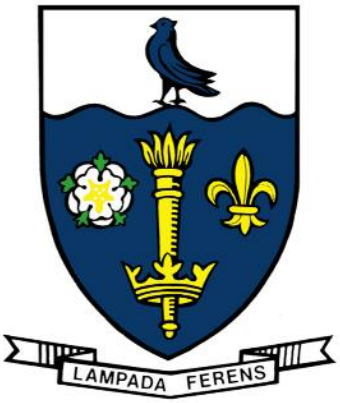
THE UNIVERSITY OF HULL



Monte Carlo Simulations of Two-Dimensional Electron Gasses in Gallium Nitride High Electron Mobility Transistors via General-Purpose Computing on Graphics Processing Units

being a Thesis submitted for the Degree of Doctor of Philosophy

in the University of Hull

By

Lee Smith MPhys (Hons)

November 2020

Summary of Thesis

The work in this thesis covers two main topics: successfully porting an Ensemble Monte Carlo (EMC) focused on bulk III-V semiconductors on to the graphics processing unit (GPU) and investigating carrier transport in a two-dimensional electron gas (2DEG) created at an Aluminium Gallium Nitride (AlGaN) and Gallium Nitride (GaN) heterojunction, specifically the effect of introducing non-equilibrium phonons.

The programming language used to be able to run on the GPU, NVIDIA CUDA, is introduced. The concept of highly parallel programming is explored, along with the challenges this poses to an EMC simulating semiconductor materials and devices. The changes made to the bulk EMC algorithm are explained, including architectural, memory strategies and execution optimisations. The performance increase related to each change is given, and it is found that the GPU algorithm has a run time that is approximately 30% of the original EMC algorithm. This is the first example of an EMC simulating electron transport in semiconductors on a GPU.

A two-dimensional EMC is created to simulate the behaviour of electrons confined in the 2DEG created at an AlGaN/GaN heterojunction. Results are presented for the electron velocity, momentum and energy relaxation times and mobility, which are compared to experimental results from AlGaN/GaN High Electron Mobility Transistors (HEMTs), and agreement is good. No velocity overshoot is observed, in agreement with experiments.

Finally, non-equilibrium phonons are introduced to the 2DEG simulation to study their effect on the electron transport. Non-equilibrium phonons are found to reduce the electron velocity due to diffusive heating. However, due to the confinement of electrons, the phonon distribution is only increased in a small volume of reciprocal space and the effects are shown to be weaker than in bulk. The consideration of electron confinement and a non-equilibrium phonon population has not been seen in the current literature.

Contents

Acknowledgements

First and foremost I would like to thank my PhD supervisors. Dr Angela Dyson for her continued support and faith throughout, aiding me with all physics related aspects over the past few years. And also for giving me this great opportunity to partake in such an interesting and exciting project on short notice. Warren Viant for his continued support, and incredible patience, whilst guiding me through my learning of new computer languages and computing techniques, in a short period of time. Also for being a much needed regular source of face-to-face contact.

For his continued help and many ideas provided during my study, massive thanks must also go to Dr Daniel Naylor. The knowledge passed on, both with respect to the original algorithm and computing in general, has been invaluable. Also for offering me a bed to sleep in and providing me company on my numerous visits to Newcastle.

Also for passing on knowledge of the algorithm, thanks must go to Dr Nick Appleyard, and special thanks for making the transition into the PhD office such a pleasant and easy one. This is also extended to the numerous students I had the pleasure of sharing an office with over the years. Dr Bethany Newton, Rajpal Theti, Steve Wilkinson, Scott Morgan and Chris Lowe. Sincere thanks must also go to Dr Addison Marshall, who regularly visited the office, especially to check up on me after the majority of the aforementioned students had left.

List of Abbreviations

Below is a list of commonly used abbreviations throughout this thesis. These abbreviations are also defined at their first instance within the body of the text.

Acronym	Meaning
2DEG	Two-dimensional electron gas
AlGaAs	Aluminium gallium arsenide
AlGaN	Aluminium gallium nitride
BTE	Boltzmann transport equation
BZ	Brillouin zone
CPU	Central processing unit
EMC	Ensemble Monte Carlo
GaAs	Gallium arsenide
GaN	Gallium nitride
GPGPU	General-purpose computing on graphics processing units
GPU	Graphics processing unit
HEMT	High electron mobility transistor
IV	Inter-valley
NPOP	Non-polar optical phonon
POP	Polar optical phonon
SM	Streaming multiprocessor

List of Figures

Below is a list of all of the figures in this thesis, along with their captions and the page number on which they appear.

List of Tables

Below is a list of tables that appear in this thesis, along with the page number.

Declaration of Authorship

I declare that the work presented in this thesis is original, my own and has not been previously submitted for examination for any other award. Where I have used the results, concepts or work of others, they have been acknowledged at the point of use.

Conference Presentations

L. Smith, D. R. Naylor, W. Viant and A. Dyson, "GPU Accelerated Monte Carlo Simulations of GaN" [Poster], UKNC, Manchester, United Kingdom, January 2018.

Lee Smith

26th November 2020

Chapter 1

Introduction

1.1 Gallium Nitride

Due to its wide, direct band gap, there has been a significant amount of research, both experimental and theoretical, on bulk GaN and device heterostructures where GaN is one of the main materials. Optical technology using GaN is fairly mature and already well understood. GaN naturally operates at a wavelength of 405 nm, and slightly longer wavelengths spanning into the green and yellow ranges of the visible light spectrum can be achieved by the formation of higher order compounds utilising GaN (such as Indium Gallium Nitride (InGaN) and Aluminium Gallium Nitride (AlGaN)). In addition to the red LED, it is possible to create a full colour range of efficient LEDs, which are now commonly used in high power applications such as traffic lights. White light LEDs have also been created using GaN with higher efficiency than conventional incandescent bulbs [1]. Blue laser diodes have also been created which are used in the now common household Blu-Ray players, and the invention of efficient blue LEDs won the Nobel Prize in Physics in 2014. There has been a lot of interest in the use of GaN for high-power, high-frequency electronics due to its high breakdown field [2, 3] and large saturation velocity [4, 5]. A specific technology that has received recent attention is the use of GaN and other nitride compounds in the High Electron Mobility Transistor (HEMT) [6].

1.2 High Electron Mobility Transistor

1.2.1 Background

GaN based HEMTs have potential for a wide range of uses, especially in high-power, high-frequency electronics, due to their large mobilities. One example is a high-gain, highly efficient power amplifier for pulsed signals, due to the high breakdown field and

high frequency response. Power outputs have been reported between 250 and 400 W at microwave frequencies [7-10]. Such devices have become commonly used in mobile network base stations, employing the high-power, high-frequency characteristics that are needed to broadcast reliable 4G network coverage, and has advanced to 5G technology. High frequency power amplifiers based on HEMTs have also been utilised in military applications, with the development of versatile, high-gain amplifiers that can be used along with a number of high power military equipment [11]. Despite great advances in GaN power devices, there are still challenges to overcome, such as suppressing the current collapse effect, which occurs when a large bias is applied to the drain, and heat dissipation [12]. The high power and high frequency demands on the devices lead to significant self-heating, resulting in rapid increases in the channel temperatures and device degradation.

Early experiments on GaN HEMTs focused on measuring the electron Hall mobility and current characteristics of the devices. Voltage-current curves, and current against applied electric field plots, are useful in determining the performance of a device and its potential for high-power and high-gain uses. Measured currents are used to estimate the electron drift velocity [13] which can then be compared to Monte Carlo simulation results. Experimental results of steady state electron drift velocity in GaN based HEMTs have shown little signs of a peak and saturation [13-17], and transient velocity results also show no signs of a velocity overshoot [18, 19]. A clear peak and saturation in the steady state, and a significant transient overshoot, are prominent features of velocity results in bulk GaN [2, 4, 20, 21]. Their absence therefore strongly indicates that the electrons are confined within a two-dimensional electron gas (2DEG), and the electron behaviour and transport properties are no longer governed by macroscopic transport physics.

In order to try and understand and replicate the results, Monte Carlo simulations of electron transport in GaN 2DEGs were created. These simulations regularly use the triangular well approximation, and when simulating narrow quantum wells, use just one

or two sub-bands. If the quantum well is narrow enough, the quantum limit can be reached where electrons only occupy a single sub-band [22]. In most devices the well is too wide for this to be the case, however, in most HEMTs the well is narrow and it is a reasonable assumption to include just two sub-bands in the simulation [23]. The infinitely deep triangular well assumption means that electrons remain confined within the quantum well even at large energies. In reality, the quantum well has a finite height above which the electrons are no longer confined in discrete energy levels and are in continuum states, behaving as they would in bulk. Electrons are initially confined in the discrete energy levels of the 2DEG, however, once they gain enough energy they can escape from the potential well and enter the continuum states where their behaviour and transport is bulk-like. This situation is difficult to simulate, so it is regularly assumed that the electrons exist purely in the 2DEG, or in bulk. Effects of the 2DEG-continuum states have potentially been seen experimentally by *Atmaca et al.* [24], who report having seen negative differential resistance in a 2DEG, which has never been reported before. Their explanation for this is the large electric fields (due to small channel lengths) cause the electrons in the 2DEG to gain enough energy to make the transition from the Γ valley (the lowest valley and where the 2DEG resides) to the L valley whose minima is 0.9 eV [25, 26] above the conduction band minima in bulk GaN. *Atmaca et al.* claim the velocity then decreases due to the 2DEG electrons in the L valley having higher effective mass, however details of the confined band structure are not present in the literature. Electrons in the L valley do have a higher effective mass [27], however, it is highly unlikely that the electrons are still confined in a 2DEG. The barrier height in the 2DEG created at an AlGaN/GaN heterojunction is typically between 0.3-0.4 eV [28], so any electrons that do transfer to the L valley at 0.9 eV will no longer be confined and will be in bulk states.

High energy electrons, or ‘hot electrons’, have been seen to cause degradation in an InAlN/GaN HEMT by *Tapajna et al.* [29]. While concentrations of hot electrons and

degradation have been investigated in AlGaN/GaN HEMTs by *Brazzini et al.* [30] with local electron temperatures reaching approximately 3900 K. Operating temperatures of AlGaN/GaN HEMTs have also been investigated. *Pomeroy et al.* [31] investigated the operating channel temperature of AlGaN/GaN HEMTs and experimentally saw local temperatures around 110°C (383 K), whilst models predicted temperatures reaching 250°C (523 K). *Tan et al.* [32] also investigated the high temperature performance of AlGaN/GaN HEMTs and found stable operation and no significant permanent degradation up to 500°C (773 K). Most early Monte Carlo simulations produce results that do not agree well with experiment. *Ardaravicius et al.* [27] compared Monte Carlo results from *Yu et al.* [18] to their own experimental results, however the Monte Carlo results significantly overestimated the electron velocity found experimentally. Similarly, *Palacios et al.* [33] also found simulation results to be much higher than experiment. An absence of hot phonon effects has been used to explain the poor match between experiment and simulation [27]. Some work has been done to include hot phonon effects in simulations and it is shown to reduce the velocity results and match more closely to experiment. Hot phonon effects were shown in bulk GaN by *Dyson et al.* [34]. *Ramonas et al.* [35] found including hot phonon effects to greatly improve the accuracy of the simulation results when compared to experiment. *Tea et al.* [36] have also shown how a phonon dedicated Monte Carlo code that treats hot phonons as a population, updates the phonon distribution function and calculates phonon relaxation over time can be coupled with a charge carrier Monte Carlo code to accurately represent the hot carrier relaxation dynamics in polar semiconductors. They found that the hot phonon population slowed relaxation when the carrier density was relatively low. Self-consistent Monte Carlo simulations have also been created that calculate and update the electric field throughout the simulation based on the distribution of the charge carriers and can also calculate heat generation [37]. Many results are limited to low applied electric fields (typically less than

50 $kVcm^{-1}$). In simulations, this is likely an attempt to keep the electron energies low, such that the two sub-band triangular well approximations remain valid. Experimentally, this can be due to limits in the measurement set up [35], but it also presumably to avoid device degradation.

1.2.2 2DEG Formation in AlGaN/GaN

Most semiconductor devices are created by doping the active semiconductor material. This generally includes HEMTs which are based on the creation of a quantum well at a heterojunction between a (typically) highly-doped wide gap material and an undoped material with a narrower band gap, and lower conduction band minima. However, this is not the case in AlGaN/GaN HEMTs, which can be created without the need for doping. In many early studies concerning the 2DEG, the existence of the 2DEG was simply assumed without any prior knowledge or proof of its existence [38-40], there was little understanding about the reasoning behind its formation. 2DEGs have been observed at the AlGaN/GaN interface when there is no doping in the AlGaN layer [41]. In AlGaN/GaN based heterostructures, the 2DEG is induced by strong polarization effects, from both spontaneous and piezoelectric polarization [38, 39, 42]. Spontaneous polarization (at zero strain) creates a significant electric field in AlGaN [43]. The piezoelectric polarization creates its own electric field which enhances this. *Ambacher et al*. [39, 44] investigated the spontaneous and piezoelectric polarization charges in AlGaN/GaN heterostructures and state that when the polarization induced sheet charge density is positive, free electrons tend to accumulate to compensate the polarization induced charge. The polarization induced sheet charge density is associated with a gradient of polarization in space. At an AlGaN/GaN interface, the polarization is different within each layer, and the sheet charge density is thus defined as the difference in polarization between the top layer and the bottom layer [44]. The total polarization in

each layer is a sum of the spontaneous polarization (no strain) and the piezoelectric polarization (strain-induced). Due to the strong spontaneous and piezoelectric polarizations in AlGaN/GaN based structures (piezoelectric polarization of the strained top layer can be more than five times larger than AlGaAs/GaAs structures [44], while the spontaneous polarization is also very large, particularly in AlN [43]), there is a significant increase in the sheet carrier concentration at the interface [45-49]. The large electric fields created by the strong polarization effects lead to a deep quantum well at the interface (large steps in the conduction band energy) and hence strong carrier confinement. For a single heterojunction, the quantum well is triangular [50, 51]. Electrons that have accumulated near the interface to compensate for the large polarization induced charge travel across the interface (effectively 'falling' over a potential cliff) and become confined in the quantum well, creating a 2DEG.

1.3 CPUs vs GPUs

As the number of transistors on computer chips increases and computers become more powerful, scientific simulations become more advanced and complex. With increasing complexity, simulations begin to require more computer resources and their run times become significantly long. Utilising multiple cores on the CPU, or multiple CPUs, where parallelisation is possible is one way to overcome these long run times. Where high amounts of parallelisation is possible, GPUs hold great potential in reducing run times further. The CPU contains several powerful cores capable of handling a wide-range of tasks. The GPU contains a significantly higher number of less powerful cores. The architectures of both the CPU and GPU are shown in figure 1.3.1, which shows the vast difference in the number of cores. The Kepler card series illustrated in figure 1.3.1 is old (released in 2012) but its architecture is indicative of more recent cards. This is the main difference between the CPU and GPU. The CPU is designed to perform a wide-range of

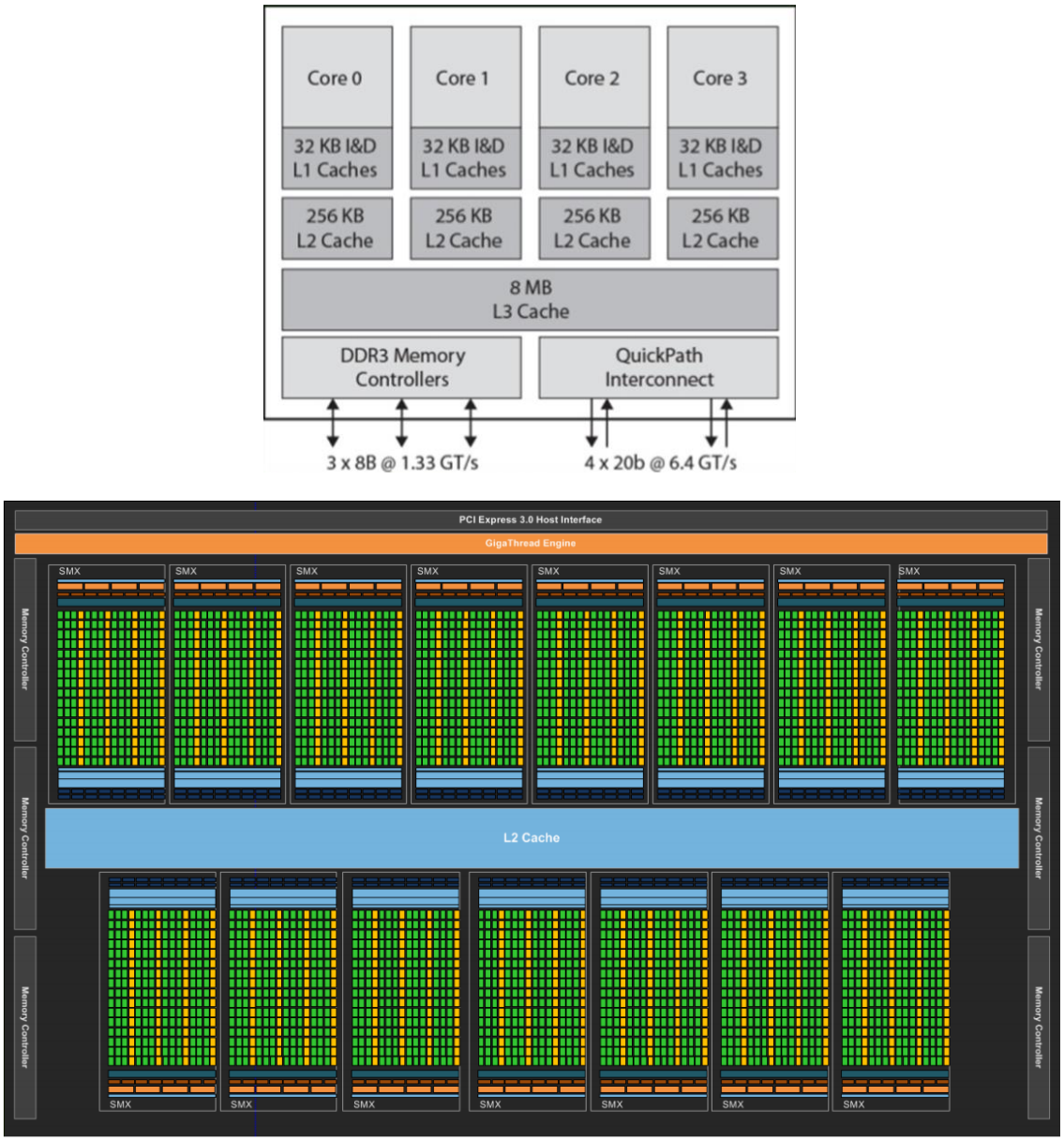


Figure 1.3.1: Illustration of the CPU and GPU architectures, showing the vast difference in cores. The CPU diagram (top) is based on the Intel Core processors, and the GPU diagram (bottom) is the block diagram of a NVIDIA Kepler GK110, where each green or yellow block represents a CUDA core [52].

different tasks as quickly as possible, but are limited by the number of tasks that can be run at the same time. The GPU is designed to render high-resolution images, at high speeds and concurrently. CPUs have broad instruction sets and manage every input and output of a computer. GPUs do not have the same wide range of instructions, however, the sheer magnitude of GPU cores and the potential for massive amounts of parallelism make them an interesting prospect for general purpose computing instructions. The use of GPUs to perform general computation tasks is known as general-purpose computing

on graphics processing units (GPGPU) and has already been utilised in many scientific disciplines [53-55].

1.4 Outline of Thesis

The next chapter of this thesis introduces the different algorithms and equations that are needed to develop the various models used throughout this work. The ensemble Monte Carlo (EMC) algorithm is described and how this premise is used to generate a model to simulate electron transport in bulk III-V semiconductor materials is then introduced. The changes required to ensure the algorithm is capable of simulating transport in a 2DEG are explained. These include changes to the scattering rates and the underlying equations. The chapter finishes by introducing how 3D non-equilibrium phonons can be included in the algorithm, and how they are tracked and updated. Chapter 3 explains how the bulk EMC algorithm is modified to be run on a GPU, the first example of an electron transport in semiconductors simulation being performed on the GPU. The challenges faced and how the algorithm was amended in an attempt to mitigate these is explained. These include architectural changes specific to the GPU, as well as some generic simulation changes that can be applied to any model. The memory strategy to best utilise the GPU memory is introduced. All changes are explained and their effect on the simulation performance are explored. These include effects on overall runtime and on the results output, when compared to the original EMC algorithm run on a CPU. This is the first example of a semiconductor Monte Carlo simulation being performed on the GPU. Chapters 4 and 5 explore the use of the 2DEG algorithm to investigate electron transport in the 2DEG created at an AlGaN/GaN interface in a HEMT. The inclusion of electron confinement in HEMT simulations has not been seen before, neither has the combination of confined electrons and a non-equilibrium phonon distribution. Chapter 4 presents results from an equilibrium 2DEG algorithm, including velocity and mobility results that

are compared to published experimental and simulation results from the literature. Chapter 5 focuses on how introducing 3D non-equilibrium phonons affects the 2DEG. The same results are investigated and compared to the equilibrium results from chapter 4 and published results. The behaviour of the phonons is explored, including the evolution of the distribution and effect on the scattering rates. The thesis ends in chapter 6 by drawing conclusions from the results presented and potential future work based on extensions of this project are discussed.

Chapter 2

Semiconductor Physics and Monte Carlo Methods for Simulating Electron Transport

Monte Carlo simulations are one of the many models that can be used in simulating electron transport in semiconductor materials and devices, the simplest being semi-classical drift-diffusion models, which make use of drift-diffusion equations derived from the Boltzmann Transport Equation. However, they are insufficient for investigating sub-micron devices and non-equilibrium behaviour. Full quantum models using numerical solutions of the Schrödinger equation can be adopted to study such devices and behaviour. The results of such models are highly accurate, however, the computational effort is very high and results can only be achieved for small numbers of particles [56]. Monte Carlo simulations are somewhere in-between drift-diffusion and full quantum models, in terms of both computational effort and accuracy. These simulations can be purely semi-classical, or if necessary, quantum corrections can be included to account for possible many body effects. However, since electron-electron interactions are ignored (including these would be a large amount of work for a relatively small effect), for the work presented in this thesis the semi-classical approach is sufficient. This chapter will introduce some key concepts in semiconductor physics and describe the base bulk model. The underlying concepts of a 2DEG are also introduced, along with how the base EMC algorithm is adapted to allow for the simulation of the electron transport within a 2DEG. Finally, a brief overview is given of how a bulk population of non-equilibrium phonons can be introduced into the algorithms. Most algorithms are based on those presented by *Tomizawa* [57], which were re-developed in FORTRAN 95 by *Naylor* [20], and the base algorithm has been reproduced in C++ to allow for the use of NVIDIA CUDA.

