DISSERTATION

Two-Dimensional Wigner Monte Carlo Simulation for Time-Resolved Quantum Transport with Scattering

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

Paul Ellinghaus

████████ ████
████████████
███████████████████

Wien, im Februar 2016 _____

Kurzfassung

Halbleiterbauelemente – hauptsächlich Logiktransistoren und Speicherzellen – wurden mittlerweile für ein halbes Jahrhundert erfolgreich in ihrer Größe reduziert, welches eine gesteigerte Rechenleistung und reduzierte Kosten mit sich brachte, wodurch Elektronik ubiquitär in unserem täglichen Leben wurde. Die seit Langem vorausgesehenen durch die Physik bedingten Grenzen der Skalierung rücken derzeit sehr rapide näher und die Erfordernis von neuen Bauelementkonzepten immer stringenter. Auf der Jagd nach neuen Konzepten für Nanoelektronik werden Simulationswerkzeuge weiter an Bedeutung gewinnen und eine fundamentale Rolle bei der Hilfe zu deren Verständnis und der Machbarkeit von Materialien, Bauelementen und Systemen spielen.

Bisher wurde beim Entwurf von nanoelektronischen Bauelementen quantenmechanische Effekte meistens nur berücksichtigt anstatt diese aktiv zu nutzen. Das bessere Verständnis und praktische Erfahrung im Umgang mit der Quantenmechanik ermöglicht es uns quantenmechanische Prinzipien für die Entwicklung von Bauelementen und Systemen gezielt einzusetzen. Die Verfügbarkeit von Simulationswerkzeugen, um beim Verständnis und Entwurf von solchen Bauelementen und Systemen zu helfen, ist von höchster Priorität. Die Simulation des Elektronentransportes in Halbleitern stellt eine fundamentale Simulationsfähigkeit für nanoelektronische Forschung dar und wird in dieser Arbeit durch die Verwendung der Wigner-Boltzmann Gleichung verwirklicht.

Der Wigner-Formalismus bietet eine intuitivere Beschreibung des quantenmechanischen Transportes im Vergleich zur Operatormechanik, da dieser im Phasenraum mit Funktionen und Variablen formuliert ist, was eine Adaption von Modellen und Analogien aus dem semiklassischen Transport erlaubt. Die wichtigste Konsequenz daraus, ist die Möglichkeit die Wigner-Quantentransportgleichung mit Boltzmann-Streuungsmodellen zu erweitern, wodurch die Wigner-Boltzmann Gleichung entsteht.

Die Berücksichtigung von Streuung im Quantentransport ist essenziell, um die Dekohärenz von verschränkten Elektronenzuständen, welche Qubits – die grundlegenden Bausteine für Quantenrechner – darstellen können, zu studieren. Die Simulation des zeitaufgelösten Quantentransports kann dabei helfen das Verhalten von stark miniaturisierten, durch quantenmechanische Effekte bestimmten Schaltkreisen zu verstehen, da diese ein Verhalten aufweisen, z.B. Oszillationen, welches mit klassischer Schaltkreistheorie nicht erklärt werden kann. Derzeit ist der einzige rechengestützte umsetzbare streuungsberücksichtigende zeitaufgelöste Quantentransportformalismus die Wigner-Boltzmann Gleichung.

In dieser Arbeit wird ein Simulationswerkzeug vorgestellt, welches die Gleichung in zwei Dimensionen löst unter Verwendung eines Monte Carlo Zuganges basierend auf Partikeln mit Affinität, die die Quanteninformation trägt – die *signed-particle* Methode. Der Fortschritt der letzten Jahre, speziell im Bezug auf Berechnungsprobleme, Algorithmen und deren praktische Implementierung, wird hier dargelegt.

Die Algorithmen, die in dieser Arbeit präsentiert werden, stellen den aktuellen Stand der Technik der *signed-particle* Methode dar und wurden im Wigner Ensemble Monte Carlo Simulationswerkzeug, als Teil der frei verfügbaren ViennaWD Simulations-Sammlung, implementiert um als Referenzimplementierung zu dienen.

Der entwickelte Simulator erlaubt das Studium von Einzelelektronen in der Wellenpaket-Darstellung. Erste Untersuchungen des dynamischen Verhaltens und der Manipulation von solchen Wellenpaketen, unter Verwendung des Konzeptes von elektrostatischen Linsen, werden gezeigt. Eine Anwendung solch einer Linse, um den Ansteuerungsstrom in einem nano-skalierten Kanal zu erhöhen, wird vorgestellt.

Abstract

Semiconductor devices – primarily logic transistors and memory cells – have now been successfully scaled down in size for half a century, which brought about the increased performance and reduced costs making electronics ubiquitous in our daily lives. However, the long-predicted limits to scaling imposed by physics are now rapidly being approached and the need for novel device concepts is becoming a pressing issue. In this pursuit of novel concepts for nanoelectronics, simulation tools will further gain in importance and play a fundamental role to help understand and explore the feasibility of new materials, devices and systems.

Up to now, the effects of quantum mechanics have mostly been accounted for, instead of harnessed, in the design of nanoelectronic devices. A better understanding and practical grasp of quantum mechanics has opened the door to using quantum principles to engineer devices and systems. The need for simulation tools to help understand and design such quantum devices is of utmost importance. The simulation of electron transport in a semiconductor presents a fundamental simulation capability for nanoelectronics research and is approached in this work using the Wigner-Boltzmann equation.

The Wigner formalism provides a more intuitive description of quantum mechanics, compared to operator mechanics, since it is formulated in the phase space with functions and variables, which allows the adoption of models and analogies from semi-classical transport. The most important consequence of this is the ability to augment the Wigner quantum transport equation with Boltzmann scattering models, which yields the Wigner-Boltzmann equation.

The consideration of scattering in quantum transport is essential to study the decoherence of entangled electron states, which can represent qubits – the fundamental building blocks for quantum computing. The simulation of time-resolved quantum transport can help to understand highly miniaturized circuits dominated by quantum effects, which exhibit behaviour, e.g. oscillations, that cannot be explained using classical circuit theory. Currently, the only computationally viable formalism for scattering-aware, time-resolved quantum transport is found in the Wigner-Boltzmann equation.

This work presents a simulation tool which solves the equation in two dimensions using a Monte Carlo approach, based on particles with an affinity which carries the quantum information – the signed-particle method. The progress that has been made in recent years regarding, especially, computational issues, algorithms and their practical implementation is reported here.

The algorithms presented in this work represent the current state of the art for the signed-particle method and have been implemented in the Wigner Ensemble Monte Carlo tool, which forms part of the freely available ViennaWD simulation suite, to serve as a reference implementation.

The developed simulator allows the study of single electrons represented as wavepackets. First investigations of the dynamic behaviour and manipulation of such wavepackets is shown, using the concept of electrostatic lenses. An application of such a lens to improve the drive-current in a nano-scaled channel is demonstrated.

Dedicated to my parents, Ewara and Paul

Acknowledgements

The writing of this thesis and the work behind it was supported by many people I wish to acknowledge:

I thank Prof. Siegfried Selberherr who laid trust in my abilities and gave me the opportunity to conduct research at the Institute for Microelectronics. He provided me with working conditions where I can truly claim that I was the only limitation to what I have achieved. The funding he ensured was there allowed me completely focus on my research and to attend multiple conferences around the world where I could interact with leading researchers. His meticulous proofreading (also shortly before deadlines) ensured that I always strived towards writing high quality manuscripts and shall continue to do so.

My scientific supervisor Mihail "Mixi" Nedjialkov put almost as much work into this thesis as I have. His research on the topic of Wigner Monte Carlo simulations over many years laid the ground-work for me to explore a stimulating and exciting research topic where he allotted me plenty of freedom. He was extremely generous to me in many ways, especially with his time (even on Friday afternoons – his favourite time of the week to face my questions and sometimes argumentative nature). It was a true pleasure for me to have had a scientific mentor with a visible passion for his research and who always was ready to crack a joke. Thanks, Mixi – you are the best!

I thank Josef Weinbub for his support in the parallelisation of my code and general support in the development of the ViennaWD software – without his help I would still be waiting for my simulations to finish! I wish to acknowledge also the contributions Matthias Glanz made to improve the usability of my Wigner Monte Carlo simulator. I thank also Thomas Windbacher for greatly refining my German version of the abstract.

I appreciated the interest and input I received on my field of research from Prof. Hans Kosina, Zlatan Stanojević and Oskar Baumgartner. The same holds true for Viktor Sverdlov who also kept an eye out for me, also outside of the office.

External to the institute, I thank Prof. Dave Ferry (Arizona State University) and Prof. Irena Knezevic (University of Wisconsin-Madison) for their valuable feedback on my work and their enthu-siasm for research relating to the Wigner formalism. I especially thank Prof. Ivan Dimov (Bulgarian Academy of Sciences) for taking an interest in my work and, ultimately, acting as the external examiner of my thesis.

Many people contributed to making the institute a pleasant place to work: I thank Prof. Eras-mus Langer who allocated a calm office space to me in which I could be productive. I also thank Ewald Haslinger and "Frau Winkler" for tolerating the constant stream of (sometimes heavy) post I had coming to the office. Manfred Katterbauer always kept all the infrastructure running at our institute with a smile, which I appreciated. I also thank Lado Filipovic who was always there to show me the intricacies of the ESS administration system before I even had the opportunity to ask.

It was a pleasure for me to work in an environment which was positive, friendly and supportive in many ways. I thank all my colleagues of present and past for this and for making the many social gatherings at the office and at conferences lots of fun.

I shared my office with Johann Cervenka who supported me with *ad hoc* solutions to all my computer problems and always was ready for a joke or a chat. Outside the office he was at least just as supportive, always ready to accompany me for round $N + 1$. I also thank Rüdiger Sonderfeld, a friend whom I made at the office, for showing me what German efficiency is all about. I thank Prateek Sharma, a colleague and now a valued friend, with whom I faced the daily onslaught of food in the canteen. He often had to endure my rants and not so funny jokes. I also fondly remember the many hours I spent cycling around Austria with my former colleague Dmitri Osintsev – it was a real joy!

Also people outside of the institute supported me, for which I am grateful: Olga and Ievgen welcomed me to Vienna with great hospitality and have been true friends. I thank Angela for allowing me to sit on her ridiculously large terrace and ensuring I do not only talk to scientists and engineers every day. I had the pleasure of many friends visiting me over the years in Vienna – some from very far away. Their visits always provided a welcome change to my daily routine. I specifically thank Alessandro, Karen, Kasia, Konrad, Devon and my *compadre* Danilo. Also, I thank my friends Soheil and Richard for inviting me to visit them and their continued friendship. I thank my flatmate, Izabela, for not causing too much trouble overall and putting up with my often 'sunny' demeanour in the mornings.

Moving further afield, I am grateful to my family and friends in Pretoria who always welcomed me back home. Especially I thank my cousins Frans and Hein who always found time for a live or virtual "multisession". I am also grateful to my cousin Rosan and her husband Vincent – my family not quite so far away from home – for their support and always making me feel welcome in their house.

Lastly, I express my deepest gratitude to my mother whose unconditional support I have enjoyed my entire life. Since I have been away from home, she took on many responsibilities which I should have shared with her. My parents have instilled qualities in me from which I have already benefited so much in my life – I am eternally thankful to them.

Paul Ellinghaus

Vienna, February 2016

Contents

Nomenclature

Acronyms and Abbreviations

BDFT	Box discrete Fourier transform
BTE	Boltzmann transport equation
DD	Drift-diffusion
DFT	Discrete Fourier transform
DOS	Density of states
EDA	Electronic design automation
ET	Energy transport
FD	Finite difference
FET	Field-effect transistor
FFT	Fast Fourier transform
ICFE	Intra-collisional field effect
LHS	Left-hand side
MFP	Mean free path
MOSFET	Metal-oxide-semiconductor field-effect transistor
MPI	Message passing interface
NEGF	Non-equilibrium Green's function
PRL	Phase relaxation length
QTBM	Quantum transmitting boundary method
RF	Radio frequency
RHS	Right-hand side
RTD	Resonant tunnelling diode
RTD	Resonant tunnelling diode
TCAD	Technology computer-aided design
UTB-SOI	Ultra-thin base silicon-on-insulator
WBE	Wigner-Boltzmann equation
WEMC	Wigner ensemble Monte Carlo

WMC   Wigner Monte Carlo

WP   Wigner potential

Globally Defined Mathematical Symbols

$\mathbf{B}$ Magnetic field

d Dimensionality of position space

d' Dimensionality of wavevector

$\mathbf{E}$ Electric field

e Magnitude of elementary charge

E Energy

$\mathbf{F}$ Force

f_w Wigner function

$\hbar$ Reduced Planck constant

$\mathbf{k}$ Wavevector

κ Dielectric permittivity

$\mathbf{L}$ Coherence length

m^* Effective mass of an electron

$\mathbf{p}$ Momentum

$\mathbf{q}$ Integer multi-index for discretized wavevector

$\mathbf{r}$ Position

V Electrostatic potential

$\mathbf{v}_g$ Group velocity

V_w Wigner potential

List of Tables

List of Figures

Chapter 1

Introduction

This chapter motivates the need for utilising technology computer-aided design in semiconductor devices and related research fields. An overview of the most common charge carrier transport models is given, along with their range of applicability and limitations. In light of this, the motivation for developing a solver based on the Wigner(-Boltzmann) quantum transport equation is given. Thereafter, an outline of the topics to be covered in this dissertation is presented.

1.1 Need for TCAD in electronics

Technology computer-aided design (TCAD) is a field of electronic design automation (EDA) activity with software, which encompasses the physics-based simulation of semiconductor devices and the associated manufacturing processes, referred to as device TCAD and process TCAD, respectively. Process TCAD simulations predict the structure and material properties resulting from a series of manufacturing steps, like etching, ion implantation and annealing. Device TCAD receives the structure and material composition of a device (provided by process TCAD and/or experimental measurements) as an input and predicts the electrical, optical, thermal and/or mechanical behaviour of the device.

The value of TCAD is two-fold: i) it serves as a predictive tool – as evidenced by its use in the international technology roadmap for semiconductors (ITRS) [1] – to verify concepts and ascertain the effects of changes in the structure and/or process on the performance of a device before manufacturing. This allows the design space, which needs to be explored with experimentally manufactured devices, to be greatly constrained, thereby saving both time and cost. ii) TCAD also aides in the understanding of process and device physics; insight into microscopic physical quantities that cannot be measured or visualized experimentally is provided.

TCAD is increasingly centred around the manufacturing process, where a concurrent optimization of the material system, manufacturing process and device design takes place [2]. Device TCAD for ultra-scaled devices has become very complex, incorporating a multitude of models which describe the electrical, optical and thermal processes at play, operating over different scales [3]. Indeed, there are very few devices that can be modelled completely using a single tool [4]. Therefore, the concurrent consideration of many physical effects – so-called multiscale, multiphysics simulations – and the robust coupling of the various tools and models is a major thrust in (especially commercial) TCAD development. On the other hand, the advancement of models to appropriately describe the physics presents another driver for TCAD development. The charge carrier transport models are fundamental to predicting the electrical performance of devices with TCAD simulation. Quantum transport models are already indispensable to appropriately model certain modern devices and will further gain in relevance.

Only a few devices have exploited quantum mechanical principles up to now, e.g. resonant tunnelling diodes (RTD), whereas other devices merely consider quantum effects to ensure the desired behaviour is retained. With end of downscaling in devices looming, new avenues must be explored to design novel devices for the *beyond CMOS* era. Quantum considerations will be fundamental in the design of nanoscale structures that will make truly ubiquitous computing and power-efficient sensor networks a reality [5]. Prototypical nano-circuits, which are formed by simple nanostructures, exhibit

electrical behaviour that cannot be explained by classical theory. Time-resolved simulations of quantum transport will help to resolve the questions that surround the frequency response of quantum capacitors and resistances.

In the following a brief overview is given of the most common transport models in use for the simulation of existing and novel devices.

1.2 Overview of carrier transport models

Semiconductor devices and nanostructures are open systems that interact with their environment, e.g. through leads/contacts, phonons or electromagnetic fields. The effect of these interactions, especially the electric field created by the voltage applied between the contacts[1], is so strong that the statistical distribution of particles (charge carriers) strongly deviates from the statistics applicable in equilibrium, e.g. the Fermi-Dirac distribution. The results of linear response theory, e.g. the relaxation time approximation, are no longer sufficient for an accurate description [6, 7]. The non-equilibrium distribution can be determined by solving a transport equation which describes the movement of particles under the influence of external forces.

There exists a plethora of transport models, each having its limitations and advantages. The choice of a model depends on the application (phenomena) to be simulated and the computational resources/time available. From a physical point of view it is attractive to choose the most fundamental models, but these may not computationally feasible, nor required, for a given task. The aim of this section is to provide an overview of the most important transport models used for semiconductor devices and nanostructures. The capabilities and limitations of the models are highlighted to be able to appreciate the position and capabilities of transport simulations based on the Wigner formalism.

In the following the most important semi-classical transport models will be covered first, because they still find application today in nanoscale devices when augmented with quantum models/corrections. Thereafter, the full-fledged quantum transport models are introduced. The considerations are presented based on intuitive concepts, like particles and distribution functions, which characterize the historical evolution of the field of transport models. Many of these classical attributes are retained in the Wigner formalism which bridges the two limiting transport regimes (semi-classical/diffusive and quantum/coherent) presented hereafter.

In the following a single-particle description is deemed sufficient to describe a particle ensemble in a statistical manner [8]. Moreover, only interactions between different kinds of particles are considered, e.g. electron with phonons. Many-particle phenomena, like Coulomb interaction, that occur between particles of the same kind are not considered explicitly and are only accounted for via the Poisson equation.

1.2.1 Classification of transport models

The physics that governs carrier transport depends on three characteristic lengths as related to the dimensions of a device [9]:

1. De Broglie wavelength: the wavelength associated to a particle with a momentum $\mathbf{p}$ by $\lambda = |\mathbf{p}|/h$. For an electron in silicon at room temperature $\lambda \approx 8\,\mathrm{nm}$ [8], whereas in GaAs $\lambda \approx 17\,\mathrm{nm}$ [10].

2. Mean free path (MFP): the mean distance a carrier travels before a momentum-altering scattering event occurs, e.g. inelastic phonon scattering. The MFP depends on the electric field but is in the order of $5 - 10\,\mathrm{nm}$ for an electron in silicon at room temperature and up to an order of magnitude longer for GaAs [11, 12].

3. Phase relaxation length (PRL): the mean distance a carrier travels before a phase-changing scattering event occurs, e.g. electron-electron scattering. In semiconductors the relation between

---

[1]To retain electrostatic control of the channel in transistors, the applied voltage should be larger than the thermal voltage and, therefore, cannot scale down at the same rate as the dimensions of the device. A continuous increase of the electric fields in transistors has been the result.
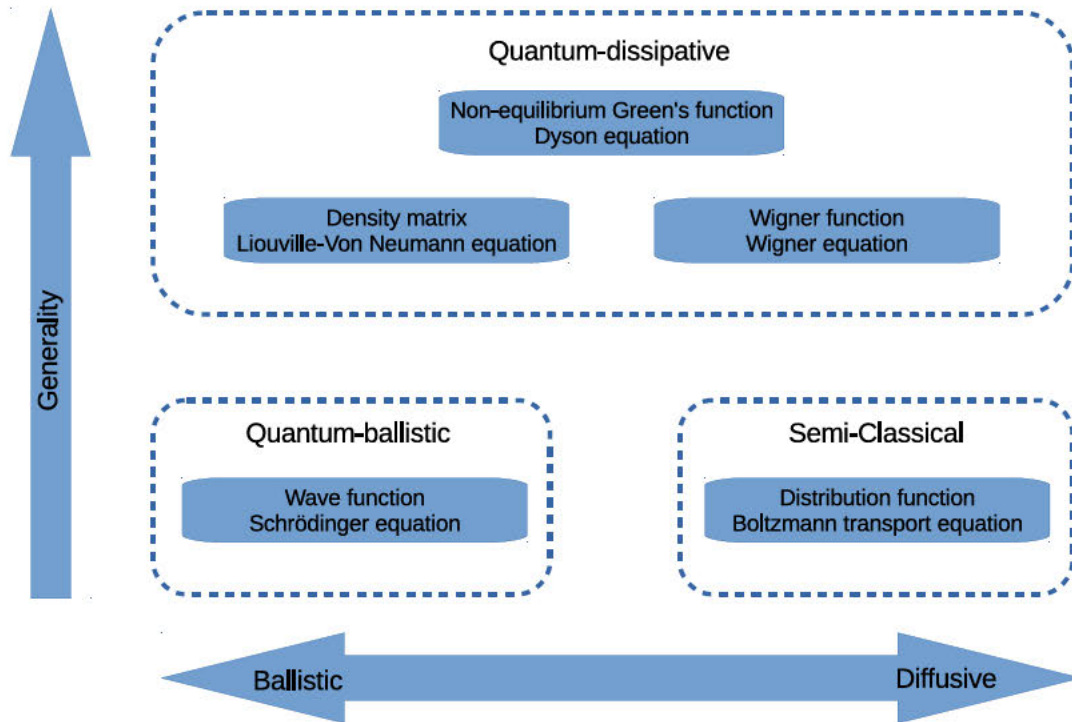
Figure 1.1: Overview of microscopic transport models. A distinction is made between models that can only describe only ballistic or diffusive transport and more general models which can describe transport in both regimes.

the PRL and the MFP depends on the type of semiconductor [13], but the PRL usually matches or exceeds the MFP [9].

Considering the above, one can distinguish between two limiting cases: the diffusive regime where the size of the device is much larger than all the characteristic lengths and carrier transport is predominantly determined by the scattering taking place; the other extreme, in devices shorter than all the characteristic lengths, is the ballistic regime where negligible scattering occurs and the transport is completely coherent[2]. Whereas transport in the diffusive regime can be appropriately described by semi-classical transport models, the ballistic regime requires models which consider the coherent wave nature of particles. The transition between the diffusive and the ballistic regime is gradual and nanostructures/devices, where both scattering and coherent transport play a role, demand the most general transport models, as is illustrated in Figure 1.1 (based on [14]).

An interplay between diffusive and coherent transport is often encountered in mesoscopic devices with at least one dimension shorter than one of the characteristic lengths mentioned above. In such devices quantum processes are at play that can be observed macroscopically, e.g. conductance fluctuations [15].

Quantization effects, due to confinement, start to appear, if at least one physical dimension of a device is below the De Broglie wavelength of the carriers. Interference effects imply coherent transport and become significant if the PRL exceeds at least one dimension of a device. Since the PRL can exceed the MFP in semiconductors, the phase coherence can be retained despite scattering taking place.

Apart from the physically-grounded distinction between transport models between the ballistic and diffusive regimes, the models can further be classified as either microscopic or macroscopic. Mi-

---

[2]Transport is termed coherent, if there exists a deterministic relation between the phase of a wave function between two points. Scattering events can be phase-preserving (elastic scattering, e.g. ionic) and phase-breaking (inelastic scattering, e.g. phonons).

croscopic transport models describe the evolution of the distribution function (semi-classical) or state (quantum) of the system, from which all physical quantities (observables) of interest can be calculated. Macroscopic models are derived from the microscopic transport equations and directly solve for physical quantities – often termed *engineering quantities* – like electron concentration or current density. The computational efficiency afforded by macroscopic models comes at the cost of a loss in generality – certain assumptions are made, which limit the scope where the models remain valid. A thorough treatment of the mathematical derivations of both semi-classical and quantum macroscopic models is given in [8].

A hierarchy of models can be set up ranging from, at the most fundamental level, a quantum microscopic model to a semi-classical macroscopic description of carrier transport. Table 1.1 presents such a hierarchy, along with the device dimensions and applications at which each model is applicable.

Table 1.1: Hierarchy of carrier transport models incorporating dissipative interactions with typical active region of devices and the physical quantity which governs the device physics.

Model	Active region	Governed by
Non-Equilibrium Green's Function	$< 10\,\mathrm{nm}$	Potential
Density matrix / Wigner	$< 100\,\mathrm{nm}$	Potential
Boltzmann	$< 1\,\mu\mathrm{m}$	Sharp field
Hydrodynamic / Energy transport	$100\,\mathrm{nm} \div 1\,\mu\mathrm{m}$	Sharp field
Drift-diffusion	$> 1\,\mu\mathrm{m}$	Smooth field

1.2.2 Poisson equation

The Poisson equation describes the electrostatics in a device based on the distribution of charges. A self-consistent solution[3] of the Poisson equation with the transport models takes the long-range Coulomb interaction – a many-identical-particle effect – into account.

The Poisson equation follows from Gauss's law, which relates the divergence of the electric displacement vector to charge density. The electric field displacement vector can be related to the electrostatic potential under the assumption of a time-independent permittivity, a negligible magnetic field and no polarization effects [16]. These assumptions are reasonable for many semiconductor devices for which the Poisson equation is defined as

$$\nabla \cdot (-\kappa \nabla V) = \rho_c, \tag{1.1}$$

where κ, V and ρ_c denote the dielectric permittivity, electrostatic potential and charge density, respectively. The charge density is typically determined by free charge carriers and ionized impurities in the semiconductor.

In semiconductor materials with very strong polarization effects, e.g. galliumnitride, or devices where piezoelectric or ferromagnetic phenomena come into play, the Poisson equation, in the form of (1.1) is no longer sufficient [16].

1.2.3 Semi-classical models

The Boltzmann transport equation (BTE) has been the cornerstone of semi-classical device TCAD for many years and appropriately describes carrier transport in the diffusive regime. The BTE is introduced here and will be referred to again in Chapter 2, where accompanying scattering models are used to augment the Wigner transport equation.

The BTE is classified as semi-classical, because the treatment of carrier transport is classical (Newtonian), but some quantum concepts are considered, like the band structure (in the group velocity) and

---

[3]The Poisson equation and the transport model are coupled through the dependence of the potential on the charge distribution and vice-versa. A self-consistent solution solves both equations such that both are satisfied simultaneously with the same potential and charge distribution.

the Pauli exclusion principle (in the collision operator). The BTE provides microscopic insight into device operation but its high dimensionality makes its solution computationally expensive. Various macroscopic models, e.g. drift-diffusion, can be derived from the BTE and directly solve macroscopic quantities, thereby greatly reducing the numerical complexity of the simulation.

1.2.3.1 Boltzmann transport equation

The BTE describes the evolution of the distribution function $f(\mathbf{r}, \mathbf{k}, t)$, which gives the number of particles per unit volume of the phase space at time t. The phase space encompasses all possible values of the position $\mathbf{r}$ and the wavevector $\mathbf{k}$ (related to the crystal momentum by $\mathbf{p} = \hbar \mathbf{k}$) that a particle can attain.

The movement of a particle in the phase space is described by its Newton trajectories, parametrized in time and initialized at the point $(\mathbf{r}_0, \mathbf{k}_0)$ at t_0:

$$\mathbf{r}(t) = \mathbf{r}_0 + \int_{t_0}^{t} \frac{\hbar \mathbf{k}}{m^*} dt; \tag{1.2}$$

$$\mathbf{k}(t) = \mathbf{k}_0 + \int_{t_0}^{t} \mathbf{F} dt. \tag{1.3}$$

The (differentiable and conservative) force is defined as

$$\mathbf{F} = \hbar \frac{d\mathbf{k}(t)}{dt} = -e(\mathbf{E} + \mathbf{v}_g \times \mathbf{B}), \tag{1.4}$$

where e, $\mathbf{E}$ and $\mathbf{B}$ denote the magnitude of an elementary charge, electric field and magnetic field, respectively. The group velocity is given by

$$\mathbf{v}_g = \frac{d\mathbf{r}(t)}{dt} = \frac{1}{\hbar} \nabla_k \epsilon(\mathbf{k}), \tag{1.5}$$

where $\epsilon(\mathbf{k})$ denotes the dispersion relation.

The Liouville theorem states that along the Newton trajectories the distribution function remains constant, i.e.

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{d\mathbf{r}(t)}{dt} \nabla_r f + \frac{d\mathbf{k}(t)}{dt} \nabla_k f = 0. \tag{1.6}$$

This is known as the Vlasov equation [8] and describes the evolution of the distribution function. The spatial gradient term $\nabla_r f$ accounts for driving forces due to diffusion, thermoelectric and thermomagnetic effects. All external forces are encompassed in the term $\frac{d\mathbf{k}(t)}{dt}$.

The Vlasov equation gives an exact description of a semi-classical carrier moving in a perfect lattice; all lattice-carrier interactions are accounted for through the group velocity $\mathbf{v}_g$, which depends on the band structure of the semiconductor lattice through the dispersion relation $\epsilon(\mathbf{k})$. Lattice imperfections, lattice vibrations (phonons) and carrier-carrier interactions, however, introduce internal forces in the lattice which perturb the motion of the carriers. These perturbations in the trajectory of a carrier are known as collisions and are described by statistical laws (in the form of a collision operator), because an explicit consideration by the laws of dynamics is not feasible [16].

The collision operator augments the right-hand side of (1.6) and describes the effect of collisions on the distribution function such that

$$\frac{df}{dt} = \hat{C}[f(\mathbf{r}, \mathbf{k}, t)]. \tag{1.7}$$

The collision operator $\hat{C}[\cdot]$ is given by

$$\hat{C}[f(\mathbf{r}, \mathbf{k}, t)] = \int \left\{ f(\mathbf{k}') S(\mathbf{k}', \mathbf{k}) - f(\mathbf{k}) S(\mathbf{k}, \mathbf{k}') \right\} d\mathbf{k}' \tag{1.8}$$

and describes the scattering rate of carriers into (from) a state with wavevector $\mathbf{k}$ ($\mathbf{k}'$) from (into) a state with a wavevector $\mathbf{k}'$ ($\mathbf{k}$). The Pauli exclusion principle (not included here) prescribes that

the in- and out-scattering only takes place if the final states are available. However, this is a many-identical-particle effect [7] and consequently is omitted in (1.8). The neglect of the exclusion principle is reasonable under non-degenerate conditions (low carrier density).

The scattering rate between the states of different wavevectors is specified by the function S, which compounds the expressions describing the various scattering mechanisms, e.g. carrier-phonon interactions, which change the momentum of a particle and make the system dissipative. The scattering rate for a specific mechanism is calculated using overlap integrals [7]. A more detailed treatment of the scattering models is given in Chapter 2.

The consideration of collisions, through the collision operator (1.8), in the Vlasov equation (1.6) yields what is known as the Boltzmann transport equation:

$$\begin{cases} \frac{\partial f}{\partial t} + \mathbf{v}_g \nabla_r f + \mathbf{F} \nabla_k f = \int \{ f(\mathbf{k}',t) S(\mathbf{k}',\mathbf{k}) - f(\mathbf{k},t) S(\mathbf{k},\mathbf{k}') \} \, d\mathbf{k}' \\ f(\mathbf{r},\mathbf{k},t_0) = f(\mathbf{r}_0,\mathbf{k}_0) \end{cases} \tag{1.9}$$

An initial condition at t_0 must be specified to have a well-posed problem.

The mean value $\langle A_T \rangle$ of any physical quantity $A(\mathbf{r},\mathbf{k})$, like carrier concentration or energy, at time $t = T$ can be calculated, if the distribution function is known:

$$\langle A_T \rangle = \int \int d\mathbf{r} d\mathbf{k} A(\mathbf{r},\mathbf{k}) f(\mathbf{r},\mathbf{k},T). \tag{1.10}$$

For instance, the carrier concentration for a volume Ω can be calculated by

$$n(\mathbf{r}) = \frac{1}{\Omega} \int d\mathbf{k} f(\mathbf{r},\mathbf{k}) \tag{1.11}$$

and the current density as

$$\mathbf{J}(\mathbf{r}) = -\frac{e}{\Omega} \int d\mathbf{k} f(\mathbf{r},\mathbf{k}) \mathbf{v}_g(\mathbf{k}). \tag{1.12}$$

The validity of the BTE rests upon the assumptions [17] that

- point-like particles are exactly localized in the phase space and their transport is accurately described by the semi-classical Newton laws;

- particle scattering takes place instantaneously at a fixed position, i.e. particles do not move during a scattering process.

From a quantum mechanical point of view the BTE fails to account for the quantum effects that appear due to the wave nature of charge carrier motion, like non-locality, interference and correlation effects. The fact that the momentum and position are defined exactly at the same time with $f(\mathbf{r},\mathbf{k},t)$ makes the classical nature of the BTE obvious.

The Heisenberg uncertainty principle imposes limits on the range of validity of the BTE but they are not (yet) reached for most devices. The uncertainty relation $\Delta r \Delta p \geq \hbar/2$ requires that the potential within a device should change only negligibly over the De Broglie wavelength of the carriers. The latter, assuming an electron in silicon at room temperature, is in the order of $8\,\text{nm}$ [8]. Once the dimensions of the active regions of a device approach the De Broglie wavelength of the charge carriers, the BTE should be substituted by a wave equation to describe carrier transport [17].

A further constraint is given by the relation, $\Delta E \Delta t \geq \hbar$, which requires that the momentum relaxation time should be much smaller compared to the period of the operating frequency – this limits the validity of BTE to operation frequencies under approximately $6\,\text{THz}$ [17].

1.2.3.2 Macroscopic models

Macroscopic transport models can be formally derived from the BTE and have been extensively treated in the literature [17, 18, 19, 20]. The basic principle is to calculate moments of the BTE, which yields equations describing the conservation and flux of physical quantities, like mass, momentum or energy. A hierarchy of macroscopic models can be derived by considering moments of increasing order (in $\mathbf{k}$), ranging from the drift-diffusion model to the six-moment model. In each case a so-called closure relation encapsulates the assumptions made on higher-order moments of the distribution needed to make the system of equations fully determined.

Drift-diffusion: The drift-diffusion (DD) model was originally proposed based on phenomenological arguments [21]. The derivation considers the first two moments of the BTE and assumes a cold Maxwellian for the distribution function as a closure relation. The resulting equations describe the conservation of mass – the carrier concentration n – and the flux of mass – the current density $\mathbf{J}_n$:

$$\partial_t n - \frac{1}{e} \nabla \mathbf{J}_n = G_n; \tag{1.13}$$

$$\mathbf{J}_n = e \left(D_n \nabla n - \mu_n n \nabla V \right), \tag{1.14}$$

where D_n and μ_n denote the material-dependent diffusion coefficient and carrier mobility, respectively. The net generation-recombination rate G_n relates to the scattering operator in the BTE. The solution of (1.13) and (1.14) directly yields the quantities calculated with the distribution function in (1.11) and (1.12).

The fast, robust numerical solvers available for the DD model in commercial and open source TCAD software makes it the (highly parametrized) workhorse of industry for daily engineering tasks. The self-consistent solution obtained with the Poisson equation is also routinely used as an initial guess to improve the convergence of more complex transport models. The DD model is suitable to simulate devices for which the mobility can be satisfactorily modelled as a function of the electric field and thermal effects (carrier heating) are unimportant. These conditions are often met in low-power devices with an active region exceeding $1\,\mu$m.

Energy transport (hydrodynamic): The energy transport (ET) model[4] was originally proposed in [22] to account for hot carrier effects, like impact ionization and velocity overshoot, occurring in devices smaller than $1\,\mu$m. The ET model considers two additional moments of the BTE, compared to the DD model, and was formally derived in [23]. A multitude of derivations and closure relations for the ET model exist [8, 22, 24]. Compared to (1.14), the gradient of the carrier temperature enters the mass flux equation as an additional driving force of the current. The additional moments introduce equations which enforce the conservation of energy and define the energy flux.

Six moments: The six-moment model considers a further two moments of the BTE [18, 19]. The conservation of a third quantity, related to the kurtosis (skewness) of the distribution function, is enforced, which allows to better model hot carrier phenomena. However, the strong coupling between the equations and the choice of appropriate closure relations has made the realization of a robust numerical solver very challenging [25]. As a result the six-moment model has not found wide-spread use and remains largely of academic interest.

1.2.4 Quantum models

The passage to quantum mechanics is opened by abandoning the idea that the state of a particle is represented by a single point in the phase space (as done in the BTE). The Heisenberg uncertainty principle does not allow the exact specification of two conjugate quantities, like position and momentum, simultaneously. The minimum uncertainty is given by

$$\Delta r \Delta p \geq \frac{\hbar}{2}, \tag{1.15}$$

where Δr and Δp refer to the uncertainty (or standard deviation) in position and momentum in a specific direction and $\hbar$ is the reduced Planck constant.

There exists a plethora of quantum transport models. Certain formalisms may be better suited, or simpler, than others for a specific application and every formalism has its (dis)advantages from a computational point of view. The focus here shall be placed on the formulations which enjoy the most widespread use, namely the Schrödinger equation, density matrix and non-equilibrium Green's

---

[4]The names energy transport and hydrodynamic are usually used synonymously in device TCAD. Strictly speaking, the ET model presents a simplification of the full hydrodynamic transport equation – a name inspired by the similarity to Euler's equations for incompressible fluid flow made to avoid difficulties in its numerical solution – and is the prevalent form.

function (NEGF). The essential aspects of each model are highlighted in the following to enable a comparison with the Wigner formulation which will be detailed in Chapter 2.

1.2.4.1 Preliminaries

Quantum mechanics relies on two important concepts to model physical systems: i) the state of the system is described by a vector in a complex Hilbert space and ii) observables which describe physical quantities, like energy, are represented by Hermitian operators which act on the state.

Vectors and Operators: The state vector can be expressed using an orthonormal basis formed by the eigenstates of the Hermitian operator (observable) under consideration:

$$|\psi\rangle = \sum_j c_j |\phi_j\rangle , \; c_j = \langle\phi_j| \psi\rangle , \tag{1.16}$$

where c_j is the projection of the state vector $|\psi\rangle$ on the basis vector $|\phi_j\rangle$. The basis vectors are obtained by solving the eigenvalue problem for the observable represented by the operator $\hat{A}$:

$$\hat{A} |\phi_j\rangle = a_j |\phi_j\rangle . \tag{1.17}$$

One can distinguish between pure states and mixed states. A pure state implies complete knowledge of the state a system is in; the classical equivalent would be to have a complete knowledge of the position and momentum of every particle in a system (i.e. no single-particle description). A mixed state, however, introduces some uncertainty and assigns a probability p_k to each possible pure state $|\psi^k\rangle$, where $\sum_k p_k = 1$. This is the quantum equivalent of a distribution function.

A mixed state is to be distinguished from a mixture (superposition/linear combination) of states. While the former conveys the probability of a system being in a certain state, the latter describes a single state composed of two (or more) other states – an entangled state. In general one has incomplete information about the quantum system under observation and the initial state from which the evolution starts. However, the probability of a state occurring is often known from statistical distributions, e.g the statistical mixture of eigenstates of energy in thermal equilibrium. A mixed state allows to model this uncertainty. It should be noted that the uncertainty about the state of a system acts in addition to the uncertainty introduced by Heisenberg's principle.

Picture and Representations: The so-called picture adopted in quantum mechanics refers to the state vectors and observables (linear operators) chosen to represent the physical state and dynamical variables of a system. The Schrödinger picture makes the state time-dependent with the evolution of the system whereas the observables remain unchanged; the Heisenberg picture is the converse[5]. In the interaction picture the time evolution is split between the observable and the state [6] – this is useful if the Hamiltonian (observable) has a time-dependent part, e.g. phonon interactions.

The choice of the basis vectors used to represent the vectors and operators corresponding to the chosen picture is known as the representation. The value of a physical observable is independent of the representation chosen. The coordinate representation is usually the most intuitive due to the complex shape of the potential profile $V(\mathbf{r})$. The state vector projected in the chosen representation is known as a wave function.

Hamiltonian: A quantum system is described by a Hamiltonian which states the energy of a system, i.e. an observable. The Hamiltonian can encapsulate all interactions in the system under observation and interactions with external systems. The complexity of the Hamiltonian can be extended as needed to model the physics of interest or as computationally feasible. If adopting the interaction picture, the Hamiltonian can be separated in a part describing time-independent effects, e.g. the band structure, and a part describing the time-dependent perturbations to the Hamiltonian from the 'outside world', e.g. interactions with time-varying electromagnetic fields:

---

[5]A 'picture' can be freely chosen and one can move between different pictures by applying a unitary transform

$$\hat{H} = \hat{H}_0 + \hat{H}_{\text{int}}, \tag{1.18}$$

where $\hat{H}_0$ is the non-interacting part, which is assumed to be exactly solvable and $\hat{H}_{\text{int}}$ contains all the interactions, such as carrier-carrier, carrier-phonon, impurity scattering and so forth.

In a coordinate representation under the assumption of a parabolic dispersion relation with effective mass m^*, the Hamiltonian is expressed as

$$\hat{H}_0 = -\frac{\hbar^2}{2m^*}\nabla_r^2 + V(\mathbf{r}) \tag{1.19}$$

and will be used throughout this text, unless specified otherwise. The potential $V(\mathbf{r})$ encapsulates the electrostatic potential (from mobile electrons, ionized impurities and externally applied voltages) and changes in the conduction band (e.g. at heterojunctions) and can be obtained from (1.1) as in the semi-classical case.

1.2.4.2 Schrödinger equation

The Schrödinger equation is the fundamental equation of motion describing the evolution of a pure quantum state $|\psi\rangle$ in a system described by the Hamiltonian operator $\hat{H}$:

$$i\hbar\frac{\partial}{\partial t}|\psi\rangle = \hat{H}|\psi\rangle. \tag{1.20}$$

The presence of an imaginary term alludes to the wave nature of the equation. The time-independent version of (1.20),

$$\hat{H}|\psi\rangle = \lambda|\psi\rangle, \tag{1.21}$$

presents a specific case of (1.17), which yields the stationary states $|\psi_j\rangle$ and associated eigenvalues λ_j (energies) of the Hamiltonian. A pure state does not necessarily correspond to an eigenvector of the observable. The state vector can expressed in terms of the resulting basis

$$|\psi\rangle = \sum_j c_j|\psi_j\rangle, \; c_j \in \mathcal{C}. \tag{1.22}$$

Since a mixed state cannot be represented by a single vector, (1.20) must be solved for each possible state separately. Equation (1.20) can be expressed in a coordinate representation as

$$i\hbar\frac{\partial}{\partial t}\psi(\mathbf{r}, t) = H\psi(\mathbf{r}, t). \tag{1.23}$$

If the state is not an eigenstate of the observable, there is an uncertainty in the value which will be measured for the observable. The average of a physical quantity represented by an operator $\hat{A}$ at time τ can be calculated by

$$\langle A_\tau\rangle = \langle\psi|\hat{A}|\psi\rangle = \int d\mathbf{r}\,\langle\psi_t|\mathbf{r}\rangle\,\langle\mathbf{r}|\hat{A}|\psi_t\rangle. \tag{1.24}$$

Macroscopic quantities, like the probability density and current density, can be obtained from the wave functions forming the mixed state:

$$n(\mathbf{r}, t) = \sum_j \lambda_j|\psi_j(\mathbf{r}, t)|^2; \tag{1.25}$$

$$\mathbf{J}(\mathbf{r}, t) = -\frac{e\hbar}{m^*}\sum_j \lambda_j\text{Im}\left[\bar{\psi}_j(\mathbf{r}, t)\nabla\psi_j(\mathbf{r}, t)\right]. \tag{1.26}$$

The Schrödinger equation as presented above is applicable to a closed (bounded) system. To be useful for semiconductor devices the system must be 'opened up' to allow the specification of boundary conditions, which gives rise to current-carrying (scattering) states. The quantum transmitting

boundary method (QTBM) [26] allows the specification of the incoming flux from a lead to a device by a boundary condition for the wave function at the device/lead interface. The QTBM can provide the solution (wave function) inside the device and the transmission and reflection coefficients at the leads. Under the assumption of a constant potential in the leads, the form of solution in the leads is known to consist of incoming and outgoing waves and exponentially decaying evanescent waves along the interface. The QTBM employs this fact to specify boundary conditions at the lead/device interface expressed in terms of this known solution.

The Schrödinger equation is only well-suited to describe ballistic transport. Models for out-scattering, inspired by the relaxation time approximation, have been theoretically proposed (by adding an imaginary potential to the Hamiltonian), but numerical implementations have proven problematic [25].

1.2.4.3 Density matrix

The density matrix provides a convenient description of mixed states by specifying the relation between the states which comprise the mixed state.

The density matrix can be derived from the Schrödinger picture following the derivation in [20]. By adding unitary operators, expressed in the coordinate basis and the observable's basis, (1.24) can be rewritten as

$$\langle A_\tau \rangle = \int d\mathbf{r}' \int d\mathbf{r} \alpha \left(\mathbf{r}, \mathbf{r}' \right) \psi_t^* \left(\mathbf{r}' \right) \psi_t \left(\mathbf{r} \right). \tag{1.27}$$

This introduces the density matrix

$$\rho \left(\mathbf{r}, \mathbf{r}', t \right) = \psi_t^* \left(\mathbf{r}' \right) \psi_t \left(\mathbf{r} \right) = \langle \mathbf{r} | \psi_t \rangle \langle \psi_t | \mathbf{r}' \rangle \tag{1.28}$$

with the corresponding density operator

$$\hat{\rho}_t \equiv |\psi_t\rangle \langle \psi_t|, \tag{1.29}$$

which can be shown to be unitary using (1.22).

It follows that (1.27) can be expressed as

$$\begin{aligned} \langle A_\tau \rangle &= \int d\mathbf{r}' \int d\mathbf{r} \langle \mathbf{r} | \hat{\rho}_t | \mathbf{r}' \rangle \langle \mathbf{r}' | \hat{A} | \mathbf{r} \rangle \\ &= \int d\mathbf{r} \langle \mathbf{r} | \hat{\rho}_t \hat{A} | \mathbf{r} \rangle \\ &= \mathrm{Tr} \left(\hat{\rho}_t \hat{A} \right). \end{aligned} \tag{1.30}$$

Therefore, the expected value of a physical observable A can be calculated by applying the trace operator, which sums the diagonal terms of the resulting matrix.

Accordingly, the probability density can be calculated by

$$n \left(\mathbf{r}, t \right) = 2\mathrm{Tr} \left(\rho_t \right) = \rho \left(\mathbf{r}, \mathbf{r}, t \right) \tag{1.31}$$

and the current density by

$$\mathbf{J} \left(\mathbf{r}, t \right) = \frac{i\hbar e}{m^*} \lim_{\mathbf{r}' \to \mathbf{r}} \left(\nabla_{r'} - \nabla_r \right) \rho \left(\mathbf{r}, \mathbf{r}', t \right). \tag{1.32}$$

The Liouville-Von Neumann equation describes the evolution of the density matrix

$$i\hbar \frac{\partial \rho}{\partial t} = [H, \rho], \tag{1.33}$$

where $[\cdot, \cdot]$ denotes the commutator bracket and H the Hamiltonian defined in (1.19), such that

$$\begin{aligned} i\hbar \frac{\partial \rho}{\partial t} &= H\rho - \rho H \\ &= -\frac{\hbar^2}{2m^*} \left(\nabla_r^2 - \nabla_{r'}^2 \right) \rho + \left(V \left(\mathbf{r} \right) - V \left(\mathbf{r}' \right) \right) \rho. \end{aligned} \tag{1.34}$$

The density operator has a complete orthonormal set of eigenfunctions and eigenvalues, where each eigenpair can be associated with the solution of a Schrödinger equation. Therefore, a set of Schrödinger equations can also be used to represent a mixed state.

The introduction of a scattering operator in (1.34) which maintains the positive-definite character (physical validity) of the density matrix under all circumstances has proven challenging. Recently, the derivation of a more robust operator has been proposed [27], which has analogies with the scattering models for the Wigner equation (Chapter 2). However, due to the non-local nature of the density matrix, the models are more difficult to interpret.

1.2.4.4   Wigner function

For completeness the Wigner function is briefly introduced here. A thorough discussion is deferred to Chapter 2.

The Wigner function is obtained by applying the Wigner transform to the density matrix:

$$f_w\left(\mathbf{r}, \mathbf{k}, t\right) = \frac{1}{(2\pi)^3} \int d\mathbf{r}' e^{-i\mathbf{k}\cdot\mathbf{r}'} \psi\left(\mathbf{r} - \frac{\mathbf{r}'}{2}\right) \psi^*\left(\mathbf{r} + \frac{\mathbf{r}'}{2}\right). \tag{1.35}$$

Similarly, a Fourier transform of the Von Neumann equation for the density matrix (1.34) yields the evolution equation for the associated Wigner function:

$$\frac{\partial}{\partial t} f_w\left(\mathbf{r}, \mathbf{k}, t\right) + \frac{\hbar \mathbf{k}}{2m^*} \frac{\partial}{\partial \mathbf{r}} f_w\left(\mathbf{r}, \mathbf{k}, t\right) = \int d\mathbf{k}' V_w\left(\mathbf{r}, \mathbf{k}' - \mathbf{k}, t\right) f_w\left(\mathbf{r}, \mathbf{k}', t\right), \tag{1.36}$$

where V_w denotes the Wigner potential.

The Wigner function uses the phase-space variables, instead of the two spatial variables in the density matrix, which allows the scattering models from the Boltzmann transport equation to be adopted (to be discussed in Chapter 2).

1.2.4.5   Non-equilibrium Green's function

The non-equilibrium Green's function provides a general framework to describe weakly interacting quantum systems. It introduces a correlation in time in addition to the space correlation considered in the density matrix / Wigner function. Consider (1.28) for the case for which $t \neq t'$: $\rho\left(\mathbf{r}, \mathbf{r}', t, t'\right) = \psi\left(\mathbf{r}, t\right) \psi^*\left(\mathbf{r}', t'\right)$. The Green's function (assuming a coordinate representation) describes how the amplitude of a wave function (an excitation) at point $\mathbf{r}'$ at time t' is propagated to a point $\mathbf{r}$ at time t.

The time evolution of a wave function – the solution to the Schrödinger equation in (1.23) – can be calculated by

$$\psi\left(\mathbf{r}, t\right) = i\hbar \int d\mathbf{r} G\left(\mathbf{r}, t, \mathbf{r}', t'\right) \psi\left(\mathbf{r}', t'\right). \tag{1.37}$$

The Green's function – also known as the propagator – can be viewed as a generalization of the scattering matrix, commonly used in radio frequency (RF) electronics, which relates the input and output at different ports, because it relates the response at an arbitrary point in the the system (device) to the excitation(s) at any other point(s).

The evolution of the time-resolved Green's function is described by the Dyson equation [28], but its high dimensionality of $2d + 2$ makes it numerically extremely challenging to solve in two and three spatial dimensions (d). Only very recently have such efforts even started [29, 30]. Therefore, the steady-state, where only the difference between the time variables appears, is considered in the following (as is routinely done in NEGF simulations).

A Fourier transform of the time difference $t - t'$ yields the energy E, which appears as parameter in the Green's function $G\left(\mathbf{r}, \mathbf{r}'; E\right)$[6].

---

[6]The steady-state semi-classical distribution function only has $2d$ dimensions, whereas the Green's correlation function has $2d + 1$. In the semiclassical case the wavevector can be clearly related to an energy through the dispersion relation. Such a relation does not exist in the quantum mechanical case, because the potential does not vary smoothly. Therefore, the energy appears explicitly as an additional variable.

The Green's function is calculated by solving the differential equation

$$[E - H(\mathbf{r})]G(\mathbf{r}, \mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}'), \tag{1.38}$$

which permits two solutions, since the inverse of a differential operator, i.e. the Hamiltonian, is not uniquely specified without boundary conditions [9]. The two solutions correspond to a propagation forward and backward in time and are known as the advanced and retarded Green's function, respectively. A unique solution arises once the boundary conditions are introduced in the form of a self-energy representing the effect of semi-infinite leads coupled to the device. In the following, the retarded Green's function is implied without introducing additional notation.

The self-energy encompasses all possible interactions between the electron and its environment. Each interaction has its own self-energy, which is calculated using many-body perturbation theory under the presumption that systems are weakly interacting [28]. The self-energy accounting for the coupling of the device to semi-infinite leads reduces the system to a finite size and follows from similar arguments as used in the QTBM, i.e. plane waves are injected from a lead assumed to be at a constant potential. The calculation of the self-energies for other interactions, like phonon scattering, can become very computationally demanding and various approximations are routinely made to keep simulations tractable..

