\mathrm{l} = -\frac{2}{L+M-N}, \quad m_\mathrm{t} = -\frac{2}{L+M+N}, \quad m_\mathrm{z} = -\frac{1}{M}, \tag{2.29}$$

as well as the deformation potentials

$$D_\mathrm{l} = \frac{l+m-n}{2}, \quad D_\mathrm{t} = \frac{l+m+n}{2}, \quad D_\mathrm{z} = m, \tag{2.30}$$

The *six-valley-model* neglects the couplings between the six ellipsoids that are present in the three-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian, which the model has been derived from. The couplings are visible as *anti-crossings* in Fig. 2.5 that connect the six ellipsoids to a single warped energy surface.

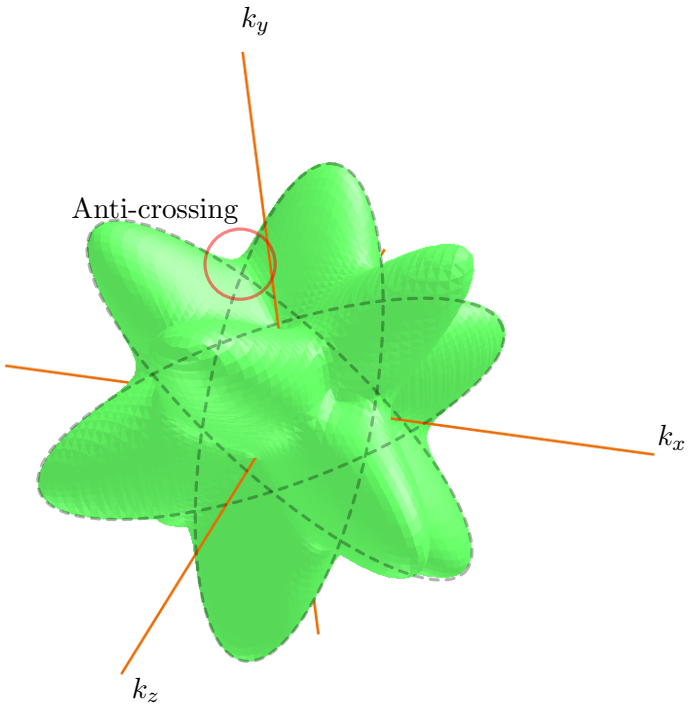


Figure 2.5: An energy-iso-surface of the valence dispersion relation in Si computed from a three-band $\mathbf{k}\cdot\mathbf{p}$ model; due to the degeneracy at the Γ-point, the valence band structure in diamond and zinc-blende crystals has a rather complex shape even close the valence band. The iso-surface exhibits a warped structure with "fingers" pointing in the ⟨110⟩-directions. These twelve fingers can be approximated by six ellipsoids, three of which have been overlaid in this figure. The ellipsoids closely approximate the energy surface everywhere except at the locations where two ellipsoids intersect; here, the coupling between the $\mathbf{k}\cdot\mathbf{p}$ bands causes an anti-crossing which cannot be represented by the ellipsoids.

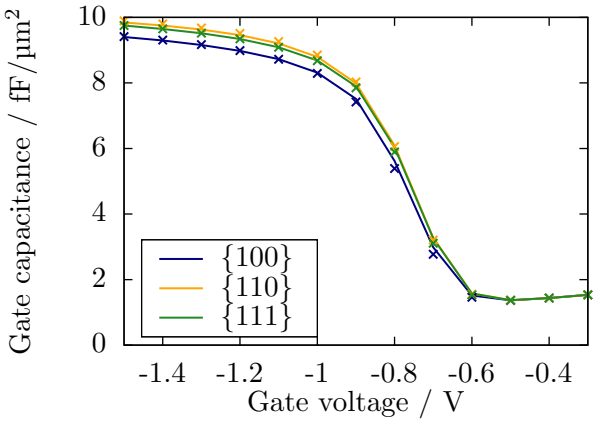


Figure 2.6: Capacitance-voltage curves for p-type MOS-capacitors with different surface orientations: Results from the six-valley-model (lines) and the three-band **k**·**p** model (symbols) are compared. The six-valley-model results agree very well with those from three-band **k**·**p** and both have the same orientation-dependence.

Each of the six decoupled valleys has a parabolic dispersion relation of its own and is independent of the other five. Neglecting the couplings makes the *six-valley-model* less accurate than the three-band or even six-band **k**·**p** models. The model does however capture the effects of confinement, orientation, and strain on the hole concentration quite well (see Fig. 2.6) while providing analytical dispersion relations for the valence band that can be handled much more efficiently in simulations.

The six-valley-model can also be used as basis for mobility calculation (c.f. Section 2.4.2). Figure 2.7 shows good agreement between results from the six-valley-model and the three-band **k**·**p**-model. The accuracy of the six-valley-model however depends on that of the three-band **k**·**p** model, i.e. both models are limited to materials with small spin-orbit-coupling energies, such as Si.

2.1.3 Electronic Structure of Confined Systems

Charge carriers may be spatially confined. Confinement can be either geometric, e.g. by putting a semiconductor between insulators, electrostatic, where the electrostatic potential forms one or more barriers holding carriers in place, or both. Confinement can be partial, which means that carriers are confined in one or two dimensions, while they are free to move in the remaining two or one dimension, respectively. Systems of carriers confined in all three spatial directions are called *quantum dots*.

Partially confined systems have an altered electronic structure, a so-called subband structure, which is the energy dispersion relation along the axes of free carrier movement. To obtain the subband structure a bulk-Hamiltonian, such as the **k**·**p** and effective mass Hamiltonians defined in the previous sections, is converted into a **k**-dependent partial differential operator, of which the eigenvalues form the subband structure.

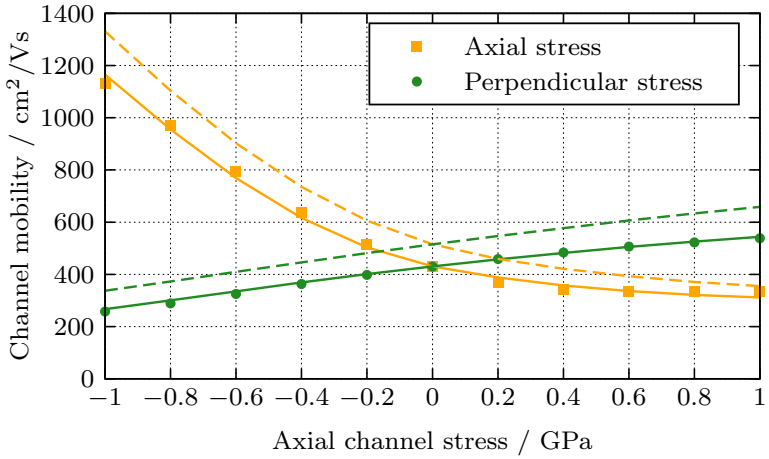


Figure 2.7: Stress-dependent hole mobility of a 5 nm thin Si channel with a (001)-surface; stress is applied in transport direction ([110]) and perpendicular to it ([1$\bar{1}$0]). Symbols were obtained from a **k**·**p**-Hamiltonian, solid lines from the six-valley-model with parameters fitted to the **k**·**p**-result, and dashed lines with analytical six-valley parameters.

The conversion to a partial differential operator is done in two steps [45]:

Rotation: As the bulk-Hamiltonian is defined in the *crystal coordinate system* it needs to be rotated in order to represent it in the *device coordinate system*, where the coordinate axes coincide with the axes of free carrier movement.

Substitution: The momentum operator in **k**-space, $\hbar k_\xi$, is replaced by its real-space representation $-\mathrm{i}\hbar\partial_\xi$, where ξ denotes a coordinate pointing in confinement direction, which is perpendicular to the axes of free carrier movement. Consequently, a term of the kinetic energy operator, $\hbar^2 k_\xi k_\eta / 2m_{\xi\eta}$ is converted into $-\hbar^2 \partial_\xi \partial_\eta / 2m_{\xi\eta}$.

An eigenvalue problem is formulated,

$$\mathbf{H}_\mathbf{k}\psi_\mathbf{k} = E_\mathbf{k}\psi_\mathbf{k}, \tag{2.31}$$

where the envelope wave function ψ is the solution variable. Geometric confinement is imposed by homogeneous Dirichlet boundary conditions. The n^{th} eigenvalue $E_{n,\mathbf{k}}$ and eigenvector $\psi_{n,\mathbf{k}}$ correspond to the energy value of subband n at **k** and the corresponding envelope wave function, respectively.

For the general case of **k**·**p**-Hamiltonians, the $E_{n,\mathbf{k}}$ values computed at different **k** define a non-trivial subband dispersion relation. For effective mass Hamiltonians, the subband dispersion relation is parabolic and can be expressed analytically. Stern and Howard derived the parabolic subband dispersion relation of electrons in the X and L valleys confined to a thin film with {100}, {110}, and {111}-oriented surfaces [46]. Here, a general procedure is given to obtain the subband dispersion relation, i.e. subband

effective mass tensor, for arbitrary orientations and arbitrary dimensionality, i.e. both films and wires. The starting point is the effective-mass Schrödinger equation,

$$H\psi = -\frac{\hbar^2}{2m_e}\boldsymbol{\nabla}\cdot\mathbf{w}\cdot\boldsymbol{\nabla}\psi + V(\mathbf{r})\psi = -\frac{\hbar^2}{2m_e}\partial_i w_{ij}\partial_j = E\psi + V(\mathbf{r})\psi. \tag{2.32}$$

The inverse effective mass tensor $\mathbf{w}$ is separated into blocks corresponding to confined ($\perp$) and unconfined ($\parallel$) bases,

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_\perp & \mathbf{w}_\mathrm{c} \\ \mathbf{w}_\mathrm{c}^\dagger & \mathbf{w}_\parallel \end{pmatrix}. \tag{2.33}$$

The potential $V(\mathbf{r})$ is assumed to be invariant (or sufficiently slow-varying) in the transport direction, leaving a potential that only depends on the confined coordinates, $V(\mathbf{r}) = V(\mathbf{r}_\perp)$. Using $\boldsymbol{\nabla}_\parallel \mapsto \mathrm{i}\mathbf{k}_\parallel$, Eq. (2.32) is rewritten as

$$-\frac{\hbar^2}{2m_e}\left[\boldsymbol{\nabla}_\perp\cdot\mathbf{w}_\perp\cdot\boldsymbol{\nabla}_\perp + 2\mathrm{i}\mathbf{k}_\parallel\cdot\mathbf{w}_\mathrm{c}\cdot\boldsymbol{\nabla}_\perp - \mathbf{k}_\parallel\cdot\mathbf{w}_\parallel\cdot\mathbf{k}_\parallel\right]\phi + V(\mathbf{r}_\perp)\phi = E\phi, \tag{2.34}$$

where ϕ is the component of the separated wave function ψ that lies in the subspace of the confined coordinates. Defining a new unknown function F as $\phi = Fe^{\mathrm{i}\mathbf{k}_\perp\cdot\mathbf{r}_\perp}$ and choosing $\mathbf{k}_\perp = -\mathbf{w}_\perp^{-1}\cdot\mathbf{w}_\mathrm{c}\cdot\mathbf{k}_\parallel$, all first-order terms vanish and Eq. (2.34) can be simplified to

$$-\frac{\hbar^2}{2m_e}\left[\boldsymbol{\nabla}_\perp\cdot\mathbf{w}_\perp\cdot\boldsymbol{\nabla}_\perp - \mathbf{k}_\parallel\cdot(\mathbf{w}_\parallel - \mathbf{w}_\mathrm{c}\cdot\mathbf{w}_\perp^{-1}\cdot\mathbf{w}_\mathrm{c})\cdot\mathbf{k}_\parallel\right]F + V(\mathbf{r}_\perp)F = EF. \tag{2.35}$$

Thus, the eigenvalue equation for F reads

$$-\frac{\hbar^2}{2m_e}\left[\boldsymbol{\nabla}_\perp\cdot\mathbf{w}_\perp\cdot\boldsymbol{\nabla}_\perp\right]F + V(\mathbf{r}_\perp)F = EF. \tag{2.36}$$

while the dispersion relation for the subbands reads

$$E(\mathbf{k}_\parallel) = E_n + \frac{\hbar^2}{2m_e}\mathbf{k}_\parallel\cdot(\mathbf{w}_\parallel - \mathbf{w}_\mathrm{c}\cdot\mathbf{w}_\perp^{-1}\cdot\mathbf{w}_\mathrm{c})\cdot\mathbf{k}_\parallel, \tag{2.37}$$

where E_n is the eigenenergy from Eq. (2.36) corresponding to the n^{th} subband. Equation (2.37) is equivalent to

$$E(\mathbf{k}_\parallel) = E_n + \frac{\hbar^2 k_i k_j}{2m_e m_{ij}}, \tag{2.38}$$

with m_{ij} being a component of the the mass tensor, i.e. $\mathbf{w}$'s inverse. This means that the confinement mass is obtained by rotating and projecting the inverse mass tensor, while the dispersion mass is obtained by rotating and projecting the mass tensor. This is consistent with the special-case formulas given in [46].

2.2 Carrier Density and Electrostatics

The carrier concentration is directly connected to the electronic structure. Every electronic state (including conduction band and valence band states) contributes to the carrier concentration according to the state density and its occupance. Each state, confined or not, is identified by a subband index n and/or a wave-vector $\mathbf{k}$, as indicated in the previous section. For partially confined systems, $\mathbf{k}$ is restricted to the subspace spanned by the axes of free movement. For fully confined systems (*quantum dots*), this subspace is a null-space and the subband index becomes the state index.

In the semi-classical picture, carriers are viewed as particles of a *carrier gas* for most of the time, their wave-like nature only being appreciated when calculating the transition probabilities or rates of a carrier scattering from one state to another. Quantum confinement is another wave-like property of carriers and is dealt with in the semi-classical framework by reducing the dimensionality of the carrier gas. The dimension of the space spanned by the axes of free movement defines the dimensionality of the carrier gas: Unconfined carriers can move in three dimension, hence they constitute a 3D carrier gas (3D electron gas/3DEG); carriers confined to a thin film or surface constitute a 2D carrier gas (2D electron gas/2DEG), and similarly, carriers confined to a thin wire form a 1D carrier gas (1DEG); fully confined carriers inside a quantum dot can be seen as a 0D carrier gas.

For the general case of a partially confined system the wave function of a state is expressed as a product of a standing wave and a plane wave,

$$\Psi_{n,\mathbf{k}}(\mathbf{r}) = \psi_{n,\mathbf{k}}(\mathbf{r}) L^{-\frac{d}{2}} e^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}, \tag{2.39}$$

where d is the dimension of the carrier gas. The factor $L^{-\frac{d}{2}}$ normalizes the plane wave component with respect to a cube/square/line segment of size L. The probability *density* of the state is

$$\rho_{n,\mathbf{k}}(\mathbf{r}) = L^{-d} |\psi_{n,\mathbf{k}}(\mathbf{r})|^2. \tag{2.40}$$

The carrier concentration is obtained by summing the density of all states weighted by their semi-classical distribution function $f_n(\mathbf{k})$,

$$n(\mathbf{r}) = \sum_{n,\mathbf{k}} \frac{g}{L^d} \rho_{n,\mathbf{k}}(\mathbf{r}) f_n(\mathbf{k}) \approx \frac{g}{(2\pi)^d} \sum_n \int_{\mathbb{R}^d} \rho_{n,\mathbf{k}}(\mathbf{r}) f_n(\mathbf{k}) \mathrm{d}^d\mathbf{k}, \tag{2.41}$$

where g denotes the degeneracy of the state, due to spin and valley multiplicity.

In equilibrium, the distribution function is a Fermi-Dirac distribution, and depends on energy rather than the state index $(n, \mathbf{k})$,

$$f^0(E) = \frac{1}{1 + e^{\frac{E - E_\mathrm{F}}{k_\mathrm{B} T}}}. \tag{2.42}$$

The equilibrium distribution is parametrized with respect to temperature T and Fermi-level E_F. For a generic (sub-)band structure, the equilibrium carrier concentration reads

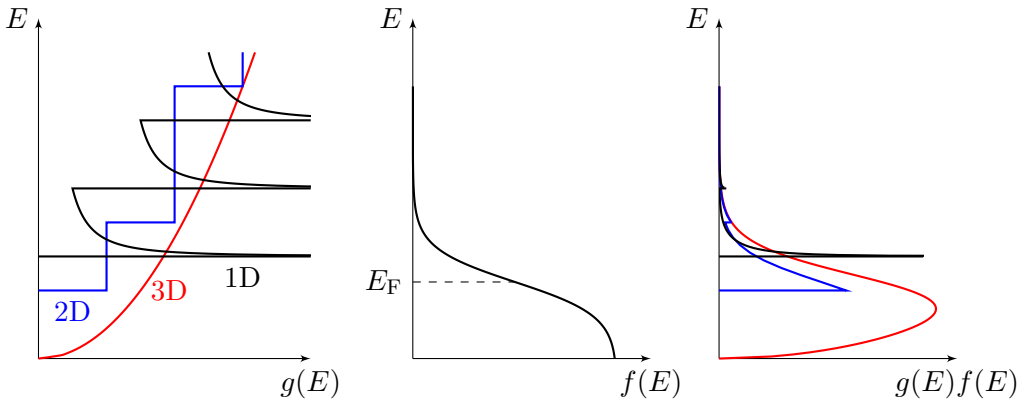


Figure 2.8: Qualitative picture of the density of states (left), equilibrium distribution (center), and spectral carrier density (right) for different carrier gas dimensionalities assuming a parabolic energy dispersion

$$n(\mathbf{r}) = \frac{g}{(2\pi)^d} \sum_n \int_{\mathbb{R}^d} \rho_{n,\mathbf{k}}(\mathbf{r}) f^0(E_n(\mathbf{k})) \mathrm{d}^d\mathbf{k}. \tag{2.43}$$

For a parabolic subband structure, i.e. for constant effective mass, the probability density $\rho_{n,\mathbf{k}}(\mathbf{r})$ is independent of the $\mathbf{k}$-vector and the $\mathbf{k}$-space integral in Eq. (2.43) can be reduced to an energy-integral using the density of states (see Fig. 2.8),

$$n(\mathbf{r}) = \sum_n \int_0^\infty \rho_{n,\mathbf{k}}(\mathbf{r}) g(E) f^0(E) \mathrm{d}E, \tag{2.44}$$

The expression can be further simplified into

$$n(\mathbf{r}) = \sum_v \sum_n \rho_{v,n}(\mathbf{r}) N^d_{\mathrm{C},v,n} \mathrm{SF}^d_{v,n}, \tag{2.45}$$

where the indices v and n denote valley and subband index. $N^d_{\mathrm{C},v,n}$ and $\mathrm{SF}^d_{v,n}$ are the effective density of states and the *supply function*, respectively; the expressions for these terms depend on the carrier gas dimensionality d and are summarized in Table 2.3.

Each electron and hole carries one negative or positive elementary charge, respectively. As such, they contribute to an overall space charge distribution, which is also comprised of ionized dopant atoms in the crystal lattice. The total space charge density reads

$$\varrho = q_0(N_\mathrm{D} - N_\mathrm{A} - n + p), \tag{2.46}$$

where n and p denote the electron and hole concentrations, while N_D and N_A represent the concentrations of *ionized* donors and acceptors. Not all dopants need to be ionized, however, at room temperature and under moderate doping concentration and bias field, it is safe to assume a complete ionization of the dopants.

Table 2.3: The supply function SF, the effective density of states N_C and the density of states mass m_dos are summarized for carrier gases of dimension d. $\mathcal{F}_i$ denotes the Fermi-Dirac integral of order i. $m_\parallel$ refers to the subband dispersion mass from Eq. (2.37)

d	SF^d	N_C^d	m_dos^d
0D	$\dfrac{1}{1+\exp\left(\frac{E_s-E_\mathrm{F}}{k_\mathrm{B}T}\right)}$	g	1
1D	$\mathcal{F}_{-\frac{1}{2}}\left(-\frac{E_s-E_\mathrm{F}}{k_\mathrm{B}T}\right)$	$g\left(\dfrac{m_\mathrm{dos}^\mathrm{1D}k_\mathrm{B}T}{2\pi\hbar^2}\right)^{\frac{1}{2}}$	$m_\parallel$
2D	$\ln\left(1+e^{\frac{E_\mathrm{F}-E_s}{k_\mathrm{B}T}}\right)$	$g\dfrac{m_\mathrm{dos}^\mathrm{2D}k_\mathrm{B}T}{2\pi\hbar^2}$	$\sqrt{\det(\underset{\sim}{m}_\parallel)}$
3D	$\mathcal{F}_{\frac{1}{2}}\left(-\frac{E_\mathrm{c}-E_\mathrm{F}}{k_\mathrm{B}T}\right)$	$g\left(\dfrac{m_\mathrm{dos}^\mathrm{3D}k_\mathrm{B}T}{2\pi\hbar^2}\right)^{\frac{3}{2}}$	$\sqrt[3]{\det(\underset{\sim}{m})}$

The charge density in Eq. (2.46) affects the electrostatic potential via the Poisson equation,

$$\boldsymbol{\nabla}\cdot\varepsilon\boldsymbol{\nabla}\varphi=-\varrho. \tag{2.47}$$

The electrostatic potential φ locally shifts the state energies in the band structure. In other words the electrostatic potential enters the Schrödinger equation as a diagonal term V,

$$\mathbf{H}_0\left|\boldsymbol{\Psi}\right\rangle-q_0\varphi\left|\boldsymbol{\Psi}\right\rangle=E\left|\boldsymbol{\Psi}\right\rangle, \tag{2.48}$$

where $\mathbf{H}$ can be any kind of effective mass or $\mathbf{k}\cdot\mathbf{p}$ (or other type of) Hamiltonian. This closes the circle between electronic structure, concentration, and electrostatics. To solve the overall problem a *self-consistent* solution needs to be found, that satisfies all three equations simultaneously.

2.3 Scattering Processes

A predictive semi-classical device modeling and simulation framework needs to include all relevant carrier scattering processes occurring in technologically important semiconductors: strained Si, Ge, SiGe, and group III/V compound semiconductors. From a physical point of view, scattering processes can be divided into three major categories: carrier-defect scattering, carrier-carrier scattering, and carrier-phonon scattering. This classification is shown in Fig. 2.9. From a modeling point of view, however, it makes more sense to divide the processes into two categories: Coulomb-like interactions and random-potential-like interactions, as will be done in this section (shown in Fig. 2.10).

The starting point for modeling scattering processes within the semi-classical framework is Fermi's golden rule, which defines the transition rate from state $n,\mathbf{k}$ to state $n',\mathbf{k}'$ as

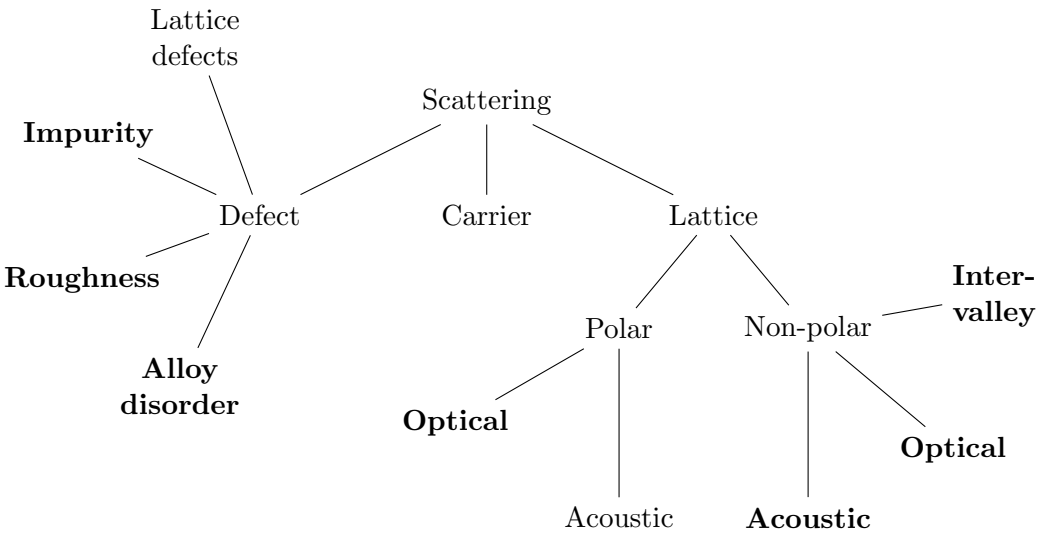


Figure 2.9: Taxonomy of scattering processes from a physical point of view. The highlighted processes dominate transport in semiconductor channels at room temperature.

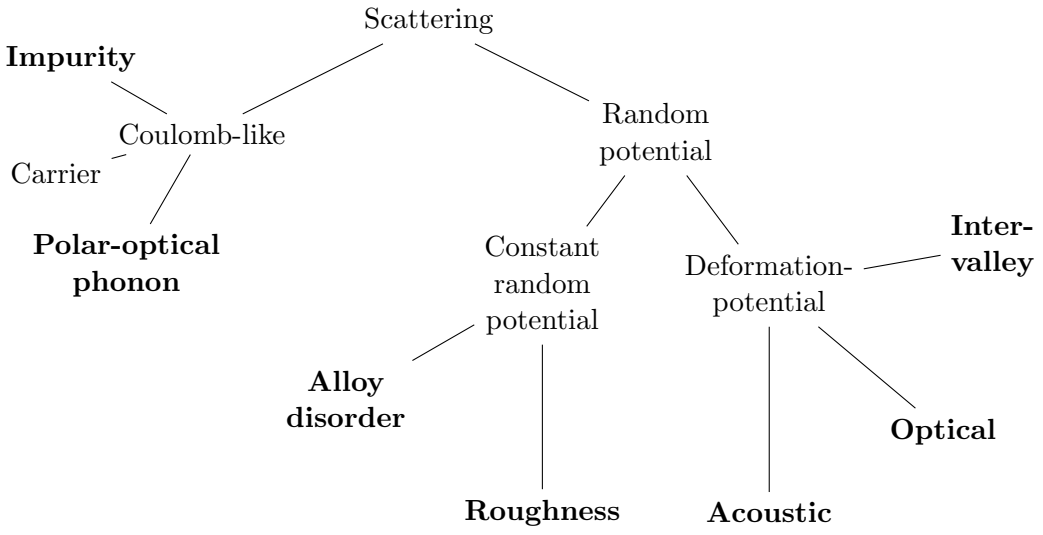


Figure 2.10: Taxonomy of scattering processes from a modeling point of view. Coulomb-like processes are due to a charge fluctuations which disturb an electronic state via electrostatic interaction. Random-potential processes are due to variations of the electronic structure which can be either caused by defects and disorder, or by deformations of the crystal lattice through phonons.

[47]

$$S_{n,n'}(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle\delta(E(\mathbf{k}) - E(\mathbf{k}') \pm \hbar\omega), \tag{2.49}$$

One of the assumptions in deriving Fermi's golden rule, is that a system is perturbed by time-dependent potential $V(\mathbf{r},t)$, where the time dependence is represented by a harmonic oscillation. Thus the time-dependent potential is written as product of a static part and a harmonic oscillation,

$$V(\mathbf{r},t) = V(\mathbf{r})e^{\mathrm{i}\omega t}, \tag{2.50}$$

where a purely static perturbation can be represented by setting $\omega = 0$. The effect of the harmonic oscillation on energy conservation is taken into account by the $\delta(E(\mathbf{k}) - E(\mathbf{k}') \pm \hbar\omega)$ term in Eq. (2.49), whereas the static part of the perturbation is represented as the *matrix element*,

$$H_{n,n';\mathbf{k},\mathbf{k}'} = \int\limits_{\mathbb{R}^3} \Psi^*_{n,\mathbf{k}}(\mathbf{r})\Psi_{n',\mathbf{k}'}(\mathbf{r})V(\mathbf{r})\mathrm{d}^3r, \tag{2.51}$$

where $V(\mathbf{r})$ is the static part of the perturbing potential. The square matrix element is central to the modeling of the various scattering processes as it contains all the physical properties of the scattering interaction. It must be noted that $V(\mathbf{r})$ is a random function of position $\mathbf{r}$ and not square-integrable in many cases, making the evaluation of the integral in Eq. (2.51) difficult. These problems can be avoided by looking at the *ensemble average* of the square matrix element,

$$\begin{aligned}\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle &= \left\langle \left| \int\limits_{\mathbb{R}^3} \Psi^*_{n,\mathbf{k}}(\mathbf{r})\Psi_{n',\mathbf{k}'}(\mathbf{r})V(\mathbf{r})\mathrm{d}^3r \right|^2 \right\rangle \\ &= \iint\limits_{\mathbb{R}^{3\times3}} \Psi_{n,\mathbf{k}}(\mathbf{r})\Psi^*_{n',\mathbf{k}'}(\mathbf{r})\Psi^*_{n,\mathbf{k}}(\mathbf{r}')\Psi_{n',\mathbf{k}'}(\mathbf{r}')\langle V^*(\mathbf{r})V(\mathbf{r}')\rangle\mathrm{d}^3r\ \mathrm{d}^3r'.\end{aligned} \tag{2.52}$$

The expression $\langle V^*(\mathbf{r})V(\mathbf{r}')\rangle =: c(\mathbf{r}-\mathbf{r}')$ is the autocorrelation function of the perturbing potential, which is not random but rather well defined and integrable.

If the random potential $V(\mathbf{r})$ is spatially uncorrelated, the autocorrelation function becomes $\langle V^*(\mathbf{r})V(\mathbf{r}')\rangle = |V_0|^2L^3\delta(\mathbf{r}-\mathbf{r}')$. As a consequence, the square matrix element for an *uncorrelated random potential* simplifies to

$$\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle = |V_0|^2L^3\int\limits_{\mathbb{R}^3} |\Psi_{n,\mathbf{k}}(\mathbf{r})\Psi_{n',\mathbf{k}'}(\mathbf{r})|^2\mathrm{d}^3r. \tag{2.53}$$

For plane wave states, $L^{-\frac{3}{2}}e^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}$ can be inserted as $\psi_{n,\mathbf{k}}$, giving a square matrix element,

$$\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle = |V_0|^2, \tag{2.54}$$

which is constant, i.e. independent from $\mathbf{k} - \mathbf{k}'$. Such scattering processes are called *isotropic* because the scattering rate,

$$S_{n,n'}(\mathbf{k}, \mathbf{k}') = \frac{2\pi}{\hbar}|V_0|^2\delta(E - E' \pm \hbar\omega), \tag{2.55}$$

is uniform for all final wave vectors $\mathbf{k}'$ on the energy surface $E(\mathbf{k}') = E(\mathbf{k}) \pm \hbar\omega$.

For partially confined systems the electron states $\Psi_{n,k}$ are separated into a bound state in the confinement cross-section and a plane wave along the axes of free movement,

$$\Psi_{n,\mathbf{k}}(\mathbf{r}) = \psi_{n,\mathbf{k}}(\mathbf{r}_\perp)L^{-\frac{d}{2}}e^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}_\parallel}. \tag{2.56}$$

Using this separation approach, Eq. (2.53) can be rewritten as

$$\begin{aligned}\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle &= |V_0|^2L^3 \int\limits_{\mathbb{R}^{3-d}} |\psi_{n,\mathbf{k}}(\mathbf{r}_\perp)\psi_{n',\mathbf{k}'}(\mathbf{r}_\perp)|^2\mathrm{d}^{3-d}r \; L^{-2d} \int\limits_0^L \underbrace{|e^{\mathrm{i}(\mathbf{k}+\mathbf{k}')\cdot\mathbf{r}_\parallel}|^2}_{=1}\mathrm{d}^d r \\ &= |V_0|^2L^{3-d} \int\limits_{\mathbb{R}^{3-d}} |\psi_{n,\mathbf{k}}(\mathbf{r}_\perp)\psi_{n',\mathbf{k}'}(\mathbf{r}_\perp)|^2\mathrm{d}^{3-d}r,\end{aligned} \tag{2.57}$$

for a d-dimensional carrier gas. The integral is called the *form factor*.

In the following sections $\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle$ shall first be derived for random-potential processes and then for Coulomb-like processes.

2.3.1 Non-Polar Lattice Scattering

Non-polar lattice scattering is divided into two types of processes: (i) near-elastic scattering by acoustic phonons (acoustic deformation potential, ADP) and (ii) inelastic scattering by optical phonons (optical deformation potential, ODP, and inter-valley scattering, IVS). Figure 2.11 shows the different scattering processes in the context of the phonon dispersion relation. As the different names suggest, non-polar lattice scattering is due to a deformation-potential-type interaction. Lattice vibrations cause a random displacement of atoms in the lattice which locally modifies the electronic structure. These local band structure changes in turn cause a perturbing potential for the electrons.

Scattering by Optical Phonons

The deformation potential due to optical phonons is proportional to the atomic displacement $\mathbf{u}(\mathbf{r})$,

$$V(\mathbf{r}) = D_{\mathrm{opt}}u(\mathbf{r}, t). \tag{2.58}$$

Assuming that the displacement direction is uniformly distributed, the autocorrelation function of the perturbing potential can be written as

$$c(\mathbf{r} - \mathbf{r}') = D_{\mathrm{opt}}^2\langle u^*(\mathbf{r})u(\mathbf{r}')\rangle. \tag{2.59}$$

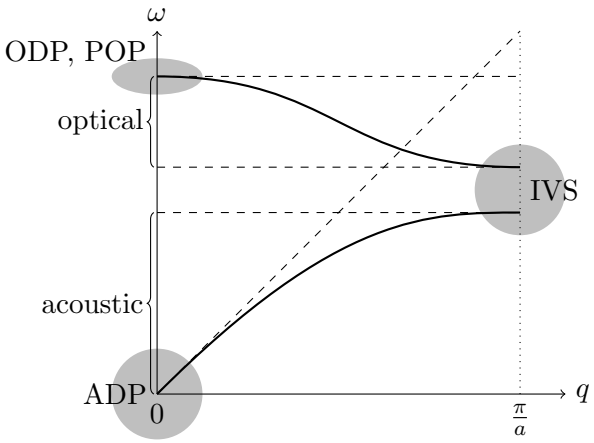


Figure 2.11: Qualitative picture of the phonon dispersion relation in most semiconductors; acoustic and optical branches are marked; also shown are those parts of the phonon dispersion responsible for the different phonon-related scattering processes.