2.1 Background Physics

2.1.1 Band Structure

If a free electron (away from any lattice) with energy, E, is treated as a wave with frequency $\omega = E/\hbar$, then through the solution of the time independent wave equation, a simple relationship between E and the electron wavevector, **k**, is found:

$$E(\mathbf{k}) = \frac{\hbar^2 \mathbf{k}^2}{2m_0} \qquad 2.1.1$$

where m_0 is the electronic mass [58] and $\hbar$ is the reduced Planck constant. It can be shown that $\hbar\mathbf{k}$ is the electron momentum, and hence equation 2.1.1 is seen to be the equivalent of the electron kinetic energy. However, if the same electron is placed in a periodic, varying potential (in a crystalline structure, for example), then the energy is no longer given by the simple expression above. The Schrödinger wave equation must now be solved, taking into account the varying potential at a given position, **r**. Semiconductors are highly ordered crystalline structures, and a characteristic of such structures is a periodic potential related to its nuclei and electrons. The potential within a semiconductor can be approximated to a predefined potential, V(**r**), giving a single electron problem [57]. For such a potential profile, applied to an electron with wave function $\psi(\mathbf{r})$ and energy E(**r**):

$$\left[-\frac{\hbar^2}{2m_0}\nabla^2 + V(\mathbf{r})\right]\psi(\mathbf{r}) = E(\mathbf{r})\psi(\mathbf{r}) \qquad 2.1.2$$

where ∇^2 is the Laplacian operator. Bloch's theorem states that the solution to the wave equation in any periodic potential must itself be periodic [59], given as:

$$\psi_{\boldsymbol{k}} = u_{\boldsymbol{k}}(\boldsymbol{r})e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \qquad 2.1.3$$

where $u_{\boldsymbol{k}}(\boldsymbol{r})$ is the Bloch lattice function, which periodically produces the same set of results. Since the solution to the Schrödinger equation is periodic, this allows the solution to be obtained within a unit cell, knowing the solution found will apply to all other unit cells. The unit cell, in reciprocal space, is the first Brillouin Zone (BZ). Kronig and Penney [60] have shown that a simple one-dimensional potential profile can be used to obtain a solution to equation 2.1.2 that contains ranges of energy where there are no physical solutions, producing forbidden energy "gaps" within the band.

2.1.1.1 Effective Mass

In three-dimensional space, the band structure is in general much more complex than the simple Kronig-Penney model. There are numerous different reasons for this, such as the difference in lattice constants in different directions, and in binary or higher order compounds, the variance in sizes and charges of the atoms. This can often lead to bands having multiple local energy minima (referred to as "valleys") throughout the BZ, and these valleys are typically not symmetric in all directions. Full band structure models can be used to solve transport problems using a wide range of numerical methods [2, 61-63]. However, mobile electrons are, generally, low in energy with respect to the rest of the conduction band and lie around the valley minima (typically populating the lowest valley). An analytic approximation for the band structure, similar to the free electron model given by equation 2.1.1, known as the parabolic band structure approximation can be used around the valley minima [58]. For a valley minima located at energy, E_c, it is given by

$$\mathrm{E}(\mathbf{k}) = \frac{\hbar^2\mathbf{k}^2}{2m^*} + \mathrm{E_c} \qquad 2.1.4$$

and is based on the concept of the effective mass of an electron, m^*, which is defined as [58, 64, 65]

$$m^* = \frac{\hbar^2}{d^2\mathrm{E}/d\mathrm{k}^2} \tag{2.1.5}$$

The term effective mass arises because an electron within a band travels as if it has a mass equal to m^*. For spherically symmetric parabolic bands, or close to the valley minima, the effective mass is a constant.

2.1.2 Electron Scattering

An electron travelling through a perfect crystal would be continuously accelerated based on the strength of an external field. However, in reality, even the smallest defect can cause the electron to be scattered, changing the direction of travel and possibly the electron energy. As an electron travels through a semiconductor, there are several scattering mechanisms that it may encounter. Not all mechanisms exist in all systems. The remainder of this subsection focuses on some of the most common mechanisms found in semiconductor transport, scattering by charged impurities and threaded dislocations, and phonon scattering. Alloy scattering is also introduced, which does not occur in pure elemental semiconductors, however, it can play a role in electron transport in 2DEG systems.

2.1.2.1 Charged Impurity Scattering

Whether they are present based on design, or otherwise, defects arise during crystal growth. Sometimes, impurities are purposely introduced through doping, in other cases it is almost inevitable that some undesirable material will affect the semiconductor during growth. Similarly, dislocations are often unavoidable. Dislocations are the inclusion, or omission, of a line of atoms in the material, which can occur when a material is grown on

a substrate material with a different lattice spacing (a problem in GaN). Both impurities and dislocations cause local ionisation, disrupting the otherwise periodic potential. Hence, any nearby electron will scatter due to the abnormal field present. The effect is greater on slowly moving (low energy) electrons that spend longer in the region of the impurity or dislocation, while electrons travelling quickly (high energy) are less affected.

2.1.2.2 Non-Polar Optical Phonon and Acoustic Scattering

For any temperature greater than 0 K, the lattice in any solid will vibrate. To account for these vibrations, the energy is quantised and the vibrations are termed "phonons". For simplicity in dealing with phonon-related events, a phonon is defined by a wave vector, $\mathbf{q}$, and frequency, $\omega_{\mathbf{q}}$. In the same way an electron can be described by a wave vector, $\mathbf{k}$, as having momentum $\hbar\mathbf{k}$, a phonon defined by a wave vector, $\mathbf{q}$, acts as though it has momentum $\hbar\mathbf{q}$. The energy of the phonon is described by its frequency, as $\hbar\omega_{\mathbf{q}}$. When a phonon interacts with an electron, the principles of conservation of momentum and energy are applied to determine the outcome of the interaction. There are two types of phonon, when neighbouring atoms are displaced in the same direction, this is termed an acoustic phonon (acoustic phonon scattering), and when atoms are displaced in opposite directions, this is an optical phonon (also known as non-polar optical phonon scattering). Each of these displacements has a different effect on the periodic potential, and hence will have a different effect on the motion of the electron.

2.1.2.3 Polar Optical Phonon and Piezoelectric Scattering

Compound materials, such as III-V semiconductors (like GaN) have a polar nature. The bonds between neighbouring atoms are slightly ionic. Any lattice vibration that causes the atoms to displace will corrupt local charge neutrality, thus causing electric polarisation. This small polarisation generates an electric field, which affects the periodic potential and hence can affect the motion of any nearby electrons. The displacement of

atoms can be in the same direction, generating an acoustic phonon (referred to as piezoelectric scattering), or in the opposite direction, generating an optical phonon (known as polar optical phonon scattering). Polar optical phonon scattering is the dominant mechanism in polar materials, whilst piezoelectric scattering is often weak and can be ignored [66].

2.1.2.4 Alloy Scattering

In semiconductor alloys ($A_xB_{1-x}C$), it is often assumed that the two types of atoms, A and B, form a uniform periodic arrangement and hence a periodic potential [67]. However, it is often the case that the actual arrangement differs slightly from this perfect periodic arrangement. Say, for example, a site that in the perfect alloy should be occupied by atom A, is in fact occupied by atom B, or vice versa. This variation from the perfect alloy lattice causes a local change in the period potential, and hence can have an effect on the motion of a nearby electron.

2.1.3 Fermi's Golden Rule

The quantum mechanical representation of the electron scattering process is Fermi's Golden Rule. It is based on the assumption that any interaction between an electron and phonon is instantaneous, and has a long lasting effect. However, since an electron is going to encounter many scattering events within any system, it is possible that the effect of any given collision is not permanent and could be relatively short-lived. Collisions are not instantaneous either, however, if the duration of each collision is much shorter than the time between scattering events, then the assumption is valid (a condition which is found to be satisfied in most cases [65, 66]). The scattering rate is then given by [66]:

$$W(\mathbf{k}) = \frac{2\pi}{\hbar} \int |M(\mathbf{k}', \mathbf{k})|^2 \delta(\mathrm{E_f} - \mathrm{E_i}) dN_f \qquad 2.1.6$$

where $M(\mathbf{k}', \mathbf{k})$ is the scattering rate matrix element (connecting the initial and final electron states), E_i and E_f are the initial and final electron energies and N_f is the number of final states. For interactions in which only one phonon process is involved, assuming the final state is unoccupied, the scattering rate can be rewritten as [65]:

$$W(\mathbf{k}) = \frac{\Omega}{(2\pi)^3}\frac{2\pi}{\hbar}\int |M(\mathbf{k}', \mathbf{k})|^2 \delta(E_{\mathbf{k}'} - E_{\mathbf{k}} \pm \hbar\omega)\delta_{\mathbf{k}'-\mathbf{k}\pm\mathbf{q}} d\mathbf{k}' \qquad 2.1.7$$

where Ω is the crystal volume, $E_{\boldsymbol{k}}$ and $E_{\boldsymbol{k}'}$ are the energies of the initial and final states with wave vectors $\mathbf{k}$ and $\mathbf{k}'$, $\hbar\omega$ is the energy of the phonon with wave vector $\mathbf{q}$. The ± arises from the two possible phonon interactions (absorption and emission), in partnership with energy and momentum conservation to ensure the delta functions are equal to zero.

2.1.4 Monte Carlo Method for Solving the Boltzmann Transport Equation

The Boltzmann transport equation (BTE) describes carrier transport evolution over time in semiconductors, in both real and momentum space (represented by the subscripts r and p respectively). For a system of electrons that is represented by the distribution function $f(\mathbf{r}, \mathbf{p}, t)$, for an electron with momentum, $\mathbf{p}$, position, $\mathbf{r}$, and at time, t, and subjected to an external force, $\mathbf{F}$, the BTE is given by [58, 65]:

$$\frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_{collision} - \left(\mathbf{v} \cdot \nabla_r f + \mathbf{F} \cdot \nabla_p f\right) \qquad 2.1.8$$

when considering collisions (electron scattering) and drift (caused by the applied field) only. Since the BTE is a partial differential equation over six dimensions (three in real space and three in momentum space), finding an analytic solution is impossible. However, the equation can also be solved using numerical methods, such as Monte Carlo methods to simulate electron transport within a system. Using path integral formulation, it can be shown that the solution can be separated into two components, carrier free-flights

and instantaneous scattering events. The probability of an electron scattering within a small time interval, dt, is

$$P = W(\mathbf{k})\mathrm{dt} \tag{2.1.9}$$

where $W(\mathbf{k})$ is the total scattering rate, and $\mathbf{k}$ varies with time due to acceleration from the external field. The probability that an electron, having scattered at time t = 0, drifts until time t without scattering is then given by

$$P(t) = \exp\left[-\int_0^t W(\mathbf{k})\mathrm{dt}\right] \tag{2.1.10}$$

This probability can then be utilised as a distribution of free flight times, by the use of a uniform random number generated between 0 and 1 to represent the probability. Since the total scattering rate changes with the wavevector, $\mathbf{k}$, the integral would need to be performed for each new wavevector an electron obtains during a simulation. This can be simplified by introducing a new scattering mechanism, labelled 'self-scattering' [57]. The self-scattering rate is set such that the total scattering rate is constant across all energies (wavevectors). This means that the calculation of the electron drift time is the same for all energies (this is explained further in section 2.2.2). Therefore, if the individual scattering rates for each of the mechanisms included in the simulation are known, a total scattering rate can be generated, and hence free flight times between scattering events. Electron transport can then be simulated by drifting the electron between scattering events, according to some external applied field, and then performing an instantaneous scatter based on one of the scattering mechanisms included. Section 2.2 explains this method in more detail, including how an electron drifts and how a scattering mechanism is chosen and the scattering event is implemented.

2.2 Bulk Ensemble Monte Carlo Simulation

An ensemble Monte Carlo (EMC) algorithm simulates a system with a large number of particles simultaneously. The many particles are simulated individually for specified time steps, allowing the evolution of the system over time to be investigated. One of the main advantages of the EMC method is that it allows for both steady state and transient data to be obtained. Steady state at the end of the simulation time (for each field step) and transient at the end of each time step.

2.2.1 Ensemble Algorithm

The EMC simulation treats each electron as an independent carrier, meaning electron-electron interactions are ignored. The scattering rates are pre-calculated just once at the start of the simulation and stored in a look-up table, saving computation time throughout the simulation. Once the scattering rates are calculated, the system and electrons initial conditions are set (by starting with a thermalized electron distribution). After this, each electron is taken in turn and simulated until the end of the first time step. Once each electron has been simulated for the same period of time, the transient data can be output. This process is repeated until all electrons have been simulated for the specified simulation time. The whole process (apart from calculating scattering rates) can also be repeated for several field strengths. A flow chart of the bulk EMC algorithm is shown in figure 2.2.1. If the simulation is being run with multiple field strengths, the loop contained within the field steps (from setting initial conditions to outputting steady state data) can be run in parallel on multiple threads. In this case, calculating the scattering tables would be performed once for all fields, before the rest of the simulation is run in parallel until all field steps are complete.

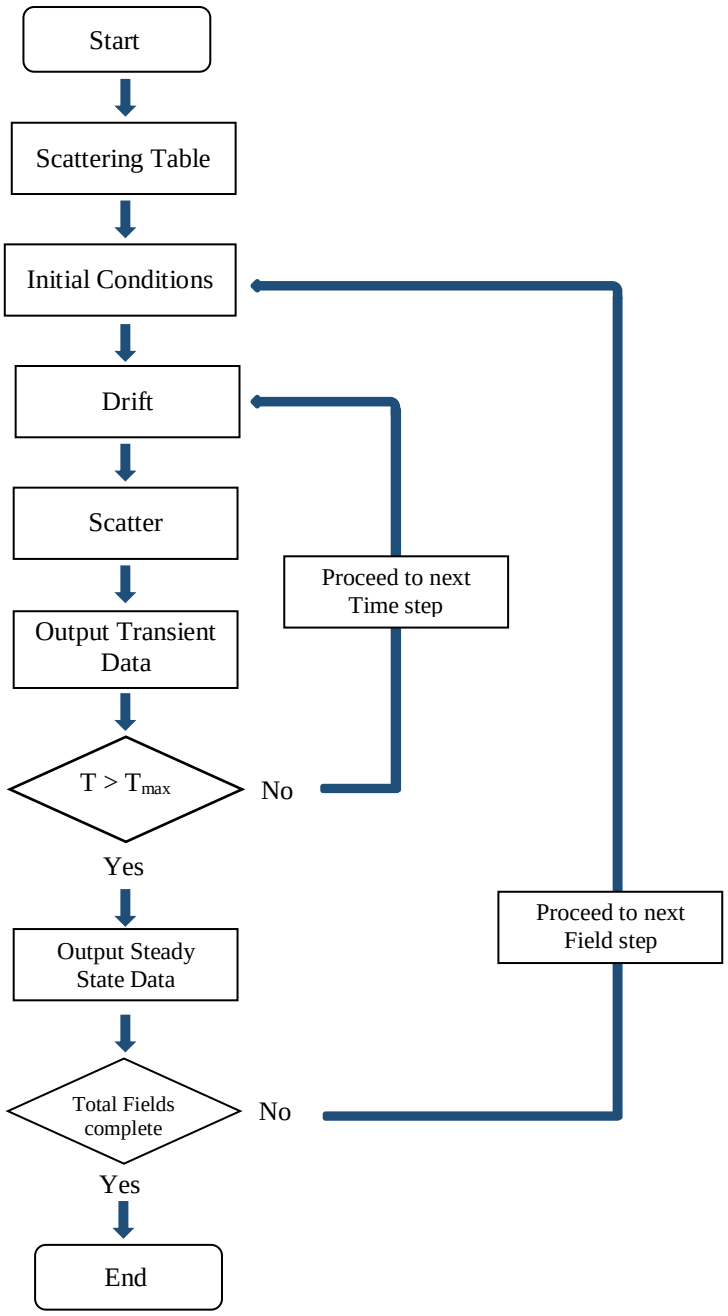


Figure 2.2.1: Flowchart showing an overview of how the base EMC algorithm simulates electron transport.

2.2.2 Scattering Rates and Drift Time

As mentioned, to increase performance scattering rates are pre-calculated at specified energy 'points' between 0 eV and some maximum, and stored in a look-up table. The maximum, and the interval step size, are material dependent (for the work using the bulk algorithm in this thesis, the material used is Gallium Nitride (GaN), and the maximum is set to 3 eV). This removes the need to calculate the scattering rate every time an electron scattering event occurs, vastly improving the run time. There is an effect on the accuracy,

however, this can be diminished by choosing an appropriate value for the interval step size such that the difference between the scattering rates at two successive energy 'points' is small. Once the scattering rates are calculated, it is possible to determine the drift time of an electron (the time between scattering events). Let $W_T(\mathrm{E}_\mathbf{k})$ be the total scattering rate for a given energy $\mathrm{E}_\mathbf{k}$, i.e. the sum of all of the individual scattering rates at a specific energy interval. *Tomizawa* states that the probability per unit time, $P(\tau)$, of an electron traveling for a period of time, τ, before being scattered is given by [57]:

$$P(\tau) = W_T(\mathrm{E}_\mathbf{k}) \exp[-\int_0^\tau W_T(E_\boldsymbol{k})dt] \qquad 2.2.1$$

As discussed in section 2.1, the electrons in the system have variable energies, and hence a different probability of scattering, leading to different drift times. These times would be required to be calculated for each energy interval. To avoid this, and to save on computer workload, 'self-scattering' is introduced. Self-scattering has no effect on the electron, it is only introduced to simplify equation 2.2.1. The self-scattering rate is set in each energy interval such that the total scattering rate, $W_T(\mathrm{E}_\mathbf{k})$ is equal across all energies. This allows equation 2.2.1 to be written as:

$$P(\tau) = \Gamma e^{-\Gamma\tau} \qquad 2.2.2$$

Where Γ is chosen as the largest total scattering rate across the energy range. Introducing self-scattering means that the calculation of the drift time is the same for all electron energies. To make use of the scattering rates when selecting a scattering mechanism, they are converted to scattering probabilities. For each energy interval, the scattering rates are stored cumulatively and normalised to Γ. This creates a series of weighted number lines, from 0-1, representing the probability of each scattering mechanism being chosen. Figure 2.2.2 represents how these scattering probabilities are stored in the code.

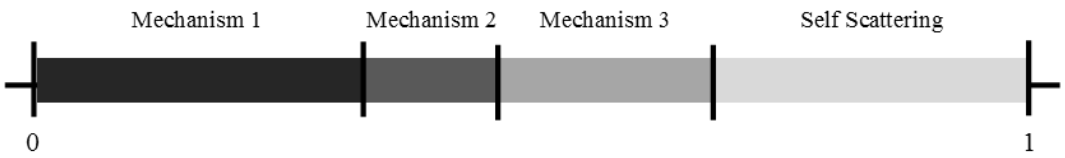


Figure 2.2.2: Illustration of the storage of scattering probabilities as a number line from 0 to 1. In this example, there are three scattering mechanisms, plus self-scattering. For mechanism 2, the probability is added to that of mechanism 1 and so on.

2.2.3 Initial Electron States

The algorithm attempts to start with a realistic electron distribution. This is achieved by using a thermal distribution based on the Fermi-Dirac function (treating the electrons as an electron gas with equipartition of energy, and allowing for three degrees of freedom by introducing a factor of 3/2 [57, 65]):

$$f(\mathrm{E}_{\mathbf{k}}) = \frac{1}{1 + \exp[(\mathrm{E}_{\mathbf{k}} - \mathrm{E}_{\mathrm{F}})/(3k_B T/2)]} \qquad 2.2.3$$

Where k_B is the Boltzmann constant and T is the lattice temperature. The Fermi energy, E_{F}, is assumed to correspond with the minima of the lowest conduction band, and is set to 0. Also assuming that $\exp[\mathrm{E}_{\mathbf{k}}/(3k_B T/2)] >> 1$ then 2.2.3 becomes,

$$f(\mathrm{E}_{\mathbf{k}}) = \frac{1}{\exp[\mathrm{E}_{\mathbf{k}}/(3k_B T/2)]} \qquad 2.2.4$$

which is rearranged to obtain,

$$\mathrm{E}_{\mathbf{k}} = -\frac{3k_B T}{2} \ln[f(\mathrm{E}_{\mathbf{k}})]. \qquad 2.2.5$$

Since $f(\mathrm{E}_{\mathbf{k}})$ is known to be a value between 0 and 1, a uniform random number generator is used to determine the energy of each electron. This energy is then used to determine the electron wavevectors. The magnitude is calculated based on the selected band-structure approximation and the direction is determined by two more randomly generated

numbers. One is used to calculate an angle between the wavevector and the x-y plane, the other used to calculate an angle between the wavevector and the z-axis. The wavevector, **k**, is then split up into its three components, k_x, k_y and k_z.

2.2.4 Electron Drift

An initial drift time is also determined, based on the solution of equation 2.2.2 in terms of τ,

$$\tau = -\frac{\ln(r)}{\Gamma} \qquad 2.2.6$$

where r is a random number between 0 and 1, chosen by a uniform random number generator. After every electron scattering event, a new drift time is determined by equation 2.2.6 and is added to the current time for that electron, to generate the time of the next scattering event. The field applied across the device is assumed to be solely in the x-direction and constant throughout. Using τ, the change in the x-component of the wavevector is then calculated using,

$$\Delta k_x = -\frac{eF}{\hbar}\tau \qquad 2.2.7$$

where e is the electronic charge and F is the applied electric field. It is also assumed that the rate of change in k_x is constant throughout the drift step. This allows for the assumption that the electron moves with a constant velocity when calculating the distance travelled by the electron in the x-direction. The constant velocity corresponds to $k_{x,(initial)} + \frac{\Delta k_x}{2}$, where $k_{x,(initial)}$ is the x-component of the wavevector at the start of the drift step.