The coordinates in equation in (1.38) can be discretized and a representation in terms of matrices follows:

$$G = [EI - H - \Sigma]^{-1}, \tag{1.39}$$

where I is the identity matrix, E denotes the energy, H is the Hamiltonian matrix and Σ is the self-energy matrix. The numerical task at hand is to invert the matrix in (1.39) for every energy value in the chosen range[7]. Therefore, the size of the matrix should be kept as small as possible, which limits the size of devices which can be practically investigated.

The correlation function[8] $G^n(\mathbf{r}, \mathbf{r}', t, t')$ is used to calculate physical observables and is related to the Green's function G by

$$G^n = G\Sigma^{\text{in}}G^\dagger, \tag{1.40}$$

where $G^\dagger$ is the Hermitian conjugate of G and Σ^{in} is the self-energy which accounts for in-scattering of electrons from leads and other interactions, like phonons. The function G^n can be regarded as the quantum mechanical equivalent of the distribution function f in the semi-classical BTE.

Following [9], the electron and current density at a certain energy can be calculated by

$$n(\mathbf{r}, E) = \frac{G^n(\mathbf{r}, E)}{2\pi}; \tag{1.41}$$

$$\mathbf{J}(\mathbf{r}, E) = \frac{ie\hbar}{2m^*}\frac{1}{2\pi}\lim_{\mathbf{r}'\to\mathbf{r}}(\nabla_r - \nabla_{r'})G^n(\mathbf{r}, E). \tag{1.42}$$

The total densities are obtained by integrating the expression (1.41) and (1.42) over all considered energy values and accounting for spin-degeneracy (multiplication by two).

The NEGF formalism is the most general of all the transport models reviewed here and is arguably the most popular formalism currently in use in the computational electronics community. However, the theoretical generality is accompanied by great computational demands and simplifications – as alluded to above – are routinely employed to make simulations numerically feasible. Therefore, the theoretical capabilities of a formalism should not be considered in isolation, but the extent to which these can be realized in simulations is equally important.

1.2.4.6 Quantum-corrected models

Quantum-corrected models refer to semi-classical models that are modified to capture important quantum effects. This allows the (relatively) computationally efficient semi-classical transport models to be used to approach the results obtained with a full-fledged quantum simulation. The modifications,

---

[7]The energy range is chosen to be a few multiples of the thermal energy $k_B T$ around the Fermi energy level
[8]this is equivalent to the quantity $iG^<(\mathbf{r}, \mathbf{r}', t, t')$

however, introduce some fitting parameters which must be appropriately calibrated to a specific device to realise the desired improvements in accuracy.

The most prominent quantum effect in transistors is the quantization of energy levels. Spatial confinement leads to the separation of sub bands and is accompanied by an increase in interurban scattering, which can be accounted for by using degenerate statistics in Monte Carlo simulations of the BTE [31]. To calculate the true occupation of the discrete sub bands requires a self-consistent solution of the Schrödinger-Poisson system, but the charge distribution can be replicated in a semi-classical model using various approaches: The density of states (DOS) can be modified using various means, e.g. introducing a spatial dependence, or artificially modifying the energy band gap [31, 32].

A further option to account for the confinement effects is to apply a quantum correction to the potential [33]. There exists a multitude of models and derivations for the quantum potential [31, 34, 35], although all of them have a similar qualitative effect. The self-consistent potential obtained from semi-classical simulation is smoothed by a convolution with a Gaussian kernel. This effective potential indirectly models the uncertainty in the position of the electrons dictated by quantum mechanics using classical calculations and, therefore, can be applied to both the BTE and macroscopic models.

Quantum versions of the macroscopic models can also be derived from the Wigner(-Boltzmann) equation using the method of moments as done in the semi-classical case with the BTE. The corrective term (Ohm potential), as defined in [31], is simply added to the classical potential in (1.14) such that

$$\mathbf{J}_n = e \left(D_n \nabla n - \mu_n n \nabla \left(V - \frac{\hbar^2}{2m^*} \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} \right) \right).$$ (1.43)

The conservation equation for mass (1.13) is retained in the quantum DD model, but the additional gradient operators in (1.43) – hence, the name density gradient method – increase the order of the resulting partial differential equation and, therefore, additional boundary conditions must be specified. The use of a generalized electron quasi-Fermi potential can, however, reduce the order of the differential equation [36], which is preferable for a solution by numerical methods.

The density-gradient method has also been used to account for tunnelling effects in the gate stack and from source to drain [37]. The tunnelling effect can also be replicated in semi-classical models by introducing an additional generation (recombination) rate in (1.14) [38].

In addition to the expected computational efficiency, a further advantage of using quantum macroscopic models is that physically relevant boundary conditions are easier to specify than for the Schrödinger equation. The quantum-corrected models have been successfully applied for device simulation [35, 39], but some numerical challenges are faced with high-frequency oscillations which occur due to the dispersive nature of the equation [40].

Many approaches have been used to model quantum effects in a computationally efficient way by combining and modifying different equations, as sketched above. However, this *ad hoc* approach works well only for specific cases and lacks the generality of a full-fledged quantum model.

1.2.5 Summary

Whereas the relation between the semi-classical transport models is rather obvious – they all stem from the BTE – the relation between the different quantum transport models is slightly obscured by the use of different terms and terminology. Before the motivation for using Wigner-based transport models is given, it is valuable to briefly summarize and evaluate the models discussed up to now.

The semi-classical transport is based on the BTE and allows an accurate description in devices where transport takes place in the diffusive regime. Macroscopic models that directly solve for physical quantities can be derived from the BTE and are much less computationally demanding. Certain quantum mechanical effects can selectively be accounted for in semi-classical models with quantum-correction terms, but the quantum physics is not treated on a fundamental level.

A fundamental description of the wave nature of particles is required to describe quantum mechanical effects. The Schrödinger equation describes the evolution of a pure state and can describe the stationary states, but the inclusion of scattering is problematic. A generalization towards mixed states requires the density matrix, which describes the correlation between different states. The inclusion of scattering in the Von Neumann equation is possible, but the non-local nature makes the

interpretation not very intuitive. The Wigner function is related to the density matrix through the unitary Fourier transform and gives a phase space description of quantum mechanics, which allows the adoption of familiar semi-classical scattering models. The most general formalism is the NEGF, which also takes time correlations into account. The high dimensionality of the Green's function imposes significant computational demands, which makes the inclusion of scattering only tractable with simplifying assumptions.

1.3   Motivation for Wigner formalism based simulation

The preceding section reviewed the transport models commonly used in device TCAD and nanoelectronics research. In light of this the main advantages of using Wigner-based simulations for quantum transport are i) the classical analogies that arise from the phase-space formalism, ii) the quite straightforward inclusion of scattering effects allowing decoherence to be investigated and iii) the ability to simulate time-resolved quantum transport with reasonable computational effort. These three points are elaborated upon in the following, whereafter some problems well-suited to be solved in this formalism are presented.

1.3.1   Classical analogies

The Wigner formulation of quantum mechanics retains many classical concepts and notions, which makes it a convenient approach to describe the transport phenomena characterizing the evolution of electrons in nanostructures. The Wigner formalism expresses quantum mechanics using functions and phase space variables, as opposed to wave functions and operators, as used in the Schrödinger equation. This phase space formulation offers a more intuitive interpretation of quantum phenomena.

The phase space representation of the Wigner formalism provides a clear analogy to classical notions. The most striking example of this is the (initially phenomenological) augmentation of the Wigner equation with the semi-classical scattering models used in the Boltzmann transport equation; a formal derivation introducing the scattering terms to the Wigner equation has now been shown [41]. Furthermore, the phase space formulation is advantageous to specify and recover classical distributions at boundaries [42].

Some caution, however, is justified when trying to apply concepts from classical physics to interpret the Wigner picture. A common mistake is interpreting the Wigner function as a true probability density function, which it is not. Unlike a distribution function, the Wigner function may attain negative values, which are a manifestation of the uncertainty relation in the phase space [43, 44]. Nonetheless, the Wigner function retains the necessary properties of a true distribution function, which allows the calculation of physical averages using the same expressions as used in the case of the Boltzmann formulation. Therefore, the Wigner function is sometimes called a quasi-distribution function. Alternative phase space formulations, like the Husimi function [45], recover a positive definite function by smoothing the Wigner function over a wavelength with a Gaussian kernel, but fail in other respects [45].

1.3.2   Decoherence and scattering

The straight-forward inclusion of scattering mechanisms in the Wigner formalism enables the description of decoherence processes which are of fundamental interest, when investigating the evolution of quantum states. A hierarchy of Wigner transport models with scattering can be derived: These begin with the simple relaxation time approximation, the Wigner-Boltzmann equation [41, 46], which accounts for scattering by phonons and impurities at the classical transport level, and end with the quite complicated Levinson and Barker-Ferry equations, which account for the quantum character of the interaction with the sources of decoherence [47]. Of central interest is the Wigner-Boltzmann equation which, as the name suggests, unifies the two theories and ensures a seamless transition between purely coherent and classical transport [48] – the Wigner function gradually turns into the Boltzmann distribution function. This transition occurs either, when phonon scattering is significant

or the potential varies very smoothly [20, 49]. The Wigner formalism bridges the gap between purely quantum (ballistic) and classical (diffusive) transport in a seamless fashion.

Currently, the Wigner-Boltzmann equation presents the only computationally attractive quantum model to consider scattering effects in multi-dimensional simulations. Nanoscale devices, e.g. silicon nanowires [50], which exploit coherent quantum phenomena, should be able to operate at room temperature to be viable as a commercial technology. Since phonon scattering increases with temperature, the inclusion of scattering in simulation is of significant importance to simulate the decoherence effects taking place.

1.3.3   Transient capability

The transient evolution of quantum states is of considerable interest in nanoelectronic devices as "Many basic concepts remain to be clarified in the area of time-dependent current flow as well as current fluctuations" [9]. Nano-circuits, which are formed by simple nanostructures, exhibit an electrical behaviour which cannot be explained by classical theory. Time-resolved simulations of quantum transport will help to resolve the questions which surround the frequency response of quantum capacitors and resistances. The Wigner formalism easily allows the time-dependent behaviour, like oscillations and switching times, to be investigated. This is an aspect which cannot currently be investigated by NEGF simulations due to the excessive computational costs.

Wigner-based simulations require the consideration of seven dimensions, whereas NEGF simulations require eight. The higher dimensionality of the NEGF simulations makes it inherently more computationally demanding, but allows the description of temporal correlations. This theoretical advantage has not been exploited up to now due to the exorbitant computational demands, although the first promising efforts are emerging [29, 30]. Transient simulations, neglecting temporal correlations, can be naturally treated in the Wigner picture. Admittedly, this is also possible with the Schrödinger equation/density matrix, however, the inclusion of scattering effects is problematic.

1.3.4   Suitable problems

Problems with (some of) the following properties are deemed well-suited to be investigated using the Wigner formalism [42, 51]:

- highly transient phenomena;

- time evolution of contact states;

- energy resolution is of lesser importance;

- many closely-spaced states participate in current flow;

- far from equilibrium where quantum effects are small;

- phase space picture is useful, e.g. chaos, boundary conditions with classical distributions.

1.4   Outline of dissertation

The body of this dissertation is arranged as follows: Chapter 2 gives an introduction to the Wigner formalism of quantum mechanics and the associated Wigner transport equation. The latter is augmented with the Boltzmann scattering models to yield the Wigner-Boltzmann equation. An overview of the existing deterministic and stochastic solvers for this equation is given. Chapter 3 presents the integral form of the Wigner-Boltzmann equation from which the signed-particle method is derived; the corresponding Monte Carlo approach and the basic building blocks of its algorithm are discussed. Chapter 4 gives a detailed presentation of the various optimizations which have been made to the signed-particle algorithms within the scope of this thesis. The improvements include algorithms with better computational efficiency, statistical enhancements and increased physical accuracy by mitigating discretization artefacts. A validation of the simulation results to illustrate the impact of the

enhancements is also shown. Chapter 5 treats the parallelization of the Wigner Monte Carlo (WMC) solver in a distributed-memory environment. The spatial domain decomposition approach is shown and its performance is analysed using a few benchmark tests. Chapter 6 illustrates some applications of the WMC simulator by investigating the various aspects the interaction of wavepackets with electrostatic lenses. Finally, in Chapter 7 the contributions which have been made in the scope of this thesis are reviewed and evaluated.

Chapter 2

Wigner Formalism of Quantum Mechanics

This chapter first discusses the Wigner transport equation and its derivation. Thereafter, the addition of scattering models to the Wigner equation is discussed, which leads to the Wigner-Boltzmann equation (WBE). The semi-discrete form of the Wigner equation is then presented. Finally, an overview of the existing solvers for the WBE is given.

2.1 Wigner equation

2.1.1 History and derivation

The Wigner function was introduced by Eugene Wigner in 1932 [52] with the motivation to add a quantum correction to the configuration of gases at low temperatures. He also derived the Wigner equation which describes the temporal evolution of the Wigner function. The Wigner formalism was originally deduced from the operator formalism. However, later a fully independent derivation was developed based on the Moyal bracket [53, 54].

The Wigner representation of an arbitrary physical quantity, like energy, is a function of the phase space coordinates. A mapping of the involved phase space functions to quantum mechanical operators is needed. Since quantum mechanical operators corresponding to position and momentum are not commutative, the ordering in which the variables appear matters in general and a simple correspondence principle is not sufficient to define a unique mapping. To remove the ambiguity in the mapping an additional 'rule' is imposed, which defines the quantization scheme. Various quantization schemes have been proposed: normal ordering in which the position operator always precedes the momentum operator, anti-normal ordering (the converse) and fully symmetrized ordering (a mixture of both).

Another quantization scheme is the Weyl transform introduced by Hermann Weyl in 1927 [55]. The inverse mapping, from operators to functions, is known as the Wigner transform and is defined as

$$A\left(\mathbf{r}, \mathbf{k}, t\right) = \int_{-\infty}^{\infty} d\mathbf{s}\, e^{-i\mathbf{k}\cdot\mathbf{s}} \left\langle \mathbf{r} + \frac{\mathbf{s}}{2} \left| \hat{A} \right| \mathbf{r} - \frac{\mathbf{s}}{2} \right\rangle. \tag{2.1}$$

The Wigner transform is applied to the density operator and yields the Wigner function

$$f_w\left(\mathbf{r}, \mathbf{k}, t\right) = \int_{-\infty}^{\infty} d\mathbf{s}\, e^{-i\mathbf{k}\cdot\mathbf{s}} \rho\left(\mathbf{r} + \frac{\mathbf{s}}{2}, \mathbf{r} - \frac{\mathbf{s}}{2}, t\right), \tag{2.2}$$

which amounts to a Fourier transform of the density matrix, expressed in the mean and difference of coordinates:

$$\mathbf{r} \equiv \frac{\mathbf{r}_1 + \mathbf{r}_2}{2}; \tag{2.3}$$

$$\mathbf{s} \equiv \mathbf{r}_1 - \mathbf{r}_2. \tag{2.4}$$

Therefore,

$$\rho\left(\mathbf{r}_1, \mathbf{r}_2, t\right) = \rho\left(\mathbf{r} + \frac{\mathbf{s}}{2}, \mathbf{r} - \frac{\mathbf{s}}{2}, t\right). \tag{2.5}$$

The Liouville-Von Neumann equation (1.34), introduced in Chapter 1, describes the evolution of the density matrix and is expressed here using the variables defined in (2.3):

$$\frac{\partial}{\partial t}\rho\left(\mathbf{r} + \frac{\mathbf{s}}{2}, \mathbf{r} - \frac{\mathbf{s}}{2}, t\right) = \frac{1}{i\hbar}\left\{-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial\mathbf{r}\partial\mathbf{s}} + \left(V\left(\mathbf{r} + \frac{\mathbf{s}}{2}\right) - V\left(\mathbf{r} - \frac{\mathbf{s}}{2}\right)\right)\right\}\rho\left(\mathbf{r} + \frac{\mathbf{s}}{2}, \mathbf{r} - \frac{\mathbf{s}}{2}, t\right). \tag{2.6}$$

Only a single differential operator remains – this is a convenience afforded by the assumption of a parabolic dispersion relation. However, the incorporation of more complicated bandstructures in the Wigner formalism has been approached by [56, 57].

The Wigner transport equation is derived by applying the Wigner transform (2.1) to (2.6); the left-hand side (LHS) immediately yields the time derivative of the Wigner function. The transform of the first term on the right-hand side (RHS), related to the spatial derivatives, gives

$$\int d\mathbf{s}e^{-i\mathbf{k}\cdot\mathbf{s}}\left\{-\frac{1}{i\hbar}\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial\mathbf{r}\partial\mathbf{s}}\rho\left(\mathbf{r} + \frac{\mathbf{s}}{2}, \mathbf{r} - \frac{\mathbf{s}}{2}, t\right)\right\} = \frac{i\hbar}{2m^*}\frac{\partial}{\partial\mathbf{r}}\int d\mathbf{s}e^{-i\mathbf{k}\cdot\mathbf{s}}\left(i\mathbf{k}\right)\rho\left(\mathbf{r} + \frac{\mathbf{s}}{2}, \mathbf{r} - \frac{\mathbf{s}}{2}, t\right)$$

$$= -\frac{\hbar\mathbf{k}}{2m^*}\frac{\partial}{\partial\mathbf{r}}f_w\left(\mathbf{r}, \mathbf{k}, t\right), \tag{2.7}$$

where the derivative $\frac{\partial}{\partial\mathbf{s}}$ corresponds to a multiplication by $i\mathbf{k}$ in the transformed space. The transform of the remaining, potential-related terms on the RHS makes use of the property $f(\mathbf{s}) = \int d\mathbf{r}_2\delta\left(\mathbf{s} - \mathbf{r}_2\right)f\left(\mathbf{r}_2\right)$ to change the variable in the density matrix and yields

$$\int d\mathbf{s}e^{-i\mathbf{k}\cdot\mathbf{s}}\left\{-\frac{1}{i\hbar}\left(V\left(\mathbf{r} + \frac{\mathbf{s}}{2}\right) - V\left(\mathbf{r} - \frac{\mathbf{s}}{2}\right)\right)\rho\left(\mathbf{r} + \frac{\mathbf{s}}{2}, \mathbf{r} - \frac{\mathbf{s}}{2}, t\right)\right\}$$

$$= \int d\mathbf{s}e^{-i\mathbf{k}\cdot\mathbf{s}}\left\{-\frac{1}{i\hbar}\left(V\left(\mathbf{r} + \frac{\mathbf{s}}{2}\right) - V\left(\mathbf{r} - \frac{\mathbf{s}}{2}\right)\right)\int d\mathbf{r}_2\delta\left(\mathbf{s} - \mathbf{r}_2\right)\rho\left(\mathbf{r} + \frac{\mathbf{r}_2}{2}, \mathbf{r} - \frac{\mathbf{r}_2}{2}, t\right)\right\}. \tag{2.8}$$

The property $\delta\left(\mathbf{s} - \mathbf{r}_2\right) = \frac{1}{(2\pi)^3}\int d\mathbf{k}'e^{i\mathbf{k}'\cdot(\mathbf{s}-\mathbf{r}_2)}$ then introduces a third integral such that (2.8) finally becomes

$$-\frac{1}{i\hbar(2\pi)^3}\left\{\int d\mathbf{k}'\int d\mathbf{s}e^{i\mathbf{s}\cdot(\mathbf{k}'-\mathbf{k})}\left(V\left(\mathbf{r} + \frac{\mathbf{s}}{2}\right) - V\left(\mathbf{r} - \frac{\mathbf{s}}{2}\right)\right)\right.$$

$$\left. \times \int d\mathbf{r}_2 e^{-i\mathbf{k}'\cdot\mathbf{r}_2}\rho\left(\mathbf{r} + \frac{\mathbf{r}_2}{2}, \mathbf{r} - \frac{\mathbf{r}_2}{2}, t\right)\right\} \tag{2.9}$$

$$= -\frac{1}{i\hbar(2\pi)^3}\int d\mathbf{k}'V_w\left(\mathbf{r}, \mathbf{k}' - \mathbf{k}, t\right)f_w\left(\mathbf{r}, \mathbf{k}', t\right),$$

where V_w is referred to as the Wigner potential and is defined as

$$V_w\left(\mathbf{r}, \mathbf{k}' - \mathbf{k}, t\right) \equiv -\frac{1}{i\hbar(2\pi)^3}\int d\mathbf{s}e^{i\mathbf{s}\cdot(\mathbf{k}'-\mathbf{k})}\left\{V\left(\mathbf{r} + \frac{\mathbf{s}}{2}\right) - V\left(\mathbf{r} - \frac{\mathbf{s}}{2}\right)\right\}. \tag{2.10}$$

Take note of the sign in the exponent, which differs from the Wigner transform defined in (2.1). These calculations are combined and finally yield the Wigner transport equation:

$$\frac{\partial}{\partial t}f_w\left(\mathbf{r}, \mathbf{k}, t\right) + \frac{\hbar\mathbf{k}}{2m^*}\frac{\partial}{\partial\mathbf{r}}f_w\left(\mathbf{r}, \mathbf{k}, t\right) = \int d\mathbf{k}'V_w\left(\mathbf{r}, \mathbf{k}' - \mathbf{k}, t\right)f_w\left(\mathbf{r}, \mathbf{k}', t\right). \tag{2.11}$$

2.1.2 Properties

The Wigner function may assume both positive and negative values, which is a manifestation of the quantum uncertainty principle [44]. Nonetheless, a critical property of a probability distribution is retained:

$$\int d\mathbf{k}\int d\mathbf{r}\, f_w\left(\mathbf{r}, \mathbf{k}, t\right) = 1, \tag{2.12}$$

which allows the Wigner function to be used (but not interpreted) in the same way as a classical distribution function to calculate physical averages. Therefore, the Wigner function is often termed a quasi-distribution function. The non-negativity of the Wigner distribution function is only guaranteed for wave functions of the form $\psi\left(\mathbf{r}\right) = \exp\left(-\mathbf{r}^T A\left(t\right)\mathbf{r} - \mathbf{r}a\left(t\right) - b\left(t\right)\right)$, where $A(t)$, $a(t)$ and $b(t)$ are a complex-valued matrix, vector and number, respectively [58].

The use of Wigner functions to describe quantum transport in semiconductor devices, which represent open quantum systems, is reviewed in [59]: the Wigner function of pure states is introduced in terms of generalized functions and, furthermore, it is proven that the mean value of an observable can be calculated as a weighted average of the mean values of the observable for the pure state. The convergence problems that arise when a potential difference occurs between the boundaries of the domain – as is the case for open quantum systems – can be treated by introducing a damping factor to the Wigner potential (2.10) [60].

The Wigner transport equation is a linear pseudo-differential equation. The Wigner equation (2.11) reduces to the Vlasov equation (collision-less Boltzmann equation) for quadratic and linear potentials [52]. This can be readily shown by expressing the potential as a Taylor expansion in (2.11) [61]. Therefore, dynamic quantum effects manifest themselves as derivatives of the potential of third-order and higher. In the semi-classical limit ($\hbar \to 0$) the Vlasov equation is recovered, regardless of the potential profile.

2.2 Wigner-Boltzmann equation

The Wigner equation introduced in Section 2.1 describes ballistic carrier transport. A widespread adoption of a semiconductor device technology requires operation at room temperature where phonon scattering plays an important role. One of the major advantages of the Wigner formalism is that the treatment of scattering is relatively simple compared to other quantum mechanical formalisms.

The phase-space description used in the Wigner formalism invites the use of the semi-classical scattering operator used in the Boltzmann equation. The addition of this semi-classical scattering model to the WTE was first proposed in [62], as an *ad hoc* solution to treating scattering in RTDs. This raised questions whether the use of semi-classical scattering is justified in the Wigner formalism. A rigorous derivation within the Wigner formalism for both phonon [41, 63] and impurity scattering [64], has shown that the semi-classical scattering models can be obtained as a limiting case of full quantum models. This derivation is outlined in the following.

2.2.1 Derivation

The Wigner equation can be generalized to account for electron-phonon interactions by considering the interactions of an electron with a quantum field; a classical field interaction is not sufficient [64]. A general system comprising both the electron and phonon populations is required, but under certain approximations the effects of the phonons on the electrons can be averaged. The phonon population is described by a collection of integers $\{n_{\mathbf{g}}\}$, giving the number of phonons (n) in the phonon mode $\mathbf{g}$. This introduces two additional variables to the Wigner function: $f_w\left(\mathbf{r}, \mathbf{k}, \{n_{\mathbf{g}}\}, \{n'_{\mathbf{g}}\}, t\right)$.

A general procedure to incorporate additional physical phenomena in the Wigner formalism is to i) express the appropriate Hamiltonian for the phenomenon to be described, ii) expand the commutator bracket in the Liouville-Von Neumann equation (2.6), iii) apply the Wigner transform and iv) perform mathematical calculations and physically-motivated approximations to obtain a tractable equation. In the following, this procedure for the derivation of the scattering models for phonon and impurity scattering in the Wigner formalism is outlined.

2.2.1.1 Phonon scattering

The Hamiltonian (1.19), describing a free electron, must be augmented to account for the free phonons and their interaction with the electron:

$$
\begin{aligned}
\hat{H} &= \hat{H}_0 + \hat{H}_{ph} + \hat{H}_{e-ph} \\
&= \left\{ -\frac{\hbar^2}{2m^*}\nabla_r^2 + V(\mathbf{r}) \right\} + \sum_{\mathbf{g}} \hbar\omega_{\mathbf{g}} \left(a_{\mathbf{g}}^\dagger a_{\mathbf{g}} + \frac{1}{2} \right) + i\hbar \sum_{\mathbf{g}} F(\mathbf{g}) \left(a_{\mathbf{g}} e^{i\mathbf{g}\cdot\mathbf{r}} + a_{\mathbf{g}}^\dagger e^{-i\mathbf{g}\cdot\mathbf{r}} \right).
\end{aligned} \quad (2.13)
$$

The second term yields the energy of all phonons: $\hbar\omega_{\mathbf{g}}$ denotes the energy of a phonon with wavevector (mode) $\mathbf{g}$; the creation and annihilation operators, denoted by $a_{\mathbf{g}}^\dagger$ and $a_{\mathbf{g}}$, form the number operator $a_{\mathbf{g}}^\dagger a_{\mathbf{g}}$. The third term describes the energy exchanged between electrons and phonons: the function $F(\mathbf{g})$ describes the coupling of electrons to phonons specific to the type of phonon scattering being considered (refer to Appendix A). Each contribution to the total Hamiltonian (2.13) can be considered separately and transformed; the additional terms arising from $\hat{H}_{ph}$ and $\hat{H}_{e-ph}$ augment the Wigner equation (2.11).

The contribution of the second term in (2.13) yields a contribution to the RHS of (2.11) of

$$
C_{ph} = \frac{1}{i\hbar} \left\{ \epsilon\left(\{n_{\mathbf{g}}\}\right) - \epsilon\left(\{n_{\mathbf{g}}'\}\right) \right\} f_w\left(\mathbf{r}, \mathbf{k}, \{n_{\mathbf{g}}\}, \{n_{\mathbf{g}}'\}, t\right), \quad (2.14)
$$

where $\epsilon\left(\{n_{\mathbf{g}}\}\right)$ is the total energy of the phonons, given by

$$
\epsilon\left(\{n_{\mathbf{g}}\}\right) = \sum_{\mathbf{g}} n_{\mathbf{g}} \hbar\omega_{\mathbf{g}}. \quad (2.15)
$$

The contribution of the third term in (2.13) gives rise to four terms of a similar form:

$$
\begin{aligned}
C_{e-ph} = \sum_{\mathbf{g}'} F(\mathbf{g}') \;\; & \left\{ e^{i\mathbf{g}'\cdot\mathbf{r}} \sqrt{n_{\mathbf{g}'}+1}\, f_w\left(\mathbf{r}, \mathbf{k} - \frac{\mathbf{g}'}{2}, \{n_{\mathbf{g}}\}_{\mathbf{g}'}^+, \{n_{\mathbf{g}}'\}, t\right) + \right. \\
& + e^{-i\mathbf{g}'\cdot\mathbf{r}} \sqrt{n_{\mathbf{g}'}'+1}\, f_w\left(\mathbf{r}, \mathbf{k} - \frac{\mathbf{g}'}{2}, \{n_{\mathbf{g}}\}, \{n_{\mathbf{g}}\}_{\mathbf{g}'}^+, t\right) - \\
& - e^{-i\mathbf{g}'\cdot\mathbf{r}} \sqrt{n_{\mathbf{g}'}}\, f_w\left(\mathbf{r}, \mathbf{k} + \frac{\mathbf{g}'}{2}, \{n_{\mathbf{g}}\}_{\mathbf{g}'}^-, \{n_{\mathbf{g}}'\}, t\right) - \\
& \left. - e^{i\mathbf{g}'\cdot\mathbf{r}} \sqrt{n_{\mathbf{g}'}'}\, f_w\left(\mathbf{r}, \mathbf{k} + \frac{\mathbf{g}'}{2}, \{n_{\mathbf{g}}\}, \{n_{\mathbf{g}}\}_{\mathbf{g}'}^-, t\right) \right\}.
\end{aligned} \quad (2.16)
$$

The first (last) two terms in the curly braces correspond to phonon creation (absorption). The notation $\{n_{\mathbf{g}}\}_{\mathbf{g}'}^{\pm}$ signifies that the integer corresponding to mode $\mathbf{g}'$ is increased/decreased by one, i.e. $n_{\mathbf{g}'} \to n_{\mathbf{g}'} \pm 1$.

The addition of (2.14) and (2.16) to (2.11) gives a generalized Wigner equation describing the interaction of a single electron with a many-phonon system in a quantum mechanical manner. This equation, however, is computationally completely intractable: each term in the summation for a given mode $\mathbf{g}'$ involves the Wigner function of mode $n_{\mathbf{g}'} \pm 1$ – a recursion is formed. Since the summation is over all modes $\mathbf{g}'$ (infinite), a closure relation is needed to obtain some tractable form of the equation.

The aim is to obtain with appropriate assumptions a reduced Wigner function, $f_w\left(\cdot, \{n_{\mathbf{g}}\}, \{n_{\mathbf{g}}\}, \cdot\right)$, which describes only the electron subsystem. The phonon subsystem is 'traced out' such that only the first off-diagonal terms of the phonon population are considered.

The weak-coupling limit considers the electron interacting with only a single phonon, i.e. the time between two consecutive scattering events is assumed to be long enough such that the first event is completed before the next one starts. In the weak-coupling limit $F(\mathbf{g})$ is assumed to be so small that $F^2(\mathbf{g})$ becomes negligible, i.e. the diagonal terms of the phonon states are considered ($n = n'$) to interact only with the the first off-diagonal elements. This implies that

$$
f_w\left(\mathbf{r}, \mathbf{k}, \{n_{\mathbf{g}}\}, \{n_{\mathbf{g}}'\}, t\right) = 0, \quad (2.17)
$$

if $|n - n'| > 1$. Correlations and many-phonon processes are thus ignored.

The phonon distribution $\{n_{\mathbf{g}}\}$ is determined using the assumption that phonons remain in equilibrium, regardless of the intensity of the interactions with the electron. The equilibrium distribution of phonons follows Bose-Einstein statistics, such that the probability of n phonons being in mode $\mathbf{g}$ is given by

$$P\left(n_{\mathbf{g}}\right) = \frac{1}{\bar{n}_{\mathbf{g}} + 1} \exp\left(-n_{\mathbf{g}} \frac{\hbar \omega_{\mathbf{g}}}{k_B T}\right), \tag{2.18}$$

where $\bar{n}_g$ denotes the mean occupation number, given by

$$\bar{n}_{\mathbf{g}} = \sum_{n_{\mathbf{g}}} n_{\mathbf{g}} P\left(n_{\mathbf{g}}\right) = \frac{1}{e^{\hbar \omega / k_B T} - 1}. \tag{2.19}$$

The generalized Wigner function can be expressed as the product

$$f_w\left(\mathbf{r}, \mathbf{k}, \{n_{\mathbf{g}}\}, \{n_{\mathbf{g}}\}, t\right) = f_w\left(\mathbf{r}, \mathbf{k}, t\right) \prod_{\mathbf{g}} P\left(n_{\mathbf{g}}\right), \tag{2.20}$$

under the assumption that the equilibrium conditions in the phonon distribution is instantaneously recovered after an interaction.

After performing a trace operation over all the phonon states, the transport equation for the reduced Wigner function can be expressed as an integral equation:

$$\begin{aligned}
f_w\left(\mathbf{r}, \mathbf{k}, t\right) &= f_w\left(\mathbf{r}, \mathbf{k}, 0\right) + \int_0^t dt' \left[\int d\mathbf{k} V_w\left(\mathbf{r}, \mathbf{k}' - \mathbf{k}\right) f_w\left(\mathbf{r}, \mathbf{k}', t'\right) + \right.\\
&\quad + \sum_{\mathbf{g}'} F^2\left(\mathbf{g}'\right) \left\{e^{i\mathbf{g}' \cdot \mathbf{r}} f_1\left(\mathbf{r}, \mathbf{k} - \frac{\mathbf{g}'}{2}, t'\right) \right.\\
&\quad \left.\left. - e^{-i\mathbf{g}' \cdot \mathbf{r}} f_2\left(\mathbf{r}, \mathbf{k} + \frac{\mathbf{g}'}{2}, t'\right) + cc\right\}\right],
\end{aligned} \tag{2.21}$$

where cc denotes the complex conjugate of the first two terms in the same braces. The auxiliary functions $f_{1,2}$ account for the phonon emission/absorption processes that start and end on the diagonal element $f_{1,2}\left(\mathbf{r}, \mathbf{k} \pm \frac{\mathbf{q}}{2}\right)$ [41]. The simplification which has been gained is obvious in comparison to (2.16), where phonon modes infinitely far away from the diagonal are considered. Indeed, a substitution of $f_{1,2}$ into (2.21) yields the desired closed equation for the reduced Wigner function.

Equation (2.21), known as the Levinson equation, is quantum mechanically correct in the weak-coupling limit and accounts for collisional broadening, intra-collisional field effect (ICFE) and retardation effects.

For the case of a constant homogeneous electric field, the auxiliary equation $f_{1,2}$ can be solved and the summation of the phonon-scattering terms in (2.21) reduces to

$$\int d\mathbf{k}' \int_0^t dt' \left(S\left(\mathbf{k}', \mathbf{k}, t, t'\right) f_w\left(\mathbf{r}_{p,q,t'}, \mathbf{k}'_{t'}, t'\right) - S\left(\mathbf{k}, \mathbf{k}', t, t'\right) f_w\left(\mathbf{r}_{p,q',t'}, \mathbf{k}'_{t'}, t'\right)\right), \tag{2.22}$$

where

$$\mathbf{k}_{t'} = \hbar \mathbf{k} - e\mathbf{E}\left(t - t'\right), \tag{2.23}$$

$$\mathbf{r}_{p,q',t'} = \mathbf{r} - \hbar \int_{t'}^t dy \frac{\mathbf{k}_y - \mathbf{g}/2}{m^*}. \tag{2.24}$$

The scattering rate is defined by

$$S\left(\mathbf{k}', \mathbf{k}, t, t'\right) = \frac{2V F^2\left(\mathbf{g}\right)}{(2\pi\hbar)^3} \left\{n\left(\mathbf{g}\right) \cos\left(\Omega\left(\mathbf{k}', \mathbf{k}, t, t'\right)\right) + \left(n\left(\mathbf{g}\right) + 1\right) \cos\left(\Omega\left(\mathbf{k}, \mathbf{k}', t, t'\right)\right)\right\}; \tag{2.25}$$

$$\Omega\left(\mathbf{k}, \mathbf{k}', t, t'\right) = \int_{t'}^t d\tau \frac{\epsilon\left(\mathbf{k}_\tau\right) - \epsilon\left(\mathbf{k}'_\tau\right) + \hbar \omega_{(\mathbf{k}' - \mathbf{k})}}{\hbar}. \tag{2.26}$$

An electron-phonon interaction does not occur instantaneously, but over a finite amount of time (as evidenced by the time integral in (2.21)). An electron slowly starts to 'feel' the oscillations of phonons. This is approximately the time an electron requires to travel one wavelength of a phonon [7]. From a classical point of view, the collision duration can be defined as the amount of time an electron feels the presence of the scattering (phonon) field it interacts with. Energy is conserved by the end of the interaction. In quantum mechanics, however, the position of a particle is not precisely known and the collision duration must be defined according to the chosen uncertainty in energy (various definitions have been proposed for this [7]). The energy of the electron after a scattering event is no longer known exactly – this is termed collisional broadening. If the collision duration is longer than the mean time between scattering events, the scattering events are no longer independent. In other words, the uncertainty in position of an electron should be less than the mean free path, such that an electron is involved in only a single scattering event at a time.

During the duration of the electron-phonon interaction, external forces still act on the electron, thereby changing its wavevector. Therefore, the state of the electron changes (and thereby the scattering rate) during the collision – this is known as the ICFE. Modern semiconductor devices are subject to electric fields which are strong enough such that ICFE becomes significant and the scattering rates of instantaneous collisions do no longer apply [7].

The time scale of the phonon interaction appears in (2.26) as

$$\frac{\epsilon\left(\mathbf{k}_\tau\right) - \epsilon\left(\mathbf{k}_\tau'\right) + \hbar\omega_{(\mathbf{k}'-\mathbf{k})}}{\hbar} \tag{2.27}$$

and can be assumed to be much faster than the time scale associated with the electron dynamics, i.e. significant changes in the value of f_w. This amounts to the assumption of instantaneous collisions, already mentioned for the Boltzmann collision operator in Chapter 1. In the classical limit ($\hbar \to 0$) the time integration can be approximated and (2.25) takes the form

$$\begin{aligned} S\left(\mathbf{k}', \mathbf{k}, t, t'\right) &= \frac{V}{\hbar^3} \frac{2\pi}{\hbar} \left\{ |\hbar F(\mathbf{g})|^2 \, n\left(\mathbf{g}\right) \delta\left(\epsilon\left(\mathbf{k}\right) - \epsilon\left(\mathbf{k}'\right) - \hbar\omega_{\mathbf{k}'-\mathbf{k}}\right) + \right. \\ &\quad \left. + |\hbar F\left(-\mathbf{g}\right)|^2 \left(n\left(-\mathbf{g}\right) + 1\right) \delta\left(\epsilon\left(\mathbf{k}\right) - \epsilon\left(\mathbf{k}'\right) + \hbar\omega_{\mathbf{k}'-\mathbf{k}}\right) \right\}, \end{aligned} \tag{2.28}$$

which recovers the Boltzmann scattering term for phonon scattering and neglects the ICFE.

The above derivations illustrate well how the Wigner picture gives the opportunity to account for electron-phonon interactions at different levels, forming a hierarchy ranging from a full quantum mechanical description of scattering to the semi-classical description of scattering as used in the Boltzmann equation, which is more practical due to the much smaller computational demands. This presents a big advantage of using the Wigner formalism for scattering-aware quantum transport simulations.

2.2.1.2 Impurity scattering

An ionized impurity (dopant) exerts a force on an electron due to the Coulomb potential associated with it. To account for Coulomb interactions in microscopic simulations requires several considerations [65]. The Poisson equation appropriately models the long-range effects of a screened Coulomb potential in a continuous distribution of impurities (dopants). However, the short-range Coulomb interaction between an electron and an impurity is not captured by the Poisson equation alone. The short-range interaction is accounted for by a scattering mechanism.

The appropriate model in the Wigner formalism can be derived – as has been shown in [64] – by calculating the Wigner potential (2.10) using the Coulomb potential instead of the Hartree potential. The Coulomb potential of a discrete distribution of ionized dopants is given by

$$V_{e-ii}\left(\mathbf{r}\right) = \frac{e^2}{4\pi\kappa} \sum_j \frac{\exp\left(-\beta\left|\mathbf{r} - \mathbf{r}_j\right|\right)}{\left|\mathbf{r} - \mathbf{r}_j\right|}, \tag{2.29}$$

where κ denotes the dielectric constant, β the screening factor (inverse of Debye length) and $\mathbf{r}_j$ the position of dopant j. The corresponding evolution term emerges as

$$\frac{e^2}{i\hbar\pi^3\kappa}\int d\mathbf{k}' f_w\left(\mathbf{r},\mathbf{k}'\right)\frac{1}{4\left(\mathbf{k}-\mathbf{k}'\right)^2+\beta^2}\times\sum_j\left(e^{-2i(\mathbf{k}-\mathbf{k}')\cdot(\mathbf{r}-\mathbf{r}_j)}-e^{2i(\mathbf{k}-\mathbf{k}')\cdot(\mathbf{r}-\mathbf{r}_j)}\right) \tag{2.30}$$

Under the assumption of a sufficiently high concentration of impurities, such that the sum can be approximated by an integral and the assumption of instantaneous collisions, this expression becomes identical to the one used in the Boltzmann equation to describe the impurity scattering (see Appendix A).

2.3 Discretization of momentum space

2.3.1 Semi-discrete Wigner equation

The Wigner function (2.2) is defined with the Wigner transform (2.1) calculated over an infinite range. However, the finite dimensions of the simulation domain impose bounds on the variables defined by (2.3) such that the maximum value that $\mathbf{s}$ can attain is limited by the dimensions of the device ($\mathbf{L}_{dev}$). Therefore, the Wigner function (2.2) must be calculated over finite dimensions:

$$f_w\left(\mathbf{r},\mathbf{k},t\right)=\frac{1}{\mathbf{L}}\int_{-\mathbf{L}/2}^{\mathbf{L}/2}d\mathbf{s}\,e^{-i2\mathbf{k}\cdot\mathbf{s}}\rho\left(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s},t\right), \tag{2.31}$$

with the substitution of variables $\frac{\mathbf{s}}{2}\to\mathbf{s}$. The value $\mathbf{L}$ is termed the coherence length and can be chosen freely subject to certain physical and computational considerations, which are investigated in Chapter 4. An isotropic coherence length is chosen such that $|\mathbf{L}|=L$.

A continuous function $f(s)$ can be written as a Fourier series:

$$f\left(s\right)=\sum_{n=-\infty}^{\infty}A_n e^{i\frac{2\pi n s}{L}}, \tag{2.32}$$

where the Fourier coefficients $\{A_n\}$ are given by

$$A_n=\frac{1}{L}\int_{-L/2}^{L/2}ds\,f(s)e^{-i\frac{2\pi n s}{L}}. \tag{2.33}$$

Therefore, a finite value of L requires $\mathbf{k}$ to become discretized to form a complete orthogonal basis set $\left\{e^{-i2\pi\mathbf{q}\Delta k\cdot\mathbf{s}}\right\}$, where $\mathbf{q}$ is an integer multi-index and $\Delta k=\frac{\pi}{L}$, which denotes the resolution of the discretized wavevector.

From the above considerations, the discrete Fourier transform (DFT) follows, such that

$$f_w\left(\mathbf{r},\mathbf{q}\Delta k,t\right)=\frac{1}{L}\sum_{\mathbf{q}}e^{-i\mathbf{q}\Delta k\cdot\mathbf{s}}\rho\left(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s},t\right), \tag{2.34}$$

where Δk is, henceforth, omitted from the function arguments for brevity. Applying the same arguments to (2.11) yields:

$$\left(\frac{\partial}{\partial t}+\frac{\hbar\mathbf{q}\Delta k}{m^*}\nabla_{\mathbf{r}}\right)f_w\left(\mathbf{r},\mathbf{q},t\right)_w=\sum_{\mathbf{q}}V_w\left(\mathbf{r},\mathbf{q}-\mathbf{q}'\right)f_w\left(\mathbf{r},\mathbf{q}',t\right), \tag{2.35}$$

The Wigner potential (which may also be time-dependent) is defined accordingly as

$$V_W\left(\mathbf{r},\mathbf{q}\right)\;\equiv\;\frac{1}{i\hbar L}\int_{-\mathbf{L}/2}^{\mathbf{L}/2}d\mathbf{s}\,e^{-i2\mathbf{q}\Delta k\cdot\mathbf{s}}\delta V \tag{2.36}$$

$$\delta V\left(\mathbf{s};\mathbf{r}\right)\;\equiv\;V\left(\mathbf{r}+\mathbf{s}\right)-V\left(\mathbf{r}-\mathbf{s}\right). \tag{2.37}$$

The equations (2.35) - (2.36) define the semi-discrete Wigner equation, which has been the subject of detailed mathematical analyses [66, 67].

2.4 Transport problems

The WTE can be used to solve stationary and transient transport problems. In both cases the specification of appropriate boundary conditions and initial conditions is paramount. This will be discussed in the following.

2.4.1 Boundary and initial conditions

To have a well-posed problem (2.35) must be specified with an initial condition at time t_0, which specifies the Wigner function at t_0 over the entire phase space. In a finite domain the initial condition is specified only in the phase space represented by the domain while the boundary conditions reflect the influence (if any) of the initial condition outside the domain under observation for all times $t \geq t_0$ [68, 69]. The specification of boundary conditions in open quantum systems has been the subject of intense investigation over many years and is a critical aspect, regardless of the formalism used to describe the problem, see e.g. [62, 70].

The phase-space formulation of the Wigner function makes the adoption of classical, equilibrium distributions at the boundaries an attractive proposition, foremost because these distributions are known *a priori* and the experience gained from semi-classical Monte Carlo simulations [71] can be built upon. However, due to the spatial correlations that exist in the WTE (through the integrand and the Wigner potential (2.36)), the boundary conditions of an equilibrium distribution should be specified infinitely far away to avoid correlations with the non-equilibrium distribution in the active region of the device. As specified in Section 2.3, based on physical and computational considerations, a finite coherence length is chosen. Therefore, once the potential changes negligibly within the coherence length, it is reasonable to assume that the distribution approaches equilibrium values within the applicable relaxation time. These conditions are given in metals (contacts), where processes are at play which destroy coherence.

Numerical studies have demonstrated that the solution of the WTE indeed depends on the size of the contact regions (extensions from the active region) [72]. What constitutes *sufficiently far away* for the classical distribution to be recovered depends on the device being simulated and a sensitivity analysis should be performed to obtain the optimal distance. A distance of 30 nm (Si) to 60 nm(GaAs) for single and double barrier structures has been suggested [73] for simulations at room temperature; at lower temperatures a larger separation is required, because the coherence length increases as scattering decreases. It is desirable to place the boundaries as close as possible to the active region to reduce the computational burden.

The use of device-specific boundary conditions has been proposed [74], where the influence of the electrostatic potential inside the device on the distribution function at the boundary is taken into consideration, thereby allowing the boundaries to be placed closer to the active region. The unbiased (equilibrium) solution is used to specify the boundary condition for the non-equilibrium case. This requires the wave function under equilibrium conditions to be obtained from analytic solutions (only possible for simple potential barriers) or a numerical solution of the Schrödinger equation, from which the Wigner function can be obtained. However, the required calculations incur additional (once-off) computational, which must be weighed against the savings gained by having a smaller domain.

It is important to note that the specification of initial and boundary conditions which are physically valid and justified is of critical importance [75]. If physical aspects are left out of consideration – as was done in [76] – the WTE can yield non-physical solutions. The question of what constitutes a physically admissible Wigner function has been investigated in [54, 75]. The appropriate initial condition for the Wigner function can be obtained from a density matrix, which can be calculated using techniques outlined in [77].

One can distinguish between three types of boundary conditions:

Absorbing: An absorbing boundary is based on the assumption that physical processes are in place, which make the boundary reflection-less. The absorbing boundary conditions for the Wigner transport equation have been mathematically derived in [78] and analysed in [79]. The implementation

of absorbing boundaries in Monte Carlo simulations is straight-forward: All particles cease to exist at the boundary and the boundary does not influence the evolution of the problem.

Injecting / in-flow: The in-flow boundary conditions retain the properties of absorbing boundaries, but also inject particles into the domain according to a specified distribution. An injecting boundary functions analogously to a black-body for radiation in that it emits (injects) electrons according to a thermal equilibrium distribution (say), regardless of the electrons that are absorbed [80]. This requires the energy relaxation in the contacts to be sufficiently fast for the contact to be regarded as memory-less and treated in a Markovian manner [81], i.e. the electron is absorbed/emitted by the boundary irrespectively of the electrons that were absorbed/emitted in the time prior. The in-flow boundary conditions for the Wigner formalism were first applied in the study of RTDs [80].

Reflecting: Particles are specularly reflected from the boundary and no particles are injected from the boundary. This approximates a boundary to an infinite potential step and assumes that no particles exist outside the domain under observation, at the time of initialization, which could enter the domain through the boundary. Reflecting boundaries are useful to approximate interfaces between semiconductors and oxide, where the wave function rapidly decays towards zero. The reflecting boundary corresponds to a zero Dirichlet boundary condition.

2.4.2 Stationary

A stationary transport problem demands a non-equilibrium solution to the transport equation (2.35) that does not change with time, i.e. $\frac{\partial f_w}{\partial t} = 0$. Stationary (steady-state) solutions are of particular interest in logic devices where the response in the 'on' and 'off' state is of primary interest and the time scales associated with the transient responses are much shorter than those associated with the operating frequency.

The relative importance of the initial and boundary conditions depends on the simulation time, domain size and the specifics of the system, e.g. the potential profile. The effect of the initial condition can become insignificant after a sufficiently long simulation time, once the particles that represented the initial condition have been absorbed by the boundaries and were replaced by particles determined solely by the boundary conditions. Conversely, if the particles that constitute the initial condition are concentrated inside the domain without ever reaching the boundary, e.g. an electron trapped in a deep quantum well, the boundary conditions are not influential and the initial condition determines the stationary solution that will be achieved.

The eigenstates of a system cannot be obtained by solving the WTE, however, the equation does 'retain' the eigensolutions, if these are correctly introduced with the initial condition.

The Wigner transport equation reduces to the Boltzmann equation for potentials of second order (quadratic) or lower. Quantum effects can be present, nonetheless, for such potentials, e.g. a quantum harmonic oscillator with a parabolic potential displays quantization effects [82]. To correctly capture these quantum effects, the boundary conditions must be specified appropriately and the time-derivative set to zero [77]. However, the time derivative $\frac{\partial f_w}{\partial t}$ term, in the WTE, can only be discarded beforehand with valid physical reasoning [75].

2.4.3 Transient

A transient transport problem investigates the time-dependent solution to the transport equation (2.35). The transient solution is of interest for quantum systems subjected to time-varying boundary conditions, e.g. RF circuits, or devices stimulated by electro-magnetic fields. While the simulation of quantum systems in the stationary state has been dealt with extensively, especially using the NEGF method, the transient problems remain largely under-investigated. This can be explained by the computational challenges faced by time-resolved quantum transport simulations using NEGF and the challenges in the metrology (photoluminscence).

In the simplest case, the transient behaviour of a quantum system can be observed by following the evolution of the initial condition while keeping the boundary conditions fixed. For systems that are

used for RF applications or are stimulated by electro-magnetic field, time-varying Dirichlet boundary conditions are used for the potential, which emulate the time-dependent potential or electromagnetic field[1].

2.5 Overview of existing solvers

The theoretical accomplishments of the Wigner formalism are accompanied by challenging and sometimes peculiar numerical aspects, when solving the associated transport equation. Several numerical methods have been explored over the years to solve the WTE. An overview of the development of the most important deterministic and stochastic solution approaches is given in the following. All the solvers mentioned below are restricted to a single spatial dimension, due to computational constraints.