Applying the Wiener-Khinchin theorem, the displacement's autocorrelation function can be represented as the inverse Fourier transform of its power spectrum,

$$\langle u^*(\mathbf{r})u(\mathbf{r}')\rangle = \sum_{\mathbf{q}} |\tilde{U}(\mathbf{q})|^2 e^{\mathrm{i}\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}')} \approx \left(\frac{L}{2\pi}\right)^3 \int\limits_{\mathbb{R}^3} |\tilde{U}(\mathbf{q})|^2 e^{\mathrm{i}\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}')} \mathrm{d}^3q, \tag{2.60}$$

with the spectrum of the potential remaining to be determined.

Optical phonons can be viewed as an ensemble of harmonic oscillators - one per unit cell of the crystal [48]. Classically, the oscillator's energy is equal to the maximum kinetic energy,

$$\max\{E_{\mathrm{kin}}\} = 2\bar{M}\omega^2|\tilde{U}(\mathbf{q})|^2, \tag{2.61}$$

where ω is the oscillator frequency and $\bar{M}$ is the reduced oscillator mass [48]. Quantum-mechanically, the oscillator energy depends on the number of energy quanta stored in the oscillator,

$$E = \hbar\omega\left(N + \frac{1}{2}\right). \tag{2.62}$$

Equating the two energies, one obtains the squared displacement amplitude,

$$|\tilde{U}(\mathbf{q})|^2 = \frac{\hbar}{2\bar{\rho}_{\mathrm{m}}L^3\omega}\left(N + \frac{1}{2}\right), \tag{2.63}$$

which is $\mathbf{q}$-independent. It can now be used to express the autocorrelation function of

the perturbing potential,

$$\begin{aligned} c(\mathbf{r}-\mathbf{r}') &= \frac{\hbar D_{\text{opt}}^2}{2\bar{\rho}_{\text{m}} L^3 \omega}\left(N+\frac{1}{2}\mp\frac{1}{2}\right)\left(\frac{L}{2\pi}\right)^3 \int\limits_{\mathbb{R}^3} e^{\mathrm{i}\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}')}\mathrm{d}^3 q \\ &= \frac{\hbar D_{\text{opt}}^2}{2\bar{\rho}_{\text{m}}\omega}\left(N+\frac{1}{2}\mp\frac{1}{2}\right)\delta(\mathbf{r}-\mathbf{r}'), \end{aligned} \tag{2.64}$$

where the sign $\mp$ distinguishes between phonon absorption and emission, respectively. We can identify $D_{\text{opt}}^2|\tilde{U}(\mathbf{q})|^2$ as $|V_0|^2$ and insert into Eq. (2.53); the square matrix element reads

$$\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle = \frac{\hbar D_{\text{opt}}^2}{2\bar{\rho}_{\text{m}}\omega}\left(N+\frac{1}{2}\mp\frac{1}{2}\right)\int\limits_{\mathbb{R}^3}|\Psi_{n,\mathbf{k}}(\mathbf{r})\Psi_{n',\mathbf{k}'}(\mathbf{r})|^2\mathrm{d}^3 r. \tag{2.65}$$

For a d-dimensional carrier gas in a partially confined system, the squared matrix element can be written as

$$\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle = \frac{\hbar D_{\text{opt}}^2}{2\bar{\rho}_{\text{m}} L^d \omega}\left(N+\frac{1}{2}\mp\frac{1}{2}\right)\int\limits_{\mathbb{R}^{3-d}}|\psi_{n,\mathbf{k}}(\mathbf{r})\psi_{n',\mathbf{k}'}(\mathbf{r})|^2\mathrm{d}^{3-d} r, \tag{2.66}$$

with the *form factor* integral known from Eq. (2.57).

Intra-valley and Inter-valley Scattering

The expression given in Eq. (2.66) can be used to model two kinds of scattering processes, intra-valley scattering and inter-valley scattering. In intra-valley scattering - also called optical deformation potential (ODP) scattering - the final states of the scattered carrier are restricted to the same valley as that of the initial state. ODP scattering occurs only for Γ-valley holes and L-valley electrons. This restriction is due to selection rules related to the symmetry of the diamond and zincblende crystals.

Contrary to ODP scattering, inter-valley scattering (IVS) can only occur between two different valleys. A slightly modified version of Eq. (2.66) can be used for modeling IVS scattering from valley v to valley v',

$$\langle|H_{v,v';n,n';\mathbf{k},\mathbf{k}'}|^2\rangle = \frac{\hbar D_{\text{iv}}^2}{2\bar{\rho}_{\text{m}} L^d \omega}\left(N+\frac{1}{2}\mp\frac{1}{2}\right)\int\limits_{\mathbb{R}^{3-d}}|\psi_{v,n,\mathbf{k}}(\mathbf{r})\psi_{v',n',\mathbf{k}'}(\mathbf{r})|^2\mathrm{d}^{3-d} r. \tag{2.67}$$

Scattering by Acoustic Phonons

The main difference between acoustic and optical phonon scattering is that the acoustic perturbation potential is related to the local strain, i.e. the derivative of displacement, rather than the displacement itself,

$$U_{\text{ac}}(\mathbf{r},t) = D_{\text{ac}}\varepsilon(\mathbf{r},t) = D_{\text{ac}}\frac{\partial u(\mathbf{r},t)}{\partial x} \tag{2.68}$$

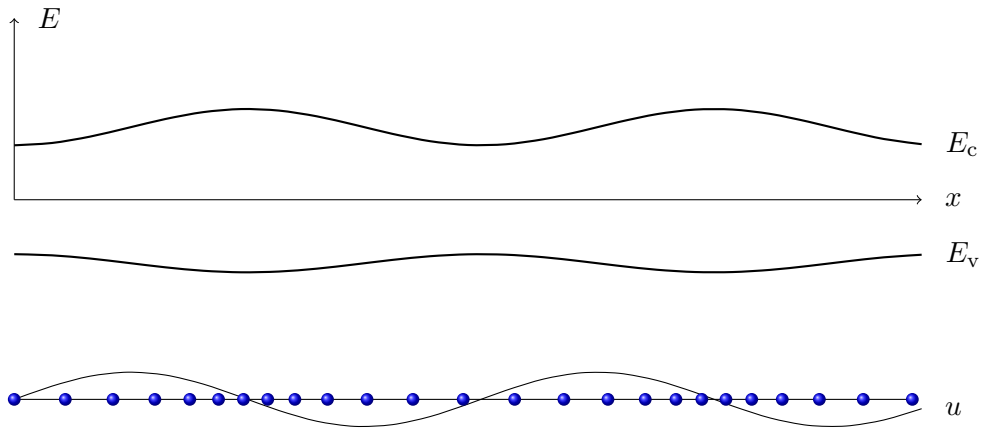


Figure 2.12: Perturbing potential due to acoustic lattice vibrations; an elastic causes local changes of mass density which translate into variations of the conduction and valence band edges. The displacement and the resulting potentials are phase-shifted by $\pi/2$ w.r.t. each other.

This is illustrated in Fig. 2.12. The autocorrelation function of the perturbing potential reads

$$c(\mathbf{r}-\mathbf{r}') = D_{\mathrm{ac}}^2 \left\langle \frac{\partial u^*(\mathbf{r})}{\partial x} \, \frac{\partial u(\mathbf{r}')}{\partial x} \right\rangle . \tag{2.69}$$

The autocorrelation of the strains can be represented in Fourier-space using the Wiener-Khinchin theorem,

$$\left\langle \frac{\partial u^*(\mathbf{r})}{\partial x} \, \frac{\partial u(\mathbf{r}')}{\partial x} \right\rangle = \sum_{\mathbf{q}} q^2 |\tilde{U}(\mathbf{q})|^2 e^{\mathrm{i}\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}')} \approx \left(\frac{L}{2\pi}\right)^3 \int\limits_{\mathbb{R}^3} q^2 |\tilde{U}(\mathbf{q})|^2 e^{\mathrm{i}\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}')} \mathrm{d}^3 q, \tag{2.70}$$

The spectral term $q^2|\tilde{U}(\mathbf{q})|^2$ can be rewritten using the displacement amplitude from Eq. (2.63),

$$q^2 |\tilde{U}(\mathbf{q})|^2 = \frac{\hbar q^2}{2\rho_{\mathrm{m}} L^3 \omega} \left(N + \frac{1}{2} \mp \frac{1}{2} \right). \tag{2.71}$$

To eliminate the $\mathbf{q}$-dependence of the term, two approximations are introduced: (i) the phonon dispersion is assumed to be linear, $\omega = v_{\mathrm{s}} q$, and (ii) the phonon energy is assumed to be small, $\hbar\omega \ll k_{\mathrm{B}}T$, so that the phonon number can be approximated as

$$N = \frac{1}{1 - e^{\frac{\hbar\omega}{k_{\mathrm{B}}T}}} \approx \frac{k_{\mathrm{B}}T}{\hbar\omega} = \frac{k_{\mathrm{B}}T}{\hbar v_{\mathrm{s}} q} \gg 1 \quad \Rightarrow \quad N + \frac{1}{2} \mp \frac{1}{2} \approx N \tag{2.72}$$

The two simplifications result in a $\mathbf{q}$-independent term for the strain spectrum,

$$q^2 |\tilde{U}(\mathbf{q})|^2 = \frac{k_{\mathrm{B}}T}{2\rho_{\mathrm{m}} L^3 v_{\mathrm{s}}^2}. \tag{2.73}$$

Finally, the square matrix element reads

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = \frac{D_{\mathrm{ac}}^2 k_{\mathrm{B}} T}{\rho_{\mathrm{m}} v_{\mathrm{s}}^2} \int\limits_{\mathbb{R}^3} |\Psi_{n,\mathbf{k}}(\mathbf{r}) \Psi_{n',\mathbf{k}'}(\mathbf{r})|^2 \mathrm{d}^3 r, \tag{2.74}$$

where an additional factor 2 was introduced in the formula to account for both absorption and emission of acoustic phonons. The square matrix element for partially confined, d-dimensional carrier gases can be written as

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = \frac{D_{\mathrm{ac}}^2 k_{\mathrm{B}} T}{\rho_{\mathrm{m}} v_{\mathrm{s}}^2 L^d} \int\limits_{\mathbb{R}^{3-d}} |\psi_{n,\mathbf{k}}(\mathbf{r}) \psi_{n',\mathbf{k}'}(\mathbf{r})|^2 \mathrm{d}^{3-d} r. \tag{2.75}$$

To simplify the treatment of ADP scattering in transport calculations, the phonon energy is ignored in Fermi's golden rule ($\hbar\omega = 0$ in Eq. (2.49)). In total, three approximations have been introduced in the model of ADP scattering:

1. A linear dispersion relation for acoustic phonons is assumed, $\omega = v_{\mathrm{s}} q$.
2. A large number of acoustic phonons is assumed and approximated as in Eq. (2.72).
3. The exchanged energy in Fermi's golden rule is neglected.

These approximations are valid for room temperature. For very cold systems, with temperatures on the order of a few tens of Kelvin or below, the assumptions do not hold. Consequently, the $\mathbf{q}$-dependence in Eq. (2.71) cannot be eliminated, and ADP scattering has to be regarded as anisotropic (and inelastic) process at very low temperatures.

2.3.2 Coulomb Scattering

Ionized impurity scattering (IIS) is of Coulomb type: The perturbing potential is due to the system's electrostatic response to a random charge distribution. The matrix element for an electrostatic potential φ is

$$H_{n,n';\mathbf{k},\mathbf{k}'} = -q_0 \int \Psi_{n,\mathbf{k}}^*(\mathbf{r}) \Psi_{n',\mathbf{k}'}(\mathbf{r}) \varphi(\mathbf{r}) \mathrm{d}V, \tag{2.76}$$

where the electrostatic potential is determined by the Poisson equation

$$\boldsymbol{\nabla} \cdot \varepsilon \boldsymbol{\nabla} \varphi + \varrho = 0. \tag{2.77}$$

For a random charge density ϱ, the potential φ is also a random function. Hence, Eq. (2.76) cannot be evaluated directly.

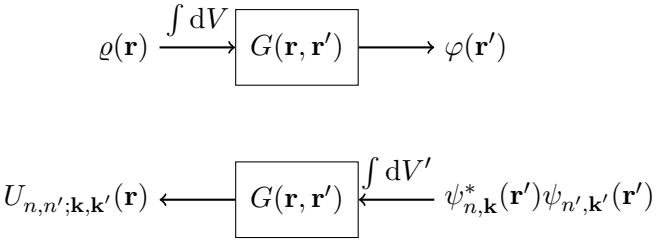


Figure 2.13: The electrostatic problem viewed as a filter; the electrostatic Green's function can be applied to a charge density to obtain the potential. Conversely, the filter can be applied in reverse to the product of two wave functions to obtain a sensitivity function that maps the influence of a charge density onto the interaction's matrix element.

The Electrostatic Green's Function

The Poisson equation is a partial differential equation that couples the potential at one point $\mathbf{r}$ to the other points $\mathbf{r}'$, thus the potential is a *correlated* random function, while the charge ϱ may be seen as an *uncorrelated* random function. The relation between charge density and potential can be viewed as a filter. The potential is the low-pass-filtered charge density, and the filter response is the electrostatic Green's function as shown in Fig. 2.13. Due to the $\mathbf{r}\leftrightarrow\mathbf{r}'$-symmetry of the Green's function, the filter can be reversed and we can rewrite Eq. (2.76) as

$$\begin{aligned} H_{n,n';\mathbf{k},\mathbf{k}'} &= -q_0 \iint \Psi^*_{n,\mathbf{k}}(\mathbf{r}')\Psi_{n',\mathbf{k}'}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')\varrho(\mathbf{r})\mathrm{d}V\mathrm{d}V' \\ &= -\int U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\varrho(\mathbf{r})\mathrm{d}V, \end{aligned} \tag{2.78}$$

where the function

$$U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r}) = q_0 \int \Psi^*_{n,\mathbf{k}}(\mathbf{r}')\Psi_{n',\mathbf{k}'}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')\mathrm{d}V' \tag{2.79}$$

is the matrix element for a single point charge at $\mathbf{r}$ or *sensitivity function* [49] for the interaction between state $(n,\mathbf{k})$ and $(n',\mathbf{k}')$.

Now the square matrix element can be evaluated. The square matrix element is in fact an ensemble average over channels with different random point charge distributions,

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = \iint U^*_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r}')\langle \varrho(\mathbf{r})\varrho(\mathbf{r}')\rangle \mathrm{d}V\mathrm{d}V'. \tag{2.80}$$

Assuming that the point charge distributions in the ensemble are uncorrelated, implying $\langle \varrho(\mathbf{r})\varrho(\mathbf{r}')\rangle \propto \delta(\mathbf{r}-\mathbf{r}')$, the square matrix element simplifies to

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = q_0^2 \int |U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})|^2 N_{\mathrm{imp}}(\mathbf{r})\mathrm{d}V, \tag{2.81}$$

where N_{imp} is the impurity concentration in the channel. For partially confined systems the squared matrix element reads

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = \frac{q_0^2}{L^d} \int |U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})|^2 N_{\mathrm{imp}}(\mathbf{r}) \mathrm{d}^{3-d} r. \tag{2.82}$$

Now the remaining question is how to compute the *sensitivity function* from Eq. (2.79) especially in the context of confined carriers. In a low-dimensional system, the cross-section coordinates where the confinement occurs ($\mathbf{r}$) and the coordinates of free propagation ($\mathbf{q} = \mathbf{k} - \mathbf{k}'$) need to be considered separately. By applying a separation analogous to the one for the wave function (Eq. (2.56)), differential operators in the transport direction, $\boldsymbol{\nabla}_{\parallel}$, can be replaced with $\mathrm{i}\mathbf{q}$. After this, the Poisson equation reads

$$[\boldsymbol{\nabla}_{\perp} \cdot \varepsilon \boldsymbol{\nabla}_{\perp} - \varepsilon q^2] \varphi_{\mathbf{q}}(\mathbf{r}) + \varrho_{\mathbf{q}}(\mathbf{r}) = 0. \tag{2.83}$$

The inverse of this operator is the reduced Green's function $G_{\mathbf{q}}(\mathbf{r}, \mathbf{r}')$. The knowledge of the actual Green's function is not required, however. What is required is a way to compute the integral in Eq. (2.79) which is equivalent to solving the equation

$$[\boldsymbol{\nabla}_{\perp} \cdot \varepsilon \boldsymbol{\nabla}_{\perp} - \varepsilon \|\mathbf{k} - \mathbf{k}'\|^2] U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r}) + q_0 \psi^*_{n,\mathbf{k}}(\mathbf{r}) \psi_{n',\mathbf{k}'}(\mathbf{r}) = 0, \tag{2.84}$$

to obtain the sensitivity function. This approach is more favorable from a numerical point of view than directly expressing the Green's function, because the Laplacian in the Poisson equation is a sparse operator and the sparsity can be leveraged for computational efficiency, whereas the Green's function is always a dense non-local operator. The approach is also more convenient from a modeling point of view, because it allows to set any kind of boundary conditions, to include channel geometry, and to include variations of the dielectric function ε.

Screening

Screening models attempt to predict a first-order response of the carrier ensemble to the Coulomb perturbation potential, since a self-consistent treatment of screening is not possible within the framework of perturbation theory employed in the derivation of Fermi's golden rule. Screening in the linear approximation is included by adding an additional term to the Poisson equation,

$$[\boldsymbol{\nabla}_{\perp} \cdot \varepsilon \boldsymbol{\nabla}_{\perp} - \varepsilon q^2] \varphi_{\mathbf{q}}(\mathbf{r}) + \varrho_{\mathbf{q}}(\mathbf{r}) + \frac{\mathrm{d}\varrho_{\mathbf{q}}}{\mathrm{d}\varphi_{\mathbf{q}}} \varphi_{\mathbf{q}} = 0. \tag{2.85}$$

Different models for screening exist, most of which are special cases of the Lindhard theory, which itself is derived from perturbation theory and the random phase approximation [50]. In the most general case, a change of potential at one point affects the carrier concentration at all other points, making the screening term non-local. Merely keeping the fully non-local screening operator in Eq. (2.85) changes the computational effort from $\mathcal{O}(n^{\frac{3}{2}})$ to $\mathcal{O}(n^2)$ for a computational mesh of n points. This does not yet

include the effort to compute the coefficients of the screening operator which can increase the effort to $\mathcal{O}(n^4)$.

Such a computationally demanding approach is not practical for a TCAD tool, where moderate run times are essential. An approximation of screening in the static limit is needed that constitutes a sparse screening operator which can be rapidly evaluated. Such an approximation can be found by looking into a related problem: When solving a coupled Schrödinger-Poisson problem (discussed in Section 2.2) an approximate response of the charge carriers to potential changes is needed to linearize the Schrödinger-Poisson equations and solve them using a Newton-Raphson scheme. Such an approximate response for confined systems has been derived in [51]; for parabolic subbands it reads

$$\frac{\mathrm{d}\varrho_{\mathbf{q}}}{\mathrm{d}\varphi_{\mathbf{q}}} = -q_0 \frac{\mathrm{d}n_{\mathbf{q}}}{\mathrm{d}\varphi_{\mathbf{q}}} \approx \sum_n |\psi_n(\mathbf{r})|^2 N_{\mathrm{C},n}^d \mathcal{F}_{\frac{d}{2}-2}\left(\frac{E_n - E_\mathrm{F}}{k_\mathrm{B}T}\right) \frac{q_0}{k_\mathrm{B}T} \tag{2.86}$$

where $N_{\mathrm{C},n}^d$ is the effective density of states for a d-dimensional carrier gas [52], and $\mathcal{F}_i$ is the complete Fermi-Dirac integral of order i. This screening operator is diagonal and can be evaluated rapidly while still taking confinement, the low-dimensional nature of the charge carriers, and the Fermi-Dirac distribution into account.

2.3.3 Polar-Optical Phonon Scattering

In III/V materials, carrier scattering by polar-optical phonons (POP) is by far the most important scattering mechanism [53]. It is caused by longitudinal-optical (LO) phonon modes, similar to ODP scattering. However, the mechanism of interaction with electrons is different. In compound semiconductors, one unit cell contains two atoms of different species, e.g. a Ga and a As atom in GaAs. Due to different valence, each of the two atoms carries a slightly different charge. When the two atoms vibrate, their displacement with respect to each other produces an oscillating electric dipole, as pictured in Fig. 2.14. Thus, in a sense, POP scattering combines the features of ODP scattering and Coulomb scattering, as shall be shown.

The semi-classical model for POP scattering in bulk crystals was first derived by H. Fröhlich [54], after whom the interaction Hamiltonian for POP scattering is named. Fröhlich's model has been reproduced in a wide range of textbooks, (e.g. [48]), but dealing with confined systems needs some adaptation of the original derivation, which shall be carried out in the remainder of this section.

The starting point is the interaction between the free electrons and the dipoles. The electrostatic potential due to a given charge density can be evaluated using the electrostatic Green's function:

$$\varphi(\mathbf{r}') = \int_{\mathbb{R}^3} G(\mathbf{r}, \mathbf{r}')\varrho(\mathbf{r})\mathrm{d}V. \tag{2.87}$$

The dipoles in the crystal lattice manifest as a random polarization field $\mathbf{P}$. To account for this in the charge density, it is decomposed into a *free* and a *bound* charge density, the latter of which is due to polarization,

$$\varrho = \varrho_\mathrm{f} + \varrho_\mathrm{b} = \varrho_\mathrm{f} - \boldsymbol{\nabla}\cdot\mathbf{P}. \tag{2.88}$$

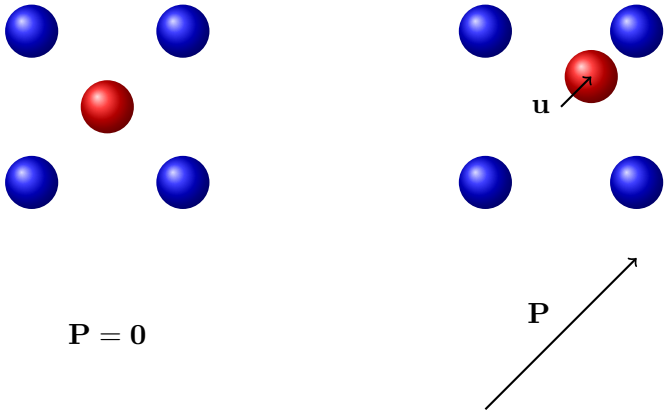


Figure 2.14: Illustration of a unit cell in a compound semiconductor; red and blue correspond to atoms of different valence carrying a different charge; when the atoms are displaced w.r.t. each other (due to a longitudinal optical phonon), they form a local variation of the polarization $\mathbf{P}$, resulting in an electric dipole.

Since the free charge density is either treated separately (when solving the global Poisson equation) or zero (in bulk crystals), only the bound contribution needs to be considered here,

$$\varphi(\mathbf{r}') = -\int_{\mathbb{R}^3} G(\mathbf{r},\mathbf{r}')\boldsymbol{\nabla}\cdot\mathbf{P}(\mathbf{r})\mathrm{d}V. \tag{2.89}$$

The matrix element for the perturbing potential is

$$\begin{aligned} H_{n,n';\mathbf{k},\mathbf{k}'} &= -q_0\int_{\mathbb{R}^3}\varphi(\mathbf{r}')\psi^*_{n,\mathbf{k}}(\mathbf{r}')\psi_{n',\mathbf{k}'}(\mathbf{r}')\mathrm{d}V' \\ &= -q_0\iint_{\mathbb{R}^3\times\mathbb{R}^3}\psi^*_{n,\mathbf{k}}(\mathbf{r}')\psi_{n',\mathbf{k}'}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')\boldsymbol{\nabla}\cdot\mathbf{P}(\mathbf{r}')\mathrm{d}V\mathrm{d}V' \\ &= \int_{\mathbb{R}^3} U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\boldsymbol{\nabla}\cdot\mathbf{P}(\mathbf{r}')\mathrm{d}V, \end{aligned} \tag{2.90}$$

where in the last step, the definition of the single-point-charge matrix element from Eq. (2.79) was used. Using the first Green's identity and the *sensitivity function* already known from Coulomb scattering (Section 2.3.2), the expression can be transformed into

$$H_{n,n';\mathbf{k},\mathbf{k}'} = -\int_{\mathscr{V}}\mathbf{P}(\mathbf{r})\cdot\boldsymbol{\nabla}U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\mathrm{d}V + \oint_{\partial\mathscr{V}} U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\mathbf{P}(\mathbf{r})\cdot\mathrm{d}\mathbf{A}, \tag{2.91}$$

where the second term vanishes for a sufficiently large volume $\mathscr{V}$, leaving the relation

$$H_{n,n';\mathbf{k},\mathbf{k}'} = -\int_{\mathbb{R}^3}\mathbf{P}(\mathbf{r})\cdot\boldsymbol{\nabla}U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\mathrm{d}V. \tag{2.92}$$

Analogously to Eq. (2.80), the squared matrix element is defined as

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = \iint \boldsymbol{\nabla} U^*_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r}) \cdot \langle \mathbf{P}^*(\mathbf{r}) \otimes \mathbf{P}(\mathbf{r}') \rangle \cdot \boldsymbol{\nabla} U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r}') \mathrm{d}V \mathrm{d}V'. \tag{2.93}$$

Assuming that the polarization fluctuation is uncorrelated, implying $\langle \mathbf{P}^*(\mathbf{r}) \otimes \mathbf{P}(\mathbf{r}') \rangle \to \|\mathbf{P}(\mathbf{r})\|^2 L^3 \delta(\mathbf{r} - \mathbf{r}')$, the expression is simplified to

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = L^3 \int \|\mathbf{P}(\mathbf{r})\|^2 \|\boldsymbol{\nabla} U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\|^2 \mathrm{d}^{3-d} r. \tag{2.94}$$

Now the amplitude of the polarization fluctuation needs to be found. The net dipole moment in a unit cell is [48]

$$\mathbf{p} = q^* \mathbf{u}, \tag{2.95}$$

where q^* is the effective charge, i.e. the charge difference between the two atoms, and $\mathbf{u}$ is the displacement vector. The polarization is given by the dipole density per unit volume,

$$\mathbf{P} = \frac{q^* \mathbf{u}}{V_{\mathrm{cell}}}, \tag{2.96}$$

with V_{cell} being the unit cell volume, yielding the square matrix element

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = L^3 \int \left(\frac{q^* \|\mathbf{u}(\mathbf{r})\|}{V_{\mathrm{cell}}} \right)^2 \|\boldsymbol{\nabla} U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\|^2 \mathrm{d}^{3-d} r. \tag{2.97}$$

For the displacement amplitude, the term derived for ODP scattering in Eq. (2.63) can be reused, giving

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = \int \frac{\hbar {q^*}^2}{2\bar{M} V_{\mathrm{cell}} \omega_{\mathrm{LO}} L^d} \left(N + \frac{1}{2} \mp \frac{1}{2} \right) \|\boldsymbol{\nabla} U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\|^2 \mathrm{d}^{3-d} r. \tag{2.98}$$

While the displacement vector $\mathbf{u}$ could be determined in the same way this was done for optical deformation potential scattering in Section 2.3.1, the direct determination of the effective charge q^* poses a difficult problem involving calculation of the electron density around the atomic cores using ab-initio methods. However, it is possible to link q^* to a different phenomenon: Most materials exhibit different permittivity at low and high (optical) frequencies. In ionic crystals such as compound semiconductors, the difference is considered to be due to the ionic contribution to susceptibility. The low-frequency permittivity of bulk semiconductors is usually a well known figure, as is the high-frequency permittivity which can be derived from the material's refractive index. Thus, the two figures can be used to determine the effective charge q^*.

The oscillating atoms in the unit cell are modeled using a classical harmonic oscillator subject to an external field [48, 55],

$$\frac{\mathrm{d}^2 \mathbf{u}}{\mathrm{d}t^2} + \omega_0^2 \mathbf{u} = \frac{\mathbf{F}}{\bar{M}} = \frac{q^* \mathbf{E}}{\bar{M}}, \tag{2.99}$$

which leads to the displacement and polarization amplitudes

$$\hat{\mathbf{u}} = \frac{q^*\hat{\mathbf{E}}}{\bar{M}} \frac{1}{\omega_0^2 - \omega^2}, \quad \hat{\mathbf{P}} = \frac{q^{*2}\hat{\mathbf{E}}}{\bar{M}V_{\text{cell}}} \frac{1}{\omega_0^2 - \omega^2}. \tag{2.100}$$

The electrical displacement field (flux density) is composed of the electronic and ionic contributions,

$$\hat{\mathbf{D}} = \epsilon_0 \left(\varepsilon_{\text{r}}^{\infty} + \frac{q^{*2}}{\bar{M}V_{\text{cell}}} \frac{1}{\omega_0^2 - \omega^2} \right) \hat{\mathbf{E}}, \tag{2.101}$$

with the dielectric function being

$$\varepsilon_{\text{r}}(\omega) = \varepsilon_{\text{r}}^{\infty} \left(1 + \frac{q^{*2}}{\bar{M}V_{\text{cell}}\varepsilon_{\text{r}}^{\infty}} \frac{1}{\omega_0^2 - \omega^2} \right), \tag{2.102}$$

as shown in Fig. 2.15. Two different modes for polar-optical waves are possible: longitudinal and transversal. Transversal waves have the electric field $\mathbf{E}$ orthogonal to the wave vector $\mathbf{q}$; thus, in the absence of magnetic fields,

$$\boldsymbol{\nabla} \times \mathbf{E} = \mathrm{i}\mathbf{q} \times \mathbf{E} = \mathbf{0} \implies \mathbf{E} = \mathbf{0}, \tag{2.103}$$

which reveals that the oscillator eigen-frequency, ω_0 where the dielectric function has a singularity, is the transversal mode frequency ω_{TO}. Longitudinal waves which are relevant for POP scattering have the wave vector parallel to the electric field; thus, for a medium with zero net charge,

$$\boldsymbol{\nabla} \cdot \mathbf{D} = \mathrm{i}\mathbf{q} \cdot \varepsilon\mathbf{E} = 0, \tag{2.104}$$

necessitating that the dielectric function becomes zero a the longitudinal mode frequency ω_{LO}, since $\mathbf{q} \cdot \mathbf{E} \neq 0$. From those two conditions and Eq. (2.102) it can be derived that $\omega_{\text{LO}}^2/\omega_{\text{TO}}^2 = \varepsilon_{\text{r}}^0/\varepsilon_{\text{r}}^{\infty}$, and eventually

$$\frac{q^{*2}}{\bar{M}V_{\text{cell}}\varepsilon_{\text{r}}^{\infty}} = \omega_{\text{LO}}^2 \left(1 - \frac{\varepsilon_{\text{r}}^{\infty}}{\varepsilon_{\text{r}}^0} \right), \tag{2.105}$$

which relates the effective charge q^* to the low- and high-frequency dielectric constants.

Equation (2.105) can now readily be inserted into Eq. (2.98) to obtain the final expression for the square matrix element of POP scattering [56, 57],

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = \frac{\hbar\omega}{2L^d} \left(N + \frac{1}{2} \pm \frac{1}{2} \right) \int \chi_{\text{ion}} \|\boldsymbol{\nabla} U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\|^2 \mathrm{d}^{3-d}\mathbf{r}, \tag{2.106}$$

where

$$\chi_{\text{ion}} = \epsilon_0 (\varepsilon_{\text{r}}^{\infty})^2 \left(\frac{1}{\varepsilon_{\text{r}}^{\infty}} - \frac{1}{\varepsilon_{\text{r}}^0} \right) \tag{2.107}$$

is the effective ionic contribution to susceptibility.

The sensitivity function $U_{n,n';\mathbf{k},\mathbf{k}'}$ can be obtained from Eq. (2.84) for *partially confined systems*. In that case the gradient operator in Eq. (2.106) is composed of two terms - one in the confinement directions, the other in the directions of free movement,

$$\boldsymbol{\nabla} = \boldsymbol{\nabla}_{\perp} + \mathrm{i}\mathbf{q}_{\parallel}, \ \|\boldsymbol{\nabla} U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\|^2 = \|\boldsymbol{\nabla}_{\perp} U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\|^2 + q^2 \|U_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r})\|^2. \tag{2.108}$$

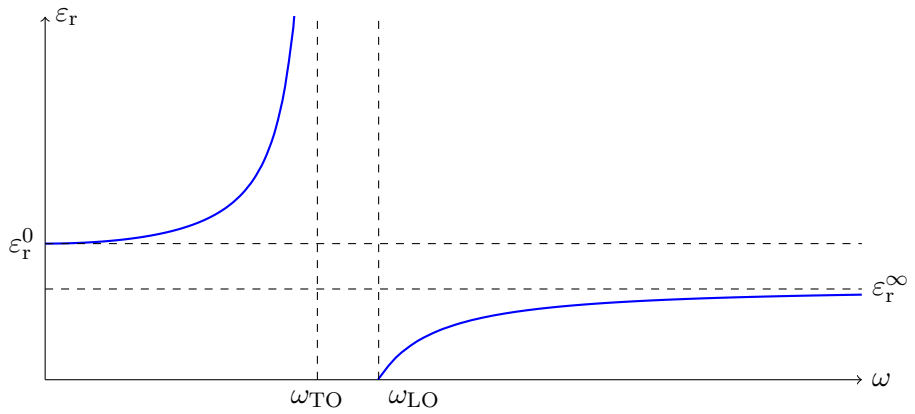


Figure 2.15: Qualitative graph of the idealized dielectric function in a ionic crystal

2.3.4 Surface and Interface Roughness Scattering

Surface or interface roughness scattering (SRS) occurs at semiconductor surfaces, hetero-interfaces, and semiconductor-dielectric interfaces. Carriers scatter off the rough surface or interface which can be seen as a fluctuation of the interface's vertical position across the interface plane. Because the its perturbation potential is static, SRS is elastic.