2.2.5 Electron Scattering

When an electron has been simulated up to the time of a scattering event, the scattering mechanism to be encountered needs to be chosen. To do this, a random number between 0 and 1 is chosen using the uniform generator. Since the algorithm has generated, and stored, cumulative scattering probabilities for all energy intervals, this is a straightforward process. First, the electron energy is calculated, and then which energy interval the electron is in is calculated by dividing by the energy step size. Once the energy interval is selected, the random number, r, is compared to the cumulative probabilities, $p_n(\mathrm{E}_{\mathbf{k}})$ in the given scattering table. The condition for selecting a scattering mechanism with index m is then,

$$p_{m-1}(\mathrm{E}_{\mathbf{k}}) < r \leq p_m(\mathrm{E}_{\mathbf{k}}) \tag{2.2.8}$$

where $p_0(\mathrm{E}_{\mathbf{k}}) = 0$. If the random number is greater than the sum of all the scattering probabilities in the given energy interval, then "self-scattering" is chosen. Once the scattering mechanism has been determined, the electron energy and wavevector are updated accordingly. In all cases, the energy change is simple (either no change, or $\pm$ a phonon energy), which allows the magnitude of the post-scatter wavevector, $\mathbf{k}'$, to be easily calculated. The direction of the wavevector is determined by two scattering angles, the polar angle, θ', and the azimuthal angle, ϕ'. The polar angle is defined as the angle between the initial and the final in-plane wavevectors, whereas the azimuthal angle is defined as the angle between the final in-plane wavevector and the x-axis. The azimuthal angle is determined by the following relation,

$$\phi' = 2\pi r \tag{2.2.9}$$

where r is again a random number uniformly generated between 0 and 1. This relationship arises from the fact that the transition rates are independent of the azimuthal angle. The polar scattering angle, however, is determined based on whether the scattering mechanism is isotropic or anisotropic.

2.2.5.1 Isotropic Scattering

For scattering mechanisms that are isotropic, the post-scatter wavevector has an equal probability of pointing in any direction. In the case of isotropic scattering, the polar angle, θ', is calculated from:

$$\cos(\theta') = 1 - 2r \tag{2.2.10}$$

Since there is an equal probability for all scattering angles to occur, the angles obtained for θ' and ϕ' can be assumed to be the polar and azimuthal angles of the new direction of $\mathbf{k}'$, as opposed to the change in these angles. Meaning the new k_x, k_y and k_z of the post-scatter wavevector can be simply calculated directly from $\mathbf{k}'$, θ' and ϕ'. Isotropic scattering mechanisms include non-polar optical phonon and acoustic phonon scattering.

2.2.5.2 Anisotropic Scattering

For scattering mechanisms that are anisotropic, it is much more difficult to determine the polar angle, a direct relationship with a random number can no longer be used. The probability of scattering from 0 to the polar angle θ' can be found by solving, in terms of $\cos(\theta')$,

$$W_t(\mathrm{E}_{\mathbf{k}})_{\theta:0-\theta'}/W_t(\mathrm{E}_{\mathbf{k}})_{\theta:0-\pi} \tag{2.2.11}$$

where the subscript represents the limits of the polar angle integration, and $W_t(\mathrm{E}_{\mathbf{k}})$ is the scattering rate of scattering mechanism, t, for an electron with energy, $\mathrm{E}_{\mathbf{k}}$. *Tomizawa* presents the solutions for polar optical phonon scattering as [57]:

$$\cos(\theta') = \frac{1 + f - (1 + 2f)^r}{f} \qquad 2.2.12$$

where r is a random number from 0 to 1, and f is defined in terms of the electron energy before ($\mathrm{E}_{\mathbf{k}}$) and after ($\mathrm{E}_{\mathbf{k}'}$) scattering as

$$f = \frac{2\sqrt{\mathrm{E}_{\mathbf{k}}\mathrm{E}_{\mathbf{k}'}}}{(\sqrt{\mathrm{E}_{\mathbf{k}}} - \sqrt{\mathrm{E}_{\mathbf{k}'}})^2} \qquad 2.2.13$$

and for impurity scattering as:

$$\cos(\theta') = 1 - \frac{2r}{1 + (1 - r)\left(\frac{2k}{q_D}\right)^2} \qquad 2.2.14$$

where q_D is the Debye length, which is given by

$$q_D = \sqrt{\epsilon_s k_B T / e^2 n}\,, \qquad 2.2.15$$

where ϵ_s is the dielectric constant and n is the electron concentration. Once the scattering angles have been calculated, it is difficult to perform any further calculations directly in the original (laboratory) frame of reference. It is easier to use a rotated frame, which is rotated around the origin of the laboratory frame such that the initial wavevector is parallel to the new z-axis (see Naylor [20] for more details of the rotated reference frame in the bulk case).

2.2.6 Output

One of the main advantage of an EMC algorithm is the ability to output transient transport properties and generate the electron distribution function. This is done at the end of specified time intervals to monitor the evolution of the system, the EMC algorithm produces ensemble averaged data at any desired point in time during the simulation. To

calculate the average electron velocity and energy in the ensemble, the instantaneous properties can be taken and averaged over all electrons in the system. For a simulation containing n electrons, the average velocity at a given time, t, is calculated as

$$\langle v \rangle_t = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\hbar} \frac{\partial \mathrm{E}(\mathbf{k})_i}{\partial k_i} \qquad 2.2.16$$

and the energy is calculated as

$$\langle \mathrm{E}(\mathbf{k}) \rangle_t = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(\mathbf{k})_i \qquad 2.2.17$$

As well as ensemble average velocity and energy, the EMC algorithm can also produce valley occupancy data and the distribution of the electrons velocities and energies which can be taken at the end of any time step. This would need to be done at multiple time steps throughout the simulation to generate time evolution properties. The output at the end of the final time step gives the steady state data, if the simulation is run for a long enough period of time.

2.3 Two-Dimensional Electron Gas Monte Carlo

The EMC method can be used to simulate electron transport in the two-dimensional electron gas (2DEG), created at a material interface. The algorithm is largely similar to the bulk EMC algorithm, the main differences being the need to replace the scattering rates with the 2D rates, and to alter the code to account for electrons being confined in one direction (assumed to be the z-direction in this work), meaning they are only free to move in the plane of the other directions (x-y plane). There are two electric field strengths used, a confining field (in the z-direction) which creates the well and the applied electric field (assumed to be solely in the x-direction, as explained in the bulk algorithm). The only movement in the confined direction comes from scattering events that cause a

transition from one sub-band to another. The simplest approach to calculate the sub-band energy levels is to assume an infinitely high triangular quantum well. In what follows, the triangular quantum well approximation, assumed to contain two sub-bands, and how this is used to simulate electron transport in the 2DEG created at a GaN-AlGaN interface are explained.

2.3.1 Triangular Well Approximation

To accurately determine the wavefunctions in a quantum well would require self-consistent numerical simulations [68], due to the electric field varying with distance from the interface. Given the complexity of the Monte Carlo algorithm, a much simpler approach is to assume that the electric field changes linearly with distance, creating a triangular potential well, as shown in figure 2.3.1. For the lowest sub-band, this is a good approximation, however it becomes less accurate for much higher sub-bands. Assuming the potential changes linearly in the confinement direction (assumed here to be the z-direction), and is given by eFz, where e is the electronic charge and F is the confining electric field, then the Schrödinger equation to be solved is [57]:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2\psi_n(z)}{\partial z^2} + eFz\,\psi_n(z) = \mathrm{E}_n\psi_n(z) \qquad 2.3.1$$

where E_n and $\psi_n(z)$ are the energy level and wavefunction of the n^{th} sub-band. This approach, considering two sub-bands, is used in the 2DEG algorithm. The wavefunctions commonly used are the Fang-Howard wavefunctions, which for the first two sub-bands are given by [69, 70]:

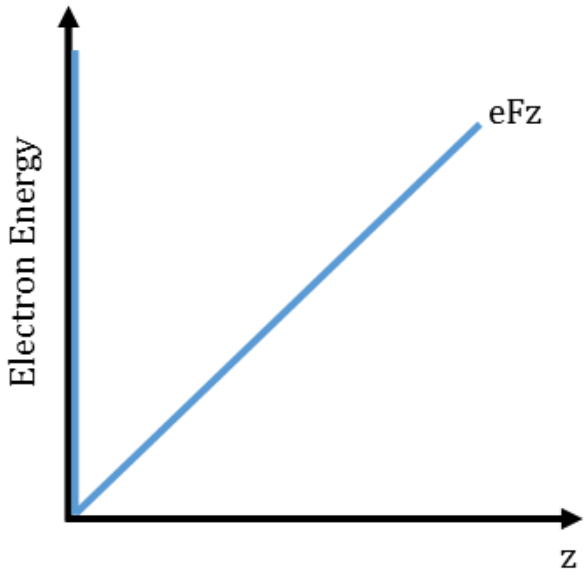


Figure 2.3.1: Illustration of the triangular potential well approximation of the conduction band, assuming the confinement is in the z-direction, for a confining electric field strength, F.

$$\Psi_0(z) = \left(\frac{b^3}{2}\right)^{\frac{1}{2}} z \exp\left[-\frac{b\,z}{2}\right] \qquad 2.3.2a$$

$$\Psi_1(z) = \left(\frac{3b^3}{2}\right)^{\frac{1}{2}} \left(z - \frac{bz^2}{3}\right) \exp\left[-\frac{b\,z}{2}\right] \qquad 2.3.2b$$

Where z is the confinement direction, and b is a constant (calculated from using the variational principle to minimise the lowest sub-band energy), given in terms of the electric field applied in the confinement direction, F_z, as [71]:

$$b = \left(\frac{14\,m^*}{\hbar^2} e\,F_z\right)^{\frac{1}{3}} \qquad 2.3.3$$

where m^* is the effective mass of the electron, e is electronic charge and $\hbar$ is the reduced Planck constant. F_z is linked to the sheet charge density (see appendix A) and its value is altered to tune the sheet density to match experiment. The minimisation parameter, b, is very important as it is used in the calculation of the 2D scattering rates. The sub-band energy levels, E_z, are also determined by minimisation and are given as [71]:

$$\mathrm{E}_{z_n} = \left(\frac{\hbar^2}{2m^*}\right)^{\frac{1}{3}} \left(\frac{3\pi A_n}{2} eF_z\right)^{\frac{2}{3}} \quad 2.3.4$$

where A_0 = 0.7587 and A_1 = 1.7540.

2.3.2 Energy and Momentum Conservation

The confinement creates discrete, quantised energy levels in the confinement direction that the electrons can occupy. Using the parabolic band approximation, the total electron energy can then be separated into two components, the sub-band energy (E_n), and the electrons kinetic energy in the plane parallel to confinement ($\mathrm{E}_{//}$). Hence

$$\mathrm{E} = \mathrm{E}_n + \mathrm{E}_{//} \quad 2.3.5$$

where

$$\mathrm{E}_n = \frac{\hbar^2 {k_z}^2}{2m^*} \quad 2.3.6$$

where $k_z$is the z-component of the wavevector, and

$$\mathrm{E}_{//} = \frac{\hbar^2 (k_x + k_y)^2}{2m^*} \quad 2.3.7$$

where k_x and k_y are the x and y components of the wavevector, and m^* is the effective mass within the well. For any intra-band electron-phonon interactions, such that the electron remains in the same sub-band, E_n and k_z remain unaffected. Hence, energy and momentum conservation is only needed to be considered within the parallel plane. For inter-band interactions, E_n and k_z will change, and hence the energy and momentum conservation rules are applied in three-dimensions.

2.3.3 2D Scattering Rates

The main differences between the derivation of the bulk and 2D scattering rates are the density of states and the introduction of a form factor in the 2D case. In the 2DEG simulation, the scattering rates used are those derived by *Yoon* [72] and are introduced below for all scattering mechanisms.

2.3.3.1 Acoustic Scattering

Using the wavefunctions given by equations 2.3.2, *Yoon* gives the acoustic deformation potential scattering rates for intra- and inter-sub-band scattering as (where the subscripts, in the form AB, refer to a scattering from sub-band A to sub-band B) [72]:

$$W_{00}(\mathbf{k}_{//}) = \frac{m^* k_B T {D_A}^2}{\hbar^3 \rho {S_l}^2} \frac{3b}{16} \qquad 2.3.8a$$

$$W_{01}(\mathbf{k}_{//}) = \frac{m^* k_B T {D_A}^2}{\hbar^3 \rho {S_l}^2} \frac{3b}{32} \qquad 2.3.8b$$

$$W_{11}(\mathbf{k}_{//}) = \frac{m^* k_B T {D_A}^2}{\hbar^3 \rho {S_l}^2} \frac{b}{8} \qquad 2.3.8c$$

where k_B is the Boltzmann constant, T is the lattice temperature, D_A is the acoustic deformation potential, ρ is the material density, S_l is the longitudinal sound velocity and b is the normalisation parameter given by 2.3.3.

2.3.3.2 Alloy Scattering

Alloy disorder also generates a short range potential which can affect the electrons in a semiconductor device, similar to the acoustic phonon. The scattering rates for intra- and inter-sub-band scattering are given by *Yoon* as (where the subscripts, in the form AB, refer to a scattering from sub-band A to sub-band B) [72]:

$$W_{00}(\mathbf{k}_{//}) = \frac{m^* x(1-x)\Omega\Delta V^2}{\hbar^3}\frac{3b}{16} \qquad 2.3.9a$$

$$W_{01}(\mathbf{k}_{//}) = \frac{m^* x(1-x)\Omega\Delta V^2}{\hbar^3}\frac{3b}{32} \qquad 2.3.9b$$

$$W_{11}(\mathbf{k}_{//}) = \frac{m^* x(1-x)\Omega\Delta V^2}{\hbar^3}\frac{b}{8} \qquad 2.3.9c$$

where x is the mole-fraction composition, Ω is the volume of the primitive cell and ΔV is the alloy disorder potential.

2.3.3.3 Polar Optical Phonon Scattering

Again, using the wavefunctions given in 2.3.2, *Yoon* derives the polar optical phonon scattering rates, for intra- and inter-sub-band scattering, as (where the subscripts, in the form AB, refer to a scattering from sub-band A to sub-band B) [72]:

$$W_{00}(\mathbf{k}_{//}) = C\int \frac{b\left[8b^2 + 9b\frac{\sqrt{2m^*}}{\hbar}\alpha + 3\left(\frac{2m^*}{\hbar^2}\right)\alpha^2\right]}{8\alpha\beta^3}\mathrm{d}\theta \qquad 2.3.10a$$

$$W_{01}(\mathbf{k}_{//}) = C\int \frac{A^2}{1944B^5}\frac{1}{\alpha}\frac{1}{\beta}\left(1 + \frac{3B}{\beta} - \frac{54B^3}{\beta^3}\right)\mathrm{d}\theta \qquad 2.3.10b$$

$$W_{11}(\mathbf{k}_{//}) = C\int \left[\alpha\left(1 + \frac{4}{b}\frac{\sqrt{2m^*}}{\hbar}\alpha\right)\right]^{-1}\mathrm{d}\theta \qquad 2.3.10c$$

where the integration range for θ is from 0 to π, and where

$$\mathrm{E}_a = 2\mathrm{E}(\mathbf{k}_{//}) \pm \hbar\omega$$

$$\mathrm{E}_b = 2\sqrt{\mathrm{E}(\mathbf{k}_{//})\left[\mathrm{E}(\mathbf{k}_{//}) \pm \hbar\omega\right]}$$

$$C = \frac{e^2\omega\sqrt{m^*}}{8\pi\sqrt{2}\hbar}\left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0}\right)\left(N(\boldsymbol{q}) + \frac{1}{2} \pm \frac{1}{2}\right)$$

$$\alpha = \sqrt{\mathrm{E}_a - \mathrm{E}_b\cos(\theta)}$$

$$\beta = b + \frac{\sqrt{2m^*}}{\hbar}\alpha$$

$$A = \frac{\sqrt{3}}{2}b^3$$

$$B = \frac{b}{3}$$

$\mathrm{E}(\mathbf{k}_{//})$ is the energy associated with the initial 2D wavevector, $\mathbf{k}_{//}$, $\hbar\omega$ is the phonon energy, ϵ_∞ and ϵ_0 are the optical and static dielectric constants, $N(\boldsymbol{q})$ is the phonon occupation number and θ is the angle between the initial and final wavevectors, $\mathbf{k}_{//}$ and $\mathbf{k}'_{//}$. To generate the polar optical phonon scattering rates requires numerical integration of the expressions in equation 2.3.10.

2.3.4 2D Electron Scattering

As in the bulk EMC algorithm, scattering rates are used to generate cumulative scattering probabilities for all energy values in the look-up table. Choosing a scattering mechanism is performed in the same way as the bulk algorithm. Similarly, the magnitude of the final wavevector is easily calculated from the resulting change in energy. However, since the electrons are only free to move in the x-y plane, the energies (and hence wavevectors) used are the in-plane energies. As the wavevector is in-plane, the direction is determined by the polar angle only, the azimuthal angle is not required.

2.3.4.1 Isotropic Scattering in 2D

As discussed in section 2.2.5, after an isotropic scattering event the post-scatter wavevector has an equal probability of pointing in all directions. This means equation 2.2.10 can be used to generate the scattering angle. Again, because of the equal probability of the final in-plane wavevector, $\boldsymbol{k}'_{//}$, pointing in all directions, the scattering angle generated is assumed to be the direction of the new wavevector, not the change in the direction. The polar angle, θ, is defined as the angle between the in-plane wavevector and the x-axis. This allows the x and y components, k_x and k_y, of the new wavevector to be simply calculated as:

$$k_x = \boldsymbol{k}'_{//} \cos(\theta) \qquad 2.3.11a$$

$$k_y = \boldsymbol{k}'_{//} \sin(\theta) \qquad 2.3.11b$$

2.3.4.2 Anisotropic Scattering in 2D

Much like in the bulk case, it is more complex to generate the scattering angle for anisotropic scattering mechanisms. There is no analytical solution to equation 2.2.11 to allow the scattering angle to be determined, instead a different approach must be taken. *Price* [73] introduces a rejection technique for determining successful polar optical phonon scattering events in 2D.

2.3.4.2.1 Polar Optical Scattering in 2D

To implement a polar optical phonon (POP) scattering event, a scattering angle is found by performing the following procedure [73], (1) generate an angle $\theta = \pi r$ where r is a random number between 0 and 1, (2) calculate the magnitude of the in-plane phonon wavevector, $q_{//}$, from:

$$q_{//}{}^2 = k_{//}{}^2 + k'_{//}{}^2 - 2k_{//}k'_{//}\cos(\theta) \qquad 2.3.12$$

where $k_{//}$ and $k'_{//}$ are the magnitudes of the initial and final in-plane wavevectors, (3) test the following condition:

$$J_*(q_{//}) > J_*(|k_{//} - k'_{//}|)\, R \qquad 2.3.13$$

where R is another random number and $J_*(q)$ is given by:

$$J_*(q) = \int_{-\infty}^{\infty} \frac{|I_*(Q)|^2}{q^2 + Q^2}\,\mathrm{d}Q \qquad 2.3.14$$

and

$$I_*(q) = \int \Psi_i \Psi_f \exp[iqx]\,\mathrm{d}x \qquad 2.3.15$$

where the subscripts i and f represent the initial and final sub-band respectively, (4) if 2.3.13 is not satisfied, repeat steps (1)-(3) until an angle is generated that does satisfy the condition. Using the wavefunctions given by 2.3.2, numerical integration is used to create another look-up table for $J_*(q)$ during the initial set up of the algorithm for use in POP scattering events. The step size for q in the look-up table was tested to optimize performance as well as obtaining a distribution of scattering angles that satisfied anisotropic scattering, i.e. a distribution that favours small angles. The angle found from the *Price* method, θ_p, is defined as the angle between the initial and final in-plane wavevectors. In order to use the angle to determine the x and y components of the final wavevector, the *Price* angle is added to the angle between the initial wavevector and the x-axis, θ, resulting in the angle between the final wavevector and the x-axis, θ', as shown in figure 2.3.2. The x and y components are then easily calculated as in equation 2.3.11, using the final angle, θ', in place of θ. The values for $\cos(\theta')$ and $\sin(\theta')$ are calculated

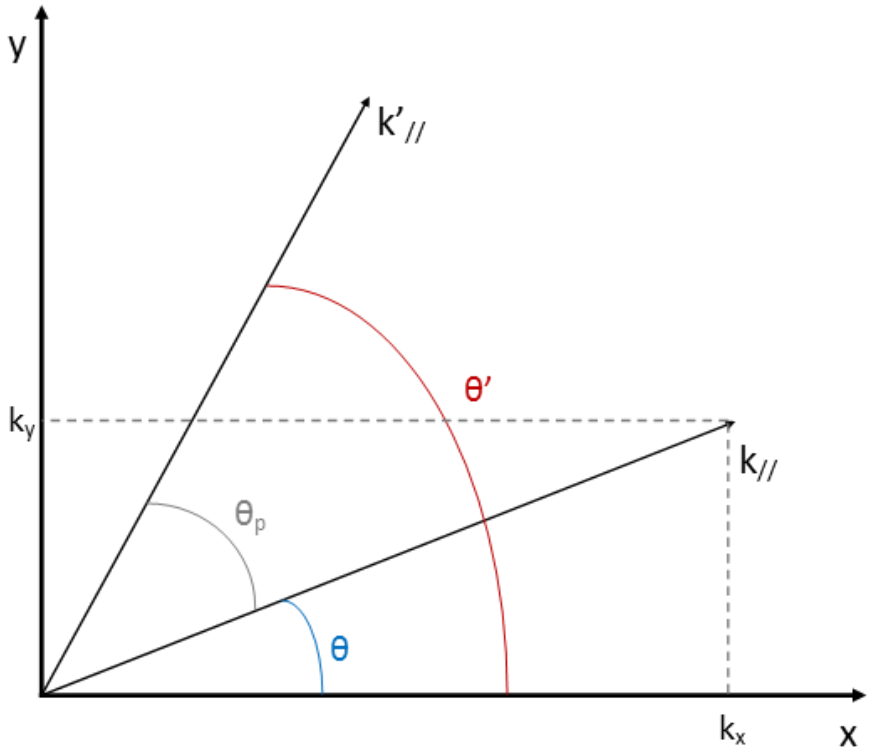


Figure 2.3.2: Illustration of the use of the Price scattering angle, θ_p, and initial wavevector direction with respect to the x-axis, θ, to generate the post-scatter angle between the final wavevector and the x-axis, θ'.

using the trigonometric functions cos(A+B) = cos(A)cos(B) – sin(A)sin(B), and sin(A+B) = sin(A)cos(B) + cos(A)sin(B). Using θ, θ_p and θ' from figure 2.3.2 this gives:

$$\cos(\theta') = \cos(\theta)\cos(\theta_p) - \sin(\theta)\sin(\theta_p) \qquad 2.3.16a$$

$$\sin(\theta') = \sin(\theta)\cos(\theta_p) + \cos(\theta)\sin(\theta_p) \qquad 2.3.16b$$

where $\cos(\theta_p)$ and $\sin(\theta_p)$ can be calculated using the *Price* scattering angle, and:

$$\cos(\theta) = \frac{k_x}{k_{//}} \qquad 2.3.17a$$

$$\sin(\theta) = \frac{k_y}{k_{//}} \qquad 2.3.17b$$

where k_x and k_y are the x and y components of the initial wavevector and $k_{//}$ is the magnitude of the initial wavevector.