2.5.1 Deterministic methods

The first deterministic solvers for the WTE applied the finite difference (FD) scheme [62, 77, 80, 83, 84] and used a relaxation time approximation for the collision operator. The numerical solution of the WTE allowed the study of physically relevant boundary conditions and demonstrated the feasibility of applying the Wigner formalism to study quantum structures, like the resonant tunnelling diode (RTD). Various refinements and additions to the FD-based solvers were made over the years [85]. However, two disadvantages of applying the finite difference scheme have become clear: The discretization of the WTE yields a dense matrix, which is numerically expensive to invert. Furthermore, the solution is sensitive to the chosen discretization of the diffusion term, due to the highly oscillatory nature of the Wigner function in regions with rapid changes in the electrostatic potential [64]. As a result, finite difference schemes remain limited in their application to single-dimensional structures of few tens of nanometres, with moderate potential variations in the active regions. Nonetheless, the high precision offered by deterministic methods remains very desirable, which motivates the continued pursuit of novel deterministic approaches.

The need to find a more efficient discretization of the non-local Wigner kernel, motivated the first alternative approach to the FD method using the spectral collocation method [86], which was later augmented with an operator-splitting technique [87]. After some years of relative inactivity, fresh efforts have started on the deterministic solvers: The spectral element method has recently been introduced [88], which intrinsically has the property of mass conservation. A weighted essentially non-oscillatory (WENO) finite difference solver [89], has tackled the numerical difficulties of the advection term with an adaptive grid in the k-space. A solution based on the integral formulation has also been demonstrated [90].

Phase space formulations always suffer under the *curse of dimensionality*, which makes a deterministic solution challenging in higher dimensions due to the fine discretization required to accommodate the highly oscillatory nature of the Wigner function. A good finite-dimensional characterization of the solution is required to reduce the size of the linear systems. Recent advancements in deterministic solvers for the BTE using wavelets may also become useful to solve the Wigner equation [91]. The application of a spherical harmonics expansion [92, 93] may also prove promising, but has only been attempted for steady-state solutions [94].

A deterministic solver for multi-dimensional problems still remains out of reach, which motivates the use of stochastic (Monte Carlo) approaches.

2.5.2 Stochastic methods

Stochastic methods offer an alternative to deterministic methods and their application to solve the Wigner equation has been inspired by the great success of the Monte Carlo approaches to the very similar Boltzmann transport equation [95, 96]. Many classical concepts have been revised and adapted

---

[1]If the dimension of the simulation domain is small compared to the wavelength of the electromagnetic stimulation, i.e. $(L_x, L_y) \lesssim \frac{\lambda}{10}$, the magnitude of the field can be approximated to be constant across the domain and only the temporal dependence of the field has to be considered. In other cases, a full coupling to a solver for the Maxwell equations is required.

to develop numerical models for computing the quantum quasi-distribution function. Nonetheless, the basis of the method remains the association of trajectories to a single or an ensemble of particles.

Wigner trajectories have been defined with the help of a quantum force [97]. They give insight in quantum phenomena like tunnelling processes, but can be created or destroyed making the important consequences of the Liouville theorem invalid for this particle model. Another particle model introduces the concept of Wigner paths [98]. Here, the action of the Wigner potential operator is interpreted as scattering, which links pieces of classical trajectories to Wigner paths.

Two more recent particle models – the affinity and signed-particle method – exhibit improved numerical efficiency and higher functionality. They unify classical and quantum regions within a single transport picture and allow the consideration of fully three-dimensional wavevector spaces in multi-dimensional devices. The affinity model represents the Wigner function as a sum of Dirac excitations in the phase-space, each weighted by an amplitude, called affinity [99]. The affinities are updated by the Wigner potential during the particle evolution and contain all the information on the quantum state of the system. The affinities can assume positive or negative values, which act as weighting factors in the reconstruction of the Wigner function and consequently in the computation of all physical averages [100]. This approach was also adopted in [64, 101] where the potential is decomposed in a 'slow' and a 'fast' varying part, representing the classical electric field (first derivative of potential) and higher-order quantum effects of the potential, respectively; only the fast-varying quantum part of the potential is used to update particle affinities.

The signed-particle method is based on the alternative interpretation of the Wigner potential as a generator of signed particles. The signed-particle method makes use of integer affinities ($+1$ and -1), which is very advantageous from a computational point of view. In all other aspects the evolution of the particle is field-less and classical. Two particles with opposite sign, which meet in the phase space, may annihilate each other, since they have an equivalent probabilistic future but make an opposite contribution in the process of averaging. Due to the ergodicity of such systems, a single particle Monte Carlo algorithm has been developed [63] and more recently the method has been generalized to also treat transient transport [49].

Indeed, currently the signed-particle method is the only computationally tractable method to solve the WTE in multiple dimensions. The signed-particle method forms the basis of this thesis and will be extensively discussed in Chapters 3 to 5.

Chapter 3

Signed-Particle Method

This chapter discusses the mathematical foundation of the signed-particle method used to solve the Wigner-Boltzmann equation using Monte Carlo techniques. The integral formulation of the WBE is introduced. The integral form is developed into a Neumann series, which is evaluated using Monte Carlo integration. Finally, the basic building blocks and architecture of the algorithm implemented in the Wigner Ensemble Monte Carlo (WEMC) simulator – as part of the open source project ViennaWD [102] – are presented.

3.1 Background

The Monte Carlo technique was developed in 1946 by Ulam and Von Neumann [103] to solve neutron transport problems in the scope of the development of the atomic bomb. Monte Carlo simulation was later also adopted for kinetic transport problems in semiconductors, around 1966, by Kurosawa [104] and others. Monte Carlo algorithms were initially developed with the idea to emulate the physical behaviour of particles rather than to solve the Boltzmann transport equation *per se*. However, as the supporting mathematical theory in statistical sampling developed, it emerged that the Monte Carlo technique can be used as a general mathematical tool with applications going far beyond solving kinetic transport equations.

Indeed, there is a well-established Monte Carlo theory to solve (especially higher-dimensional) integrals or large systems of linear equations and integral equations efficiently [105]. Monte Carlo algorithms are better suited, compared to deterministic methods, to solve problems with a low regularity (smoothness) [106] and can be much more memory efficient, if large systems/domains are considered.

One can distinguish between grid and grid-free Monte Carlo algorithms to solve an equation: The former approach entails a discretization of the equation, whereafter the resulting system of linear algebraic equations is solved using Monte Carlo techniques. In certain cases, e.g. large integration domains, grid Monte Carlo algorithms show a superior computational complexity compared to grid-free algorithms [107]. Grid-free Monte Carlo algorithms, on the other hand, consider the integral form of the equation – this approach is applied to solve the WBE in this chapter.

The Monte Carlo approach is by now very well established in the field of semiconductor transport simulations; the *de facto* standard reference literature on the topic [108, 109, 110] can be consulted for an in-depth treatment of the topic. Gradually Monte Carlo algorithms have moved beyond a direct emulation of particle behaviour and issues of statistical enhancement were approached by adding a statistical weight to particles [111], or using backward-in-time trajectories [112]. These techniques were quickly recognised to be special cases of solving the BTE expressed as an integral equation and applying numerical Monte Carlo integration techniques [113]. This generalized approach to devising Monte Carlo algorithms has been used to develop the novel signed-particle method to solve the semi-discrete Wigner-Boltzmann equation and will be discussed in the remainder of this chapter.

3.2 Outline

In the following an evolution problem posed by an initial condition is considered, where the boundaries of the simulation domain are assumed to be far enough away to have no influence on the evolution. If a bounded domain is considered, the initial condition must be replaced by a boundary condition when a particle reaches the boundary (refer to [114] for a rigorous treatment). Injecting boundary conditions, which are used in stationary problems, are not considered here, but only boundaries of a computational nature, i.e. the finite range of the phase space. The case for stationary problems is analysed in [63].

First the WBE is written as a Fredholm integral equation. The adjoint equation of the latter is then used to express the mean value of an arbitrary physical quantity as a Neumann series. The value of this series represents the solution to the computational problem and can be determined by a stochastic sampling. This sampling procedure is performed using numerical particles, which follow Newton trajectories and are scattered between different states as determined by the Wigner potential and other scattering mechanisms.

3.3 Integral formulation

The WBE, introduced in Chapter 2, is written here as

$$\left(\frac{\partial}{\partial t} + v_g\left(\mathbf{k}\right)\right) f_w\left(\mathbf{r},\mathbf{k},t\right) = \int d\mathbf{k}' \left\{ S\left(\mathbf{r},\mathbf{k},\mathbf{k}'\right) + V_w\left(\mathbf{r},\mathbf{k}'-\mathbf{k}\right) \right\} f_w\left(\mathbf{r},\mathbf{k}',t\right) +$$
$$-\lambda\left(\mathbf{r},\mathbf{k}\right) f_w\left(\mathbf{r},\mathbf{k},t\right); \tag{3.1}$$

$$\lambda\left(\mathbf{r},\mathbf{k}\right) \equiv \int d\mathbf{k}' S\left(\mathbf{r},\mathbf{k}',\mathbf{k}\right). \tag{3.2}$$

The first term on the RHS of (3.1) denotes the in-scattering and the second term denotes the out-scattering (at a rate λ). The Wigner potential will later emerge to be interpretable as a scattering mechanism. Although the scattering rates and the Wigner potential are assumed to be time-independent, introducing a time-dependence does not present a conceptual problem.

The Wigner potential can be decomposed as

$$V_w\left(\mathbf{r},\mathbf{k}\right) = V_w^+\left(\mathbf{r},\mathbf{k}\right) - V_w^-\left(\mathbf{r},\mathbf{k}\right);$$
$$V_w^+\left(\mathbf{r},\mathbf{k}\right) \equiv \max\left(0,\,V_w\right); \tag{3.3}$$
$$V_w^-\left(\mathbf{r},\mathbf{k}\right) \equiv \max\left(0,\,-V_w\right).$$

The quantity, which will emerge to be the scattering rate associated with the Wigner potential, is defined as

$$\gamma\left(\mathbf{r}\right) \equiv \int d\mathbf{k} V_w^+\left(\mathbf{r},\mathbf{k}\right) \geq 0. \tag{3.4}$$

The term $\gamma\left(\mathbf{r}\right) f_w\left(\mathbf{r},\mathbf{k},t\right)$ is added to both sides of (3.1), such that

$$\left(\frac{\partial}{\partial t} + v_g\left(\mathbf{k}\right) + \lambda\left(\mathbf{r},\mathbf{k}\right) + \gamma\left(\mathbf{r}\right)\right) f_w\left(\mathbf{r},\mathbf{k},t\right) =$$
$$\int d\mathbf{k}' \left\{ S\left(\mathbf{r},\mathbf{k},\mathbf{k}'\right) + V_w^+\left(\mathbf{r},\mathbf{k}'-\mathbf{k}\right) - V_w^-\left(\mathbf{r},\mathbf{k}'-\mathbf{k}\right) \right\} f_w\left(\mathbf{r},\mathbf{k}',t\right) + \gamma\left(\mathbf{r}\right) f_w\left(\mathbf{r},\mathbf{k},t\right). \tag{3.5}$$

This is expressed in the more compact form

$$\left(\frac{\partial}{\partial t} + v_g\left(\mathbf{k}\right) + \mu\left(\mathbf{r},\mathbf{k}\right)\right) f_w\left(\mathbf{r},\mathbf{k},t\right) = \int d\mathbf{k}' \Gamma\left(\mathbf{r},\mathbf{k},\mathbf{k}',t\right) f_w\left(\mathbf{r},\mathbf{k}',t\right); \tag{3.6}$$

$$\mu\left(\mathbf{r},\mathbf{k}\right) \equiv \lambda\left(\mathbf{r},\mathbf{k}\right) + \gamma\left(\mathbf{r}\right); \tag{3.7}$$

$$\Gamma\left(\mathbf{r},\mathbf{k},\mathbf{k}',t\right) \equiv S\left(\mathbf{r},\mathbf{k},\mathbf{k}'\right) + V_w^+\left(\mathbf{r},\mathbf{k}'-\mathbf{k}\right) +$$
$$-V_w^-\left(\mathbf{r},\mathbf{k}'-\mathbf{k}\right) + \gamma\left(\mathbf{r}\right)\delta\left(\mathbf{k}-\mathbf{k}'\right). \tag{3.8}$$

The partial differential equation represented in (3.6) can be transformed into an ordinary differential equation by introducing the characteristics of the Wigner function, which correspond to Newton trajectories in the physical sense. The trajectory of position (assuming parabolic bands and no electro-magnetic fields) can be parametrized with τ as

$$\mathbf{R}(\tau; \mathbf{r}_n, \mathbf{k}_n, t_n) = \mathbf{R}_n(\tau) = \mathbf{r}_n + \int_{t_n}^{\tau} v_g(\mathbf{k}_n)\, d\ell = \mathbf{r}_n + \frac{\hbar \mathbf{k}_n}{m^*}(\tau - t_n);$$

$$\mathbf{R}_n(\tau = t_n) = \mathbf{r}_n. \tag{3.9}$$

The trajectory $\mathbf{R}_n$ is initialized by the values $(\mathbf{r}_n, \mathbf{k}_n, t_n)$ (which represent values in this context and not variables). Depending on the choice of the initialization time t_n, the trajectory is either forward in time $(\tau > t_n)$ or backward in time $(\tau < t_n)$. The trajectory of the wavevector is constant, since the LHS of (3.1) does not have a force term that accelerates the particle and is formally introduced as

$$\mathbf{K}(\tau; \mathbf{r}_n, \mathbf{k}_n, t_n) = \mathbf{K}_n(\tau) = \mathbf{k}_n \,\forall \tau. \tag{3.10}$$

In the following the initialization parameters are only explicitly stated where it aides clarity.

The LHS of (3.6) can represent the full derivative of f_w (with respect to parameter τ). To achieve this, the following property, for an arbitrary function g, is used:

$$\frac{d}{dx}e^{\int_b^x g(y)dy} = \frac{d}{dx}e^{-\int_x^b g(y)dy} = g(x)$$

such that

$$\frac{d}{dx}\left(f(x)e^{\int_b^x g(y)dy}\right) = \frac{df}{dx}e^{\int_b^x g(y)dy} + g(x)e^{\int_b^x g(y)dy}f(x) = \left\{\frac{\partial}{\partial x} + g(x)\right\}f(x)\,e^{\int_b^x g(y)dy}.$$

The value of b (integration limit) can be chosen freely and is chosen as t_0 in the following.

Both sides of (3.6) are multiplied by the term $\exp\left(-\int_\tau^{t_0} \mu(\mathbf{R}(y),\mathbf{k})\,dy\right)$ to yield

$$\frac{d}{d\tau}\left(f_w(\mathbf{R}(\tau;\mathbf{r},\mathbf{k},t_0),\mathbf{k},\tau)\,e^{-\int_\tau^{t_0}\mu(\mathbf{R}(y),\mathbf{k})dy}\right) =$$

$$\int d\mathbf{k}'\Gamma(\mathbf{R}(\tau),\mathbf{k},\mathbf{k}',\tau)\,f_w(\mathbf{R}(\tau),\mathbf{k}',\tau)\,e^{-\int_\tau^{t_0}\mu(\mathbf{R}(y),\mathbf{k})dy}. \tag{3.11}$$

This equation can be formally integrated on the interval $\tau \in (t, t_0)$ to finally yield the integral form of (3.6):

$$f_w(\mathbf{r},\mathbf{k},t_0) = f_{w,i}(\mathbf{r},\mathbf{k})\,e^{-\int_t^{t_0}\mu(\mathbf{R}(y),\mathbf{k})dy} +$$

$$+ \int_t^{t_0} dt'\left\{\int d\mathbf{k}'\Gamma(\mathbf{R}(\tau),\mathbf{k},\mathbf{k}',\tau)\,f_w(\mathbf{R}(t'),\mathbf{k}',t')\,e^{-\int_{t'}^{t_0}\mu(\mathbf{R}(y),\mathbf{k})dy}\right\}. \tag{3.12}$$

This equation evaluates the Wigner function at the phase space point $(\mathbf{r}, \mathbf{k})$ at time t_0 using the initial condition $f_{w,i}(\mathbf{r},\mathbf{k}) = f_w(\mathbf{R}(t;\mathbf{r},\mathbf{k},t_0),\mathbf{k},t)$, which is assumed to be known at time t.

The integral equation (3.12) has a form analogous to the Chamber's path integral [115], but with the contributions of the Wigner potential added. The introduction of the exponential term in (3.11) actually relates to an analytical summation of all out-scattering – the term μ could, alternatively, also be retained on the RHS of (3.1) and integrated otherwise, thereby suppressing the exponential term. However, the form of (3.12) is preferable as it gives a clear physical interpretation, which is useful when devising the Monte Carlo algorithm (Section 3.6).

The exponential term $e^{-\int_t^{t_0}\mu(\mathbf{R}(y),\mathbf{k})dy}$ gives the probability of a particle to remain on its trajectory, i.e. not be scattered, from time t until time t_0. Therefore, the first term of (3.12) gives the contribution of particles initialized at $(\mathbf{R}(t;\mathbf{r},\mathbf{k},t_0),\mathbf{k})$ to reach point $(\mathbf{r},\mathbf{k})$ without being scattered, whereas the second term gives the contribution of all particles scattered into the appropriate trajectory (according to in-scattering rate Γ) at time t' to remain on the trajectory to reach point $(\mathbf{r},\mathbf{k})$ at time t_o.

3.3.1   Fredholm integral form

Fredholm integral equations of the second kind[1] have the form:

$$f(Q) = f_i(Q) + \int dQ_1\, K(Q, Q_1) f(Q_1),\tag{3.13}$$

where the free term f_i and the kernel K are known and describe the initial/boundary conditions and the propagation of particles, respectively; Q is a multi-variable representing e.g. $(\mathbf{r}, \mathbf{k}, t)$.

To express (3.12) in the form of (3.13), the integral must be augmented by the variable[2] $\mathbf{r}_1$ to complete $Q_1 = (\mathbf{r}_1, \mathbf{k}_1, t_1)$:

$$
\begin{aligned}
f_w(\mathbf{r}_0, \mathbf{k}_0, t_0) &= f_i(\mathbf{r}_0, \mathbf{k}_0, t_0) + \\
&\quad + \int_{-\infty}^{t_0} dt_1 \int d\mathbf{r}_1 \int d\mathbf{k}_1 K\left(\mathbf{R}_0(t_1), \mathbf{k}_0, t_0, \mathbf{r}_1, \mathbf{k}_1, t_1\right) f_w(\mathbf{r}_1, \mathbf{k}_1, t_1) \quad (3.14)
\end{aligned}
$$

$$f_i(\mathbf{r}_0, \mathbf{k}_0, t_0) \equiv f(\mathbf{R}_0(t_1), \mathbf{k}_0, t_1) e^{-\int_{t_1}^{t_0} \mu(\mathbf{R}_0(y), \mathbf{k}) dy}\tag{3.15}$$

$$K(\mathbf{r}_0, \mathbf{k}_0, t_0, \mathbf{r}_1, \mathbf{k}_1, t_1) \equiv \Gamma(\mathbf{r}_1, \mathbf{k}_0, \mathbf{k}_1, t_1)\, \delta(\mathbf{r}_1 - \mathbf{R}_0(t_1))\, e^{-\int_{t_1}^{t_0} \mu(\mathbf{R}_0(y), \mathbf{k}) dy} \theta(t_0 - t_1).\tag{3.16}$$

This is a linear Fredholm integral equation of the second kind, which has a unique solution (if one exists) [116]. The step function θ is added to the kernel to retain the limits of the time integral. Similarly, an indicator function can be added to limit the spatial domain, but is omitted here for brevity.

A very wide variety of physical phenomena can be described by integral equations of the form (3.13) and a strong theory has evolved surrounding the solution of such Fredholm integral equations using Monte Carlo algorithms [117].

It is possible to formulate an integral representation of the WBE for the entire (global) domain. Sometimes, however, the operator of the equation under consideration is too complicated to be able to formulate a global integral representation. In such a case, local integral representations are used based on the Green's function[3] Monte Carlo algorithm [105]. The theory of Green's function Monte Carlo algorithms keeps on developing [118, 119, 120] and is often applied to solve non-linear physical problems, e.g. [121].

3.3.2   Adjoint equation

The adjoint form of integral equations, of the form as in (3.13), is often easier to solve [106, 122].

The integral form of the WBE in (3.12) and (3.14) describes a backward-in-time equation as seen from the limits of the time integral. To obtain the corresponding forward-in-time equation the adjoint equation is required [123].

The kernel of (3.13) can be interpreted as a propagator: $K(Q, Q_1)$ describes the propagation from Q_1 to Q. The adjoint equation to (3.13) solves for the function g (to which no particular physical meaning is attached at this stage) and uses the (self-)adjoint kernel $K^\dagger(Q_1, Q) = K(Q, Q_1)$, which exchanges the position of the integration variable:

$$g(Q_1) = g_i(Q_1) + \int dQ\, K(Q, Q_1) g(Q),\tag{3.17}$$

where g_i is the free term.

Using the newly introduced adjoint equation it can be shown that

$$\int dQ f_i(Q)\, g(Q) = \int dQ_1 g_i(Q_1)\, f(Q_1).\tag{3.18}$$

---

[1]A Fredholm equation of the second kind is identified by the function f_w appearing both inside and (as an initial value) outside of the integral.

[2]The property used is: $f(x) = \int dx' f(x') \delta(x - x')$.

[3]The Green's function should be interpreted here purely in the mathematical sense and does not imply the use of the quantum mechanics formalism based on the non-equilibrium Green's function (NEGF).

This result is obtained, if (3.13) is multiplied by $g(Q)$ and integrated by $\int dQ$ and (3.17) is multiplied by $f(Q_1)$ and integrated by $\int dQ_1$; the two resulting equations are subtracted. It follows

$$
\begin{aligned}
\int dQ\, f_i(Q)\, g(Q) &= \int dQ_1\, g_i(Q_1)\, f(Q_1) \text{ iff} \\
\int dQ_1\, f_i(Q_1)\, g(Q_1) &= \int dQ\, g_i(Q)\, f(Q).
\end{aligned}
\tag{3.19}
$$

This formulation will emerge to be very convenient for solving the computational task, described in Section 3.6.

3.4 Neumann series

To develop a Neumann series of a general Fredholm equation, the function appearing in the integrand is expanded recursively. The first expansion yields

$$
\begin{aligned}
f(Q_0) &= f_i(Q_0) + \int dQ_1\, K(Q_0, Q_1) \left\{ f_i(Q_1) + \int dQ_2\, K(Q_1, Q_2) f(Q_2) \right\} \\
&= \underbrace{f_i(Q_0)}_{f^0} + \underbrace{\int dQ_1\, K(Q_0, Q_1) f_i(Q_1)}_{f^1} + \underbrace{\int dQ_1 \int dQ_2\, K(Q_0, Q_1) K(Q_1, Q_2) f(Q_2)}_{f^2 + f^3 + \cdots + f^n}.
\end{aligned}
\tag{3.20}
$$

The Neumann series is formed by an iterative application of the kernel K to the free term f_i. The term f^n corresponds to n applications of the kernel:

$$
f = \sum_{n=0}^{\infty} f^n;
\tag{3.21}
$$

$$
f^n(Q_n) = \begin{cases} \int dQ_n\, K(Q_{n-1}, Q_n) f^{n-1}(Q_n) & \text{if } n > 0 \\ f_i(Q_0) & \text{if } n = 0. \end{cases}
\tag{3.22}
$$

The evaluation of a Neumann series formally solves the Fredholm integral equation [124].

3.5 Monte Carlo integration

The terms comprising a Neumann series contain high-dimensional integrals which can be efficiently calculated using stochastic sampling by Monte Carlo techniques.

Consider an integral

$$
I \equiv \int_a^b dx\, \phi(x) = \int dx\, p(x)\, \psi(x),
\tag{3.23}
$$

where p is a probability density function such that $\int_a^b dx\, p(x) = 1$ and $\psi(x) = \frac{\phi(x)}{p(x)}$. The integral I corresponds to the mean value of $\psi(x)$.

The choice of the distribution p determines various qualities of the Monte Carlo algorithm [106], namely the computational efficiency, the convergence rate and the associated trade-off with reliability (variance in the result). Often physical considerations are used to choose the distribution p.

Consider the random variables X and $\Psi[X]$: A sequence of N numbers $\{x_i\}$ is generated according to p and is used to sample $\Psi[X]$, thereby approximating the mean value by an expected value:

$$
I \approx E[\Psi[X]] = \frac{1}{N} \sum_{i=1}^{N} \psi(x_i).
\tag{3.24}
$$

This establishes the link between the Neumann series and the Monte Carlo algorithm, which will be discussed in Section 3.7.

3.6   Computational task

The computational task at hand is to calculate the statistical mean of an arbitrary physical quantity, represented by $A(\mathbf{r}, \mathbf{k})$, at time T using the Wigner function:

$$\langle A_T \rangle = \int d\mathbf{r} \int d\mathbf{k} f_w(\mathbf{r}, \mathbf{k}, T) A(\mathbf{r}, \mathbf{k}). \tag{3.25}$$

The mean of a physical quantity is often of greater interest than the Wigner function itself (one can solve for f_w by setting $A(\mathbf{r}, \mathbf{k}) = \delta(\mathbf{r})$).

It is formally proven in [117] that a linear functional of the form (3.25) can be solved by calculating the statistical expectation value of a random variable, which is calculated using the procedure discussed in Section 3.5.

The RHS of (3.18) represents an inner product between g_i and f, which is reminiscent of the inner product (3.25). Therefore, either f or g can be solved to obtain the mean value of a physical quantity. In the latter case the free term g_i of the adjoint equation (3.17) is chosen to correspond to the physical quantity A, augmented with the time variable, such that:

$$g_i(Q) = A(\mathbf{r}, \mathbf{k}) \delta(t - T) = A_T(Q). \tag{3.26}$$

The statistical mean of A_T can be expressed as:

$$\langle A_T \rangle = \int dQ\, g(Q) f_i(Q), \tag{3.27}$$

which can be developed into a Neumann series of the form

$$\langle A_T \rangle = \sum_{n=0}^{\infty} \langle A_T \rangle_n. \tag{3.28}$$

The first expansion of g as a Fredholm equation, with the free term (3.26) yields:

$$\int dQ_1 f_i(Q_1) g(Q_1) = \int dQ_1\, f_i(Q_1) A_T(Q_1) + \int dQ_1 f_i(Q_1) \int dQ\, K(Q, Q_1) A_T(Q) +$$
$$+ \int dQ_1 f_i(Q_1) \int dQ \int dQ_2\, K(Q, Q_1) K(Q_2, Q) g(Q_2). \tag{3.29}$$

The first two terms of (3.28) are written here to convey the principle. The zero-th term simply contains the free term (3.26):

$$\begin{aligned}
\langle A_T \rangle_0 &= \int dQ\, f_i(Q) A_T(Q) \\
&= \int_{t_0}^{\infty} dt_1 \int d\mathbf{k}_1 \int d\mathbf{r}_1 f_i(\mathbf{r}_1, \mathbf{k}_1) e^{-\int_{t_0}^{t_1} \mu(\mathbf{R}_1(y), \mathbf{k}_1) dy} A(\mathbf{R}_1(t_1), \mathbf{k}_1) \delta(t_1 - T) \\
&= \int d\mathbf{k}_1 \int d\mathbf{r}_1 f_i(\mathbf{r}_1, \mathbf{k}_1) e^{-\int_{t_0}^{T} \mu(\mathbf{R}_1(y), \mathbf{k}_1) dy} A(\mathbf{R}_1(T), \mathbf{k}_1).
\end{aligned} \tag{3.30}$$

The meaning of the term is as follows: a particle initialized at $(\mathbf{r}_1, \mathbf{k}_1, t_0)$ follows a trajectory $\mathbf{R}_1$ and reaches the point $(\mathbf{R}_1(T), \mathbf{k}_1, T)$ with a probability of $e^{-\int_{t_0}^{T} \mu(\mathbf{R}_1(y), \mathbf{k}_1) dy}$ and makes a contribution of $A(\mathbf{R}_1(T), \mathbf{k}_1)$ to the statistical mean of A. If the particle is scattered from its trajectory, it contributes through another term in the series.

$$\begin{aligned}
\langle A_T \rangle_1 &= \int dQ_1 f_i(Q_1) \int dQ\, K(Q, Q_1) A_T(Q) \\
&= \underbrace{\int_{t_0}^{\infty} dt_1 \int d\mathbf{k}_1 \int d\mathbf{r}_1 f_i(\mathbf{r}_1, \mathbf{k}_1) \left\{ \mu(\mathbf{R}_1(t_1), \mathbf{k}_1) e^{-\int_{t_0}^{t_1} \mu(\mathbf{R}_1(y), \mathbf{k}_1) dy} \right\}}_{I} \\
&\quad \times \underbrace{\theta_D(\mathbf{r}_1) \int_{t_1}^{\infty} dt \int d\mathbf{k} \left\{ \frac{\Gamma(\mathbf{r}_1, \mathbf{k}, \mathbf{k}_1)}{\mu(\mathbf{R}_1(t_1), \mathbf{k}_1)} \right\} e^{-\int_{t_1}^{t} \mu(\mathbf{R}_1(y), \mathbf{k}) dy} A(\mathbf{R}_1(t), \mathbf{k}) \delta(t - T)}_{II}. \tag{3.31}
\end{aligned}$$

As before, a particle initialized at $(\mathbf{r}_1, \mathbf{k}_1, t_0)$ follows a trajectory $\mathbf{R}_1$, but only until time t_1 where it is scattered. By extracting μ as a common factor, the term in the first set of braces can be interpreted as the probability for a particle to not be scattered until time t_1 (the exponential) and then to be scattered in the interval dt_1 thereafter ($\mu(\mathbf{R}_1(t_1), \mathbf{k}_1) dt_1$). The term in the braces of term II describes the probability of scattering from $\mathbf{k}_1$ to $\mathbf{k}$ and the exponential term again gives the probability that the scattered particle will not be scattered again until time T, where it contributes $A(\mathbf{R}_1(T), \mathbf{k})$ to the statistical mean of A. If the particle is scattered again before time T, it contributes through another term in the series.

Further terms of the series can be written down in a similar manner to reveal the general structure: A particle that scatters n times before time T makes a contribution to $\langle A_T \rangle$ through the term $\langle A_T \rangle_n$ in the series. The braced terms in the integrals of I and II represent probabilities, which clearly establishes the link to the Monte Carlo integration introduced in Section 3.5.

3.6.1 Wigner potential as a scattering mechanism

The term $\frac{\Gamma(\mathbf{r}_1, \mathbf{k}, \mathbf{k}_1)}{\mu(\mathbf{R}_1(t_1), \mathbf{k}_1)}$ appearing in term II represents the total scattering probability, including both phonons and the 'scattering' (particle generation) associated to the Wigner potential. Phonon scattering occurs with a probability $\frac{\lambda}{\mu}$ and is described in Section 3.7.5; the Wigner-related term is selected with a probability $\frac{\gamma}{\mu} = \left(1 - \frac{\lambda}{\mu}\right)$ and can then be written as

$$3 \left\{ \frac{1}{3} \frac{V_w^+(\mathbf{r}, \mathbf{k}' - \mathbf{k})}{\gamma(\mathbf{r})} - \frac{1}{3} \frac{V_w^-(\mathbf{r}, \mathbf{k}' - \mathbf{k})}{\gamma(\mathbf{r})} + \frac{1}{3}\delta(\mathbf{k} - \mathbf{k}') \right\}. \tag{3.32}$$

Each term is normalized by gamma to represent a probability, which follows from the definition

$$\gamma(\mathbf{r}) \equiv \int d\mathbf{k} V_w^+(\mathbf{r}, \mathbf{k}) \geq 0.$$

Two interpretations of (3.32) are possible: If considered as a scattering mechanism, one of the terms is selected, each with a probability $\frac{1}{3}$ and the particle is generated with a weight of ± 3 at wavevector $\pm \mathbf{k}'$. Alternatively, all three terms can be chosen simultaneously and take a weight of ± 1, i.e. two additional particles are created with wavevector $\pm \mathbf{k}'$ and weight ± 1; the original particle, associated with the δ, persists unchanged. The algorithm to select the wavevectors of the generated particles is discussed in Section 3.7 and revisited in Chapter 4, which considers the semi-discrete form of the WBE.

3.7 Algorithm

The preceding sections introduced the Neumann series with which the statistical mean of an arbitrary physical quantity can be represented. The value of the latter can be obtained by stochastic sampling of the Neumann series using numerical particles (each particle represents a sampling). A particle is evolved through free-flight and scattering up to time T, which selects one of the terms in the series. The contribution of the selected term is determined by sampling the integral associated with that term. The mean value of all the sampled values approximates the value of the Neumann series.

An algorithm to perform this task uses the physical interpretation of the various terms to propagate numerical particles along trajectories and scatter them to different wavevectors or spawn additional particles with different wavevectors. This physical interpretation corresponds to the notion of free-flight and scattering used in semi-classical Monte Carlo simulation and allows the adoption of many established algorithms.

The conditions for obtaining a unique solution to (3.25) using this Monte Carlo algorithm and the associated error analysis are formally discussed in [122].

The implementation of this algorithm is discussed in the following subsections, starting with an outline of the basic structure of the algorithm. Thereafter, the basic aspects of the important steps are discussed. A detailed discussion of certain algorithmic aspects is postponed to Chapter 4.

3.7.1 Structure

The basic structure of the algorithm to solve the WBE is illustrated in Figure 3.1 and highlights the aspects differentiating the Wigner ensemble Monte Carlo simulation from semi-classical equivalents.

The simulation commences with an initialization step, which receives inputs describing the geometry, potential profile and parameters (e.g. time step, simulation time) used for the simulation and computing the (stationary) Wigner potential. An ensemble of particles representing the initial condition of the evolution problem is initialized. Thereafter, the time loop commences which tracks the trajectories of the particles over time. The position and wavevector of each particle in the ensemble is recorded on a histogram, taking the particle signs into account, at regular time intervals Δt (the observation time), which approximates the distribution function $f_w(\mathbf{r}, \mathbf{k}, t)$.

The time loop consists of the evolution and annihilation modules, which are executed alternately until the total simulation time is reached. The evolution module entails the drift (free-flight) and scattering/generation of particles. Particles are drifted classically according to their momentum value. If a particle experiences a generation event, two new particles are generated and added to the particle ensemble. If a scattering event occurs, the wavevector of the particle is modified according to the selected Boltzmann scattering mechanism. The processes of drift and generation/scattering are repeated in an iterative fashion for all particles in the (growing) ensemble until the end of the time-step is reached.

The annihilation procedure is only performed when needed, i.e. if the size of the particle ensemble will exceed the set maximum in the next time step. This is preferable to performing it at every time step since the annihilation introduces approximations, which can have undesirable effects (discussed in Section 4.3.1). After the annihilation, the remaining particles are regenerated and the time loop continues until the final time is reached.

3.7.2 Discretization

The integral form for the semi-discrete version of the Wigner equation follows directly from (3.12) by omitting all terms related to phonon scattering and substituting the integral $\int d\mathbf{k}$ with the summation $\sum_{\mathbf{q}}$ and the wavevector with its discretized form $\mathbf{k} \to \mathbf{q}\Delta k$:

$$f_w(\mathbf{r}, \mathbf{q}, t) = e^{-\int_0^t \gamma(\mathbf{r}(y))dy} f_i(\mathbf{r}(0), \mathbf{q}) + \int_0^t dt' \sum_{q'} f_w(\mathbf{r}(t'), \mathbf{q}', t') \Gamma(\mathbf{r}(t'), \mathbf{q}, \mathbf{q}') e^{-\int_{t'}^t \gamma(\mathbf{r}(y))dy}. \quad (3.33)$$

Real-valued wavevectors can be retained for the Boltzmann scattering terms, given certain considerations discussed in Appendix B.

The calculation of the Wigner potential requires the discretization of the position coordinate, which will be stipulated in Section 4.1.1.

3.7.3 Particle initialization

The initial condition from which the evolution starts is represented by an ensemble of N particles. Each particle in the ensemble carries the following attributes:

- position (vector of positive real numbers);

- momentum (vector of integer multipliers of Δk);

- space cell/node closest to particle (vector of positive integers);

- particle sign ($a = \pm 1$) (integer);

- free-flight time (τ) (positive real number);

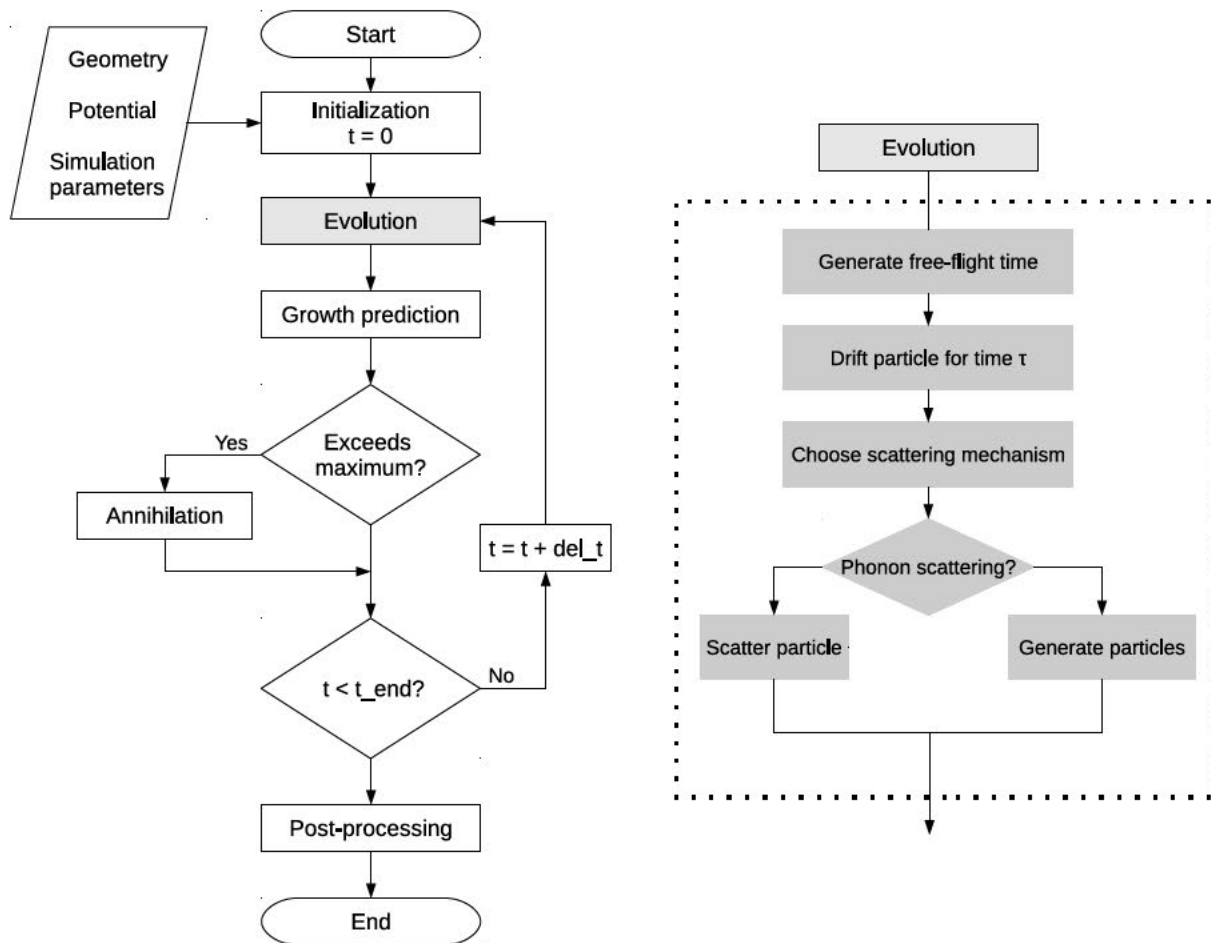- time remaining in time step(δt) (positive real number).

Figure 3.1: Flowchart of a Monte Carlo algorithm to solve the Wigner-Boltzmann equation using the signed-particle method. Particle drift and scattering and/or generation occur alternately until the end of the simulation is completed.

The position and momentum of the particles is initialized according to a chosen distribution or a previously calculated initial condition. The particle signs should only take negative values if the Wigner function, representing the initial condition, is calculated from a physically valid wavefunction.

The Gaussian minimum uncertainty wavepacket is often used as an initial distribution and is defined by

$$f_w\left(\mathbf{r},\mathbf{q}\right) = \mathcal{N}e^{-\frac{(\mathbf{r}-\mathbf{r_0})^2}{\sigma^2}}e^{-(\mathbf{q}\Delta k-\mathbf{k_0})^2\sigma^2}, \tag{3.34}$$

where $\mathbf{r}_0$ and $\mathbf{k}_0$ represent the mean position and the mean wavevector, respectively; σ is the standard spatial deviation and $\mathcal{N}$ represents a normalization constant. Each wavepacket, consisting of many numerical particles, represents a single electron and captures both the particle- and wave-like properties of the electron [125].

3.7.4 Free-flight

3.7.4.1 Drift

Particle drift or free-flight refers to the movement of a carrier, according to Newtonian laws, between scattering events. The electromagnetic forces do not appear explicitly in the Wigner(-Boltzmann) equation, therefore a particle experiences no acceleration due to forces during free-flight and its wavevector remains constant. The position of a particle changes according to

$$\delta\mathbf{r} = \frac{\hbar(\mathbf{q}\Delta k)}{m^*}t^*, \tag{3.35}$$

where t^* is given by

$$t^* = \min(\tau, \delta t). \tag{3.36}$$

Here, δt represents the time remaining in the time step Δt. Therefore, the particle drifts up to the end of the current time step or the next scattering event – whichever comes first. The value of δt is updated continuously and the drift and scattering processes alternately repeat until δt is zero, i.e. the end of the time interval Δt is reached.

3.7.4.2 Free-flight time

A particle that has completed a scattering event at time t_0 will not be scattered again until time τ, i.e. it will undergo free-flight between t_0 and τ, with a probability given by

$$P(\tau; \mathbf{r}_0, \mathbf{k}_0) = e^{-\int_{t_0}^{\tau} \mu(\mathbf{R}_0(y), \mathbf{k}_0)\mathrm{d}y}. \tag{3.37}$$

The total scattering rate μ, given in (3.7), is a function of time, since the scattering rate changes with position due to the spatial dependence of the Wigner potential and should be integrated along the trajectory $\mathbf{R}$ initialized by $(\mathbf{r}_0, \mathbf{k}_0)$. Since the particle is not accelerated, the wavevector $\mathbf{k}$ is only modified by scattering and remains constant during the free-flight.

Using the concept of self-scattering [126] a value is added to the scattering rate such that it is kept constant (θ) over space and time. Thereby, the need to calculate the integral in the exponent is avoided and the probability (P) is given by

$$r = e^{-\theta\tau} \tag{3.38}$$

where $r \in [0, 1]$ is a uniformly distributed random number. This considerably simplifies the determination of the duration of free-flight:

$$\tau = -\frac{\ln r}{\theta}. \tag{3.39}$$

3.7.5 Scattering

3.7.5.1 Scattering rates

The total scattering rate Γ is calculated once (for a given potential), according to the generation rate (3.4) and the scattering rates given in Appendix A, during the initialization of the simulator and is retained in a look-up table. The scattering mechanism to be used for a scattering event is randomly selected from the normalized scattering Table [109] by generating a uniformly distributed random number $r \in [0, 1]$. The selected mechanism can either be one of the phonon scattering mechanisms or a particle generation event.

3.7.5.2 Particle generation

A generation event entails the creation of two additional particles with complementary signs and offsets ℓ_1 and ℓ_2, with respect to the wavevector $\mathbf{k}$ of the generating particle; one particle will have a wavevector $\mathbf{k}+\ell_1$ and the other particle, with the complementary sign (affinity), will have a wavevector $\mathbf{k}+\ell_2$. The Wigner potential (V_w; defined in Chapter 2) defines the rate at which particles are generated and the probability distributions used to select momentum offsets.

The two momentum offsets, ℓ_1 and ℓ_2, are determined by sampling the quantities $\frac{V_w^+(\mathbf{r},\ell_1)}{\gamma(\mathbf{r})}$ and $\frac{V_w^-(\mathbf{r},\ell_2)}{\gamma(\mathbf{r})}$, which represent probabilities, as discussed in Section 3.6.1. Due to the antisymmetry of the Wigner potential[4]:

$$V_w^-(\cdot, \ell) = V_w^+(\cdot, -\ell). \tag{3.40}$$

both distributions can be sampled by a single random number generation, which yields a symmetric momentum offset $\pm\ell$. A uniformly distributed random number $r \in [0, 1]$ is generated to determine the offset ℓ such that

$$r = \int_0^\ell dy \frac{V_w^+(\mathbf{r}, y)}{\gamma(\mathbf{r})} \tag{3.41}$$

holds true.

The single sampling is attractive from a computational point of view and also a valid approach, if a sufficiently large number of particles is considered in the simulation (law of large numbers). The finite range of the wavevectors considered, however, requires a closer analysis, which is deferred to Section 4.2.

The process of particle generation leads to an exponential increase in the number of particles. The increase in particles within a time step is described by

$$N_{t_{n+1}} = N_{t_n} e^{2\gamma(t_{n+1}-t_n)} \tag{3.42}$$

where $N_{t_{n+1}}$ and N_{t_n} represent the number of particles at times t_{n+1} and t_n, respectively, and γ is the generation rate given by (3.4). The generation rate is in the order of $10^{15}\,\mathrm{s}^{-1}$ for potential profiles where the largest potential differences are in the order of $100\,\mathrm{meV}$. Therefore, the particle number increases so rapidly that simulations beyond a few femtoseconds become computationally infeasible.

This numerically debilitating increase in the number of particles is counteracted by the notion of particle annihilation, which keeps the number of particles under control, as discussed in the next section.

3.7.5.3 Phonon scattering

The treatment of the conventional phonon scattering mechanisms follows the standard references for Monte Carlo simulation [109, 110]. The equations for the considered scattering mechanisms (specified in Appendix A) theoretically are only valid in bulk semiconductors with a continuous spectrum of energy and wavevectors. The use of the phonon scattering models with a discretized k-grid necessitates some rounding of values, which can introduce an increase in energy. However, experiments have established that the discretization errors accumulate to an insignificant amount over simulation times of interest (several picoseconds). Appendix B can be consulted for details.

---

[4]The anti-symmetry of the Wigner potential implies $V_W(\mathbf{r}, \mathbf{k}) = -V_W(\mathbf{r}, -\mathbf{k})$
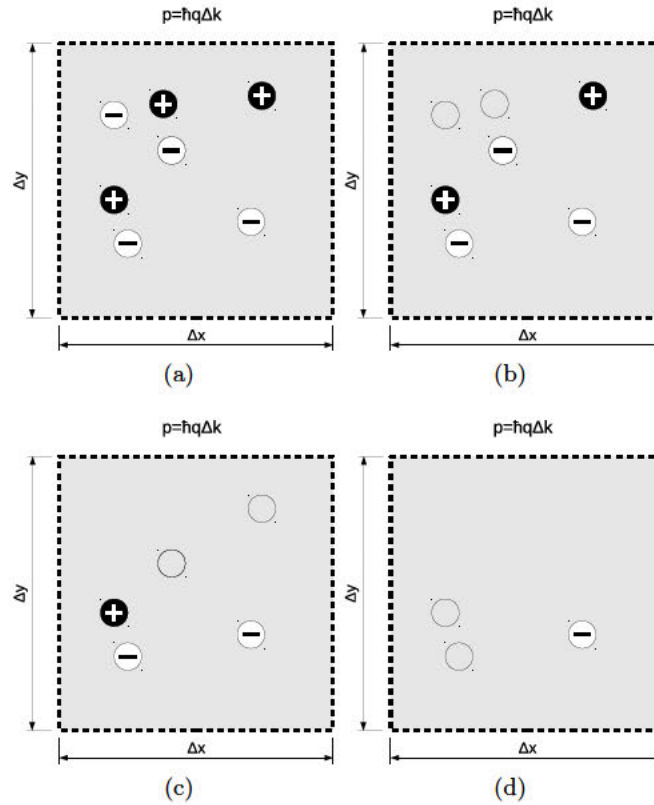
Figure 3.2: Illustration of the pair-wise annihilation of particles in a cell of the phase space (a) - (d), representing a two-dimensional area $\Delta x \Delta y$ and a fixed momentum $\mathbf{p} = \hbar \mathbf{q} \Delta k$. All particles within a cell are considered indistinguishable and any two particles with complementary signs can annihilate each other.

3.7.6 Annihilation process

The annihilation concept entails dividing the phase space into many cells. Each cell represents a volume $(\Delta r)^d (\Delta k)^{d'}$ of the phase space, where d (d') is the dimension of space (wavevector). Since the wavevector is discretized in the semi-discrete Wigner equation a single value is associated with each cell.

   The particles within a given cell are considered to have the same probabilistic future: On average, two particles of opposite sign within a cell will end up in approximately the same place in the phase space and make a contribution to the Wigner function (or any physical quantity), which is equal in magnitude but opposite in sign. Since the particles' contributions cancel out in the future, they immediately annihilate each other and cease to exist. All the particles within a cell are deemed to be identical and indistinguishable; any positive particle can annihilate any negative particle (and vice-versa). By considering the particles to be identical and indistinguishable avoids the need to keep track of the attributes of each individual particle; instead, the attributes are associated with the cells of the phase space and only the number of particles within the cell is important.

   Using the concept of particle annihilation the number of particles and the associated computational burden, is considerably reduced. This is of great importance as it has made the simulations computationally feasible, also in multiple dimensions. The annihilation procedure is illustrated in Figure 3.2 for a two-dimensional cell with a fixed wavevector.

   Consider a cell with A particles with a positive sign and B particles with a negative sign, which are summed up to yield a remainder of particles, $R = A - B$; $|R|$ particles, each carrying the sign of R, are regenerated within the cell. The $|R|$ particles which survive the annihilation procedure should, ideally, recover the information represented by the $(A + B)$ particles before the annihilation took place. Since the wavevector is quantized and a single value is shared amongst all particles within a cell, the distribution in the k-space is recovered after annihilation. The positions of the particles, however, are

real-valued and require additional consideration, which will be given in Section 4.3.1. New values for the free-flight time are generated for each particle using the procedure discussed in Section 3.7.4. Due to the Markovian character of the evolution the values the particles had prior to annihilation need not be considered.

3.7.7 Implementation

The models and algorithms presented in the preceding sections (and elsewhere in this work) have been implemented in the WEMC module of the ViennaWD suite of particle-based simulation tools [102]. In addition to the WEMC simulator, ViennaWD includes two additional simulators: i) The Phonon Decoherence (PD) simulator, which uses the single particle Monte Carlo method to simulate the evolution of an entangled quantum state under the influence of phonon scattering in one dimension and ii) the Classical Ensemble Monte Carlo (CEMC) simulator, which allows the self-consistent solution of the BTE for two-dimensional MOSFET structures.

The WEMC simulator is written in C and uses the Message Passing Interface (MPI) for parallel execution in a distributed-memory computing environment (see Chapter 5). ViennaWD is an open source project, publicly available on SourceForge [102], and the development of the WEMC simulator presents a major contribution to enabling other researchers to use and investigate Wigner-based simulation for time-resolved quantum transport.

User interaction with the WEMC simulator is facilitated through input files, which are specified as command line arguments at execution. The simulation parameters, like time step and simulation domain, are specified using a text-based input file, implemented using scripts written in Lua [127]. The potential profile and/or initial condition can also be specified with text files, using a simple text format. Alternatively, the potential profile and the initial conditions can be generated directly in the simulator with analytical functions and distributions.

Post-processing of the output data is handled by Python scripts, which merge data files (if needed) and execute further scripts, which generate the requested graphical output. Plots of output data, e.g. density or k-distribution, are automatically generated (if selected) using gnuplot scripts.

The complete documentation on the use of the simulator has been made available through the ViennaWD user manual [102].

Chapter 4

Optimized Algorithms for the Signed-Particle Method

This chapter highlights the improvements and contributions that have been made to optimize the algorithms of the signed-particle method. The contributions encompass algorithms for increased computational efficiency, statistical enhancement and address certain discretization effects which can lead to erroneous behaviour in simulations. Finally, the accuracy of the signed-particle method with the optimized algorithms is demonstrated by a comparison of numerical results to the exact solution of a physical problem.