For two-dimensional carrier gases the most widely used model for surface and interface roughness scattering was initially formulated by Prange and Nee [58]. Here, the squared matrix element $|H_{n,n';\mathbf{k},\mathbf{k}'}|^2$ from Eq. (2.49) is effectively an average over an ensemble of rough channels, $\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle$. For a system of 2DEG carriers it reads

$$\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle = \frac{C(\mathbf{q})}{A}|F_{n,n';\mathbf{k},\mathbf{k}'}|^2, \tag{2.109}$$

where $\mathbf{q} = \mathbf{k}-\mathbf{k}'$, $C(\mathbf{q})$ is the *roughness power spectrum*, and $F_{n,n';\mathbf{k},\mathbf{k}'}$ are the form-factors due to confinement. The form-factors account for the "closeness" of the states to the interface. They are commonly approximated by the derivatives of the wave functions at the interface,

$$F_{n,n';\mathbf{k},\mathbf{k}'} = \frac{\hbar^2}{2m}\frac{\mathrm{d}\psi^*_{n,\mathbf{k}}}{\mathrm{d}x}\frac{\mathrm{d}\psi_{n',\mathbf{k}'}}{\mathrm{d}x}. \tag{2.110}$$

Roughness is a random process and hence can only be characterized by its autocorrelation function $c(\mathbf{r}) = \langle\Delta(\mathbf{r}')\Delta(\mathbf{r}'+\mathbf{r})\rangle$, where $\Delta(\mathbf{r}')$ is the actual fluctuation of the interface position. The 2D-Fourier transform of the roughness autocorrelation function is the aforementioned roughness power spectrum. The autocorrelation function is frequently modeled either as Gaussian [58]

$$c(\mathbf{r}) = \Delta^2 e^{-\frac{r^2}{\Lambda^2}}, \qquad C(\mathbf{q}) = \pi\Delta^2\Lambda^2 e^{-\frac{q^2\Lambda^2}{4}} \tag{2.111}$$

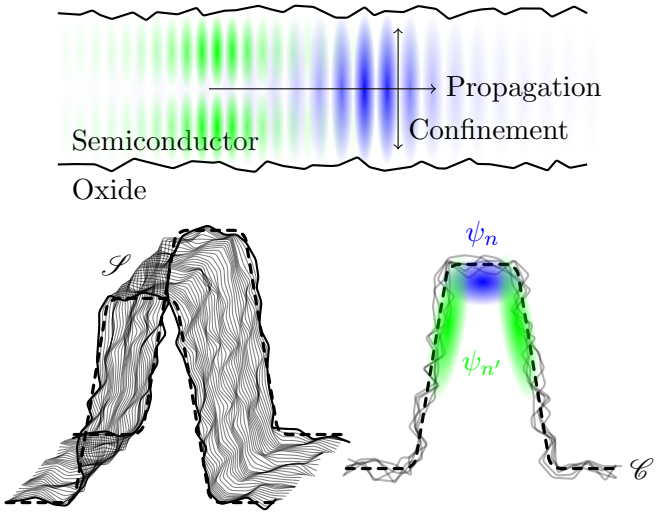


Figure 2.16: Comparison of surface roughness scattering in a planar 2DEG (top) and a non-planar 1DEG (bottom) channel; in the 2DEG case the rough planes are parallel to the direction of free propagation in which the electronic state is described as plane wave. The interacting states are selected via the plane-wave component only, resulting in a $\mathbf{q} = \mathbf{k} - \mathbf{k}'$ dependence for the square matrix element. In the 1DEG case roughness appears both along the axis and in the cross-section, thus both plane-wave and the standing wave component contribute to the state selection.

or exponential [59]

$$c(\mathbf{r}) = \Delta^2 e^{-\frac{\sqrt{2}r}{\Lambda}}, \qquad C(\mathbf{q}) = \frac{\pi\Delta^2\Lambda^2}{\left(1 + \frac{q^2\Lambda^2}{2}\right)^{\frac{3}{2}}}. \tag{2.112}$$

The roughness amplitude Δ and autocorrelation length Λ are parameters of the rough surface or interface.

The extension of SRS to the 1DEG is more involved than it was in the case of phonon scattering. The reason for this is shown in Fig. 2.16. In a 2DEG the rough interface is always parallel with the direction of propagation. Wavefunctions are separated into a 1D standing wave perpendicular and a plane wave parallel to the interface. Hence, one only needs to be concerned with the plane wave part in the derivation of the SRS square matrix element, and the standing wave enters Eq. (2.109) only as a form-factor. In a 1DEG, the wave functions are separated into a 2D standing wave in the channel cross section and a plane wave along the device axis. Now the roughness has to be taken into account not only in the plane-wave part but also in the standing wave part.

The starting point for evaluating the SRS matrix element $H_{n,n';\mathbf{k},\mathbf{k}'}$ for a 1DEG, shall be a look at the perturbing potential in Fig. 2.17. The position of an abrupt potential step of height ΔV fluctuates by the value of the function $\Delta(\mathbf{r})$. The resulting perturbing

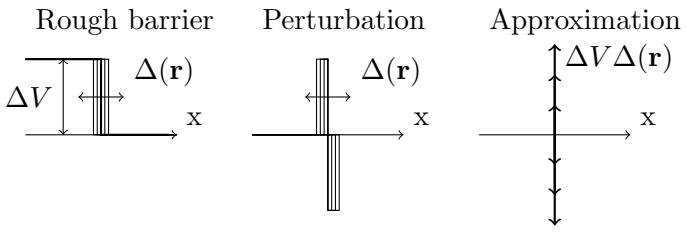


Figure 2.17: The potential across the interface is modeled as a step function. The interface roughness causes the position of the abrupt potential step to fluctuate. Subtracting the potential of an ideal surface results in the perturbing potential which is either a thin barrier or a thin well depending in the sign of the fluctuation. For roughness amplitudes much smaller than the electron wavelength the thin barrier/well can be approximated as $\Delta V \Delta(\mathbf{r})$-weighted δ-distribution.

potential is either a very thin barrier or well (depending on the sign of $\Delta(\mathbf{r})$) of height ΔV and width $\Delta(\mathbf{r})$. The perturbing potential is approximated by a weighted surface-delta-distribution $\Delta V \Delta(\mathbf{r}) \delta(\mathbf{r} \in \mathscr{S})$, where $\mathscr{S}$ represents the set of points on the ideal surface (Fig. 2.16). This allows to convert the evaluation of the matrix element $H_{n,n';\mathbf{k},\mathbf{k}'}$ from a volume integration to a surface integration,

$$H_{n,n';\mathbf{k},\mathbf{k}'} = \Delta V \int_{\mathscr{S}} \Psi^*_{n,k}(\mathbf{r}) \Psi_{n',k'}(\mathbf{r}) \Delta(\mathbf{r}) \, \mathrm{d}A. \tag{2.113}$$

This matrix element cannot be evaluated directly since $\Delta(\mathbf{r})$ is a random function. The ensemble average of the square magnitude of Eq. (2.113), however, can be evaluated:

$$\left\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \right\rangle = \iint_{\mathscr{S}} \mathrm{d}A \mathrm{d}A' \Psi_{n,k}(\mathbf{r}) \Psi^*_{n',k'}(\mathbf{r}) \Psi^*_{n,k}(\mathbf{r}') \Psi_{n',k'}(\mathbf{r}') \Delta V^2 \left\langle \Delta(\mathbf{r}) \Delta(\mathbf{r}') \right\rangle. \tag{2.114}$$

So far no assumptions about the electron states $\Psi_{n,k}$ have been made. Recalling Eq. (2.56), the electron states are decomposed into a two-dimensional bound state in the cross-section and a plane wave along the channel axis,

$$\Psi_{n,\mathbf{k}}(\mathbf{r}_\perp) = \psi_{n,\mathbf{k}}(\mathbf{r}) \frac{1}{\sqrt{L}} e^{\mathrm{i}kz} = \psi_{n,\mathbf{k}}(x,y) \frac{1}{\sqrt{L}} e^{\mathrm{i}kz}. \tag{2.115}$$

Using this separation ansatz, Eq. (2.114) can be rewritten as

$$\left\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \right\rangle = \frac{1}{L^2} \iint_{\mathscr{C}} \int_0^L f_{n,n';\mathbf{k},\mathbf{k}'}(s) f^*_{n,n';\mathbf{k},\mathbf{k}'}(s') \, e^{\mathrm{i}(k-k')(z-z')} \left\langle \Delta(\mathbf{r}) \Delta(\mathbf{r}') \right\rangle \mathrm{d}z \mathrm{d}z' \mathrm{d}s \mathrm{d}s'. \tag{2.116}$$

The integration across surface $\mathscr{S}$ was separated into integrations along curve $\mathscr{C}$, i.e. the intersection of $\mathscr{S}$ with the cross-section plane, and a normalization length L along the channel direction; s denotes the path coordinate along the curve $\mathscr{C}$ and z the axial coordinate. We introduced the form-functions $f_{n,n';\mathbf{k},\mathbf{k}'}(s)$ which are defined as

$$f_{n,n';\mathbf{k},\mathbf{k}'}(s) = \psi^*_{n,k}\psi_{n',k'}\Delta V. \tag{2.117}$$

The effect of different effective masses in the materials on either side of $\mathscr{S}$ can be included in the form-functions as

$$\begin{aligned} f_{n,n';\mathbf{k},\mathbf{k}'}(s) = {} & \psi^*_{n,k}\psi_{n',k'}(V_- - V_+) \\ & - \frac{\hbar^2}{2}\boldsymbol{\nabla}\psi^*_{n,k;-}\cdot\mathbf{m}_-^{-1}\cdot\boldsymbol{\nabla}\psi_{n',k';-} \\ & + \frac{\hbar^2}{2}\boldsymbol{\nabla}\psi^*_{n,k;+}\cdot\mathbf{m}_+^{-1}\cdot\boldsymbol{\nabla}\psi_{n',k';+}, \end{aligned} \tag{2.118}$$

where the subscripts $_+$ and $_-$ indicate limits at either side of $\mathscr{S}$. In the limit of high potential barriers (e.g. dielectrics), the wave functions $\psi_{n,k}$ do not penetrate from one medium into the other but vanish at the interface. In that case the expression in Eq. (2.118) can be approximated by

$$f_{n,n';\mathbf{k},\mathbf{k}'}(\mathbf{r}) \approx \frac{\hbar^2}{2}\boldsymbol{\nabla}\psi^*_{n,k}\cdot\mathbf{m}^{-1}_{\text{well}}\cdot\boldsymbol{\nabla}\psi_{n',k'}. \tag{2.119}$$

The autocorrelation function $\big\langle\Delta(\mathbf{r})\Delta(\mathbf{r}')\big\rangle =: c(\mathbf{r})$ in Eq. (2.116) can be represented as inverse 2D Fourier transform of the *roughness power spectrum*,

$$c(\mathbf{r}) = \frac{1}{4\pi^2}\iint\limits_{\mathbb{R}} C(\mathbf{q})e^{\mathrm{i}q_\perp(s-s')}e^{\mathrm{i}q_\parallel(z-z')}\mathrm{d}q_\perp\mathrm{d}q_\parallel, \tag{2.120}$$

separating the roughness "wave vector" $\mathbf{q}$ into an axial component $q_\parallel$ and a component $q_\perp$ along $\mathscr{C}$. Inserting Eq. (2.120) into Eq. (2.116), one arrives at

$$\begin{aligned} \big\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\big\rangle = \frac{1}{4\pi^2L^2}\iint\limits_{\mathscr{C}}\mathrm{d}s\mathrm{d}s'\iint\limits_{\mathbb{R}}\mathrm{d}q_\perp\mathrm{d}q_\parallel\int\limits_{-\infty}^{\infty}\!\!\int\limits_{-\infty}^{\infty}\mathrm{d}z\mathrm{d}z' \\ f_{n,n';\mathbf{k},\mathbf{k}'}(s)f^*_{n,n';\mathbf{k},\mathbf{k}'}(s')C(\mathbf{q})e^{\mathrm{i}q_\perp(s-s')}e^{\mathrm{i}(k-k'+q_\parallel)(z-z')}. \end{aligned} \tag{2.121}$$

The double axial integration of the plane wave term $e^{\mathrm{i}(k-k'+q_\parallel)(z-z')}$ for a sufficiently large L leads to

$$\begin{aligned} \int\limits_0^L\!\!\int\limits_0^L e^{\mathrm{i}(k-k'+q_\parallel)(z-z')}\mathrm{d}z\mathrm{d}z' &= \int\limits_0^L\mathrm{d}z\int\limits_0^L e^{\mathrm{i}(k-k'+q_\parallel)z''}\mathrm{d}z'' \\ &\approx L\lim_{L\to\infty}\int\limits_0^L e^{\mathrm{i}(k-k'+q_\parallel)z''}\mathrm{d}z'' = 2\pi L\delta(k-k'+q_\parallel) \end{aligned} \tag{2.122}$$

allowing to simplify the previous equation to

$$\left\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \right\rangle = \frac{1}{2\pi L} \iint\limits_{\mathscr{C}} \mathrm{d}s\mathrm{d}s' \int\limits_{\mathbb{R}} \mathrm{d}q_\perp f_{n,n';\mathbf{k},\mathbf{k}'}(s) f^*_{n,n';\mathbf{k},\mathbf{k}'}(s') C(\mathbf{q}) e^{\mathrm{i}q_\perp(s-s')}. \quad (2.123)$$

A change of variables $s' - s =: s''$ gives

$$\left\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \right\rangle = \frac{1}{2\pi L} \int\limits_{\mathbb{R}} C(\mathbf{q})\mathrm{d}q_\perp \int\limits_{\mathscr{C}} \left[\int\limits_{\mathscr{C}} f_{n,n';\mathbf{k},\mathbf{k}'}(s) f^*_{n,n';\mathbf{k},\mathbf{k}'}(s+s'')\mathrm{d}s \right] e^{\mathrm{i}q_\perp s''} \mathrm{d}s''. \quad (2.124)$$

The term in square brackets is the autocorrelation of the form-functions $f_{n,n';\mathbf{k},\mathbf{k}'}(s)$ and the integration surrounding it is a Fourier transform $s \mapsto q_\perp$. Using the Wiener-Khinchin theorem the Fourier transform of the autocorrelation of $f_{n,n';\mathbf{k},\mathbf{k}'}(s)$ can be expressed as square magnitude of its Fourier transform $F_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)$ obtaining the final expression for the square matrix element [49, 60, 61],

$$\left\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \right\rangle = \frac{1}{2\pi L} \int\limits_{\mathbb{R}} |F_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)|^2 C(\mathbf{q})\mathrm{d}q_\perp. \quad (2.125)$$

A few assumptions are contained within this last step of our derivation:

1. For closed curves $\mathscr{C}$ (e.g. in a gate-all-around channel) the Fourier transform is in fact a Fourier series expansion.

2. For open curves $\mathscr{C}$, such as the tri-gate channel in Fig. 2.16 the Fourier transform is effectively *windowed* by the finite length of the curve $\mathscr{C}$. However, due to electrostatic confinement the wave functions, and hence the form functions decay exponentially towards both ends of $\mathscr{C}$. The windowing effect is therefore negligible.

3. The roughness power spectrum is assumed to be isotropic, $C(\mathbf{q}) = C(q)$.

The integral in Eq. (2.125) represents momentum conservation in the cross-section plane. It can be summarized that in a planar geometry with a 2DEG, carrier momentum conservation is characterized by a $\delta(\mathbf{k} - \mathbf{k}' + \mathbf{q})$ term. In a non-planar structure with a 1DEG, there is still a $\delta(k_\parallel - k'_\parallel + q_\parallel)$ term for the axial direction. However, the cross-section momentum conservation is not sharply defined but is now accounted for by the integral in Eq. (2.125).

2.3.5 Alloy Disorder Scattering

Alloy disorder scattering occurs in alloyed semiconductors, such as binary (e.g. $Si_{1-x}Ge_x$), ternary (e.g. $In_{1-x}Ga_xAs$) and quaternary alloys. This work follows the discussion in [62],

in which alloy disorder scattering is considered as an elastic, isotropic intra-valley process modeled using the semi-empirical expression

$$\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle = \frac{U_{\text{alloy}}^2 V_{\text{cell}}}{L^d} \int x(1-x) |\psi_{n,\mathbf{k}}(\mathbf{r})|^2 |\psi_{n',\mathbf{k}'}(\mathbf{r})|^2 \mathrm{d}^{3-d} r, \tag{2.126}$$

with $V_{\text{cell}} = a^3/8$ being the unit cell volume. The square matrix element depends on the local material composition x; for pure materials, x is either 0 or 1, which results in the transition rate being zero. The effective scattering potential U_{alloy} is used to fit the measured data.

2.4 Transport and Mobility

This section looks at carrier transport, which, in the semi-classical framework, is governed by the Boltzmann transport equation (BTE).

$$\frac{\partial f_n(\mathbf{r},\mathbf{k},t)}{\partial t} + \mathbf{v}_n(\mathbf{k}) \cdot \boldsymbol{\nabla}_{\mathbf{r}} f_n(\mathbf{r},\mathbf{k},t) + \frac{\mathbf{F}(\mathbf{r})}{\hbar} \cdot \boldsymbol{\nabla}_{\mathbf{k}} f_n(\mathbf{r},\mathbf{k},t) = - \left[\frac{\partial f_n(\mathbf{r},\mathbf{k},t)}{\partial t} \right]_{\text{scatt.}} \tag{2.127}$$

The solution variable of the BTE is the *distribution function* $f_n(\mathbf{r},\mathbf{k},t)$. The BTE is based on semi-classical mechanics, treating carriers as particles. The wave properties of the carriers were appreciated in the previous section to calculate the subband structure, the probability density, and the transition rates due to scattering. These steps essentially provide the coefficients for the free streaming operator on the left hand side and the scattering operator on the right-hand side of Eq. (2.127).

In the most general case, the scattering operator reads

$$\begin{aligned} \left[\frac{\partial f_n(\mathbf{r},\mathbf{k},t)}{\partial t} \right]_{\text{scatt.}} = & \sum_{n',\mathbf{k}'} S_{n',n}(\mathbf{k}',\mathbf{k}) f_{n'}(\mathbf{k}') [1 - f_n(\mathbf{k})] \\ & - S_{n,n'}(\mathbf{k},\mathbf{k}') f_n(\mathbf{k}) [1 - f_{n'}(\mathbf{k}')], \end{aligned} \tag{2.128}$$

where $S_{n,n'}(\mathbf{k},\mathbf{k}')$ denotes the transition rate given by Fermi's golden rule in Eq. (2.49). Scattering does not affect the position of a particle and therefore the spatial coordinate of the distribution function has been omitted in above equation.

2.4.1 Linearizing the Boltzmann Transport Equation

For a three-dimensional problem, the BTE constitutes a seven-dimensional integro-differential equation. Solving such an equation seems like a daunting task, but one does not necessarily need to include all seven dimensions. For confined systems, the $\mathbf{k}$-space is reduced from three to two (2DEG) or one (1DEG) dimension. For stationary problems, the time-dependence can dropped to obtain the time-independent or *steady-state* version of the BTE,

$$\mathbf{v}_n(\mathbf{k}) \cdot \boldsymbol{\nabla}_{\mathbf{r}} f_n(\mathbf{r},\mathbf{k}) + \frac{\mathbf{F}(\mathbf{r})}{\hbar} \cdot \boldsymbol{\nabla}_{\mathbf{k}} f_n(\mathbf{r},\mathbf{k}) = - \left[\frac{\partial f_n(\mathbf{r},\mathbf{k})}{\partial t} \right]_{\text{scatt.}}. \tag{2.129}$$

The equation is further simplified when a constant force field $\mathbf{F}$ is assumed and the spatial coordinate can be dropped,

$$\frac{\mathbf{F}}{\hbar} \cdot \boldsymbol{\nabla}_{\mathbf{k}} f_n(\mathbf{k}) = -\left[\frac{\partial f_n(\mathbf{k})}{\partial t}\right]_{\text{scatt.}}. \tag{2.130}$$

The BTE for a constant field can be used to determine macroscopic properties of a material or a channel based on microscopic information about electronic structure and carrier scattering. This equation can now be linearized to obtain the linear response of the carrier ensemble to a *small* force field $\mathbf{F}$. To do this, the distribution function is separated into an equilibrium part and a first-order response part,

$$f_n(\mathbf{k}) = f^0(E) + f_n^1(\mathbf{k}), \tag{2.131}$$

where the equilibrium part only depends on energy as stated in Eq. (2.42). Inserting the expression back into Eq. (2.130), one obtains

$$\frac{\mathbf{F}}{\hbar} \cdot \boldsymbol{\nabla}_{\mathbf{k}} E_n(\mathbf{k}) \frac{\mathrm{d}f^0(E)}{\mathrm{d}E} + \underbrace{\frac{\mathbf{F}}{\hbar} \cdot \boldsymbol{\nabla}_{\mathbf{k}} f_n^1(\mathbf{k})}_{\text{neglected}} = \underbrace{-\left[\frac{\partial f^0}{\partial t}\right]_{\text{scatt.}}}_{=0} - \left[\frac{\partial f_n^1}{\partial t}\right]_{\text{scatt.}}, \tag{2.132}$$

The second-order acceleration term on the left-hand side is negligible for small force fields. The scattering operator on the right-hand side contains the equilibrium distribution in its kernel. In other words: The equilibrium distribution is unaffected by scattering and it holds that

$$\left[\frac{\partial f^0}{\partial t}\right]_{\text{scatt.}} = 0. \tag{2.133}$$

Omitting these terms, we arrive at the *linearized Boltzmann transport equation* (LBTE),

$$\mathbf{F}(\mathbf{r}) \cdot \mathbf{v}_n(\mathbf{k}) \frac{\mathrm{d}f^0(E)}{\mathrm{d}E} = -\left[\frac{\partial f_n^1}{\partial t}\right]_{\text{scatt.}}, \tag{2.134}$$

where the definition of the group velocity, $\mathbf{v}_n = \boldsymbol{\nabla}_{\mathbf{k}} E_n / \hbar$, has been used. The solution variable is now the *linear distribution response*.

Linearization also applies to the scattering operator. Here, it is makes sense to split the scattering operator into an elastic an an inelastic part,

$$\left[\frac{\partial f_n}{\partial t}\right]_{\text{scatt.}} = \left[\frac{\partial f_n}{\partial t}\right]_{\text{scatt.}}^{\text{el}} + \left[\frac{\partial f_n}{\partial t}\right]_{\text{scatt.}}^{\text{inel}}. \tag{2.135}$$

For *elastic* processes the transition rate in Eq. (2.128) is *symmetric*, $S_{n,n'}(\mathbf{k}, \mathbf{k}') = S_{n',n}(\mathbf{k}', \mathbf{k})$, and the elastic scattering operator simplifies to

$$\left[\frac{\partial f_n}{\partial t}\right]_{\text{scatt.}}^{\text{el}} = \sum_{n', \mathbf{k}'} S_{n,n'}(\mathbf{k}, \mathbf{k}') \left[f_n^1(\mathbf{k}) - f_{n'}^1(\mathbf{k}')\right]. \tag{2.136}$$

For *inelastic* processes the scattering operator is *not symmetric*, and the full expression in Eq. (2.128) must be used. The factors $[1 - f_{n'}(\mathbf{k}')]$ and $[1 - f_n(\mathbf{k})]$, which take the Pauli-principle of exclusion into account, make the inelastic scattering operator non-linear and it needs to be linearized as well.

$$\begin{aligned}\left[\frac{\partial f_n^1}{\partial t}\right]_{\text{scatt.}}^{\text{inel}} = \sum_{n',\mathbf{k}'} & S_{n,n'}(\mathbf{k},\mathbf{k}') \left\{ f_n^1(\mathbf{k})[1 - f^0(E_{n'}(\mathbf{k}'))] - f^0(E_n(\mathbf{k}))[1 - f_{n'}^1(\mathbf{k}')] \right\} \\ & - S_{n',n}(\mathbf{k}',\mathbf{k}) \left\{ f_{n'}^1(\mathbf{k}')[1 - f^0(E_n(\mathbf{k}))] - f^0(E_{n'}(\mathbf{k}'))[1 - f_n^1(\mathbf{k})] \right\}.\end{aligned} \tag{2.137}$$

2.4.2 Conductivity and Mobility Extraction

The linearized Boltzmann transport equation (LBTE) can now be utilized to extract macroscopic properties of a channel such as the conductivity or mobility, which relate current and carrier velocity to a small electric field. Besides electric fields other driving forces lead to other macroscopic properties: Using for example a temperature gradient as driving force leads to the Seebeck coefficient.

The General Case

With a semi-classical distribution function available, the electrical current density in a system of carriers is determined by

$$\mathbf{J}_n = \frac{\mathrm{q}g}{(2\pi)^d} \int_{\mathbb{R}^d} \mathbf{v}_n(\mathbf{k}) f_n^{\mathrm{a}}(\mathbf{k}) \mathrm{d}^d\mathbf{k}, \tag{2.138}$$

where $\mathrm{q} = \pm q_0$ denotes the carrier charge, g the degeneracy due to to spin and valley multiplicity, $\mathbf{v}_n(\mathbf{k})$ the group velocity of the n^{th} subband, and f_n^{a} the *anti-symmetric* part of the distribution function f_n. In the limit of low driving force, the asymmetric part is given by the linear distribution response f_n^1 with respect to the field $\mathbf{E}$. The linear distribution response depends on the electric field via the LBTE from Eq. (2.134),

$$\left[\frac{\partial f_n^1}{\partial t}\right]_{\text{scatt.}} = \mathrm{q}\mathbf{E} \cdot \mathbf{v}_n(\mathbf{k}) \frac{\mathrm{d}f^0(E)}{\mathrm{d}E}. \tag{2.139}$$

with the force replaced by $\mathrm{q}\mathbf{E}$. The response f_n^1 is linear with respect to the modulus of the electric field $\|\mathbf{E}\|$. Consequently the linearity-relation also applies to the current from Eq. (2.138). The modulus of the electric field can be factored out of the equation to give

$$\left[\frac{\partial \tilde{f}_n^1}{\partial t}\right]_{\text{scatt.}} = \mathrm{q}\mathbf{e}_{\mathbf{E}} \cdot \mathbf{v}_n(\mathbf{k}) \frac{\mathrm{d}f^0(E)}{\mathrm{d}E}, \tag{2.140}$$

where $f_n^1 = \tilde{f}_n^1 \|\mathbf{E}\|$. The field is assumed to point in the direction given by the unit vector $\mathbf{e}_{\mathbf{E}}$. For a different field direction $\mathbf{e}'_{\mathbf{E}}$, a different version of Eq. (2.140) needs to be solved and a different response $\tilde{f}_n^{1\prime}$ needs to be inserted into Eq. (2.138). While the distribution

response and current density are linear with respect to the modulus of the field, their dependence on the direction of the field $\mathbf{e_E}$ is non-linear. An anisotropic linear relation,

$$\mathbf{J}_n = \boldsymbol{\sigma}_n \cdot \mathbf{E}, \tag{2.141}$$

is thus not sufficient to represent the angular dependence between $\mathbf{E}$ and $\mathbf{J}_n$. For a complete representation, one could expand the function $\boldsymbol{\sigma}_n(\mathbf{E})$ using spherical harmonics in 3D or Fourier harmonics in 2D. Equation (2.141) would be a first-order approximation to such an expansion.

For practical purposes however, one is mainly interested in the current flow projected along the same axis along which the field is applied. In this case one can define the direction-dependent conductivity

$$\sigma_n[\mathbf{e_E}] := \frac{\mathbf{e_E} \cdot \mathbf{J}_n[\mathbf{e_E}]}{\|\mathbf{E}\|} = \frac{\mathrm{q}g}{(2\pi)^d} \int\limits_{\mathbb{R}^d} \mathbf{e_E} \cdot \mathbf{v}_n(\mathbf{k}) \tilde{f}_n^1[\mathbf{e_E}](\mathbf{k}) \mathrm{d}^d\mathbf{k}, \tag{2.142}$$

where the distribution response, the current density, and the conductivity are labeled using the field unit vector in brackets, $[\mathbf{e_E}]$, denoting them to be associated with this particular field direction only. From this, mobility is derived as

$$\mu_n[\mathbf{e_E}] := \frac{\sigma_n[\mathbf{e_E}]}{\mathrm{q}n} = \frac{\int \mathbf{e_E} \cdot \mathbf{v}_n(\mathbf{k}) \tilde{f}_n^1[\mathbf{e_E}](\mathbf{k}) \mathrm{d}^d\mathbf{k}}{\int f^0(E(\mathbf{k})) \mathrm{d}^d\mathbf{k}}. \tag{2.143}$$

The Effective Mass Case

The complications involving the angular dependence of conductivity discussed above can be simplified for parabolic subbands. This is facilitated by introducing a microscopic relaxation time tensor $\boldsymbol{\tau}$ [63], such that

$$f_n^1(\mathbf{k}) = -\mathrm{q}\mathbf{v}_n(\mathbf{k}) \cdot \boldsymbol{\tau}_n(\mathbf{k}) \cdot \mathbf{E} \frac{\mathrm{d}f^0}{\mathrm{d}E}, \tag{2.144}$$

with f^0 denoting the equilibrium Fermi-Dirac distribution. Inserting Eq. (2.144) into Eq. (2.138), the conductivity tensor can be obtained:

$$\boldsymbol{\sigma}_n = -\frac{\mathrm{q}^2}{(2\pi)^d} \int\limits_{\mathbb{R}^d} \mathbf{v}_n(\mathbf{k}) \otimes \mathbf{v}_n(\mathbf{k}) \boldsymbol{\tau}_n(\mathbf{k}) \frac{\mathrm{d}f^0}{\mathrm{d}E} \mathrm{d}^d k. \tag{2.145}$$

For parabolic bands, the integration over $\mathbf{k}$ is replaced by an integration over energy to obtain

$$\boldsymbol{\sigma}_n = -\mathrm{q}^2 \ \mathbf{m}_n^{-1} \cdot \frac{2}{d} \int\limits_{E_n}^{\infty} \boldsymbol{\tau}_n(E) \frac{\partial f^0}{\partial E} E g_n^d(E) \mathrm{d}E, \tag{2.146}$$

using the d-dimensional density of states $g_n^d(E)$. Note that the relaxation time in both Eq. (2.144) and Eq. (2.146) is a tensor. This is necessary to correctly account

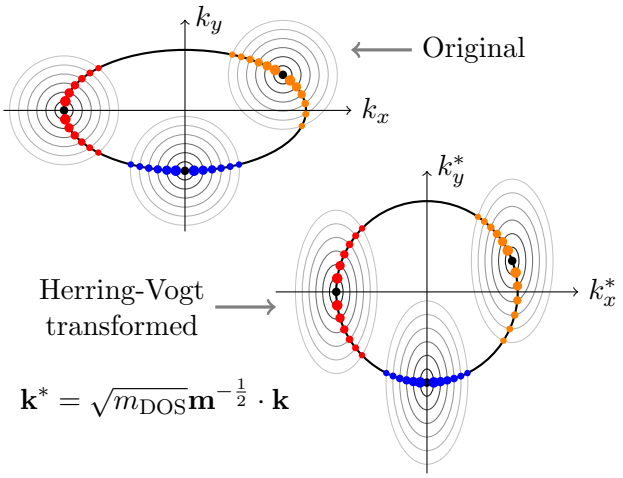


Figure 2.18: Anisotropic scattering in anisotropic subbands: Different positions of the initial state in **k**-space result in different probability distributions of final states. Applying a Herring-Vogt transform eliminates band anisotropy but distorts the **q**-dependence of the scattering process at the same time. In both cases the relaxation time can only be described as a tensor.

for anisotropic scattering processes, such as Coulomb scattering or surface-roughness scattering, which prefer small deflections between initial and final momenta as illustrated in Fig. 2.18. Due to symmetry, the tensors $\mathbf{m}_n$ and $\boldsymbol{\tau}_n(E)$ have the same principal directions, which implies that they commute. The expression in Eq. (2.146) can be separated into principal components,

$$\sigma_{n,\xi} = -\mathrm{q}^2 \, \frac{1}{m_{n,\xi}} \cdot \frac{2}{d} \int_{E_n}^{\infty} \tau_{n,\xi}(E) \frac{\partial f^0}{\partial E} E g_n^d(E) \mathrm{d}E, \tag{2.147}$$

where ξ denotes each principal direction. Finally, subband mobility $\boldsymbol{\mu}_n$ and total mobility $\boldsymbol{\mu}$ are computed as

$$\boldsymbol{\mu}_n = \frac{\boldsymbol{\sigma}_n}{q_0 n_n}, \qquad \boldsymbol{\mu} = \frac{1}{q_0 n} \sum_n \boldsymbol{\sigma}_n. \tag{2.148}$$

The linear distribution response f_n^1 is governed by the linearized Boltzmann transport equation. Inserting Eq. (2.144), we obtain an equation for the microscopic relaxation time,

$$\sum_{n',\mathbf{k}'} S_{n,n'}(\mathbf{k},\mathbf{k}') \left[\mathbf{v}_n(\mathbf{k}) \cdot \boldsymbol{\tau}_n(\mathbf{k}) \cdot \mathbf{E} - \mathbf{v}_{n'}(\mathbf{k}') \cdot \boldsymbol{\tau}_{n'}(\mathbf{k}') \cdot \mathbf{E} \right] = \mathbf{v}_n(\mathbf{k}) \cdot \mathbf{E}. \tag{2.149}$$

Fortunately, for parabolic bands, the different principal components τ_ξ are not coupled

by Eq. (2.149), so we can write

$$\sum_{n',\mathbf{k}'} S_{n,n'}(\mathbf{k},\mathbf{k}')\left[\tau_{n,\xi}(E)\frac{\hbar k_\xi}{m_{n,\xi}} - \tau_{n',\xi}(E)\frac{\hbar k'_\xi}{m_{n',\xi}}\right] = \frac{\hbar k_\xi}{m_{n,\xi}}. \tag{2.150}$$

for each principal direction ξ. The symbol k_ξ denotes the projection of $\mathbf{k}$ along the principal direction ξ. We recall from Eq. (2.49) that for elastic processes $S_{n,n'}(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\delta(E(\mathbf{k}) - E(\mathbf{k}'))$. The energy-conserving δ-distribution decouples the scattering operator in Eq. (2.150) for different energies. Since $\tau_{n,\xi}$ itself depends on the energy but not the direction of $\mathbf{k}$, a system of equations can be formulated for every energy as

$$L_n\tau_{n,\xi} - M_{n,n'}\tau_{n',\xi} = 1, \tag{2.151}$$

where the coefficients are

$$L_n = \sum_{n'} g_n^d(E)J_{n,n',\xi} \tag{2.152}$$

$$M_n = g_n^d(E)\sqrt{\frac{m_{n,\xi}(E-E_n)}{m_{n',\xi}(E-E_{n'})}}J'_{n,n',\xi}. \tag{2.153}$$

The integrals $J_{n,n',\xi}$ and $J'_{n,n',\xi}$ are defined as

$$J_{n,n',\xi} = \frac{d}{\Omega_d}\int\limits_{\Omega_d\times\Omega_d}\frac{\pi}{\hbar}|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\cos^2\vartheta_\xi \mathrm{d}\Omega'_d\mathrm{d}\Omega_d \tag{2.154}$$

$$J'_{n,n',\xi} = \frac{d}{\Omega_d}\int\limits_{\Omega_d\times\Omega_d}\frac{\pi}{\hbar}|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\cos\vartheta_\xi\cos\vartheta'_\xi \mathrm{d}\Omega'_d\mathrm{d}\Omega_d, \tag{2.155}$$

with ϑ_ξ denoting the angle between the Herring-Vogt transformed [64] vector $\mathbf{k}^*$ and the principal direction ξ (see Fig. 2.18),

$$\cos\vartheta_\xi = \frac{\mathbf{k}^*\cdot\mathbf{e}_\xi}{k^*}. \tag{2.156}$$

Ω_d denotes the surface of a d-dimensional "unit sphere" which measures 2 for a 1DEG and 2π for a 2DEG.[1] The integrals in Eqs. (2.154) and (2.155) can be understood as

$$\int\limits_{\Omega_d\times\Omega_d} y(\vartheta,\vartheta')\mathrm{d}\Omega'_d\mathrm{d}\Omega_d = \begin{cases} y(0,0)+y(0,\pi)+y(\pi,0)+y(\pi,\pi) & : \text{1DEG} \\ \int\limits_0^{2\pi}\int\limits_0^{2\pi} y(\vartheta,\vartheta')\mathrm{d}\vartheta'\mathrm{d}\vartheta & : \text{2DEG.} \end{cases} \tag{2.157}$$

[1] A two-dimensional "unit sphere" is a unit circle. A one-dimensional "unit sphere" are the two points 1 and -1 on the number line; the two points of the one-dimensional "unit sphere" are denoted by their the pseudo-angles 0 and π.