2.3.5 2DEG Algorithm

The 2DEG algorithm follows the same procedure as the bulk algorithm explained in section 2.2.1, and illustrated in the flow diagram in figure 2.2.1. Before calculating the scattering table, the minimisation parameter and sub-band energy levels need to be determined. This is followed by filling the scattering table, with the use of numerical integration, where the energies used are the in-plane energies, as opposed to the full electron energy. In the 2DEG algorithm the scattering table is split into two sub-bands, where each sub-band has its own version of each scattering rate included. So, as in the bulk algorithm each valley has its own self-scattering constant, in the 2DEG algorithm each sub-band has its own self-scattering constant to reduce self-scattering. Since self-scattering has no effect on the electrons properties, when a self-scatter event is chosen, the computer wastes computational time. Therefore, reducing the number of self-scattering events optimises the algorithm in terms of run time, since less time is needlessly spent implementing a self-scatter, and also improves the accuracy in terms of calculating the time between two scattering events. The initial states are then set, again using a thermal electron distribution. Each electron is then simulated in turn. The electron drift is the same as described in section 2.2.4, as the applied field is still assumed to be along the x-axis. Choosing a scattering mechanism is the same as in bulk, but scattering is implemented in 2D as described in section 2.3.4. Transient data is output at the end of each time step, and steady state data at the end of each field step.

2.4 Simulating Non-Equilibrium Phonon Effects

To derive polar optical phonon scattering rates, a static and thermal phonon population is assumed. This implicitly assumes that any energy produced in a phonon emission process will decay instantly, which is unrealistic. When a phonon is emitted, this energy would be stored in the lattice and could possibly be reabsorbed. In an area with a significant

population of phonons, phonon emission can cause a shift away from equilibrium. When this occurs, 'hot phonons' are introduced, metaphorically relating an increase in phonons with an increase in the electron temperature. The process of reabsorption will increase the carrier relaxation rates, for both energy and momentum [65]. The effects of non-equilibrium phonons can be incorporated into an EMC algorithm using the method proposed by *Jacaboni* [74]. Non-equilibrium phonons can be included in the bulk EMC, however, since they are not used in the simulations for the work in this thesis this is not discussed here, instead how non-equilibrium phonons are included in the 2DEG EMC, assuming bulk phonons interacting with 2D electrons is explained in the next section.

2.4.1 2DEG Non-Equilibrium Phonon Algorithm

To implement non-equilibrium phonons, a phonon occupation histogram is created, defined as a grid in momentum space. After each POP scattering event the phonon is recorded. To record a phonon, a scattering angle must be determined and the wavevector calculated. This is performed using the *Price* rejection technique. The angle is determined using the procedure introduced in section 2.3.4.2.1, and the phonon wavevector is calculated from equation 2.3.12. Since the phonon distribution changes as the simulation proceeds, the pre-calculated POP scattering rates in the scattering table no longer apply and must be updated according to the changed phonon distribution. The non-equilibrium phonon distribution is relaxed back towards thermal after a length of time has passed that is equal to the phonon lifetime, which is known to be dependent on material, carrier density, temperature, phonon distribution and mechanism [65]. The carrier density is constant throughout the simulation, allowing a constant phonon lifetime to be used. Lifetimes have been measured in the range 0.1 to 2.5 ps [75]. The initial phonon distribution is set at thermal equilibrium, calculated using:

$$N = \frac{1}{(\exp\left[\frac{\mathrm{E}_q}{k_b T}\right] - 1)} \qquad 2.4.1$$

where N is the phonon occupation number and E_q is the phonon energy. The flowchart in figure 2.2.1 is updated for the non-equilibrium phonon algorithm and is shown in figure 2.4.1. In the non-equilibrium phonon simulation, transient and steady state velocity and energy can be output and compared to the equilibrium result, to investigate the effects of non-equilibrium phonons on electron transport characteristics. Phonon distributions, maximum phonon occupation number, average phonon occupation number and POP scattering rates can all be output as a function of time to investigate non-equilibrium phonon behaviour.

2.4.2 Phonon Occupancy Table

In the 2DEG algorithm, although the electrons are confined the phonons are assumed to be bulk and are treated as such. The bulk phonon model is employed for non-equilibrium phonons, following *Jacoboni* [74] and the approach used by *Ramonas et al.* [35] for non-equilibrium phonon effects in 2DEG channels. The phonon momentum is split into two directions, a component parallel to the applied electric field, q_x, and a component in the plane perpendicular to this, q_t, which can be written as:

$$q_t{}^2 = q_y{}^2 + q_z{}^2 \qquad 2.4.2$$

and

$$\boldsymbol{q}^2 = q_x{}^2 + q_t{}^2. \qquad 2.4.3$$

The grid in momentum space is then a set of concentric rings, shown in figure 2.4.2, where each ring corresponds to a section of the phonon occupation table.

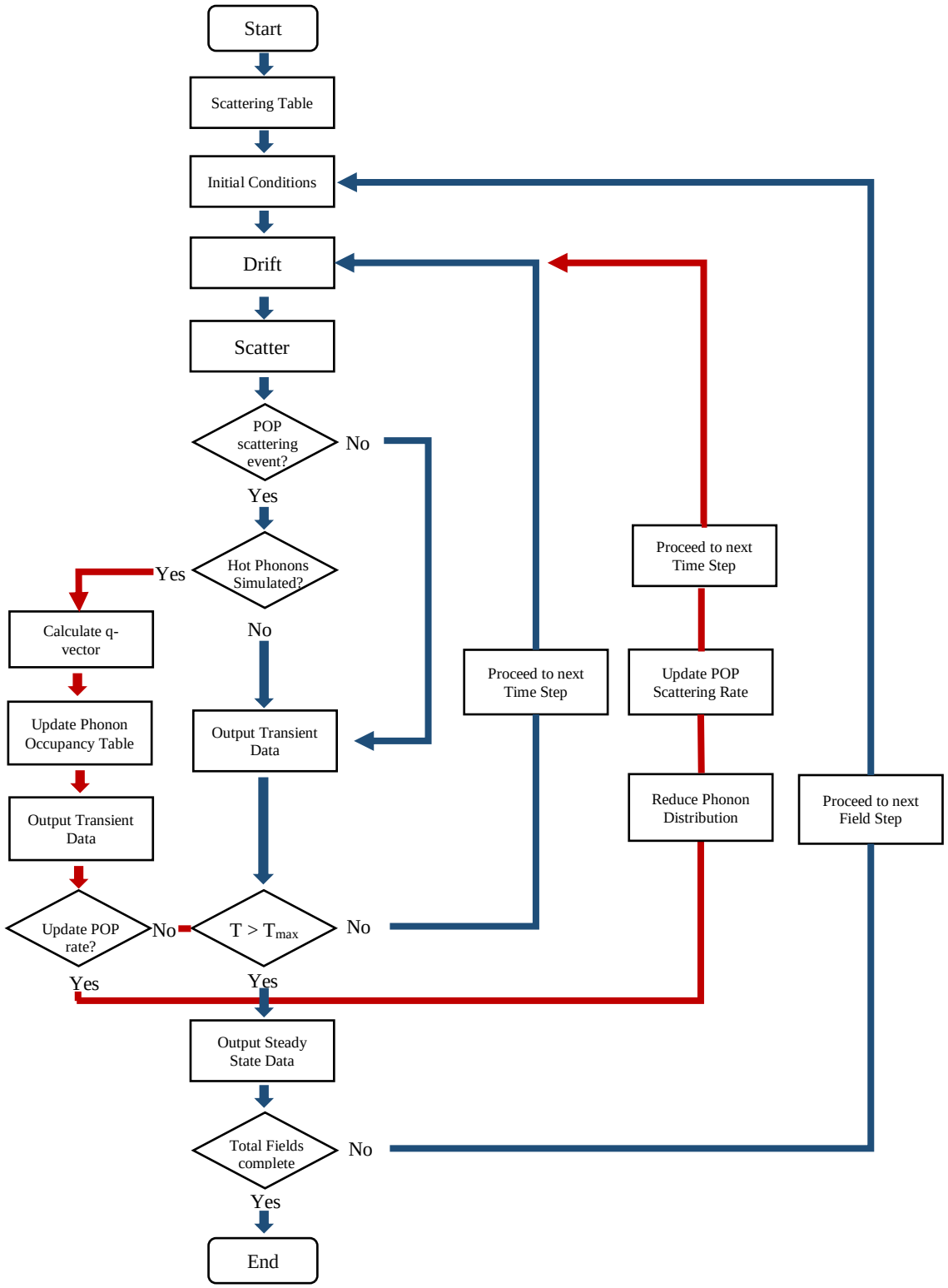


Figure 2.4.1: Flowchart diagram showing how non-equilibrium phonons are added to an EMC algorithm.

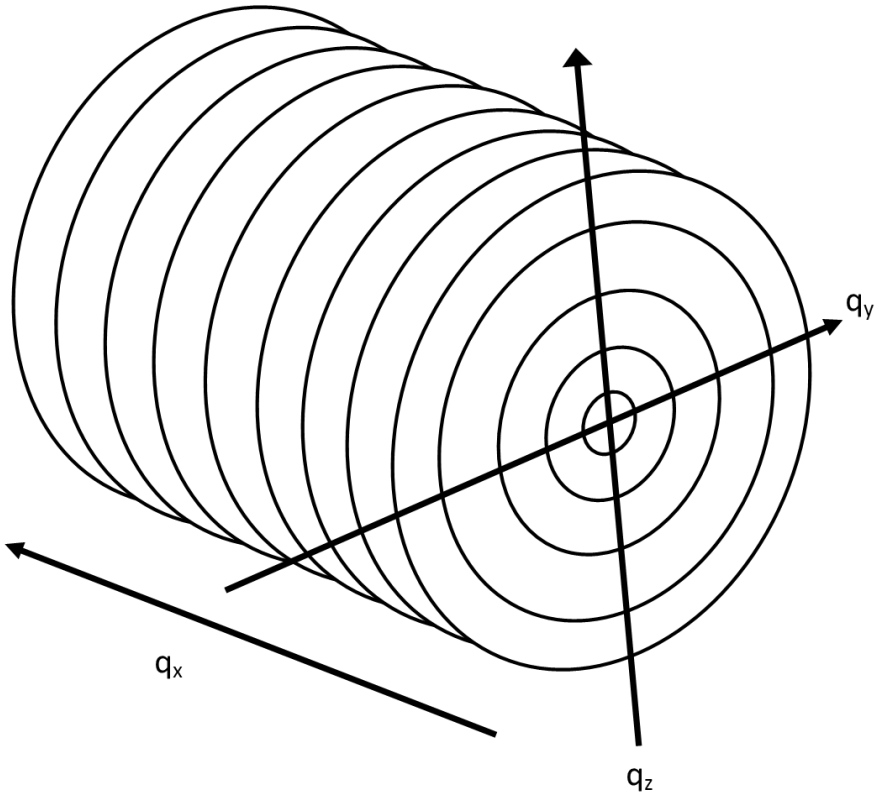


Figure 2.4.2: Schematic diagram showing the q-space grid, consisting of concentric circles, centred along the q_x axis and with increasing radii in the y-z plane.

2.5 Summary

This chapter has introduced how the ensemble Monte Carlo method can be implemented to simulate electron transport in various semiconductor structures. The algorithm to simulate transport in bulk semiconductor materials was presented first. The different stages of the simulation and the vast range of data available were explained. The range of data available is the main reason an ensemble Monte Carlo is used over a single electron Monte Carlo. The extra data being calculated and output has an effect on the run time, though this is mitigated by introducing parallelisation. The changes and additions required to simulate transport in a two-dimensional electron gas were then introduced, where a confining field creates a triangular quantum well. With the electrons now only free to move in the 2D plane perpendicular to confinement, the new scattering rates and their implementations were described. Finally, the process of adding bulk non-equilibrium phonon effects into the algorithms was introduced. This includes recording

the phonon scattering events to generate the true phonon distribution, and how this is used to recalculate the phonon scattering rates throughout the simulation.

Chapter 3

Bulk Ensemble Monte Carlo on a GPU

Computational simulations are an extremely popular method of scientific research and have been used in a wide range of disciplines for many decades. As computer technology advances, so do the simulations, allowing for more complex systems to be modelled. Of course, adding more complexity leads to more computationally expensive simulations. Complicated equations and calculations demand longer computation times, causing considerable run times. Assumptions and simplifications can be made to decrease complexity, at the expense of accuracy when compared to a real, physical system. Parallelisation is another common method of reducing run times, provided the simulation can be decomposed into components that can be run in parallel. This then allows multiple cores on a central processing unit (CPU) to run similar sections of the algorithm simultaneously. However, parallel computation can also be performed on the many (less powerful) cores on the graphics processing unit (GPU), this is known as general-purpose computing on graphics processing units (GPGPU), and gives a much higher level of parallelisation. In this chapter, GPGPU is introduced with an overview of the GPU architecture. The computer language used to run sections of the algorithm on the GPU is introduced and the properties of an ideal algorithm are explained, followed by the challenges that arise from attempting to run a Monte Carlo simulation on a GPU. The changes made to the bulk Monte Carlo algorithm (described in section 2.2) to be run on a GPU are then presented. These begin with architectural changes to the algorithm before moving on to GPU specific optimisations, such as memory and branching. Finally, more general changes that affect the underlying physics and can be performed on any algorithm are investigated. All changes and optimisations are presented with their effect on the performance, and where necessary, outputs are compared to those from the original bulk Monte Carlo to show any effect on the overall results. This is the first example, to my

knowledge, of an electron transport in semiconductor Monte Carlo simulation being performed on a GPU.

3.1 Introduction to NVIDIA CUDA & GPGPU

There are several languages that can be used for GPGPU implementations, OpenCL, OpenMP and OpenAAC all have the capability to run code on the GPU. CUDA is NVIDIA's own parallel computing platform and programming model, and is written specifically for the NVIDIA GPU hardware. CUDA can be used as an extension to several languages including C, C++ and FORTRAN and it allows for computationally intensive sections of code to be run on the GPU. With the original bulk ensemble Monte Carlo (EMC) algorithm written in C++, CUDA is the chosen language for this project for the similarities in the languages (CUDA is essentially the C language with added extensions to allow for highly parallel programming). In this section, CUDA is briefly introduced followed by an overview of GPGPU programming and the architecture. The characteristics of an 'ideal' GPU algorithm are presented and linked to the bulk EMC. Finally, this is followed by the problems that arise from transferring the EMC to the GPU.

3.1.1 NVIDIA CUDA

CUDA is a parallel computing programming language developed by NVIDIA to be used for GPGPU, specifically on NVIDIA GPU hardware. NVIDIA have created a number of CUDA-accelerated libraries that are readily available for developers to use, as well as a C/C++ compiler, nvcc. C++ algorithms can be transformed to use CUDA C++ by using the nvcc compiler along with the relevant CUDA libraries. A function to be invoked on the GPU is called a kernel.

3.1.2 Thread Blocks and Warps

Due to the enormous number of threads available on the GPU, as discussed in section 1.3 and shown in figure 1.3.1, threads are divided and execute instructions in a unique way. The NVIDIA GPU architecture is an array of Streaming Multiprocessors (SMs). Each SM contains a large number of threads (architecture dependent), as well as having its own shared memory, L1 cache and thousands of 32-bit registers. A CUDA application consists of CPU and GPU sections, the CPU section handles all of the function/kernel calling while the GPU section contains one or more kernels. Each kernel is executed across a large number of the GPUs threads, and all threads will be running identical code. CUDA algorithms are split into sub-problems that can be solved independently, and in parallel, by a block of threads. Threads within a block can cooperate when solving each sub-problem, by using shared memory. Each block is scheduled on to any available SM, in any order, either concurrently or sequentially depending on availability, such that the compiled CUDA program completes the whole algorithm on any number of SMs. The threads within a block always execute concurrently on one SM, while multiple blocks can also execute concurrently on one SM (dependent on number of threads and memory availability). SMs are designed to execute hundreds of threads simultaneously, and are a part of the SIMT (Single-Instruction, Multiple-Thread) taxonomy. The SM executes threads in groups of 32 called warps. Threads within a warp begin at the same instruction, but are free to branch and execute independently. A warp executes one single instruction at a time, meaning full efficiency is obtained when all 32 threads have the same execution path. If threads diverge due to a conditional branch, the warp executes each branch path taken sequentially, disabling threads that are not on the current path. Branch divergence only occurs within a warp, since all warps execute independently.

3.1.3 GPU Memory

There are a number of memory types on a GPU. Registers were already mentioned when introducing the SMs. In an ideal algorithm, all data would be stored in registers and other memory types would only be used for transferring data to and from the CPU, and for sharing data within a block. The remaining types, in order of performance (from fastest to slowest) are: constant, shared, local and global. GPUs also contain texture memory, but since there are no large matrices in the EMC, texture memory is not used. Constant memory is a read-only memory type. Reading data from constant memory is fast but has a small bandwidth, meaning only a small amount of data can be read at once. Reading too much data leads to several sequential read requests as the local cache becomes full. Each block of threads running on a SM has its own allocated shared memory that each thread has access to. This shared memory is divided into equally-sized modules, referred to as banks [76], which can be accessed simultaneously. Therefore, if multiple threads make a read or write request, where the memory addresses fall in separate banks, then they can be performed simultaneously. However, if multiple addresses are within the same memory bank, the access has to be serialised, and the memory request is split into as many separate 'conflict-free' requests as are required. Each thread has its own local memory, which is effectively a back-up space for registers. If a thread uses more registers than it has available, any further variables defined are placed in local memory, this is known as register spilling. Any large structures or arrays that would consume too much register space are also likely to be stored in local memory. Global memory has global scope, meaning all threads in all SMs have access to it. Local memory resides in device memory, so both local and global memory have high latency. Shared memory is on-chip, therefore has much higher bandwidth and lower latency than local and global memory.

3.1.4 Ideal GPGPU Algorithm

To fully utilise the parallel potential of the GPU and its architecture, an 'ideal' algorithm must have some specific features. Firstly, the algorithm must be highly parallelisable to make use of the vast amount of threads available. The parallelisable component must also be significant, in terms of run time, compared to any serial components (described by Amdahl's law [77]). As explained in section 3.1.2, an algorithm is split into thread blocks, and these thread blocks are run in warps of 32 threads. All threads in a warp run the same instruction simultaneously, any threads that follow a different branch are disabled and all branched instructions are run sequentially. Therefore, a perfect algorithm would ideally be able to be split into a number of blocks that have no branching conditions, meaning the threads follow the same flow. In terms of the bulk EMC algorithm, the highly parallel characteristic is present since a large number of electrons are simulated. For an ideal GPU algorithm, the electrons would all follow the same flow from start to finish of the algorithm, or at least be split in individual problems that have no branches, so all threads within a warp can run in unison. If the EMC algorithm consisted of a huge number of electrons that all follow the same path, then a GPU algorithm would see an enormous performance increase. Unfortunately, this is not the case for the bulk EMC algorithm and some problems arise.

3.1.5 Problems with Bulk Monte Carlo

The bulk EMC algorithm includes a large number of electrons (ideally 10,000s for accurate statistical analysis) and since electron-electron interactions are ignored each electron can run simultaneously, creating a highly parallel problem. It is relatively simple to parallelise the algorithm at the electron level, however, this introduces a number of issues for an 'ideal' GPGPU algorithm. Firstly, the large number of scattering mechanisms included in the simulation creates an equally large number of conditional

branches (12 when considering absorption/emission). As explained in section 3.1.2, all 32 threads in a warp execute the same instruction at the same time, if threads diverge due to a conditional branch, the warp must execute each branch taken, disabling threads not on this path. Hence, there is the possibility that in a single warp, all 12 scattering mechanisms occur and all 12 paths must be executed sequentially. As well as the high number of branches, the paths an electron can take are long and can involve a number of complex operations. Monte Carlo simulations are based on the random number generator creating random scattering events, therefore it would be very difficult to alleviate the issue created from the large number of branches, as it is possible for the electron to follow each path. The probability of each path being taken is weighted based on the relevant scattering mechanism for a given electron energy, hence some paths are more likely to be taken than others, but it is not possible to know when each path will be taken.

3.2 Algorithm Implementation on a GPU

In the previous section, an overview of the CUDA language and how an algorithm is divided and performed on the GPU architecture were introduced, along with some issues that arise from porting the bulk EMC algorithm onto the GPU. In this section, the changes made to the algorithm are presented along with their effect on performance. Firstly, architectural changes are introduced, such as where parallelism is included and the number of threads per block. This is followed by an explanation of the memory strategy taken, which includes the effects of copying the data from host (CPU) to device (GPU) and vice versa, how the different memory locations explained in section 3.1.3 are used and how the data is rearranged in memory.

3.2.1 Architectural Changes

The original bulk EMC algorithm, run on a CPU, is parallelised at the field level (as is the non-equilibrium phonon variant), meaning each core of the CPU is given a field step

to perform (see section 2.2.1). With the large number of threads available on the GPU, the parallelisation is moved to the electron level, which is each thread is given a single electron to execute, utilising the greater number of cores available. Each field step is run sequentially, however, it would be possible to run multiple field steps simultaneously on multiple GPU devices. The flow diagram shown in figure 2.2.1 needs to be changed, and the updated algorithm flow diagram is shown in figure 3.2.1.