4.1 Wigner potential

The Wigner potential (WP) is of central importance in the signed-particle method as it dictates the particle generation statistics (see Section 3.7.5). This section summarizes the contributions made in this thesis related to the WP, aiming at an optimized computational implementation. First, the full discretization of the WP is shown, which introduces the computational task (Section 4.1.1). Thereafter, a highly efficient algorithm to calculate the two-dimensional (2D) WP is presented (Section 4.1.2), which is of importance in the pursuit of self-consistent solutions. Finally, the implications the choice of the coherence length has on physical and computational aspects is discussed (Section 4.1.3); a physically relevant example clearly illustrates how the encountered discretization effects can be mitigated.

4.1.1 Full discretization

The semi-discrete WP

$$V_W\left(\mathbf{r},\mathbf{q}\right) \equiv \frac{1}{i\hbar\mathbf{L}} \int_{-\mathbf{L}/2}^{\mathbf{L}/2} d\mathbf{s}\, e^{-i2\mathbf{q}\Delta\mathbf{k}\cdot\mathbf{s}}\delta V \tag{4.1}$$

$$\delta V\left(\mathbf{s};\mathbf{r}\right) \equiv V\left(\mathbf{r}+\mathbf{s}\right) - V\left(\mathbf{r}-\mathbf{s}\right), \tag{4.2}$$

considers $\mathbf{s}$ to be bounded by a finite coherence length, $\mathbf{L}$ (Section 2.3). The length and position vectors are discretized and are defined here as

$$\begin{aligned}
\mathbf{r} &\equiv (x\Delta r, y\Delta r) \\
\mathbf{s} &\equiv (m\Delta s, n\Delta s) \\
\mathbf{L} &\equiv (M\Delta s, N\Delta s) \\
\mathbf{q}\Delta\mathbf{k} &\equiv \left(p\frac{\pi}{M\Delta s}, q\frac{\pi}{N\Delta s}\right).
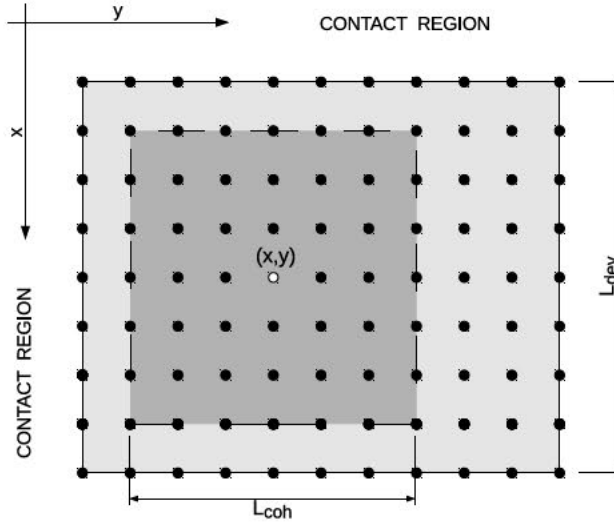\end{aligned} \tag{4.3}$$

Figure 4.1: Coherence box of size $\mathbf{L} = (M'\Delta x, N'\Delta y)$, centred at node (x, y) in the discretized domain, of size $\mathbf{L}_{dev} = (M\Delta x, N\Delta y)$, surrounded by semi-infinite contact regions.

This yields the 2D computational domain, as depicted in Figure 4.1, for which the fully discretized WP,

$$V_W(x, y, p, q) = \frac{1}{i\hbar MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-i2\left(pm\frac{\pi}{M} + qn\frac{\pi}{N}\right)} \ldots$$
$$\delta V\left(x \pm \left(m - \frac{M}{2}\right)\Delta s, y \pm \left(n - \frac{N}{2}\right)\Delta s\right),$$

(4.4)

must be calculated at each node of the mesh.

4.1.2   Efficient calculation of two-dimensional Wigner potential

The WP has to be computed at every mesh point in the domain for a given potential profile. The calculation of the 2D WP can consume considerable computation time, if it must be recalculated many times, e.g. if the potential is time-dependent or a self-consistent solution is pursued. This section presents an algorithm, first introduced in [128], named the box discrete Fourier transform (BDFT), which reduces the computational effort of calculating the 2D WP by at least a factor of five.

Equation (4.4) is akin to a 2D DFT of (4.2), conventionally calculated using a row-column decomposition scheme, which entails the successive application of a one-dimensional (1D) DFT algorithm: With reference to Figure 4.2, consider a $M' \times N'$ matrix of values representing the calculated potential differences, as per (4.2). First the 1D DFT of each row of values is calculated, which yields a $M' \times N'$ matrix of Fourier coefficients. Thereafter, the 1D DFT of each column of the latter matrix is calculated, the result of which corresponds to the 2D DFT.

The fast Fourier transform (FFT) algorithm has a computational complexity, for a problem size N, of $\mathcal{O}(N \log_2 N)$ and presents the *de facto* standard algorithm for calculating 1D DFTs, thanks to flexible, highly optimized implementations, which are freely available in libraries, like FFTW3 [129]. Algorithms which directly calculate multi-dimensional DFTs exist, e.g. [130, 131], and have a reduced computational complexity, which should theoretically result in superior computational performance. However, these advantages are most often completely eroded in practical implementations, which can be attributed to the fact that the cache complexity of algorithms implemented on modern hardware architectures plays an equally important role as the computational complexity. Moreover, these multi-dimensional algorithms require extensive 'tailoring', e.g. to the number of dimensions or problem size (N), making general purpose implementations difficult, thereby hampering a wide-spread adoption of such algorithms.
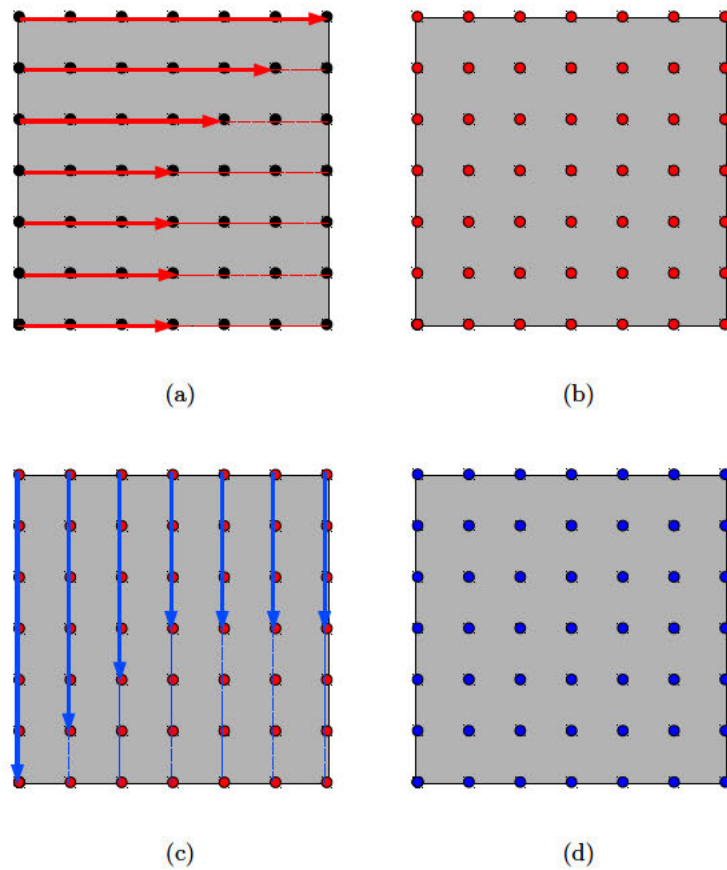
Figure 4.2: Illustration of the calculation of a 2D DFT using the successive application of 1D DFTs in the orthogonal dimensions: (a) the 1D DFT is calculated for each row, yielding (b) a matrix of Fourier coefficients (red). The (c) 1D DFT of each column of the prior result yields (d) the final result (blue).

4.1.2.1 Algorithm

The BDFT algorithm expands on the idea of the 1D sliding DFT [132, 133]: The sliding DFT calculates the Fourier coefficients of a sequence $\{x_c..x_{c+N-1}\}$ using the coefficients calculated for $\{x_{c-1}..x_{c+N-2}\}$:

$$X_c(p) = e^{i\frac{2\pi p}{N}}\left(X_{c-1}(p) + x_{c+N-1} - x_{c-1}\right). \tag{4.5}$$

Each application of (4.5) consists of two real additions and two complex multiplications, which have to be repeated for each value of p. Therefore, the sliding DFT has a computational complexity of $\mathcal{O}(N)$. To apply the sliding DFT, as in (4.5), the two sequences under consideration must differ by only a single value; the potential values of each row (column) of the coherence boxes associated with two horizontally (vertically) adjacent nodes in the domain also differ only by a single value. This observation is exploited to calculate the 2D WP in an efficient manner.

Unlike the potential values, all the values of the potential difference (4.2) change between adjacent nodes. To allow a direct application of (4.5) to calculate (4.4), (4.1) is reformulated using a substitution of variables (Fourier shift theorem), such that

$$V_W\left(\mathbf{r}, \mathbf{q}\right) = \frac{2}{\hbar \mathbf{L}}\text{Im}\int_{-\mathbf{L}/2}^{\mathbf{L}/2} d\mathbf{s}\, e^{-i2\mathbf{q}\Delta\mathbf{k}\cdot\mathbf{s}}V\left(\mathbf{r} + \mathbf{s}\right). \tag{4.6}$$

This formulation has the additional advantage that it avoids the calculation of the potential difference, saving further computation time. It can be noted that (4.5) allows, unlike the FFT, to easily compute only selected momentum (p, q) values, which do not have to be uniformly spaced. This can be of interest under certain physical considerations, e.g. uniformly spaced energy grid, and offers an additional possibility to reduce computational costs.

The BDFT algorithm is applied to calculate the WP at each node in the domain, using the following procedure (as visualized in Figure 4.3): First, the 1D DFT of the first N' potential values in each of the M rows in the domain are calculated, using an FFT algorithm (Figure 4.3(b)); the resulting Fourier coefficients are retained in an array of size $N' \times M$. Thereafter, the 1D DFTs of the first M' Fourier coefficients of each of the N' columns are calculated (Figure 4.3(c)), which yields an $M' \times N'$ matrix of Fourier coefficients representing the WP for the top-left node, $V_W(0, 0, p, q)$. After this initialization, the coherence box is moved downwards to the next node for which the WP is calculated by simply applying (4.5) to calculate the DFTs of the columns (Figure 4.3(d),(e)). Once the WP has been calculated for each node in the first column of the domain, the $N' \times M$ array is developed to the right (Figure 4.3(f)), again using (4.5). From there the same procedure as used for the first column (Figure 4.3(a)-(e)) is repeated for the second column of nodes (Figure 4.3(g),(h)) etc., until the entire domain has been covered. Variations of the initialization approach can be envisioned, but the presented procedure shows favourable serial performance and cache complexity; a parallelized implementation would require multiple (modified) initializations.

4.1.2.2 Performance evaluation

The BDFT algorithm was benchmarked against an FFT implementation using the FFTW library [129], with a setup detailed in Table 4.1. The FFT implementation was optimized by exploiting the fact that (4.2) is real-valued and anti-symmetric and therefore must yield a purely imaginary output with conjugate symmetry.

Table 4.2 makes it evident that the BDFT reduces the computation time by at least a factor of five over a range of (plausible) domain sizes, as visualized in Figure 4.4. Table 4.2 also reveals that the performance of the FFT implementation strongly depends on the transform size (which is proportional to the coherence length), because the algorithms selected by the FFTW library perform best with transform sizes that are products of small prime numbers. The BDFT algorithm, on the other hand, is insensitive to the transform size and scales at a constant rate with size.

It is concluded that the presented box discrete Fourier transform (BDFT) algorithm is an efficient approach to compute the WP in a 2D domain, with a significant reduction in computation time. The BDFT algorithm can easily be extended to three dimensions and makes self-consistent solutions of the Wigner equation more feasible.
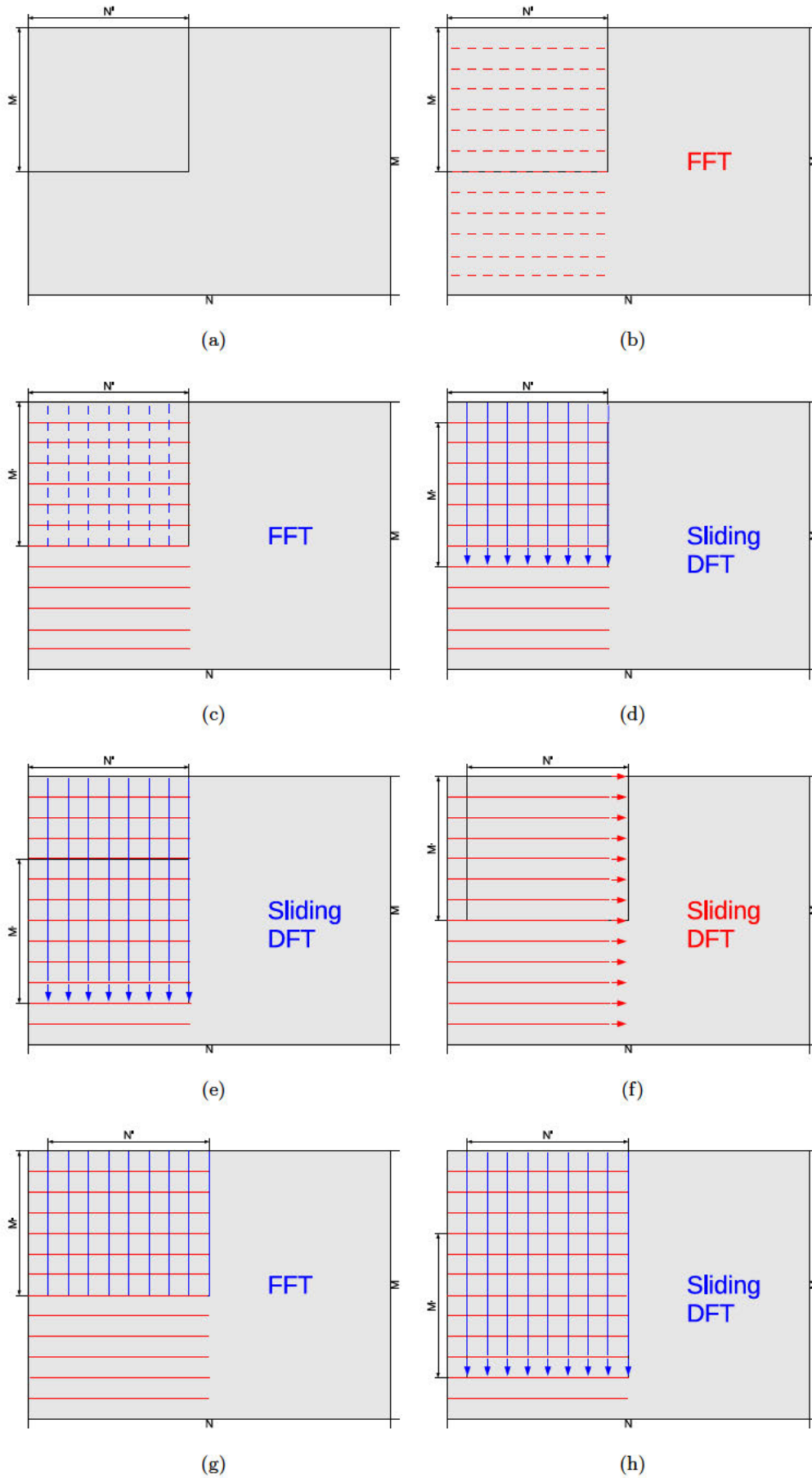
Figure 4.3: Initialization and application of the BDFT algorithm, used to sequentially compute the WP at every node in the computational domain by successively applying the FFT and sliding DFT algorithms.
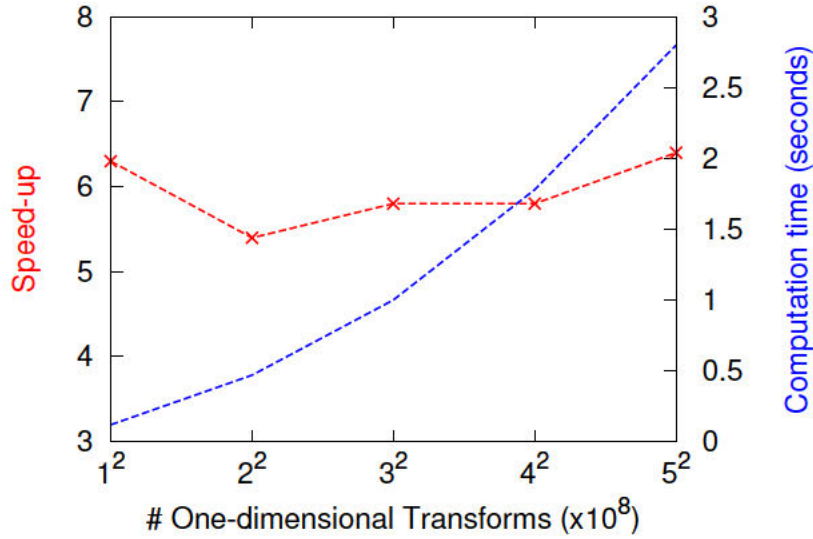
Figure 4.4: Speed-up (as per Table 4.2) of the BDFT algorithm versus a (pure) FFT implementation, to calculate the 2D WP for devices of various sizes, using a coherence box containing 100×100 potential values.

4.1.3 Choice of coherence length

The semi-discrete formulation of the WTE (2.35) and the WP (4.1) rely on a finite coherence length to be chosen. The second important contribution made surrounding the WP concerns an analysis of the physical interpretation attributed to a chosen coherence length and its computational implications.

4.1.3.1 Physical interpretation

The Wigner theory is formulated for a coherence length approaching infinity. However, in a device, with dimensions $\mathbf{L}_{dev}$, there does not exist coherence between any two contacts, which limits the maximum coherence length to the dimensions of the device (separation of the contacts), i.e $\mathbf{L} = \mathbf{L}_{dev}$. Furthermore, in practical computations the coherence length also cannot be taken to be infinite. When calculating (4.1) at any point other than the centre of the device, i.e. $V_W\left(\mathbf{r} \neq \frac{\mathbf{L}}{2}, \cdot\right)$, however, the boundaries of the device are exceeded and therefore the potential is unknown. Two different approaches to treat this situation are discussed in the following.

i) Regard any point outside the device to be in a contact region, which is assumed to maintain equilibrium conditions, and extend the potential value at the boundary as a constant or to ii) assume no coherence exists between the device and the contact region outside and progressively reduce the coherence length as the boundary is approached. Figure 4.5 reveals that the difference between these two approaches is significant. Approach ii) implies a decreasing value of Δk as the boundary is approached. To efficiently interpolate the values onto the fixed, finer, momentum grid, the potential difference in (4.1) is zero-padded at the front and the back. A further consideration when

Table 4.1: Setup for benchmark of BDFT algorithm

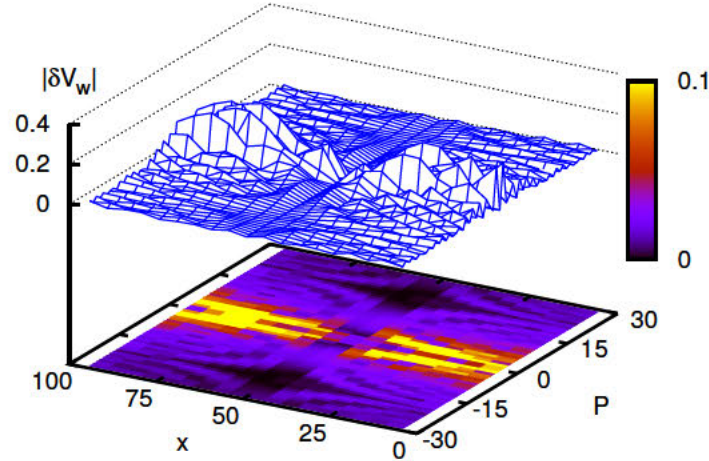| Hardware	Intel Core 3110M; 8 GB (dual channel)
OS	Ubuntu 13.10 (64 bit)
Compiler	gcc 4.8.1
flags	-O3 -fastmath -march=native
FFT library	FFTW 3.3 (SIMD enabled)
interface \ flags	dft_r2c_2d \ FFTW_MEASURE

Figure 4.5: The absolute value of the difference of the (normalized) WPs corresponding to the different boundary treatments (discussed in text), using Profile 2 (blue) in Figure 4.6.

using approach ii) is that a shrinking coherence length implies that the portion of the potential profile considered for calculating the WP gradually reduces as the boundary is approached, up to the point, where the WP is no longer predominantly determined by the potential profile, but rather the length/width of the 'coherence box' – a rectangular window with undesirable numerical effects, as will be discussed in the following section.

The coherence length is often associated with the distance over which quantum information, in the form of an electron state, is conveyed. Due to inelastic phonon scattering the electron state is changed and this information is lost. Based on this argument, an increase in the scattering rate suggests a decrease in the coherence length. However, the coherence length is often wrongly considered to be equivalent to the inelastic mean free path, as pointed out in, in two orthogonal dimensions, [134].

Further insight into the meaning of the coherence length can be gained from the *scaling theorem* – an analysis of a dimensionless, scaled version of the WBE, which considers the scales of the involved physical quantities: electron energy, electrostatic potential, phonon energy and electron-phonon coupling [49]. The analysis reveals that an increase of the electron-phonon coupling effectively reduces the distance over which the WP is 'felt'. Therefore, decoherence effects can be modelled through the choice of $\mathbf{L}$ itself, e.g. [135] considers an exponential damping of the WP, which essentially reduces $\mathbf{L}$ to model decoherence. Choosing a shorter coherence length can considerably reduce the computational burden of the simulation and is also justified, if the modelled decoherence processes reduce the effective coherence length sufficiently.

The physical considerations made to choose the coherence do not (significantly) depend on orientation. For computational convenience an isotropic coherence length is chosen such that $|\mathbf{L}| = L$.

Table 4.2: Comparison of computation times for two-dimensional Wigner potential

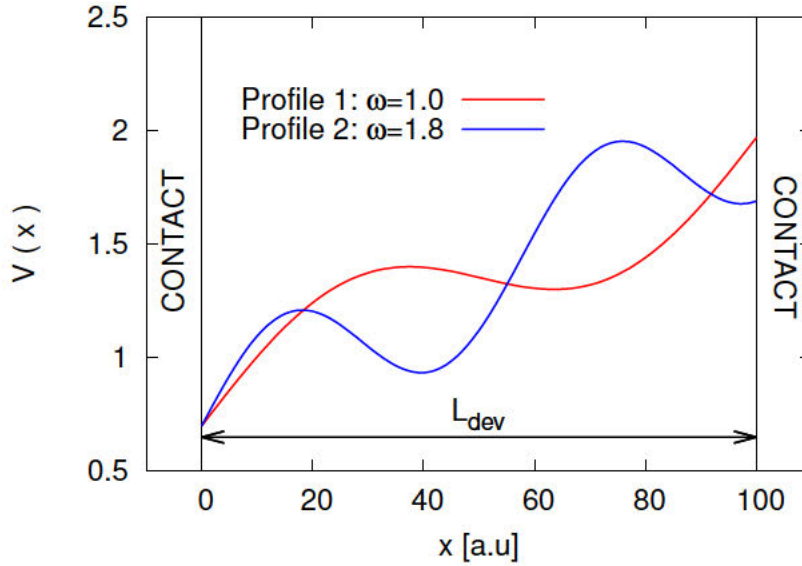| L_{dev} [a.u.]	L_{coh} [a.u.]	BDFT [s]	FFT [s]	Speed-up [1]
100	100	0.12	0.75	6.3
200	100	0.47	2.53	5.4
300	99	0.96	10.66	11.1
300	100	1.00	5.77	5.8
300	101	1.04	61.25	55.9
400	100	1.78	10.27	5.8
500	100	2.80	17.84	6.4

Figure 4.6: Two analytic potential profiles, of the form (4.7), with known spectral content (ω) and a bias between the left and the right contacts of 1.3 V.

4.1.3.2   Computational aspects

It is desirable to choose the coherence length as short as possible to limit the computational load of calculating the WP, which can be substantial in multi-dimensional simulations. However, the coherence length also determines the momentum resolution, $\Delta k = \frac{\pi}{L}$, therefore, **L** must be chosen sufficiently large to be able to differentiate the spectral content of different potential profiles. However, the attainable momentum resolution is limited by the dimensions of the device as pointed out in the previous section.

Figure 4.6 shows different analytical potential profiles of the form

$$V(x) = V_0 + mx + V_1 \sin \omega x, \tag{4.7}$$

which suggest a sinusoidal potential variation, with a spatial frequency ω, superimposed on a potential bias $(V_0 + mx)$.

Since ω can be chosen, the values of the momentum index p (bins) of a 1D version of (4.4), which should be non-zero are known. If ω is chosen ($\omega = 1.0, 1.8$; *cf.* Figure 4.6) such that only the bins 1 and 2 of the corresponding WPs (at a fixed position) should be non-zero, Figure 4.7 reveals that this is not the case – there appear non-zero values at much higher-valued bins. These values at higher bins are not attributable to the physical profile, but are due to adverse effects, often termed 'spectral leakage', inherent, when calculating a DFT of a non-periodic 'function' over a finite length [136]. This situation is further exacerbated by the biasing condition which introduces a big discontinuity, because, when calculating the DFT, the potential profile is implicitly assumed to repeat periodically.

4.1.4   Tapering window

The preceding section has demonstrated situations where the components of the WP appear at values which cannot be attributed to a physical profile. Since the WP is normalized to represent a probability distribution for the generation of particles (see Section 3.7.5), the accumulated probability of the higher-valued bins (Figure 4.7) becomes significant and causes the generation of particles with very high momenta. The generation of these particles leads to a dramatic non-physical behaviour in the simulation. This section presents an approach that avoids this problem and represents an important step forward in attaining physically accurate simulation results with the WMC method. The results are based on [137] and are elaborated upon in the following.

To suppress the generation of particles which are not physically sound, a Tukey window function [138] is applied to taper the potential profile at the boundaries towards zero (or to the average
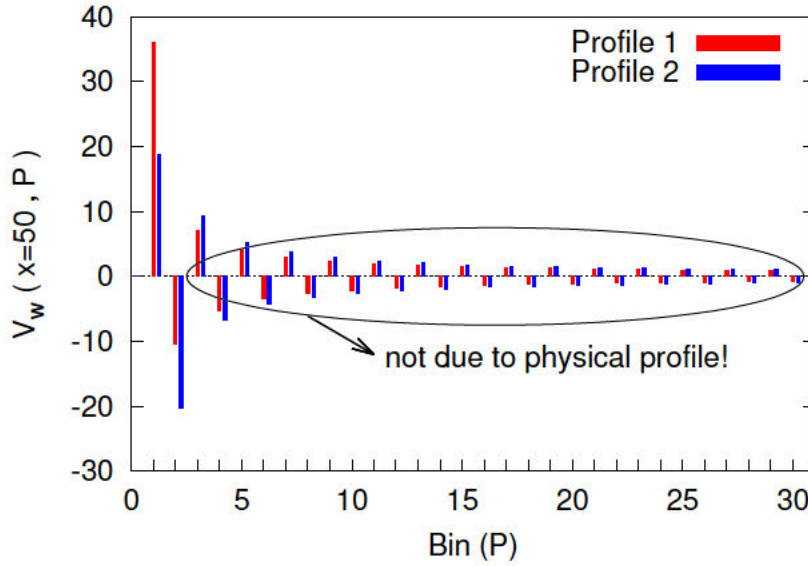
Figure 4.7: Wigner potential at a fixed position, $V_W(x = 50, P)$, for the profiles in Figure 4.6, calculated with a coherence length of $\mathbf{L} = \mathbf{L}_{dev}$. The components appearing the higher bins ($P \geq 5$), are not attributed to the physical potential profile, defined in (4.7).

value between the two boundary values), which yields a probability distribution that corresponds much better to the physical profiles. The smoothing mitigates the discontinuities introduced by the periodic repetition of the potential profile, which is implied when calculating a DFT.

Amongst the plethora of available window functions [136] – also known as tapering functions – the Tukey window (with $\alpha = 0.2$; *cf.* Figure 4.8) was chosen with the physical motivation that it does not alter the physical profile inside the device noticeably. The Tukey window gives a coefficient with which sample n is multiplied and defined as

$$
w(n) = \begin{cases}
\frac{1}{2}\left[1 + \cos\left(\pi\left(\frac{2n}{\alpha(N-1)}\right) - 1\right)\right] & \text{if } 0 \leq n < \frac{\alpha(N-1)}{2} \\
1 & \text{if } \frac{\alpha(N-1)}{2} \leq n < (N-1)\left(1 - \frac{\alpha}{2}\right) \\
\frac{1}{2}\left[1 + \cos\left(\pi\left(\frac{2n}{\alpha(N-1)}\right) - \frac{2}{\alpha} + 1\right)\right] & \text{if } (N-1)\left(1 - \frac{\alpha}{2}\right) \leq n \leq (N-1)
\end{cases} \tag{4.8}
$$

The parameter α controls the number of samples, of the total N, which are altered by the smoothing.

Figure 4.9 shows that by applying the Tukey window the (non-physical) higher bin values are suppressed, while the actual spectral components of the profile, at bins 1 and 2, remain pronounced. This results in a relative change of the probability distribution, which better reflects the physical situation in the device as illustrated by the following example.

A practical example showing the importance of the tapering window is shown in Figure 4.10. A wavepacket travels, from the left, along a potential profile suggesting the channel (in transport direction) of a field-effect transistor (FET) with a k-distribution initially centred around $12\Delta k$.

As the small potential barrier is approached a part of the wavepacket is slightly decelerated/reflected while the other part is accelerated by the big potential drop. An investigation of the (blue) k-distribution in Figure 4.10, however, reveals three peaks instead of the two suggested by physical considerations. The third (smallest) peak indicates a very strong reflection with a change in energy much larger than the potential difference between the left and the right contacts (boundaries). The latter is not physically grounded, but rather a consequence of the high-momenta particles being generated, if no tapering window is applied to treat the discontinuity introduced by the potential difference between the boundaries.

The application of a Tukey window function, which smooths the potential towards zero at the limits of the region of coherence, clearly remedies this problem (shown in red in Figure 4.10) and also avoids negative 'probabilities' appearing in the k-distribution, leading to physically consistent results.
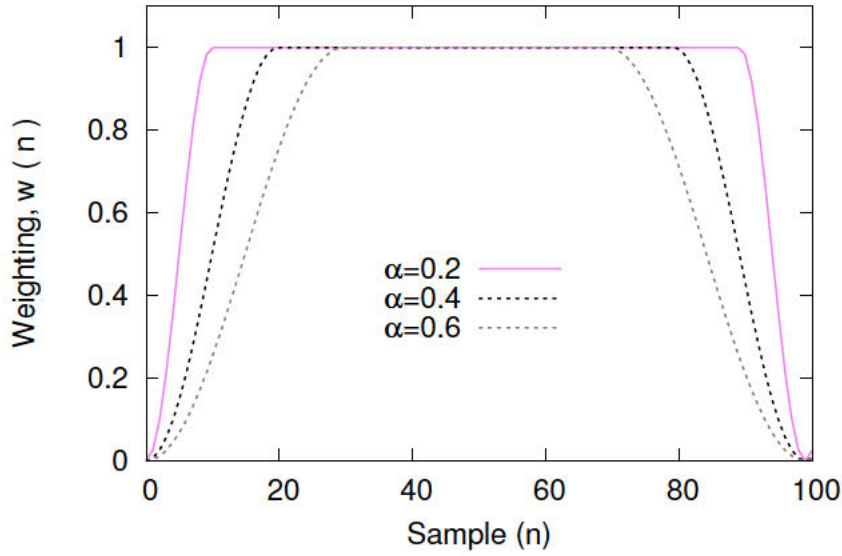
Figure 4.8: Tukey window function used to smooth the potential values toward zero at the boundaries by assigning an appropriate weighting to the samples (potential difference values) used to compute the DFT in the Wigner potential. The number of samples affected by the tapering is controlled by the parameter α.
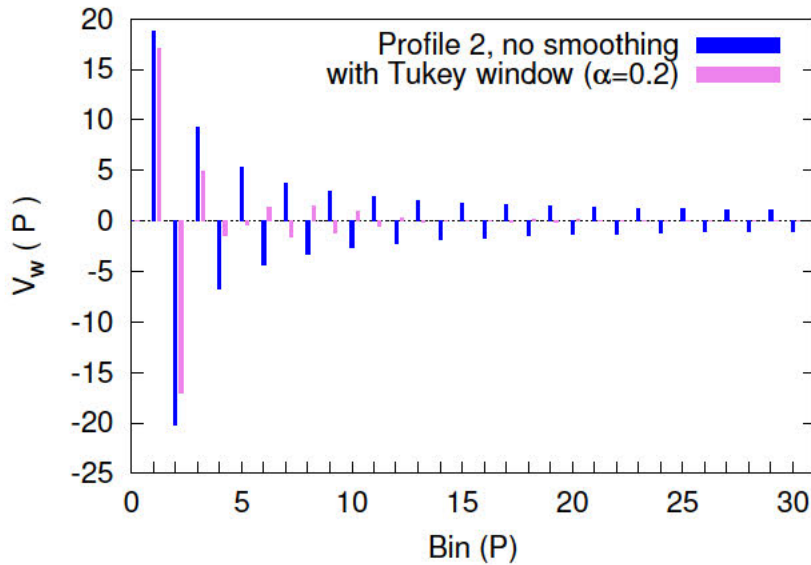

Figure 4.9: The unnormalized Wigner potential, $V_W (x = 50, P)$, calculated at the centre of the device with and without the application of a smoothing Tukey window ($\alpha = 0.2$), using Profile 2 (blue) in Figure 4.6, and a coherence length of $\mathbf{L} = \mathbf{L}_{dev}$. The non-physical values revealed in Figure 4.7 are suppressed by the application of the tapering window.
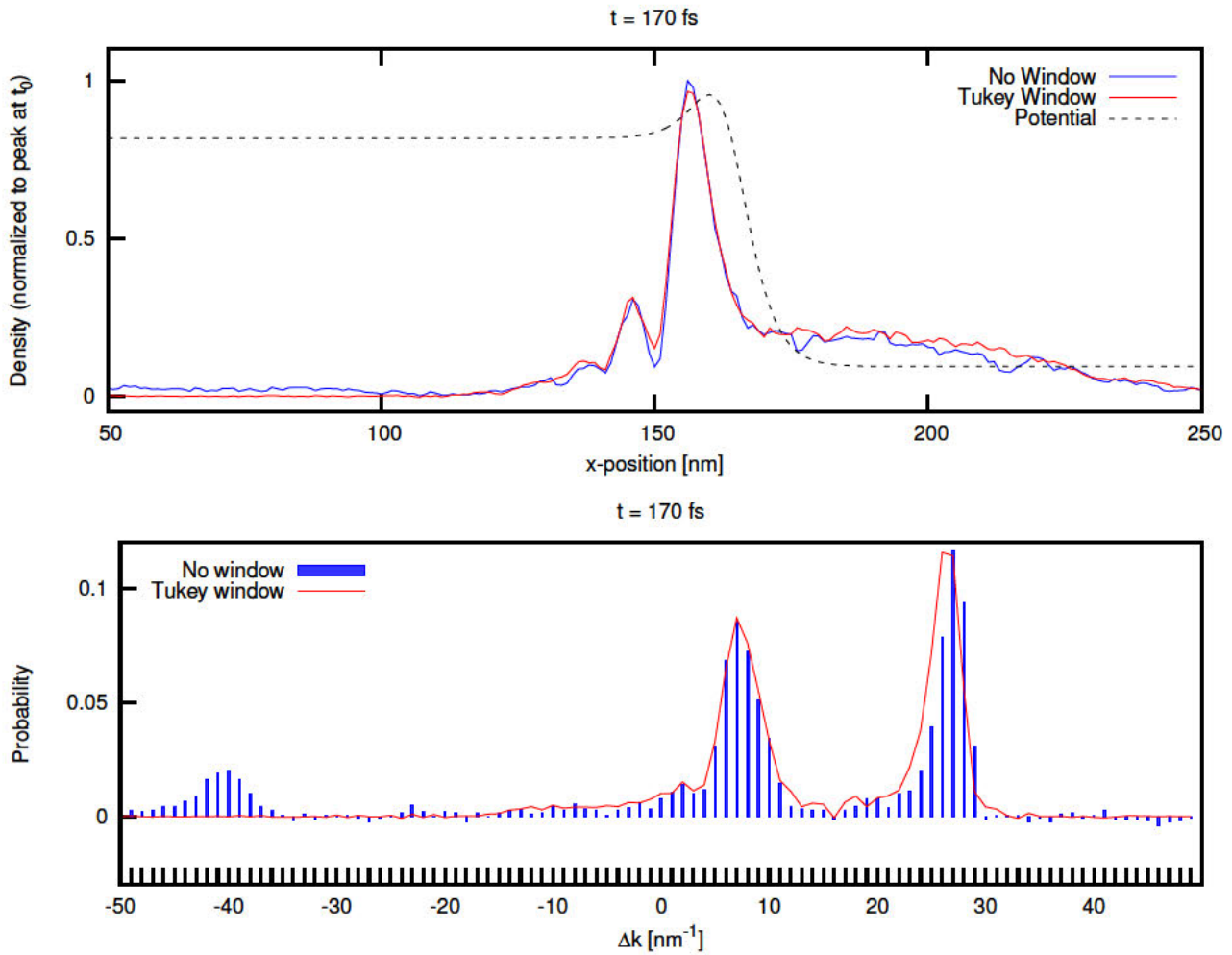
Figure 4.10: The density and k-distributions of a wavepacket that is propagated from the left along a potential profile, suggesting a biased transistor. If no tapering window is applied when calculating the Wigner potential, non-physical reflections (around $-40\Delta k$) and negative densities in the k-distribution appear.

4.2   Generation process

The value of the wavevector for newly generated particles using a single random number sampling has been discussed in Section 3.7.5. A problem in this approach arises, if the finite range of the multi-index $\mathbf{q}$ is considered.

In (4.1) the variable $\mathbf{s}$ is bounded by the finite coherence length $\mathbf{L}$. To ensure that the Wigner transform and its inverse are unitary, the transformed variable $\mathbf{q}$ must also be bounded. The range of (any component of) the multi-index $\mathbf{q}$ is restricted by

$$\pm K = \pm \frac{L}{2\Delta s},\tag{4.9}$$

where $\Delta s = \Delta r$ is usually chosen to avoid an interpolation between the grids. An estimation of the error introduced in the density by the choice of K was made in [73]. However, the value of K, which results for reasonable values of the coherence length $\mathbf{L}$ and the mesh-spacing Δs, is large enough to accommodate the oprticle momenta which can be expected from physical considerations.

The momentum offsets of the newly generated particles (ℓ_1 and ℓ_2) should be such that their momenta remain within the set bounds, i.e. $(q + \ell_1, q + \ell_2) \in [-K, K]$. The problem arises if particles are generated in pairs with momenta $q \pm \ell$, using a single sampling of the distribution function $\frac{V_w^+}{\gamma}$ (or $\frac{V_w^-}{\gamma}$). To maintain a balance of positive- and negative-signed particles, it seems reasonable to reject

both particles and sample the distribution again, until a momentum offset appears which renders both momenta valid simultaneously. Such a rejection technique clearly influences the statistics. Moreover, the probability for a valid particle pair to be generated decreases, the closer the momentum of the generating particle is to the limit $\pm K$, as illustrated in Figure 4.11. In such a case, a small momentum offset is more likely to produce a pair of generated particles with valid momenta, thereby unfairly promoting the generation of particles with small offsets in momenta and influencing the momentum distribution of the particle ensemble as a whole. Due to this biasing, particles with a high momentum persist much longer, since it becomes impossible for particles to be generated with significantly different momenta.
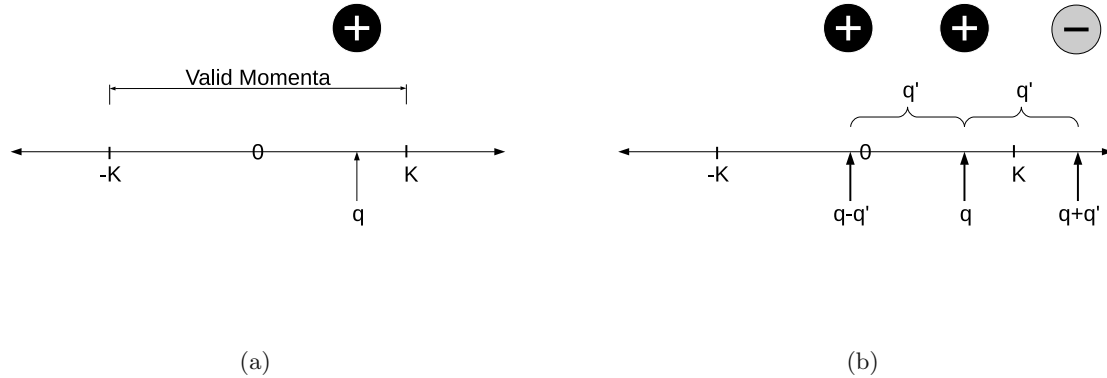

(a)   (b)

Figure 4.11: A particle can generate a particle pair only with momenta in the valid range $[-K, K]$ (a). If a symmetric offset (q') is use for both generated particles the statistics are biased towards smaller offsets such that the new momenta remain in the valid range (b).

To avoid this systematic biasing of the statistics, the generated particles should not be rejected in pairs, if one is assigned a momentum which is out of bounds, but rather only a single momentum offset for the invalid particle should be regenerated until a valid momentum (inside the finite bounds) is obtained. In practice, if no valid momentum can be obtained after a set number of attempts, both of the generated particles are 'destroyed'. This ensures that the balance between positive and negative particles is maintained at all times.

This optimized particle generation algorithm enables particles with a high momentum to again return to a lower momentum; the persistence of high-momentum particles is no longer promoted by the generation statistics.

4.3 Annihilation process

The concept of particle annihilation has been introduced in Section 3.7.6. The annihilation procedure is the crucial aspect of the signed-particle method, which has made multi-dimensional simulations computationally feasible. For this reason the effects of the approximations that are made are given careful consideration in this section. Moreover, two alternative implementations of the annihilation algorithm are presented, which reduce or completely eliminate the huge memory demands of the conventional annihilation algorithm.

4.3.1 Numerical diffusion

The phenomena of numerical diffusion brought about by the regeneration of particles has been identified in [139] and is reviewed in the following.

After the annihilation of particles within a cell has taken place, the remaining particles have to be regenerated. The conventional approach to selecting the position of the regenerated particles is to spread the particles uniformly across the cell. This approach, however, may lead to a 'numerical
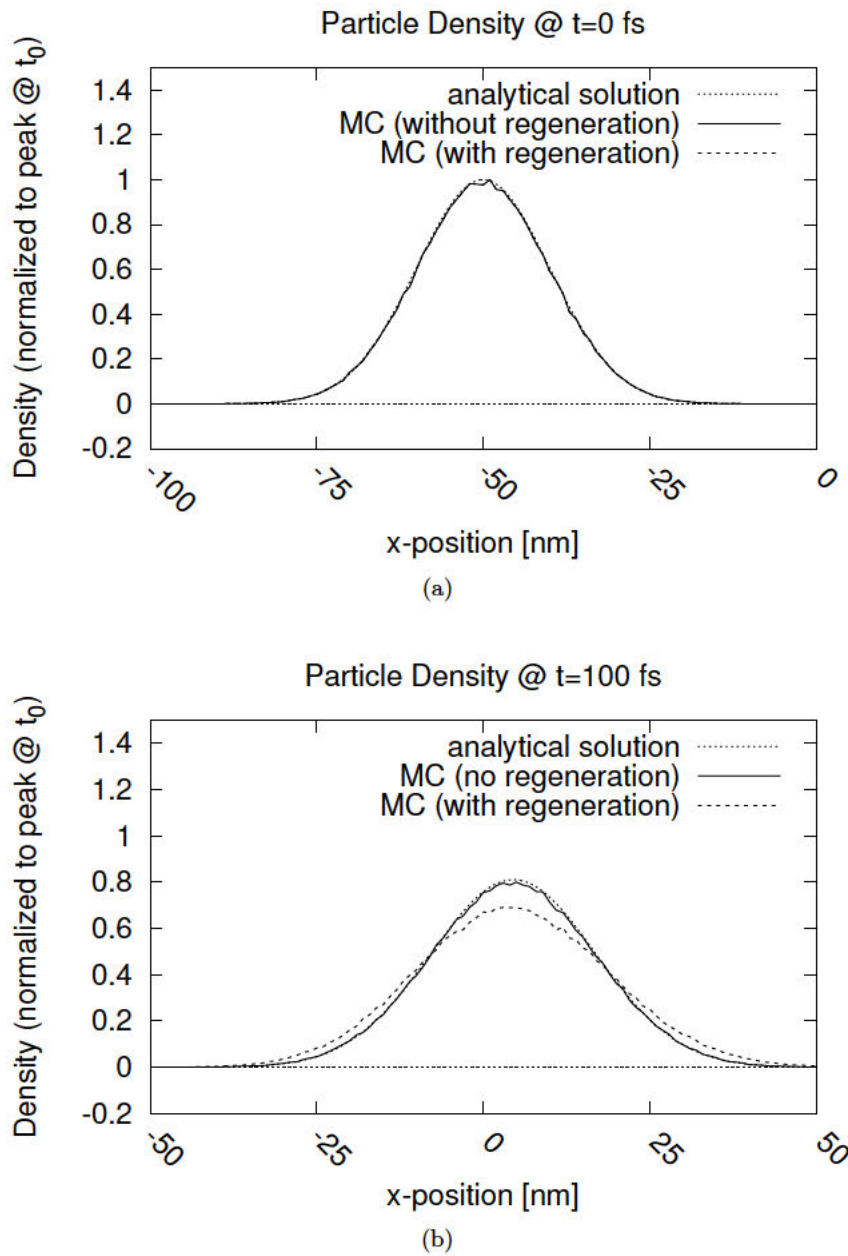
Figure 4.12: Comparison of a wavepacket evolved from (a) 0 fs to (b) 100 fs, using an analytical solution and a Monte Carlo (MC) approach without and with the regeneration process (repeated every 0.1 fs).

diffusion' of particles, which causes the global particle ensemble to propagate at a different rate than dictated by its k-distribution. To demonstrate this numerical artefact the following example is considered.

A 1D minimum uncertainty wavepacket of the form (3.34) with $r_o = -50\,\mathrm{nm}$, $k_0 = 6 \left(\frac{\pi}{50}\right)\,\mathrm{nm}^{-1}$ and $\sigma = 10\,\mathrm{nm}$ propagates in a domain with zero potential. The evolution of the wavepacket is compared in Figure 4.12 using three different approaches: i) an analytical solution, ii) a Monte Carlo approach without any regeneration and iii) a Monte Carlo approach with a (forced) regeneration procedure at each time step (a typical value for the annihilation of 0.1 fs is chosen). It is evident that the approaches i) and ii) correspond exactly, however, the wavepacket which is subjected to the regeneration procedure spreads out faster. This discrepancy must be attributed to the regeneration procedure and is analysed in the following.

Consider an ensemble of N particles, with positions $\{r_j\}$ $j = 1 \ldots N$, $r_j \in \Omega_i$, within the cell (i, q) at time t_0. The mean position of the ensemble at time t_0 is

$$
\begin{aligned}
\bar{r}_{t_0} &= \frac{1}{N} \sum_{j=1}^{N} r_j \\
&= r_i + \frac{1}{N} \sum_{j=1}^{N} \delta r_j,
\end{aligned}
\tag{4.10}
$$

where the position is expressed as $r_j = r_i + \delta r_j$, $\delta r_j \in [0, \Delta r]$. The particles of the ensemble evolve (drift; *cf.* Section 3.7.4) for a time period Δt, whereafter the mean position of the ensemble at time t_1 is

$$
\begin{aligned}
\bar{r}_{t_1} &= \frac{1}{N} \sum_{j=1}^{N} r_j + v_j \Delta t \\
&= r_i + \frac{1}{N} \sum_{j=1}^{N} \delta r_j + v_j \Delta t,
\end{aligned}
\tag{4.11}
$$

where v_j denotes the velocity of particle j, which is assumed to be small enough such that the particle remains within the bounds of the cell for one time step. Since only a single discrete momentum value is associated with the cell, all particles within the cell have the same velocity (v_m). Therefore,

$$
\bar{r}_{t_1} = \bar{r}_{t_0} + v_m \Delta t.
\tag{4.12}
$$

Now, suppose that before the particle evolution commences an annihilation step is performed, whereafter N' particles are regenerated within the cell with positions $\left\{r'_j\right\}$ $j = 1 \ldots N' \le N$. If the particles are uniformly distributed over the cell, one imposes

$$
\bar{r}'_{t_0} = \frac{r_i + r_{i+1}}{2} = r_i + \frac{\Delta r}{2}.
\tag{4.13}
$$

Consequently, the mean position of the ensemble at time t_1 will be

$$
\bar{r}'_{t_1} = r_i + \frac{\Delta r}{2} + v_m \Delta t,
\tag{4.14}
$$

which, when compared to (4.12), introduces an artificial propagation/retardation depending on the spatial distribution of particles before the annihilation procedure.