For inelastic scattering, energy changes to a value $\hbar\omega$ above or below E, requiring to take the density of states at $E \pm \hbar\omega$ and the reduction of available states due to Pauli exclusion into account. To properly account for the change in energy, Eq. (2.152) needs to be modified to

$$L_n = \sum_{n'} g_n^d(E \pm \hbar\omega) J_{n,n',\xi} \frac{1 - f^0(E \pm \hbar\omega)}{1 - f^0(E)}. \tag{2.158}$$

CHAPTER 3 Computational Foundation

Contents

3.1 Model Concept 55
3.2 Data Level 56
3.2.1 Geometry and Topology 56
3.2.2 Data Storage 57
3.2.3 Discretization 57
3.2.4 I/O and Configuration 61
3.3 Modeling Level 61
3.3.1 Expressions 61
3.3.2 Problem Specification and Assembly 63
3.3.3 Contour Integration 66
3.4 Algebraic Level 67
3.4.1 Abstraction of Linear Operations 67
3.4.2 Working with Expressions 67
3.4.3 Solvers 68
3.4.4 Fast Fourier Transform 70
3.5 Extension Through Modules 70
3.5.1 Module Loading 70
3.5.2 Software Development Kit 71
3.5.3 Literate Modeling 71

This chapter lays out the computational infrastructure used to implement all the models required to tackle the physics explained in Chapter 2. The infrastructure has been implemented in the course of this work as part of the Vienna Schrödinger-Poisson (VSP) simulation framework [52], which has been made available commercially within GTS Framework [65, 66]. The computational infrastructure comprises building blocks which form the base for the implementation of the models described in Chapter 4 which are used to simulate the nanoelectronic devices presented in Chapter 5.

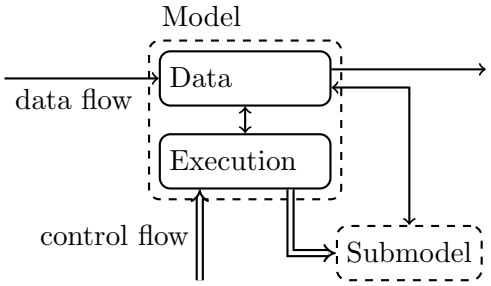


Figure 3.1: Conceptual schematic of a model in VSP; solid arrows represent data flow, double arrows control flow.

3.1 Model Concept

The software design of VSP is based on five concepts which shall be elaborated in detail in the following sections:

1. Flexibility
2. Automation
3. Efficiency
4. Consistency
5. Customization

From a user point of view *Flexibility* means that VSP can be conveniently and intuitively configured to perform a vast multitude of simulation tasks. Models serve as building blocks of a simulation flow. *Automation* is closely related to *Flexibility* and means providing interfaces for parameterizing the simulation flow, including scriptable input decks, parameter sweeps, and interfacing with meshing tools, device simulators etc. *Automation* also allows VSP to be used in automated device design optimization and parameter calibration. *Efficiency* is an enabling concept for *Automation*; strong emphasis has been put on *Efficiency* in the design of VSP. The model design in VSP is based on broad general approaches to solving certain classes of problems and avoids functions specifically implemented to cover special cases. *Consistency* is a result of those general approaches; it ensures consistent results with respect to changes in dimensionality, materials, or computational methods used in the simulation. Finally, *Customization* allows the VSP to be extended in several ways, like adding new models, materials, or numerical libraries.

The core design philosophy in VSP can be summarized as follows: *Everything is a model.* In the sense of VSP's design a model is an object similar to a class in computer programming. A model can be instantiated and the instance can be invoked, which may result in equations being solved, expressions evaluated and so forth. A model may have submodels for certain sub-tasks; the model may invoke its submodels during execution.

Every model instance can store data relevant to the model and may expose the data, so that it can be passed to and from other model instances. Figure 3.1 shows the architecture of a VSP model.

Model data is organized into attributes. Three types of attributes are considered: *parameters*, *properties*, and *quantities*. *Parameters* represent numerical, or non-physical entities: error tolerance, number of iteration steps, and so on. *Properties* represent concentrated physical quantities: contact voltages, subband energies, etc. *Quantities* represent distributed physical quantities – in real space (electrostatic potential, charge density) and **k**-space (e.g. band/subband structure). All attributes have identifiers that can be used to access them. *Properties* and *quantities* have physical units.

The VSP modeling framework provides a number of facilities tailored to tackle computational problems encountered in nanoelectronic device simulation. The main purpose of these facilities is to foster code reuse which keeps both development time low and reduces the probability of introducing errors during development. The facilities are structured in three *levels of abstraction*: (i) the data level, (ii) the modeling level, and (iii) the algebraic level.

3.2 Data Level

Data and geometry form the foundation of numerical device modeling and simulation. This level can be regarded as the *low level* of VSP's infrastructure, whereas the modeling level to be described in the next section would be the *high level*.

3.2.1 Geometry and Topology

The simulation domains of VSP (in real space or **k**-space) are called *devices*. A *device* is organized in *segments*. *Segments* consist of *elements* which are simplices spanned between their *vertices* – also contained in the respective *segment*. *Segments* are non-overlapping, i.e. no *elements* or *vertices* are shared between *segments*. Global, i.e. *device*-wide connectivity is provided by *nodes*. Figure 3.2 shows a sketch of a VSP device, highlighting the role of *nodes*. Figure 3.3 displays the topological relations between *segment*, *element*, *vertex*, and *node*. Note that most relations are bidirectional, allowing to go from any topological entity to any other by following the references; some additional references that serve as shortcuts (e.g. Element → Node) have been omitted for clarity.

VSP can store structured data on all of the aforementioned topological entities, except on *nodes*, in the form of *quantities*. *Quantities* are one form of model *attributes*, the other two being *parameters* and *properties* – discussed already in Section 3.1. The relation between different kinds of *attributes* is shown in Fig. 3.4. All *attributes* of a model are available for data exchange; the user can instruct VSP to transfer *attributes* between model instances, demonstrated in Section 4.2.

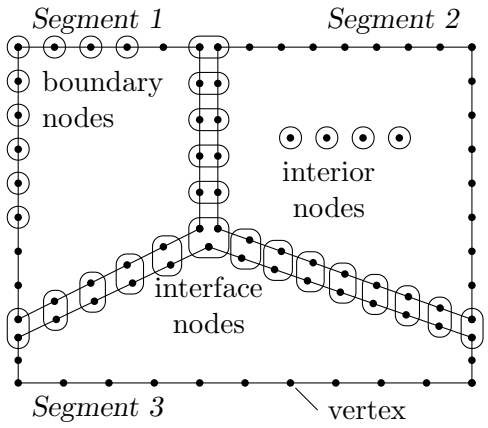


Figure 3.2: The role of *nodes* in VSP becomes clear when looking at devices with more than one *segment*. Every *segment* contains its own set of *vertices*, rather than sharing one set of *vertices* for the whole domain. At the interface between two *segments* we may have two (or more) *vertices* representing the same point in space. *Nodes* resolve this ambiguity by referencing the interface nodes.

3.2.2 Data Storage

Quantities may store any kind of position-dependent or k-dependent data; the data may be be scalar (`double`, `complex`, `int`), vector-valued (`Tuple<T>`), or tensor-valued (`Transform<T>`). A *quantity* may be represented in arrays of any number of dimensions: zero-dimensional (e.g. potential, carrier concentration), one-dimensional (e.g. single-band wave functions), two-dimensional (e.g. multi-band wave functions), four-dimensional (e.g. multi-band wave functions, k-resolved), and so on. The *quantity* storage features a smart allocation system, that reduces the memory footprint as well as the number of system calls to allocate memory. Allocations are deferred until a quantity is accessed for the first time, and the default value of a quantity (usually zero) is represented without using any memory.

On construction, every *attribute* must be provided with a data type, an identifier, a brief description, and a tag indicating the *attribute's* usage such as input, output, and internal. *Properties* and *quantities* must also be provided with a physical unit. The given information (type, identifier, description, tag, and unit) is used to refer to the *attribute*, verify data flow, and as part of the automated model documentation generation referred to as *literate-modeling* (cf. Section 3.5.3).

3.2.3 Discretization

VSP uses a finite volume discretization scheme, thus avoiding the weak formulation fundamental to finite elements and relying instead on a formulation based on the conservation of fluxes in each of the finite volumes. Unlike to most finite volume codes, the fluxes are treated in their full vectorial form and not as projections along the edge between two

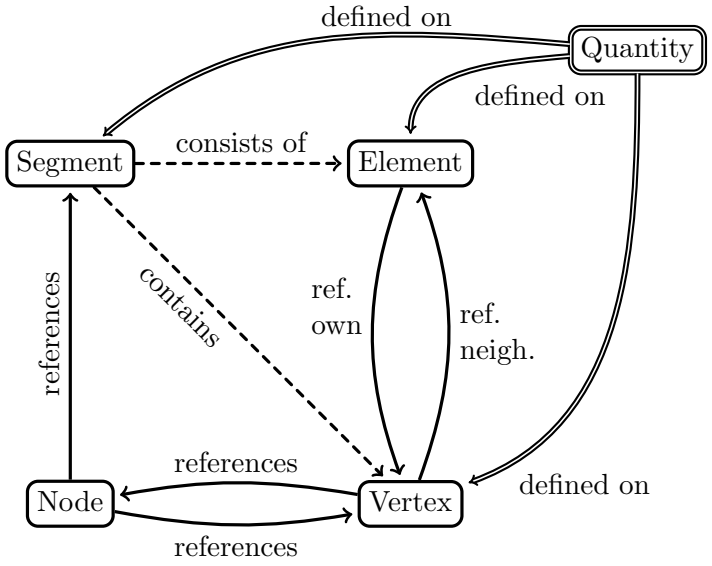


Figure 3.3: Topological relations between objects in VSP; *segments* contain both *elements* and *vertices*. *Vertices* contain their location in real/**k**-space, while *elements* contain geometrical data for computing couplings; *elements* and *vertices* reference each other. *Nodes* reference one or more *vertices* along with their corresponding *segments*. *Quantities* can be defined on *vertices*, *elements*, or *segments*.

points of the mesh. This is important because it is the only way material anisotropy can be introduced within a finite volume scheme. The discretization was demonstrated in [67], where the valence band states of a quantum dot were calculated using a highly anisotropic six-band **k**·**p** Hamiltonian.

Most physical laws are laws of conservation. Conservativity, therefore, serves as a common basis for the numerical modeling in the VSP simulation framework. The finite volume method (FVM) possesses the inherent property of conservativity and is therefore well-suited as a common discretization formalism for virtually all problems occurring in nanoelectronic devices [52].

Traditional FVM codes (Fig. 3.5 left) are edge-based (see e.g. [68]); a mesh node (i) couples to its neighbors (j) via the edges of the mesh graph. Each edge stores a length d_{ij} and a coupling area A_{ij}, each node stores its Voronoi cell volume V_i. The projection of the field, i.e. the derivative of a quantity φ along an edge, is approximated by $(\varphi_j - \varphi_i)/d_{ij}$. Some material property (permittivity, effective mass, ...) relates the field to a flux density which is multiplied by A_{ij} to obtain the partial flux leaving the cell. This approach has one major shortcoming: The field obtained by $(\varphi_j - \varphi_i)/d_{ij}$ is not the gradient of φ but only its projection along $\mathbf{e}_{ij}$ which implicitly assumes that the flux density caused by the field is parallel to $\mathbf{e}_{ij}$ as well. This restricts the discretization to isotropic media, i.e. ones with scalar field-flux relations.

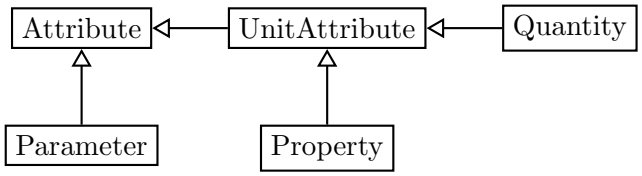


Figure 3.4: Relations between attribute types in VSP; the base type for data storage and exchange is the *Attribute*; *Parameter* is a direct descendant of *Attribute*; *UnitAttribute* stores a physical unit along with the raw data; *Property* and *Quantity* are its descendants.

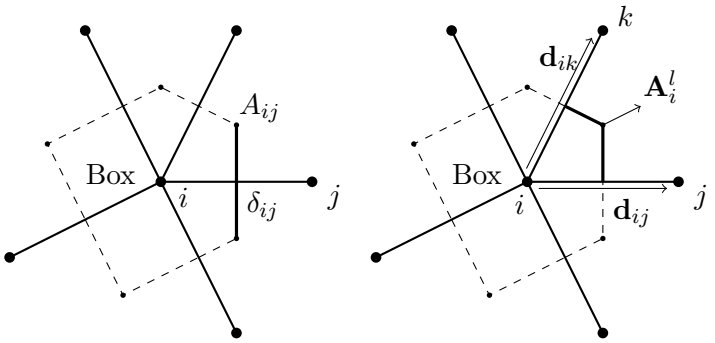


Figure 3.5: Comparison of edge-based and element-based finite volume methods; in element-based FVM both gradient and coupling surface area are vector-valued entities, while in edge-based FVM they are scalar, assuming implicit projection along the edge.

The FVM approach used in VSP is element-based [67]. Instead of looking at the neighbor nodes (j) of node i we look at its neighbor elements (l) as shown in Fig. 3.5 (right). By looking at the field in the element, we can now obtain the projection of the gradient of φ not only along one edge but along two edges in a two-dimensional mesh or three edges in a three-dimensional mesh. This allows to reconstruct the approximate gradient of φ which is assumed constant within the element. The reconstruction is done by inverting $\mathbf{U}^l := [\mathbf{d}_{ij},\ \mathbf{d}_{ik},\ \ldots]$ which is a matrix containing the edge vectors of the element with respect to node i as columns. Consequently, the approximate gradient of φ can be calculated as follows:

$$[\boldsymbol{\nabla}\varphi]^l \approx (\mathbf{U}^l)^{-1} \begin{bmatrix} \varphi_j - \varphi_i \\ \varphi_k - \varphi_i \\ \vdots \end{bmatrix}. \tag{3.1}$$

However, $\mathbf{U}$ may not be invertible. This could be the case when dealing with a two-dimensional surface in a three-dimensional coordinate space. In such a case $\mathbf{U}$ would be

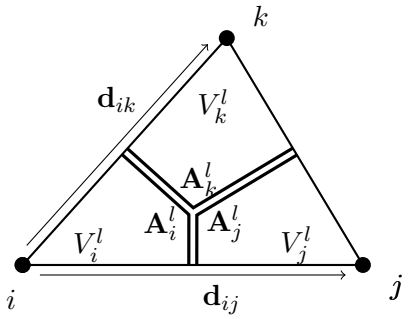


Figure 3.6: Element-centric discretization; partial fluxed are evaluated according to one of the rules in Table 3.2. The resulting $n_\mathrm{v} \times n_\mathrm{v}$ (n_v = number of vertices per simplex) partial fluxes are added to the system matrix in the appropriate rows and columns.

Table 3.1: Continuous operators and their discrete counterparts

	Continuous	Discrete, element l
Gradient	$\boldsymbol{\nabla}$	$\mathbf{Z}^l = \left[-\mathbf{Y}^l[1\ 1\ \ldots]^T,\ \mathbf{Y}^l\right]$
Divergence	$\mathrm{d}V\mathrm{div}$	$\mathbf{A}^l = [\mathbf{A}_i^l,\ \mathbf{A}_j^l,\ \ldots]^T$
Control volume	$\mathrm{d}V$	$\mathbf{V}^l = \mathrm{diag}(V_i^l,\ V_j^l,\ \ldots)$
Scalar quantity q	$q(\mathbf{r})$	$\mathbf{q}^l = \mathrm{diag}(q_i,\ q_j,\ \ldots)$

a 3×2 matrix which cannot be inverted. In such a case a pseudo-inverse can be used,

$$\mathbf{Y}^l := \mathbf{U}^l \left((\mathbf{U}^l)^T \mathbf{U}^l \right)^{-1}. \tag{3.2}$$

The field is now available inside the element l in its vectorial form. As such it can be manipulated by a second-order tensor to produce the flux density, allowing to fully account for anisotropy of the medium. The dot product of the flux density, which is also constant within the element, and coupling area vector $\mathbf{A}_i^l$ gives the partial flux leaving the cell i via element l.

The discretization can also be viewed from an element-centric point of view, as illustrated in Fig. 3.6. Here the element l is composed of n_v *partial volumes* V_i^l – the intersections between the element and each of the vertices' Voronoi cell. The parts of the Voronoi cell's surface that lie within the element are the *partial areas* or coupling areas $\mathbf{A}_i^l$ mentioned before. Along with the element's edge vectors $\mathbf{d}_{ij}$ those entities contain all the information necessary to discretize an (anisotropic) PDE on the mesh. Table 3.1 shows common operators and PDE terms along with a recipe to express each of them in a discretized form.

During VSP's initialization, the matrices $\mathbf{Z}^l$, $\mathbf{A}^l$, and $\mathbf{V}^l$ are precomputed for each element of the input-mesh and are provided to the assembly process described in Section 3.3.2.

3.2.4 I/O and Configuration

VSP is controlled by files written in the Input Deck (IPD) language [69]. IPD is a hierarchically structured configuration language organized in sections. Each section may contain nested sections and variables. Variables can be physical quantities containing units. IPD allows the use and evaluation of mathematical and logical expressions that are evaluated when needed. Any section may be derived from one or more parent sections by which it inherits all the parent sections' content.

VSP-IPDs contain three top-level sections: `Device`, `Materials`, and `Simulation`. `Device` defines the base (real-space) device; additional devices (also **k**-space) can be specified in their respective sections. `Materials` contains a nested section for each material known to VSP. A *material database* is included with VSP and contains parameter values for common semiconductors (Si, Ge, GaAs, ...) and insulators (SiO_2, HfO_2, ...) as well as metals. The `Simulation` section contains all the data and control flow information of the simulation work flow. The subsections of the `Simulation` section are used to provide configuration to each of the models instances in the simulation work flow, also allowing to specify data to be transferred between model instances.

Device files containing geometry information, meshes, and input data can be read by VSP. The files must be one of the DEV formats (DEVA, DEVAZ, DEVB, DEVBZ[1]) supported by GTS Framework [66], which also provides structure generation and visualization facilities. GTS Framework also features a graphical front-end to VSP, which facilitates VSP usage by writing syntactically and semantically correct IPD files.

3.3 Modeling Level

Section 3.2 introduced low-level terms such as *topology*, *data*, and *model attributes* grouped together in the data level. The modeling level described in this section is concerned with higher-level items, such as *mathematical expressions*, *(differential/integral) equations*, or *boundary conditions*. VSP provides a high-level interface to deal with these items – the `ModelExtended` class.

A model derived from `ModelExtended` inherits all the required infrastructure required for translation between the modeling and the topology or data levels. A typical model layout is shown in Fig. 3.7. It shows two additional modules: `Problem` (cf. Section 3.3.2) and `Assembler` (cf. Section 3.3.2); they are required when partial differential equations need to be solved numerically, but are optional otherwise.

3.3.1 Expressions

VSP allows the use of symbolic mathematical expressions for specifying equations, evaluation of expressions, integration, and similar tasks. The following code lines serve as illustration:

[1] Suffixes A and B denote ASCII and binary encoding, respectively. Suffix Z denotes compression using the DEFLATE algorithm.

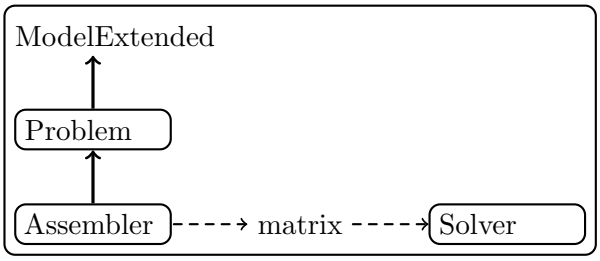


Figure 3.7: Typical algorithmic layout of a VSP model; a `Problem` instance is generated that uses the topological structure provided by the `ModelExtended` instance. The different types of boundary conditions are handled by the `Problem` instance. An `Assembler` instance uses the information provided by the `Problem` instance along with geometrical information from the model to discretize the equation and to assemble a matrix. The matrix is processed by a solver instance.

```
// vector potential
A   = 0.5 * B0 * cross(ez, position);

// density from wave function
rho = magsq(psi);

// electric field
E   = -grad(phi);

// calculating the centroid
Tuple<> center =
   integrate(position) / integrate(1.0);
```

Principle of Operation

Every combination of terms has its own C++ type as shown in Fig. 3.8; nested within are the type to which the expression evaluates, `EvalType`, and a tag to represent the location of evaluation, `Location`. Operators or functions applied to expressions are aware of both evaluation and location type and may modify them according to specific rules. For instance `magsq(...)`, the square-magnitude function will change the `EvalType` from `double`, `complex`, or `Tuple<T>` to `double`, while `grad(...)` will change `EvalType` to `Tuple<EvalType>` and `Location` from `Vertex` to `Element`. The static typing system in C++ serves as a formal correctness check for all expressions.

Material properties

The expression system is attached to the *material database* via the `MaterialDBAccess` class:

```
Quan<double> permittivity;
MaterialDBAccess poissonmat(this, "PoissonMaterialIpd");
permittivity = esp0 * poissonmat.prop<double>("epsr");
```

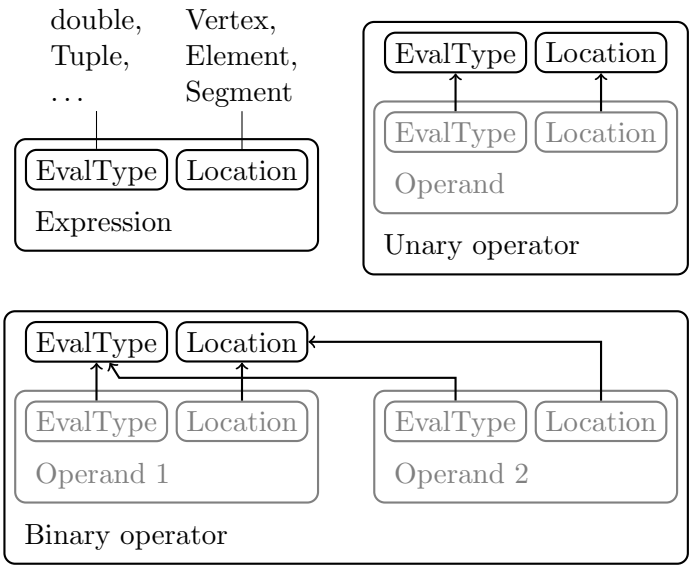


Figure 3.8: Expression typing scheme in VSP; each expression contains its evaluation type (`double`, `complex`, `Tuple`, `Tensor`, ...) and location tag (`Vertex`, `Element`, `Segment`) as nested types. Unary and binary operations derive their evaluation type and location based on the operands' evaluation and location types.

```
SegmentMask conductors;
MaterialDBAccess mattype(this, "MaterialTypeMaterialIpd");
conductors = selectSegments(
      mattype.param<MaterialType>("materialtype") == CONDUCTOR);
```

The read-out of the *material database* is handled by *material models*. The `MaterialDB-Access` class acts as an interface layer to the material model class specified in its constructor, and wraps the attributes of the material model into expressions, which are retrieved using the `param` and `prop` methods.

3.3.2 Problem Specification and Assembly

A `Problem` instance takes topological information about the boundary conditions, i.e. where and of what type they are. The information is processed using the low-level topological information provided by `ModelExtended` to pre-determine the rank and structure of the discrete equation system. The boundary conditions can be passed to an `Assembler` instance in equation form, along with the equations for the interior points of the domain. The `Assembler` instance can then generate a system matrix, which may be passed to a numerical solver. The most important aspect of the `Problem` class is that it provides a mapping between the *nodes* (Figs. 3.2 and 3.3) and the rows/columns of the matrix to be assembled.

Table 3.2: Second-order PDE terms in their discretized form

	Continuous	Discrete, element l
Laplacian	$\mathrm{d}V\nabla^2$	$\mathbf{A}^l\mathbf{Z}^l$
Anisotropic Laplacian	$\mathrm{d}V\boldsymbol{\nabla}\cdot\underset{\sim}{\tau}\cdot\boldsymbol{\nabla}$	$\mathbf{A}^l\boldsymbol{\tau}^l\mathbf{Z}^l$

Ordering for Sparsity

VSP heavily relies on sparse direct linear solvers to perform the bulk of the computational burden. These solvers most commonly use sparse-LU, sparse-LDL, or sparse-Cholesky factorization and their performance is greatly affected by the sparsity pattern of the matrix to be factored. The pattern is subject to the ordering of the rows and columns of the matrix, and the optimal pattern in terms of memory and operations required to perform the factoring, is obtained through ordering by nested dissection [70].

The `Problem` class applies nested-dissection-ordering to the *node* $\mapsto$ row/column mapping where it is appropriate. The nested-dissection algorithm itself is implemented as a *meta-model*, which uses the framework provided by `ModelExtended` but does not model any physics; it is automatically instantiated and invoked by the `Problem` class.

Assembly

In a 2D or 3D mesh the number of elements is several times greater than the number of nodes. To reduce the number of times a particular element has to be evaluated when building the system matrix, the assembly is element-centric: A loop iterates over the elements of the simulated structure. In each iteration, the partial fluxes between an element's vertices are evaluated and added to the appropriate elements of the system matrix.

This kind of assembly also allows the discretization of the problem's constitutive partial differential equations (PDE) to be broken down on a per-element basis. Continuous operators and operands can be directly translated into discrete ones which are represented by matrices. Table 3.1 shows how continuous vector-analytic operators (gradient and divergence) as well as continuous quantities are related to their discrete per-element representations as matrices. Operand matrices are diagonal and each diagonal entry corresponds to the operand's value at each of the element's vertices, hence for an n-dimensional simplex with $n_v = n + 1$ vertices, the element operands are n_v-dimensional diagonal matrices. Operators in contrast are full matrices; $\mathbf{A}^l$ is a $n_v \times 3$ matrix and contains the area vectors of the coupling surfaces between the element's vertices (see Fig. 3.6) as rows; $\mathbf{Z}^l$ is a $3 \times n_v$ matrix which relates the values at nodes to the gradient vector on the element.

Second order operators are discretized by multiplying the corresponding matrices as shown in Table 3.2. This also allows the assembly of mixed derivatives such as $\partial^2/\partial x\partial y$ which are just a special case of an anisotropic Laplacian with $\underset{\sim}{\tau} = \mathbf{e}_x \otimes \mathbf{e}_y$. First order

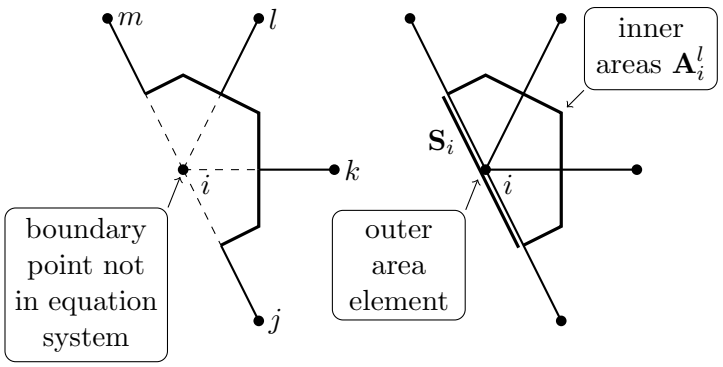


Figure 3.9: Topological treatment of boundary nodes; Dirichlet nodes (left) are not represented in the system matrix but the coupling to their neighbor nodes is computed nevertheless. Neumann and Robin nodes (right) are kept in the system adding an outer self-coupling area element.

derivatives are constructed using the relation

$$\boldsymbol{\nabla} = \frac{1}{2}[\nabla^2, \mathbf{r}] = \frac{1}{2}(\nabla^2\mathbf{r} - \mathbf{r}\nabla^2) \tag{3.3}$$

which is guaranteed to have an anti-symmetric discrete representation. In quantum mechanics, the self-adjointness property of Hamiltonians implies that comprising differential operators of even order must be must be symmetric and while operators of odd order must be anti-symmetric. This must also hold for the operators' discretized form.

VSP's assembly process discriminates between two types of boundary conditions, (i) Dirichlet and (i) Neumann boundary conditions, shown in Fig. 3.9. The type of boundary condition at each segment must be declared during initialization of the `Problem`-instance.

For Dirichlet boundary conditions, the `Problem` class pre-eliminates each *node* that is part of a Dirichlet boundary segment. The row and column corresponding to the element do not appear in the matrix, which means that the dimension of the matrix is reduced by the number of Dirichlet-nodes. For each interior node coupling to one or more boundary nodes, the terms of the linear equation belonging to the boundary nodes are transferred to the right-hand side of the equation.

A special case of Dirichlet boundary conditions are *floating* boundary conditions. For floating segments, all the nodes are lumped together into a single element of the solution vector, hence occupying only one line in the matrix. The total flux entering or exiting the floating segments is imposed via the right-hand side.

For Neumann boundary conditions, the boundary nodes are kept in the system. By default, a zero-flux Neumann boundary condition is imposed. A non-zero flux is imposed by multiplying the boundary flux density with the boundary element surface area $\mathbf{S}_i$ and adding the product to the right-hand side of the linear equation. Robin boundary conditions,

$$\mathbf{n} \cdot \boldsymbol{\nabla}\varphi + \alpha\varphi = \beta, \tag{3.4}$$

are constructed from Neumann boundary conditions by adding $\alpha_i \mathbf{S}_i$ to the matrix diagonal and β_i to the right-hand side. Plane wave boundary conditions,

$$\mathbf{n} \cdot \boldsymbol{\nabla}\psi - \mathrm{i}k\psi = 0, \tag{3.5}$$

used for open-boundary conditions in quantum transport simulations are a special case of Robin boundary conditions.