3.2.1.1 Threads per Block

As explained in sections 3.1.1 and 3.1.2, the GPU divides the algorithm into thread blocks, and these thread blocks are scheduled on to available SMs. The number of threads in a block can be user defined, however, threads within a block are also separated into a number of warps (group of 32 threads) so it is common practice to have the thread block size as a multiple of 32, such that you have a block of full warps, rather than having to create a warp to run less than 32 threads. The number of threads per SM is architecture dependent, therefore the optimal number of threads per block to ensure the full usage of the SMs will vary depending on the graphics card used. The number of threads per block is varied from 24 (to study the effect of having a block size that is not a multiple of 32) to 192 to investigate the effect on simulation run time.

The results are shown in table 3.1, where the effect on performance is quantified as the percentage increase/decrease in run time compared to a block size of 32. A block size of 24 gives the longest run time, 11% longer than for a block size of 32. This is likely due to the number of disabled threads that occur from creating a warp of 32 threads to run blocks of 24, leaving 8 unused threads for each block that is created. The fastest run time is achieved for a block size of 64, which gives a 5% speed increase. Any further increase in block size results in a slower run time, and for block sizes exceeding 160, the run times exceed the result for a block size of 32. This is possibly due to the large number of

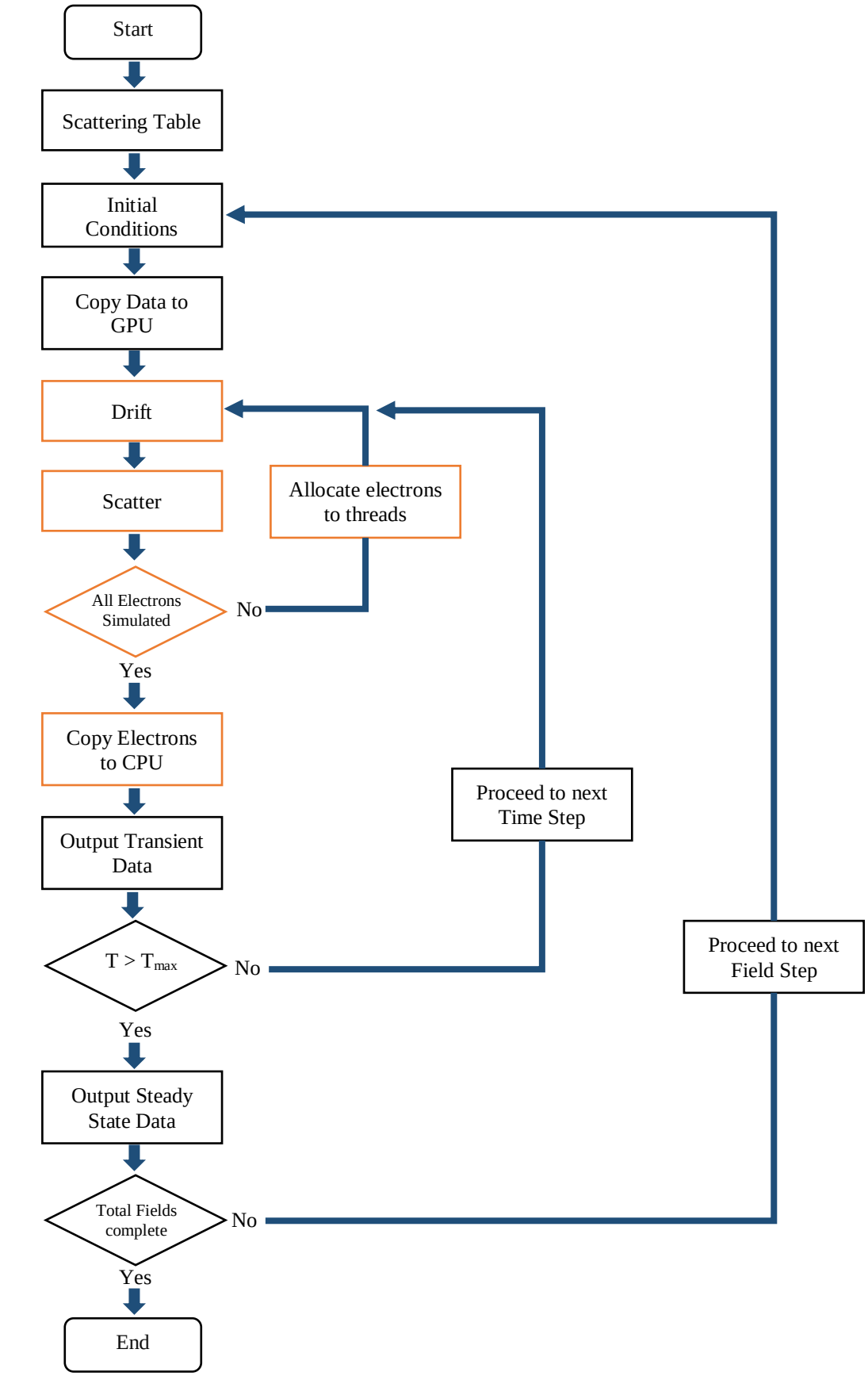


Figure 3.2.1: Flowchart showing an overview of how the base EMC algorithm was redesigned to run on a GPU. Orange borders represent sections performed on the GPU.

branches, as explained in section 3.1.5. The larger the block size, the more likely it is that more, or all, of the possible execution paths must be taken in a given warp.

Threads per block	Performance effect, % of 32 threads per block
24	+11.4
32	0
64	-5.1
96	-4.0
128	-2.3
160	+1.1
192	+8.6

Table 3.1: Investigation into the effect of number of threads per block on simulation run time.

3.2.1.2 CUDA Math Functions

The majority of NVIDIA GPUs are optimised for single precision floating point (floats) operations. Double precision (doubles) calculations take approximately twice as long as single precision calculations, but this performance is only available on a GPU that has been specified as part of a compute engine. Double precision performance on all GPUs designed specifically for graphics use has been restricted to 10% of the single precision performance. Therefore, all double precision floating point variables (doubles) in the algorithm were changed to single precision (floats). Another alteration made revolves around the mathematical functions used in the GPU sections of the code. NVIDIA CUDA has its own fast math library designed and optimised for use on the GPU [76]. For all sections of the code that run on the GPU, these CUDA functions are used. The simulation run time was investigated and it was found that including CUDA math functions produces a 14.3% speed increase compared to the original C++ math functions, a significant performance increase for a simple change.

3.2.2 Memory Strategy

With the many different memory locations explained in section 3.1.3, a strategy was created to best utilise these wherever possible. This section will explain the various methods used and their effect on performance. It begins with an experiment to investigate the effect of having to copy memory from the host (CPU) to device (GPU), and how the performance changes if we minimise the amount of memory copies required. This is followed by examples of how different memory locations are used. Firstly, how local memory was used to create a local copy of each electron. Then how constant memory is used throughout the simulation, including variables that are constant throughout the whole simulation as well as variables that are constant for each field step. Finally, modifications made to the structure of the data and how it is stored in memory are introduced, including electron parameters and the scattering table.

3.2.2.1 Memory Copy from Host to Device

To allow for sections of the algorithm to be performed on the device (GPU), it needs its own version of the relevant data. This data is copied from the host (CPU), and likewise, the data must be copied back from the device once all instructions are performed, such that the host has the updated data for the remainder of the simulation. In the bulk EMC, transient data is output at the end of each time step to investigate how electron properties evolve over time. Thus the electron data must be copied from the device to host at the end of each time step for the same transient data to be output. For accurate transient results, data is typically output at 500 stages throughout the simulation time, for each field. The effect of having transient output switched on/off was investigated. Removing transient output resulted in a 25% decrease in the overall run time. Therefore, it would be possible to obtain a performance increase by varying how frequently transient data is output from the simulation, the cost being the amount of transient data available to investigate (less

frequent transient output may result in the sparse data not showing the true evolution of electron properties over time).

3.2.2.2 Localise Electrons

Due to the large number of electrons, the GPU copy of the electron array is stored in global memory. As discussed in section 3.1.3, global memory has high latency, meaning any reads or writes are slow. To lessen this issue, when a thread is assigned an electron, each thread creates its own local copy from the global array and all calculations and updates are then performed on the local electron. The local electron is then used to update the global array at the end of the kernel execution. Creating a local electron produced a 15% speed increase.

3.2.2.3 Constant Memory

As discussed in section 3.1.3, reading from constant memory is fast but has low bandwidth so only a small amount of data can be read at any one time. Constant memory space is cached, and thus a read from constant memory costs one read from the constant cache, or one read from device memory on a cache miss. If threads within a warp attempt to access different addresses, these are serialised and as a result the cost scales linearly with the number of serialised requests. Hence, the constant cache is best when threads within a warp access only a few distinct memory locations. If all threads in a warp access the same location, i.e. require the same variable, constant memory can be extremely fast. Therefore, all variables in the simulation that are constant for all electrons, and are required concurrently, are stored in constant memory. These include such things as the number of electrons in the simulation, the energy step size in the scattering table and the thermal energy. In every instance of an electron using π in a calculation, it is multiplied by 2, thus the value of 2π is calculated once and stored in constant memory.

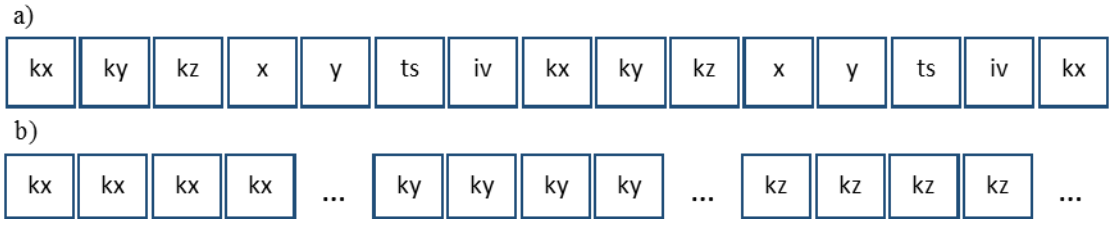


Figure 3.2.2: Illustration of electron parameters stored in memory, a) shows original structure from CPU bulk algorithm with electrons stored one by one, b) shows new design for CUDA algorithm with parameters stored one by one, … represents parameters being continued.

3.2.2.4 Memory Coalescing

Global memory is accessed via 32-, 64-, or 128-byte memory transactions. This means that when a thread reads from global memory, it does not read in a single variable, it reads the 32-, 64- or 128-byte segment in which the variable is aligned. In the bulk EMC algorithm, all reads are for an integer or a float (4 bytes), meaning every read would be a 32-byte read. If the next thread requires a variable that is in this 32-byte segment, the data is readily available. However, if the next thread requires a variable that is not in this segment, another 32-byte read is required and if this memory is not already in the cache, a cache miss occurs, there is then a delay until the new 32-byte section of memory is loaded into the cache and is ready to be read. Therefore, it is extremely important, and beneficial, to coalesce the data in memory such that the data required by alternate threads is adjacent in memory locations. In the CPU algorithm, a structure is created (Electron) that contains all of the electron parameters (kx, ky, kz etc.), and then an array of Electrons is created. This means that in memory, electrons are stored one after another. All the parameters for one electron are stored together, then the parameters for the next electron are stored and so on. Hence, the same parameter for two neighbouring electrons are separated in memory, therefore multiple reads from global memory would be required. A new array was designed which stores each parameter continuously, one after another, e.g. all electron kx's, followed by ky's and so on, as illustrated in figure 3.2.2. This ensures that in the kernel, when an electron parameter is read from memory, the 32-byte segment

will contain the same parameter for the next electron. This reduces the number of cache misses immensely and improves the performance of data acquisition. It has the same effect at the end of a kernel when the electron parameter is updated, the memory where the parameter is to be written is also readily available. The combination of storing variables in constant memory, and coalescing electron parameters, produced a 10% performance increase.

3.2.3 Execution Optimisations

As discussed in section 3.1.2, the GPU separates threads into warps, and all 32 threads within a warp perform the same instruction simultaneously. When threads within a warp follow different execution paths, this is known as divergence and each path is performed sequentially. Threads not on the current execution path are disabled, drastically reducing occupancy and performance. Of course, this makes flow control instructions (if, switch, while, for example) detrimental to performance due to the creation of different execution paths. The different execution paths within a flow control instruction are known as branches. Branches with just a few instructions generally result in marginal performance losses. However, branches with many instructions, or flow control instructions with many branches, tend to result in significant performance losses. Unfortunately, the bulk EMC algorithm contains flow control instructions with a large number of branches, due to the high number of scattering mechanisms included. The remainder of this subsection explains changes made to the flow control instructions in the algorithm in an attempt to improve occupancy and minimise the performance losses.

3.2.3.1 Zero Branching

The best case scenario would be to have all electrons follow the exact same execution path. This would require all electrons (within a warp) to choose and implement the same scattering mechanism. To perform a test of this scenario, the scattering function was

replaced by a new function which only implemented acoustic scattering. Acoustic scattering causes no change to the electron energy and is isotropic, meaning the final wavevector calculations use the least, and simplest, code. The new routine included no branching statements, purely the code required to implement an acoustic scattering event, meaning all electrons within a warp now perform the same instruction concurrently for the whole scattering implementation. Utilising the new scattering mechanism produced a 40% performance increase compared to the simulation including all scattering mechanisms. This gives an upper bound to the execution optimisations, having only one scattering mechanism, and hence only one execution path, is the ideal case for GPU performance. All electrons within a warp perform the same instructions and there are no conditional branches to cause threads to be disabled, reducing occupancy and performance.

3.2.3.2 If-then-else Statement

When choosing a scattering mechanism, a random number between 0 and 1 is compared to the weighted scattering table for the current electron energy (see section 2.2.2). There are twelve possible scattering mechanisms, including self-scattering, therefore this is performed in the original EMC in a 12-branch if-then-else statement, starting with scattering mechanism 1, and continuing until the scattering mechanism is chosen. Hence, if a large number (or potentially all 12) of the scattering mechanisms are chosen in a single warp, multiple execution paths are created and there are several instances of threads being disabled. However, the implementation of certain scattering mechanisms are the same, or very similar, meaning they can be grouped together. For example, piezoelectric and acoustic scattering both cause no change in the electron energy and are grouped together, the only difference is acoustic scattering is isotropic whereas piezoelectric scattering is anisotropic. For scattering mechanisms such as non-polar/polar optical phonon, the magnitude of the energy change is the same (phonon energy), but since the

scatter is either by absorption or emission, the only difference is whether the phonon energy is added or subtracted from the initial energy. On each thread, a local float is created to store the sign of the energy change (either 1.0 for absorption or -1.0 for emission), as well as a local Boolean to store whether the scattering is isotropic or anisotropic (stored as either true or false). The scattering mechanisms are collected into six groups, polar optical phonon (POP), non-polar optical phonon (NPOP), piezoelectric and acoustic, inter-valley to lower band, inter-valley (IV) to upper band and impurity and self-scattering.

When choosing a scattering mechanism, the 12-branch if-then-else statement was replaced by a binary subdivision if statement. Rather than comparing the random number against each scattering mechanism (or scattering group) individually, the binary subdivision dissects the scattering table each step. Firstly, checking whether the random number selects one of the first six, or last six, scattering mechanisms. Then by splitting each six into four and two. Since the scattering mechanisms are in six groups of pairs, if the random number corresponds to the group of two, then the scattering group is selected after just two steps. If the random number corresponds to the group of four, a third step is needed to determine the scattering group. In each scattering group, whether the scattering is absorption or emission, or isotropic or anisotropic (if required), is decided and stored in the relevant local variable. Consequently, the 12-branch if-then-else statement has been transformed to a 6-branch binary if statement, where the minimum number of steps required to make a decision is two, and the maximum is three. This can be compared to the minimum of one, but maximum of twelve originally. This lowers the number of possible execution paths and thus reduces the amount of time spent with disabled threads. In order for the scattering groups and binary subdivision to be used, the order in which the scattering mechanisms appear in the scattering table must be rearranged.

3.2.3.3 Switch Statement

Similarly, when implementing the energy change caused by a scattering event in the original EMC, a 12-branch switch statement is used, with one case for each scattering mechanism. Therefore, if this method were used on the GPU, if a large number of scattering mechanisms are encountered in a single warp, each implementation is performed sequentially leading to several instances of threads being disabled. However, since scattering mechanisms with similar implementations are grouped together, the 12-branch switch statement is replaced by a new switch statement with only five branches. The six groups from the previous section are reduced further to five groups. Acoustic, piezoelectric and impurity scattering all have no effect on the electron energy, and so can be implemented together. The remaining scattering groups (POP, NPOP and IV) are implemented individually, with absorption and emission scattering events implemented at the same time based on the predetermined sign as discussed in section 3.2.3.2. The combination of reordering the scattering table into groups, using a binary subdivision approach to choosing a scattering mechanism, and implementing like scattering mechanisms simultaneously, caused an approximately 4% performance increase. In section 3.2.3.1, the ideal scenario of having zero branches and only one execution produced a performance increase of 40%. This shows that limiting the number of execution paths is vitally important in optimising the performance of the algorithm. Unfortunately, this means that including more scattering mechanisms (hence more execution branches) will have a negative effect on performance as is seen by the relatively small 4% increase (when compared to the ideal scenario) obtained when all scattering mechanisms are included.

3.2.4 General Physics Simulation Changes

As well as the changes made based on the specific architecture of the GPU and its memory layout, there are also some general changes that were made. Although these changes were influenced by the GPU architecture, they could be made to any physics simulation to obtain a performance increase. This subsection introduces the changes made along with their effect on the performance, and output, whenever applicable. Experiments were also performed that changed the underlying physics of the simulation to investigate the effect on performance and accuracy of the output and these changes are also explained.

3.2.4.1 Doubles to Floats

The first general change was already mentioned in section 3.2.1.2 and is highly influenced by the GPU architecture, but is a more general change that could be made in physics simulations to potentially obtain a performance increase. Changing between double (doubles) and single (floats) precision floating point numbers is essentially a trade-off between performance and precision. Doubles have twice the precision of floats but are also double the size in memory, to be able to store the extra precision. In the GPU case, section 3.2.1.2 explained how CUDA has optimised mathematical functions for float operations and so it is extremely beneficial to use floats on a GPU. Double precision operations can be up to 32x slower than single precision operations [76], on a consumer GPU card. Double operations are 2x slower on a GPU compute card, however, these are vastly more expensive. It is found that the change from doubles to floats has minimal effect on the simulation output, as shown in figure 3.2.3. The steady state outputs from the C++ EMC, using doubles, are compared to the CUDA EMC, using floats. The velocity output from the CUDA EMC is slightly higher at low fields, however, as the field increases the outputs begin to match more closely. The energy outputs are extremely close

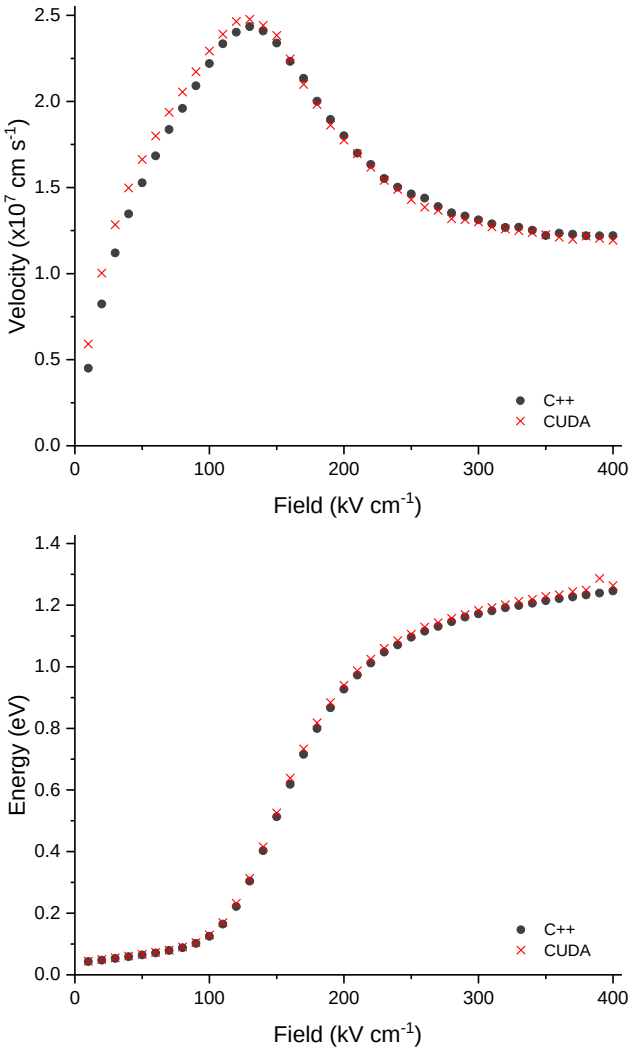


Figure 3.2.3: Steady state velocity and energy results, comparing results from the original C++ code (using doubles) to the CUDA code (using floats).

for all field strengths. The minor differences may be due to the change in the random number generator on the GPU and the order in which the numbers are accessed.

3.2.4.2 Time Step Duration

The total EMC simulation time is separated into a number of individual time steps. The duration of these time steps is dependent on how often the transient data is required, and for a simulation including hot phonons also depends on the phonon lifetime and how regularly phonons need to be updated. Hot phonons are not included in the GPU algorithm so these dependencies can be ignored. In some simulations, such as device algorithms, a mesh is used and the electrons positions are monitored in order to regularly calculate the

charge distribution, and the time step must be considered accordingly. However, in the bulk EMC, the time step is purely used to take snapshots of electron properties for transient data. In the original EMC, the time step was set to 0.1 fs, however, transient data was output every 1 fs (or 100 time steps). Using this time step on the GPU led to relatively poor performance, i.e. the speed increase in comparison to the CPU algorithm was low. Investigating this with the NVIDIA Visual Profiler, it was found that the occupancy and thread utilisation was extremely low. The reason was discovered to be that with a short time step, only a small percentage of electrons would scatter, meaning a large percentage of threads were disabled while these few scattering events occurred. The time step was increased to 1 fs, and transient output was now performed after every time step. This ensured that a much higher percentage of electrons encountered a scattering event, meaning fewer threads were disabled. This led to an enormous 80% performance increase, and the change in time step had minimal effect on the output. The longer time step was vital for improved performance on the GPU, however, the same change produced a similar performance increase on the CPU algorithm. Evidently, having a longer time step, meaning fewer stop-start points, produces the best performance. This is logical, given that the end point is the same, it is simply reached in fewer steps.