The original spatial distribution of the particles within a cell can be perfectly recovered, if all (infinitely many) of the moments of the distribution before the annihilation are known (and the Carleman's condition for uniqueness is satisfied [140]). The mean position represents the first moment of the local distribution and already retains the most important information. By uniformly distributing the particles over a distance Δr around the pre-annihilation mean, the 'numerical diffusion' is effectively

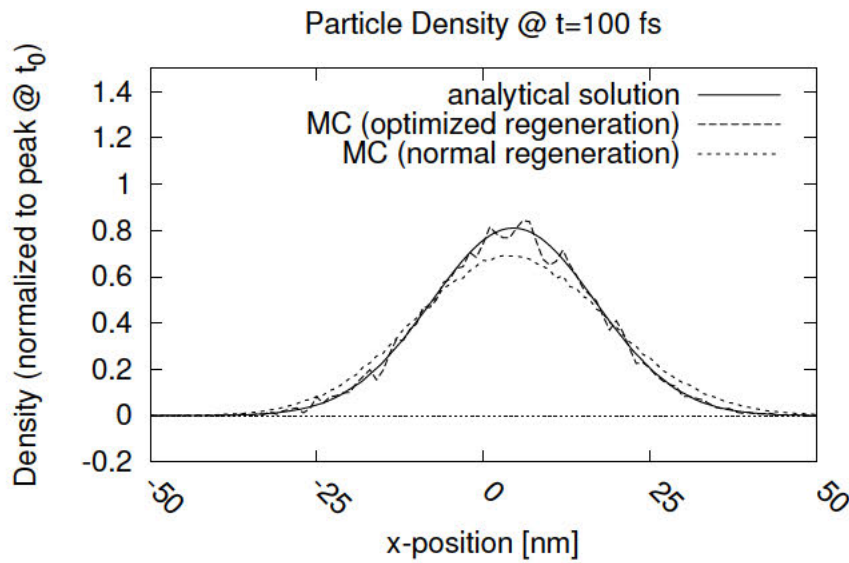Figure 4.13: Comparison of wavepackets evolved for 100 fs using an optimized regeneration scheme and the conventional regeneration process; the analytical solution is shown by the solid line.
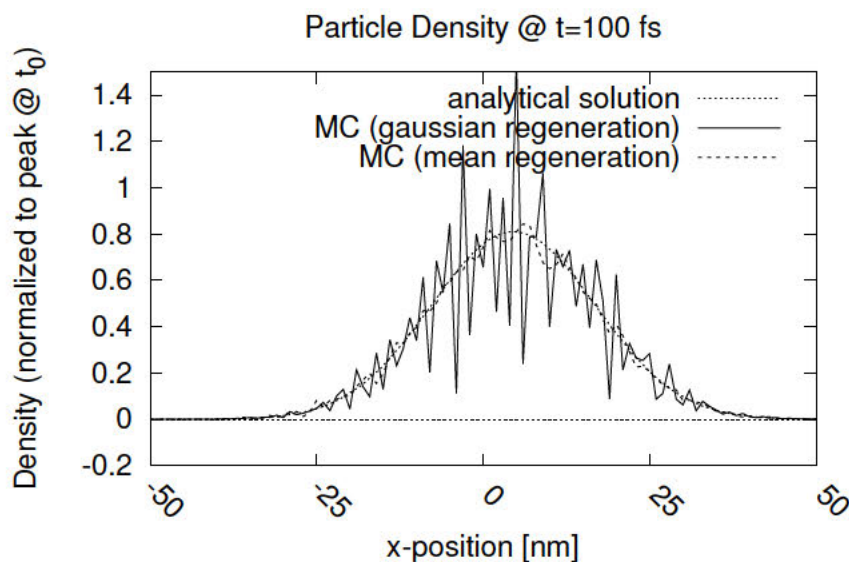
Figure 4.14: Comparison of wavepackets evolved for 100 fs using regeneration schemes based on a Gaussian distribution and a mean with uniform distribution (as in Figure 4.13); the analytical solution is shown for comparison.
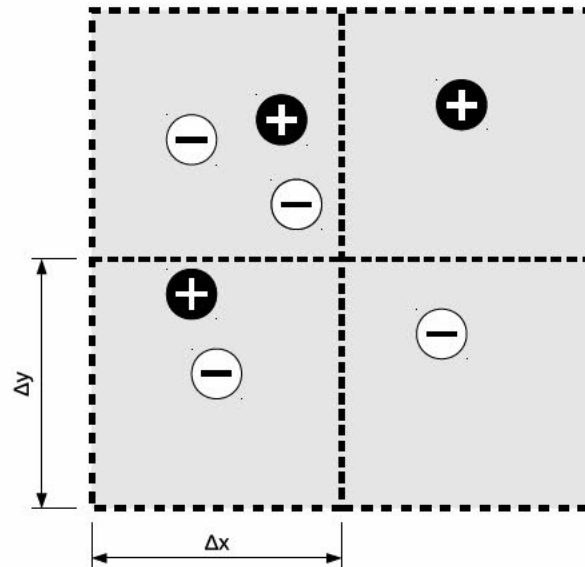
Figure 4.15: A distribution of + and − particles in four spatial cells, indicated by the dashed lines. The grey cell, encapsulating all four cells, represents a cell of the coarsened phase-space grid used to perform annihilation. It is enlarged by a factor 2 in both directions, reducing the array size by a factor of four. The effectiveness of annihilation is improved from 3 to only 1 surviving particle.

remedied, albeit with some added 'noise', as shown in Figure 4.13. This 'noise' is attributed to the fact that the uniform distributions of neighbouring cells overlap.

If, in addition to the mean, the second moment of the distribution – the standard deviation – is also calculated the particles can be regenerated using e.g. a Gaussian distribution. The result, shown in Figure 4.14, is very noisy, however, since a Gaussian distribution poorly models the actual distribution in each cell: the Gaussian distribution approaches zero in both directions, whereas the actual particle distribution has a non-zero value at the boundaries of the cell. All the common statistical distributions (Gaussian, exponential, beta etc.) tend towards zero on at least one tail, which makes them unsuitable to obtain a satisfactory fit with the particle distribution in a single cell.

In summary, an artificial propagation/retardation of particles ('numerical diffusion') arises, when the spatial distribution of particles, within a phase-space cell, is not considered prior to the annihilation for the subsequent regeneration. By calculating the mean position of the particles in a cell before annihilation takes place presents an effective way to avoid this 'numerical diffusion'. The computational costs of this approach remains almost negligible (<1% for the presented cases) and, therefore, is well-suited for problems where computational penalty of simply decreasing Δr would be intolerable. The proposed solution presents another optimization which makes the WMC simulations more computationally efficient and thereby more accessible for users with limited computational resources.

4.3.2   Spatial scaling

The memory required to represent the phase space grid, on which particles are recorded for annihilation, quickly becomes exorbitant in multi-dimensional simulations with a fine spatial resolution, since the number of cells increases with the power of the dimensionality of the phase space (the resolution also affects the number of k-values which must be retained to ensure a unitary Fourier transform). A high spatial resolution is required is required to investigate certain problems, e.g. the direct modelling of surface roughness discussed in Appendix C.

A $30\,\text{nm} \times 30\,\text{nm}$ domain and a three-dimensional k-space with a coherence length of $30\,\text{nm}$ results in an array size exceeding $141\,\text{GB}$ at a resolution of $0.2\,\text{nm}$. Chapter 5 presents a distributed-memory (MPI) parallelization approach to address these large memory requirements, using a spatial domain decomposition. However, the computational demands of the signed-particle WMC simulator allow it to be run on a typical desktop computer, therefore, its memory demands should also follow suit.
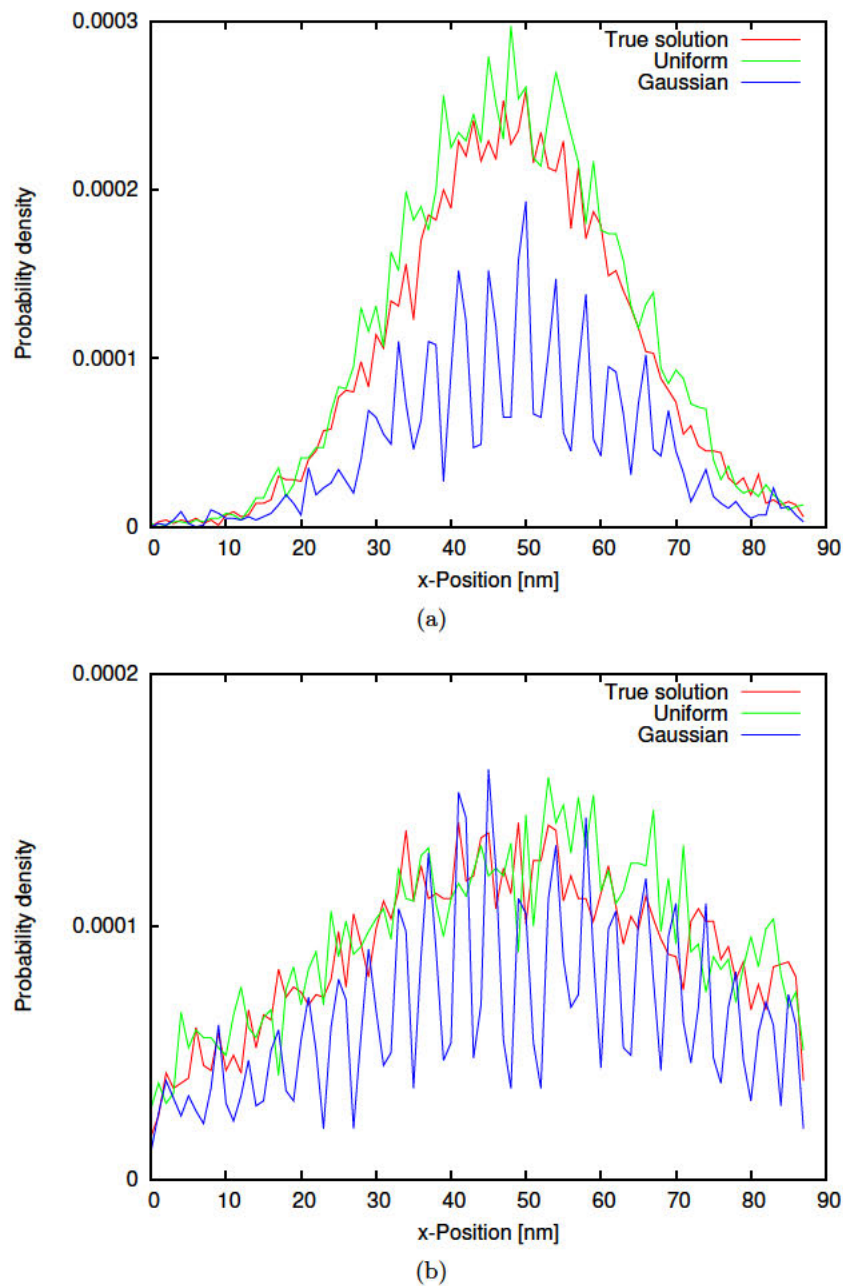
Figure 4.16: A slice of the two-dimensional probability density of a wavepacket, evolving freely in a domain with a spatial resolution of 0.25 nm, after (a) 40 and (b) 80 forced annihilation steps. The annihilation is performed on a coarsened grid with a 1 nm resolution and the particles are regenerated using uniform and Gaussian distributions. The 'true solution' – the evolution when the annihilation step is omitted – is followed the best when particles are regenerated by a uniform distribution; the Gaussian distribution artificially introduces information, which leads to unsatisfactory results.

A possibility to reduce the memory requirements of the annihilation algorithm is to reduce the spatial resolution of the grid on which the particles are recorded for annihilation; the resolution of the k-values and the potential mesh remain unaltered. The concept is depicted in Figure 4.15. To counteract the loss in resolution, the spatial distribution of the particles in the enlarged cell is fitted to a statistical distribution before annihilation ensues. The obtained distribution is then used to regenerate the particles which remain after the annihilation – essentially, the approach presented in Section 4.3.1 to counteract numerical diffusion is applied here for a bigger spatial area to reduce the memory requirements.

Figure 4.16 compares the regeneration of particles, annihilated on a coarsened grid, using a uniform distribution and a Gaussian distribution around the pre-annihilation mean position of particles. The former follows the true solution much better than the Gaussian distribution which provides a poor approximation of the distribution in each cell; this is consistent with the observations made in the preceding section. The use of a Gaussian distribution artificially re-introduces information which conflicts with the assumption made to perform annihilation, namely that all particles in a cell are considered to be indistinguishable regardless of their position. The uniform distribution best reflects this state of information.

Regenerating particles over an area corresponding to one annihilation cell, centred on the mean position of all particles (both positive and negative) before the annihilation, is justified by the observation that the area of the cell that is highly populated has better statistics (lower variance). Therefore, the variance is minimized in some sense by generating particles across the area of a cell centred at the pre-annihilation mean, instead of uniformly across the cell (i.e. a mean forced to be at the centre of the cell).

In summary, a reduction of the spatial resolution of the phase space grid used for annihilation reduces the memory requirements. By fitting the pre-annihilation spatial distribution of particles in a cell to a statistical distribution, the loss in resolution can be mitigated. Under the assumptions made for annihilation the uniform distribution is best-suited; other common statistical distributions, like Gaussian distributions, are ill-suited for the fitting and require some extra memory and computation to calculate the additional moment of the distribution (the standard deviation).

4.3.3 Ensemble sorting

The representation of the phase-space as an array to record the signs of particles is a direct reflection of the physical concept underlying particle annihilation. However, the annihilation concept can be also realized with an algorithm which avoids representing the phase space by an array, thereby completely avoiding the huge memory demands associated with it. This novel algorithm presents a significant contribution to making WMC simulations computationally more accessible and is presented in the following.

4.3.3.1 Algorithm

An integer index can be associated to the position and momentum attributed to each particle. These indices are mapped to a single integer H, uniquely identifying the cell of the phase space in which the particle resides:

$$(i_x, j_y, q_x, q_y, q_z) \rightarrow H. \tag{4.15}$$

All the particles with the same value of H are in the same cell of the phase space and their signs must be accumulated. To search an ensemble consisting of N particles, to find particles with matching values of H, requires an algorithm with an $\mathcal{O}\left(N^2\right)$ time complexity. As N can be several millions, the additional computation time is not tolerable.

The situation is greatly improved by first sorting the array, representing the particle ensemble, according to the values H. This can be efficiently performed by a quicksort algorithm which has a $\mathcal{O}\left(N \log_2 N\right)$ time complexity [141]. In the sorted particle ensemble, all particles in the same cell of the phase space (i.e. the same value of H) now appear consecutively in the array – this is also beneficial for the memory access speed. The sorted array allows the sum of signs and mean positions of the particles to be calculated in-place without the need of any additional memory. A flow-chart of
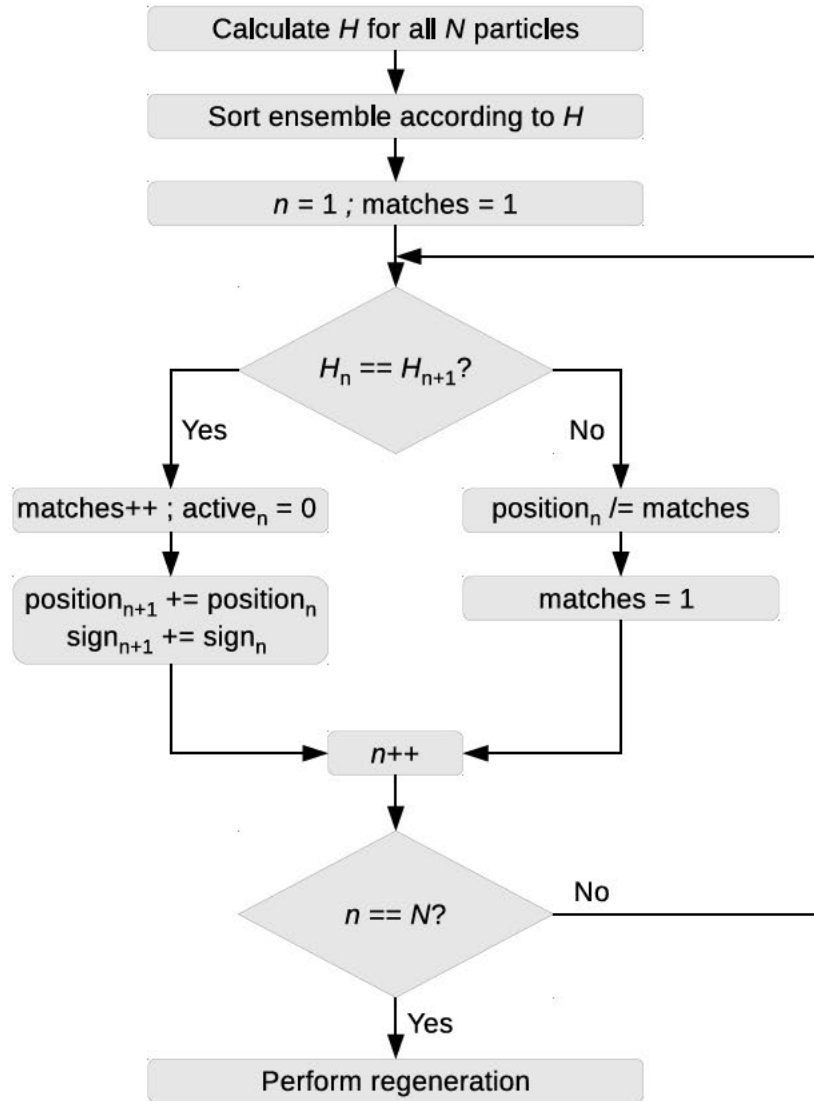
Figure 4.17: Flow-chart of annihilation algorithm based on ensemble sorting.

the algorithm is shown in Figure 4.17. The sorted array is iterated, if the value of H for particle n and $n+1$ are identical, a 'matches' counter is incremented, particle n is deactivated (by a flag) and its sign and position is added to the values of particle $n+1$. This process continues until $H_n \neq H_{n+1}$, then the mean position of all the particles in the cell corresponding to H_n is calculated by dividing the sum of the positions by the counter which is then reset.

Once the entire ensemble has been covered, the regeneration process commences. The information required for regeneration is stored in the fields for position and sign for particles which are still marked active ($active_n = 1$; $cf.$ Figure 4.17). The new particles are regenerated and stored in the original array, overwriting the particles in the array, which have been deactivated during the annihilation process. This allows the entire annihilation and regeneration process to be completed with an insignificant amount of additional memory.

4.3.3.2 Performance

The trade-off for the small memory footprint of this annihilation algorithm is the additional computation time required to sort the array of the particle ensemble. The calculation of the indices and the regeneration of particles is essentially identical between all the annihilation algorithms presented here.

Table 4.3: Simulation parameters of benchmark examples for the annihilation algorithm

	A.1	A.2	B.1	B.2
Max. ensemble size [$\times 10^6$]	20	10	5	5
Annihilations performed	23	210	24	33
Simulation time [fs]	100	100	100	100
Domain [nm]	100×150	100×150	100	100
$\Delta x \, (= \Delta y)$ [nm]	1	1	1	1
Coherence length L [nm]	30	30	30	60

Table 4.4: Simulation times for different annihilation algorithms

	Reference [s]	Sort-based [s]	Change
A.1	2041	2425	+19%
A.2	1717	2028	+18%
B.1	427	455	+7%
B.2	508	548	+8%

The impact of the sorting on the overall computation time depends on i) the regularity of the annihilation (the generation rate) and ii) the threshold value for the number of particles in the ensemble (N) at which annihilation (sorting) occurs. To characterize the performance impact of the algorithm, the simulation time of two examples are compared with different parameters, as listed in Table 4.3.

The results in Table 4.4 reveal the sorting-based annihilation algorithm increases the computation time between 7% and 19%. In the investigated examples, the peak memory demand of the simulation is not strongly affected by the type of annihilation algorithm, since the chosen resolution of the phase space is coarse and, therefore, the representation of the phase space in memory is small. Indeed, if sufficient memory is available, there are no advantages to using the sort-based algorithm. However, it should be noted that in certain practical cases – as exemplified in Chapter 5 – the memory demands of the conventional annihilation algorithm exceed the capacities of a workstation. Therefore, this algorithm makes two-dimensional Wigner Monte Carlo simulations accessible to users that have limited computational resources. This presents a great step towards a wider adoption of Wigner Monte Carlo simulations.

In summary, the annihilation algorithm based on ensemble sorting eradicates the memory demands associated with particle annihilation almost entirely. The trade-off is a slight increase in computation time, which depends on the parameters of the particular simulation problem. The sort-based annihilation algorithm allows two-dimensional Wigner Monte Carlo simulations to be performed on conventional workstations with limited memory resources, also when using high-resolution meshes, which significantly improves the accessibility of multi-dimensional time-dependent quantum transport simulations.

4.3.4 Particle growth prediction

The concept of particle annihilation relies on some approximations and, as has transpired from Section 4.3.1, the process may also introduce undesired numerical side-effects. It is therefore desirable not to perform the annihilation step unless needed (to ensure the total number of particles do not exceed the chosen maximum number).

The conventional approach is to choose (guess) an annihilation frequency (as a multiple of the time step Δt) such that the ensemble maximum is never exceeded. This approach requires some trial-and-error in choosing the most appropriate value, which is very unattractive from a usability point of view. Moreover, a fixed frequency of the annihilation may induce annihilation when it is not needed.

A superior manner to handle the problem is to induce an annihilation only when needed. For this purpose the increase in the number of particles, due to generation (Section 3.7.5), is predicted using

the current number of particles and the generation rate (which is known from the WP):

$$N_{t+\Delta t} = N_t \left(1 + \sum_{i=1}^{N_t} \gamma \left(\mathbf{r}_i \right) \Delta t \right), \tag{4.16}$$

where $\mathbf{r}_i$ in this case refers to the position of the i-th particle.

Since the particle generation is a stochastic process (which depends on the generation of random numbers), the exact number of generated particles can only be estimated. The estimate improves as the size of the particle ensemble increases. However, smaller ensemble sizes may occur, e.g. in subdomains of the spatial domain decomposition in parallelized code (discussed in Chapter 5), and then the estimate can deviate considerably from the actual change. Depending on the specific implementation of the algorithm this can be problematic, e.g. when allocating memory in accordance to the growth prediction. It is advisable to overestimate the particle increase. For this reason the reference implementation in ViennaWD uses only the maximum value of γ for all particles, such that

$$N_{t+\Delta t} = N_t \left(1 + \max_i \gamma \left(\mathbf{r}_i \right) \Delta t \right). \tag{4.17}$$

This presents an upper bound on the particle growth, which would only occur if all particles are in the region where the generation rate is at its maximum. A further advantage is that no summation has to be performed, allowing a faster calculation of the prediction.

4.4 Validation

The improvements which have been made to the algorithms of the signed-particle method, presented in this chapter, have increased the accuracy of WMC simulations. To illustrate this fact (and for validation purposes) a comparison is made here between the exact solution of the time-dependent Schrödinger equation and results obtained by the WMC method.

A general, exact solution of the time-dependent Schrödinger equation can be calculated by

$$\psi \left(x, t \right) = \int dx' K \left(x, t; x', t_0 \right) \psi \left(x', t_0 \right), \tag{4.18}$$

where $K \left(\cdot \right)$ is the propagator function associated to the stationary Hamiltonian describing the system and $\psi \left(\cdot, t_0 \right)$ is the initial condition of the wave function.

As a benchmark problem a wavepacket travelling towards a square potential barrier, to its right, within a closed system is considered here. The minimum uncertainty wavepacket, which serves as an initial condition, is defined as

$$\psi \left(x, t_0 \right) = \left(2\pi\sigma^2 \right)^{-1/4} e^{-\frac{(x-x_0)^2}{4\sigma^2}} e^{ik_0 z}, \tag{4.19}$$

with parameters given in Table 4.5. The potential considered here (square barrier) gives rise to analytic expressions for the propagator [142], allowing an exact solution of the Schrödinger equation to be obtained by a numerical integration of (4.18). This exact solution avoids approximations associated to numerical treatments and serves as a reliable basis for validation of the results obtained by the WMC simulator.

The Wigner transform is applied to (4.19) to obtain the corresponding Wigner function, which serves as an initial condition for the WMC simulation:

$$f \left(x, m, t_0 \right) = \frac{1}{\pi} e^{-\frac{(x-x_0)^2}{2\sigma^2}} e^{-(m\Delta k - k_0)^2 2\sigma^2}. \tag{4.20}$$

Table 4.5: Simulation parameters for validation example

x_0 [nm]	σ [nm]	L_{coh} [nm]	k_0 [nm^{-1}]	Δx [nm]
-29.5	10	100	$12\Delta k$	0.1
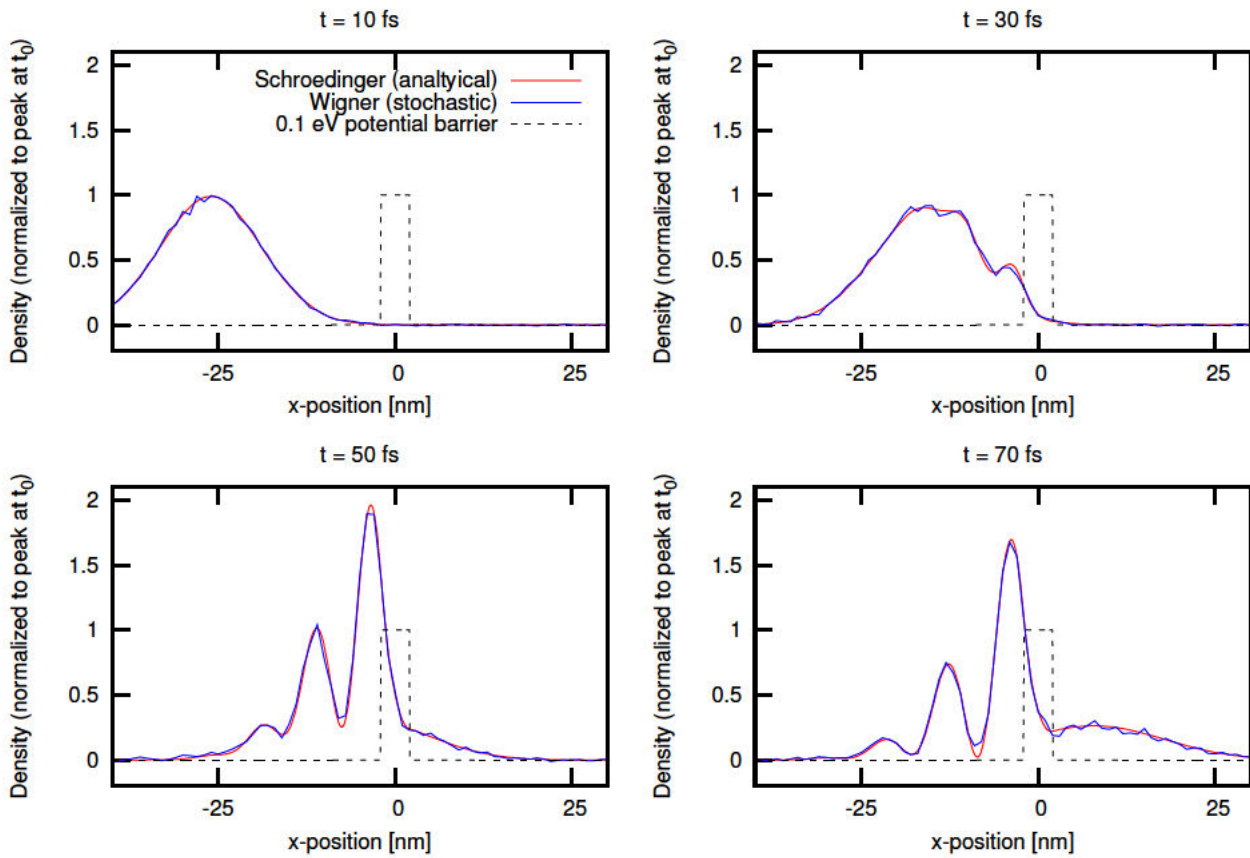
Figure 4.18: Charge density at various time steps for a wavepacket (mean energy of 0.067 eV) approaching a 4 nm wide, 0.1 eV high barrier.
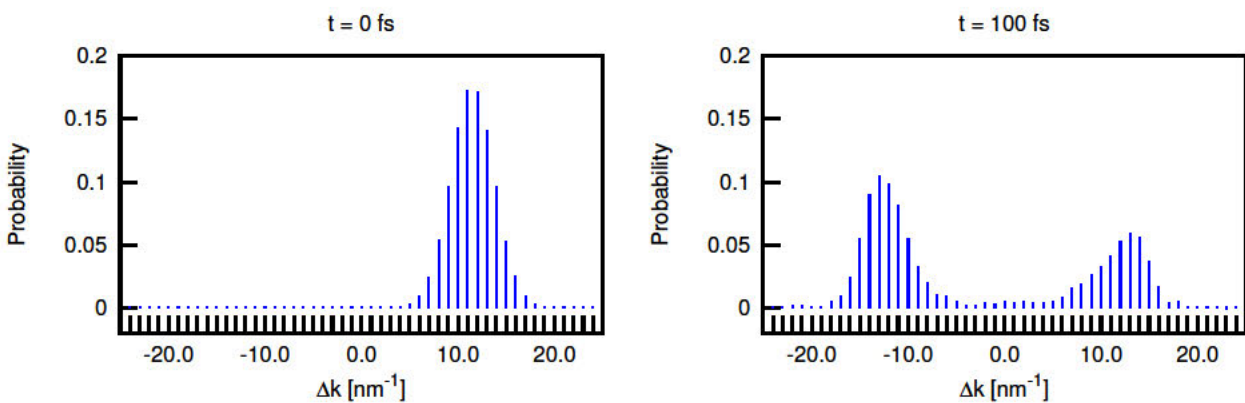
Figure 4.19: The k-distributions, corresponding to Figure 4.18, at 0 fs and 100 fs showing the reflection and transmission from/through the barrier of a wavepacket initially centred around $12\Delta k$.
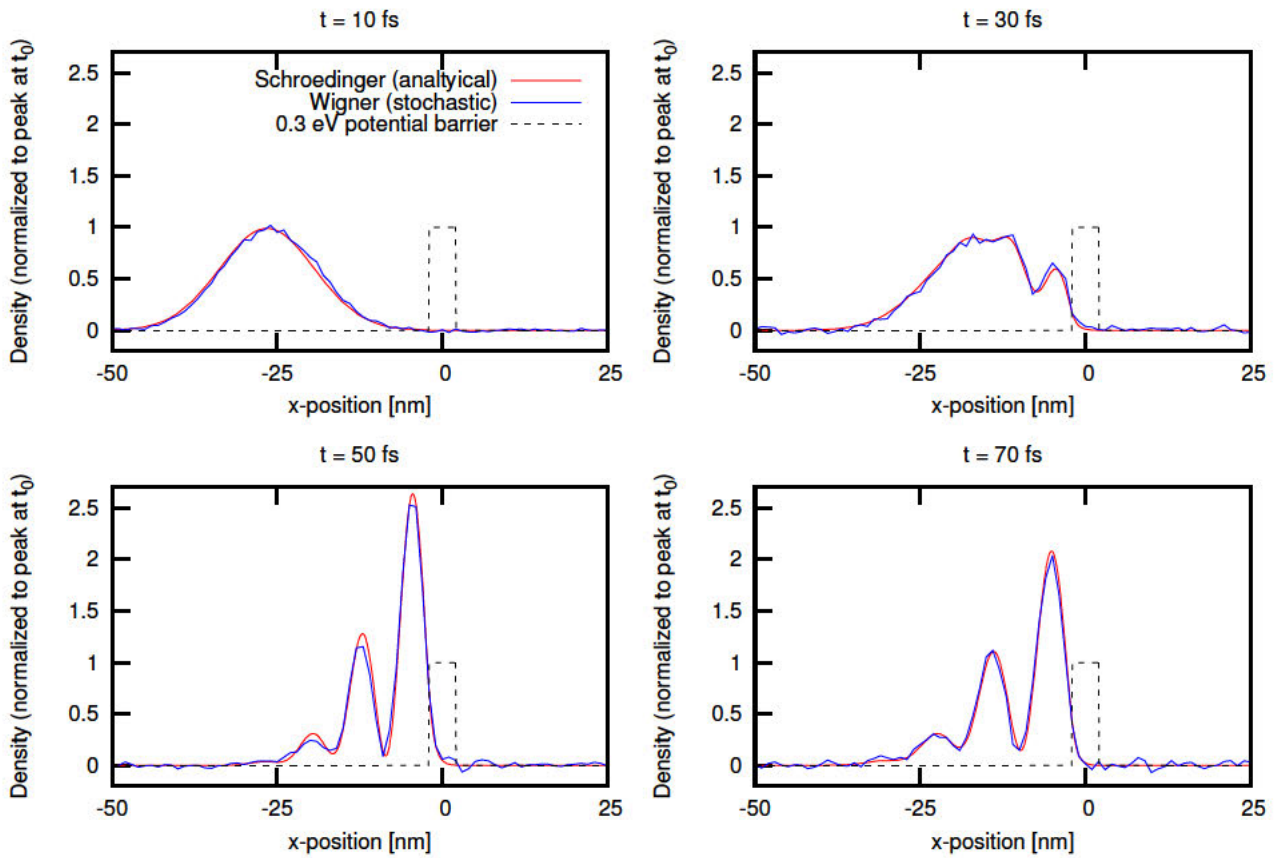
Figure 4.20: Charge density at various time steps for a wavepacket (mean energy of $0.067\,\mathrm{eV}$) approaching a $4\,\mathrm{nm}$ wide, $0.3\,\mathrm{eV}$ high barrier.
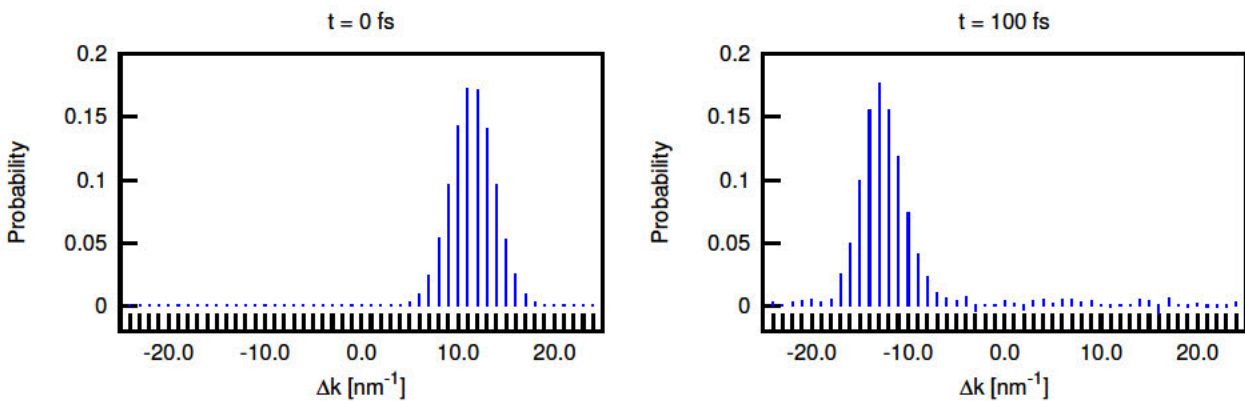
Figure 4.21: The k-distributions, corresponding to Figure 4.20, at $0\,\mathrm{fs}$ and $100\,\mathrm{fs}$, showing almost complete reflection from the barrier.

Figure 4.18 compares the solution obtained by the WMC method and the exact solution of the corresponding Schrödinger equation, over a time sequence, for a 4 nm wide, 0.1 eV barrier – the mean energy of the wave package is 0.067 eV. The transmitted and reflected components of the wavepacket are evident, as supported by the k-distributions in Figure 4.19. Previous comparisons between results obtained by the Schrödinger equation and Wigner Monte Carlo simulations have failed to show a truly quantitative match [143]. Here, an excellent match between the stochastic solution of the Wigner equation and an exact solution of the Schrödinger equation is evident (as was first reported in [144]) thanks to the optimized algorithms and considerations presented in this chapter. The Monte Carlo solution shows some noise, due to the stochastic nature of the method, especially the particle generation process.

Figure 4.20 and Figure 4.21 show the result of the same wave package approaching a 0.3 eV barrier, which leads to almost complete reflection. The simulation required a fine spatial resolution (0.1 nm) to appropriately represent the sharp edges of the square barrier along with an appropriately chosen coherence length (*cf.* Table 4.5).

Unlike for the discussed analytical method, the ability to include an arbitrary, time-dependent potential in the WMC simulator is a big advantage. This, however, requires the WP, V_w, to be recomputed at each time step. The computational cost of the latter can be high (especially in higher dimensions), but can be significantly reduced by using the specialised, box discrete Fourier transform (discussed in Section 4.1.2), which exploits the correlation between the values of V_w amongst adjacent nodes. As a validation of the WMC results, incorporating a time-dependent potential, the potential barrier is made to rapidly oscillate between 0.04 eV and 0.1 eV with a period of 20 fs, and the result is compared to the bounds set by the exact solutions of the limiting (static) cases in the spatial domain, as illustrated in Figure 4.22. Figure 4.23 shows the k-distributions of the stochastic solutions for the dynamic potential and the two limiting, static potentials – the solution for the oscillating barrier remains within the set static bounds. Moreover, a reflection of the wavepacket (0.067 eV) against the static 0.04 eV barrier can be identified, which is consistent with the expected quantum behaviour.

The increased accuracy is attributed primarily to improved generation statistics (Section 4.2) and the use of a sufficient wavevector resolution (through the choice of the coherence length). The increase in the computational demands of a large coherence length can be significant in 2D simulations but can be alleviated by the algorithms presented in Sections 4.3 and 4.1.2.

The presented benchmark tests show that the WMC method has been matured to the point of providing highly accurate results, thanks to appropriate handling of the particle-generation statistics (Section 4.2) and the appropriate choice for the mesh resolution. The presented considerations conceptually extend to higher dimensions, thereby paving the way for the accurate numerical analysis of mesoscopic semiconductor devices using the Wigner formalism.
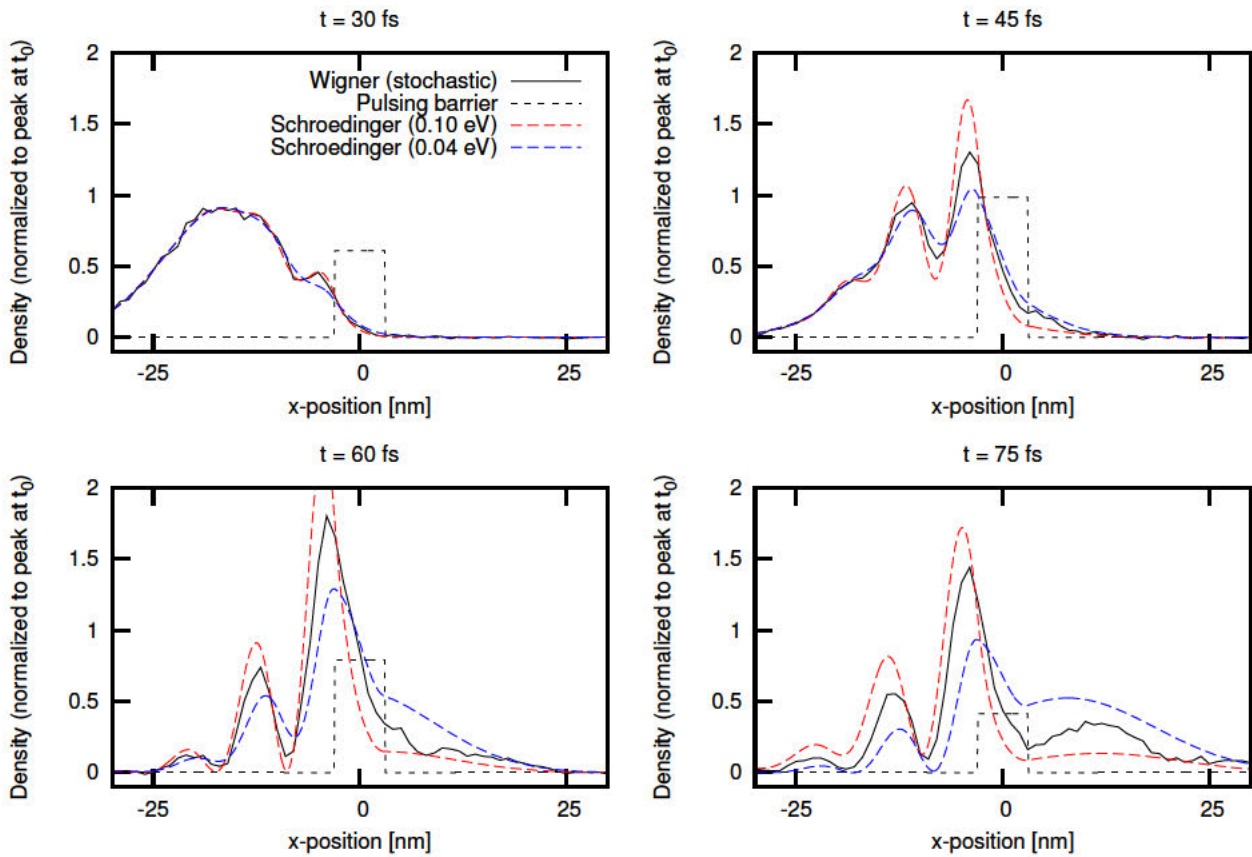
Figure 4.22: Charge density at various time steps for a 6 nm wide barrier oscillating between 0.04 eV and 0.1 eV. The simulated solution (Wigner) with a dynamic potential remains between the bounds given by exact solutions of the Schrödinger equation for the static, limiting cases.
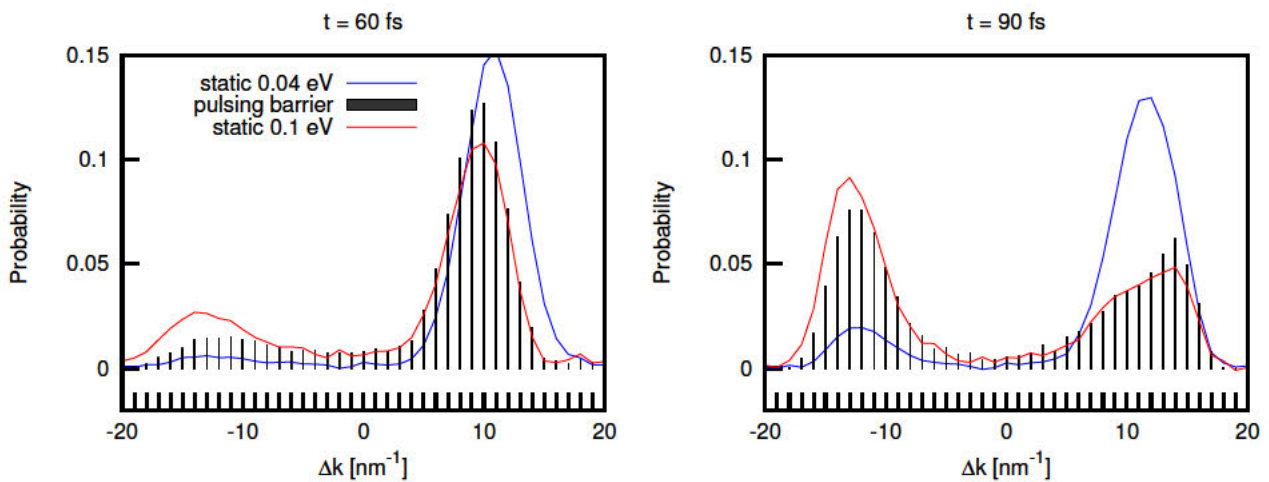

Figure 4.23: Comparison of the k-distributions (obtained by WMC) of the pulsating barrier and the static barriers of the limiting cases, as in Figure 4.22.

Chapter 5

Parallelization of the Wigner Monte Carlo Simulator

This chapter first outlines the possibilities to parallelize the Wigner Ensemble Monte Carlo (WEMC) algorithm to motivate the parallelization approach. Thereafter, a detailed discussion of the spatial domain decomposition approach is given, along with results demonstrating its parallel scaling performance in one- and two-dimensional simulations.

5.1 Background

Scientific computation has become an integral part in most fields of modern research. Apart from the derivation and analysis of the models underlying the computation – as has been done in Chapters 3 and 4 for the WEMC method – the feasibility and optimization of the numerical calculations is paramount to obtain simulation results in a reasonable amount of time. Parallel computation is now the primary method[1] to speed up, or even enable, computations on systems ranging from single workstations with multi-core CPUs to large supercomputers consisting of thousands of nodes[2]. An efficient utilization of high performance computation resources requires consideration to be given to the architecture of the system on both a physical (hardware) and a logical (software) level. The most common system architectures for high-performance computing are outlined in the following.

5.1.1 Shared memory

A shared-memory system refers to an architecture in which multiple CPUs share a common memory space, i.e. every CPU can access the entire memory. The simplest scenario is a single workstation with a multi-core CPU where each core has access the same memory space. In terms of hardware, a shared-memory architecture is usually limited to a single computation node, which limits the total number of CPUs and memory which can be used. Although there exist supercomputers that facilitate a (logical) shared memory space distributed between many computation nodes, they are not widespread.

The advantage of a shared memory space is that no replication of data is required and data can be accessed/exchanged between different CPUs with little to no additional communication latency. The disadvantage is that special care must be taken to avoid race conditions and that the computations of one CPU do not inadvertently affect the data/variables another CPU is using – the computations are not isolated.

The parallelization of code in a shared-memory architecture usually makes use of computation threads, which can be implemented with OpenMP [146], for instance. This allows a simple parallelization of for-loops and other control statements commonly used in programming.

---

[1] In the course of the 2000's the increase in clock frequencies of CPUs, which ensured a direct scaling in performance, gradually stopped [145] due to limitations imposed by thermal power dissipation.

[2] In this context a node refers to a computer, consisting of several CPUs (cores), which is part of a larger cluster.

5.1.2 Distributed memory

A distributed memory system refers to an architecture where each CPU has its own dedicated memory space, which cannot be directly accessed by other CPUs. Distributed memory systems are most common in large-scale supercomputers, which consist of multiple computation nodes connected by network interfaces. A distributed memory environment can also be emulated on a single workstation, where the same physical memory is logically separated.

The advantage of a distributed memory system is that vast computational resources can be utilized to solve problems which would otherwise be computationally intractable on a single node/workstation. The disadvantage is that the network interfaces are very slow (compared to the CPU-RAM bus) and communication to access data on other nodes is accompanied by a latency which can be a significant factor in the performance of parallel code.

The parallelization of code for distributed memory systems requires special consideration for the data structures and the timing of data communication. The communication of data between processes[3] is commonly handled using the message passing interface (MPI) [147]. The programmer is forced to explicitly consider these aspects in the design of the algorithms, which makes the code more robust for scaling up to many CPUs.

5.1.3 Hybrid systems

The shared- and distributed-memory approaches can be combined to what is referred to as a hybrid system. One can, for instance, use parallelization by threads on a single node (shared memory) and use an MPI communication between the nodes (distributed memory). The optimal combination depends on the computational problem and the system on which it is run.

5.1.4 Accelerator cards

Accelerator cards are enjoying increased popularity in scientific computing as they offer significant raw computation power at low cost/energy. Examples of accelerator cards are dedicated graphical adapters, like the nVidia Tesla, or co-processors, like the Intel Xeon Phi. Accelerators are especially useful for problems where many calculations have to be performed with relatively little data; communication to obtain data is slow due to the limited bandwidth of the bus connecting the cards to the RAM.

Various frameworks are available for programming with accelerator cards in mind, like OpenCL.

5.2 Parallelization approaches for Monte Carlo simulation

Monte Carlo code is often termed *embarrassingly parallel* as a high parallel efficiency is easily achieved in general: If the particles of an ensemble are independent of each other, smaller subensembles can be handled by separate computational units[4] without the need for further communication. The results of the subensembles are later merged to yield the global solution. An independent computation necessitates that the entire simulation domain is available on each computational unit.

The parallelization of the Wigner Monte Carlo code, however, is complicated by the annihilation step, which hinders the independent treatment of subensembles for two reasons: i) The annihilation step must be performed on the entire (global) ensemble of particles since (here) the subensembles are not regarded to be big enough to be statistically representative[5]. The latter necessitates some communication and/or synchronization between the computational units.

The second obstacle the annihilation step presents to parallelization is ii) the exorbitant memory demands of the annihilation algorithm when treating higher-dimensional problems. Although the alternative sort-based annihilation algorithm (Section 4.3.3) effectively avoids this problem, the faster,

---

[3]A process, in the context of MPI, refers to a logical computation entity. A process can be associated with a single core/CPU/node.

[4]A computational unit in this general context can be e.g. a separate computer(s), a core of a CPU or a graphics card.

[5]If a subensemble is big enough to yield a statistically representative solution to the simulation task, the 'parallelization' simply amounts to a simultaneous repetition of the same experiment on different computational units, the results of which are averaged.

conventional algorithm is considered here, in the spirit of high-performance computing this chapter follows. The annihilation step requires the phase-space to be represented in the memory using an array of integers, each storing the sum of particle signs inside one cell $((\Delta r)^d (\Delta k)^{d'})$ of the phase space grid. While for one-dimensional simulations the memory footprint of this array remains small, the memory consumption grows rapidly for higher-dimensional simulations. Consider a 2D spatial domain of $100\,\text{nm} \times 100\,\text{nm}$ with a resolution of $\Delta x = 1\,\text{nm}$ and a 3D k-space with 100 k-values per direction. The associated phase-space grid would consist of $100^2 \times 100^3$ cells, each represented by an integer of (at least) 2 bytes. This would demand a total memory consumption of $\mathcal{O}\left(2^{10}\right)$ bytes, i.e. approximately $20\,\text{GB}$. A higher resolution increases the memory demands dramatically because they grow to the power of $(d + d')$, as discussed in Section 4.3.2.

The need for synchronization/communication amongst subensembles and the high memory demands of the annihilation algorithm make the parallelization of a Wigner Monte Carlo code much more challenging than other Monte Carlo codes. Possible parallelization approaches are evaluated in the following with reference to the architectures introduced in Section 5.1.

5.2.1   Shared-memory

In a shared-memory setting the particle ensemble is partitioned amongst the threads and only a single instance of the simulation (the domain and all subensembles) exists in memory and all threads have shared access to it. The communication required to perform the annihilation on the global ensemble is thereby avoided to a major extent, but a synchronization amongst the threads is still needed. By using appropriate parallel loop-scheduling techniques the computation load can be well balanced amongst the threads, which ensures that no thread is left idle for long periods of time before the annihilation is performed on the global ensemble.

Although simple to implement, a pure shared-memory approach is confined to a single computation node with a limited number of CPU cores and memory. The latter restricts the simulation problems that can be investigated, based on acceptable run-times and the memory demands of the simulation. Especially the high-memory demands of the annihilation algorithm limit the simulations to large-memory nodes[6], which are less prevalent in shared computation facilities.

Therefore, a pure shared-memory approach is best suited for small-scale parallelization cases.

5.2.2   Distributed-memory

A large-scale, MPI-based parallelization approach, is not restricted by the computational resources of a single node, thereby considerably expanding the scope of the simulations, which can be handled from a computational point of view. The particle ensemble is split into many subensembles, each of which is assigned to a separate MPI process[7] for computation. It is more challenging to dynamically react to computational load imbalances amongst processes, compared to the loop-scheduling techniques for threads available with OpenMP. However, physical insight into the problem and experience can help mitigate this disadvantage.

5.2.2.1   Domain replication

Since the processes do not share a single memory space, the simulation environment must be replicated for each process; this is referred to as domain replication. Due to the high memory demands of the annihilation algorithm domain replication – as is common for classical Monte Carlo simulation – is problematic in a distributed-memory environment: Today's large-scale clusters typically provide between $2 - 4\,\text{GB}$ of memory per CPU core. If each process is assigned to one CPU core – this is desirable for an optimal utilization of the computational resources – the memory available to each process ($2 - 4\,\text{GB}$) can be insufficient for a complete representation of the phase-space array in multi-dimensional problems ($20\,\text{GB}$ in the example presented before).

---

[6]In supercomputers large-memory nodes refer to nodes that contain significantly more memory than the common nodes, but are less prevalent.

[7]For the remainder of this work, the term *process* refers to an *MPI process*.

One possibility to circumvent this problem is using a large-memory node to perform the annihilation of the global particle ensemble. In such a case, the subensemble of every process must be communicated to the master node at each time step, where they are all combined and the annihilation step is performed. If a sufficient number of worker processes are in operation, the communication bandwidth of the master process' node will quickly saturate – the worker processes remain idle while waiting for all other processes to complete their communication and, thereafter, for the annihilation step to be completed. The post-annihilation particle ensemble is split up again and distributed amongst the processes. All this communication severely impacts parallel efficiency. Indeed, without domain replication, achieving good parallel efficiencies becomes more challenging.

5.2.2.2   Domain decomposition

An alternative to domain replication is domain decomposition, which entails splitting up the simulation domain amongst the processes. Each process represents a subdomain (i.e. a part of the global domain) and only handles particles, which fall within its own subdomain. Thereby, the memory requirements to represent the (localized) phase-space, as well as all other space-dependent quantities, are scaled down with the number of processes (subdomains) used. This makes the approach very attractive to large-scale parallelization as it avoids the problems of the aforementioned approaches, i.e. a large memory footprint, centralized communication and limited scalability/accessibility.

In light of the above, the domain decomposition approach is found to be best-suited for a future-proof parallelization of the WEMC method and its implementation is discussed in the remainder of this chapter.

5.3   Domain decomposition for Wigner Monte Carlo simulator

The adoption of the domain decomposition approach requires consideration to be given to exactly how the domain decomposition is performed and how the process of annihilation is affected.

5.3.1   Domain decomposition

There are three design choices that must be made for the domain decomposition: i) the physical quantity to be decomposed, ii) the number of dimensions in this quantity to be decomposed and iii) the size of the subdomains. These three issues are discussed below.

5.3.1.1   Decomposed quantity

The decomposition of the domain (phase space) can be either according to the spatial position or the wavevector (or a combination of both). The WBE presents a transport problem where the particles move in the domain (phase space) and have to be passed between subdomains, which necessitates a regular communication between the processes representing the subdomains. Moreover, in the signed-particle method (Chapter 3) particles are scattered/generated to/with new wavevectors, which can differ considerably from the original value/generating particle.