The assembly process is automated by the `Assembler` class. It allows specification of the PDE system as a set of discrete operator equations for each *device segment*. The `Assembler` object can then extract a sparse matrix from the defined equation. The following code snippet illustrates the process for a simple Poisson equation:

```
Assembler<double> assembler(problem);
Variable var_phi;

assembler.defineEquation(partvol * rho +
   eps0 * area(epsr * grad(var_phi)));

Sparse<double> matrix(problem.size());
Full<double> rhs(problem.size(), 1);
assembler.assembleLinear(matrix, rhs);
```

The symbolic `Sparse` matrix is converted to a CSR/CSC format [71] (`ConstSparse`) which can be processed using matrix operations and solvers displayed in Section 3.4.

3.3.3 Contour Integration

Contour integration is a specific but recurring task in the models presented in this work. Contour integration is mainly used in three contexts: (i) density-of-states calculation, (ii) assembly of the scattering operator of the Boltzmann transport equation, and (iii) evaluation of the effective generation/recombination rates for tunneling transport (not discussed in this work).

The `ContourIntegrator` class evaluates the integral of an *expression* ξ along a curve of surface $\mathscr{S}$ defined by the contour (iso-level ϕ_0) of a scalar function ϕ:

$$I[\xi, \phi, \phi_0] = \int_{\mathscr{S}} \xi \mathrm{d}S, \text{ where } \mathscr{S} = \{\mathbf{r} \in \mathbb{R}^d | \phi(\mathbf{r}) = \phi_0\}. \tag{3.6}$$

For each *element* that intersects the contour at ϕ_0 the intersecting segment of the contour surface is generated; the integrand ξ is interpolated onto that segment and multiplied by the area or the length of the segment, respectively. The sum of the products yield the approximated value of the integral.

The following snippet illustrates the evaluation of the density of states at energy E_0 using

$$g(E_0) = \int_{E(\mathbf{k})=E_0} \frac{\mathrm{d}S_{\mathbf{k}}}{\|\boldsymbol{\nabla}_{\mathbf{k}} E\|}, \tag{3.7}$$

from a band structure provided by *quantity* E :

```
Quan<double> E;
double E_0;
ContourIntegrator contour_integrator(this);

contour_integrator.generate(expr(E));

double dos_value =
      contour_integrator.integrate(
      E_0, 1.0 / norm(grad(expr(E))));
```

The `ContourIntegrator` class is instantiated using a `ModelExtended` instance before passing the contour-generating function by calling the `generate` method. This pre-processing step sorts the *elements* and *vertices* of the mesh for fast access, thereby reducing the run-time of multiple `integrate` calls.

Rather than performing the integration, the `ContourIntegrator` can also supply a list of pairs containing the *vertex* indices and their respective integration weights. This mode of operation is called *symbolic contour integration*.

3.4 Algebraic Level

The *algebraic level* is detached from the low-level picture of the topological and data level, and the high-level picture of the modeling level. It provides abstraction of entities such as matrices, solvers, and projections to a generic finite-dimensional linear operator, called `MatrixInterface`.

3.4.1 Abstraction of Linear Operations

A `MatrixInterface` object has the property of dimension and provides various methods for multiplication by a vector (or multiple vectors) from left and right, as well as evaluation of bilinear forms. Derivate classes of `MatrixInterface` are required to at least implement left and right multiplication as a minimal set of operations. The remaining methods can be constructed by `MatrixInterface` automatically. Figure 3.10 shows all the algebraic operator classes and their relation to `MatrixInterface`.

3.4.2 Working with Expressions

Interoperability between *quantities*, *expressions*, and matrix storage is simplified by three means of transferring data between these representations provided by VSP:

1. The `Assembler` converts the system of defined equations to a sparse matrix.
2. The function `assignToMatrix` evaluates an *expression* and writes the values using mapping provided by a `Problem` instance to the column of a full matrix:

   ```
   Problem problem;
   Quan<double> rho;
   Full<double> rhs;
   assignToMatrix(rhs, problem, expr(rho) * volume, false);
   ```

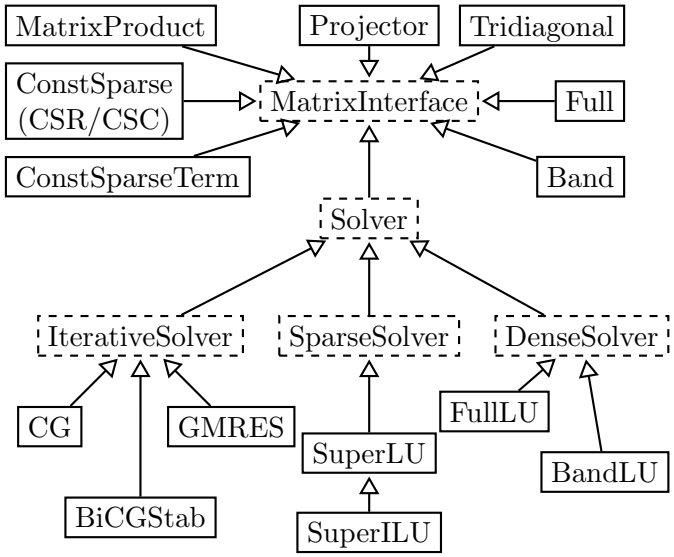


Figure 3.10: Relation map of the algebraic operators in VSP; the base type is `MatrixInterface`. The common interface allows to represent dense or sparse matrices, projections, solvers (direct, sparse direct, iterative), and combinations of these as generic algebraic operators.

Variants of `assignToSparse` and `assignToDiag` exist, which write the evaluated *expression* to the diagonal of a matrix.

3. The contents of a matrix can be wrapped into an *expression* using the function `matexpr`. The expression can then be used in a term or assigned to a *quantity*:

```
Problem problem;
Quan<double> phi;
Quan<double> phi_0;
Full<double> x;
phi = matexpr(x, problem) - phi_0;
```

3.4.3 Solvers

Two classes of numerical solvers are crucial for quantum-electronic simulation: linear solvers and eigenvalue solvers. Both are provided by a number of software packages that are stable and efficient. VSP links to several numerical libraries and tries to select the most appropriate solver for a problem at hand during run-time.

Linear Solvers

Available linear solvers in VSP are divided into (i) dense solvers provided by LAPACK or an interface-compatible library, (ii) direct sparse solvers such as SuperLU [72] or

PARDISO [73], and (iii) iterative solvers such as CG, GMRES, or BiCGStab [71]. The linear solver selection follows the rules:

- Full LAPACK for very small systems
- Banded LAPACK for 1D problems
- Direct sparse for 2D problems
- ILU-preconditioned iterative for 3D problems

Eigenvalue Solvers

Available eigenvalue solvers fall into two categories: "direct" solvers provided by the LAPACK library and subspace solvers such as the *Implicitly Restarted Arnoldi Method* (IRAM) provided by ARPACK [74] or the Jacobi-Davidson method [75]. The *efficiency* of the eigenvalue solver is crucial, since most of the simulation time in quantum problems is spent there. The eigenvalue solver is selected based on the following rules:

- Full LAPACK for very small systems (e.g. bulk $\mathbf{k}\cdot\mathbf{p}$)
- Banded LAPACK for 1D systems with few variables
- ARPACK (shift-invert) for 1D/2D problems
- ARPACK (plain) for definite 3D problems
- Jacobi-Davidson for large/indefinite 3D problems

Shift-invert is a technique that considerably improves performance of subspace solvers. It is based on the spectral transformation

$$\mathbf{A} \mapsto (\mathbf{A} - \sigma\mathbf{I})^{-1} \;\Rightarrow\; \lambda \mapsto \frac{1}{\lambda - \sigma}, \tag{3.8}$$

which makes convergence for eigenvalues close to the pole σ more favorable than for the remaining spectrum. For IRAM, the matrix inversion in Eq. (3.8) needs to be exact. For this reason, direct sparse linear solvers are used for inversion.

One challenge in quantum-electronic carrier models (cf. Sections 4.3.2 and 4.3.3) is that the exact number of eigenvalues needed is unknown beforehand; instead, the carrier model requires all eigenvalues within an energy interval, usually between a band edge and a cut-off energy, E_{lim}. While most direct solvers are capable of doing an interval search, subspace solvers require the number of sought eigenvalues to be known in advance. In VSP, this issue is resolved by developing a technique called *subspace deflation* [52] explained in Fig. 3.11. The method wraps around an existing subspace-based eigenvalue solver, such as ARPACK, thus the solver code needs not be altered. The wrapper repeatedly invokes the subspace solver until the search interval is exhausted, but eliminates already found eigenvalues from the matrix using a projection technique, resulting in negligible overhead compared to knowing the exact number of eigenvalues in the interval beforehand.

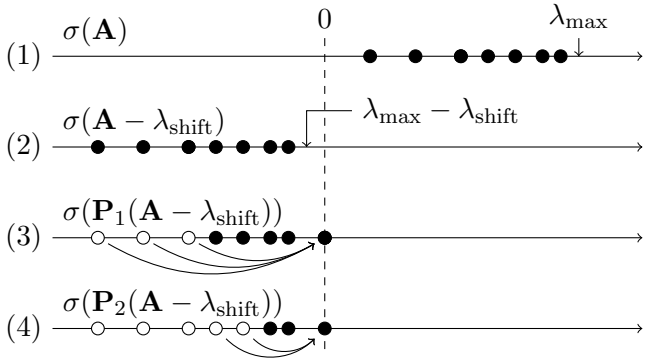


Figure 3.11: Searching for eigenvalues up to to $\lambda_{\max}$; (1) shows the spectrum of a positive definite Hermitian matrix $\mathbf{A}$. $\mathbf{A}$ is first shifted to the left by $\lambda_{\text{shift}} > \lambda_{\max}$ (2) and the first $n_{\text{ev}} = 3$ eigenvalues computed by a subspace solver (e.g. ARPACK); a projection matrix $\mathbf{P}_1 = \mathbf{I} - \mathbf{v}_i\mathbf{v}_i^{\text{H}}$ is constructed from the eigenvectors $\mathbf{v}_i$. The subspace solver is invoked again on the projected system $\mathbf{P}_1(\mathbf{A} - E_{\text{shift}})$. The projection (3) moves the found eigenvalues to 0 and effectively prevents the solver to converge on already found eigenvalues. The process is repeated (4) until all eigenvalues $< \lambda_{\max}$ are found.

3.4.4 Fast Fourier Transform

The Fast Fourier Transform (FFT) is most commonly used in scattering models to transform *form functions* from real into momentum space, where the square matrix element can be evaluated more efficiently. The commercial libraries MKL by Intel and ACML by AMD both provide their own FFT-implementation – VSP can use either.

3.5 Extension Through Modules

Due to its modular and model-oriented design, VSP can be readily extended by new models. With the proper interface new models can easily be integrated into the existing set of models.

3.5.1 Module Loading

VSP provides a module loading mechanism for external *modules*. *Modules* are shared objects or libraries that contain model classes VSP can use. When provided with a list of module names during invocation

```
$ vsp -M module1,module2 input.ipd
```

VSP locates and loads the given modules during startup; during the load code is executed which causes the models contained in the *modules* to be recognized by VSP.

3.5.2 Software Development Kit

Users can develop their own specific models using the VSP Software Development Kit (SDK). The SDK contains all headers necessary to access all the infrastructure presented in this chapter, along with examples and a CMake project [76] pre-configured for module building.

In order to make a model class recognizable to VSP, it must be *registered*. This is done using a simple macro, as shown below:

```
struct MyModel : ModelExtended
{
   ...
};

REGISTER_MODEL(MyModel);
```

The macro generates code which is executed during module load and registers the model class with VSP's *model server*.

3.5.3 Literate Modeling

In [77] the authors of NEMO5 point out: "Being a research code employed by changing generations of students, documentation, clarity, and modularity of the code are essential. Only when all these criteria are fulfilled, can junior researchers act as builders of individual modules and the code endure multiple generations of developers."

VSP goes even further by introducing the notion of *literate modeling*. It borrows from the concept of literate programming by D. Knuth [78] in which the program and its description are written as one document from which code and documentation can be extracted. The VSP code provides facilities to embed documentation into the models themselves. Their structure is based on *topic-oriented authoring* [79]. The following code snippet serves as illustration:

```
struct Schroedinger : ModelExtended
{
   ...
};

EXTERN_DOCUMENT(Topic, models_unstr)
DECLARE_DOCUMENT(
   ModelNode<vsp::Schroedinger>,
   Schroedinger)
DOCUMENT(Schroedinger,
   topic = &models_unstr,
   description = "This model solves the "
      "closed boundary single band "
      "Schroedinger equation ...")
```

The example only shows how to add a brief description to a model but the model description can be structured into several paragraphs. The documentation is organized in nodes representing sections and subsections of the documentation. The node for the `Schroedinger` model is added to the `Topic` node `models_unstr` which represents the

section containing descriptions of all VSP models operating on unstructured grids. The information provided with the model's *attributes* (Section 3.1) is automatically compiled into its documentation node. Also, every of the important IPD sections (`Device`, `Simulation`, `Logging`, `WriteQuans`, `WriteParams`) are documented in this manner.

The documentation is contained within the VSP binary and can be accessed by the user. The user can obtain formatted output for any documentation node by running VSP in documentation mode. This is useful as a quick reference for models and IPD sections and lowers the learning barrier of VSP especially for new users. Another usage is that the output of the entire documentation can be compiled into a manual using a document preparation system such as LaTeX. Additionally, the documentation system features automated generation of IPD defaults that can be included in simulation IPDs.

CHAPTER 4 Model Implementation

Contents

4.1 Basic Models 74
4.1.1 Poisson 74
4.1.2 Self-Consistent Loop 75
4.1.3 Strain 77
4.2 Model Chains 77
4.3 Carrier Models 78
4.3.1 Classic 3D Carrier Gas with Parabolic Band Structure 78
4.3.2 Confined Carrier Gas with Parabolic Band Structure 79
4.3.3 Confined Carrier Gas with Non-Parabolic ($\mathbf{k}\cdot\mathbf{p}$) Band Structure 81
4.4 Mobility Models 84
4.4.1 Mobility Calculation for Parabolic Bands 84
4.4.2 Mobility Calculation for Non-Parabolic Bands 85
4.5 Scattering Models 88
4.5.1 Scattering Model Interface 88
4.5.2 Coulomb Scattering Template 89
4.5.3 Non-Polar Phonon Scattering 90
4.5.4 Alloy Disorder Scattering 90
4.5.5 Ionized Impurity Scattering 90
4.5.6 Polar-Optical-Phonon Scattering 91
4.5.7 Surface and Interface Roughness Scattering 91

This chapter deals with the implementation details of the models implemented in VSP. The underlying physics of the models were already discussed in Chapter 2. This chapter draws on the computational methods laid out in Chapter 3. In a sense, the chapter combines Chapter 2 and Chapter 3 and discusses the implementation of each VSP *model* involved in simulating mobility in planar and non-planar devices. Throughout this chapter the word *model* refers specifically to a VSP model as described in Section 3.1 rather than a model in the common sense of the word.

The chapter is divided into five sections, each dealing with a different class of models: (i) basic models dealing with basic physical aspects such as electrostatics, (ii) model chains for consecutive model execution, (iii) carrier models providing electronics structure, carrier density, and concentration, (iv) mobility models, and (v) scattering models used by the mobility models.

4.1 Basic Models

This is an umbrella section for models involved in the overall simulation process but not belonging to any larger class of models.

4.1.1 Poisson

The **Poisson** model solves the linearized version of the Poisson equation in Eq. (2.47),

$$\boldsymbol{\nabla} \cdot \varepsilon \boldsymbol{\nabla} \varphi = -\varrho - \frac{\mathrm{d}\varrho}{\mathrm{d}\varphi}\varphi. \tag{4.1}$$

The second term on the right hand side serves as (approximative) derivative of the space charge density with respect to the potential and used to stabilize the self-consistent process. The equation is discretized using the finite volume method discussed in Section 3.2.3. Dirichlet boundary conditions may be imposed (ideal conductors), all other boundaries are zero-flux Neumann conditions. A special kind of boundary conditions are floating boundaries, where a boundary charge rather than an applied voltage is imposed, and the voltage becomes a solution variable.

Input quantities are the space charge density, interface charge density, and their respective derivatives, as well as initial potential, boundary voltages, and boundary charges. The output quantity is the potential difference between a provided initial potential and the solution, which can then be used as an update in a non-linear iteration scheme (cf. Section 4.1.2).

The **Poisson** model accesses the *material database* through the **MaterialTypeMaterialIpd** model, to determine which segments are metallic, and through the **PoissonMaterialIpd** model, to retrieve the dielectric constants of the materials present on the device object. The material models access the `Materials` section of the IPD to obtain the necessary information:

```
Materials
{
   ...
   Si
   {
      MaterialTypeModel
      {
         type = "semiconductor";
      }
      PoissonModel
      {
         epsr = 3.9;
      }
      ...
   }
   ...
}
```

4.1.2 Self-Consistent Loop

The self-consistent loop model **SCLoop** is used to iterate between the **Poisson** model and the carrier models, which will be discussed in Section 4.3, until the electrostatic potential converges. The **SCLoop** model takes care of setting the correct *parameters*, *properties*, and *quantities* of the submodels, and of supplying the *quantities* calculated in each step, such as potential and space charge density, to the respective models.

The **Poisson** model is hardwired into the **SCLoop** model but carrier models can be specified freely. This is illustrated in Fig. 4.1. The **SCLoop** model also supports multiple carrier models in different regions of the device. A typical application would be the examination of a gate stack with a poly-Si gate, where an accurate description of the electronic structure is sought in the channel, but a simpler (and faster) model suffices for the poly-Si gate.

SCLoop not only combines the electron, hole, and doping concentrations into a total charge density using Eq. (2.46) but also computes the space charge derivative w.r.t. the electrostatic potential using

$$\frac{\mathrm{d}\varrho}{\mathrm{d}\varphi} = q_0\left(\frac{\mathrm{d}N_\mathrm{D}}{\mathrm{d}\varphi} - \frac{\mathrm{d}N_\mathrm{A}}{\mathrm{d}\varphi} - \frac{\mathrm{d}n}{\mathrm{d}\varphi} + \frac{\mathrm{d}p}{\mathrm{d}\varphi}\right). \tag{4.2}$$

The derivative is used in the **Poisson** model to stabilize the self-consistent loop and to improve convergence rate. An exact derivative effectively results in a Newton-Raphson scheme, which has second-order convergence. However, for practical purposes it is sufficient to merely approximate the derivative and fast convergence can still be achieved. An update for the electrostatic potential is obtained from the **Poisson** model and applied to the current electrostatic potential estimate according to

$$\boldsymbol{\varphi}_n = \boldsymbol{\varphi}_{n-1} + d\delta\boldsymbol{\varphi}_n, \tag{4.3}$$

where $d \in [0, 1]$ is the damping parameter. **SCLoop** automatically lowers and increases

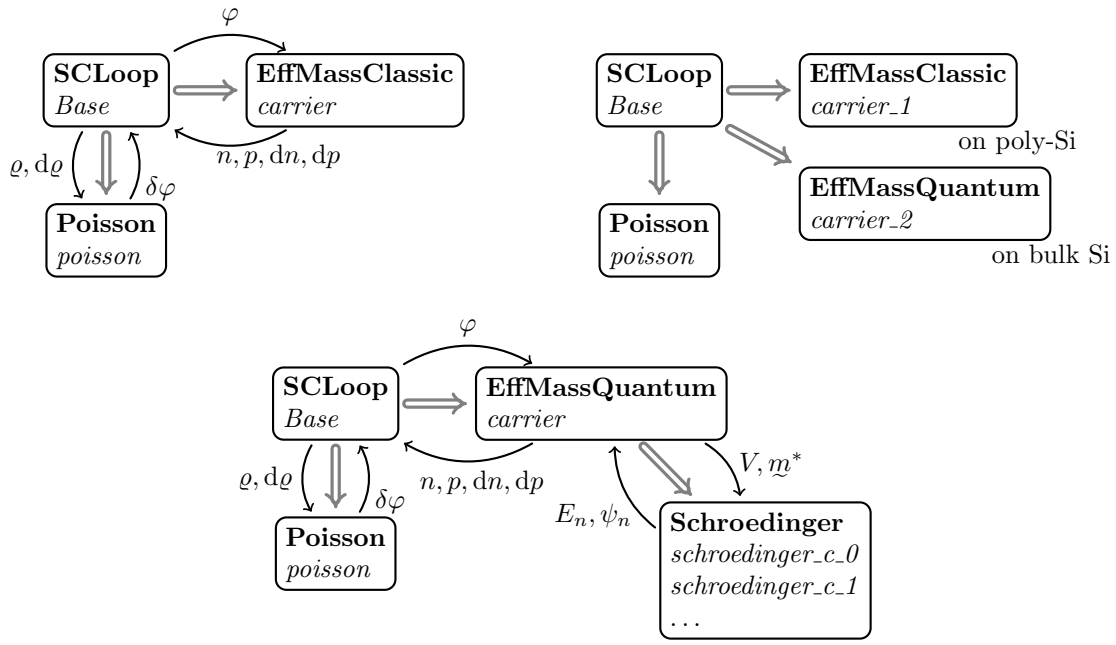


Figure 4.1: Common usage scenarios of the SCLoop model; **top left: SCLoop** configuration involving a **Poisson** model and the **EffMassClassic** model for classical equilibrium carrier distributions (cf. Section 4.3.1); ⇛ indicates submodel invocation, → indicates data flow; **bottom: SCLoop** model with the **EffMassQuantum** model for equilibrium distribution of confined carrier states (cf. Section 4.3.2); the carrier model instance invokes one closed-boundary Schrödinger model for every conduction/valence band valley (c_0, c_1, ..., v_0, ...); **top right:** different carrier models can be used on different segments, where appropriate; here, a classic carrier distribution is sufficient to model the accumulation/depletion effects in the poly-Si gate, while quantum confinement is accounted for in the channel.

d from iteration to iteration depending of the convergence behavior. Different strategies for computing d are described in [52].

The **SCLoop** model also defines the settings of the contact regions through the **Contact** model. In the **Contact** model the *voltage* or *Fermi* boundary conditions are set to the user-defined values. Voltage-type conditions impose a Dirichlet boundary condition in the Poisson equation, while Fermi-type conditions keep the Fermi-energy at a constant value while applying Neumann boundary conditions in the Poisson equation. By self-consistent iteration, the potential in the Fermi-type contacts will approach the built-in potential in order to reach charge neutrality. The **Contact** model also includes the work function difference for metals. At the same time, *contacts* define the regions where space charge is summed up and thereby allows the calculation of capacitance-voltage characteristics, which the **SCLoop** model can automatically extract.

4.1.3 Strain

Device files obtained from process simulators often contain the mechanical stress field rather than the strain field. However, strain is required as input for electronic structure models as discussed in Section 2.1. The **Strain** model converts the stress into strain, by inverting the elastic relation

$$\boldsymbol{\sigma} = \mathbf{C} : \boldsymbol{\varepsilon}, \tag{4.4}$$

with $\boldsymbol{\sigma}$, $\boldsymbol{\varepsilon}$, and $\mathbf{C}$ being the stress, strain, and stiffness tensors, respectively. The **Strain** model first rotates the stress tensor from device coordinates into crystal coordinates, depending on the orientation of the crystal. Then the strain is obtained in crystal coordinates using relation in Eq. (4.4), and rotated back into device coordinates.

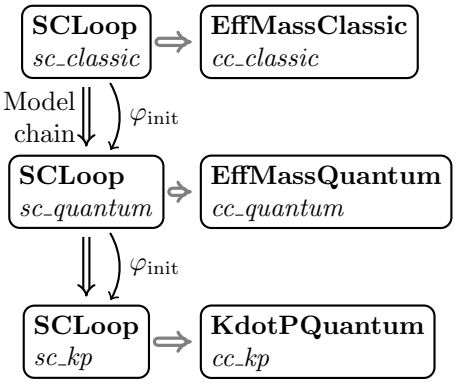
The **Strain** model accesses the *material database* through the **ElasticityMaterialIpd** model, to obtain the component of the stiffness tensor in Eq. (4.4) for each material present on the device object. The `Materials` section of the IPD provides the necessary information:

```
Materials
{
   ...
   Si
   {
      ElasticityModel
      {
         C11 = 166.0 "GPa";
         C12 =  64.0 "GPa";
         C44 =  79.6 "GPa";
      }
      ...
   }
   ...
}
```

4.2 Model Chains

The **Chain** model is different from the other models in the sense that it does not model any physics but serves as an utility to instantiate a *chain* of models and execute the instances in sequence. Together with VSP's ability to arbitrarily pass attributes between models, the **Chain** model provides the foundation to build complex simulation work flows and execute them.

A typical example is shown in Fig. 4.2 where the **Chain** runs **SCLoop** instances with increasingly complex (and computationally expensive) carrier models. The converged result from a **SCLoop** with a simpler carrier model is used as initial guess for a **SCLoop** with a more complex one. Using the the simpler model's result as initial guess, rather than starting with complex one right away, saves iterations and thus simulation time.




```
Simulation {
  model = "Chain";
  Base {
    models = ["SCLoop:sc_classic",
              "SCLoop:sc_quantum",
              "SCLoop:sc_kp"];

    sc_classic {
      ccd = "~Device.AcceptorConcentration";
                  ... }
    sc_quantum { phi = "^sc_classic.phi";
                  ... }
    sc_kp      { phi = "^sc_quantum.phi";
                  ... }
} }
```

Figure 4.2: A *chain* instantiates a list of models and invokes each of them in sequence when run; *attributes* can be passed between the model instances. A snippet of the IPD configuration is shown in the lower part.

4.3 Carrier Models

Carrier models compute the electron and hole concentrations for a given potential, as well as their respective derivatives or approximations thereof. They all have a common interface consisting of four input quantities, and four output quantities. The input quantities are electrostatic potential (φ/`phi`), electron and hole quasi-Fermi energies, ($E_{\mathrm{F}n}$/`Efn` and $E_{\mathrm{F}p}$/`Efp`), and temperature (T). The output quantities are the electron and hole concentrations (n/`ccn` and p/`ccp`) and the derivatives of the electron and hole concentration w.r.t. the electrostatic potential.

4.3.1 Classic 3D Carrier Gas with Parabolic Band Structure

The simplest carrier model is **EffMassClassic**. It evaluates the carrier concentration using the 3D formulas from Table 2.3 for parabolic bands. It does not include quantum

confinement, strain, and non-parabolicity effects.

The **EffMassClassic** model retrieves information about the electronic structure from the **EffectiveMassMaterialIpd** model, which references the following section in the *material database*:

```
Materials
{
   ...
   Si
   {
      EffectiveMassModel
      {
        Ec0 =  1.12 "eV";          // conduction band edge
        Ev0 =  0.00 "eV";          // valence band edge

        ConductionBand
        {
           ValleyX
           {
              ml = 0.916;          // longitudinal effective mass
              mt = 0.196;          // transversal effective mass
              symmetry = "X";      // symmetry: X, L, G(amma), K
              gv = 2;              // valley degeneracy
              shift = 0.0 eV;      // valley energy shift
              Dt = 1.1 "eV";       // transversal deformation potential
              Dl = 9.2 "eV" + Dt;  // longitudinal deformation potential
           }
           ...
        }

        ValenceBand
        {
           ...
        }
      }
      ...
   }
   ...
}
```

4.3.2 Confined Carrier Gas with Parabolic Band Structure

The **EffMassQuantum** model computes the concentration of a partially or fully confined carrier system, using the appropriate formula from Table 2.3 depending on the dimensionality of the device geometry. The states' energies, wave functions, and densities are computed by invoking the **Schroedinger** model, which solves the single-band effective-mass Schrödinger equation, Eq. (2.48), for each valley found in the *material database* for the semiconductor materials on the device.

Before having the confined states calculated by the **Schroedinger** model, the **EffMassQuantum** model first analyzes the profile of the confining potential, as shown ind Fig. 4.3. Here, we need to distinguish between two situations: *shallow* and *deep*

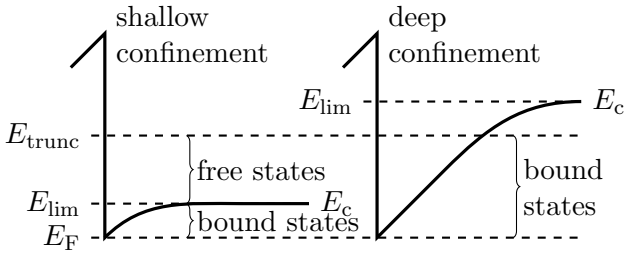


Figure 4.3: The shallow (left) and deep (right) confinement regimes; in the former mobile charge is comprised of bound carriers with energies up to E_{lim} and free carriers above E_{lim} modeled as a classic 3D carrier gas. Shallow confinement is typical of carrier accumulation. In deep confinement E_{lim} lies above truncation energy E_{trunc} and charge is comprised of bound carriers only. Deep confinement is typical of inversion and fully-depleted channels.

confinement. The two cases roughly correspond to carrier accumulation and inversion, respectively. In shallow confinement, bound states can only exist up to the energy E_{lim}, at which carrier can leak out of the potential well. In this case, bound states below E_{lim} are occupied using a low-order Fermi-Dirac integral from Table 2.3, while the free states above E_{lim} are treated as 3D carrier gas and are populated using the *incomplete* 3D Fermi-Dirac-Integral [52].

EffMassQuantum retrieves its material parameters from the same *material database* sections as **EffMassClassic** (Section 4.3.1). In contrast to **EffMassClassic**, **EffMassQuantum** uses the anisotropic masses, and deformation potentials provided by the *material database*. Additional attributes are added to the **EffMassQuantum** interface for setting strain and crystal orientation.

Single-Band Schrödinger Solver

The **Schroedinger** model is typically used as a submodel of **EffMassQuantum**, discussed above, although stand-alone invocation is also possible. The model solves the eigenvalue problem of the stationary single-band Schrödinger equation in real-space representation,

$$\left[-\frac{\hbar^2}{2}\boldsymbol{\nabla}\cdot\mathbf{w}\cdot\boldsymbol{\nabla}+V(\mathbf{r})+\sum_{\xi,\eta}\varepsilon_{\xi\eta}D_{\xi\eta}\right]\psi(\mathbf{r})=E\psi(\mathbf{r}), \tag{4.5}$$

where $\mathbf{w}=\mathbf{m}^{-1}$ is the inverse effective mass tensor, V the external potential, and $\varepsilon_{\xi\eta}$ and $D_{\xi\eta}$ the strain and deformation potential components, respectively. The equation is discretized using finite volumes discussed in Section 3.2.3. The resulting algebraic eigenvalue problem reads

$$\mathbf{H}\boldsymbol{\psi}=E\mathbf{V}\boldsymbol{\psi}, \tag{4.6}$$

where $\mathbf{H}$ is the discretized Hamiltonian, and $\mathbf{V}$ a diagonal matrix containing the volumes of the boxes associated with each *vertex* on the grid. The generalized eigenvalue equation can be transformed into an ordinary one,

$$\mathbf{V}^{-\frac{1}{2}}\mathbf{H}\mathbf{V}^{-\frac{1}{2}}\mathbf{x} = E\mathbf{x}, \quad \mathbf{x} = \mathbf{V}^{\frac{1}{2}}\boldsymbol{\psi}. \tag{4.7}$$

The **Schroedinger** model solves the discretized and transformed Schrödinger equation, finding either a given number of smallest or greatest eigenvalues, or all eigenvalues within a given energy interval using the solvers and methods discussed in Section 3.4.3.

4.3.3 Confined Carrier Gas with Non-Parabolic (k·p) Band Structure

KdotPQuantum is a carrier model that calculates the subband structure and carrier concentration based on the **k·p** model of the electronic structure discussed in Section 2.1.1. **KdotPQuantum** is interface-compatible with **EffMassQuantum**, such that the models can easily replace each other. Being based on a multi-band **k·p** model of the electronic structure, **KdotPQuantum** can represent confinement, orientation, and strain more accurately than **EffMassQuantum** and also describes band non-parabolicity.

Multi-Band Schrödinger Solver

KdotPQuantum is hard-wired to the **SchroedingerMulti** model, which solves the Schrödinger sub-problem, very much in the same sense as the **Schroedinger** model does for **EffMassQuantum**. Here, we deal with the multi-band Schrödinger equation,

$$\sum_j \hat{H}^{ij}_{\mathbf{k}}\psi^j_{\mathbf{k}}(\mathbf{r}) + V(\mathbf{r})\psi^i_{\mathbf{k}}(\mathbf{r}) = E_{\mathbf{k}}\psi^i_{\mathbf{k}}(\mathbf{r}), \tag{4.8}$$

where the indices i and j refer to the bands of the **k·p** band structure model. Each of the coupling Hamiltonians $\hat{H}^{ij}_{\mathbf{k}}$ contains terms of second, first, and zeroth order, similar to the operator in Eq. (4.5),

$$\hat{H}^{ij}_{\mathbf{k}} = -\frac{\hbar^2}{2}(\boldsymbol{\nabla} + \mathbf{k}) \cdot \mathbf{w}^{ij} \cdot (\boldsymbol{\nabla} + \mathbf{k}) - \hbar\mathbf{v}^{ij} \cdot (\boldsymbol{\nabla} + \mathbf{k}) + U^{ij} + \sum_{\xi,\eta} \varepsilon_{\xi\eta} D^{ij}_{\xi\eta} \tag{4.9}$$

with inverse masses $\mathbf{w}^{ij}$, velocities $\mathbf{v}^{ij}$, potentials U^{ij}, and deformation potentials $D^{ij}_{\xi\eta}$, which describe couplings of different order between bands i and j.