3.2.4.3 Limit Scatters per Time Step

With the time step increased, and the vast majority of electrons encountering a scattering event, the occupancy and thread utilisation was still found to be low during some periods of the simulation. Investigating this, the low occupancy was found to be caused by a small percentage of electrons scattering a high number of times, compared to the rest of the electrons. The majority of electrons would encounter two or three scattering events during a time step, whereas a very small percentage of electrons scattered six or seven times. To investigate the effect on performance, a maximum number of scatter events per time step was set and experimented with. Logically, limiting the number of scattering events stops

the small percentage of electrons scattering more than the majority, preventing threads being disabled, hence should lead to a speed increase. However, by limiting the number of scattering events, the simulation is made less physical as the scattering events are produced based on the scattering rates and related probabilities. Setting the maximum number of scattering events to five produced a 2% performance increase, whereas limiting the scattering events to four produced a 5% performance increase. In both cases, the steady state results were within 1% of the original results. However, the transient results saw on average a 4% change when the limit was five, and a 5.5% change when the limit was four. This shows that setting a limit on the number of scattering events does have an effect on performance, but also has an effect on the simulation output, as expected. Due to the effect on the results and the unphysical nature of having a limit on scattering events, for a relatively small performance increase (the 2% increase is minimal, whereas the 5% increase comes at the expense of a much higher effect on the results), the scattering limit is removed from the final GPU simulation.

3.2.5 Timing Results

So far, the results of implementing changes have been presented as a relative percentage performance improvement. Calculated as the percentage change in run time from the previous version of the algorithm (i.e. a 5% performance increase represents a reduction of 5% of the initial run time). Here, the run times of certain main stages are presented in seconds to give some perspective. The initial C++ algorithm, running on a single core of the CPU, had a run time of 336 s. A direct switch to the GPU, and after changing the block size to 64, which was found to have the best performance, the CUDA algorithm had a run time of 166 s, approximately 50% of the C++ run time. After changing the duration of the time step from 0.1 fs to 1 fs, the C++ algorithm had a run time of 56.5 s, while the CUDA algorithm had a run time of 23.9 s. The CUDA algorithm now has a run

time approximately 40% of the C++ run time, showing that the increase in time step duration, leading to more electrons scattering in the current time step (hence improving occupancy), produces a greater performance increase. After implementing further changes, including swapping all C++ math functions with CUDA math functions, grouping like scattering mechanisms and adding the binary subdivision if-then-else statement, the final CUDA run time was 18.0 s, approximately 30% of the C++ run time. The CPU used for these timings was an Intel Core i5-3570K 3.40 GHz and the GPU used was an NVIDIA GeForce GTX 550 Ti, both of which are considered mid-range for a standard desktop PC.

Algorithm Status	Time (s)
Original C++ (CPU)	336
Original CUDA (GPU, blocksize = 64)	166
Time Step Change, (CPU)	56.6
Time Step Change, (GPU)	23.9
Final Version, (GPU)	18.0

Table 3.2: Table of run times at various significant stages throughout the CUDA algorithm development.

3.3 Summary

This chapter has introduced how general purpose computing on the graphics card is run using NVIDIA CUDA. It is shown how the algorithm is separated into individual blocks of threads, consisting of warps of 32 threads, to perform a set of instructions, and the problems this creates for a Monte Carlo algorithm. The changes required to utilise the GPU and overcome some of these problems were introduced and it was proven that CUDA can be used to improve the performance of Monte Carlo simulations. A direct move from CPU to GPU, with the only change being where the parallelisation occurs, saw an approximate 45% performance increase, and varying the number of threads per

block improved this to a 50% performance increase. Further changes inspired by the GPU architecture, memory layout and NVIDIAs optimised math libraries were introduced to show how the GPU algorithm can begin to be optimised. The simple change of replacing the mathematical functions with NVIDIAs own optimised math library produced a 14% performance increase. Creating a memory strategy to store parameters that are constant for all electrons in constant memory, and reordering electron parameters to coalesce them in memory, also produced a significant performance increase of approximately 10%. The biggest performance increase of 80% came from a more general physics simulation change of increasing the length of the time step. Although this was inspired by investigations into optimising the GPU occupancy, it was found to have an equally substantial effect on the CPU algorithm. It was also shown that changing the simulation variables from doubles to floats had minimal effect on the output, suggesting that in any physics simulation a performance increase could be obtained by using floats instead of doubles, so long as the accuracy of the results remains consistent.

Chapter 4

2DEG Scattering in Gallium Nitride

The algorithm described in section 2.3 has been used to investigate the transport properties of a two-dimensional electron gas (2DEG) created at a Gallium Nitride/Aluminium Gallium Nitride (GaN/AlGaN) interface. A triangular well of infinite height is assumed with the two lowest sub-bands considered, as shown in figure 4.1. A uniform electric field is assumed across the whole system, in the x-direction. A confining field is assumed in the z-direction (direction of confinement) that creates the triangular well and determines the effective well width and the minimisation parameter. The sub-band energy levels for each confining field and the material parameters are shown in tables 4.1 and 4.2 respectively. The scattering mechanisms included are non-polar optical phonon, polar optical phonon, acoustic phonon, and alloy scattering. All mechanisms can cause an inter-band scattering event (see section 2.3.3 for more details on the inter-band scattering rates). Scattering due to impurities is not included as the structure is assumed to be intrinsic. Piezoelectric scattering is also not included as it has been reported that the piezoelectric component of acoustic scattering in AlGaN/GaN 2DEGs is much weaker than the deformation potential scattering and thus can be neglected [78]. For all results in this chapter, the 2DEG Ensemble Monte Carlo (EMC) code is run using a parabolic band approximation, to allow for the separation of the in-plane (x-y) and confined (z) energies and wavevectors. The simulation parameters are shown in table 4.3. The chapter begins with investigating the steady state results, comparing velocity results to published experimental data. Relaxation times and low-field mobility results are then compared to published experimental and theoretical data. Transient properties are compared to bulk GaN results, specifically an investigation into whether there is the presence of a velocity overshoot, as is seen in bulk GaN [2, 4, 20, 21]. The chapter finishes with an investigation of the effect of the alloy disorder potential on the velocity and mobility results. The

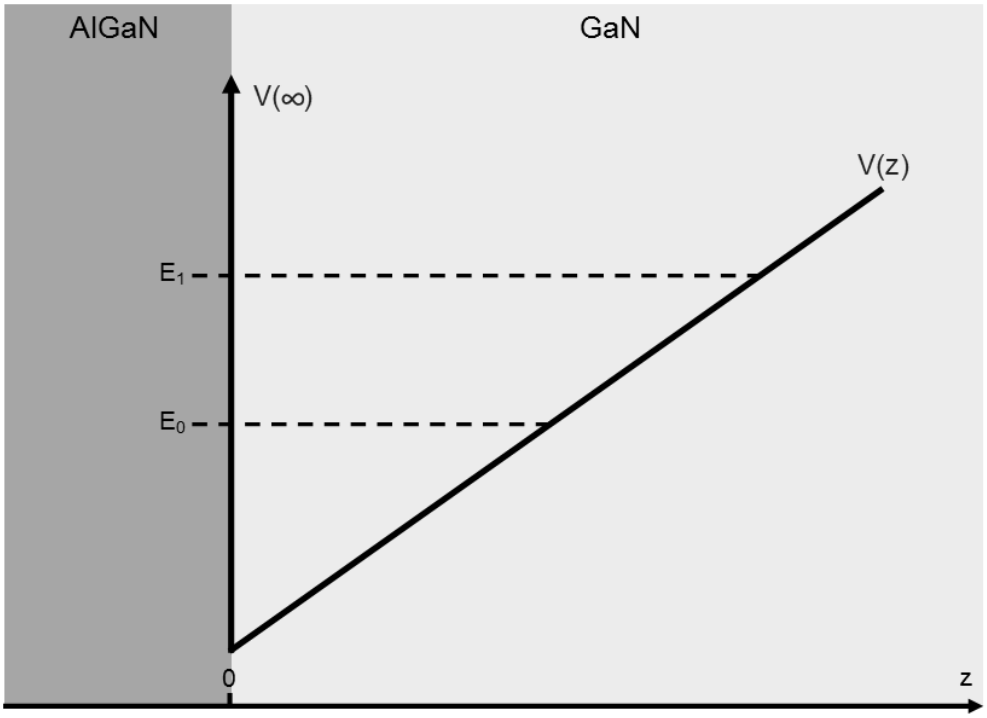


Figure 4.1: Schematic diagram of a triangular potential well at an AlGaN/GaN interface. E_0 and E_1 represent the sub-band energy levels, V(∞) represents the infinite potential assumed at the interface (z=0), V(z) is the potential as a function of z, i.e. the depth into the GaN layer.

inclusion of electron confinement for AlGaN/GaN HEMT simulations is novel compared to current literature.

Confining Field (kV m^{-1})	Equivalent Sheet Density (x10^{12} cm^{-2})	E_0 Sub-band Energy Level (eV)	E_1 Sub-band Energy Level (eV)
250	0.434	0.115	0.201
500	0.868	0.183	0.319
750	1.30	0.239	0.418
1000	1.74	0.290	0.507

Table 4.1: Sub-band energy levels, E_0 and E_1, for varying confining field strengths, F_z, and the equivalent sheet densities (sheet density-electric field conversion shown in Appendix A).

4.1 Steady State

Perhaps the easiest, and most important, data to obtain from an EMC is the steady state ensemble average velocity and energy characteristics. It is simple to output at the end of

the total simulation time (as explained in chapter 2). Steady state velocity results are regularly obtained from experiment and are readily available for comparison.

Parameter (units)	GaN
Density (kgm^{-3})	6150
Longitudinal sound velocity (ms^{-1})	6560
Static dielectric constant (ε_0)	8.9
High frequency dielectric constant (ε_0)	5.35
Effective mass (m_e)	0.2
Acoustic deformation potential (eV)	8.3
Non-polar optical deformation potential coupling constant (eVm^{-1})	10^{11}
Alloy disorder potential (eV)	1.5
Volume of primitive cell ($x10^{-30}$ m^3)	46.943
Polar optical phonon energy (meV)	91.2
Non-polar optical phonon energy (meV)	91.2
Mole-fraction composition (in AlGaN layer)	0.3

Table 4.2: Gallium Nitride parameters, at 300 K, used in simulations. Parameters obtained from [4, 21, 79-83].

Parameter (units)	Value
Number of electrons	100,000
Number of sub-bands	2
Time per time step, dt (fs)	2
Number of time steps	1000
Lattice temperature (K)	300
Confining field strength, F_z ($kVcm^{-1}$)	250-1000
Applied field strength, F_x, step size ($kVcm^{-1}$)	1/5
Number of field steps	25

Table 4.3: Simulation parameters used for all AlGaN/GaN 2DEG simulations.

4.1.1 Velocity

Steady state velocity characteristics are one of the results that can be obtained from experiment and can be compared to EMC code results to verify that the algorithm is giving physically sensible results. Here, the velocity field characteristics are compared with published experimental results and other 2DEG algorithms, for verification, as well as including a comparison to the bulk results. The steady state velocity-field characteristics generated for a GaN 2DEG are shown in figure 4.1.1, it is evident that

alloy scattering is limiting the velocity and this will be addressed in section 4.5.1. Results for a range of confining fields from 250 to 1000 $kVcm^{-1}$ are compared to results obtained from the bulk EMC simulation, employing non-parabolic bands. As shown in figure 4.1.1, the velocities output from the 2DEG EMC (for varying confining fields) are all similar to the bulk results. Figure 4.1.2 shows the 2DEG results for a confining field strength of 500 $kVcm^{-1}$, along with experimental results from *Matulionis et al.* [16] and *Palacios et al.* [33]. Both experimental set ups have significantly higher carrier densities than the simulation using a 500 $kVcm^{-1}$ confining field. *Matulionis et al.* report a carrier density of $1.2x10^{13}$ cm^{-2}, while *Palacios et al.* report $1.46x10^{13}$ cm^{-2}, compared to the value of $0.868x10^{12}$ cm^{-2} from table 4.1. This shows that the 2DEG simulation results compare remarkably well, given the differences in sheet densities, with the *Palacios et al.* results at low applied electric fields, and begin to compare well with the *Matulionis* results at larger field strengths (50-60 $kVcm^{-1}$). However, the 2DEG results continue to increase after 60 $kVcm^{-1}$ and begin to exceed the experimental results by a considerable margin. The steady state velocity-field results are compared to more experimental results in chapter 5 when non-equilibrium phonon effects are introduced to the algorithm.

4.1.2 Energy

The steady state ensemble average electron (total) energy vs applied electric field results are shown in figure 4.1.3. The energy remains around the sub-band energy levels at low fields, increasing slightly with field. Once the applied electric field passes 60 $kVcm^{-1}$, the energy begins to increase rapidly and very quickly reaches the limits of the scattering table. The maximum applied electric field is limited in order to mitigate the effects of scattering events with electron energies that are outside of the scattering table. Capping the field at 100 $kVcm^{-1}$, where the average electron energy begins to approach the scattering table limit, 2.5 eV. When the confining field strength is 250 $kVcm^{-1}$, the

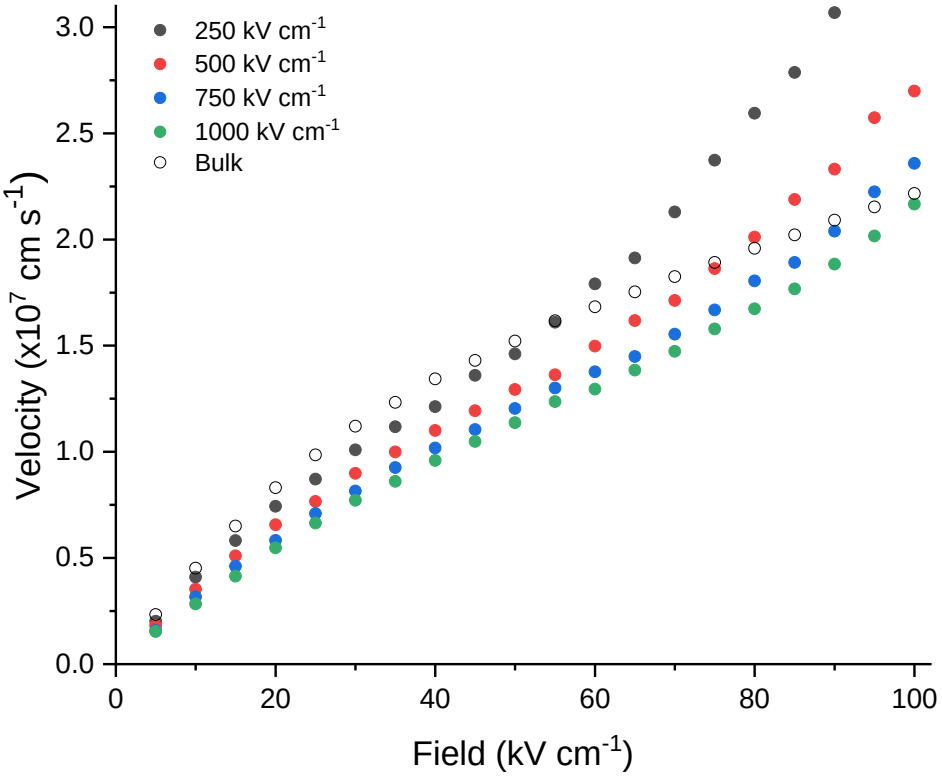


Figure 4.1.1: Average electron velocity vs applied electric field, confining fields from 250-1000 kVcm^{-1}, compared to results from the bulk Monte Carlo.

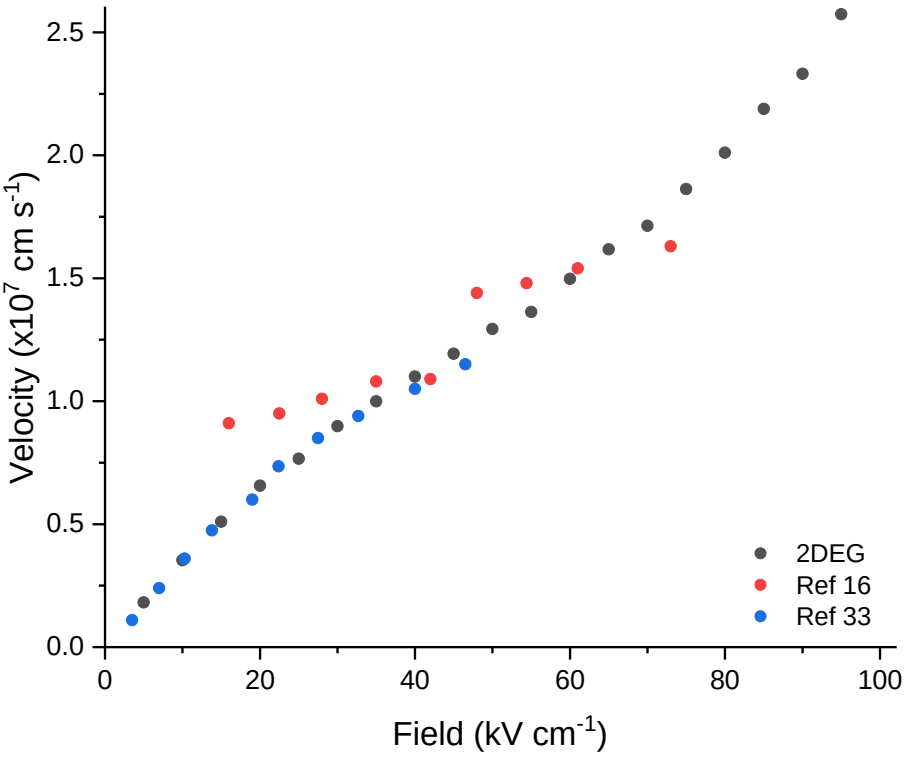


Figure 4.1.2: Average electron velocity vs field for the 2DEG (500 kVcm^{-1} confining field), compared to experimental results from Matulionis [16] and Palacios [33].

average electron energy surpasses the maximum energy of the scattering table at high applied field strengths. This could be a possible explanation for the hint of a runaway in the velocity-field plot at this confining field strength in figure 4.1.1. For a confining field strength of 500 $kVcm^{-1}$, the average electron energy rapidly approaches the scattering table limit at higher applied fields, which could explain the slight runaway in the velocity in figure 4.1.1, and be the reason for the velocity at this confining field exceeding the experimental data from *Matulionis* in figure 4.1.2. The energy increase is much slower for higher confining fields, which is thought to be linked to the sub-band occupancies. Higher confining fields create a larger separation between the sub-band energies, meaning fewer electrons make the transition to the excited sub-band at lower fields. Figure 4.1.4 shows the sub-band occupancies vs applied field. It shows that for a confining field of 1000 $kVcm^{-1}$, the occupancy of the second sub-band only begins to increase at larger fields. The percentage of electrons in the second sub-band at the highest applied field (100 $kVcm^{-1}$) is ~20% for a confining field of 1000 $kVcm^{-1}$, as opposed to ~40% when the confining field is 250 $kVcm^{-1}$. The energy runaway has been explained by *Ridley* [65] as an effect caused by 'hot electrons', meaning their average energy rises above thermal equilibrium. Of course, this occurs at all field strengths, but becomes increasingly noticeable at higher field strengths. *Ridley* [65] showed that assuming a purely parabolic conduction band, and that the energy dependence of the relaxation times can be represented as a simple analytic solution (AE^p, where A and p are constants, E is the electron energy), the energy will increase with the square of the field. A parabolic band approximation is used in this algorithm, and figure 4.1.3 shows a parabolic nature to the energy increase past an applied electric field of 60 $kVcm^{-1}$. Results from this point onwards, including future chapters, are presented up to 60 $kVcm^{-1}$, to more clearly investigate the electron behaviour before the energy runaway causes unphysical behaviour.

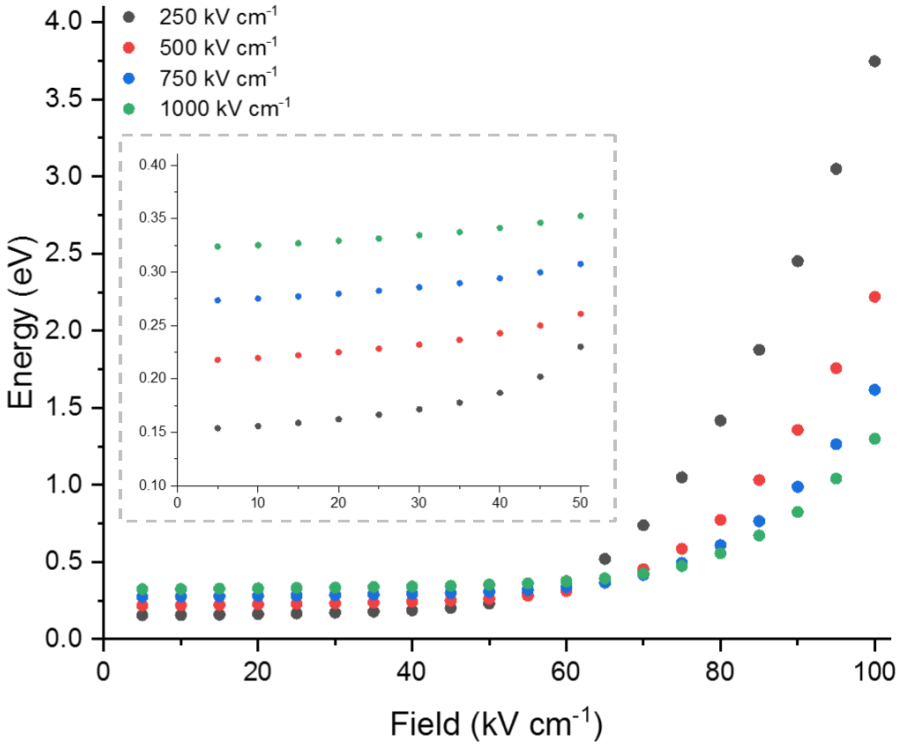


Figure 4.1.3: Steady state ensemble average total electron energy vs field plots for a range of confining electric fields, showing a rapid increase in electron energy at high fields. Subset shows the low field range.