A decomposition of the k-space is not attractive from a performance point of view, because it is very likely that the new wavevector of the generated/scattered particle lies in a part of the k-space which is represented on another process and requires the particles to be transferred to this process. Considering the high particle generation rate (*cf.* Section 3.7.5) the additional communication required would be debilitating for parallel efficiency.

A decomposition of the spatial domain avoids such problems, since the position of particles change in a continuous, gradual fashion as they propagate; the position of newly generated particles also corresponds to that of the generating particle. As the particles propagate through the domain they are passed between neighbouring subdomains as they cross their boundaries. It should be pointed out that the boundaries between the subdomains are only of a computational nature and inconsequential to the mathematical formulation of the evolution, presented in Chapter 2.

5.3.1.2 Dimension of decomposition

For two- or three-dimensional problems one can choose the number of spatial dimensions to be partitioned. In the 2D case one has the choice between slab- and block-decomposition, which correspond to partitioning in one and two directions, respectively. Which decomposition method is best-suited depends on the computational resources and computational demands of the investigated problem; the physical behaviour of the problem also plays a deciding role. The case for 2D WEMC simulations is investigated here.

Slab-decomposition: A slab-decomposition splits the domain along a single direction (Figure 5.1a). The advantage of this approach is that the transfer of particles only has to be handled on two boundaries, thereby the communication is limited to only two other processes, which is advantageous for the parallel efficiency. Slab-decomposition will work well in simulation problems where there is a large degree of uniformity in one direction, e.g. the flow of particles is predominantly in one direction. In such a case, the domain is partitioned in the direction orthogonal to the flow. A judicious choice is important to achieve good load-balancing.

Since each subdomain/slab has a minimum width (one mesh cell), the maximum number of subdivisions/subdomains is limited. This places a maximum on the number of processes which can be used. Furthermore, the 'thinner' the slabs are, the less particles can be accommodated in the subdomain – this gives rise to the situation where the computational load becomes too small in relation to the communication overhead incurred to transfer the particles between adjacent domains. Although these limits arise for all decomposition approaches, the slab-decomposition reaches these limits first. Nonetheless, the slab-decomposition method has been successfully applied to one- and two-dimensional Wigner Monte Carlo simulations [148, 149], as evidenced by the results in Section 5.5.

Block-decomposition: A partitioning in two spatial dimensions – a block decomposition – holds the promise to delay the onset of the limitations set out for the slab decomposition. Overall one attains greater granularity in the decomposition and can accommodate many more MPI processes. The number of communication links increases from 2 (slab-decomposition) to 8 per process since movement in the the diagonal directions must also be accounted for (Figure 5.1b). The additional communication channels which need to be set up with each time step introduce a significant overhead for the MPI communication back-end. Moreover, additional logic is required to identify to which subdomain particles have to be transferred to.

5.3.1.3 Subdomain size

The computational load a process must handle is proportional to the number of particles in its subdomain. The domain decomposition approach allows a process to only treat the particles that are physically located in its subdomain. This complicates the task of load-balancing, since in a transport problem, by definition, there is a non-uniform distribution of particles moving about in the domain.

The size of the subdomains can be chosen such that the particles are more equally distributed (on average over time) between the processes, but this requires some heuristics as the optimal decomposition will differ considerably between different simulation problems. The situation is complicated by the fact that particles are also generated non-uniformly across the domain.

A possibility to make an *a priori* estimate of the particle distribution is to use the particle generation rate (assuming a time-independent potential is used; see Section 3.7.5) to weigh the size of the subdomains. The reasoning is that the most particles will end up being in the regions where there is the highest generation rate. This is only an approximation based on the assumption of a homogeneous initial distribution of particles across the entire domain. It is important to note that the distribution of the numerical particles does not correspond to the physical density, which takes the sign of the numerical particles into account.

The issue of load-balancing will not be treated further here; the remaining discussion and presented results (Section 5.5) assume a uniform decomposition of the spatial domain (and achieve reasonable scaling nonetheless).
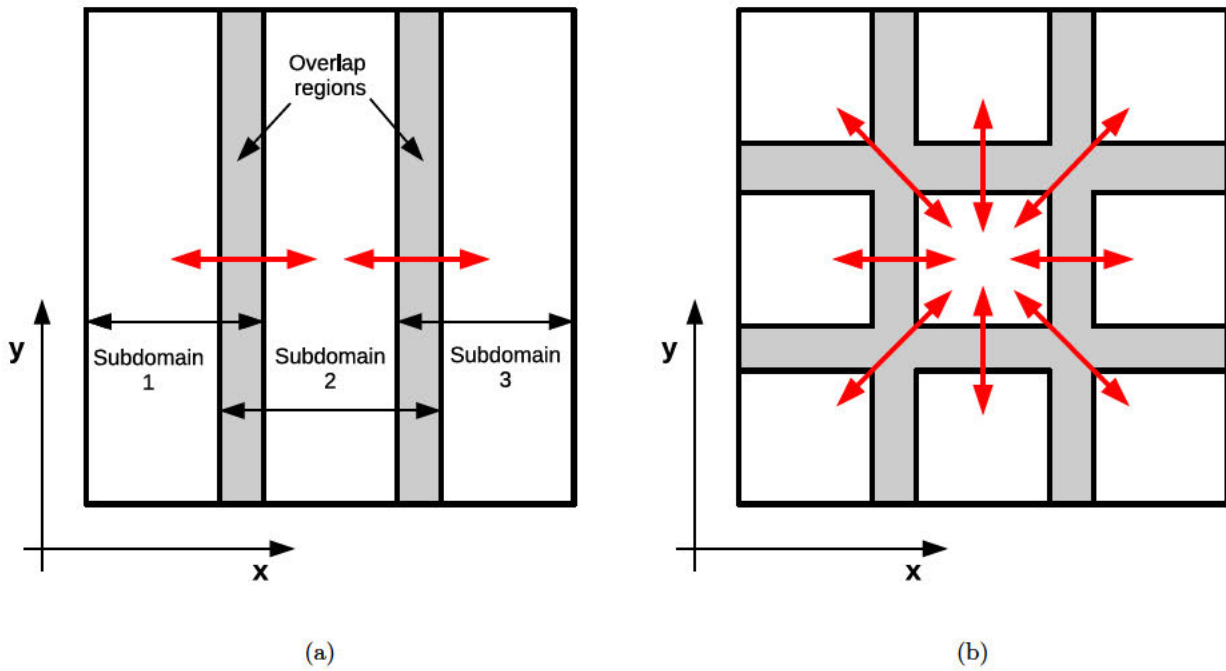
Figure 5.1: A uniform (a) slab- and (b) block-decomposition approach. Overlap areas shared by neighbouring subdomains are indicated in grey. The red arrows indicate the communication links needed to transmit particles between subdomains.

5.3.2 Localized annihilation

Due to the domain decomposition, the portion of the phase-space associated with each process can be represented in the memory typically available for a single process $(2-4\,\text{GB})$. The annihilation step can be performed locally by each process for the particles in its subdomain – it is not required to perform the annihilation on a single process/node for the global ensemble. The only additional requirement is that the annihilation step must be performed amongst all the processes at a certain time step, i.e. if one process requires an annihilation step – as determined by its local growth prediction (Section 4.3.4) – all other processes should perform an annihilation, irrespective of whether it is required locally (according to the local growth prediction). This approach ensures that the global statistics (Wigner function) are not falsified. If an annihilation step would be performed in one sub-domain but not in the others, the number of particles representing the local distribution is reduced and thereby also its statistical weighting in the global ensemble.

5.4 Algorithm

This section discusses the implementation of the domain decomposition approach, presented in Section 5.3, using MPI. The discussion is made against the backdrop of the serial algorithm, which has been presented in Chapter 3, and focuses on the additions/changes needed for the parallelization.

5.4.1 MPI topology

Consider N MPI processes – a master process (process 0) with worker processes (Process 1..$(N-1)$) – each assigned to one CPU core. The spatial domain considered in the simulation, i.e. the dimensions of the structure/device, is divided into N uniformly sized subdomains, one for each process. A slab-decomposition is used here, as illustrated in Figure 5.1a, but the concepts extend to other space decompositions or higher dimensions. The subdomains are assigned to processes in a sequential order, thereby inherently allowing each process to 'locate' the processes treating its spatially neighbouring subdomains, e.g. Process 2 would be responsible for the subdomain to the left of the subdomain handled by Process 3, etc.

The MPI communication takes place between each process and its spatially neighbouring subdomains, along with some minimal communication (one character (the flag) per time step) to the master process for coordination of the annihilation step. Such a decentralized approach avoids a constant querying of the master process, which – due to increased latency and bandwidth limitations – would impede scaling for increasing numbers of processes. The transfer (communication) of particles between processes only occurs once at the end of each time-step. This necessitates a small overlap between adjacent subdomains, which serves as buffer to accommodate particles travelling towards a neighbouring subdomain, until they get transferred to the subdomain at the end of the time step. The exact extent of the overlap should consider the maximum distance a particle can travel within the chosen time-step as well as its direction of travel. It is desirable to make the overlap between the subdomains as small as possible to avoid data redundancy which negatively affects the parallel efficiency (in terms of memory scaling).

5.4.2 Initialization

As illustrated in Figure 5.2, the master process performs the initialization of the simulation environment, which entails receiving external input data (just like in the serial case), performing the discussed domain decomposition and finally communicating this data to the worker processes.

The initial condition for the simulation is given by an (arbitrary) ensemble of particles, which is distributed by the master process amongst the various worker processes by assigning each particle to an appropriate subdomain based on its position. The particles associated to each subdomain are first collected and then communicated to the associated worker process by the master process. Furthermore, the master process broadcasts the potential profile and global parameters, needed for a localized simulation setup, to all worker processes.

After receiving setup parameters and its initial particles ensemble (in case of worker processes), each process initializes localized versions of the required data structures, specific to its subdomain. Thereby, the memory demands of each process scale down with the number of processes/subdomains. Moreover, the localization of the WP allows its computation to be distributed amongst the processes, which is beneficial when problems with time-dependent potentials are considered.

5.4.3 Time loop

After the initialization phase, each process performs the evolution of its ensemble of particles for a single time-step – this is identical to the serial case discussed in Section 3.7. After the time-step is completed, each process performs a growth prediction for its subensemble of particles, the result of which is communicated to the master process in the form of an annihilation flag (1-byte character) in order to facilitate a synchronized annihilation amongst all processes. After the master process has received the flags from all worker processes, it broadcasts a global annihilation flag back to the worker processes. The global annihilation flag is true, if the annihilation flag of at least one process is true, otherwise it is false. The annihilation step ensues (or not) locally within each subdomain, depending on the global annihilation flag received. The communication step associated with the annihilation flag implicitly serves as a synchronization point between the processes, which is required anyway due to the need to transfer the boundary particles at the end of each time step. Therefore, communicating the annihilation flag does not impede parallel efficiency.

After the (possible) annihilation step, each process identifies the particles in its subdomain, which qualify for transfer to its adjacent subdomains. These particles are collected and sent to the appropriate process, which is implicitly known due to the sequential ordering discussed above. Likewise, particles are also received from the neighbour processes. This communication is non-blocking, however, a synchronization barrier is used to ensure all transfers are complete before the next time-step commences. Since the processes already will have been synchronized shortly before by the annihilation communication, and the fact that the execution time of the annihilation procedure does not vary significantly between the processes, this second synchronization is not as detrimental to the efficiency of the parallelization as it initially appears. The reason for performing the transfer after the annihilation, is that after an annihilation step the size of the particle ensemble will be significantly smaller,
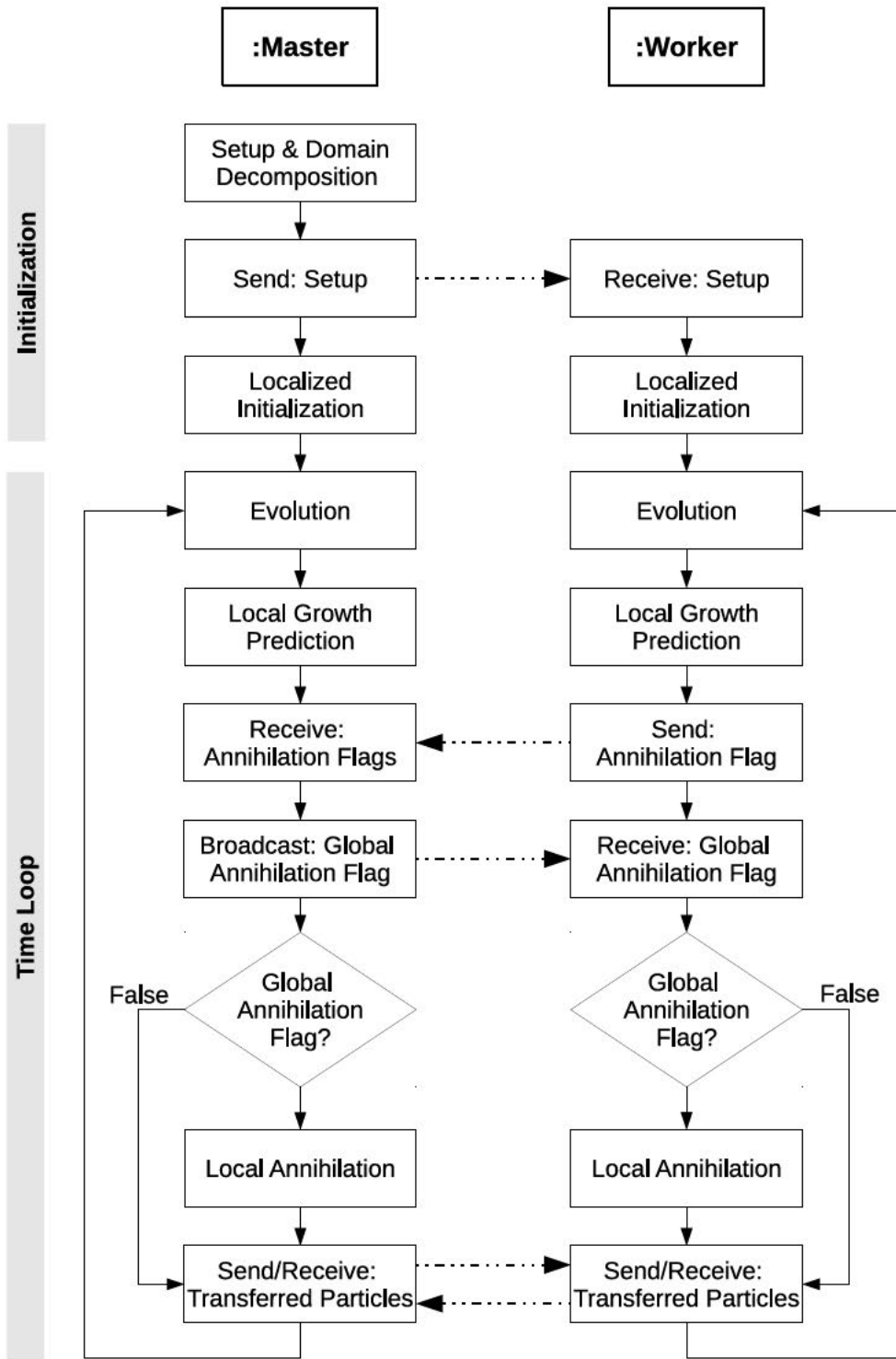
Figure 5.2: Flowchart illustrating the time sequence of the steps in the parallel algorithm performed by the master process and a worker process. The initialization and one step of the time-loop are shown. The dashed arrows indicate communication between the process (communication to other (possible) worker processes is not shown)

consequently the number of particles to be transferred will have been reduced.

This sequence of evolution, annihilation and transfer is repeated until the total simulation time has been reached. The simulation results of each process are written to disks locally by each process, which increases efficiency by avoiding a global reduction step issued by the master process. The simulation results are merged in a straightforward manner via a separate post-processing step (e.g. on a personal workstation) after the simulation has ended and the reserved computational resources have been released.

5.5 Evaluation

This section presents results obtained by the parallel algorithm introduced in the preceding section. First the parallelized algorithm is validated, whereafter its performance is evaluated with physical representative examples. The most important results obtained in [148] and [150], for 1D and 2D cases, are summarized here.

5.5.1 Validation

The spatial-decomposition approach must be validated to ensure that it yields the same results as the serial algorithm, regardless of the number of processes used, and does not introduce some (obvious) systematic errors, when the domain is split up.

Figure 5.3 shows the solution of a validation example – similar to the one described in Section 4.4 – for 16, 32 and 64 processes and compares it to the analytical solution. The parameters for the wavepacket and simulation are given in Table 5.1. An exact correspondence between the simulated results (within the bounds of the stochastic noise) is evident, irrespective of the number of processes used. This ensures the validity of the spatial domain decomposition approach. The slight deviation from the analytical solution – unlike the solution in Section 4.4 – is attributed to the much lower spatial resolution, which effectively gives the potential barrier a trapezoidal shape.

Table 5.1: Simulation parameters for validation example of parallelization scheme

| x_0 [nm]	σ [nm]	L_{coh} [nm]	k_0 [nm^{-1}]	Δx [nm]
40	7	100	$18\Delta k$	0.1

5.5.2 Performance

The parallel efficiency is evaluated based on the execution times of representative, physical examples in one and two dimensions. The simulations are first run with a single process, to acquire a baseline, and then repeated using 16, 32, 64 and 128 (2D case) processes. This procedure is repeated for different values for the maximum allowed ensemble size (8, 16 and 32 million particles). The maximum number of particles per process is scaled with the number of processes, e.g. a set maximum of 32 million particles for a simulation using 32 process, implies a maximum of 1 million particles per process. This scaling is necessary to allow a fair comparison. The execution time is recorded from the point where the master process starts the serial initialization and ends, when all process have completed the parallel time-loop (*cf.* Figure 5.2). All file output is disabled during the benchmarking.

The presented simulation results were obtained using (a part of) the VSC-3 supercomputer [151], which consists of 2020 nodes. Each node provides 16 cores (two 8-core Intel Xeon Ivy Bridge-EP E5-2650v2, 2.6 GHz, Hyperthreading) and 64 GB of system memory; the nodes are connected via an Intel QDR-80 dual-link high-speed InfiniBand fabric.

5.5.2.1 One-dimensional results

The parallel efficiency has been investigated in [148] at the hand of two examples presented in the following.
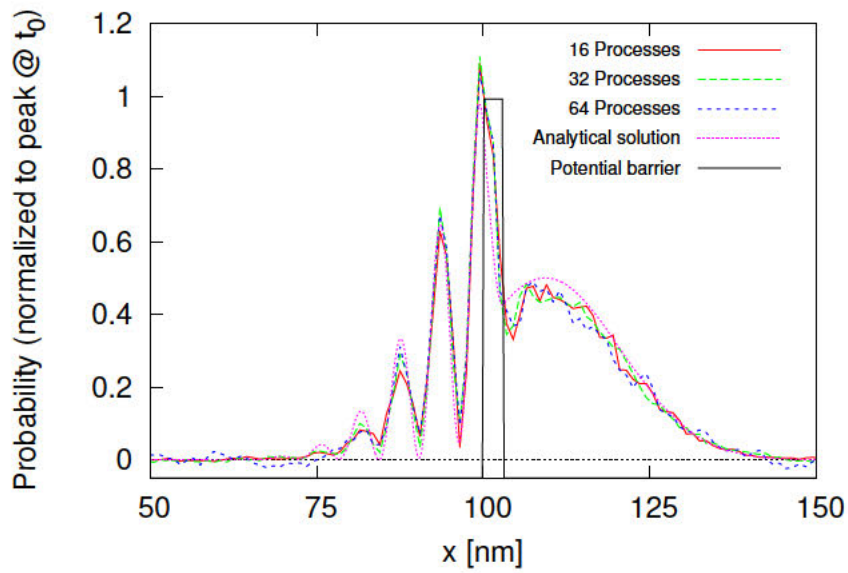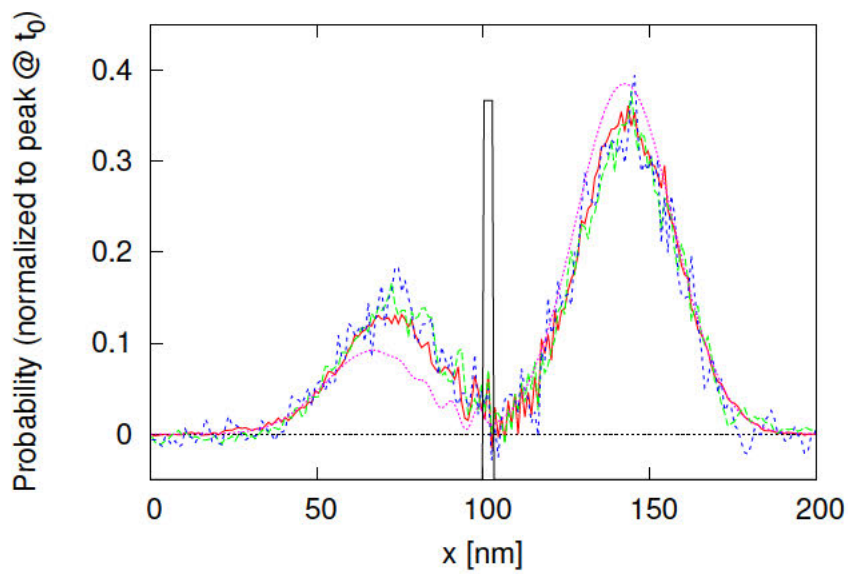
Figure 5.3: Comparison of the density of a wavepacket impinging on a 0.1 eV, 3 nm wide potential barrier for different numbers of MPI processes and an analytical solution after (a) 85 fs and (b) 125 fs.

Figure 5.4: Parallel speed-up and efficiency of the single-barrier problem for different maximum particle ensemble sizes
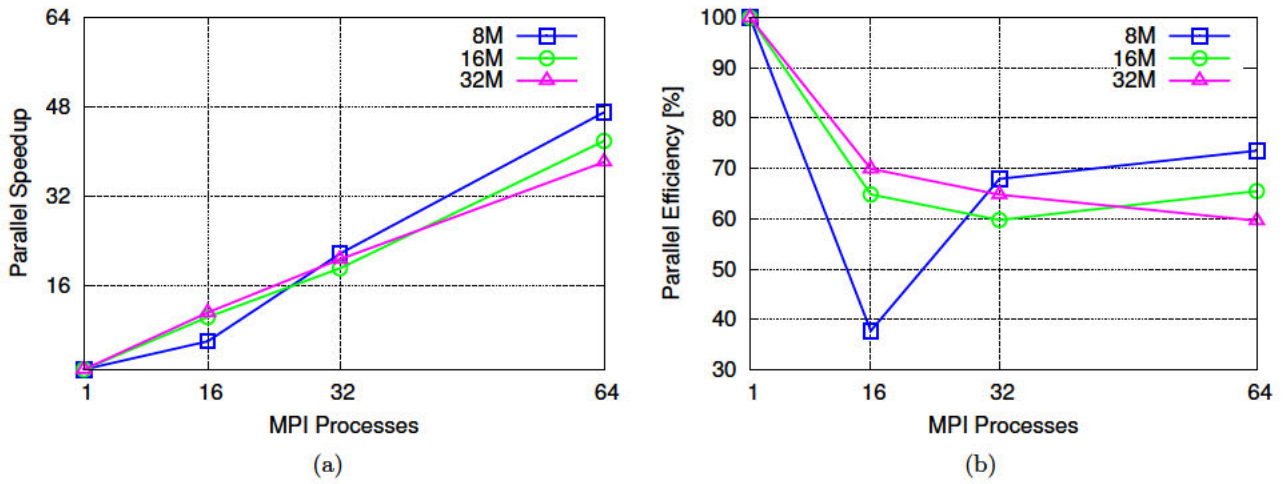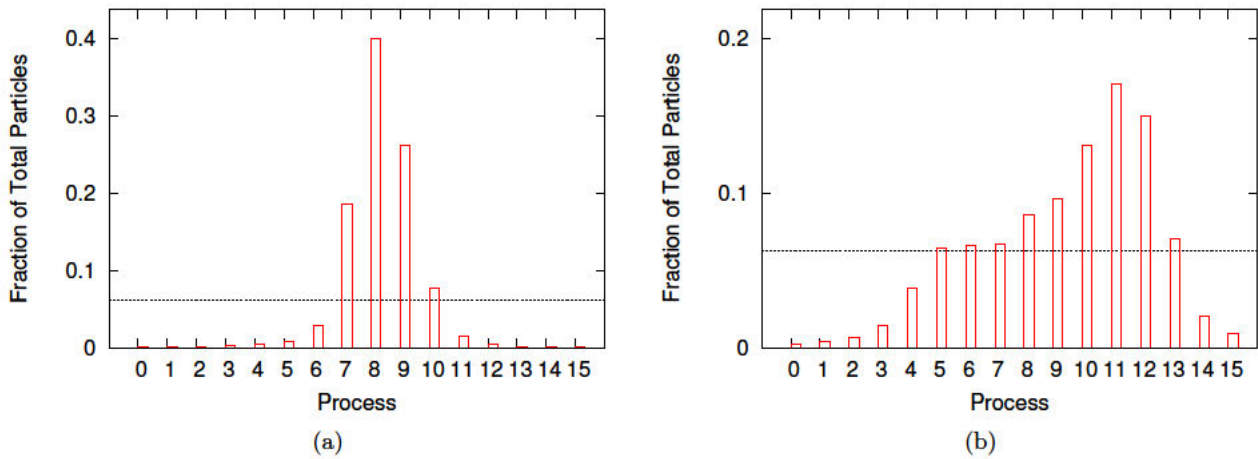

Figure 5.5: Distribution between 16 processes of the total number of numerical particles in the simulation representing the distribution of the computational load at (a)85 fs and (b)125 fs; the dashed horizontal line indicates the ideal load distribution.

Single potential barrier: The first benchmark problem is the square potential barrier as used for validation purposes before. The parallel scaling is shown in Figure 5.4. A parallel efficiency of at least 60% is achieved for all cases. The only outlier is the case with a global particle maximum of 8 million particles, which shows a big jump in efficiency from 16 to 32 processes. This can be attributed to the annihilation process, which uses the maximum subensemble size per process as a criterion for performing an annihilation step, depending on the outcome of the growth prediction. For 16 processes each process is allowed a maximum of 500 000 particles, whereas for 32 processes it is only 250 000. This specific example shows an imbalanced concentration of particles in the form of the initial wavepacket; the larger subensemble maximum for 16 processes, allows for more growth before annihilation takes place and therefore can lead to a larger computational load overall.

The absolute number of (numerical) particles within a subdomain serves as an indicator of the computational load a process experiences (the true load is also a function of the generation rate). Figure 5.5 shows the number of numerical particles at the same two time instances as in Figure 5.3.

Figure 5.6 illustrates the time evolution of the number of particles on the processes. The evolution of the wavepacket travelling across the 200 nm domain is clearly demonstrated by the initial imbalance of the workload on the processes. Initially, when the wavepacket is still narrow, the particles are passed on between the processes, but as the packet spreads the load distribution becomes more uniform. After
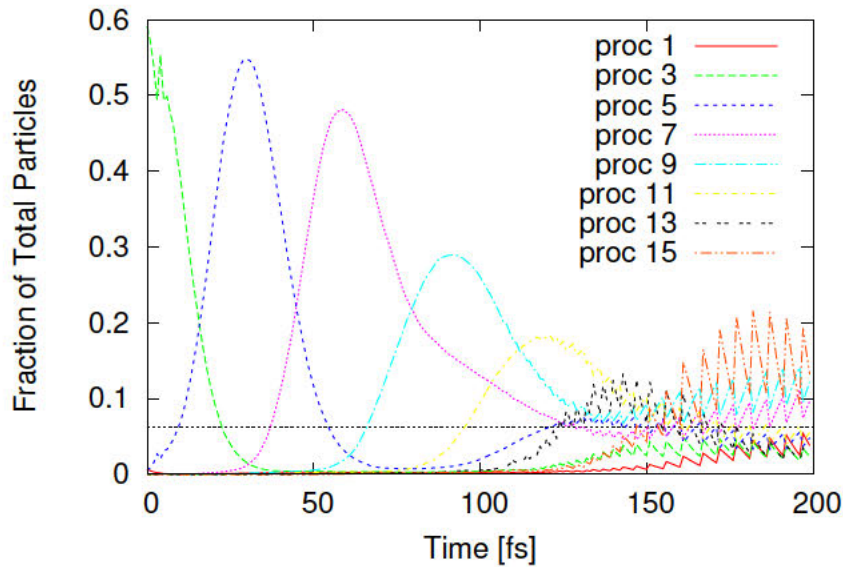
Figure 5.6: Time evolution of numerical particles on (selected) processes representing the changing computational load. Particles are moving from left to right and eventually spread out, making the load distribution better ($> 150\,\mathrm{fs}$); the dashed horizontal line indicates the ideal load per process

approximately $150\,\mathrm{fs}$ the distribution remains almost constant with some oscillations. The oscillations are due to the particle growth and annihilation taking place at regular intervals.

An important observations can be made from Figure 5.6: the ideal load per process is approached, when the evolution approaches a stationary regime. In light of this consideration, the next example starts with an initial condition which is already broadly spread out.

Double potential barrier: This example considers two wavepackets travelling towards each other in a domain with two potential barriers. The setup is the same as in the first example, apart from the location of the barriers and the initial position of the wavepackets, as stated in Table 5.2. A particle ensemble obtained after $80\,\mathrm{fs}$ of evolution is used as an initial condition to start the simulation already with a good load-balancing between the processes.

Figure 5.7 shows the parallel efficiency curves for the double barrier example, which is improved with respect to the single barrier case. This can be attributed to the improved load balancing amongst the processes from the initial condition. The evolution of the numerical particle distributions in Figure 5.8 shows how the load amongst the processes spreads out faster than for the single barrier problem. These peculiarities make the spatial decomposition method the ideal candidate for simulations of the stationary state, an example of which is shown in Section 6.2.

Table 5.2: Simulation parameters for double barrier benchmark example

Parameter	Value	Unit
x_0^1	40.0	nm
k_0^1	$18\Delta k$	nm^{-1}
x_0^2	150.0	nm
k_0^2	$-18\Delta k$	nm^{-1}
Barrier 1 width	3.0	nm
Barrier 1 left edge	70	nm
Barrier 1 height	0.15	eV
Barrier 2 width	3.0	nm
Barrier 2 left edge	130	nm
Barrier 2 height	0.05	eV
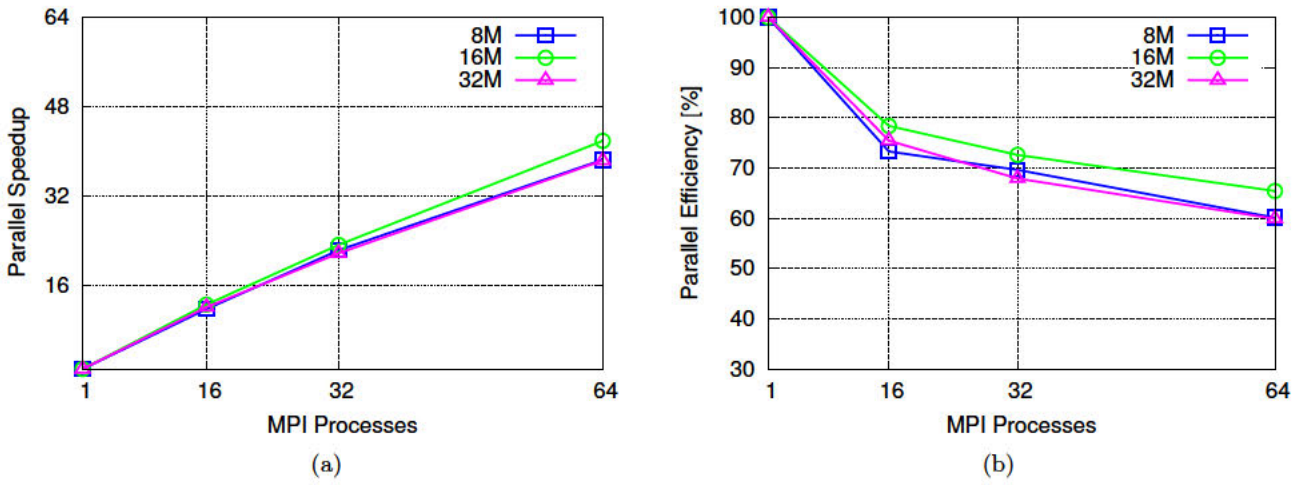
Figure 5.7: Parallel speed-up and efficiency of the double-barrier problem for different maximum particle ensemble sizes
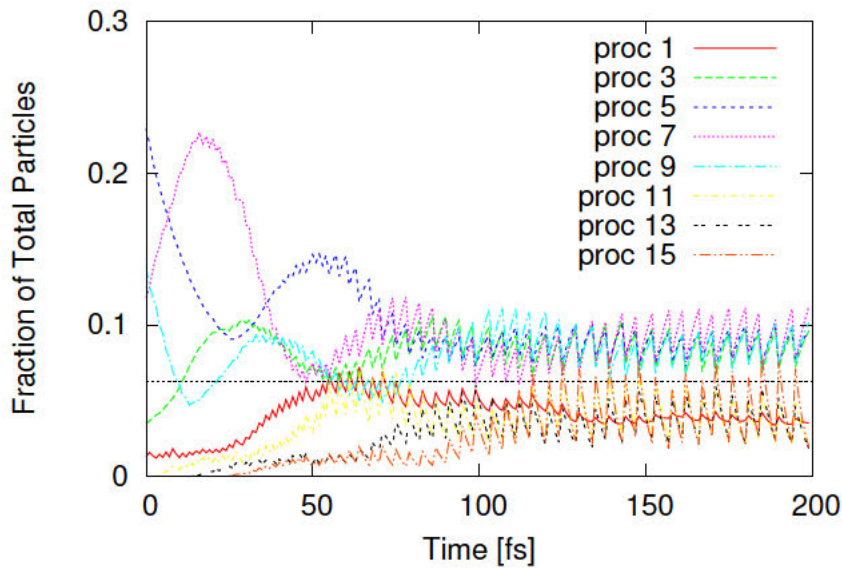
Figure 5.8: Time evolution of numerical particles on (selected) processes to represent changing computational load. The simulation starts with a precomputed initial condition, where particles are already spread out. A reasonable load distribution is achieved at early evolution times: the dashed horizontal line indicates the ideal load per process.

It is important at this point to underline the actual benefits of utilizing the presented parallelization approach in day to day research: the single barrier example, running on 64 processes with 60% efficiency, translates into a speed-up of around 40 times. In terms of execution time, the serial runtime of around 47 minutes was reduced to just 70 seconds. In the same vein, a simulation problem (of sufficient complexity), which would normally require two days of computation can be completed in approximately one hour. This aspect opens up a new realm of simulation problems, which can be investigated using the Wigner formalism.

#### 5.5.2.2 Two-dimensional results

The impact of the decomposition approach, slab- versus block-decomposition has been investigated in [150]. The main results are summarized here.

An example problem is considered with three minimum-uncertainty wave packages evolving over a $128\,\mathrm{nm} \times 128\,\mathrm{nm}$ spatial domain. Sixteen non-identical acceptor dopants (positively charged) are spread across the domain, which yields the potential profile shown in Figure 5.9a and the corresponding particle generation rate γ depicted in Figure 5.9b. The particle generation rate is concentrated around each dopant within the region corresponding to the coherence length of $30\,\mathrm{nm}$. All boundaries are reflective, therefore, no particles leave the simulation domain. The maximum number of particles for the entire domain is limited to $64 \cdot 10^7$ particles; the local maximum for each process depends on the total number of processes used.

Figure 5.10 depicts the parallel execution performance for the slab and the block decomposition approach. The slab decomposition technique offers a significantly better parallel execution performance. The block decomposition method introduces a significant communication and requires additional logic, resulting in inferior performance relative to the slab decomposition approach. Especially for 128 MPI processes, the overhead triggers a stagnation of the scalability as the number of particles (work load) per process is too small relative to the additional communication overhead.

The memory consumption of the block-decomposition approach is slightly higher than for the slab-decomposition approach, due to the greater overhead for the communication and the related logic. However, this is negligible compared to memory requirements for sorting the particle ensemble and performing the annihilation algorithm.

Although the load balance with the block decomposition appears to be better at the various time steps shown in Figure 5.11, the method's performance is inferior to the approach using slab-decomposition. This is attributed to the additional communication and logic overhead incurred when using the block decomposition, which is between 1.5 to 4 times greater than in the case of slab-decomposition [150]. This result is particularly interesting, as the load balance does not appear to have a significant impact on the parallel scalability.

The annihilation procedure considerably reduces the number of particles (load) in a subdomain and is performed on an 'as-needed' basis, which will differ for the various decomposition approaches. An average of the load over the entire simulation time paints the most realistic picture. Nonetheless, the snapshots of load balance at various times shown in Figure 5.11 already are useful, as they make clear that such differences exist.

5.5.3 Summary

This chapter has described the parallelization of the WEMC simulator using spatial domain decomposition, which has been deemed to be the best-suited for a large-scale parallelization in a distributed-memory system. The performance of the developed code has been evaluated and the slab-decomposition approach shows excellent parallel efficiency for two-dimensional simulations. This result has enabled the efficient use of high performance computing for two-dimensional Wigner Monte Carlo quantum simulations which facilitated the applications shown in Chapter 6.
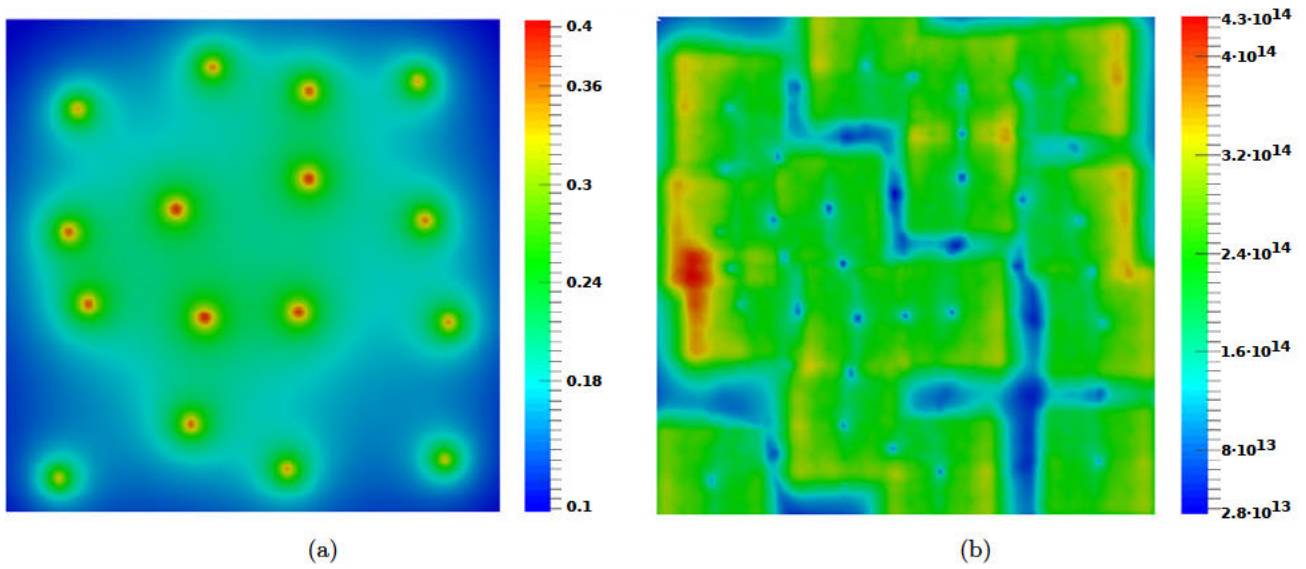
Figure 5.9: Sixteen (non-identical) acceptor dopants forming a potential profile (a) with peaks between 0.3 eV and 0.4 eV. The corresponding generation rate ($\gamma \, [\mathrm{s}^{-1}]$) (b) attains its highest value around the placed dopants (figures taken from [150]).
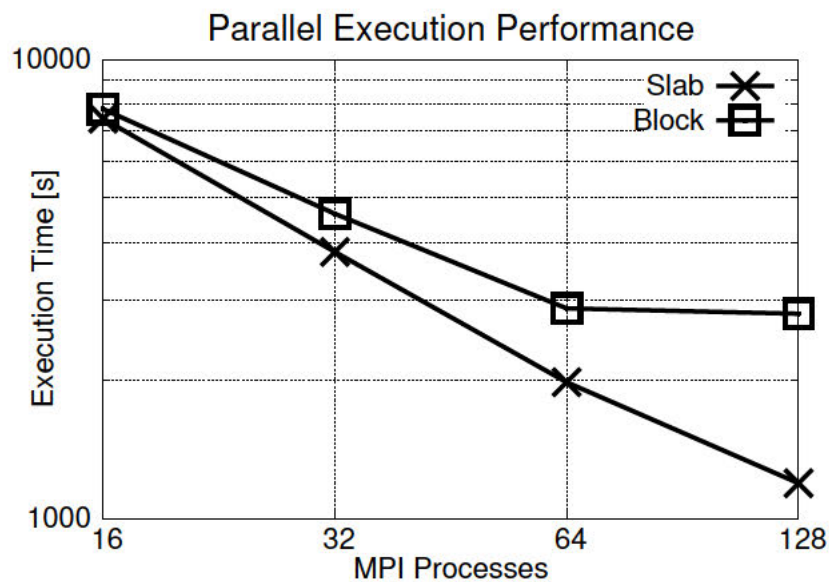
Figure 5.10: Comparison of the execution times between the slab- and block-decomposition approaches.
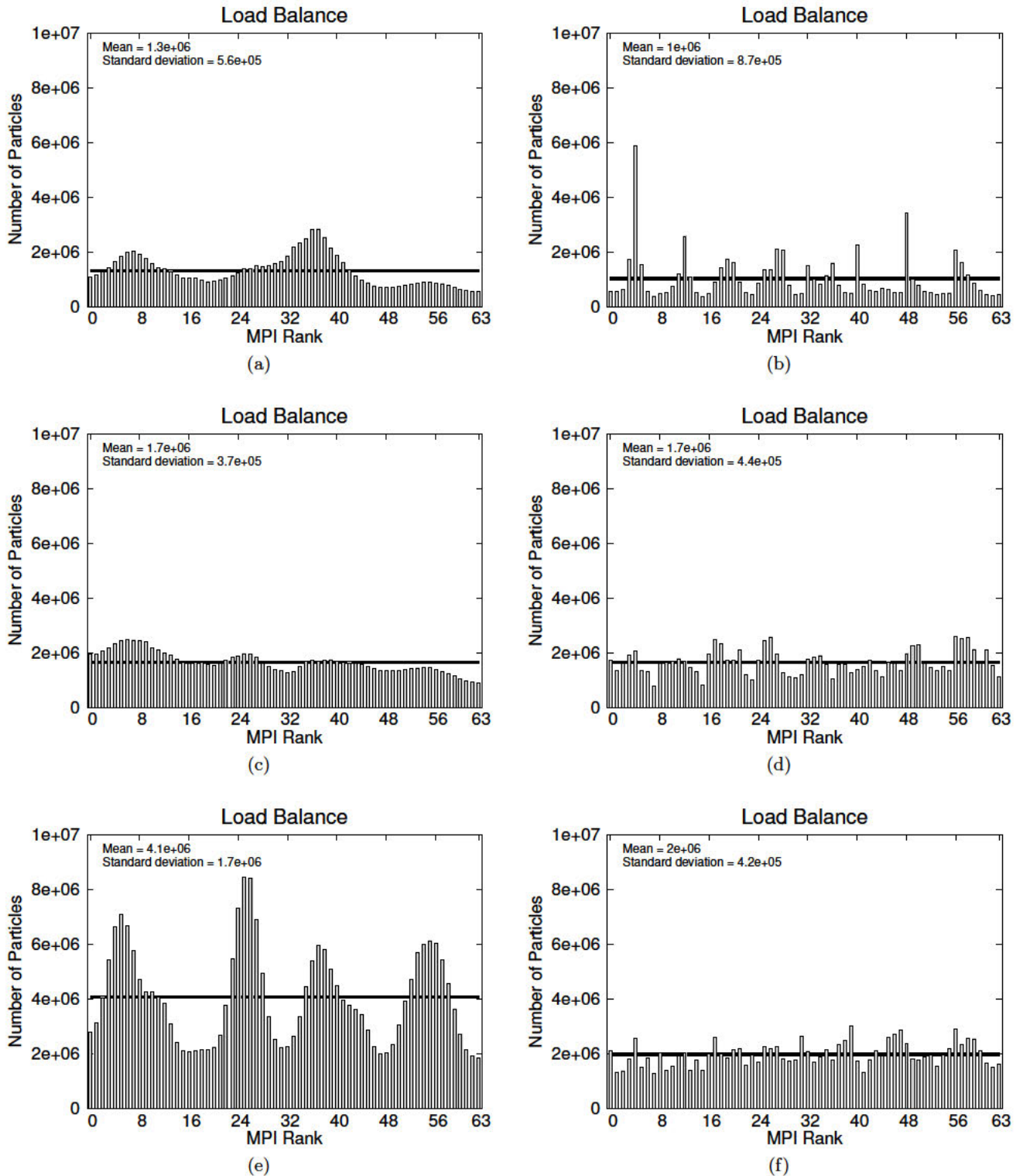
Figure 5.11: Comparison of the load balance for the slab decomposition (a,c,e) and block decomposition (b,d,f) approaches for 64 processes at 100 fs, 200 fs and 300 fs. The horizontal lines denote the mean number of particles. (Figures taken from [150])

Chapter 6

Applications

This chapter demonstrates the application of the Wigner Monte Carlo simulator which has been discussed in the preceding chapters. The simulation of electrostatic lenses is presented in Section 6.1 along with investigations into the emerging field of quantum control of electron wavepackets. An application of electrostatic lenses to improve the performance of actual devices is demonstrated in Section 6.2. The calculation of a steady-state solution using 2D WMC simulation is shown, which presents a novel result that has been made possible by the optimized algorithms (Chapter 4) and the parallelization (Chapter 5) presented before.

6.1 Electrostatic lenses

6.1.1 Introduction

An electrostatic lens refers to a specially shaped potential with convex/concave features, similar to optical lenses, used to steer coherent electrons. The concept was first demonstrated experimentally in 1990 in [152, 153], in low-temperature, high-mobility semiconductors, which ensured that the coherent electrons had a sufficiently long mean free path to conduct experiments with structures made with the lithographic capabilities at that time. An experimental realization of the electrostatic lenses is illustrated in Figure 6.1.

The astounding decrease of the feature sizes in semiconductor devices, along with novel materials like graphene, has made (semi-)ballistic electron transport applicable at room temperature [154]. This has sparked new interest in applying electrostatic lenses in nanoelectronic devices, e.g. [155] suggests the use of lenses to focus electrons to the centre of nanowires, thereby avoiding rough interfaces and increasing mobility.

6.1.2 Law of refraction

Electrostatic lenses use analogous concepts from geometrical optics: Snell's law describes the refraction of a light beam traversing an interface between two different media of propagation, e.g. air and glass. An equivalent law of refraction can be derived for electrostatic lenses using the principle of energy conservation.

A particle with a wavevector $\mathbf{k}$ has a kinetic energy

$$E_k = \frac{\hbar^2 \left| \mathbf{k} \right|^2}{2m^*}, \tag{6.1}$$

where $\hbar$ denotes the reduced Planck constant and m^* the effective mass. As a particle traverses the interface between regions at different potentials (illustrated in Figure 6.2), its kinetic and potential energies change. The change in kinetic energy is attributed only to the change of the component of the wavevector normal to the interface (red); the component parallel to the interface (blue) is left unchanged. It follows that

$$\left| \mathbf{k}_1 \right| \sin \theta_1 = \left| \mathbf{k}_2 \right| \sin \theta_2, \tag{6.2}$$
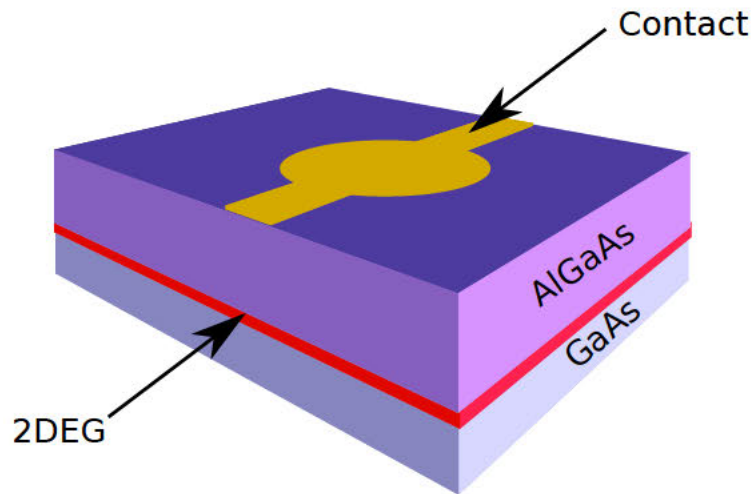
Figure 6.1: A GaAs/AlGaAs heterostructure, which forms a two-dimensional electron gas (2DEG) at the material interface. A potential is applied to a contact at the top of the structure, which is shaped to induce the desired potential barrier (electrostatic lens) in the plane of the 2DEG.

where θ_1 (θ_2) is the angle of incidence (refraction) with respect to the normal of the interface. The magnitude of the wavevector is proportional to the square root of the kinetic energy:

$$\frac{\sin\theta_2}{\sin\theta_1} = \frac{|\mathbf{k}_1|}{|\mathbf{k}_2|} = \frac{\sqrt{E_{k_1}}}{\sqrt{E_{k_2}}}. \tag{6.3}$$

Therefore, the square root of the kinetic energy of a particle is analogous to the refractive index used in geometrical optics and this value can be dynamically modified by changing the value of the potential energy in the region of the lens. This concept is illustrated in the following section.

6.1.3 Converging lens

Optical lenses which operate in a medium (air) with a lower refractive index use the familiar double-convex shape to focus light. Rays are refracted towards the normal of the interface as they enter the lens. The law of refraction for electrostatic lenses (6.2) reveals that if a positive potential step is used for the lens, the kinetic energy (refractive index) decreases and the trajectory of the electron is bent away from the normal to the interface. This dictates that a double-concave shape is needed to form a converging lens with a positive potential [153]. Conversely, a negative potential, i.e. a well, would require a double-concave shape.

The potential shape used to form the electrostatic lens is shown in Figure 6.3. The electrostatic lens has a peak potential energy of 40 meV and the wavepacket is initialized with a kinetic energy of 180 meV, moving rightwards. The electrostatic lens clearly focuses the wavepacket (density) after 150 fs of evolution. A comparison to the evolution without a lens is shown in Figure 6.4, using a similar setup in the vertical orientation. A further consideration for designing the lens is that the size of the lens should be larger than the De Broglie wavelength of the electron (determined by its energy), otherwise an effective focusing will be distorted by diffraction effects.