As in Section 4.3.2, the multi-band Schrödinger equation is discretized using the finite volume method. The anisotropic approach described in Section 3.2.3 is particularly important here, since mixed derivative operators, such as $\hbar^2\partial_x\partial_y/M$ often occur in **k·p** Hamiltonians. The Hamiltonian blocks in Eq. (4.9) are **k**-dependent and so are the energies and the wave function solutions of the multi-band Schrödinger equation. This **k**-dependence can be also represented in the discretized equation as

$$\mathbf{H} = \left(\sum_{\xi,\eta} \mathbf{A}_{\xi\eta} k_\xi k_\eta + \sum_\xi \mathbf{B}_\xi k_\xi + \mathbf{C}\right)\mathbf{x} = E\mathbf{x}, \tag{4.10}$$

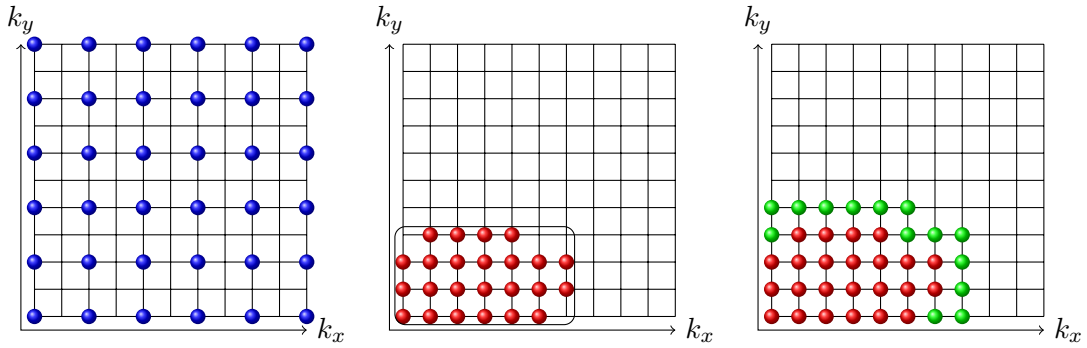


Figure 4.4: **k**-space scanning process; first, the **k**-grid is mapped coarsely and the states that lie within the energy range of interest are selected, while the remaining ones are discarded; second, the states of the **k**-points between the ones from the coarse run are computed; third, a layer of **k**-points surrounding the already computed region is tested for states that lie within the energy region if interest. The last step is repeated until the search returns no new states.

where each of the matrices $\mathbf{A}_{\xi\eta}$, $\mathbf{B}_{\xi}$, and $\mathbf{C}$ can be assembled separately and then combined using sparse matrix addition for different values of $\mathbf{k}$, without invoking the assembler over and over again.

The **SchroedingerMulti** model uses the algebraic methods described in Section 3.4.3 to solve the multi-band Schrödinger equations.

k-Space Integration

Because the solutions of the multi-band Schrödinger equation depend on $\mathbf{k}$ in a non-trivial way, a numerical **k**-space integration is needed to compute the integral for the carrier concentration in Eq. (2.43). **KdotPQuantum** uses a **k**-space grid to represent the subband structure and to perform the necessary integration. Unless provided with a specific **k**-grid by the user, **KdotPQuantum** constructs an ortho-product grid spanning the portion of **k**-space where the subbands of interest are expected to appear.

To make the integration computationally efficient, a search algorithms was devised that narrows the selection of **k**-grid points to include only the ones that significantly contribute to the integral. The search is done in multiple passes until the set of the contributing **k**-points is exhausted, as shown in Fig. 4.4.

Extraction of k·p-Parameters from the Material Database

The **KdotPQuantum** model has an interface to the *material database*. The interface consists of the **KPMaterialIpd** material model and a series of specialized *expressions* classes wrapped around it (cf. Section 3.3.1). The corresponding section in the *material database* reads similar to this:

```
Materials {
```

```
   Si {
      KPModel {
         models = ["TwoBandConduction", "ThreeBandValence"];

         TwoBandConduction {
            symmetry = "X"; // symmetry: X, L, G(amma), K
            degeneracy = 2; // spin

            H1 // definition of diagonal Hamiltonian block
            {
               type = "conduction";
               ml = 0.916;
               mt = 0.196;
               aux k0 = 0.15 * 2.0 * pi / 5.431 "Angstrom";
               vl = ^k0 / ml;
               Dt = 1.1 "eV";
               Dl = 9.2 "eV" + Dt;
            }

            H2 : H1 { vl = -^H1.vl; }

            HC // definition of off-diagonal Hamiltonian block
            {
               type = "coupling";
               inv_mtt = -2.0 * (1.0 / ^mt - 1.0);
               Dtt = 7.0 "eV";
            }

            // definition of Hamiltonian using blocks above
            H = [["H1", "HC"],
                 ["HC", "H2"]];
         }

         ThreeBandValence { ... }
   }
   ...
}
```

The example represents the configuration of the two-band $\mathbf{k}\cdot\mathbf{p}$ model in Eq. (2.20) taken from [39, 40]. The model parameters are structured in the same way as suggested by Eq. (4.8). The Hamiltonian consists of blocks coupling the individual bands of the $\mathbf{k}\cdot\mathbf{p}$ model; the block structure is reflected in above example by

```
H = [["H1", "HC"],
     ["HC", "H2"]];
```

allowing an arbitrary number of blocks to be set and, therefore, an arbitrary number on bands to be modeled. Each element references a subsection where the block masses, velocities, and potentials can be set following Eq. (4.9). The inverse coupling masses are provided in units of $1/m_e$ and the coupling velocities in unit of $\hbar/m_e$.

4.4 Mobility Models

Mobility models calculate the low field mobility of a confined carrier system. A mobility model needs to be attached to a carrier model instance from which it obtains the electronic structure: subbands, wave functions, and densities. Two mobility models are implemented in VSP, one for each type of electronic structure: **EffMassLowField** attaches to instances **EffMassQuantum**, whereas **KdotPLowField** attaches to instances of **KdotPQuantum**. The two models shall be described in this section.

Both mobility models rely on scattering models, which will be discussed later in Section 4.5, to compute the *square matrix elements*. Those are required to evaluate the transition rates via Fermi's golden rule, Eq. (2.49), for each of the various scattering processes. Mobility models instantiate scattering models as their submodels. The user has the freedom to select the scattering processes to be included in a simulation. The effect of multiple scattering processes is accounted for *microscopically*, by adding the transition rates, rather than macroscopically via Mathiessen's rule.

4.4.1 Mobility Calculation for Parabolic Bands

The **EffMassLowField** model implements the model described in *the effective mass case* of Section 2.4.2. The mobility computation is a linear process consisting of four steps: (i) transport mass calculation, (ii) evaluation of the scattering models, (iii) solving the linearized Boltzmann transport equation for the microscopic scattering rates, and (iv) extraction of channel mobility and conductivity.

Transport Mass Calculation

The **EffMassQuantum** model obtains effective masses for each material present in the device, rotates the effective mass tensor into the device coordinate system, and passes the rotated mass tensors to the **Schroedinger** submodels. The **EffMassLowField** model obtains the rotated mass tensors and applies the procedure in Eq. (2.37) to compute the effective mass component relevant for transport.

Scattering Model Invocation

Following the transport mass calculation, each of the selected scattering models is invoked. Each scattering model is expected to pre-calculate the square matrix elements for each possible transition between two states. The square matrix elements are retrieved in the next step via the `getSqMatrixElem` method. In the single-band effective mass approximation the square matrix elements of isotropic scattering processes only depend on the indices of the two subbands. For anisotropic processes, they also depend on the momentum transfer $\mathbf{q} = \mathbf{k} - \mathbf{k}'$; for better performance, models of anisotropic scattering processes tabulate the square matrix element for a range of $\mathbf{q}$-values and use a polynomial interpolation formula when the `getSqMatrixElem` method is called.

Conductivity and Mobility Calculation

Eq. (2.147) involves an energy-integration to compute the conductivity of each subband, the integrand being the microscopic relaxation time tensor, which is only defined implicitly via Eq. (2.151). The energy-integration is performed numerically, where the coefficients in Eqs. (2.152) to (2.155) are evaluated for every energy-grid value and Eq. (2.151) is solved to obtain the relaxation time tensor components.

Eq. (2.151) represents a system of equations the rank of which is equal to the number of subbands intersecting energy E. It is a small dense system and can be readily solved using direct methods. While the integrals Eqs. (2.154) and (2.155) can be reduced to simple analytical expressions for isotropic scattering processes, they need to be evaluated numerically for anisotropic processes, necessitating two nested loops for Ω_d and Ω'_d.

Finally, the subband conductivities are readily inserted into Eq. (2.148) to obtain the mobility of the individual subbands and the channel as a whole.

4.4.2 Mobility Calculation for Non-Parabolic Bands

The **KdotPLowField** model calculates mobility and conductivity from the linearized Boltzmann transport equation (LBTE), which is more general than the approach used in **EffMassLowField** model. The starting point is the reduced LBTE in Eq. (2.140),

$$\left[\frac{\partial \tilde{f}_n^1}{\partial t}\right]_{\text{scatt.}} = \mathrm{q}\mathbf{e}_{\mathbf{E}} \cdot \mathbf{v}_n(\mathbf{k}) \frac{\mathrm{d}f^0(E)}{\mathrm{d}E}. \tag{4.11}$$

where $\mathbf{e}_{\mathbf{E}}$ is a unit vector specifying the direction of the driving field and $\tilde{f}^1 = f^1/F$ is the reduced distribution response.

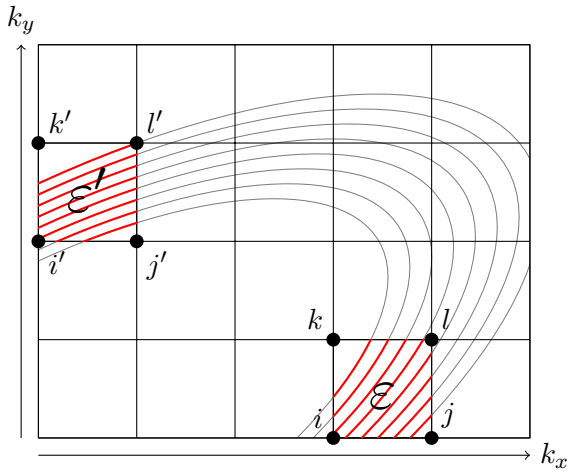


Figure 4.5: Energy contours that pass trough **k**-grid elements ε and ε' couple the elements' vertices, i, j, k, l, i', j', k', and l'.

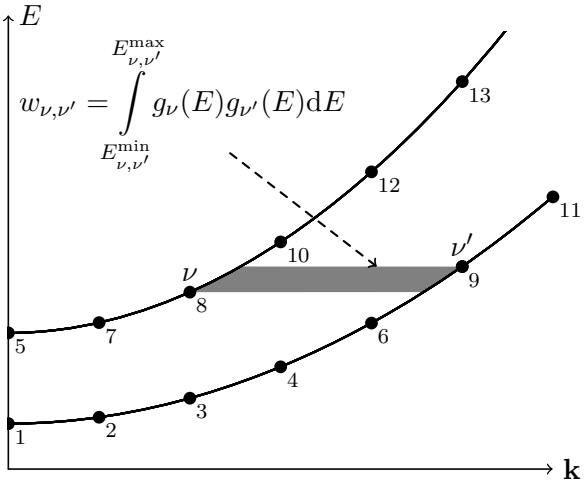


Figure 4.6: Calculation of the coupling weights for an elastic scattering operator; $w_{\nu,\nu'}$ is obtained by integrating the product of the density of states of state ν and ν' over the energy interval where ν and ν' overlap. Multiplied by a transition rate it gives the probability flux between ν and ν'.

k-Space Discretization

Equation (4.11) is discretized using the same **k**-space grid as that employed to obtain the subband structure in the first place, i.e. diagonalizing a **k**·**p** Hamiltonian. The **k**-grid imposes the concept of discrete **k**-cells that are coupled by probability fluxes due to scattering [80, 81]. The general form of the discretized equation is thus

$$\sum_{\nu'} \hat{S}_{\nu,\nu'} w_{\nu,\nu'} \left[\tilde{f}^1_\nu - \tilde{f}^1_{\nu'} \right] = -q_0 \mathbf{e}_\mathbf{E} \cdot \mathbf{v}_\nu \frac{\mathrm{d}f^0}{\mathrm{d}E} V_\mathbf{k}, \tag{4.12}$$

where $\nu = (n, \mathbf{k})$ is a global index that denotes the index of each **k**-grid cell in each subband and $V_\mathbf{k}$ is the volume, area, or length of a **k**-grid cell depending on the dimensionality of the carrier gas. The discrete cells are coupled through $\hat{S}_{\nu,\nu'} w_{\nu,\nu'}$, where

$$\hat{S}_{\nu,\nu'} = \frac{2\pi}{\hbar} \langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle \tag{4.13}$$

is the transition rate due to Fermi's golden rule without the energy-conserving $\delta(E - E' \pm \hbar\omega)$ and $w_{\nu,\nu'}$ are weights that arise from the discretization of the scattering operator. For inelastic processes, additional weights $(1 - f_\nu)$ and $(1 - f_{\nu'})$ appear, and the matrix $\hat{S}_{\nu,\nu'} \neq \hat{S}_{\nu',\nu}$ is no longer symmetric.

$$\sum_{\nu'} \hat{S}_{\nu,\nu'} w_{\nu,\nu'} \tilde{f}^1_\nu (1 - f_{\nu'}) - \hat{S}_{\nu',\nu} w_{\nu',\nu} \tilde{f}^1_{\nu'} (1 - f_\nu) = -q_0 \mathbf{e}_\mathbf{E} \cdot \mathbf{v}_\nu \frac{\mathrm{d}f^0}{\mathrm{d}E} V_\mathbf{k} \tag{4.14}$$

To compute the coupling weights $w_{\nu,\nu'}$ we need to consider the total probability flux from one cell ν to another ν'. When an equi-energy contour passes through two elements

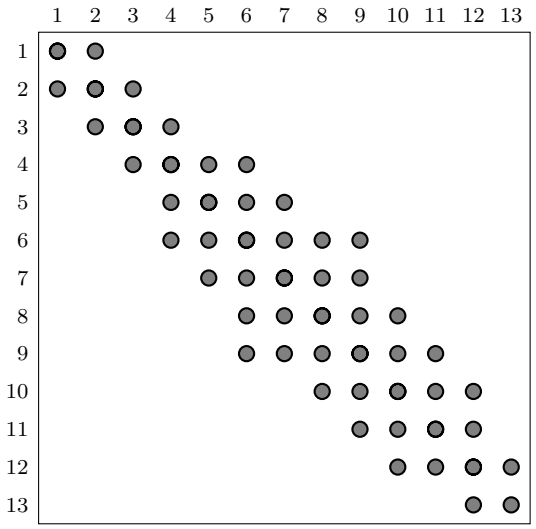


Figure 4.7: The resulting non-zero pattern of the discretized scattering operator for the example subband structure from Fig. 4.6; sorting all states by absolute energy produces dense symmetric skyline matrix, thus eliminating storage overhead.

on the **k**-grid it couples each element's vertices, as shown in Fig. 4.5. A numerical integration $\int \mathrm{d}E$ is performed, where the contribution of each contour is accumulated

$$\mathrm{d}w_{\nu,\nu'} = g_\nu(E) g_{\nu'}(E \pm \hbar\omega)\mathrm{d}E, \tag{4.15}$$

with $g_\nu(E) = 1/\hbar\|\mathbf{v}_\nu\|$ being the local density of states. The integration is preformed using the `ContourIntegrator` class described in Section 3.3.3. Figs. 4.5 and 4.6 visualize the integration procedure for elastic scattering ($\hbar\omega = 0$) in two-dimensional and one-dimensional subband structures, respectively.

Energy conservation in the scattering operator makes the resulting matrix $w_{\nu,\nu'}$ sparse. Ordering the discrete states by their absolute energy will result in a more dense skyline-type arrangement of the non-zero elements in the matrix [82] as shown in Fig. 4.7. This is important since the number of non-zero elements in a realistic device can easily reach tens or hundreds of millions; dense storage has virtually no memory overhead and results in fast access times, since the elements don't have to be *searched for* in maps.

Scattering Model Invocation

The sparsity of $w_{\nu,\nu'}$ also means that $\hat{S}_{\nu,\nu'}$ only needs to be computed for the non-zero elements. This is crucial since the computation of the transition rates in confined systems is by far the most time demanding task and the effort must be kept to a minimum.

The **KdotPLowField** model invokes each of the activated scattering models and passes the information about the sparsity pattern of $w_{\nu,\nu'}$ by calling the scattering model's method `setSqMatrixElemMaps`. A `StateIndexer` object is passed during the call which maps each pair of states for which $w_{\nu,\nu'}$ is nonzero to a unique transition index and

vice versa. The scattering model is then expected to process each of the transitions and store the values of $\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle$ of each transition in a list. This list is obtained by the **KdotPLowField** model by calling the `setSqMatrixElemMaps` method. The `StateIndexer` is then used to retrieve the information from the list.

LBTE-Solution

Once the scattering operator has been computed, the **KdotPLowField** model needs to solve Eq. (4.12). Being a large, sparse system, the discretized LBTE is ideally solved using iterative methods. However, it needs to be considered that the LBTE is singular, and necessarily so: In the absence of any fields, the homogeneous Boltzmann transport equation Eq. (2.130) reduces to

$$\left[\frac{\partial f_n(\mathbf{k})}{\partial t}\right]_{\text{scatt.}} = 0. \tag{4.16}$$

Thus, at least the equilibrium distribution function is within the kernel of the scattering operator. This also carries forward to its discretized version. To remedy this problem, an iterated Tikhonov regularization technique is applied [83].

4.5 Scattering Models

Scattering models are a category of models used by low-field models **EffMassLowField** and **KdotPLowField**. Each scattering model describes a specific type of scattering process based on Fermi's golden rule (Eq. (2.49))

$$S_{n,n'}(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle\delta(E(\mathbf{k}) - E(\mathbf{k}') \pm \hbar\omega). \tag{4.17}$$

Each scattering model provides the square matrix elements, $\langle|H_{n,n';\mathbf{k},\mathbf{k}'}|^2\rangle$, for each possible transition based on the input consisting of the wave functions along with process-specific parameters.

The physics of each scattering model has been described in detail in Section 2.3. This section will establish the connection between the mathematical models and the implemented VSP models, and also fill in some of the relevant implementation details.

4.5.1 Scattering Model Interface

Before discussing the individual scattering models in detail, a short description of the programming interface common to all scattering models is shown. The base class for scattering models has the following interface:

```
struct ScatteringBase : ModelExtended
{
   // constructor, destructor
   ScatteringBase(ObjectUnstructured *obj);
   ~ScatteringBase();
```

```
    // for querying properties of model
    virtual bool isIsotropic() const = 0;
    virtual bool isElastic() const = 0;
    virtual bool isIntraValley() const = 0;

    // relaxation energy ħω
    virtual double getRelaxationEnergy() const = 0;

    // Is process activated for a specific band structure valley?
    bool isEnabled(size_t ival) const;

    // square matrix element for a specific transition
    // called by EffMassLowField only
    virtual double getSqMatrixElem(const tensor::Tuple<> &q,
          size_t ival, size_t jval, size_t isb, size_t jsb,
          int relax = 0) = 0;

    // defines the transitions that are possible due to energy
    // conservation by means of a StateIndexer object
    // called by KdotPLowField only
    void setSqMatrixElemMaps(size_t ival, const StateIndexer &indexer,
          const valarray<tensor::Tuple<> > &kvec);
    void setSqMatrixElemMaps(size_t ival, const StateIndexer &indexer);

    // retrieves list of square matrix elements
    // called by KdotPLowField only
    const Full<double> &getHsqMatrix(size_t ival) const;
    Full<double> &getHsqMatrix(size_t ival);
}
```

From `ScatteringBase`, a template class is derived which allows to predefine the return values of `isIsotropic`, `isElastic`, and `isIntraValley` based on tags:

```
// tags for ScatteringTemplate
struct Elastic;
struct Inelastic;
struct Isotropic;
struct Anisotropic;
struct IntraValley;
struct InterValley;

template <class Isotropy, class Elasticity, class Valley = IntraValley>
struct ScatteringTemplate : ScatteringBase
{
};
```

4.5.2 Coulomb Scattering Template

A special case is the class

```
template <class Elasticity>
struct CoulombScatteringTemplate :
      ScatteringTemplate<Anisotropic, Elasticity>
```

```
{...};
```

which contains facilities for defining the electrostatic Green's function and evaluating the sensitivity function from Eq. (2.79).

The Poisson equation from Eq. (2.84) is discretized using the finite volume method developed in Section 3.2.3 which also takes care of device geometry, interface conditions, and boundary conditions. The discretized system is solved using direct sparse methods mentioned in Section 3.4.3.

It is of advantage, to factor the matrix of the discretized Poisson equation only once for each discrete value of $q = \|\mathbf{k} - \mathbf{k}'\|$ and to apply only the much cheaper solve-step for each wave function product $\psi_n^* \psi_{n'}$. To make use of this optimization, the system matrix is only factored on a grid of predefined q-values. The values of $\langle |H_{n,n';\mathbf{k},\mathbf{k}'}|^2 \rangle$ at a specific $q = \|\mathbf{k} - \mathbf{k}'\|$ are then obtained by executing the solve-step at the two adjacent q-grid points and interpolating between the results.

This template is the base for the models **ScatteringIIS** and **ScatteringPOP**.

4.5.3 Non-Polar Phonon Scattering

The models **ScatteringADP**, **ScatteringODP**, and **ScatteringIVS** compute the square matrix elements for acoustic, intra-valley optical, and inter-valley optical phonon scattering by means of the integrals given in Eq. (2.75), Eq. (2.66), and Eq. (2.67), respectively.

Scattering model parameters, such as phonon energy $\hbar\omega$, sound velocity v_s, and mass density $\bar{\rho}_\mathrm{m}$, as well as the individual deformation potentials D_ac, D_opt, and D_iv can be provided via their respective IPD sections.

4.5.4 Alloy Disorder Scattering

The **ScatteringADS** model computes the square matrix element for alloy disorder scattering using integral Eq. (2.126).

Scattering potentials for electrons and holes are provided via the model's IPD section, while the cell volume is obtained from the *material database* for each material found on the device. Material composition x can be either set via IPD or read from the device object.

4.5.5 Ionized Impurity Scattering

The **ScatteringIIS** model implements the integral in Eq. (2.82) to compute the square matrix element for a given density distribution of charged impurities. It is based on the `CoulombScatteringTemplate` class described in Section 4.5.2 which provides the efficient solution of the Poisson equation, which in turn provides the sensitivity function, $U_{n,n';\mathbf{k},\mathbf{k}'}$.

Both volume and interface-densities of charged impurities can be passed to **ScatteringIIS**, as well as screening functions for electrons and holes. The dielectric constants

for the materials found in the device are obtained from the *material database* (see Section 4.1.1). The **ScatteringIIS** model can be used for all kinds of charged impurities, such as dopants, interface charges, or remote charges.

4.5.6 Polar-Optical-Phonon Scattering

The **ScatteringPOP** model follows the derivation in Section 2.3.3. It implements the integral from Eq. (2.106) to evaluate the square matrix element using the gradient of the sensitivity function as defined in Eq. (2.108). As **ScatteringIIS**, the **ScatteringPOP** models is based on the `CoulombScatteringTemplate` class which provides the means to obtain the sensitivity function.

The model allows one to set the phonon energy $\hbar\omega$, while the low and high-frequency dielectric constants of the materials found in the device are obtained via the *material database*.

The **ScatteringPOP** model can be used for both local polar-optical phonon scattering which is typically present in compound semiconductors, as well as for remote phonon scattering that occurs in combination with high-k dielectric gate stacks.

4.5.7 Surface and Interface Roughness Scattering

The **ScatteringSRS** model implements the procedure laid out in Section 2.3.4. First, the form-functions are computed along segment interfaces according to Eqs. (2.118) and (2.119) using the wave functions provided to the model. The **ScatteringSRS** model automatically obtains the effective band edge potentials and effective mass tensors from the **EffMassQuantum** and **KdotPQuantum** models used to compute the states.

Seccond, the spectral form-functions need to be computed from the form-functions $F_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)$ required for the integral in Eq. (2.125). A computationally efficient procedure was devised to compute the spectral form-functions, depicted in Fig. 4.8. The form-functions, defined along the interface curve $\mathscr{C}$, are resampled onto an equidistant s-grid and fast-Fourier-transformed to obtain their spectral counterparts $F_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)$.

Having found the spectral form-functions $F_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)$ the square matrix element in Eq. (2.125) is obtained via $q_\perp$-integration. In the case of a parabolic subband structure the wave functions and thus the form-functions are independent of $\mathbf{k}$ and $\mathbf{k}'$. So $F_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp) = F_{n,n'}(q_\perp)$ allows $|H_{n,n';\mathbf{k},\mathbf{k}'}|^2$ to be tabulated for different subband pairs n, n' and $q_\perp$-values further reducing computational cost.

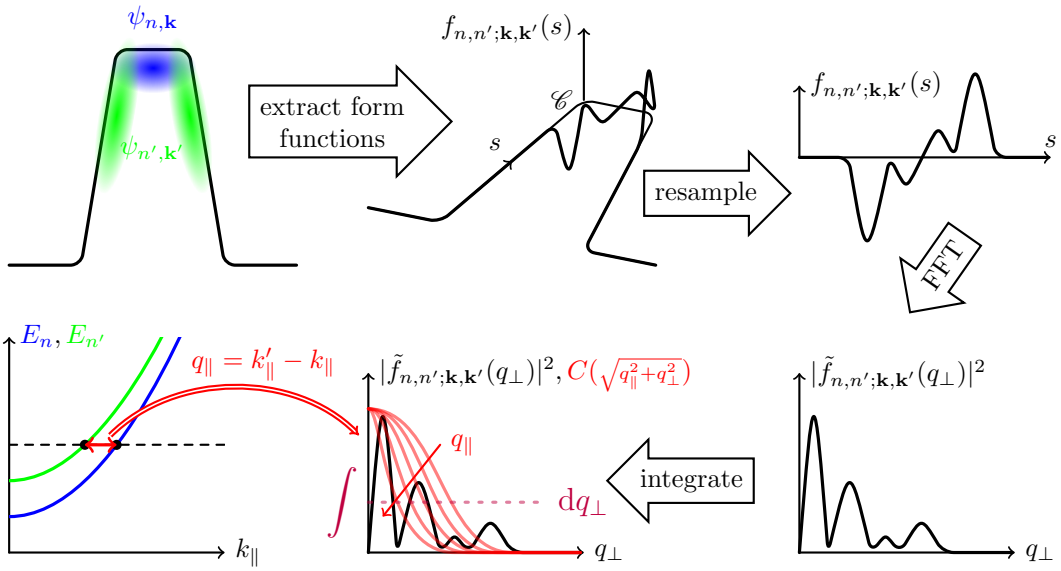


Figure 4.8: Computational procedure to obtain the form-functions $f_{n,n'}(s)$ and the spectral form-functions $F_{n,n'}(q_\perp)$: For each two cross-section wave functions ψ_n and $\psi_{n'}$ the expression in Eq. (2.118) is evaluated along the interface curve $\mathscr{C}$ on the mesh used for computing the states. The form-function $f_{n,n'}(s)$ is interpolated onto an equidistant s-grid and padded with zeros if $\mathscr{C}$ is open. The spectral form-function $F_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)$ is computed using the fast Fourier transform (FFT). Calculation of the square matrix element for a transition from subband n to subband n' in Eq. (2.125) is done as follows: For each energy value the difference of axial k-vectors is evaluated which represents the axial momentum transfer $q_\parallel$. The roughness *power spectrum* $C(q)$ is offset using $\sqrt{q_\parallel^2+q_\perp^2}$ and its product with the spectral form-function $F_{n,n';\mathbf{k},\mathbf{k}'}(q_\perp)$ integrated.

CHAPTER 5 Results

Contents

5.1 Inversion-mode Channels 93
5.1.1 22 nm n-Type Silicon FinFET 94
5.1.2 22 nm p-Type Silicon FinFET 99
5.1.3 InGaAs-Based Devices 101
5.2 Junction-Less Channels 107

This chapter presents a number of case studies to demonstrate the models and computational methods developed in this work. The results focus mainly on inversion mode devices with a brief section on junction-less devices at the end. The presented inversion-mode devices, including silicon-based n-type and p-type devices as well as InGaAs n-type devices, will be analyzed by low-field simulations of the cross-section. These will provide channel mobility and conductivity. While these low-field quantities are not directly observable due to the very short gate lengths of the devices, they serve as a meaningful metric to capture the amount of scattering the carriers undergo in the respective channels. In this way, trends of channel mobility with respect to crystal orientation, geometry, stress, and material composition can be established.

5.1 Inversion-mode Channels

The bulk of the case studies presented in this chapter focus on inversion-mode devices. Without a gate, inversion-mode devices would normally be in the off-state. The channel is either intrinsic or doped to produce majority carriers of opposite type to the type of the carriers conducting current in the on-state. So, a n-type device would have p-type doping in the channel, whereas a p-type device would have n-type. While the conducting carriers are minority carriers in the channel, they are majority carriers in the source and drain regions which are highly doped. The doping profile produces p/n-junctions between channel and source, and channel and drain, which form a potential barrier in the channel limiting the flow of current from source to drain. Thus, in a inversion-mode device, the gate potential actively reduces the barrier and allows current to flow.

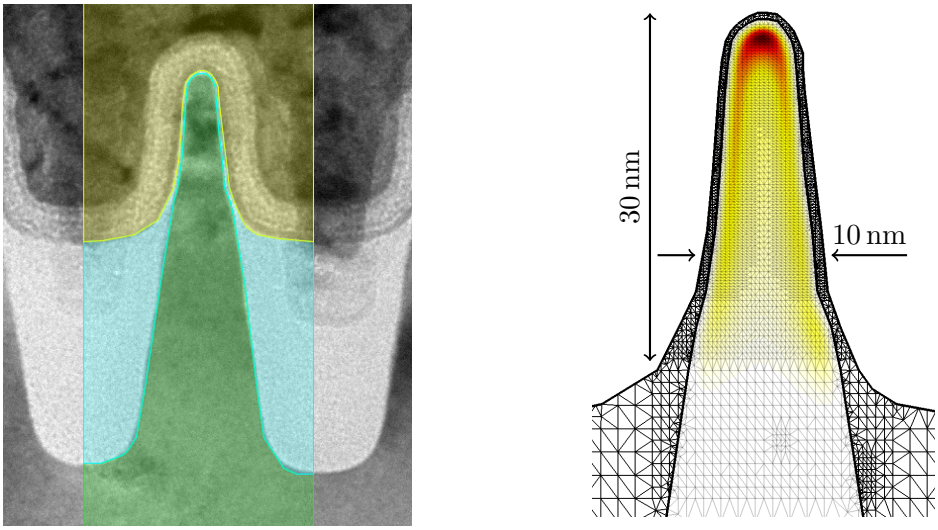


Figure 5.1: Left: TEM image of an NMOS fin structure fabricated by Intel [6], [84]; segments of the simulation domain are overlaid. Right: electron concentration in the fin from a self-consistent Schrödinger-Poisson calculation at gate bias $V_G = 1\,V$; the computational grid is shown.

5.1.1 22 nm n-Type Silicon FinFET

The first device to investigate will be Intel's tri-gate NMOS transistor [6] a cross-section of which is shown in Fig. 5.1. While the exact fabrication process of these FinFETs is not publicly known, the structure shown here was obtained by dismantling a commercially available chip and exposing the FinFET cross-sections [84]. The teardown has revealed that the fins are not perfectly rectangular but have inclined sidewalls and are rounded at the top. Due to the confidentiality of the fabrications process it is not known whether these modifications are deliberate or involuntarily induced by the process. A closeup of the TEM image [84] reveals that the FinFET channel is oriented along the {110}-axis and fabricated on a {100}-surface wafer. The devices are reported to be mechanically stressed to enhance performance [6]. However, no quantitative figures on the strain conditions in the devices are known.

In the following paragraphs different parameter variations will be applied to the device design and their impact analyzed.

Channel and Substrate Orientation

To investigate the transconductivity for different combinations of channel and substrate orientation, self-consistent Schrödinger-Poisson simulations are performed using the **SCLoop** (Section 4.1.2) and **EffMassQuantum** models (Section 4.3.2), combined with low-field calculations using the **EffMassLowField** model (Section 4.4.1). The geometry

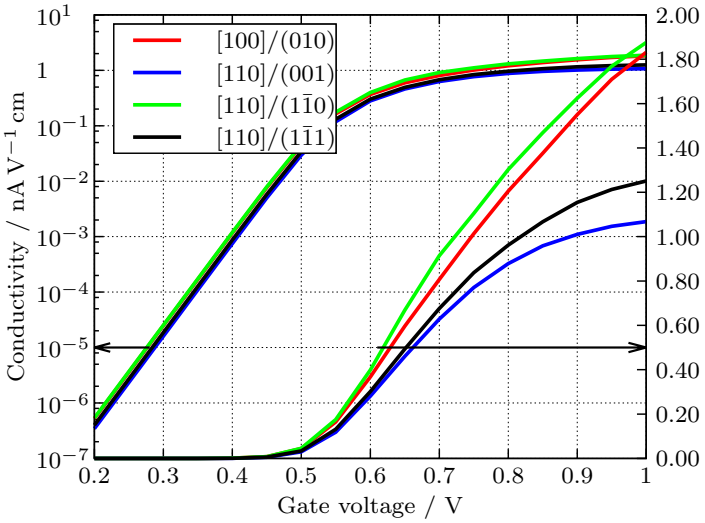


Figure 5.2: Fin channel transfer characteristics for four different channel/substrate orientations of the device shown in Fig. 5.1; degradation of the characteristic can be observed for [110]/(001) and [110]/(1$\bar{1}$1) orientations, but not for [110]/(1$\bar{1}$0) which has about the same drive current as [100](010), the traditional orientation in Si MOSFET fabrication.

of the simulation domain was extracted directly from the TEM-image of the fabricated device, as shown in Fig. 5.1.