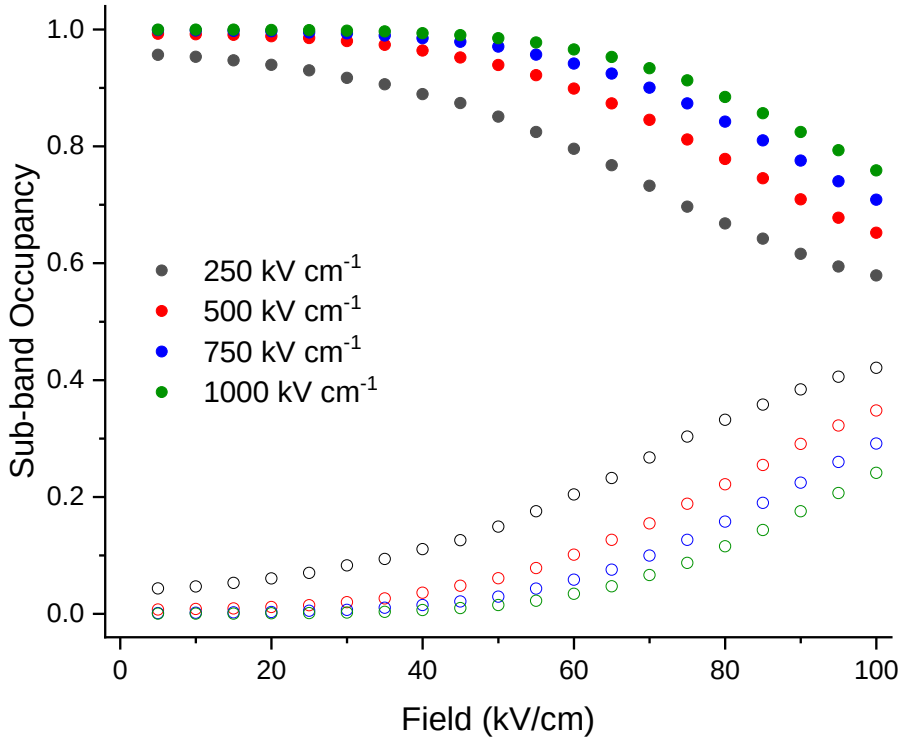


Figure 4.1.4: First sub-band (solid circles) and second sub-band (open circles) occupancy vs field for a range of confining fields.

4.2 Relaxation Times

In a semiconductor, the distribution of electrons have a tendency to approach equilibrium due to scattering events. There are several relaxation processes, and the momentum relaxation time is considerably lower than the energy relaxation time. Almost all scattering mechanisms will relax momentum, whereas only polar optical phonon scattering relaxes energy [84]. Relaxation rates are a regularly investigated characteristic. Energy, τ_e, and momentum, τ_m, relaxation times can be calculated using [85]:

$$\tau_e = \frac{\langle \mathrm{E} \rangle - \mathrm{E}_0}{eF\langle v \rangle}, \tag{4.2.1}$$

$$\tau_m = \frac{\langle p \rangle}{eF} \tag{4.2.2}$$

where $\langle\ \rangle$ represents the ensemble average, E is the electron energy, v is velocity, p is momentum, e is the electronic charge, F is the applied electric field and E_0 is the zero-field average electron energy. Since all electrons in the simulation have the same effective mass, the ensemble average momentum, $\langle p \rangle$, is simply calculated as $m^*\langle v \rangle$.

4.2.1 Momentum Relaxation

The momentum relaxation times, calculated for each field step using equation 4.2.2, are shown in figure 4.2.1, for a confining field of 500 kVcm^{-1}. At low applied electric field strengths, the relaxation time is at a maximum of 41 fs. As the applied field increases, the relaxation times steadily decreases to approximately 28 fs. After 50 kVcm^{-1}, the decrease begins to slow. *Bulutay* plots relaxation times versus electron energy, for bulk GaN [85]. The momentum relaxation time is lowest at high energies, and gradually increases as energy decreases. Figure 4.1.3 shows that, with a confining field of 500 kVcm^{-1}, the electron energy very slowly increases between applied electric fields of 5-60 kVcm^{-1}.

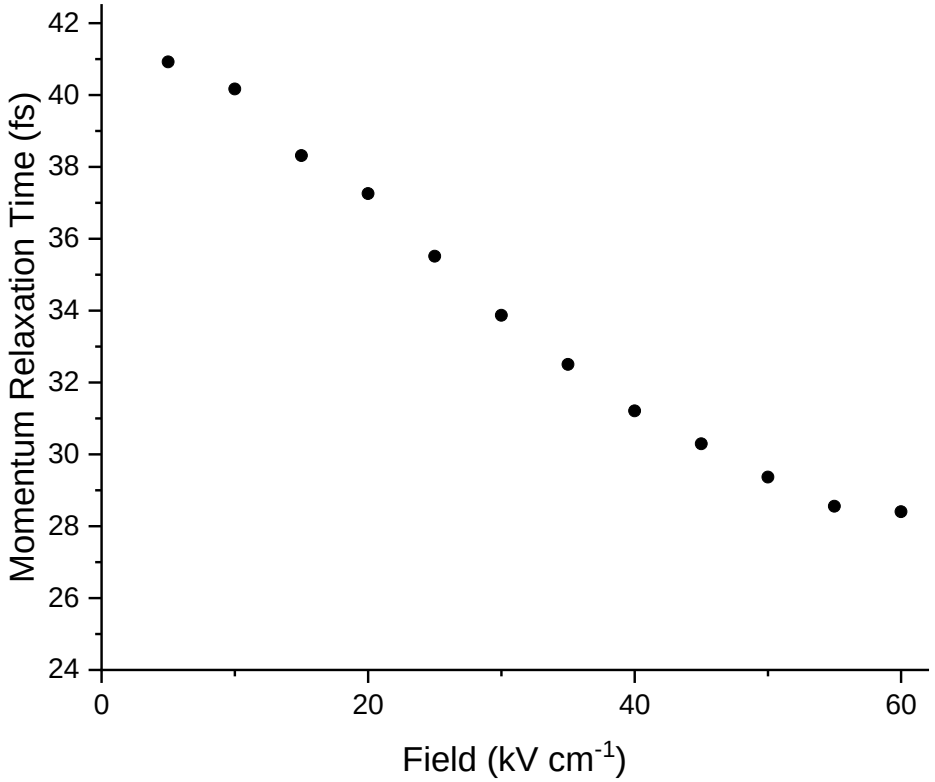


Figure 4.2.1: Momentum relaxation times vs field, with a confining field of 500 kVcm^{-1}.

This would mean a steady decrease in the momentum relaxation time, as is seen in figure 4.2.1. The average electron energy in this field range is between 0.2-0.4 eV, which in *Bulutay's* results yields relaxation times between 10 and 30 fs. These results are therefore consistent with bulk. *Suntrup et al.* [86] measured the momentum relaxation rate of hot electrons using a GaN/AlGaN hot electron transistor (HET). For devices with an injection energy of approximately 1 eV, a momentum relaxation rate of 16 fs was calculated, which is consistent with the results presented here.

4.2.2 Energy Relaxation

Energy relaxation times for each field step are calculated from equation 4.2.1 and shown in figure 4.2.2, again with a confining field of 500 kVcm^{-1}. The energy relaxation times rapidly increase as the applied electric field increases. As the electric field increases further, the energy relaxation times continue to increase but at a slower rate. As the field passes 50 kVcm^{-1}, the energy relaxation times begin to rapidly increase again which is

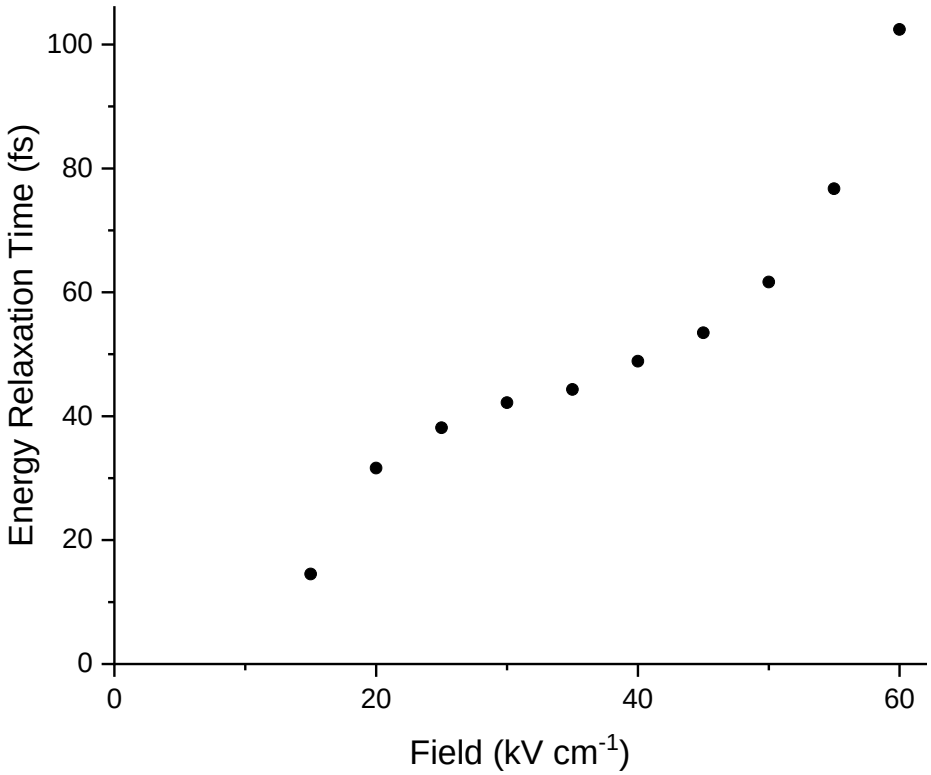


Figure 4.2.2: Energy relaxation times vs field, with a confining field of 500 kVcm^{-1}, subset showing the low field range.

likely due to the energy runaway. *Bulutay* again plots relaxation time versus electron energy for bulk GaN [85]. At low energies, the energy relaxation time increases rapidly with an increase in electron energy. As the electron energy increases further, the increase in relaxation time slows. The range of the energy relaxation times in *Bulutay's* results, in the energy range 0.2-0.4 eV, is between 20-80 fs. Figure 4.2.2 shows a similar trend in the field range of 15-50 kVcm^{-1}. The initial increase in relaxation time between 15-25 kVcm^{-1} is much more rapid than between 25-45 kVcm^{-1}. At 15 kVcm^{-1}, the energy relaxation time is 15 fs, and at 50 kVcm^{-1} is 60 fs. Which are in a similar range to the *Bulutay* relaxation times of 20-80 fs. Again, results are consistent with reported bulk results.

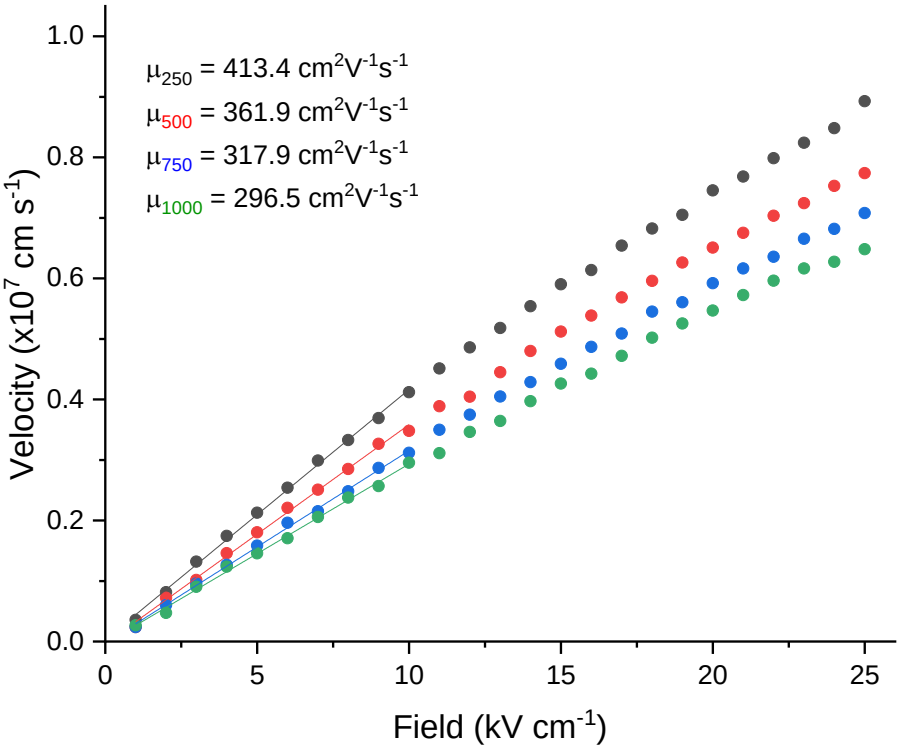


Figure 4.3.1: Linear fit analysis, performed within the OriginPro software, of the low-field steady state velocity vs field results, generating the electron mobility for each confining field strength. Colour and number in subscript corresponds to the colour of the data plot and confining field strength in kVcm^{-1}.

4.3 Low Field Mobility

Another characteristic in semiconductor devices that is well researched, both experimentally and theoretically, is the electron mobility. Electron mobility, μ, is defined by

$$v_d = \mu F \qquad 4.3.1$$

Where v_d is the electron drift velocity and F is the applied electric field. At low applied electric fields, the drift velocity is directly proportional to the field, hence, using equation 4.3.1 the electron mobility can be simply calculated as the gradient at the low-field region of the steady state velocity field results shown in figure 4.1.1. The simulation is performed with smaller field steps of 1 kVcm^{-1} and linear analysis is performed on the first ten field steps, from 1-10 kVcm^{-1}. The results are shown in figure 4.3.1. The highest mobility

obtained is 413 $cm^2V^{-1}s^{-1}$ for a confining field of 250 $kVcm^{-1}$. *Bajaj et al* [14], who measured the Hall mobility in an AlGaN/GaN HEMT, present a result of 445 $cm^2V^{-1}s^{-1}$ with a measured sheet density of $7.8x10^{11}$ cm^{-2}. Converting the sheet density to confining electric field strength (see Appendix A), this sheet density corresponds to a confining field of 450 $kVcm^{-1}$. The closest confining field used is 500 $kVcm^{-1}$, which gives a mobility of 362 $cm^2V^{-1}s^{-1}$. This is lower than the experimental result, however for a slightly higher confining field. In these results a higher confining field leads to a lower mobility. However, these results are significantly lower than those presented by *Matulionis et al.* [16] and *Palacios et al.* [33], who alongside the velocity results used for comparison in figure 4.1.2 also presented measured Hall mobilities of 1500 and 1670 $cm^2V^{-1}s^{-1}$ respectively. Later in the chapter, the mobility will be revisited in light of the alloy scattering parameters.

4.4 Transient

One of the main advantages of using the ensemble Monte Carlo method is the ability to output transient data, this allows the electron properties evolution over time in the simulation to be seen. This is particularly interesting when investigating how the electron velocity evolves over time. In the bulk case, there is a velocity overshoot at the beginning of the simulation where the electron velocity increases rapidly and significantly exceeds the saturation velocity for a short period of time. The presence of this velocity overshoot is of particular interest for use in devices, for example, it would be extremely beneficial if the electrons only travelled for a short period of time through the device, such that they only ever experience the velocity overshoot, and never reach velocity saturation. Electron properties are output at the end of each time step, this subsection shows the transient velocity and energy characteristics, for a confining field of 500 $kVcm^{-1}$.

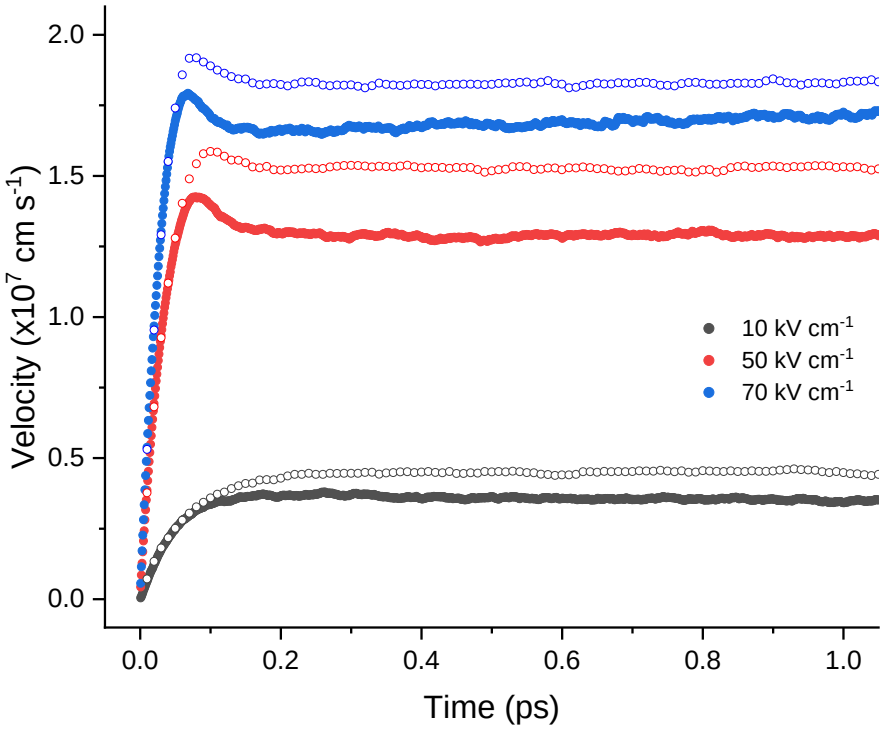


Figure 4.4.1: Transient velocity results for a low applied electric field, mid-electric field just before the energy runaway, and just after the energy runaway in the 2DEG (solid circles) and corresponding data from the bulk EMC (open circles).

4.4.1 Velocity

Figure 4.4.1 shows the velocity evolution over time for three different applied electric fields. A low electric field (10 $kVcm^{-1}$), a mid field just before the energy runaway (50 $kVcm^{-1}$) and one just after the energy runaway (70 $kVcm^{-1}$). At low applied electric fields, the velocity quickly rises towards the saturation velocity and then remains constant at this value. At higher applied electric fields, the velocity rises rapidly to a peak velocity just above the saturation value, showing a very small hint of a novel velocity overshoot. For 70 $kVcm^{-1}$, the velocity reaches a peak of $1.8x10^7$ cms^{-1} before dropping to a saturation value of $1.65x10^7$ cms^{-1}, however, as time continues to increase, the velocity starts to slowly increase due to the energy runaway, reaching a value of $1.73x10^7$ cms^{-1} at 1 ps. Again, the 2DEG velocity results are limited by the alloy scattering and are lower than the bulk results, when the 2DEG velocity is expected to be greater. The effect of the alloy scattering is investigated in section 4.5.

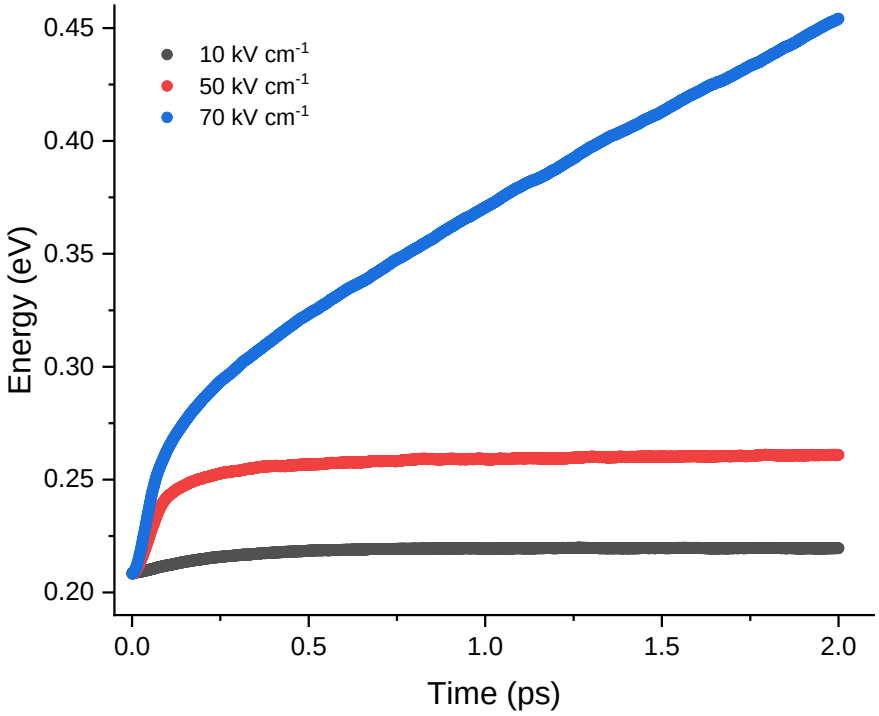


Figure 4.4.2: Transient energy results for a low applied electric field, mid-electric field just before the energy runaway, and just after the energy runaway in the 2DEG.

4.4.2 Energy

The transient energy characteristics were investigated for the same applied electric fields, and are shown in figure 4.4.2. Again at low applied fields, the energy has a small increase before reaching a saturation level. At mid applied fields the energy has a more significant, and more rapid, increase before reaching saturation. At fields beyond 60 kVcm^{-1}, where the energy runaway begins, the energy has the same significant and rapid initial increase, however instead of reaching a saturation value, the energy continues to steadily increase. In the 70 kVcm^{-1} case, the initial rapid increase sees the energy reach a value of 0.27 eV, then the energy continues to increase and after 2 ps has reached a value of 0.45 eV.