The refractive index of the lens, and thereby its focal length, can be modified by varying the magnitude of the potential with which it is formed; a larger value of the potential energy for the lens (relative to its surrounding region) invokes a stronger refraction ('bending'). Figure 6.5 compares the effect of different values: It can be clearly seen that the barrier with the higher potential energy focuses the wavepacket more sharply (at the distance observed at the time instance shown). The applied potential can thereby control, when and at which distance, a wavepacket is focused (on a detector, for instance).
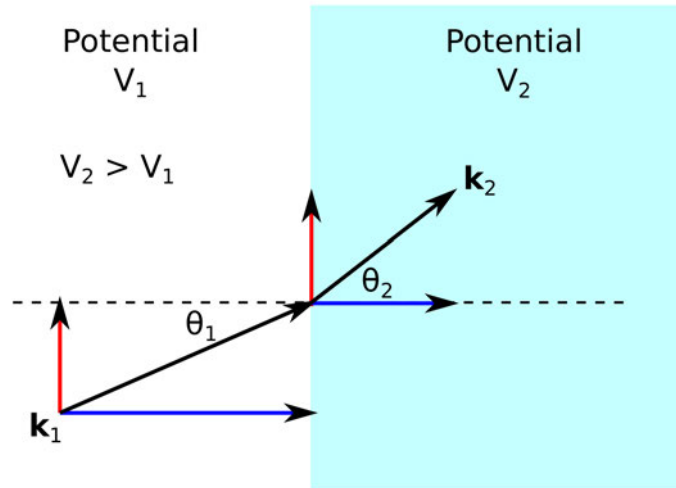
Figure 6.2: The wavevector of a particle is changed from $\mathbf{k}_1$ to $\mathbf{k}_2$ as the interface between regions at different potentials is traversed. The decomposition of each wavevector into its components normal (blue) and parallel (red) to an interface is shown. The normal component is modified according to the potential change, whereas the parallel component remains unchanged. The illustrated case assumes a positive potential step, where $V_2 > V_1$, such that the particle is refracted away from the normal to the interface.

6.1.4   Wavepacket control

The converging lens demonstrated in the preceding section, follows from classical analogues in geometrical optics. It is possible, however, to extend the concept of electrostatic lenses much further: engineered potential profiles can be used to dynamically control wavepackets, e.g. creating entangled states by splitting up wavepackets. It should be noted that the electron is not physically split; it is a single electron in an entangled state.

Figure 6.6 shows a rhomboid-like potential shape, along with the corresponding generation rate, which forms a lens that is able to realize such a function. The behaviour of the lens can be manipulated by changing the magnitude of the potential. This can be done dynamically, if the lens is realized with a structure similar to the one shown in Figure 6.1, using a time-dependent potential bias on the contact. Figure 6.7 illustrates the effect of the lens at different potential values. The density peaks indicate regions with a higher probability to find an electron. In Figure 6.7b (peak potential energy 70 meV) the wavepacket almost fully traverses the lens and is split into two parts. The same lens shape, but with a potential energy of 120 meV, splits the wavepacket into four parts (Figure. 6.7b): The front edges splits off a portion of the wavepacket by reflection, while the concave-shaped rear edges focus the transmitted parts again (applying the principle of Section 6.1.3). In the first case, with two peaks (the most-probable components of the state), the y-component of the wavevector remains positive, whereas for the second case, at a higher potential energy, the wavevector of the scattered state also has a negative y-component. This example clearly illustrates how specially shaped potentials can be used to influence the scattering pattern of an electron wavepacket. By varying the potential energy the electron can be guided in a certain direction with a controllable probability. This can be of use in the field of quantum computing to generate a (modifiable) entangled state and direct it to other computing elements.

6.2   Drive-current enhancement

The application of electrostatic lenses to focus electrons can be used to increase the performance of actual devices. This novel concept is presented here (based on [156]).
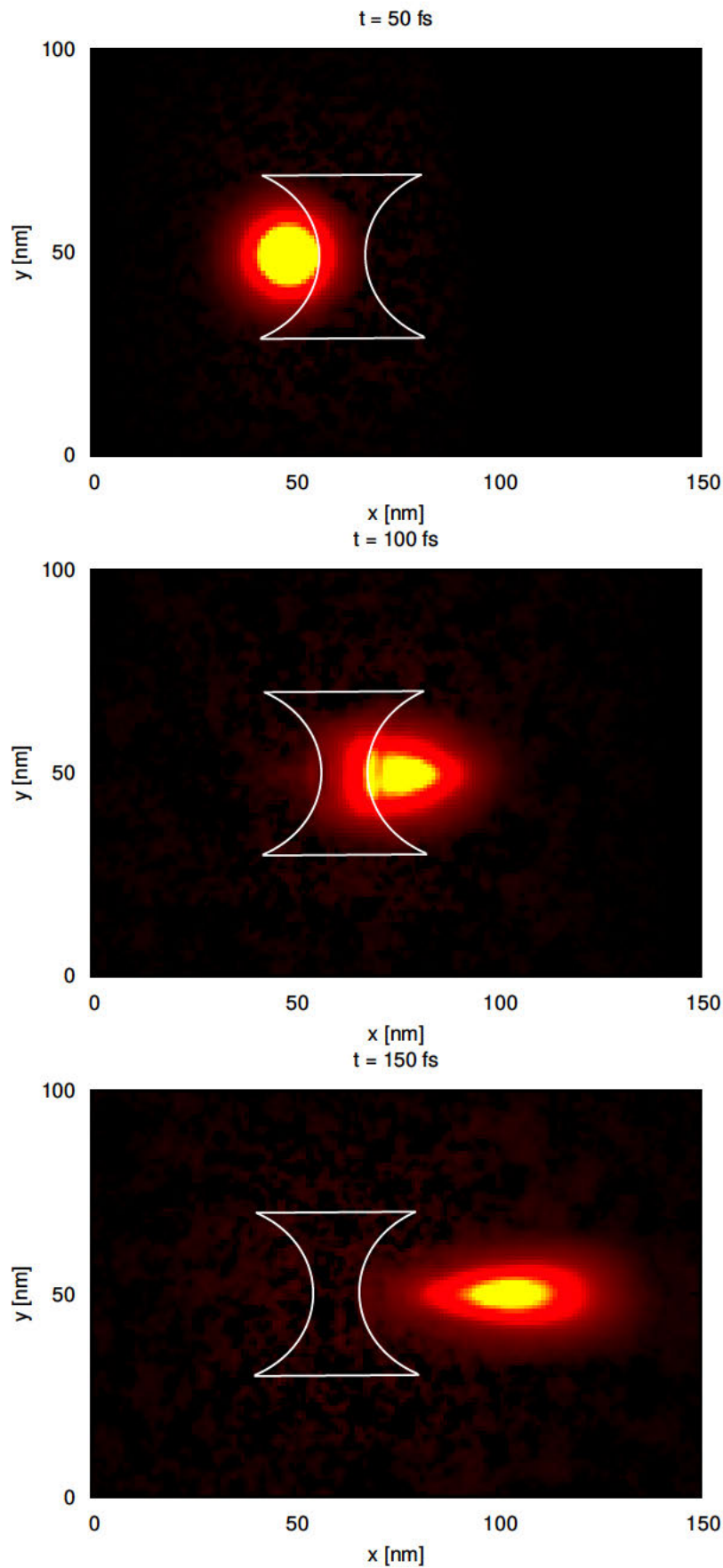
Figure 6.3: The density of a Gaussian wave packet, over a sequence of time steps, is focused by a converging electrostatic lens with a double-concave shape (as indicated by the annotation in white). The wavepacket has a kinetic energy of 180 meV and the lens has a potential energy of 40 meV.

(a) 0 meV

(b) 40 meV
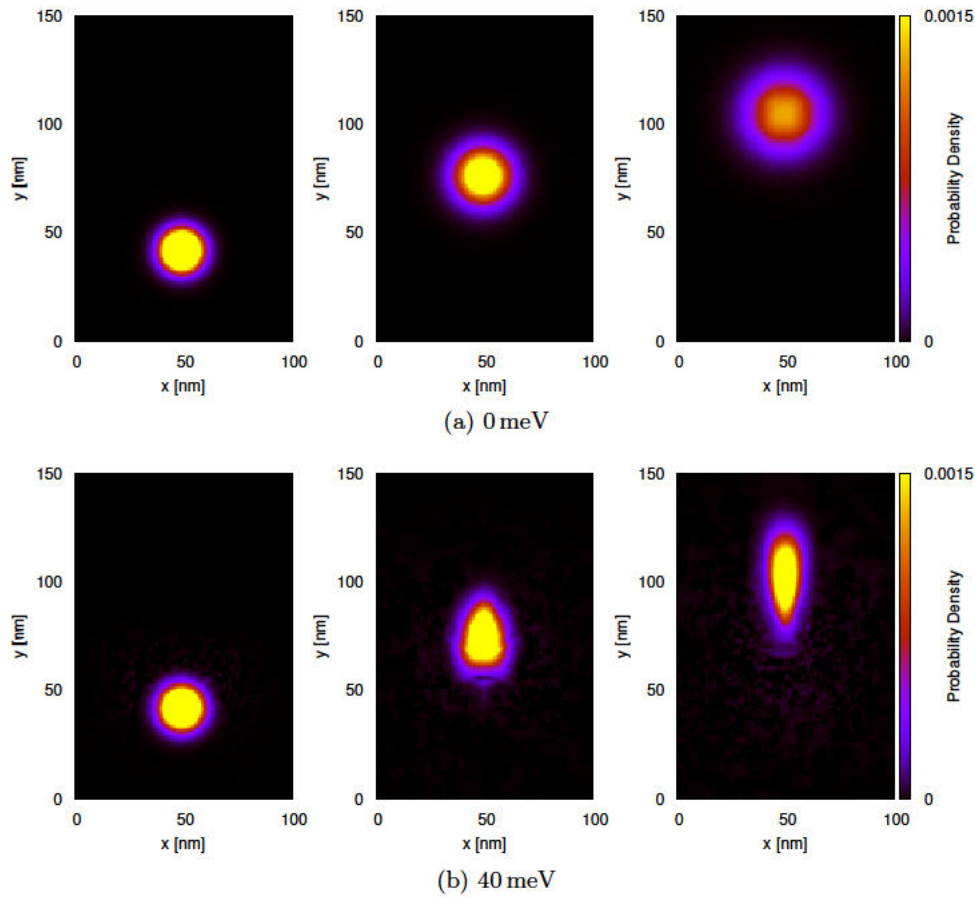
Figure 6.4: Wavepacket evolving freely (top sequence) and interacting with a double concave electrostatic lens (bottom sequence); the time steps (from left to right) correspond to 40 fs, 100 fs and 150 fs.

(a) 10 meV   (b) 20 meV   (c) 40 meV
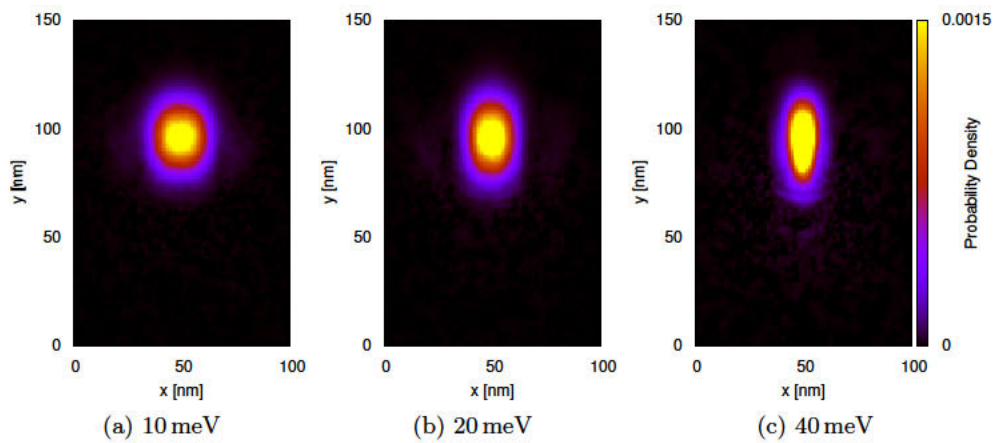
Figure 6.5: Comparison of the density of a wavepacket evolved for 135 fs in the presence of an electrostatic lens at various potential energy values to show different focussing.
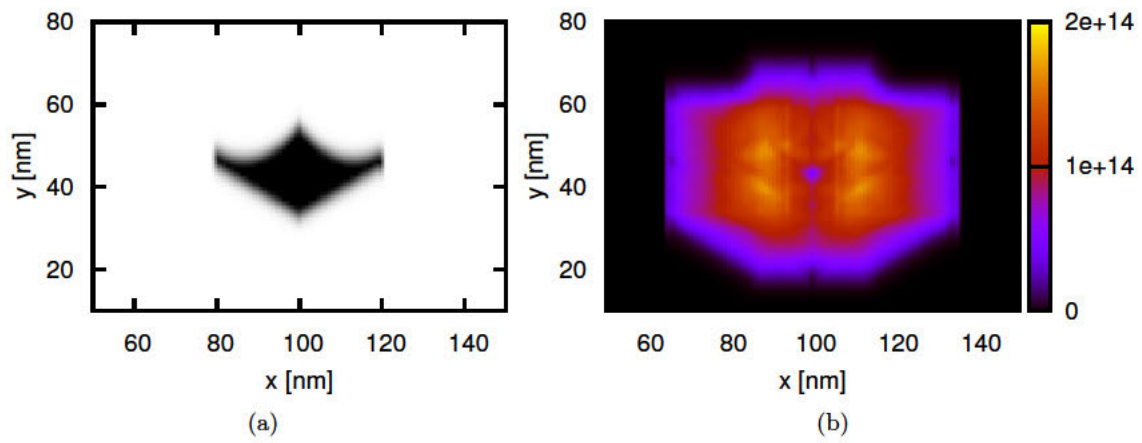
Figure 6.6: Two-dimensional potential (a) with rhomboid-like shape and concave-shaped rear edges, forming an electrostatic lens to scatter an electron wavepacket in various directions. The potential value of the lens is constant; it has no three-dimensional features. The corresponding particle generation rate γ is shown in (b).
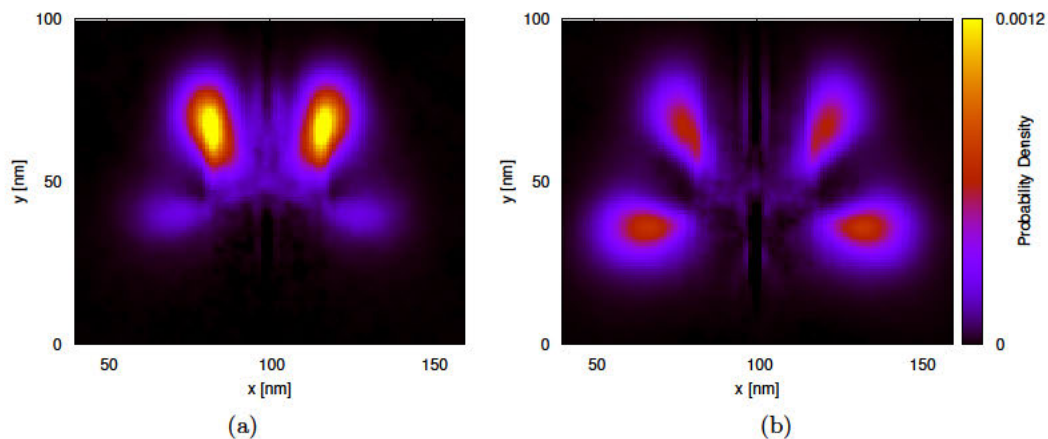
Figure 6.7: Wavepacket is split either (a) into two or (b) into four parts, after 90 fs evolution, by a rhomboid-like potential profile with concave rear edges with a peak potential energy of 70 meV and 120 meV, respectively.

(a)   (b)
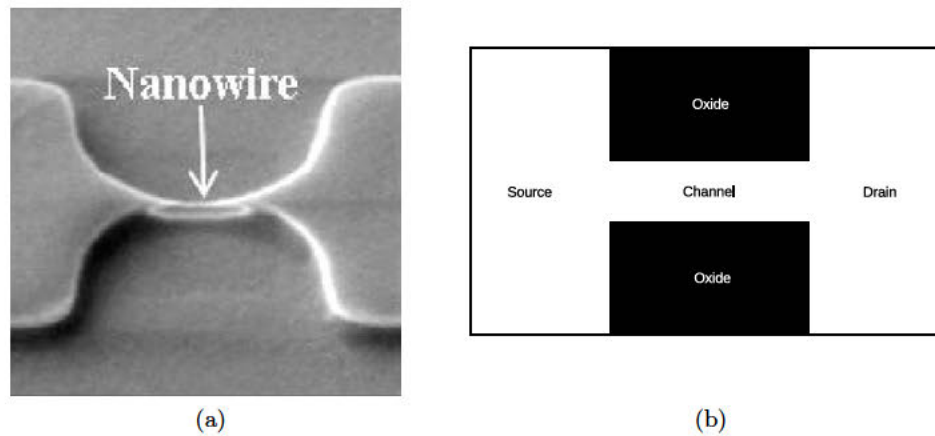
Figure 6.8: A scanning electron microscope (SEM) image of a Si nanowire (a) shows the narrow channel region connected to wide access regions/contacts (Figure taken from [160] and adapted). This geometry is approximated by the structure in (b) and is representative of a DG-MOSFET structure (or with a change in orientation, a UTB-SOI MOSFET with raised drain and source regions).

The dimensions of the channel in modern transistor architectures, like multigate-FETs or UTB-SOI, have progressively been scaled down to retain adequate electrostatic control of the channel. The connecting source/drain extensions, however, remain relatively large and have started to play a significant role in the overall performance of the device [157]. Therefore, considerations are made to optimize these regions, e.g. to reduce the contact and access resistances [158, 159]. An electrostatic lens in the source region can be used to increase the drive-current through the channel by focusing the electrons into the channel aperture and thereby reducing reflections from the adjacent oxide.

6.2.1   Geometry

A geometry representative for many current transistor structures in which a narrow channel ($< 10\,\text{nm}$) is surrounded/sandwiched by oxide and is extended to a larger drain/source contact region is shown in Figure 6.8. The geometry considered for the simulations is shown in Figure 6.9 omits the drain region. The oxide is approximated by a $0.4\,\text{eV}$ barrier, which is sufficient to constrain the wavepackets to the channel and source regions without inducing excessive particle generation (the generation rate is related to the magnitude of the potential differences).

Wavepackets are periodically injected from the left boundary towards the channel (discussed further hereafter). A converging electrostatic lens is placed before the aperture of the channel to focus the wavepackets into the channel. The De Broglie wavelength of the wavepackets considered here is in the order of a few nanometres, which requires the lens to be at least $10\,\text{nm}$ wide to avoid diffraction effects.

6.2.2   Steady-state current

The feasibility of investigating time-resolved quantum transport presents one of the biggest advantages of Wigner formalism based simulations. However, the steady-state behaviour of devices is of considerable importance in many applications. The possibility to obtain a steady-state solution is also of interest to be able to strike a comparison with the results obtained by the NEGF approach. The approach presented here to calculate the current is intended to make a qualitative investigation possible. A validation of this approach was done by counting the charge crossing the absorbing boundaries (the left source and the right drain) over a time interval.

In order to calculate the steady-state current in a device, minimum uncertainty wavepackets (3.34) are periodically injected from the left boundary representing the source contact. Here, identical wavepackets, with the parameters defined in Table 6.1 with a mean energy of approximately $200\,\text{meV}$ are injected. A distribution function, e.g. Maxwell-Boltzmann shaped, can also be sampled to select
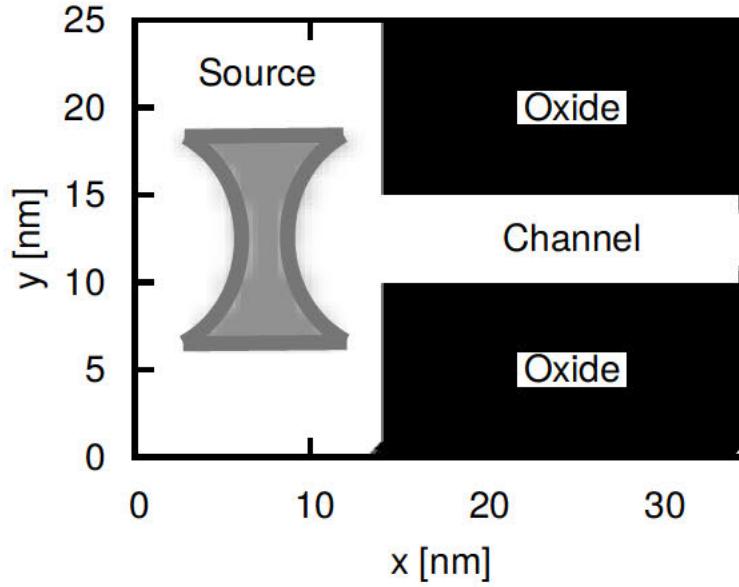
Figure 6.9: A 5 nm wide channel sandwiched between two oxide layers (black) with an adjacent source region with a double-concave lens (grey), at 40 meV, positioned in front of the aperture to the channel. The left and right boundaries of the shown domain are considered as contacts (absorbing/emitting particles), whereas reflecting boundary conditions are imposed on the top and bottom boundaries. Wavepackets are injected from the left 'contact' periodically every 10 fs.

Table 6.1: Simulation parameters for injected wavepackets and domain

r_0 [nm]	σ [nm]	L_{coh} [nm]	k_0 [nm^{-1}]	Δx [nm]
$(\cdot, 12.5)$	$3/5$	30	$(9\Delta k)$	0.5

the mean energy (wavevector) of each individual wavepacket injected from the source contact.

The drive-current (through the channel) is calculated by the Ramo-Shockley theorem [161], adapted to account for the signed numerical particles used in the WMC simulator:

$$I = -\frac{1}{L_x}\frac{e}{N_{wp}}\sum s_i v_i, \tag{6.4}$$

where L_x, N_{wp} and e represent the device length in the x-direction (35 nm), the number of numerical particles representing one injected wavepacket and the unit charge, respectively. The summation is performed over all particles in the device, taking the sign (s_i) and the x-velocity (v_i) of each numerical particle into account. The electric field is assumed to be uniform across the channel, directed rightwards, i.e. $\mathbf{E} = (0, E_x)$.

Since each wavepacket represents a single electron, the magnitude of the injected current can be set by the period of injection T_{inj}, i.e.

$$I = \frac{\Delta Q}{\Delta t} = \frac{e}{T_{inj}}. \tag{6.5}$$

6.2.3 Results

Simulations, using the geometry in Figure 6.9, are run with and without the addition of a converging lens in front of the aperture to the channel. Figure 6.10 compares the evolution of the probability density towards a steady state with and without a lens. Without the lens a large portion of the wavepacket is reflected from the oxide adjacent to the aperture of the channel, thereby reducing the drive-current. With the addition of a converging lens, the wavepackets can be focused to move into the channel, thereby reducing the reflections from the oxide barriers next to the aperture of the channel.

This observation is supported by the associated difference of the k_x-distributions. Figure 6.11 (a) shows the change in k_x-distributions corresponding to Figure 6.10 (a); the reduction in the kinetic energy of the wavepackets due to the barrier presented by the lens is clearly shown. However, once the steady-state is reached (at approximately $t = 140$ fs; Figure 6.11 (b)) a reduction in the probability of left-moving (negative k_x) particles is observed, indicating the lens leads to reduced reflections overall; the forward-moving particles (around $5\Delta k$) are enhanced.

The steady-state current is calculated for wavepackets with a standard deviation of 3 nm and 5 nm, as shown in Figure 6.12. The addition of the lens consistently increases the drive current; the gains made by focusing the wavepackets into the channel are larger than the losses from reflections by the lens, which presents a small barrier to the electrons. The lens shows a better effect, when a (spatially) broader wavepacket is focused and increases the channel current by 15% if $\sigma = 5$ nm, compared to an 8% increase if $\sigma = 3$ nm. Some uncertainty exists about the true value of the standard deviation for the wavepacket [34]. Moreover, the Gaussian wavepacket spreads out as it propagates. The current gradually rises as the domain is filled with particles before it converges to a steady-state value after approximately 140 fs.

It is concluded that the addition of a converging electrostatic lens in the source region of a transistor can be used to effectively focus electron wavepackets into a nanoscale channel by reducing reflections from the oxide surrounding the aperture. Moreover, the results illustrate for the first time how a steady-state current is obtained with a two-dimensional WEMC simulator, allowing it to be applied to investigate practical issues of semiconductor devices.

(a) $t = 40\,\mathrm{fs}$

(b) $t = 60\,\mathrm{fs}$

(c) $t = 100\,\mathrm{fs}$
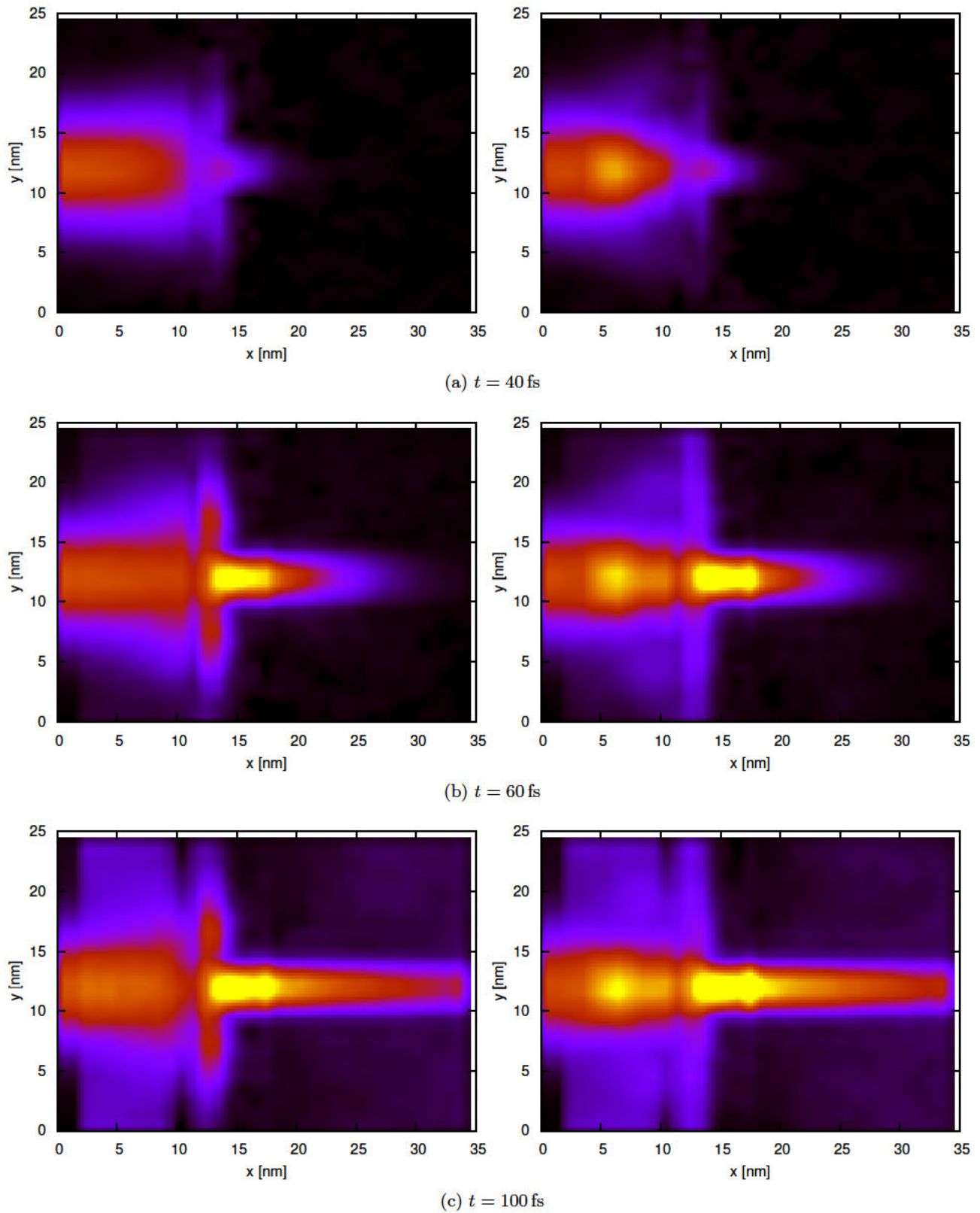
Figure 6.10: Comparison of the evolution of the particle density for the structure in Figure 1 without (left) and with (right) a lens. A wavepacket with $\sigma = 5\,\mathrm{nm}$ is injected every $10\,\mathrm{fs}$. The density at the sides of the aperture to the channel is reduced by the addition of the lens.

(a) $t = 40\,\text{fs}$

(b) $t = 140\,\text{fs}$
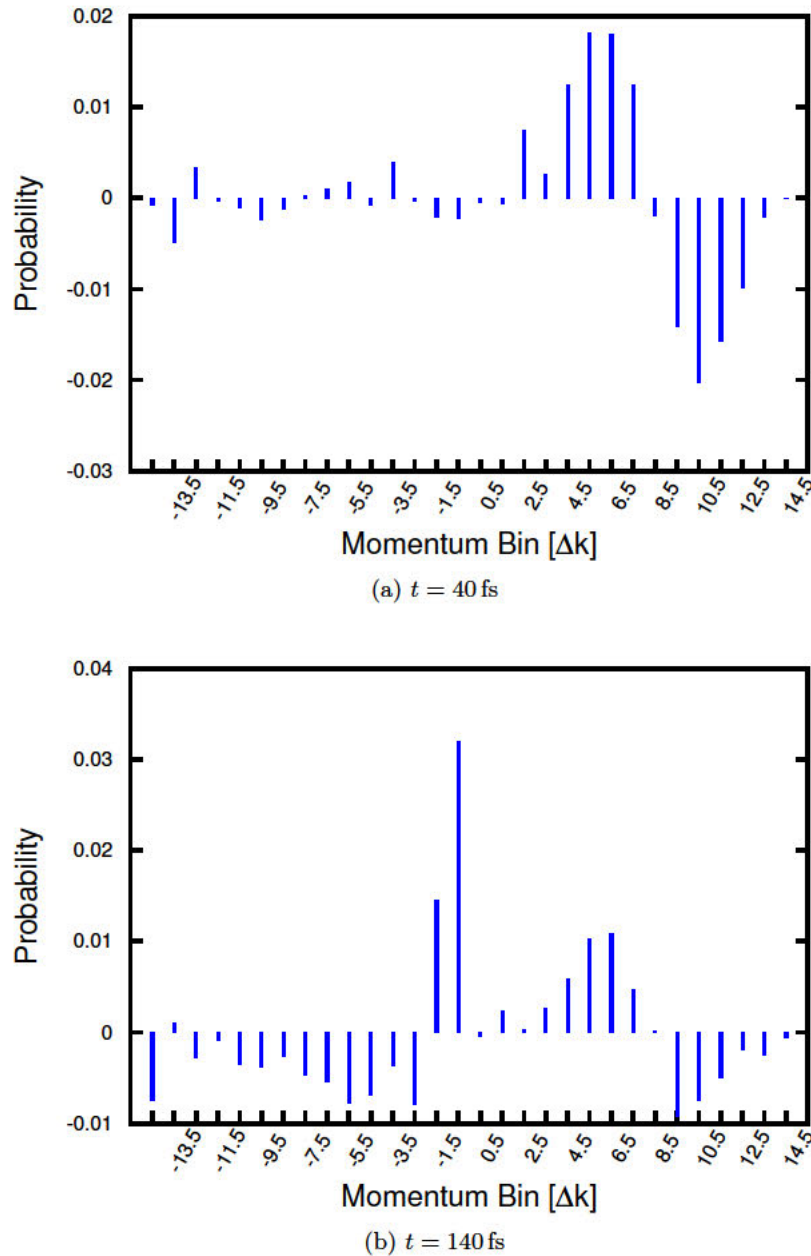
Figure 6.11: Difference in k_x distributions with and without the addition of a lens after (a) 40 fs and (b) 140 fs of evolution. A positive value indicates components enhanced by the lens; negative values indicate a suppression. The initial reduction in the kinetic energy of the wavepacket is seen in (a). Once steady-state is reached (b), an overall reduction in the negative k_x (i.e. leftwards moving) components is evident.
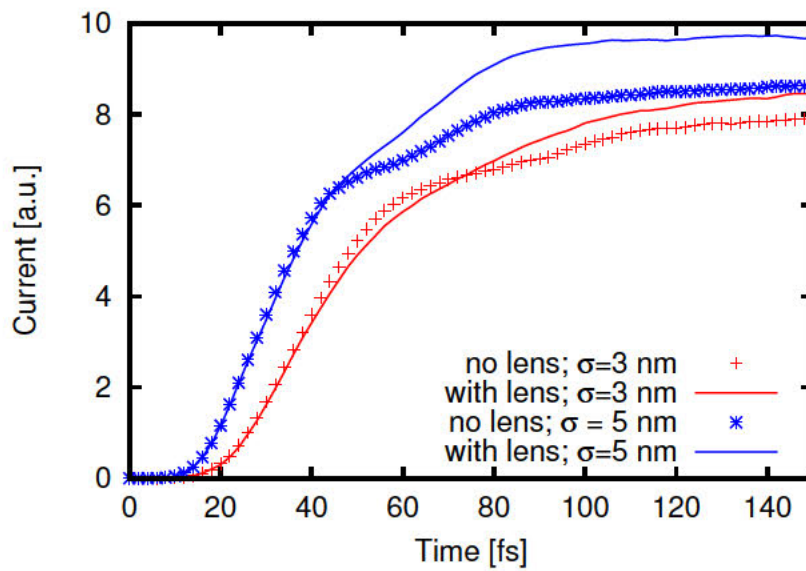
Figure 6.12: Comparison of the channel currents achieved with and without the addition of a 40 meV electrostatic lens for wavepackets with a spatial standard deviation of 3 nm and 5 nm.

Chapter 7

Summary and Evaluation

This chapter briefly summarizes the work which has been presented in the preceding chapters. Thereafter, the major contributions made in the scope of this thesis to advance the field are summarized. Finally an overall conclusion is drawn.

7.1 Summary of content

The preceding chapters have introduced all the aspects relevant to the simulation of time-resolved quantum transport, in two dimensions, using a Wigner Monte Carlo approach. The work has been presented in six chapters.

Chapter 1 motivated the need for utilizing TCAD in the research and design of nanoelectronics. An overview of the most widespread transport models was given with the qualities and limitations of each. It was found that the Wigner-Boltzmann transport model offers some unique qualities in that it makes time-resolved quantum transport with scattering computationally feasible, while providing many classical analogies to be used. This presented the motivation for the development of a solver for the Wigner-Boltzmann transport equation.

Chapter 2 presented the Wigner formalism of quantum mechanics, which gives a formulation in the phase space. From this background the derivation of the Wigner transport equation was presented. Thereafter, the augmentation of the Wigner equation with semi-classical phonon scattering models was shown to be justified by outlining the rigorous derivation, which finally yielded the Wigner-Boltzmann equation. The semi-discrete form of the Wigner equation, which follows from considering a finite spatial domain, was presented. The various transport problems that can be approached with assigned simulations and the handling of boundary conditions was outlined. Finally, an overview of existing solvers for the Wigner-Boltzmann equation, for one-dimensional problems, was given.

Chapter 3 introduced the signed-particle method – the Monte Carlo approach which has made the solution of the Wigner equation in two spatial dimensions computationally feasible. The mathematical foundation of the signed-particle method was presented to be based on the integral formulation of the Wigner-Boltzmann equation, which is developed into a Neumann series that can be evaluated using Monte Carlo integration. Armed with this theoretical background, the basic building blocks and architecture of an algorithmic implementation of the signed-particle method was shown.

Chapter 4 highlighted the improvements and contributions made by the author to the signed-particle method's algorithms. The contributions encompass optimized algorithms for increased computational efficiency, statistical enhancements and treatment of discretization effects, which can lead to non-physical behaviour. Finally, the considerably improved accuracy of the signed-particle method, using the optimized algorithms, was demonstrated by a comparison of the numerical results to an exact solution of a physical problem.

Chapter 5 treated the parallelization of the Wigner Ensemble Monte Carlo code, which was developed in the scope of this thesis, using a spatial domain-decomposition approach, which is well-positioned for large scale parallelization. The latter was motivated by evaluating the possible parallelization approaches to Wigner Monte Carlo simulations with due consideration to the typical hardware architecture of supercomputers. Finally, the parallel efficiency of the selected approach was

demonstrated using one- and two-dimensional examples.

Chapter 6 demonstrated an application of the Wigner Monte Carlo simulator to investigate electrostatic lenses, which were applied to manipulate the dynamics of wavepackets. The latter is of interest in the emerging field of quantum control, but also for actual devices. The improvement of the drive-current in a nano-scaled channel was shown by focussing electrons using a converging lens. The calculation of a steady-state solution was shown for the first time and has been made practical by the parallelization presented in Chapter 5.

7.2 Summary of contributions

Chapters 1 to 3 provide the contextual and theoretical background of this thesis. The existing body of literature has been reviewed and summarized to establish the state of the art from where this research started from. In Chapters 4 to 6 the main contributions and achievements made in the scope of this thesis are reported and are reiterated here:

1. Optimized algorithms for the most fundamental building blocks of the signed-particle method were designed and implemented. This entails the Wigner potential, the particle generation and the particle annihilation algorithms:

    (a) An algorithm to speed up the calculation of the two-dimensional Wigner potential by at least a factor of five was presented.

    (b) An analysis of the physical and computational implications of choosing a finite coherence length was given.

    (c) The origin of non-physical behaviour in certain simulations was discovered to be attributable to the Wigner potential, which governs the statistics of particle generation. The application of a tapering window to the potential has been demonstrated to mitigate this problem and greatly improved the quality of the Wigner Monte Carlo simulation results.

    (d) A potential statistical biasing when generating the momentum offsets for generated particles was identified and an approach to avoid this was shown.

    (e) It was discovered that the particle annihilation process can introduce a numerical diffusion and a method to counteract this has been devised.

    (f) Two alternative realizations of the annihilation algorithm have been shown, which greatly reduce, or completely eradicate, the huge memory demands of the annihilation step.

    (g) A method to anticipate the increase in the number of particles in a time step was introduced, which has enabled an automatic activation of the annihilation process without user input, greatly simplifying the use of the simulator for non-expert users.

2. A Wigner Ensemble Monte Carlo simulator was developed, which implements the signed-particle method with all the optimized algorithms presented. This entailed:

    (a) Writing over 8000 lines of code in C and various scripts for data post-processing and plotting functionality. The usability of the simulator has been considerably improved by facilitating control through input files for simulation parameters, potential profiles, initial conditions and the automatic selection of (some) simulation parameters.

    (b) The developed simulator now forms part of the open source suite of ViennaWD tools, which is freely available online and has been published on a website with examples and a user manual for the simulator[1]. The code serves as a reference implementation for the state-of-the-art of the signed-particle method.

    (c) A parallelized version of the code for high performance computing was designed. The data structures and logic needed for the parallelization of the code using MPI were implemented and the performance of the parallel code was characterized.

---

[1]The ViennaWD code and documentation is freely available under http://viennawd.sourceforge.net

3. The developed simulator has been applied to the simulation of electrostatic lenses, used for electron state control. The most important results include:

   (a) Demonstrating the control of the dynamics of wavepackets, making it possible to focus them or split them to create an entangled state, using special potential profiles.
   (b) Applying a converging lens to increase the drive-current through a nano-scaled channel.
   (c) Approximating a steady-state solution by the periodic injection of wavepackets from a boundary and calculating the corresponding current.

7.3 Conclusion

The preceding two sections encapsulate the research which has been performed and the contributions that have been made by the author of this thesis to advance the state of the art in Wigner Monte Carlo simulations over the course of approximately three years. The optimization of the algorithmic implementation of the signed-particle method has significantly increased its accuracy. Moreover, the development of a reliable simulation tool which can be run in a high performance computing environment presents a major step forward. These two facts combined, now enable the investigation of many interesting problems through the ability to simulate time-resolved quantum transport in two-dimensional structures with scattering mechanisms.

7.4 Outlook

The future research on simulations using the Wigner formalism can be pursued along three avenues, which all go hand in hand: Theoretical aspects of the Wigner formalism and their implementation with the signed-particle method, algorithmic and computational issues, and the application of the simulator to investigate quantum processes in nanostructures.

The extension of the Wigner Monte Carlo simulator to three spatial dimensions and the capability to calculate a self-consistent solution with the Poisson equation presents the most straight-forward continuation the work presented in this thesis. The challenge is of a computational nature since the existing algorithms must be extended for this purpose.

To garner a wider interest in Wigner Monte Carlo simulation within the semiconductor device research community, the development of features which would enable real device geometries to be simulated with common current-voltage curves as an output would be very beneficial. For this the treatment of boundaries and especially contacts should be further investigated; the consideration of non-uniform meshes will become a necessity to treat complex geometries.

The transient behaviour of quantum point contacts, which essentially form a nanocircuit with quantum resistance and capacitance, is not correctly explained by existing circuit theory. This is a problem which could be investigated with two-dimensional Wigner Monte Carlo simulations, using time-dependent boundary conditions. Moreover, the typical size of these structures could make phonon scattering important.

The inclusion of the vector potential (magnetic field) in the Wigner formalism, and specifically the signed-particle method, is very desirable, since an increasing number of devices is utilizing magnetic fields to operate, e.g. magnetic tunnel junctions. Furthermore, the consideration of magnetic fields will give the capability to investigate the quantum mechanical behaviour demonstrated in many experiments, e.g. Aharanov-Bohm rings, to improve our understanding and interpretation.

The capability to simulate time-resolved quantum transport with phonon scattering, magnetic fields and electrostatics taken into account, will provide the possibility to research a wide range of problems in nanoelectronics. Topics of immediate interest are an investigation of structures like nanowires and how discrete dopants and scattering affect the dynamics of single electrons. More novel concepts, like qubits, where entanglement and decoherence are of primary interest, can also be readily investigated, since phonon scattering can be accounted for.

It is an exciting time to be involved in research of nanoelectronics and in light of the above, simulations based in the Wigner formalism can make a valuable contribution to this undertaking.

Appendix A

Phonon Scattering Models

In the following the most important scattering mechanisms for silicon are recollected along with the associated equations for the scattering rates, as are used in the semi-classical Boltzmann transport equation.

The provided equations stem from [110] and assume:

- deformation potential interaction mechanism (covalent materials, e.g. Si);

- ellipsoidal, parabolic bands;

- electrons;

- isotropic phonon scattering.

The values for the scattering parameters for silicon are very well-established. The values presented here are taken from [162], which references the standard literature [108, 163], and provides a complete summary of the scattering parameters for both silicon and germanium.

Phonons are the dominant scattering mechanism at room temperature and can be divided into intervalley and intravalley mechanisms. The total scattering rates for the latter are required to construct a scattering table for Monte Carlo simulations.

Intravalley scattering

Intravalley scattering refers to processes where the initial and final valley of the scattered electron is the same. The scattering can take place through acoustic or optical phonons.

Acoustic phonons

The scattering by acoustic phonons in the same valley is assumed to be elastic by the equipartition approximation. The total scattering rate is given by

$$P_{e^-,ac}(E) \quad = \quad \frac{\sqrt{2}m_{DOS}k_B T \Xi^2}{\pi \hbar^4 u^2 \rho_D}\sqrt{E}; \tag{A.1}$$

$$u \quad = \quad \frac{1}{3}\left(2u_t + u_l\right); \tag{A.2}$$

$$m_{DOS} \quad = \quad \left(m_l m_t^2\right)^{\frac{1}{3}}. \tag{A.3}$$

The meaning and value of the parameters is given in Table A.1.

Table A.1: Parameters for intravalley acoustic phonon scattering

Symbol	Meaning	Value
m_{DOS}	effective mass for density of states	–
m_l	effective mass in longitudinal direction	$0.98\,\mathrm{m_o}$
m_t	effective mass in transverse direction	$0.19\,\mathrm{m_o}$
m_0	mass of an electron	$9.11 \times 10^{-31}\,\mathrm{kg}$
ρ_D	material (silicon) density	$2338.0\,\mathrm{kg \cdot m^{-3}}$
u	average sound (acoustic phonon) velocity	–
u_l	sound velocity in transverse direction	$5.410 \times 10^{7}\,\mathrm{m \cdot s^{-1}}$
u_t	sound velocity in longitudinal direction	$9.033 \times 10^{7}\,\mathrm{m \cdot s^{-1}}$
Ξ	scalar representing average of potential deformation tensor	$7.2\,\mathrm{eV}$

Table A.2: Parameters for intravalley optical phonon scattering

Symbol	Meaning	Value
$D_t K$	optical coupling constant	$2.2 \times 10^{10}\,\mathrm{eV \cdot m^{-1}}$
ω_{op}	radial frequency of optical phonon	–
$\hbar\omega_{op}$	optical phonon energy	$0.0612\,\mathrm{eV}$

Optical phonons

Intravalley scattering by optical phonons is inelastic; energy is lost (gained) through the emission (absorption) of an optical phonon.

The total scattering rate as a function of energy is given by

$$P_{e^-,op}(E) = \frac{(D_t K)^2 m_{DOS}}{\sqrt{2}\pi\hbar^3 \rho_D \omega_{op}} \left[\begin{array}{c} N_{op} \\ N_{op}+1 \end{array} \right] \sqrt{E \pm \hbar\omega_{op}}, \tag{A.4}$$

where the top branch corresponds to the rate for absorption and the bottom branch corresponds to the rate for emission.

The value N_{op} denotes the mean occupancy number for optical phonons under equilibrium conditions and is given by a Bose-Einstein distribution:

$$N_{op} = \left(\frac{\hbar\omega_{op}}{k_B T} - 1 \right)^{-1}. \tag{A.5}$$

The meaning and value of the other newly-introduced parameters are given in Table A.2.

Intravalley scattering with optical phonons is forbidden in the X- and Γ-valleys of silicon, due to symmetry considerations; only the L-valley electrons can undergo intravalley scattering in silicon [164].

Intervalley scattering

Intervalley scattering takes place between different valleys by both acoustic and optical phonons. The total scattering rate, as a function of energy, is given by

$$P_{e^-,op}(E) = \frac{(D_t K)_i m_{DOS} Z_f}{\sqrt{2}\pi\hbar^3 \rho \omega_i} \left[\begin{array}{c} N_i \\ N_i+1 \end{array} \right] \sqrt{E \pm \hbar\omega_{op} - \Delta E_{fi}}; \tag{A.6}$$

$$N_i = \left(\frac{\hbar\omega_i}{k_B T} - 1 \right)^{-1}. \tag{A.7}$$

The equation is formally equivalent to the case of intravalley scattering by optical phonons, when setting Z_f and ΔE_{fi} equal to 1 and 0, respectively. As for the intravalley optical phonons, the top (bottom) branch of the equation yields the rate for absorption (emission).

Scattering can take place between equivalent valleys, e.g. X-X or L-L, or between non-equivalent valleys, e.g. X-L – different scattering parameters apply in each respective case. For a given transition one can distinguish between g-type phonons, which induce transitions between opposite valleys on the same axis in space, and f-type phonons, which induce transitions among orthogonal axes. The case for scattering between equivalent X-valleys is currently implemented, using the values in Table A.3.

Table A.3: Parameters for intervalley scattering (equivalent X-valleys)

Symbol	Meaning	Value
Z_f	number of equivalent final valleys (f-process)	4
Z_f	number of equivalent final valleys (g-process)	1
Δ_{fi}	difference in energy between the minima of the initial and final valley	$0.0\,\mathrm{eV}$
$\hbar\omega_i$	transition energy (f-process)	$0.01896\,\mathrm{eV}$
$\hbar\omega_i$	transition energy (g-process)	$0.01206\,\mathrm{eV}$

Appendix B

Scattering in Discretized k-Space

The scattering models presented in Section 2.2 assume a bulk semiconductor with a continuous spectrum of energy and wavevectors. It is well established that the spatial confinement in nanostructures leads to energy quantization and thereby also influences the scattering rates. However, these effects only become appreciable, if the confinement is below 10 nm. Since the coherence length chosen for a simulation typically easily exceeds this value, the effects of introducing a discretized k-space in the Wigner-Boltzmann equation are subtle and are discussed in the following.

Elastic scattering

An elastic scattering mechanism conserves the kinetic energy of an electron: The magnitude of the k-vector is maintained, whereas its direction changes. The after-scattering components are determined using trigonometric relations, which yield real values for the components of the k-vector that cannot necessarily be represented as an integer multiple of Δk. These real values are rounded up/down to the nearest integer multiple of Δk. These 'residuals' of each component of the k-vector are uniformly distributed over Δk. However, since the kinetic energy of a particle is proportional to $|\mathbf{k}|^2$, rounding up adds more energy than rounding downwards. Thereby, a systematic increase in energy with every elastic scattering event is introduced. Therefore, on average, the 'loss' of values rounded up will match the 'gain' of values rounded down.

Inelastic scattering

An inelastic scattering mechanism, like phonon scattering, makes an electron emit/absorb energy of $\hbar\omega_{\mathbf{q}}$. The number of unique energy values representable by the discrete k-values is determined by the number of unique sum of squares:

$$S = n_1^2 + n_2^2 + n_3^2, \ n_{1,2,3} \in [0, K-1]. \tag{B.1}$$

The resulting energy grid is non-uniform, but remains approximately uniform for energies under $\sim 0.8\,\mathrm{eV}$ (for $\Delta k = \frac{\pi}{100}\,\mathrm{nm}^{-1}$). At higher energies the grid becomes more non-uniform and coarser.

The energy gained/lost through phonon absorption/emission ($\hbar\omega_{\mathbf{g}}$) in general will not exactly correspond to an interval in the discretized energy grid and therefore will be rounded up/down. The approximation introduced by this rounding is below 1 meV for most points in the energy grid. However, at higher energy values, where the energy grid is coarser, the energy gained/lost by optical phonon can be under- or overestimated.

Magnitude of discretization error

Practical computations have shown that, on average, the additional energy introduced per scattering event lies in the order of $10^{-10}\,\mathrm{eV}$, when using a coherence length of 100 nm. This equates to about $2-3\,\mathrm{meV}$ over 100 ps using scattering rates for silicon at room temperature. This error will increase

for coarser k-grids (shorter coherence lengths). Nonetheless, the error can be considered negligible for typical simulation times, which do not surpass a few picoseconds.

Appendix C

Modelling of Surface Roughness

To retain adequate electrostatic control of the conduction channel in a transistor, different architectures, finFETs or UTB-SOI structures are used. The increased surface-to-volume ratio in these architectures has made the effects of scattering due to surface roughness between material interfaces a significant factor in the electrical performance. Surface scattering is now one of the dominant scattering mechanisms in devices with quasi-ballistic transport.

Interface roughness can be modelled by adding random perturbations to a smooth interface. The perturbations are statistically characterized by an autocorrelation function with parameters signifying the mean offset and the correlation length. Measurements have revealed that Si/SiO_2 interfaces can be characterized by an exponential autocorrelation function with a variance and a correlation length. The mean offset can range between $0.1\,nm$ and $0.3\,nm$ for a Si/SiO_2 interface [165, 166].

A numerical synthesis of the one-dimensional roughness, exhibiting the desired statistical properties, can be obtained by first generating a sequence of (uncorrelated) random numbers, which is then convolved with a impulse response (to be chosen) such that a random number sequence is produced with the desired degree of correlation. The Wiener-Kinchin theorem states that the spectral components of an autocorrelation function are given by the power spectral density (PSD) of the sequence used to obtain said function:

$$\mathcal{F}\{r_{xx}\} = P(k).$$
(C.1)

It is desired that $\mathcal{F}\{r_{zz}\}$ has a PSD of an exponential autocorrelation function. Since r_{xx} is generated, its PSD is known (or can be calculated); the transfer function $H(k) = \mathcal{F}\{h\}$, such that

$$\mathcal{F}\{r_{zz}\} = \mathcal{F}\{r_{xx}\} |\mathcal{F}\{h\}|^2.$$
(C.2)

A sequence of perturbations $\{\Delta x\}$ is generated by which the ideal (smooth) interface is displaced. Figure C.1 shows an example of a roughened interface with the parameters given in the caption. To appropriately resolve such perturbations requires a very fine spatial resolution for the potential profile. The latter leads to very high memory demands for the annihilation algorithm, which can be handled by the algorithms presented in the Section 4.3.
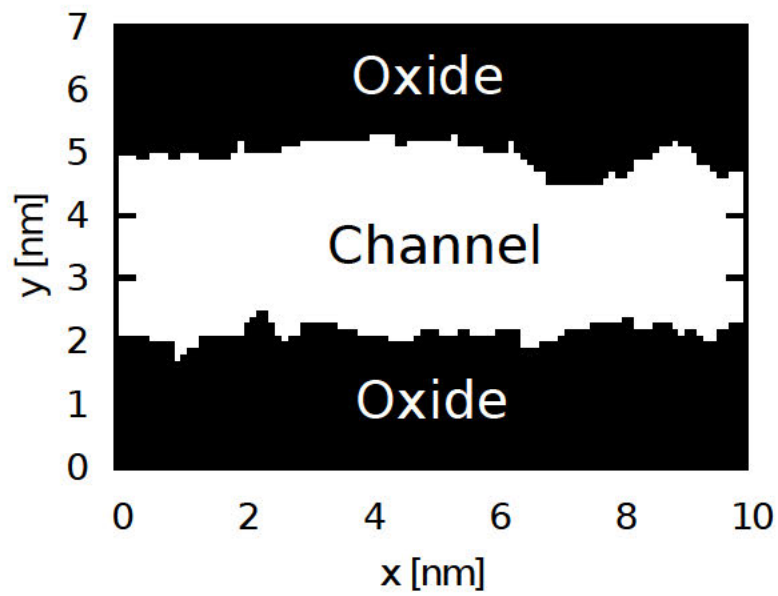
Figure C.1: The geometry represents a 3 nm wide silicon channel between two oxide layers at a resolution of 0.1 nm. The roughness of the Si/SiO$_2$ interface is characterized by an exponential autocorrelation function with a mean displacement, obtained from experiments in [165].