Figure 5.2 shows that channel conductivity behaves differently for each combination of orientations. The transconductivity is severely degraded for the orientations [110]/(001) and [110]/(1$\bar{1}$1) but not for [110]/(1$\bar{1}$0) and [100]/(010), the latter being the traditional channel orientation for Si MOSFET devices. The orientation of the fabricated device is [110]/(001), which is not optimal, but is likely due to the unavailability of wafers with surfaces other than (001) for large-scale production.

Sidewall Inclination

As previously mentioned, the fabricated 22 nm FinFETs have inclined sidewalls, i.e. the fins are tapered. To find a possible reason why and whether tapering has any influence on performance, a systematic study on an ensemble of different channel shapes, ranging from triangular via trapezoidal to rectangular is performed. The ensemble of devices is shown in Fig. 5.3. The simulation procedure is the same as for the previous analysis.

Figure 5.4 shows the fin-shape dependence of the mobility and of the contributing scattering mechanisms. It reveals that certain orientations show a significant variation of mobility with respect to fin shape; [110]/(001) clearly shows optimal mobility for a slightly tapered fin.

The analysis also reveals that surface roughness scattering is the source of mobility

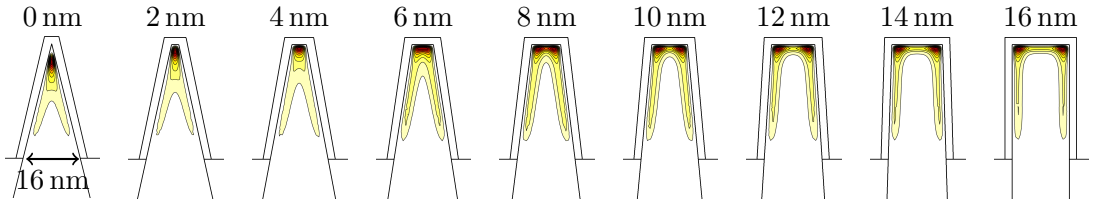


Figure 5.3: To shed light on what influences current degradation, an ensemble of channel cross-section shapes is generated, ranging from a triangular fin to rectangular one. Transfer characteristics are calculated for each shape and orientation; electron concentration is shown for the [110]/(001) orientation (degraded) at $V_G = 1\,V$ (strong inversion). We note that electrons preferably occupy the top of the fin and the device corners with slight inversion close to the sidewalls.

degradation, while phonon scattering shows little variation with respect to orientation or shape. This also indicates that the variation of transconductivity for different orientations shown in the previous analysis is due to surface roughness scattering. The observation can be explained as follows: In the rectangular channel electrons interact with both top and sidewall roughness, while in the triangular channel they are squeezed between the inclined sidewalls. Proximity to rough {110} or {111} surfaces is known to reduce electron mobility, compared to {100} [85]. Hence, if the interacting surfaces are not close to {100}, mobility degradation will occur.

It seems plausible from our observations that a tapered fin might have been found to give the best performance during technology development and was therefore deliberately chosen for the device design.

Strain

The last analysis of the 22 nm NMOS FinFET is concerned with strain and its impact on the device performance. Strain is caused by mechanical stress induced by the fabrication process and acts on carrier transport via two mechanisms: (i) valley re-population and (ii) effective mass change [45].

Valley re-population happens due to a strain-induced energy shift of individual valleys in the band structure, and is a first-order effect. Since proccess-induced strain is either uniaxial or biaxial, differently oriented valleys experience different shifts, causing a carrier population in each valley different from the unstrained case. Strain is commonly applied in such way that the valleys with light transport mass (see Section 2.1.3) are more populated than the others.

This impact of valley re-population for FinFETs is shown in Fig. 5.5. Due to two-dimensional confinement, the light-mass valleys are already well separated from the heavy-mass valles, so there is not much to be gained from strain engineering by means of valley re-population.

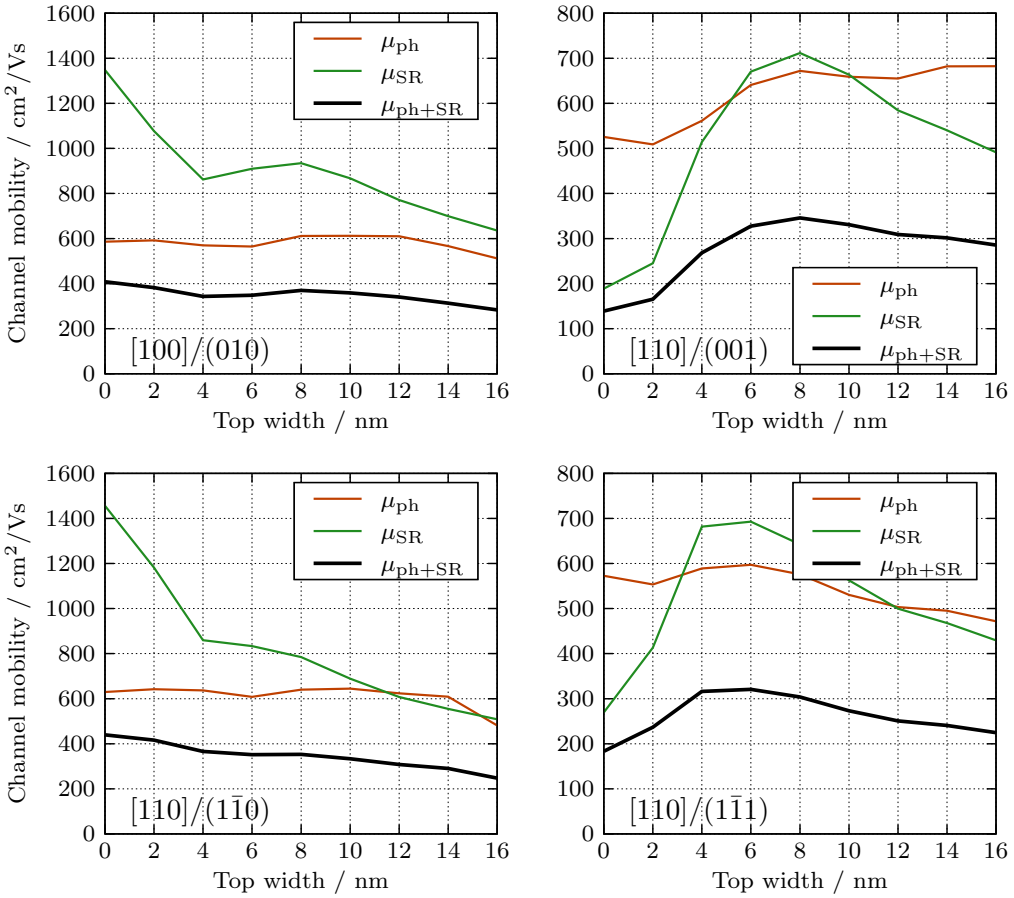


Figure 5.4: Breaking down the mobility into the contributing scattering mechanisms reveals that surface roughness scattering (SRS) is mainly responsible for the orientation-dependent behavior. SRS-limited mobility increases with thinning of the fin top for [100]/(010). In [110]/(001) direction the picture is quite different: SRS-limited mobility increases with tapering, reaching a local maximum and then plummeting as the triangular shape is approached; [110]/(1$\bar{1}$0) does not appear to suffer from increased SRS and exhibits almost the same behavior as [100]/(010). The explanation for this is that electrons scatter more off rough {110} and {111} surfaces than they do off {100} surfaces [85]. In the [100]/(010) and [110]/(1$\bar{1}$0) channels electrons face the rough sidewalls roughly at {100}, while in the [100]/(001) channel sidewall are approximately {110}-oriented.

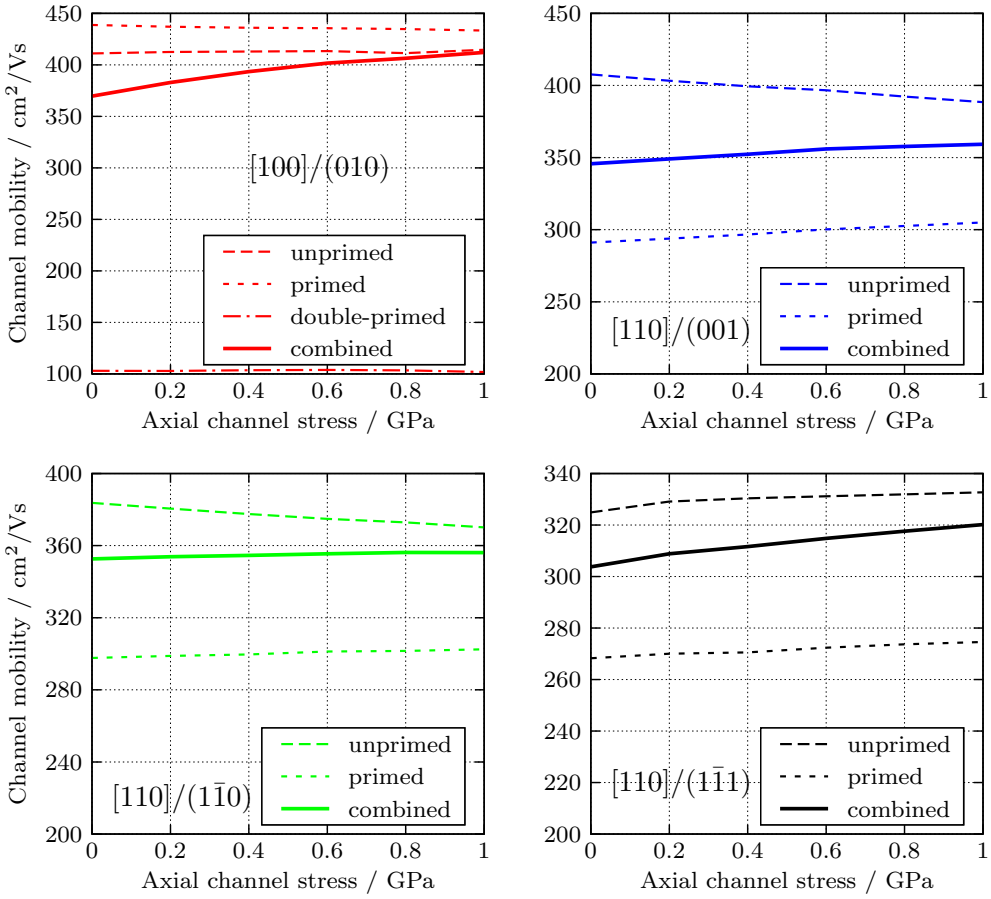


Figure 5.5: One way to increase mobility is valley re-population which can be achieved by confinement or strain. Valleys experience different energy shifts due to strain depending on their orientation, which alters their population. Enhancement through re-population is possible for ⟨100⟩ channels only, where tensile stress causes the heavy-transport-mass valley (double-primed) to shift upwards and become less populated. In non-planar channels such as fins the valley with heavy transport mass is already shifted almost out of reach due to confinement and the mobility is close to its maximum possible value. Thus, mobility enhancement by re-population is diminished in fins compared to planar MOS and UTB channels.

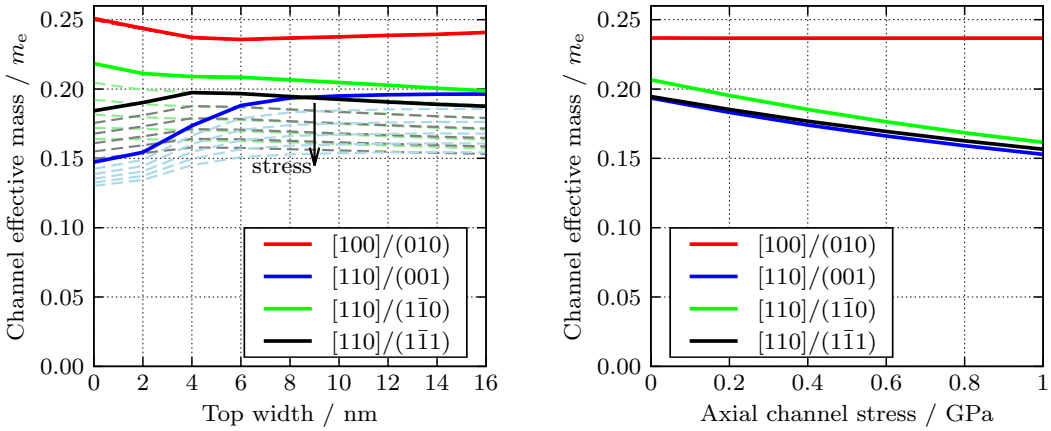


Figure 5.6: The key figure is the transport effective mass; a lighter effective mass directly results in higher mobility due to a $\mu \propto m_{\text{eff}}^{-3/2}$ relation in one-dimensional electron gases [45]. The increased transport mass in the [100]/(010) channel results in a mobility reduction of $\approx 25\%$ which cannot be reversed using strain. Channels in $\langle 110 \rangle$ direction do not have this penalty; in fact, their transport mass can be reduced below the bulk value by tensile stress along the channel. The resulting mobility enhancement can be as big as $\approx 30\%$ for 600 MPa which is in excellent agreement with experimental observations from [86].

Effective mass change is a second-order effects and can only be modeled using higher-order band structure models, such as **k·p**. Shear strain can affect the coupling energies between bands (see Section 2.1.1) which causes a change in the resulting effective mass of the (sub-)bands. A smaller effective mass results in higher mobility and injection velocity, both of which are figures of merit in device design.

A **k·p**-analysis of the subband structure reveals on one hand that [100]/(010)-oriented devices have a larger transport mass due to confinement and, hence, a lower mobility. The mass does not change with uniaxial $\langle 100 \rangle$ stress because it does not produce any shear strain. The combination [110]/(1$\bar{1}$0) on the other hand not only shows confinement-induced transport mass decrease but also that mass can be reduced below bulk level (0.196) using tensile stress along the channel (Fig. 5.6). The stress-induced mass reduction results in a mobility enhancement of $\approx 30\%$ for 600 MPa which is in excellent agreement with experimental data [86].

5.1.2 22 nm p-Type Silicon FinFET

The second example device studied is a 22 nm PMOS tri-gate transistor, the cross-section of which can be seen in Fig. 5.7. The simulation here is fully **k·p**-based [87], i.e. the models used are **SCLoop** and **KdotPQuantum** (Section 4.3.3), along with **KdotPLowField** (Section 4.4.2). This is necessary due to the warped structure of the valence band. The

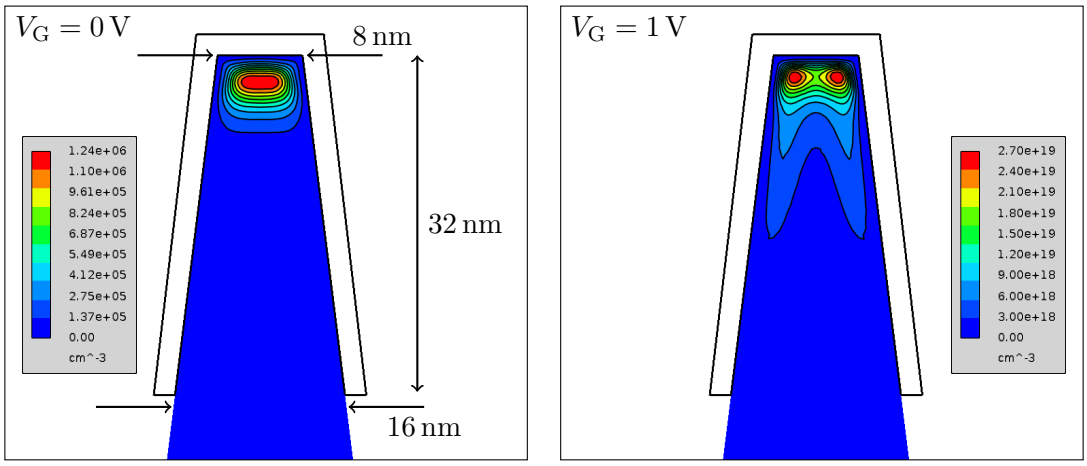


Figure 5.7: Hole concentration profile at low (left) and high (right) gate voltage for channel/substrate orientation [100]/(010); the holes exhibit true two-dimensional confinement in the channel, and are centered below the fin top at low inversion densities. At high inversion densities the holes form regions of quasi-one-dimensional inversion near the sidewalls.

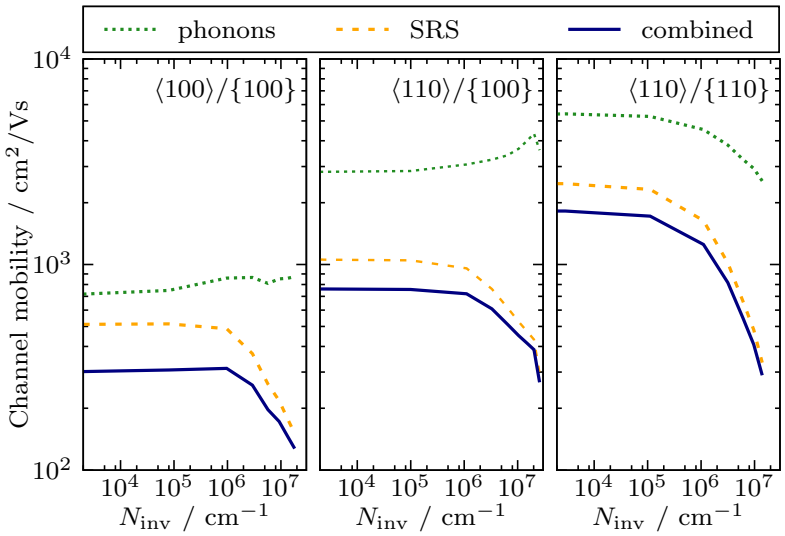


Figure 5.8: Mobility vs. hole inversion density; [100]/(010) channel mobility is below [110]/(1$\bar{1}$0). Mobility is mostly limited by SRS which causes all three curves to decrease at high inversion densities. Interestingly, phonon-limited mobility increases slightly for [100]/(010) and the [110]/(001) channel due to energy-separation of the heavy and light hole bands by the gate field.

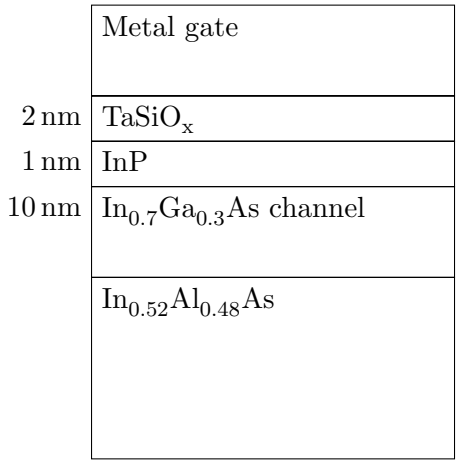


Figure 5.9: Structure of the MIS capacitor fabricated in [88]; the layers are assumed to be undoped.

six-valley effective mass model for holes (Section 2.1.2) could also be applied here, but it would need to be calibrated against **k**·**p**-results first. The hole concentration computed by the self-consistent loop is shown in Fig. 5.7.

Similar to what was done for the NMOS FinFET, we look at different channel/substrate orientations of the PMOS FinFET and evaluate the respective performances. The investigated channel/substrate orientations are [100]/(010), [110]/(001), and [110]/(1$\bar{1}$0). Channel mobility was computed for various gate voltages and is displayed as function of inversion density in Fig. 5.8. There are significant differences in the mobility curves of the different orientations with [110]/(1$\bar{1}$0) exhibiting the best and [100]/(010) the worst transport properties.

The favorable performace trend for [110]/(001) and [110]/(1$\bar{1}$0) orientations is consistent with the performance benefits observed in the NMOS device. Again, [110]/(1$\bar{1}$0) orientation shows the best performance, suggesting that developing a high-performance CMOS process based on {110}-surface wafers may be worthwhile.

5.1.3 InGaAs-Based Devices

Quantum-well MISFETs based on III/V materials are promising candidates for low-power high-performance logic applications [88]. They benefit from the high electron mobilities of the III/V materials and exhibit good channel control due the substrate material acting as a back-barrier – similar to a SOI MOSFET – but without degradation of thermal conductivity in the substrate.

In this section, two simulation studies are presented – a planar InGaAs MISFET and a InGaAs FinFET. Both are based on the same layered structure containing a $In_{0.7}Ga_{0.3}As$-quantum well and a gate stack corresponding to the one fabricated in [88] and shown in Fig. 5.9.

Table 5.1: Electronic structure parameters for the materials of the MIS capacitor

	InGaAs	InP	InAlAs
m_Γ	0.0364	0.0795	0.0733
$m_{\mathrm{l,L}}$	1.478	1.5	1.297
$m_{\mathrm{t,L}}$	0.203	0.408	0.195
$m_{\mathrm{l,X}}$	1.788	1.5	1.493
$m_{\mathrm{t,X}}$	0.249	0.438	0.229
α_Γ	2.0		
α_L	7.0		
α_X	5.0		

Electronic Structure

The electronic structure is modeled using a multi-valley effective mass approach with non-parabolic correction, i.e. the dispersion of each valley is defined implicitly as

$$E_{\mathrm{kin}}(1 + \alpha_v E_{\mathrm{kin}}) = \frac{\hbar}{2}\mathbf{k} \cdot \mathbf{m}_v^{-1} \cdot \mathbf{k}, \tag{5.1}$$

where $\mathbf{m}_v$ and α_v are the valley-specific effective mass tensor and a non-parabolicity correction factor, respectively. Table 5.1 shows a summary of the electronic structure parameters used. The effective masses are based on interpolation rules from [38], the non-parabolic correction factors are interpolated from the values in [89].

The non-parabolic correction is used in the calculation of the subband edge energies by linearizing the relation in Eq. (5.1) and calculating the eigenenergy shifts using first-order perturbation. It is also used to correct the dispersion relation of each subband individually. Numerical quadrature is used to compute the Fermi-Integrals with the non-parabolic correction [90]. The non-parabolic correction also affects electron transport via the density of states and group velocity. The models **EffMassQuantum** from Section 4.3.2 and **EffMassLowField** from Section 4.4.1 were extended accordingly to include the non-parabolic correction.

Dielectric Material

The dielectric used in the presented gate stack is $TaSiO_x$, a high-k dielectric suitable for III/V MISFETs. The material is found in numerous experimental works, often in conjunction with Ta_2O_5. Although [88] does not disclose the exact nature and growth process of the dielectric, one may conclude from [91, 92] that $TaSiO_x$ is in fact a layered structure composed of Ta_2O_5 and SiO_2 layers. Both SiO_2 and Ta_2O_5 are materials with well-known bulk properties, thus the dielectric is modeled as $(Ta_2O_5)_x(SiO_2)_{1\text{-}x}$ and the resulting dielectric constant approximated as

$$\varepsilon_\mathrm{r}^{\mathrm{TaSiO_x}} = \varepsilon_\mathrm{r}^{\mathrm{Ta_2O_5}} x + \varepsilon_\mathrm{r}^{\mathrm{SiO_2}}(1 - x), \tag{5.2}$$

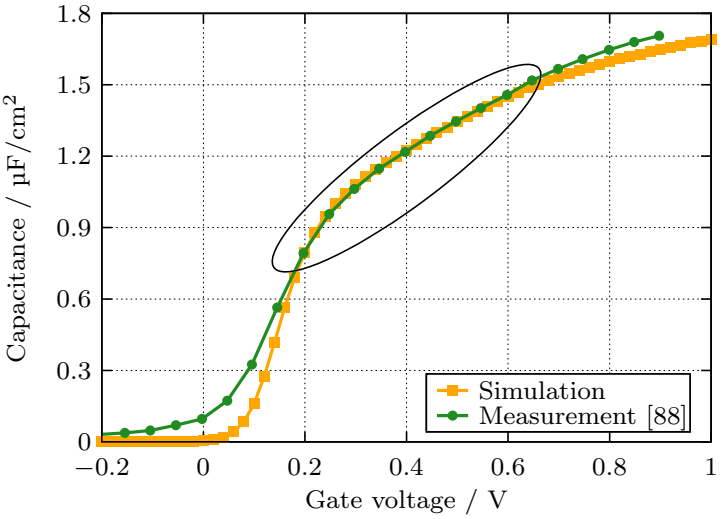


Figure 5.10: Capacitance-voltage curve of the MIS capacitor used to fit the work-function and the composition of the dielectric; the error in the circled region was minimized to avoid the influence of interface traps.

with the layer composition x being treated as a fitting parameter. The values used are $\varepsilon_r^{Ta_2O_5} = 25$ and $\varepsilon_r^{SiO_2} = 3.9$.

Planar InGaAs MISFET

A MIS capacitor reported in [88] has been analyzed (see Fig. 5.9). First, a capacitance-voltage (C/V) curve is calculated to determine the gate-work-function difference and the composition of the dielectric. Then, using these results, the channel conductivity is calculated using the **EffMassLowField** models along with the appropriate scattering models.

Capacitance-voltage curve These curves are simulated using a self-consistent Schrödinger-Poisson loop. The simulations are then fed into an automated optimizer to fit the simulated C/V curve against the measured one. The optimizer is configured to match the region above the threshold-*knee* in the C/V curve, thereby omitting the contribution of interface traps which we have not included in our model.

The C/V curve was successfully fitted yielding a metal-gate work function difference of $-0.4575\,\mathrm{eV}$ with respect to the valence band edge of the InGaAs-layer, and a dielectric composition of $x = 0.2404$, resulting in a $\varepsilon_r^{TaSiO_x} = 8.9722$.

Transconductance curve Using the fitted parameters, the transconductance of a long channel device ($L_G = 40\,\mu\mathrm{m}$) is simulated. The transconductance was also characterized in [88]. The roughness power spectrum at the $TaSiO_x$/InP-interface is assumed to be exponential with an RMS-amplitude of 3 Å and an autocorrelation length of 40 Å. The

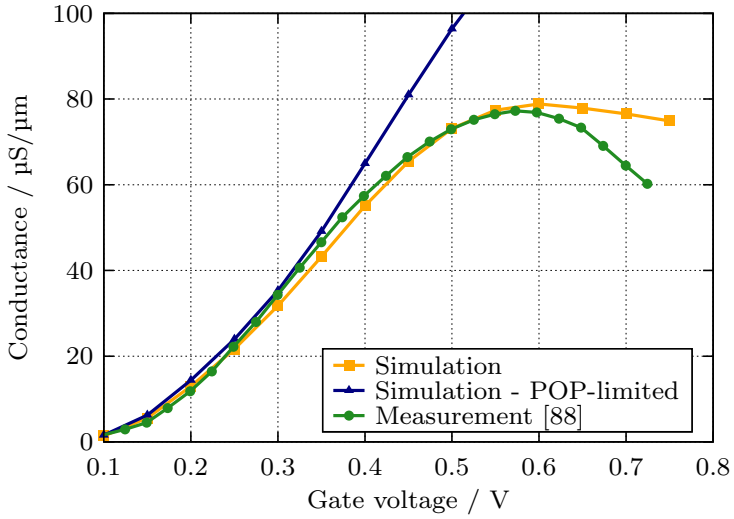


Figure 5.11: Transconductance versus gate voltage for a 40 µm long channel; the simulated curve closely matches the measured one. Up to 0.4 V gate voltage the channel current is limited almost exclusively by POP scattering. For higher gate voltages SRS becomes also important.

simulated result is in good agreement with the measured curve as shown in Fig. 5.11. A conductivity simulation with only POP scattering enabled shows, that POP scattering is by far the dominant and thus conductivity-limiting scattering process. This result is achieved without any parameter-fitting for POP scattering, leaving the interface roughness as the only source of uncertainty.

Figure 5.12 shows the contribution of all scattering processes to the overall channel mobility. This confirms that electron transport in the channel is limited by POP scattering and SRS on the $TaSiO_x$/InP interface, whereas non-polar phonons and scattering on the InP/InGaAs interface can be neglected. The latter can be explained by the relatively low electron barrier between InGaAs and InP of 0.4 eV, resulting in a much smaller SRS matrix element than for the high $TaSiO_x$/InP-barrier. Furthermore, the InP-layer is readily penetrated by the electron wave function as can be seen in Fig. 5.13.

InGaAs FinFET

The same modeling framework used for the planar InGaAs MISFET is also applied to a FinFET. The device uses the same layer structure as shown in Fig. 5.9. However, the InGaAs layer is 20 nm thick and a 10 nm wide fin is formed from the InGaAs and InP layers grown on top of the InAlAs layer. $TaSiO_x$ and a metal gate are deposited isotropically onto the free surfaces.

The long-channel properties of the InGaAs FinFET are analyzed by computing the transconductance curve of a 40 µm long gated fin. The transconductance, shown in Fig. 5.14, is seen to be dominated by POP scattering and surface-roughness scattering. The mobilities due to each scattering process in Fig. 5.15 give a more detailed picture:

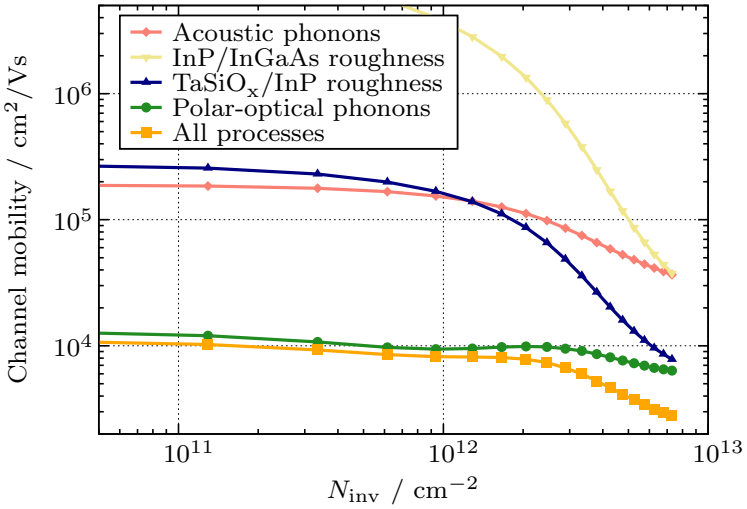


Figure 5.12: Electron mobility of the MIS capacitor plotted against the inversion density; POP scattering clearly dominates electron transport. SRS on the TaSiO$_x$/InP-interface is the second most important process.

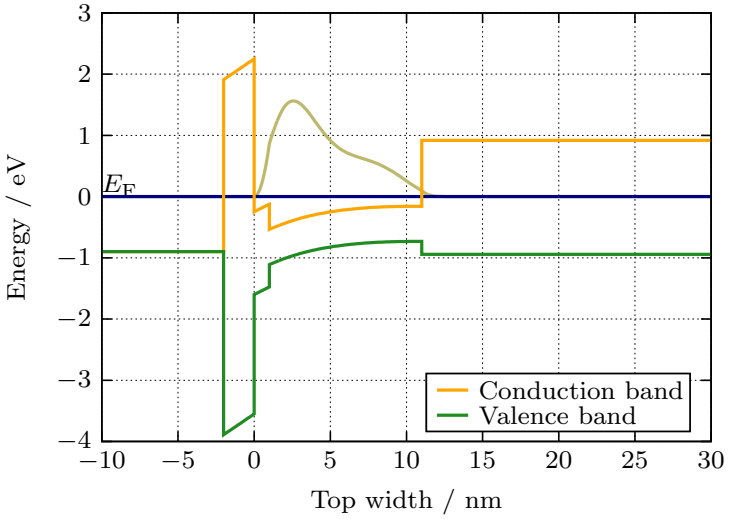


Figure 5.13: Band edges and electron concentration profile in the MIS capacitor at high gate bias (1 V)

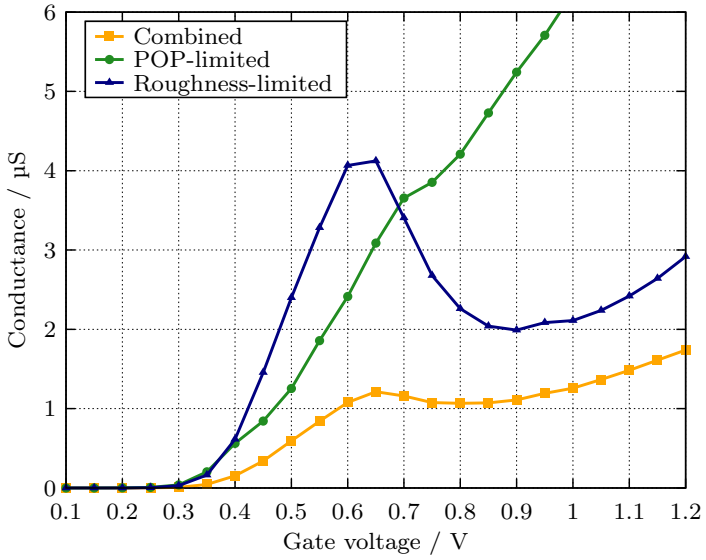


Figure 5.14: Transconductance curve for the 40 µm long InGaAs FinFET; the pi-gated fin has a rectangular shape, 10 nm wide, 20 nm high with a 1 nm InP-layer on top, and is surrounded by 2 nm-layer of $TaSiO_x$ and a metal gate. POP scattering and SRS dominate transport in the channel. Starting from 0.6 V conductance becomes severely degraded by SRS.