Bibliography

[1] *International Technology Roadmap for Semiconductors (ITRS)*. 2015. URL: http://www.itrs2.net.

[2] C. Tavernier, F. Pereira, O. Nier, D. Rideau, F. Monsieur, G. Torrente, M. Haond et al. 'TCAD Modeling Challenges for 14nm Fully-Depleted SOI Technology Performance Assessment'. In: *Simulation of Semiconductor Processes and Devices (SISPAD), International Conference on*. 2015, pp. 4–7. DOI: 10.1109/SISPAD.2015.7292244.

[3] V. Moroz, L. Smith, J. Huang, M. Choi, T. Ma, J. Liu, Y. Zhang et al. 'Modeling and Optimization of Group IV and III FinFETs and Nano-Wires'. In: *Electron Devices Meeting (IEDM), IEEE International*. 2014, pp. 7.4.1–7.4.4. DOI: 10.1109/IEDM.2014.7047004.

[4] M. Lundstrom. 'Drift-Diffusion and Computational Electronics - Still Going Strong After 40 Years!' In: *Simulation of Semiconductor Processes and Devices (SISPAD), International Conference on*. 2015, pp. 1–3. DOI: 10.1109/SISPAD.2015.7292243.

[5] E. Pop. 'Energy Dissipation and Transport in Nanoscale Devices'. In: *Nano Research* 3.3 (2010), pp. 147–169. DOI: 10.1007/s12274-010-1019-z.

[6] D. K. Ferry and C. Jacoboni, eds. *Quantum Transport in Semiconductors*. Physics of Solids and Liquids. Plenum Press, 1992. ISBN: 0306438534.

[7] C. Jacoboni. *Theory of Electron Transport in Semiconductors; a Pathway from Elementary Physics to Nonequilibrium Green Functions*. Solid-State Sciences; 165. Springer, 2010. ISBN: 9783642105852.

[8] A. Jüngel. *Transport Equations for Semiconductors*. Lecture Notes in Physics ; 773. Springer, 2009. ISBN: 3540895256. DOI: 10.1007/978-3-540-89526-8.

[9] S. Datta. *Electronic Transport in Mesoscopic Systems*. 1. paperback ed. (with corr.), reprint. Cambridge Studies in Semiconductor Physics and Microelectronic Engineering. Cambridge University Press, 2002. ISBN: 0521416043.

[10] D. Dolgos. 'Full-Band Monte Carlo Simulation of Single Photon Avalanche Diodes'. PhD thesis. Eidgenössische Technische Hochschule Zürich, 2011. URL: https://iis.ee.ethz.ch/~schenk/theses/dolgos.pdf.

[11] M. V. Fischetti and S. E. Laux. 'Monte Carlo Analysis of Electron Transport in Small Semiconductor Devices including Band-Structure and Space-Charge Effects'. In: *Physical Review B* 38.14 (1988), pp. 9721–9745. DOI: 10.1103/PhysRevB.38.9721.

[12] T. Ouisse. *Electron Transport in Nanostructures and Mesoscopic Devices: an Introduction*. John Wiley & Sons, 2013. ISBN: 9781848210509.

[13] E. Schöll. *Theory of Transport Properties of Semiconductor Nanostructures*. Vol. 4. Springer Science & Business Media, 2013. ISBN: 9780412731006.

[14] A. Gehring and S. Selberherr. 'Evolution of Current Transport Models for Engineering Applications'. In: *Journal of Computational Electronics* 3.3-4 (2004), pp. 149–155. DOI: 10.1007/s10825-004-7035-z.

[15] U. Mishra and J. Singh. *Semiconductor Device Physics and Design*. Springer Science & Business Media, 2007. ISBN: 9781402064814.

[16]   S. Selberherr. *Analysis and Simulation of Semiconductor Devices.* Springer, 1984. ISBN: 3211818006.

[17]   M. Lundstrom. *Fundamentals of Carrier Transport.* Cambridge University Press, 2009. ISBN: 9780521637244.

[18]   T. Grasser, H. Kosina, M. Gritsch and S. Selberherr. 'Using Six Moments of Boltzmann's Transport Equation for Device Simulation'. In: *Journal of Applied Physics* 90.5 (2001), pp. 2389–2396. DOI: 10.1063/1.1389757.

[19]   A. Jüngel, S. Krause and P. Pietra. 'A Hierarchy of Diffusive Higher-Order Moment Equations for Semiconductors'. In: *SIAM Journal on Applied Mathematics* 68.1 (2007), pp. 171–198. DOI: 10.1137/070683313.

[20]   D. Vasileska and S. Goodnick. *Nano-Electronic Devices: Semiclassical and Quantum Transport Modeling.* SpringerLink. Springer, 2011. ISBN: 9781441988409.

[21]   W. Van Roosbroeck. 'Theory of the Flow of Electrons and Holes in Germanium and other Semiconductors'. In: *Bell System Technical Journal* 29.4 (1950), pp. 560–607. DOI: 10.1002/j.1538-7305.1950.tb03653.x.

[22]   R. Stratton. 'Diffusion of Hot and Cold Electrons in Semiconductor Barriers'. In: *Physical Review* 126 (6 1962), pp. 2002–2014. DOI: 10.1103/PhysRev.126.2002.

[23]   N. B. Abdallah, P. Degond and S. Génieys. 'An Energy-Transport Model for Semiconductors Derived from the Boltzmann Equation'. In: *Journal of Statistical Physics* 84.1-2 (1996), pp. 205–231. DOI: 10.1007/BF02179583.

[24]   K. Blotekjaer. 'Transport Equations for Electrons in Two-Valley Semiconductors'. In: *Electron Devices, IEEE Transactions on* 17.1 (1970), pp. 38–47. DOI: 10.1109/T-ED.1970.16921.

[25]   R. Kosik. 'Numerical Challenges on the Road to NanoTCAD'. PhD thesis. Technische Universität Wien, 2004. URL: http://www.iue.tuwien.ac.at/phd/kosik/.

[26]   C. S. Lent and D. J. Kirkner. 'The Quantum Transmitting Boundary Method'. In: *Journal of Applied Physics* 67.10 (1990), pp. 6353–6359. DOI: 10.1063/1.345156.

[27]   F. Rossi and T. Kuhn. 'Theory of Ultrafast Phenomena in Photoexcited Semiconductors'. In: *Reviews of Modern Physics* 74.3 (2002), pp. 895–950. DOI: 10.1103/RevModPhys.74.895.

[28]   M. Pourfath. *The Non-Equilibrium Green's Function Method for Nanoscale Device Simulation.* Springer, 2014. ISBN: 9783709117996.

[29]   B. Gaury, J. Weston, M. Santin, M. Houzet, C. Groth and X. Waintal. 'Numerical Simulations of Time-Resolved Quantum Electronics'. In: *Physics Reports* 534.1 (2014), pp. 1–37. DOI: 10.1016/j.physrep.2013.09.001.

[30]   B. Novakovic and G. Klimeck. 'Atomistic Quantum Transport Approach to Time-Resolved Device Simulations'. In: *Simulation of Semiconductor Processes and Devices (SISPAD), International Conference on.* 2015, pp. 8–11. DOI: 10.1109/SISPAD.2015.7292245.

[31]   V. Sverdlov, E. Ungersboeck, H. Kosina and S. Selberherr. 'Current Transport Models for Nanoscale Semiconductor Devices'. In: *Materials Science and Engineering: Reports* 58.6 (2008), pp. 228–270. DOI: 10.1016/j.mser.2007.11.001.

[32]   M. Wagner, M. Karner and K.-T. Grasser. 'Quantum Correction Models for Modern Semiconductor Devices'. In: *Proceedings of the 13th International Workshop on Semiconductor Devices.* Vol. 1. 2005, pp. 458–461.

[33]   D. Ferry, R. Akis and D. Vasileska. 'Quantum Effects in MOSFETs: Use of an Effective Potential in 3D Monte Carlo Simulation of Ultra-Short Channel Devices'. In: *Electron Devices Meeting (IEDM), IEEE International.* 2000, pp. 287–290. DOI: 10.1109/IEDM.2000.904313.

[34]   D. Ferry, S. Ramey, L. Shifren and R. Akis. 'The Effective Potential in Device Modeling: The good, the Bad and the Ugly'. In: *Journal of Computational Electronics* 1.1-2 (2002), pp. 59–65. DOI: 10.1023/A:1020763710906.

[35] Y. Li, T.-w. Tang and X. Wang. 'Modeling of Quantum Effects for Ultrathin Oxide MOS Structures with an Effective Potential'. In: *Nanotechnology, IEEE Transactions on* 1.4 (2002), pp. 238–242. DOI: 10.1109/TNANO.2002.807386.

[36] A. Asenov, A. Brown and J. Watling. 'Quantum Corrections in the Simulation of Decanano MOSFETs'. In: *Solid-State Electronics* 47.7 (2003), pp. 1141–1145. DOI: 10.1016/S0038-1101(03)00030-3.

[37] M. Ancona. 'Multi-dimensional Semiconductor Tunneling in Density-Gradient Theory'. In: *Book of Abstracts of the 17th International Workshop on Computational Electronics (IWCE)*. 2006, pp. 9–10. ISBN: 3901578161.

[38] O. Baumgartner, L. Filipovic, H. Kosina, M. Karner, Z. Stanojevic and H. Cheng-Karner. 'Efficient Modeling of Source/Drain Tunneling in Ultra-Scaled Transistors'. In: *Simulation of Semiconductor Processes and Devices (SISPAD), International Conference on*. 2015, pp. 202–205. DOI: 10.1109/SISPAD.2015.7292294.

[39] O. Badami, N. Kumar, D. Saha and S. Ganguly. 'Quantum Drift-Diffusion and Quantum Energy Balance Simulation of Nanowire Junctionless Transistors'. In: *Silicon Nanoelectronics Workshop (SNW), IEEE*. 2012, pp. 1–2. DOI: 10.1109/SNW.2012.6243303.

[40] M. Vasicek. 'Advanced Macroscopic Transport Models'. PhD thesis. Technische Universität Wien, 2009. URL: http://www.iue.tuwien.ac.at/phd/vasicek/.

[41] M. Nedjalkov. 'Book Chapter: Wigner Transport in Presence of Phonons: Particle Models of the Electron Kinetics'. In: *From Nanostructures to Nanosensing Applications, Proceedings of the International School of Physics Enrico Fermi*. Ed. by A. Paoletti, A. D'Amico and G. Ballestrino. Vol. 160. IOS Press, 2005, pp. 55–103. ISBN: 1586035274. DOI: 10.3254/1-58603-527-4-55.

[42] F. A. Buot. *Nonequilibrium Quantum Transport Physics in Nanosystems; Foundation of Computational Nonequilibrium Physics in Nanoscience and Nanotechnology*. World Scientific, 2009. ISBN: 9812566791.

[43] V. I. Tatarskii. 'The Wigner Representation of Quantum Mechanics'. In: *Sov. Phys. Usp.* 26 (1983), pp. 311–327. DOI: 10.1070/PU1983v026n04ABEH004345.

[44] D. Leibfried, T. Pfau and C. Monroe. 'Shadows and Mirrors: Reconstructing Quantum States of Atom Motion'. In: *Print edition* 51.4 (1998), pp. 22–28. DOI: 10.1063/1.882256.

[45] E. Colomés, Z. Zhan and X. Oriols. 'Comparing Wigner, Husimi and Bohmian Distributions: Which One is a True Probability Distribution in Phase Space?' In: *Journal of Computational Electronics* 14.4 (2015), pp. 894–906. DOI: 10.1007/s10825-015-0737-6.

[46] D. Querlioz and P. Dollfus. *The Wigner Monte Carlo Method for Nanoelectronic Devices - A Particle Description of Quantum Transport and Decoherence*. ISTE-Wiley, 2010. ISBN: 9781848211506.

[47] M. Nedjalkov, D. Vasileska, D. K. Ferry, C. Jacoboni, C. Ringhofer, I. Dimov and V. Palankovski. 'Wigner Transport Models of the Electron-Phonon Kinetics in Quantum Wires'. In: *Physical Review B* 74 (3 2006), p. 035311. DOI: 10.1103/PhysRevB.74.035311.

[48] M. Nedjalkov, D. Querlioz, P. Dollfus and H. Kosina. 'Review Chapter: Wigner Function Approach'. In: *Nano-Electronic Devices: Semiclassical and Quantum Transport Modeling*. Ed. by D. Vasileska and S. Goodnick. Springer, 2011, pp. 289–358. ISBN: 9781441988393. DOI: 10.1007/978-1-4419-8840-9_5.

[49] M. Nedjalkov, S. Selberherr, D. Ferry, D. Vasileska, P. Dollfus, D. Querlioz, I. Dimov et al. 'Physical Scales in the Wigner-Boltzmann Equation'. In: *Annals of Physics* 328 (2012), pp. 220–237. DOI: 10.1016/j.aop.2012.10.001.

[50] S. Barraud. 'Dissipative Quantum Transport in Silicon Nanowires based on Wigner Transport Equation'. In: *Journal of Applied Physics* 110.9 (2011), p. 093710. DOI: 10.1063/1.3654143.

[51] I. Knezevic. Personal Communication and Presentation at Internal Meeting of Wigner Initiave, Vienna (Austria). http://www.iue.tuwien.ac.at/wigner-wiki. 22-24 July 2015.

[52] E. Wigner. 'On the Quantum Correction For Thermodynamic Equilibrium'. In: *Physical Review* 40 (5 1932), pp. 749–759. DOI: 10.1103/PhysRev.40.749.

[53] J. E. Moyal. 'Quantum Mechanics as a Statistical Theory'. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 45 (01 1949), pp. 99–124. DOI: 10.1017/S0305004100000487.

[54] N. C. Dias and J. N. Prata. 'Admissible States in Quantum Phase Space'. In: *Annals of Physics* 313.1 (2004), pp. 110–146. DOI: 10.1016/j.aop.2004.03.008.

[55] H. Weyl. 'Quantenmechanik und Gruppentheorie'. In: *Zeitschrift für Physik* 46.1-2 (1927), pp. 1–46. DOI: 10.1007/BF02055756.

[56] K.-Y. Kim and B. Lee. 'Wigner Function Formulation in Nonparabolic Semiconductors using Power Series Dispersion Relation'. In: *Journal of Applied Physics* 86.9 (1999), pp. 5085–5093. DOI: 10.1063/1.371484.

[57] L. Demeio, P. Bordone and C. Jacoboni. 'Multiband, Non-Parabolic Wigner-Function Approach to Electron Transport in Semiconductors'. In: *Transport Theory and Statistical Physics* 34.7 (2005), pp. 499–522. DOI: 10.1080/00411450508951151.

[58] R. Hudson. 'When is the Wigner Quasi-Probability Density Non-negative?' In: *Reports on Mathematical Physics* 6.2 (1974), pp. 249–252. DOI: 10.1016/0034-4877(74)90007-X.

[59] M. Nedjalkov, I. Dimov, P. Bordone, R. Brunetti and C. Jacoboni. 'Using the Wigner Function for Quantum Transport in Device Simulation'. In: *Mathematical and Computer Modelling* 25.12 (1997), pp. 33–53. DOI: 10.1016/S0895-7177(97)00093-9.

[60] M. Nedjalkov, I. Dimov, F. Rossi and C. Jacoboni. 'Convergency of the Monte Carlo Algorithm for the Solution of the Wigner Quantum-Transport Equation'. In: *Mathematical and Computer Modelling* 23.8–9 (1996), pp. 159–166. DOI: 10.1016/0895-7177(96)00047-7.

[61] N. Kluksdahl, A. Kriman, C. Ringhofer and D. Ferry. 'Quantum Tunneling Properties from a Wigner Function Study'. In: *Solid-State Electronics* 31.3 (1988), pp. 743–746. DOI: 10.1137/S0036142901388366.

[62] W. Frensley. 'Boundary Conditions for Open Quantum Systems Driven far from Equilibrium'. In: *Reviews of Modern Physics* 62.3 (1990), pp. 745–789. DOI: 10.1103/RevModPhys.62.745.

[63] M. Nedjalkov, H. Kosina, S. Selberherr, C. Ringhofer and D. Ferry. 'Unified Particle Approach to Wigner-Boltzmann Transport in Small Semiconductor Devices'. In: *Physical Review B* 70.11 (2004), p. 115319. DOI: 10.1103/PhysRevB.70.115319.

[64] D. Querlioz and P. Dollfus. *The Wigner Monte Carlo Method for Nanoelectronic Devices - A Particle Description of Quantum Transport and Decoherence*. ISTE-Wiley, 2010. ISBN: 9781848211506.

[65] S. Barraud, P. Dollfus, S. Galdin and P. Hesto. 'Short-range and Long-range Coulomb Interactions for 3D Monte Carlo Device Simulation with Discrete Impurity Distribution'. In: *Solid-State Electronics* 46.7 (2002), pp. 1061–1067. DOI: 10.1016/S0038-1101(02)00042-4.

[66] A. Arnold, H. Lange and P. F. Zweifel. 'A Discrete-Velocity, Stationary Wigner Equation'. In: *Journal of Mathematical Physics* 41.11 (2000), pp. 7167–7180. DOI: 10.1063/1.1318732.

[67] T. Goudon. 'Analysis of a Semidiscrete Version of the Wigner Equation'. In: *SIAM Journal on Numerical Analysis* 40.6 (2002), pp. 2007–2025. DOI: 10.1137/S0036142901388366.

[68] P. Bordone, M. Pascoli, R. Brunetti, A. Bertoni, C. Jacoboni and A. Abramo. 'Quantum Transport of Electrons in Open Nanostructures with the Wigner-Function Formalism'. In: *Physical Review B* 59.4 (1999), pp. 3060–3069. DOI: 10.1103/PhysRevB.59.3060.

[69] A. Baute, I. Egusquiza and J. Muga. 'Sources of Quantum Waves'. In: *Journal of Physics A: Mathematical and General* 34.20 (2001), pp. 4289–4299. DOI: 10.1088/0305-4470/34/20/303.

[70] A. Arnold. 'Mathematical Concepts of Open Quantum Boundary Conditions'. In: *Transport Theory and Statistical Physics* 30.4-6 (2001), pp. 561–584. DOI: `10.1081/TT-100105939`.

[71] T. González and D. Pardo. 'Physical Models of Ohmic Contact for Monte Carlo Device Simulation'. In: *Solid-State Electronics* 39.4 (1996), pp. 555–562. DOI: `10.1016/0038-1101(95)00188-3`.

[72] H. Jiang, W. Cai and R. Tsu. 'Accuracy of the Frensley Inflow Boundary Condition for Wigner Equations in Simulating Resonant Tunneling Diodes'. In: *Journal of Computational Physics* 230.5 (2011), pp. 2031–2044. DOI: `10.1016/j.jcp.2010.12.002`.

[73] A. Savio and A. Poncet. 'Study of the Wigner Function at the Device Boundaries in One-Dimensional Single-and Double-Barrier Structures'. In: *Journal of Applied Physics* 109.3 (2011), p. 033713. DOI: `10.1063/1.3526969`.

[74] H. Jiang, T. Lu and W. Cai. 'A Device Adaptive Inflow Boundary Condition for Wigner Equations of Quantum Transport'. In: *Journal of Computational Physics* 258 (2014), pp. 773–786. DOI: `10.1016/j.jcp.2013.11.007`.

[75] I. Dimov, M. Nedjalkov, J.-M. Sellier and S. Selberherr. 'Boundary Conditions and the Wigner Equation Solution'. In: *Journal of Computational Electronics* 14.4 (2015), pp. 859–863. DOI: `10.1007/s10825-015-0720-2`.

[76] R. Rosati, F. Dolcini, R. C. Iotti and F. Rossi. 'Wigner-Function Formalism Applied to Semiconductor Quantum Devices: Failure of the Conventional Boundary Condition Scheme'. In: *Physical Review B* 88.3 (2013), p. 035401. DOI: `10.1103/PhysRevB.88.035401`.

[77] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry and C. Ringhofer. 'Self-Consistent Study of the Resonant-Tunneling Diode'. In: *Physical Review B* 39.11 (1989), pp. 7720–7735. DOI: `10.1103/PhysRevB.39.7720`.

[78] C. Ringhofer, D. Ferry and N. Kluksdahl. 'Absorbing Boundary Conditions for the Simulation of Quantum Transport Phenomena'. In: *Transport Theory and Statistical Physics* 18.3-4 (1989), pp. 331–346. DOI: `10.1080/00411458908204692`.

[79] A. Arnold. 'On Absorbing Boundary Conditions for Quantum Transport Equations'. In: *RAIRO - Modélisation Mathématique et Analyse Numérique* 28.7 (1994), pp. 853–872.

[80] W. R. Frensley. 'Wigner-function Model of a Resonant-Tunneling Semiconductor Device'. In: *Physical Review B* 36.3 (1987), pp. 1570–1580. DOI: `10.1103/PhysRevB.36.1570`.

[81] I. Knezevic. 'Decoherence due to Contacts in Callistic Nanostructures'. In: *Physical Review B* 77 (12 2008), p. 125301. DOI: `10.1103/PhysRevB.77.125301`.

[82] J. Sellier and I. Dimov. 'Wigner Functions, Signed Particles, and the Harmonic Oscillator'. In: *Journal of Computational Electronics* 14.4 (2015), pp. 907–915. DOI: `10.1007/s10825-015-0722-0`.

[83] U. Ravaioli, M. A. Osman, W. Pötz, N. Kluksdahl and D. K. Ferry. 'Investigation of Ballistic Transport through Resonant-Tunnelling Quantum Wells using Wigner Function Approach'. In: *Physica B+C* 134.1-3 (1985), pp. 36–40. DOI: `10.1016/0378-4363(85)90317-1`.

[84] N. Kluksdahl, W. Pötz, U. Ravaioli and D. Ferry. 'Wigner Function Study of a Double Quantum Barrier Resonant Tunnelling Diode'. In: *Superlattices and Microstructures* 3.1 (1987), pp. 41–45. DOI: `10.1016/0749-6036(87)90175-3`.

[85] K.-Y. Kim and B. Lee. 'On the High Order Numerical Calculation Schemes for the Wigner Transport Equation'. In: *Solid-State Electronics* 43.12 (1999), pp. 2243–2245. DOI: `10.1016/S0038-1101(99)00168-9`.

[86] C. Ringhofer. 'A Spectral Method for the Numerical Simulation of Quantum Tunneling Phenomena'. In: *SIAM Journal on Numerical Analysis* 27.1 (1990), pp. 32–50. DOI: `10.1137/0727003`.

[87] A. Arnold and C. Ringhofer. 'Operator Splitting Methods Applied to Spectral Discretizations of Quantum Transport Equations'. In: *SIAM Journal on Numerical Analysis* 32.6 (1995), pp. 1876–1894. DOI: `10.1137/0732084`.

[88] S. Shao, T. Lu and W. Cai. 'Adaptive Conservative Cell Average Spectral Element Methods for Transient Wigner Equation in Quantum Transport'. In: *Communications of Computational Physics* 9 (2011), pp. 711–739. DOI: `10.4208/cicp.080509.310310s`.

[89] A. Dorda and F. Schürrer. 'A WENO-Solver Combined with Adaptive Momentum Discretization for the Wigner Transport Equation and its Application to Resonant Tunneling Diodes'. In: *Journal of Computational Physics* 284 (2015), pp. 95–116. DOI: `10.1016/j.jcp.2014.12.026`.

[90] J. Cervenka, P. Ellinghaus and M. Nedjalkov. 'Deterministic Solution of the Discrete Wigner Equation'. In: *Numerical Methods and Applications.* Ed. by I. Dimov, S. Fidanova and I. Lirkov. Springer International Publishing, 2015, pp. 149–156. ISBN: 9783319155845. DOI: `10.1007/978-3-319-15585-2_17`.

[91] V. Peikert and A. Schenk. 'A Wavelet Method to Solve High-Dimensional Transport Equations in Semiconductor Devices'. In: *Simulation of Semiconductor Processes and Devices (SISPAD), International Conference on.* 2011, pp. 299–302. DOI: `10.1109/SISPAD.2011.6035029`.

[92] S.-M. Hong, A.-T. Pham and C. Jungemann. *Deterministic Solvers for the Boltzmann Transport Equation.* Springer Science & Business Media, 2011. ISBN: 9783709107775.

[93] K. Rupp. 'Deterministic Numerical Solution of the Boltzmann Transport Equation'. PhD thesis. Technische Universität Wien, 2011. URL: `http://www.iue.tuwien.ac.at/phd/rupp/`.

[94] N. Goldsman, C.-K. Lin, Z. Han and C.-K. Huang. 'Advances in the Spherical Harmonic-Boltzmann-Wigner Approach to Device Simulation'. In: *Superlattices and Microstructures* 27.2 (2000), pp. 159–175. DOI: `10.1006/spmi.1999.0810`.

[95] P. Vitanov, M. Nedjalkov, C. Jacoboni, F. Rossi and A. Abramo. 'Unified Monte Carlo Approach to the Boltzmann and Wigner Equations'. In: *Advances in Parallel Algorithms.* Ed. by B. Sendov and I. Dimov. IOS Press, 1994, pp. 117–128.

[96] F. Rossi, C. Jacoboni and M. Nedjalkov. 'A Monte Carlo Solution of the Wigner Transport Equation'. In: *Semiconductor Science Technology* 9 (1994), pp. 934–936. DOI: `10.1088/0268-1242/9/5S/143`.

[97] R. Sala, S. Brouard and J. G. Muga. 'Wigner Trajectories and Liouville's Theorem'. In: *The Journal of Chemical Physics* 99.4 (1993), pp. 2708–2714. DOI: `10.1063/1.465232`.

[98] P. Bordone, A. Bertoni, R. Brunetti and C. Jacoboni. 'Monte Carlo Simulation of Quantum Electron Transport based on Wigner Paths'. In: *Mathematics and Computers in Simulation* 62.3-6 (2003), pp. 307–314. DOI: `10.1016/S0378-4754(02)00241-0`.

[99] L. Shifren and D. Ferry. 'A Wigner Function Based Ensemble Monte Carlo Approach for Accurate Incorporation of Quantum Effects in Device Simulation'. In: *Journal of Computational Electronics* 1.1-2 (2002), pp. 55–58. DOI: `10.1023/A:1020711726836`.

[100] L. Shifren, C. Ringhofer and D. Ferry. 'A Wigner Function-based Quantum Ensemble Monte Carlo Study of a Resonant Tunneling Diode'. In: *Electron Devices, IEEE Transactions on* 50.3 (2003), pp. 769–773. DOI: `10.1109/TED.2003.809434`.

[101] D. Querlioz, P. Dollfus, V.-N. Do, A. Bournel and V. L. Nguyen. 'An Improved Wigner Monte Carlo Technique for the Self-Consistent Simulation of RTDs'. In: *Journal of Computational Electronics* 5.4 (2006), pp. 443–446. DOI: `10.1007/s10825-006-0044-3`.

[102] *ViennaWD - Wigner Ensemble Monte Carlo Simulator.* URL: `http://viennawd.sourceforge.net/`.

[103] R. Eckhardt. 'Stan Ulam, John von Neumann, and the Monte Carlo Method'. In: *Los Alamos Science* 15 (1987), pp. 131–136.

[104] T. Kurosawa. 'Monte Carlo calculation of Hot Electron Problems'. In: *Journal of the Physical Society of Japan* 21 (1966), p. 424.

[105] I. Dimov. *Monte Carlo Methods for Applied Scientists*. World Scientific, 2008. ISBN: 9810223293.

[106] I. Dimov. 'Optimal Monte Carlo Algorithms'. In: *Modern Computing, John Vincent Atanasoff International Symposium on*. 2006, pp. 125–131. DOI: 10.1109/JVA.2006.37.

[107] I. Dimov and R. Georgieva. 'Complexity of Monte Carlo Algorithms for a Class of Integral Equations'. In: *International Conference on Computational Science (ICCS), Proceedings of the 7th*. Ed. by Y. Shi, G. D. Albada, J. Dongarra and P. M. A. Sloot. Springer, 2007, pp. 731–738. DOI: 10.1007/978-3-540-72584-8_97.

[108] C. Jacoboni and L. Reggiani. 'The Monte Carlo Method for the Solution of Charge Transport in Semiconductors with Applications to Covalent Materials'. In: *Reviews of Modern Physics* 55.3 (1983), p. 645. DOI: 10.1103/RevModPhys.55.645.

[109] K. Tomizawa. *Numerical Simulation of Submicron Semiconductor Devices*. Artech House, 1993. ISBN: 0890066205.

[110] C. Jacoboni and P. Lugli. *The Monte Carlo Method for Semiconductor Device Simulation*. Springer Science & Business Media, 2012. ISBN: 9783709174531.

[111] P. Poli, L. Rota and C. Jacoboni. 'Weighted Monte Carlo for Electron Transport in Semiconductors'. In: *Applied Physics Letters* 55.10 (1989), pp. 1026–1028. DOI: 10.1063/1.101723.

[112] C. Jacoboni, P. Poli and L. Rota. 'A New Monte Carlo Technique for the Solution of the Boltzmann Transport Equation'. In: *Solid-State Electronics* 31.3 (1988), pp. 523–526. DOI: 10.1016/0038-1101(88)90332-2.

[113] M. Nedjalkov and P. Vitanov. 'Iteration Approach for Solving the Boltzmann Equation with the Monte Carlo Method'. In: *Solid-State Electronics* 32.10 (1989), pp. 893–896. DOI: 10.1016/0038-1101(89)90067-1.

[114] H. Kosina, M. Nedjalkov and S. Selberherr. 'The Stationary Monte Carlo Method for Device Simulation. I. Theory'. In: *Journal of Applied Physics* 93.6 (2003), pp. 3553–3563. DOI: 10.1063/1.1544654.

[115] R. Chambers. 'The Kinetic Formulation of Conduction Problems'. In: *Proceedings of the Physical Society. Section A* 65.6 (1952), pp. 458–468. DOI: 10.1088/0370-1298/65/6/114.

[116] H. ZhiMin, Y. ZaiZai and C. JianRui. 'Monte Carlo Method for Solving the Fredholm Integral Equations of the Second Kind'. In: *Transport Theory and Statistical Physics* 41.7 (2012), pp. 513–528. DOI: 10.1080/00411450.2012.695317.

[117] I. Dimov and T. Gurov. 'Monte Carlo Algorithm for Solving Integral Equations with Polynomial Non-linearity: Parallel Implementation'. In: *Pliska Studia Mathematica Bulgarica* 13.1 (2000), pp. 117–132.

[118] I. Dimov and T. Gurov. 'Estimates of the Computational Complexity of Iterative Monte Carlo Algorithm Based on Green's Function Approach'. In: *Mathematics and Computers in Simulation* 47.2–5 (1998), pp. 183–199. DOI: 10.1016/S0378-4754(98)00102-5.

[119] T. Gurov, P. Whitlock and I. Dimov. 'A Grid Free Monte Carlo Algorithm for Solving Elliptic Boundary Value Problems'. In: *Numerical Analysis and Its Applications (NAA), Second International Conference on*. Ed. by L. Vulkov, P. Yalamov and J. Waśniewski. Springer, 2001, pp. 359–367. DOI: 10.1007/3-540-45262-1_42.

[120] I. Dimov and R. Papancheva. 'Green's Function Monte Carlo Algorithms for Elliptic Problems'. In: *Mathematics and Computers in Simulation* 63.6 (2003), pp. 587–604. DOI: 10.1016/S0378-4754(03)00094-6.

[121] K. Chatterjee, J. R. Roadcap and S. Singh. 'A New Green's Function Monte Carlo Algorithm for the Solution of the Two-Dimensional Nonlinear Poisson–Boltzmann Equation: Application to the Modeling of the Communication Breakdown Problem in Space Vehicles during Re-Entry'. In: *Journal of Computational Physics* 276 (2014), pp. 479–485. DOI: 10.1016/j.jcp.2014.07.042.

[122] I. Dimov. 'Monte Carlo Algorithms for Linear Problems'. In: *Pliska Studia Mathematica Bulgarica* 13.1 (2000), pp. 55–77.

[123] H. Kosina, M. Nedjalkov and S. Selberherr. 'Theory of the Monte Carlo Method for Semiconductor Device Simulation'. In: *IEEE Transactions on Electron Devices* 47.10 (2000), pp. 1898–1908. DOI: 10.1109/16.870569.

[124] F. W. Byron and W. Fuller. *Mathematics of Quantum and Classical Physics.* Dover, 1992. ISBN: 9780486671642.

[125] Y. Fu and M. Willander. 'Electron Wave-Packet Transport through Nanoscale Semiconductor Device in Time Domain'. In: *Journal of Applied Physics* 97.9 (2005), p. 094311. DOI: 10.1063/1.1890452.

[126] H. Rees. 'Calculation of Distribution Functions by Exploiting the Stability of the Steady State'. In: *Journal of Physics and Chemistry of Solids* 30.3 (1969), pp. 643–655. DOI: 10.1016/0022-3697(69)90018-3.

[127] *Lua Scripting Language.* URL: http://www.lua.org/.

[128] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Efficient Calculation of the Two-Dimensional Wigner Potential'. In: *Computational Electronics (IWCE), International Workshop on.* 2014, pp. 1–3. DOI: 10.1109/IWCE.2014.6865812.

[129] M. Frigo and S. Johnson. 'The Design and Implementation of FFTW3'. In: *Proceedings of the IEEE* 93.2 (2005), pp. 216–231. DOI: 10.1109/JPROC.2004.840301.

[130] H.-Y. Huang, Y.-Y. Lee and P.-C. Lo. 'A Novel Algorithm for Computing the 2D Split-Vector-Radix FFT'. In: *Signal Processing* 84.3 (2004), pp. 561–570. DOI: 10.1016/j.sigpro.2003.11.018.

[131] Z. Chen and L. Zhang. 'Vector Coding Algorithms for Multidimensional Discrete Fourier Transform'. In: *Journal of Computational and Applied Mathematics* 212.1 (2008), pp. 63–74. DOI: 10.1016/j.cam.2006.11.025.

[132] E. Jacobsen and R. Lyons. 'The Sliding DFT'. In: *Signal Processing Magazine, IEEE* 20.2 (2003), pp. 74–80. DOI: 10.1109/MSP.2003.1184347.

[133] E. Jacobsen and R. Lyons. 'An Update to the Sliding DFT'. In: *Signal Processing Magazine, IEEE* 21.1 (2004), pp. 110–111. DOI: 10.1109/MSP.2004.1516381.

[134] D. Ferry and S. M. Goodnick. *Transport in Nanostructures.* 6. Cambridge University Press, 1997. ISBN: 9780521663656.

[135] C. Jacoboni and P. Bordone. 'Wigner Transport Equation with Finite Coherence Length'. In: *Journal of Computational Electronics* 13.1 (2014), pp. 257–263. DOI: 10.1007/s10825-013-0510-7.

[136] F. Harris. 'On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform'. In: *Proceedings of the IEEE* 66.1 (1978), pp. 51–83. DOI: 10.1109/PROC.1978.10837.

[137] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Implications of the Coherence Length on the Discrete Wigner Potential'. In: *Computational Electronics (IWCE), International Workshop on.* 2014, pp. 1–3. DOI: 10.1109/IWCE.2014.6865852.

[138] J. W. Tukey. 'An Introduction to the Calculation of Numerical Spectrum Analysis'. In: *Spectral Analysis of Time Series* (1967). Ed. by B. Harris, pp. 22–46.

[139] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Optimized Particle Regeneration Scheme for the Wigner Monte Carlo Method'. In: *Lecture Notes in Computer Science.* Ed. by I. Dimov, S. Fidanova and I. Lirkov. Vol. 8962. Springer International Publishing, 2015, pp. 27–33. ISBN: 9783319155845. DOI: 10.1007/978-3-319-15585-2_3.

[140] N. Akhiezer. *The Classical Moment Problem: and Some Related Questions in Analysis.* University Mathematical Monographs. Out of Print. Oliver & Boyd, 1965.

[141] C. A. R. Hoare. 'Algorithm 64: Quicksort'. In: *Communications of the ACM* 4.7 (1961), pp. 321–322. DOI: 10.1145/366622.366644.

[142] V. Los and N. Los. 'Exact Solution of the One-Dimensional Time-Dependent Schrödinger Equation with a Rectangular Well/Barrier Potential and its Applications'. In: *Theoretical and Mathematical Physics* 177.3 (2013), pp. 1706–1721. DOI: 10.1007/s11232-013-0128-8.

[143] J. M. Sellier, M. Nedjalkov, I. Dimov and S. Selberherr. 'A Benchmark Study of the Wigner Monte Carlo Method'. In: *Monte Carlo Methods and Applications* 20.1 (2014), pp. 43–51. DOI: 10.1515/mcma-2013-0018.

[144] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'The Wigner Monte Carlo Method for Accurate Semiconductor Device Simulation'. In: *Simulation of Semiconductor Processes and Devices (SISPAD), International Conference on.* 2014, pp. 113–116. DOI: 10.1109/SISPAD.2014.6931576.

[145] C. Moore. 'Data Processing in Exascale-Class Computer Systems'. In: *The Salishan Conference on High Speed Computing.* 2011. URL: http://www.lanl.gov/conferences/salishan/.

[146] L. Dagum and R. Enon. 'OpenMP: an Industry Standard API for Shared-Memory Programming'. In: *Computational Science & Engineering* 5.1 (1998), pp. 46–55. DOI: 10.1109/99.660313.

[147] M. Snir, S. Otto, S. Huss-Lederman, D. Walker and J. Dongarra. *MPI-The Complete Reference, Volume 1: The MPI Core.* 2nd. (Revised). MIT Press, 1998. ISBN: 0262692155.

[148] P. Ellinghaus, J. Weinbub, M. Nedjalkov, S. Selberherr and I. Dimov. 'Distributed-Memory Parallelization of the Wigner Monte Carlo Method using Spatial Domain Decomposition'. In: *Journal of Computational Electronics* 14.1 (2015), pp. 151–162. DOI: 10.1007/s10825-014-0635-3.

[149] J. Weinbub, P. Ellinghaus and S. Selberherr. 'Parallelization of the Two-Dimensional Wigner Monte Carlo Method'. In: *Large-Scale Scientific Computing.* Ed. by I. Lirkov, S. D. Margenov and J. Waśniewski. Vol. 9374. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 309–316. ISBN: 9783319265193. DOI: 10.1007/978-3-319-26520-9_34.

[150] J. Weinbub, P. Ellinghaus and M. Nedjalkov. 'Domain Decomposition Strategies for the Two-Dimensional Wigner Monte Carlo Method'. In: *Journal of Computational Electronics* 14.4 (2015), pp. 922–929. DOI: 10.1007/s10825-015-0730-0.

[151] *Vienna Scientific Cluster - VSC-3.* URL: http://vsc.ac.at/.

[152] U. Sivan, M. Heiblum, C. P. Umbach and H. Shtrikman. 'Electrostatic Electron Lens in the Ballistic Regime'. In: *Physical Review B* 41 (11 1990), pp. 7937–7940. DOI: 10.1103/PhysRevB.41.7937.

[153] J. Spector, H. L. Stormer, K. W. Baldwin, L. N. Pfeiffer and K. W. West. 'Electron Focusing in Two-Dimensional Systems by means of an Electrostatic Lens'. In: *Applied Physics Letters* 56.13 (1990), pp. 1290–1292. DOI: 10.1063/1.102538.

[154] R. Wang, H. Liu, R. Huang, J. Zhuge, L. Zhang, D.-W. Kim, X. Zhang et al. 'Experimental Investigations on Carrier Transport in Si Nanowire Transistors: Ballistic Efficiency and Apparent Mobility'. In: *Electron Devices, IEEE Transactions on* 55 (2008), pp. 2960–2967. DOI: 10.1109/TED.2008.2005152.

[155] M. Muraguchi and T. Endoh. 'Size Dependence of Electrostatic Lens Effect in Vertical MOS-FETs'. In: *Japanese Journal of Applied Physics* 53.4S (2014), 04EJ09. DOI: 10.7567/JJAP.53.04EJ09.

[156] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Improved Drive-Current into Nanoscaled Channels using Electrostatic Lenses'. In: *Simulation of Semiconductor Processes and Devices (SISPAD), International Conference on.* 2015, pp. 24–27. DOI: 10.1109/SISPAD.2015.7292249.

[157] A. Dixit, A. Kottantharayil, N. Collaert, M. Goodwin, M. Jurczak and K. De Meyer. 'Analysis of the Parasitic S/D Resistance in Multiple-Gate FETs'. In: *Electron Devices, IEEE Transactions on* 52.6 (2005), pp. 1132–1140. DOI: 10.1109/TED.2005.848098.

[158] A. Villalon, C. Le Royer, S. Cristoloveanu, M. Casse, D. Cooper, J. Mazurier, B. Previtali et al. 'High-Performance Ultrathin Body c-SiGe Channel FD-SOI pMOSFETs Featuring SiGe Source and Drain: V_{th} tuning, variability, access resistance, and mobility issues'. In: *Electron Devices, IEEE Transactions on* 60.5 (2013), pp. 1568–1574. DOI: 10.1109/TED.2013.2255055.

[159] S. Berrada, M. Bescond, N. Cavassilas, L. Raymond and M. Lannoo. 'Carrier Injection Engineering in Nanowire Transistors via Dopant and Shape Monitoring of the Access Regions'. In: *Applied Physics Letters* 107.15 (2015), p. 153508. DOI: 10.1063/1.4933392.

[160] Y. Jiang, T. Liow, N. Singh, L. Tan, G. Lo, D. Chan and D. Kwong. 'Performance Breakthrough in 8 nm Gate Length Gate-All-Around Nanowire Transistors using Metallic Nanowire Contacts'. In: *VLSI Technology, Symposium on.* 2008, pp. 34–35. DOI: 10.1109/VLSIT.2008.4588553.

[161] H. Kim, H. Min, T. Tang and Y. Park. 'An Extended Proof of the Ramo-Shockley Theorem'. In: *Solid-State Electronics* 34.11 (1991), pp. 1251–1253. DOI: 10.1016/0038-1101(91)90065-7.

[162] S. Smirnov. 'Physical Modeling of Electron Transport in Strained Silicon and Silicon-Germanium'. PhD thesis. Technische Universität Wien, 2004. URL: http://www.iue.tuwien.ac.at/phd/smirnov/.

[163] M. V. Fischetti. 'Monte Carlo Simulation of Transport in Technologically Significant Semiconductors of the Diamond and Zinc-Blende Structures. I. Homogeneous Transport'. In: *Electron Devices, IEEE Transactions on* 38.3 (1991), pp. 634–649. DOI: 10.1109/16.75176.

[164] D. Ferry. *Semiconductors*. Macmillan Publishing Company, 1991. ISBN: 9780023371301.

[165] S. M. Goodnick, D. K. Ferry, C. W. Wilmsen, Z. Liliental, D. Fathy and O. L. Krivanek. 'Surface Roughness at the Si(100)-SiO$_2$ Interface'. In: *Physical Review B* 32 (12 1985), pp. 8171–8186. DOI: 10.1103/PhysRevB.32.8171.

[166] D. Vasileska and D. K. Ferry. 'Scaled Silicon MOSFETs: Universal Mobility Behavior'. In: *Electron Devices, IEEE Transactions on* 44.4 (1997), pp. 577–583. DOI: 10.1109/16.563361.

Own Publications

Journals

[1] P. Ellinghaus, J. Weinbub, M. Nedjalkov, S. Selberherr and I. Dimov. 'Distributed-Memory Parallelization of the Wigner Monte Carlo Method using Spatial Domain Decomposition'. In: *Journal of Computational Electronics* 14.1 (2015), pp. 151–162. DOI: 10.1007/s10825-014-0635-3.

[2] J. Weinbub, P. Ellinghaus and M. Nedjalkov. 'Domain Decomposition Strategies for the Two-Dimensional Wigner Monte Carlo Method'. In: *Journal of Computational Electronics* 14.4 (2015), pp. 922–929. DOI: 10.1007/s10825-015-0730-0.

[3] M. Nedjalkov, J. Weinbub, P. Ellinghaus and S. Selberherr. 'The Wigner Equation in the Presence of Electromagnetic Potentials'. In: *Journal of Computational Electronics* 14.4 (2015), pp. 888–893. DOI: 10.1007/s10825-015-0732-y.

Book Contributions

[4] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Optimized Particle Regeneration Scheme for the Wigner Monte Carlo Method'. In: *Lecture Notes in Computer Science*. Ed. by I. Dimov, S. Fidanova and I. Lirkov. Vol. 8962. Springer International Publishing, 2015, pp. 27–33. ISBN: 9783319155845. DOI: 10.1007/978-3-319-15585-2_3.

[5] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'The Influence of Electrostatic Lenses on Wave Packet Dynamics'. In: *Large-Scale Scientific Computing*. Ed. by I. Lirkov, S. D. Margenov and J. Waśniewski. Vol. 9374. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 277–284. ISBN: 9783319265193. DOI: 10.1007/978-3-319-26520-9_30.

[6] J. Cervenka, P. Ellinghaus and M. Nedjalkov. 'Deterministic Solution of the Discrete Wigner Equation'. In: *Numerical Methods and Applications*. Ed. by I. Dimov, S. Fidanova and I. Lirkov. Springer International Publishing, 2015, pp. 149–156. ISBN: 9783319155845. DOI: 10.1007/978-3-319-15585-2_17.

[7] J. Cervenka, P. Ellinghaus, M. Nedjalkov and E. Langer. 'Optimization of the Deterministic Solution of the Discrete Wigner Equation'. In: *Lecture Notes in Computer Science*. Ed. by I. Lirkov, S. Margenov and J. Wasniewski. Vol. 9374. Springer International Publishing, 2015, pp. 269–276. ISBN: 9783319265193. DOI: 10.1007/978-3-319-26520-9{_}29.

[8] J. Weinbub, P. Ellinghaus and S. Selberherr. 'Parallelization of the Two-Dimensional Wigner Monte Carlo Method'. In: *Large-Scale Scientific Computing*. Ed. by I. Lirkov, S. D. Margenov and J. Waśniewski. Vol. 9374. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 309–316. ISBN: 9783319265193. DOI: 10.1007/978-3-319-26520-9_34.

Conference Contributions (Full Proceedings)

[9] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Implications of the Coherence Length on the Discrete Wigner Potential'. In: *Computational Electronics (IWCE), International Workshop on*. 2014, pp. 1–3. DOI: 10.1109/IWCE.2014.6865852.

[10] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Efficient Calculation of the Two-Dimensional Wigner Potential'. In: *Computational Electronics (IWCE), International Workshop on.* 2014, pp. 1–3. DOI: `10.1109/IWCE.2014.6865812`.

[11] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'The Wigner Monte Carlo Method for Accurate Semiconductor Device Simulation'. In: *Simulation of Semiconductor Processes and Devices (SISPAD), International Conference on.* 2014, pp. 113–116. DOI: `10.1109/SISPAD.2014.6931576`.

[12] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Memory-Efficient Particle Annihilation Algorithm for Wigner Monte Carlo Simulations'. In: *Computational Electronics (IWCE), International Workshop on.* 2015, pp. 1–4. DOI: `10.1109/IWCE.2015.7301955`.

[13] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Improved Drive-Current into Nanoscaled Channels using Electrostatic Lenses'. In: *Simulation of Semiconductor Processes and Devices (SISPAD), International Conference on.* 2015, pp. 24–27. DOI: `10.1109/SISPAD.2015.7292249`.

Conference Contributions (Book of Abstracts)

[14] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Implications of the Coherence Length on the Discrete Wigner Potential'. In: *Book of Abstracts of the 17th International Workshop on Computational Electronics (IWCE).* 2014, pp. 155–156. ISBN: 9782954785806.

[15] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Efficient Calculation of the Two-Dimensional Wigner Potential'. In: *Book of Abstracts of the 17th International Workshop on Computational Electronics (IWCE).* 2014, pp. 19–20. ISBN: 9782954785806.

[16] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Optimized Particle Regeneration Scheme for the Wigner Monte Carlo Method'. In: *Eighth International Conference on Numerical Methods and Applications.* 2014, p. 19.

[17] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'The Influence of Electrostatic Lenses on Wave Packet Dynamics'. In: *Abstracts International Conference on Large-Scale Scientific Computations (LSSC).* 2015, pp. 39–40.

[18] P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Improved Particle Annihilation for Wigner Monte Carlo Simulations on a High-Resolution Mesh'. In: *Book of Abstracts 18th International Workshop on Computational Electronics (IWCE).* 2015, pp. 93–94. ISBN: 978-0-692-50554-0.

[19] P. Ellinghaus, J. Weinbub, M. Nedjalkov and S. Selberherr. 'ViennaWD - Applications'. In: *Booklet of the 1st International Wigner Workshop (IW2).* 2015, p. 9.

[20] J. Cervenka, P. Ellinghaus and M. Nedjalkov. 'Deterministic Solution of the Discrete Wigner Equation'. In: *Eighth International Conference on Numerical Methods and Applications.* 2014, p. 36.

[21] J. Cervenka and P. Ellinghaus. 'Preconditioned Deterministic Solver for the Wigner Equation'. In: *Abstracts International Conference on Large-Scale Scientific Computations (LSSC).* 2015, p. 31.

[22] M. Nedjalkov, P. Ellinghaus and S. Selberherr. 'The Aharanov-Bohm Effect from a Phase Space Perspective'. In: *Abstracts International Conference on Large-Scale Scientific Computations (LSSC).* 2015, pp. 59–60.

[23] J. Weinbub, P. Ellinghaus and S. Selberherr. 'Parallelization of the Two-Dimensional Wigner Monte Carlo Method'. In: *Abstracts International Conference on Large-Scale Scientific Computations (LSSC).* 2015, p. 73.

[24] J. Weinbub, P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'ViennaWD - Status and Outlook'. In: *Booklet of the 1st International Wigner Workshop (IW2).* 2015, p. 8.

[25] J. Weinbub, P. Ellinghaus, M. Nedjalkov and S. Selberherr. 'Comparison of Slab and Block Decomposition Strategies for the Two-Dimensional Wigner Monte Carlo Method'. In: *Abstracts International Symposium on Advanced Nanodevices and Nanotechnology (ISANN).* 2015.

Curriculum Vitae

Date of Birth: 22 May 1986
Place of Birth: Johannesburg, South Africa

2/2005 − 11/2008
Bachelor of Engineering (Electronic Engineering)
University of Pretoria, South Africa

12/2008 − 8/2009
Engineer, INSiAVA (Pty) Ltd
Pretoria, South Africa

9/2009 − 9/2011
Master of Science (Mathematical Engineering)
University of L'Aquila, Italy

10/2011 − 12/2012
Research Assistant, IHP Microelectronics
Frankfurt (Oder), Germany

1/2013
Commenced Doctoral Studies, Institute for Microelectronics
Technische Universität Wien, Austria