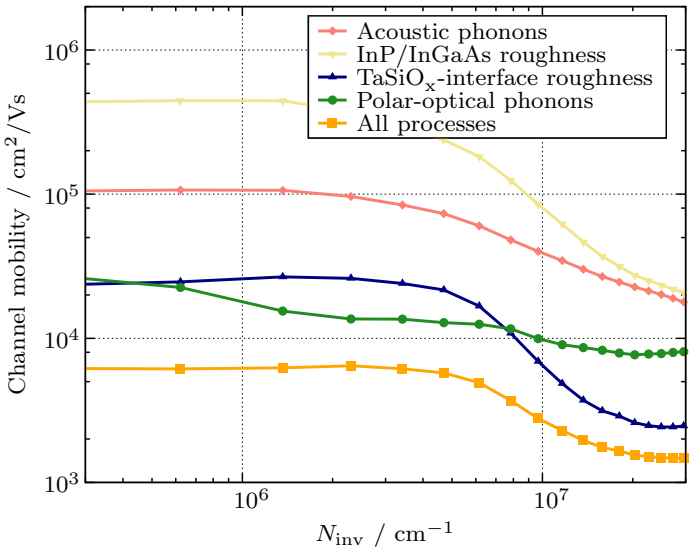


Figure 5.15: Electron mobility versus inversion density in the channel of the InGaAs FinFET; POP scattering and SRS clearly dominate electron transport. SRS on the $TaSiO_x$-interfaces is far more important than SRS on the InGaAs/InP-interface.

POP limits the mobility in the low-density regime, while scattering off the rough $TaSiO_x$ interface limits mobility at higher densities. In the FinFET, surface roughness scattering also occurs at the sidewalls, becoming the main mobility-limiting process.

5.2 Junction-Less Channels

Contrary to the inversion-mode device, the junction-less device [13] has a high, uniform doping in source, drain, and channel. Without a gate, the junction-less transistor is simply a resistor with enough majority carriers in the channel for current to flow, resulting in a normally-on device. Thus, the gate potential actively squeezes the majority carriers out of the channel, creating a potential barrier that controls current flow from source to drain.

While an inversion-mode transistor attracts carriers towards the semiconductor/insulator-interface, the junction-less transistor pinches the channel, repelling carriers from the interface. This difference in operation is shown in Fig. 5.16 for a pi-gate channel. The different distribution of carriers in the channel cross-section results in the carrier transport in the junction-less transistor being affected much less by surface roughness scattering. However, at the same time, the junction-less transistor has a higher doping concentration in the channel, resulting in a higher impurity scattering rate. A computational study is performed to further investigate the differences [49].

The analyzed devices are (i) a inversion-mode NMOS transistor with $3 \times 10^{17}\,\mathrm{cm}^{-3}$ acceptor doping in the channel and (ii) a junction-less NMOS transistor with $1 \times 10^{19}\,\mathrm{cm}^{-3}$ donor doping. The channel of both devices is a Si slab with a maximum width of 30 nm and a height of 10 nm and is surrounded by SiO_2. The channel and substrate orientations are [100] and (010), respectively. In both devices, the gate consists of poly-Si and is doped with $1 \times 10^{19}\,\mathrm{cm}^{-3}$ donors for the inversion-mode device and $1 \times 10^{19}\,\mathrm{cm}^{-3}$ acceptors for the junction-less device.

A quantitative comparison between the inversion-mode and a junction-less device is shown in Fig. 5.17. The gate work-function has been adjusted such that both devices have the same sub-threshold response and matching off-currents. Above threshold, the inversion-mode transistor has a much sharper response to the gate voltage, while saturating a high gate voltage due to increasing surface roughness scattering, also visible in Fig. 5.18.

The junction-less device shows a shallow response right above threshold voltage but is not affected by surface roughness - at least not in the shown voltage range. Figure 5.18 shows that mobility is limited by impurity scattering with the impurities being effectively unscreened for most of the operating regime. At high saturation, screening of the impurities becomes effective. But in this regime the junction-less transistor operates in accumulation-mode which causes a sharp increase in surface roughness scattering.

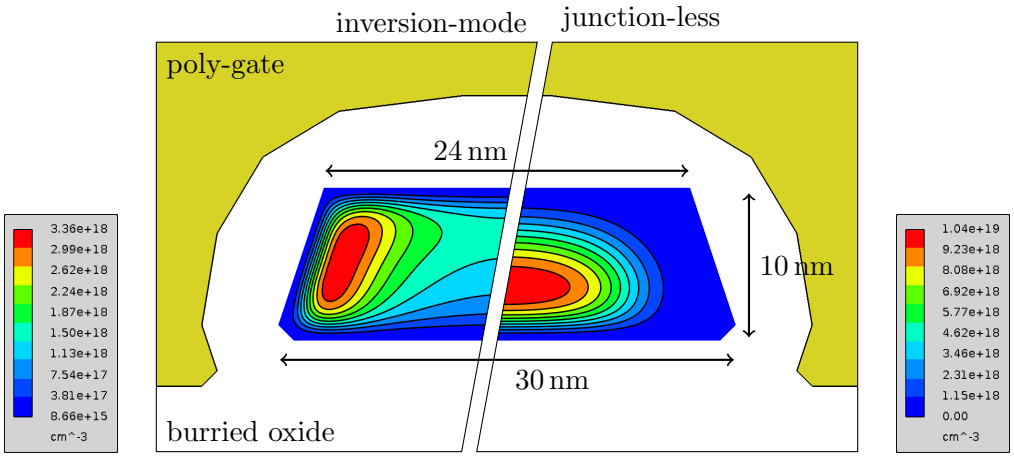


Figure 5.16: Electron concentration in a pi-gate Si channel cross-section; left – inversion-mode device, right – junction-less device; both devices are biased above threshold. In the inversion-mode FET electrons are pushed towards the top and sidewalls while in the junction-less FET electrons remain centered.

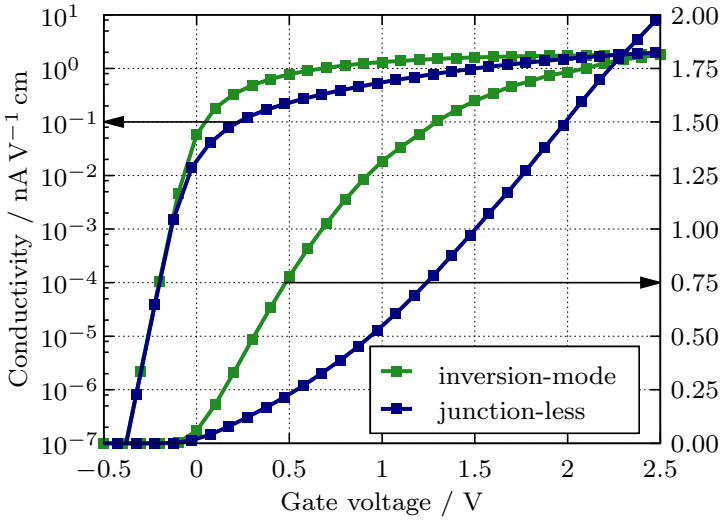


Figure 5.17: Transconductivity for channels from Fig. 5.16; gate work-functions have been adjusted so that both the inversion-mode and the junction-less device have the same threshold voltage. In inversion-mode channels the conductivity saturates due to SRS. In junction-less devices conductivity is lower initially but shows no saturation.

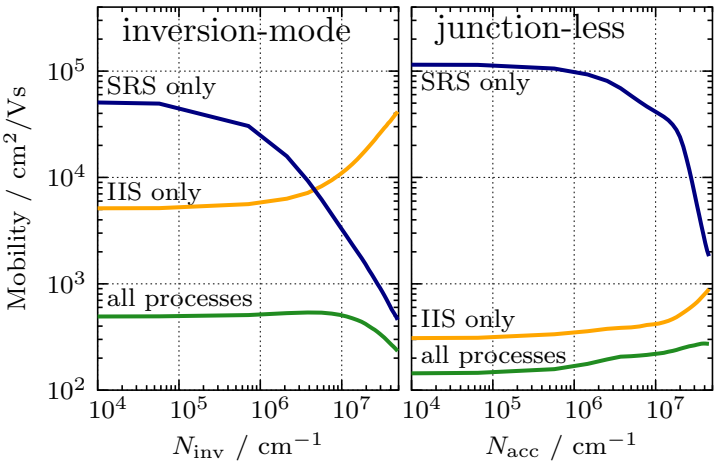


Figure 5.18: Mobility versus inversion/accumulation charge for the channels from Fig. 5.16; left – inversion-mode device, right – junction-less device; the dominant scattering processes are reversed for the two device types. In inversion-mode channels IIS is negligible and SRS dominates when the device is switched on. In the junction-less FET IIS is stronger by an order of magnitude, limiting the mobility, while SRS becomes effective at high densities only.

CHAPTER 6 Conclusion and Outlook

6.1 Conclusion

In the presented work a comprehensive theoretical basis for mobility modeling in nano-scale field-effect transistors has been established, which treats planar and non-planar technologies, n-type and p-type channels, as well as inversion and accumulation-based devices on equal footing. This consistency of modeling increases the confidence in the validity of simulation results. The work comprises two major blocks of innovation.

First, numerous models were contributed in this work to fill gaps that were left open by previous works, such as (i) the six-valley effective-mass model for holes, (ii) a generic model for carrier scattering with polar-optical phonons using the electrostatic Green's function, (iii) a generalized surface-roughness scattering model suitable for non-planar channels of any shape, (iv) an approach to calculating the low-field mobility of a channel by solving the linearized Boltzmann transport equation, and (v) a method to include effective mass anisotropy and scattering process anisotropy in mobility modeling.

Second, a computational framework has been developed to process the models. The main requirement for the framework was to be TCAD-compatible, i.e. to be fast, reliable, and flexible enough for use as a device engineering tool. The developed framework is capable of processing arbitrary device geometries with unstructured meshes. High-efficiency algorithms have been developed that adapt to the problem at hand, e.g. automatically selecting the necessary number of subbands and **k**-points to satisfy a given error tolerance. The modular and layer-based design of the framework allows for relatively easy extension with new models, such as those described in the following section.

6.2 Outlook

This work represents a snapshot of a work in progress. It forms the basis for further works on physical device modeling. Calculating mobilities already required to deal with all major aspects of semiconductor device physics, such as electronic structure modeling, carrier density and self-consistent electrostatics, a comprehensive set of scattering models based on the semi-classical approach, covering most semiconductor materials at room temperature, and, finally, mobility modeling based on the linearized Boltzmann transport

equation. The next development steps point towards extending the presented modeling framework to simulate entire devices instead of device cross-sections.

6.2.1 Boltzmann Transport Equation in Phase Space

To calculate the current in a narrow-channel devices, such as FDSOI FETs, nanowire FETs, and FinFETs, the steady-state subband Boltzmann transport equation from Eq. (2.129) needs to be solved. The right-hand side of the equation containing the scattering operator has already been dealt with. The methods for evaluating the scattering operator have been presented in this work. The *free-streaming operator* at left-hand side governs the ballistic transport in the BTE. It consists of two differentials, one acting in real-space, the other in $\mathbf{k}$-space. The velocity coefficient $\mathbf{v}_n(\mathbf{k})$ contains the subband structure information.

The solution variable, i.e. the distribution function, is a function of both real and $\mathbf{k}$-space coordinates. The Cartesian product of real and $\mathbf{k}$-space is called the *phase space*. To solve the steady-state BTE, the phase space of the device's channel needs to be constructed, along with a grid, and the free-streaming operator discretized within it.

Such a phase-space BTE solver is able to simulate carrier transport in devices at high fields, including effects such as velocity overshoot, carrier heating, and quantum resistance. Being an extension of the work presented here, it also allows to include all the aspects covered in physical mobility modeling, such as doping, channel geometry, crystal orientation, strain, and different channel materials. The phase-space transport can either be based on effective mass or on a $\mathbf{k}\cdot\mathbf{p}$ model of the electronic structure. A prototype of a phase-space BTE solver is presented in [93].

6.2.2 Quantum Transport

A full solution of the Boltzmann transport equation as discussed in the previous section is able to exhaust the capabilities of the semi-classical modeling framework. Going beyond semi-classical transport leads to the domain of quantum transport, where the wave-nature of electrons and holes is fully appreciated in both carrier propagation and carrier scattering. Rather than solving a particle equation in phase-space, a wave equation is solved in real space using *open boundary conditions* at the devices contacts. The two most popular methods for quantum transport are the Quantum Transmitting Boundary Method (QTBM) which operates in the *Schrödinger picture* and is wave-function-based, and the Non-Equilibrium Green's Function method (NEGF) which operates in the *interaction picture* and is operator-based. A third method would be the *Wigner method* based on the Wigner equation, which can be seen as an extension the Boltzmann transport equation.

The general-purpose computational framework described in Chapter 3 contains components that can be used to model quantum transport based on the QTBM and NEGF methods. The facilities for dealing with geometry, topology, and discretization can be re-used. The expression system, problem specification, and automatic assembly allow to specify both parabolic and $\mathbf{k}\cdot\mathbf{p}$-based Hamiltonians, as done for the closed-boundary problems in this work. The aforementioned open boundary conditions are based on the

plane wave boundary condition in Eq. (3.5) which is readily imposed using the existing assembly components. The wave-function-based QTBM benefits from infrastructure for solving sparse systems. A prototype of a QTBM solver for intra-band and band-to-band source-drain tunneling is presented in [93–95].

The NEGF method allows to include self-consistent scattering in carrier transport, which is its most important advantage. However, naïvely implemented, NEGF calculations quickly become computationally intractable for devices with technologically relevant dimensions. A practical, TCAD-compatible extension to NEGF requires additional work regarding efficient matrix-inversion techniques.

Bibliography

[1] G. Moore. "Cramming More Components Onto Integrated Circuits". In: *Proceedings of the IEEE* 86.1 (Jan. 1998), pp. 82–85. DOI: 10.1109/JPROC.1998.658762.

[2] S. Thompson et al. "A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 µm^2 SRAM cell". In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 2002, pp. 61–64. DOI: 10.1109/IEDM.2002.1175779.

[3] P. Bai et al. "A 65 nm logic technology featuring 35 nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 µm^2 SRAM cell". In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 2004, pp. 657–660. DOI: 10.1109/IEDM.2004.1419253.

[4] K. Mistry et al. "A 45 nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193 nm Dry Patterning, and 100% Pb-free Packaging". In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 2007, pp. 247–250. DOI: 10.1109/IEDM.2007.4418914.

[5] S. Natarajan et al. "A 32 nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and 0.171 µm^2 SRAM cell size in a 291 Mb array". In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 2008, pp. 1–3. DOI: 10.1109/IEDM.2008.4796777.

[6] C. Auth et al. "A 22 nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors". In: *2012 Symposium on VLSI Technology, Digest of Technical Papers*. 2012, pp. 131–132. DOI: 10.1109/VLSIT.2012.6242496.

[7] H.-S. Wong, K. Chan, and Y. Taur. "Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel". In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 1997, pp. 427–430. DOI: 10.1109/IEDM.1997.650416.

[8] M. Jurczak et al. "SON (silicon on nothing)-a new device architecture for the ULSI era". In: *1999 Symposium on VLSI Technology, Digest of Technical Papers*. June 1999, pp. 29–30. DOI: 10.1109/VLSIT.1999.799324.

[9] J.-H. Lee et al. "Super self-aligned double-gate (SSDG) MOSFETs utilizing oxidation rate difference and selective epitaxy". In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 1999, pp. 71–74. DOI: 10.1109/IEDM.1999.823849.

[10] K. Guarini et al. "Triple-self-aligned, planar double-gate MOSFETs: devices and circuits". In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 2001, pp. 19.2.1–19.2.4. DOI: 10.1109/IEDM.2001.979527.

[11] S. Harrison et al. "Highly performant double gate MOSFET realized with SON process". In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 2003, pp. 18.6.1–18.6.4. DOI: 10.1109/IEDM.2003.1269319.

[12] S. Bangsaruntip et al. "High performance and highly uniform gate-all-around silicon nanowire MOSFETs with wire size dependent scaling". In: *International Electron Devices Meeting (IEDM) Technical Digest*. 2009, pp. 1–4. DOI: 10.1109/IEDM.2009.5424364.

[13] J.-P. Colinge et al. "Nanowire transistors without junctions". In: *Natutre Nano* 5.3 (Mar. 2010), pp. 225–229.

[14] J. E. Lilienfeld. "Method and apparatus for controlling electric currents". Patent US 1745175. 1930.

[15] J. E. Lilienfeld. "Device for controlling electric current". Patent US 1900018. 1933.

[16] M. Lundstrom and Z. Ren. "Essential physics of carrier transport in nanoscale MOSFETs". In: *IEEE Transactions on Electron Devices* 49.1 (Jan. 2002), pp. 133–141. DOI: 10.1109/16.974760.

[17] G. L. Bir and G. Pikus. *Symmetry and Strain-Induced Effects in Semiconductors*. Wiley, 1974.

[18] P. Y. Yu and M. Cardona. *Fundamentals of Semiconductors*. Springer, 2005.

[19] G. Dresselhaus, A. F. Kip, and C. Kittel. "Cyclotron Resonance of Electrons and Holes in Silicon and Germanium Crystals". In: *Physical Review* 98 (2 Apr. 1955), pp. 368–384. DOI: 10.1103/PhysRev.98.368.

[20] J. M. Luttinger. "Quantum Theory of Cyclotron Resonance in Semiconductors: General Theory". In: *Physical Review* 102 (4 May 1956), pp. 1030–1041. DOI: 10.1103/PhysRev.102.1030.

[21] T. Manku and A. Nathan. "Valence energy-band structure for strained group-IV semiconductors". In: *Journal of Applied Physics* 73.3 (1993), pp. 1205–1213. DOI: 10.1063/1.353287.

[22] J. R. Chelikowsky and M. L. Cohen. "Electronic structure of silicon". In: *Physical Review B: Condensed Matter and Materials Physics* 10 (12 Dec. 1974), pp. 5095–5107. DOI: 10.1103/PhysRevB.10.5095.

[23] J. P. Dismukes, L. Ekstrom, and R. J. Paff. "Lattice Parameter and Density in Germanium-Silicon Alloys". In: *The Journal of Physical Chemistry* 68.10 (1964), pp. 3021–3027. DOI: 10.1021/j100792a049.

[24] K. Takeda, A. Taguchi, and M. Sakata. "Valence-band parameters and hole mobility of Ge-Si alloys-theory". In: *Journal of Physics C: Solid State Physics* 16.12 (1983), p. 2237.

[25] P. Lawaetz. "Valence-Band Parameters in Cubic Semiconductors". In: *Physical Review B: Condensed Matter and Materials Physics* 4 (10 Nov. 1971), pp. 3460–3467. DOI: 10.1103/PhysRevB.4.3460.

[26] J. S. Kline, F. H. Pollak, and M. Cardona. "Electroreflectance in Ge-Si alloys". In: *Helvetica Physica Acta* 41 (1968), p. 968.

[27] T. Ebner et al. "Electroreflectance spectroscopy of strained $Si_{1-x}Ge_x$ layers on silicon". In: *Physical Review B: Condensed Matter and Materials Physics* 57 (24 June 1998), pp. 15448–15453. DOI: 10.1103/PhysRevB.57.15448.

[28] D. Aspnes and A. Studna. "Direct observation of the E_0 and $E_0 + \Delta_0$ transitions in silicon". In: *Solid State Communications* 11.10 (1972), pp. 1375–1378. DOI: http://dx.doi.org/10.1016/0038-1098(72)90546-7.

[29] J. A. Van Vechten. "Quantum Dielectric Theory of Electronegativity in Covalent Systems. I. Electronic Dielectric Constant". In: *Physical Review* 182 (3 June 1969), pp. 891–905. DOI: 10.1103/PhysRev.182.891.

[30] F. Schäffler. "Silicon-Germanium". In: *Properties of Advanced Semiconductor Materials*. Ed. by M. E. Levinstein, S. L. Rumyantsev, and M. S. Shur. Wiley, 2001.

[31] J. A. Van Vechten. "Quantum Dielectric Theory of Electronegativity in Covalent Systems. II. Ionization Potentials and Interband Transition Energies". In: *Physical Review* 187 (3 Nov. 1969), pp. 1007–1020. DOI: 10.1103/PhysRev.187.1007.

[32] F. L. Madarasz, J. E. Lang, and P. M. Hemeger. "Effective masses for nonparabolic bands in p-type silicon". In: *Journal of Applied Physics* 52.7 (1981), pp. 4646–4648. DOI: 10.1063/1.329345.

[33] C. G. Van de Walle. "Band lineups and deformation potentials in the model-solid theory". In: *Physical Review B: Condensed Matter and Materials Physics* 39 (3 Jan. 1989), pp. 1871–1883. DOI: 10.1103/PhysRevB.39.1871.

[34] M. Levinshtein and S. Rumyantsev. "Silicon". In: *Handbook Series on Semiconductor Parameters*. Ed. by M. Levinstein, S. Rumyantsev, and M. Shur. Vol. 1. World Scientific, 2000. Chap. 1.

[35] J. Wiley. "Chapter 2 Mobility of Holes in III-V Compounds". In: *Semiconductors and Semimetals*. Ed. by R. Willardson and A. C. Beer. Vol. 10. Semiconductors and Semimetals. Elsevier, 1975, pp. 91–174. DOI: http://dx.doi.org/10.1016/S0080-8784(08)60332-4.

[36] M. Levinshtein and S. Rumyantsev. "Germanium". In: *Handbook Series on Semiconductor Parameters*. Ed. by M. Levinstein, S. Rumyantsev, and M. Shur. Vol. 1. World Scientific, 2000. Chap. 1.

[37] E. O. Kane. "Energy band structure in p-type germanium and silicon". In: *Journal of Physics and Chemistry of Solids* 1.1-2 (1956), pp. 82–99. DOI: http://dx.doi.org/10.1016/0022-3697(56)90014-2.

[38] I. Vurgaftman, J. R. Meyer, and L. R. Ram-Mohan. "Band parameters for III-V compound semiconductors and their alloys". In: *Journal of Applied Physics* 89.11 (2001), pp. 5815–5875. DOI: http://dx.doi.org/10.1063/1.1368156.

[39] J. C. Hensel, H. Hasegawa, and M. Nakayama. "Cyclotron Resonance in Uniaxially Stressed Silicon. II. Nature of the Covalent Bond". In: *Physical Review* 138.1A (Apr. 1965), A225–A238. DOI: 10.1103/PhysRev.138.A225.

[40] V. A. Sverdlov et al. "Mobility Modeling in Advanced MOSFETs with Ultra-Thin Silicon Body under Stress". In: *JICS* 4.2 (2009), pp. 55–60.

[41] R. L. Aggarwal, M. D. Zuteck, and B. Lax. "Nonparabolicity of the L_1 Conduction Band in Germanium from Magnetopiezotransmission Experiments". In: *Physical Review Letters* 19 (5 July 1967), pp. 236–238. DOI: 10.1103/PhysRevLett.19.236.

[42] G. Dresselhaus. "Electronic energy bands in semiconductors with cubic crystal structure". unpublished. PhD thesis. University of California, 1955.

[43] P. Li, Y. Song, and H. Dery. "Intrinsic spin lifetime of conduction electrons in germanium". In: *Physical Review B: Condensed Matter and Materials Physics* 86 (8 Aug. 2012), p. 085202. DOI: 10.1103/PhysRevB.86.085202.

[44] M. Karner et al. "Bringing physics to device design – A fast and predictive device simulation framework". In: *2015 IEEE Silicon Nanoelectronics Workshop (SNW)*. June 2015.

[45] Z. Stanojevic et al. "Subband engineering in n-type silicon nanowires using strain and confinement". In: *Solid-State Electronics* 70 (2012), pp. 73–80. DOI: 10.1016/j.sse.2011.11.022.

[46] F. Stern and W. Howard. "Properties of Semiconductor Surface Inversion Layers in the Electric Quantum Limit". In: *Physical Review* 163 (3 Nov. 1967), pp. 816–835. DOI: 10.1103/PhysRev.163.816.

[47] M. Lundstrom. *Fundamentals of carrier transport*. Cambridge University Press, 2009.

[48] B. Ridley. *Quantum Processes in Semiconductors*. Oxford University Press, 2013.

[49] Z. Stanojević et al. "Consistent low-field mobility modeling for advanced MOS devices". In: *Solid-State Electronics* 112 (2015), pp. 37–45. DOI: http://dx.doi.org/10.1016/j.sse.2015.02.008.

[50] D. K. Ferry, S. M. Goodnick, and J. Bird. *Transport in Nanostructures*. Cambridge University Press, 2009.

[51] A. Trellakis et al. "Iteration scheme for the solution of the two-dimensional Schrödinger-Poisson equations in quantum structures". In: *Journal of Applied Physics* 81.12 (1997), pp. 7880–7884. DOI: 10.1063/1.365396.

[52] O. Baumgartner et al. "VSP–a quantum-electronic simulation framework". In: *Journal of Computational Electronics* 12 (2013), pp. 701–721. DOI: 10.1007/s10825-013-0535-y.

[53] O. Baumgartner, Z. Stanojevic, and H. Kosina. "Efficient simulation of quantum cascade lasers using the Pauli master equation". In: *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. Sept. 2011, pp. 91–94. DOI: 10.1109/SISPAD.2011.6035057.

[54] H. Fröhlich. "Theory of Electrical Breakdown in Ionic Crystals". In: *Proceedings of the Royal Society A*. Vol. 160. 1937, pp. 230–241.

[55] D. Vasileska, S. Goodnick, and G. Klimeck. *Computational Electronics: Semiclassical and Quantum Device Modeling and Simulation*. CRC Press, 2010.

[56] Z. Stanojevic et al. "Predictive physical simulation of III/V quantum-well MISFETs for logic applications". In: *Solid State Device Research Conference (ESSDERC), 2015 45th European*. Sept. 2015, pp. 310–313. DOI: 10.1109/ESSDERC.2015.7324776.

[57] Z. Stanojevic et al. "New computational perspectives on scattering and transport in III/V channel materials". In: *2015 International Workshop on Computational Electronics (IWCE)*, Sept. 2015. DOI: 10.1109/IWCE.2015.7301985.

[58] R. E. Prange and T. Nee. "Quantum Spectroscopy of the Low-Field Oscillations in the Surface Impedance". In: *Physical Review* 168 (3 Apr. 1968), pp. 779–786. DOI: 10.1103/PhysRev.168.779.

[59] S. M. Goodnick et al. "Surface roughness at the Si(100)-SiO_2 interface". In: *Physical Review B: Condensed Matter and Materials Physics* 32 (12 Dec. 1985), pp. 8171–8186. DOI: 10.1103/PhysRevB.32.8171.

[60] Z. Stanojevic and H. Kosina. "Modeling Surface-Roughness-Induced Scattering in Non-Planar Silicon Nanostructures". In: *2013 IEEE Silicon Nanoelectronics Workshop (SNW)*. 2013, pp. 132–133.

[61] Z. Stanojevic and H. Kosina. "Surface-Roughness-Scattering in Non-Planar Channels – the Role of Band Anisotropy". In: *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. 2013, pp. 352–355. DOI: 10.1109/SISPAD.2013.6650647.

[62] M. V. Fischetti and S. E. Laux. "Band structure, deformation potentials, and carrier mobility in strained Si, Ge, and SiGe alloys". In: *Journal of Applied Physics* 80.4 (1996), pp. 2234–2252. DOI: http://dx.doi.org/10.1063/1.363052.

[63] C. Jungemann and B. Meinerzhagen. *Hierarchical device simulation: The Monte-Carlo perspective*. Springer, 2003.

[64] C. Herring and E. Vogt. "Transport and Deformation-Potential Theory for Many-Valley Semiconductors with Anisotropic Scattering". In: *Physical Review* 101 (3 Feb. 1956), pp. 944–961. DOI: 10.1103/PhysRev.101.944.

[65] *Vienna Schrödinger-Poisson*. http://www.globaltcad.com/vsp.

[66] *GTS Framework*. http://www.globaltcad.com/framework.

[67] Z. Stanojevic et al. "A versatile finite volume simulator for the analysis of electronic properties of nanostructures". In: *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. Sept. 2011, pp. 143–146. DOI: 10.1109/SISPAD.2011.6035089.

[68] K. M. Kramer and W. N. G. Hitchon. *Semiconductor Devices*. Ed. by J. Czerwinski. Prentice Hall, 1997.

[69] R. Klima et al. "Controlling TCAD Applications with an Object-Oriented Dynamic Database". In: *15th European Simulation Multiconference*. 2001, pp. 161–165.

[70] T. A. Davis. *Direct Methods for Sparse Linear Systems (Fundamentals of Algorithms 2)*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2006.

[71] Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, 1996.

[72] X. S. Li. "An Overview of SuperLU: Algorithms, Implementation, and User Interface". In: *ACM Transactions on Mathematical Software* 31.3 (Sept. 2005), pp. 302–325.

[73] O. Schenk, M. Bollhöfer, and R. Römer. "On Large-Scale Diagonalization Techniques for the Anderson Model of Localization". In: *SIAM Review* 50.1 (2008), pp. 91–112. DOI: 10.1137/070707002.

[74] R. Lehoucq, D. Sorensen, and C. Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, 1998.

[75] J. Demmel et al. *Templates for the solution of algebraic eigenvalue problems: a practical guide*. Ed. by Z. Bai. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2000.

[76] *CMake*. https://cmake.org/.

[77] S. Steiger et al. "NEMO5: A Parallel Multiscale Nanoelectronics Modeling Tool". In: *IEEE Transactions on Nanotechnology* 10.6 (2011), pp. 1464–1474. DOI: 10.1109/TNANO.2011.2166164.

[78] D. E. Knuth. "Literate programming". In: *The Computer Journal* 27.2 (1984), pp. 97–111.

[79] *OASIS Darwin Information Typing Architecture (DITA) Version 1.2 Specification*. OASIS, 2010.

[80] Z. Stanojevic et al. "On the Validity of Momentum Relaxation Time in Low-Dimensional Carrier Gases". In: *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. 2014, pp. 181–184. DOI: 10.1109/SISPAD.2014.6931593.

[81] Z. Stanojevic et al. "Full-Band Modeling of Mobility in p-Type FinFETs". In: *2014 IEEE Silicon Nanoelectronics Workshop (SNW)*. 2014, pp. 83–84.

[82] Z. Stanojevic et al. "Advanced Numerical Methods for Semi-Classical Transport Simulation in Ultra-Narrow Channels". In: *Abstracts of The 18th European Conference on Mathematics for Industry (ECMI)*. 2014.

[83] A. Neumaier. "Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization". In: *SIAM Review* 40.3 (1998), pp. 636–666. DOI: 10.1137/S0036144597321909.

[84] C. T. Blog. *Intel's 22-nm Tri-gate Transistors Exposed*. Chipworks. Apr. 2012. URL: http://www.chipworks.com/blog/technologyblog/2012/04/23/intels-22-nm-tri-gate-transistors-exposed/.

[85] S. Takagi et al. "On the universality of inversion layer mobility in Si MOSFET's: Part II-effects of surface orientation". In: *IEEE Transactions on Electron Devices* 41.12 (1994), pp. 2363–2368. DOI: 10.1109/16.337450.

[86] S. Bangsaruntip et al. "Gate-all-around silicon nanowire 25-stage CMOS ring oscillators with diameter down to 3 nm". In: *VLSIT*. 2010, pp. 21–22. DOI: 10.1109/VLSIT.2010.5556136.

[87] Z. Stanojevic et al. "Full-Band Transport in Ultra-Narrow p-Type Si Channels: Field, Orientation, Strain". In: *Proceedings of the 15th International Conference on Ultimate Integration on Silicon*. 2014, pp. 69–72.

[88] M. Radosavljevic et al. "Non-planar, multi-gate InGaAs quantum well field effect transistors with high-K gate dielectric and ultra-scaled gate-to-drain/gate-to-source separation for low power logic applications". In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 2010, pp. 6.1.1–6.1.4. DOI: 10.1109/IEDM.2010.5703306.

[89] D. Lizzit et al. "Performance Benchmarking and Effective Channel Length for Nanoscale InAs, $In_{0.53}Ga_{0.47}As$, and sSi n-MOSFETs". In: *IEEE Transactions on Electron Devices* 61.6 (June 2014), pp. 2027–2034. DOI: 10.1109/TED.2014.2315919.

[90] O. Baumgartner et al. "Adaptive Energy Integration of Non-Equilibrium Green's Functions". In: *NSTI Nanotech Proceedings*. 3, 2007, pp. 145–148.

[91] C. Adelmann et al. "Atomic-layer-deposited tantalum silicate as a gate dielectric for III-V MOS devices". In: *Microelectronic Engineering* 88.7 (2011), pp. 1098–1100. DOI: http://dx.doi.org/10.1016/j.mee.2011.03.135.

[92] C. Adelmann et al. “Atomic Layer Deposition of Tantalum Oxide and Tantalum Silicate from Chloride Precursors”. In: *Chemical Vapor Deposition* 18.7-9 (2012), pp. 225–238. DOI: 10.1002/cvde.201106967.

[93] Z. Stanojevic et al. “Physical modeling – A new paradigm in device simulation”. In: *International Electron Devices Meeting (IEDM) Technical Digest*. Dec. 2015, pp. 5.1.1–5.1.4. DOI: 10.1109/IEDM.2015.7409631.

[94] O. Baumgartner et al. “Investigation of quantum transport in nanoscaled GaN high electron mobility transistors”. In: *Simulation of Semiconductor Processes and Devices (SISPAD), 2014 International Conference on*. Sept. 2014, pp. 117–120. DOI: 10.1109/SISPAD.2014.6931577.

[95] L. Filipovic et al. “BTB Tunneling in InAs/Si Heterojunctions”. In: *Proceedings of the 19th International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. 2014, pp. 245–248. DOI: 10.1109/SISPAD.2014.6931609.