4.5 Effect of the Alloy Disorder Potential

The alloy disorder potential used in calculating the alloy scattering rates in this work is taken as 1.5 eV [84], however, this generates a rate that is comparable to the peak polar optical phonon emission scattering rate, making alloy scattering one of the dominant

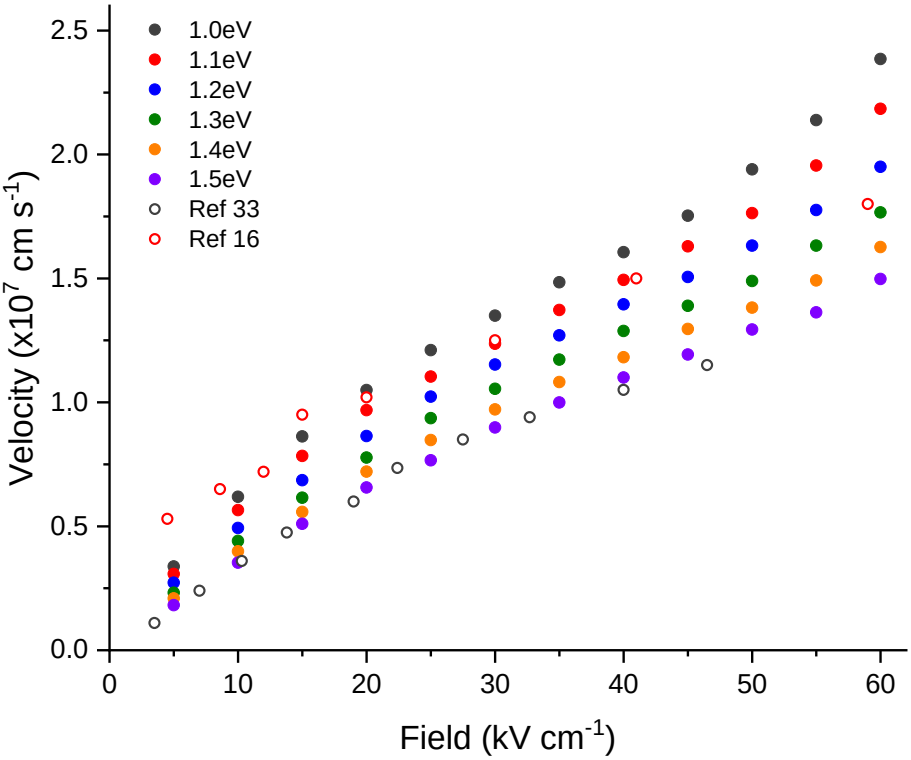


Figure 4.5.1: Steady state velocity results for varying alloy disorder potential, compared to experimental results from Palacios [33] and Matulionis [16].

scattering rates, especially at low energies when polar optical phonon emission is not yet active. Alloy scattering is not expected to be a dominant scattering mechanism, and in the case of an AlGaN/GaN interface, the electron gas is localised in the binary material (GaN) and would only slightly experience effects of alloy scattering from the ternary material (AlGaN). In this case, the alloy scattering is expected to play a negligible role [70]. The expressions for the alloy scattering rates given in equation 2.3.9(a-c), show that the alloy disorder potential, ΔV, is in the numerator. This means a lower alloy disorder potential would result in a lower scattering rate, and given that the potential is squared this effect is enhanced for a small change in its value. Therefore, the effect of small changes in the alloy disorder potential on the results is investigated, specifically, the steady state velocity and low field mobility.

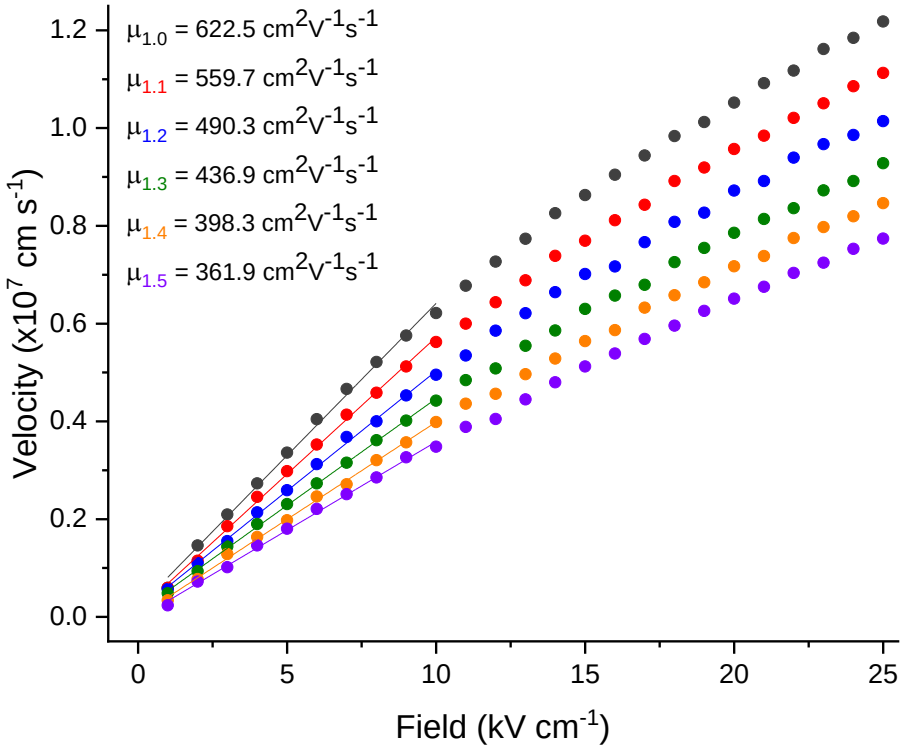


Figure 4.5.2: Linear fit analysis, performed within the OriginPro software, of the low-field steady state velocity results, generating the electron mobility for each alloy disorder potential. Colour and number in subscript corresponds to the colour of the data plot and alloy disorder potential in eV.

4.5.1 Steady State Velocity

The simulation generates low velocities and mobilities compared to experimental data (seen in figure 4.1.2 and section 4.3). This could be due to the high alloy scattering rate causing more scattering events, limiting the mobility. The simulation is repeated with alloy disorder potentials ranging from 1.0-1.5 eV, in 0.1 eV steps (for a confining field of 500 kVcm^{-1}), and the velocity results are shown in figure 4.5.1. It is evident that a lower alloy disorder potential, resulting in a lower alloy scattering rate, generates larger steady state velocities. The increase in velocity becomes more prominent as the applied electric field increases. At low applied electric fields the velocities are similar, but by 50 kVcm^{-1} the difference in the velocities is much more significant. This further increase in velocity as the applied field increases causes a steeper velocity-field curve, which leads to higher mobilities, since low field mobility is calculated as the gradient at the low fields.

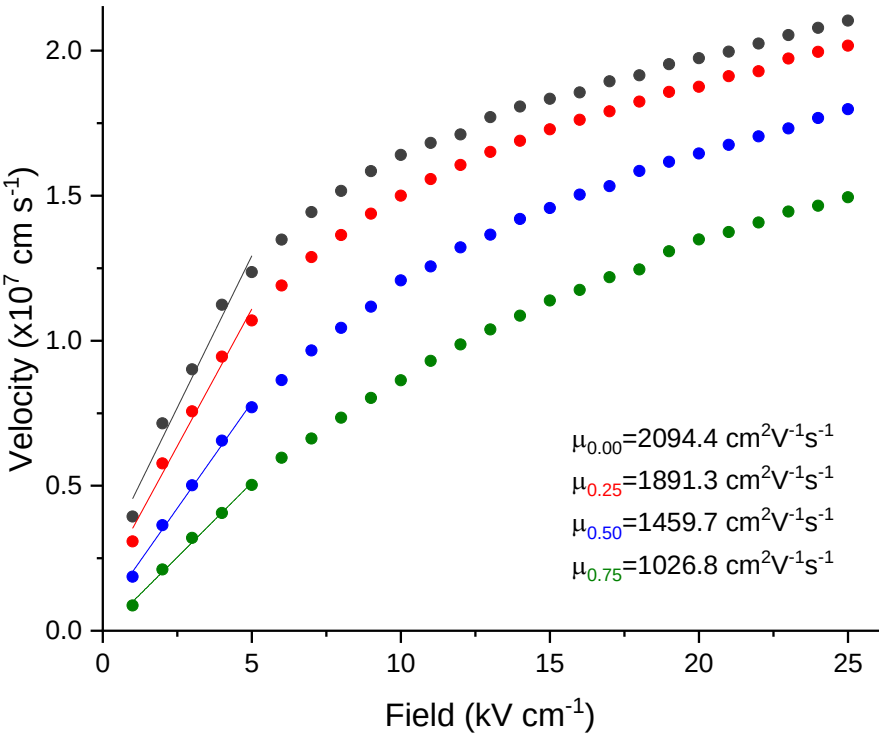


Figure 4.5.3: Linear fit analysis, performed within the OriginPro software, of the low-field steady state velocity results, generating the electron mobility for further reduced alloy disorder potential. Colour and number in subscript corresponds to the colour of the data plot and alloy disorder potential in eV.

4.5.2 Low Field Mobility

The same simulations are run with field steps of 1 kVcm^{-1} to generate the low field velocity, to allow for the calculation of the mobility for varying alloy disorder potentials. The increased velocity shown in figure 4.5.1 that becomes more prominent as the applied electric field increases results in a steeper gradient velocity-field plot, hence larger mobilities. Linear analysis is performed on the first 10 field steps, from 1-10 kVcm^{-1}, and the results are shown in figure 4.5.2. The maximum mobility obtained is 622.5 cm^{2}V^{-1}s^{-1} for an alloy disorder potential of 1.0 eV, a 72% increase on the 1.5 eV result. The increase in mobility is investigated further by running simulations with additional reductions to the alloy disorder potential. Alloy disorder scattering is completely removed by setting the disorder potential to 0 eV to calculate a maximum mobility, along with values of 0.25, 0.50 and 0.75 eV. The results are shown in figure 4.5.3. The enhanced mobility with reduced alloy disorder potential becomes stronger, with a rapid velocity increase at low

fields for lower potentials. Linear analysis is performed on the first 5 field steps, before the velocity results begin to saturate, so the low field mobility equation, equation 4.3.1, remains valid. Removing alloy scattering generates a maximum mobility value of 2094.4 $cm^2V^{-1}s^{-1}$, which exceeds the experimental values of *Matulionis* (1500 $cm^2V^{-1}s^{-1}$) and *Palacios* (1670 $cm^2V^{-1}s^{-1}$). Including alloy scattering with a low alloy disorder potential generates mobility values which are close to these experimental values. An alloy disorder potential of 0.25 eV generates a mobility of 1891.3 $cm^2V^{-1}s^{-1}$, while an alloy disorder potential of 0.50 eV generates 1459.7 $cm^2V^{-1}s^{-1}$.

4.5.3 Experimental Results

It is shown in figure 4.1.2 that the velocity results from this work are comparable to the experimental results of *Palacios et al.* [33]. However, this was for a confining field of 500 $kVcm^{-1}$ which generates a much lower sheet density value ($0.868x10^{12}$ cm^{-2}) than that reported by *Palacios* ($1.46x10^{13}$ cm^{-2}) [33]. Converting a sheet density of $1.46x10^{13}$ cm^{-2} to a confining field strength (see Appendix A) generates a value of 8400 $kVcm^{-1}$. The simulation was repeated with a confining field of 8400 $kVcm^{-1}$ corresponding to an effective well width of 2 nm. The alloy disorder potential is varied until the velocity output is closest to the experimental results. It is found that an alloy disorder potential of 0.9 eV generates velocity results that match closely to the experimental results, as shown in figure 4.5.4.

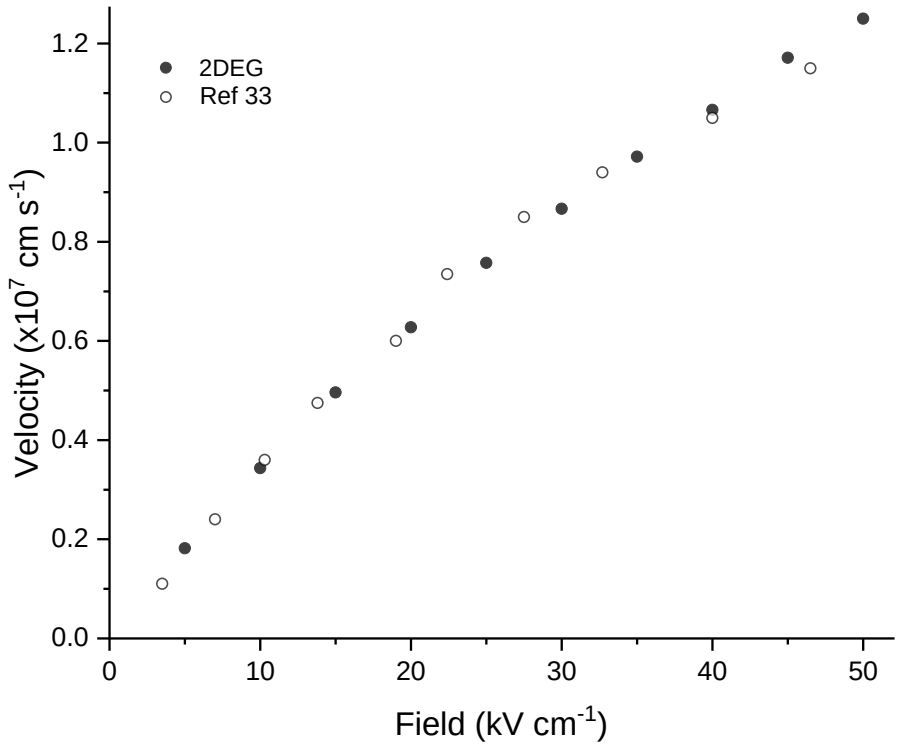


Figure 4.5.4: Steady state velocity results for a confining field strength of 8400 kVcm^{-1} (corresponding to an electron sheet density of 1.46x10^{13} cm^{-2}) and an alloy disorder potential of 0.9 eV, compared to experimental results from Palacios [33].

4.6 Summary

In this chapter, the steady state and transient characteristics of a two-dimensional electron gas created at a Gallium Nitride/Aluminium-Gallium Nitride interface were investigated. The chapter began by comparing the steady state velocity results to published experimental results [16, 33]. It was shown that the average electron velocity agreed well with the published data. The 2DEG results were also compared to results from the bulk EMC, these showed how the average electron velocity continued to increase with applied electric field, unlike the peak and saturation seen in bulk GaN. It was also shown that increasing the applied electric field leads to an energy runaway for fields greater than 60 kVcm^{-1}. The runaway is explained by *Ridley* as an effect caused by 'hot' electrons [65]. For the parabolic band approximation, as used in this simulation, the energy is shown to increase with the square of the applied field. It is shown that at high applied fields the energy increase has a parabolic nature.

The chapter then went on to investigate momentum and energy relaxation times, and compared their values and behaviour with published data. Momentum relaxation time is found to drop off as the applied electric field increases, hence the average electron energy increasing, which agrees with published data for bulk GaN [85]. The values of 28-42 fs also agree with the magnitude of the published data. The energy relaxation times are found to increase with increasing applied electric field (average electron energy), increasing more significantly at low fields, then beginning to slow as the field increases, matching the behaviour of published results for bulk GaN [85]. Next, the low-field electron mobility and how it changes with confining electric field strength (effectively changing the well width and carrier density) was investigated. For a confining field strength of 500 $kVcm^{-1}$, the mobility result of 362 $cm^2V^{-1}s^{-1}$ is close to the value of 445 $cm^2V^{-1}s^{-1}$ found by *Bajaj* [14] for a two-dimensional sheet density of $7.8x10^{11}$ cm^{-2} (which corresponds to a confining field of 450 $kVcm^{-1}$). A lower confining field in the 2DEG simulation leads to a higher mobility, with the highest mobility obtained being 413 $cm^2V^{-1}s^{-1}$. Therefore, for a confining field of 450 $kVcm^{-1}$, the mobility would be in the range of 362-413 $cm^2V^{-1}s^{-1}$.

The chapter then investigated the transient properties, it was found that the average electron velocity increases rapidly initially before saturating. For low applied electric fields, there is no sign of a velocity overshoot. For higher applied electric fields, there were signs of very minor novel overshoots, likely caused when the average electron energy increased beyond the phonon energy and polar optical phonon emission becomes dominant, causing the electrons to lose energy and the velocity to drop slightly. For applied fields greater than 60 $kVcm^{-1}$, after the peak velocity is reached and saturation should occur, the velocity continues to slowly increase as time increases. Similar behaviour was seen in the average electron energy results. For low applied electric fields, there is a small initial increase in energy over time before saturating. For mid fields below

energy runaway, there is a more significant initial increase in energy before saturation. When the energy runaway is present, the initial rapid increase in energy is followed by a steady increase in energy instead of saturation.

The chapter ends with an investigation of the effect of the alloy disorder potential on the velocity and mobility results. It was found that lower alloy disorder potentials, which result in smaller alloy scattering rates, generate higher velocities which match experimental results. The velocity increase becomes more prominent as the field increases, which produces a steeper velocity-field curve and thus larger low field mobilities. The mobility sees a 72% increase from 361.9 $cm^2V^{-1}s^{-1}$ when the alloy disorder potential is 1.5 eV to 622.5 $cm^2V^{-1}s^{-1}$ when the alloy disorder potential is 1.0 eV. For an alloy disorder potential of 1.3 eV the mobility is 436.9 $cm^2V^{-1}s^{-1}$, which is much closer to the value of 445 $cm^2V^{-1}s^{-1}$ reported by *Bajaj et al*. [14]. It was also shown that removing alloy scattering from the simulation produces a maximum mobility value of 2094.4 $cm^2V^{-1}s^{-1}$. Whereas including low alloy disorder potentials of 0.25 and 0.50 eV generated mobilities of 1891.3 and 1459.7 $cm^2V^{-1}s^{-1}$, which are much closer to the *Matulionis* [16] and *Palacios* [33] experimental values of 1500 and 1670 $cm^2V^{-1}s^{-1}$ respectively.

Chapter 5

Non-Equilibrium Phonons in a Gallium Nitride 2DEG

The algorithm described in section 2.4 is used to investigate the effect of non-equilibrium phonons on the transport properties of the two-dimensional electron gas (2DEG) created at a Gallium Nitride/Aluminium Gallium Nitride (GaN/AlGaN) interface. The same triangular well as depicted in figure 4.1 is considered. A uniform electric field across the whole device, in the x-direction (perpendicular to the confinement), is assumed. The sub-band energy levels for each confining field remain the same as in table 4.1. No additional material parameters are introduced, so they remain as given in table 4.2. Non-equilibrium phonon scattering is considered for both intra- and inter-band polar-optical phonon (POP) scattering events. With the addition of non-equilibrium phonons, new simulation parameters are introduced, these are given in table 5.1. The 2DEG phonon lifetime is not well characterised for GaN, the value used is based on that typically measured for HEMTs [13]. The chapter begins with investigating the effect of non-equilibrium phonons on the steady state results, comparing velocity field results with those from chapter 4, along with experimental results. The effects on relaxation times and low-field mobility are then studied and are also compared to experimental and theoretical data. A brief look at the non-equilibrium effects on the transient data follows. The chapter concludes with an investigation into the distribution of the non-equilibrium phonons, and how the confinement in one dimension and constraints due to momentum conservation in two dimensions constrict the non-equilibrium phonons to a small area of the non-equilibrium table (or q-space). The inclusion of a non-equilibrium phonon population interacting with confined phonons in an AlGaN/GaN simulation has not been seen in the current literature.

Parameter (units)	Value
Phonon lifetime (ps)	1
Phonon update time (fs)	25

Table 5.1: Additional simulation parameters used for all non-equilibrium AlGaN/GaN 2DEG simulations.

5.1 Steady State

The steady state ensemble average velocity and energy data is output at the end of the total simulation time (as explained in chapter 2). The effect of non-equilibrium phonons on the steady state energy, and the energy runaway seen in chapter 4, are investigated. The effects of non-equilibrium phonons on the steady state velocity are then studied, firstly by comparing to the equilibrium results from chapter 4, which is followed by a comparison with published experimental results. Investigations are limited to 60 $kVcm^{-1}$, before the energy runaway begins in the equilibrium results, to allow for the focus to be on the effects of the non-equilibrium phonons.

5.1.1 Energy

The steady state ensemble average total electron energy vs applied electric field results are shown in figures 5.1.1 and 5.1.2, and the results are compared to the equilibrium results from figure 4.1.3. For all confining fields, the non-equilibrium energy results are greater than the equilibrium results. The non-equilibrium results follow the same trend as the equilibrium results, with signs of an energy runaway as the applied electric field increases beyond 50 $kVcm^{-1}$. The non-equilibrium energies are slightly higher at low electric fields, while at higher electric fields the difference becomes marginally more pronounced. This increase in energy due to the introduction of non-equilibrium phonons is caused by an increase in the POP absorption rate, which is the dominant scattering mechanism at low energies (before an electron has enough energy to emit a phonon), exceeding acoustic scattering. The phonon relaxation time in bulk is 3 ps [87], while here it is 1 ps, therefore non-equilibrium phonon effects would be expected to be stronger in bulk than in the 2DEG, since the 2DEG phonons relax faster.

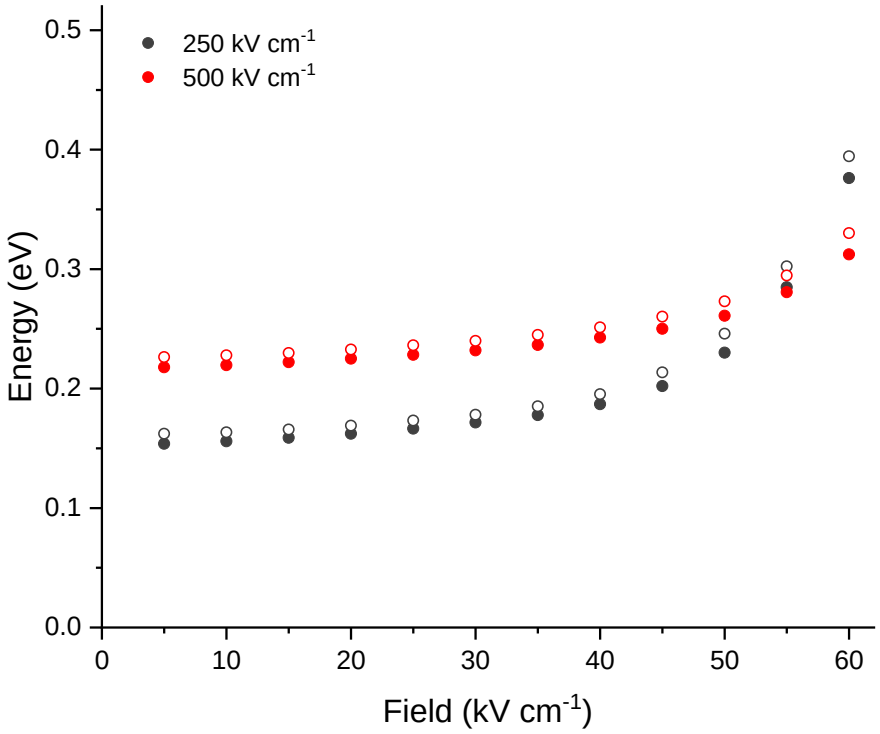


Figure 5.1.1: Average electron energy vs applied electric field for confining field strengths of 250 and 500 kVcm^{-1}. Results from both the equilibrium, figure 4.1.3, (solid circles) and non-equilibrium (open circles) simulations are shown for comparison.

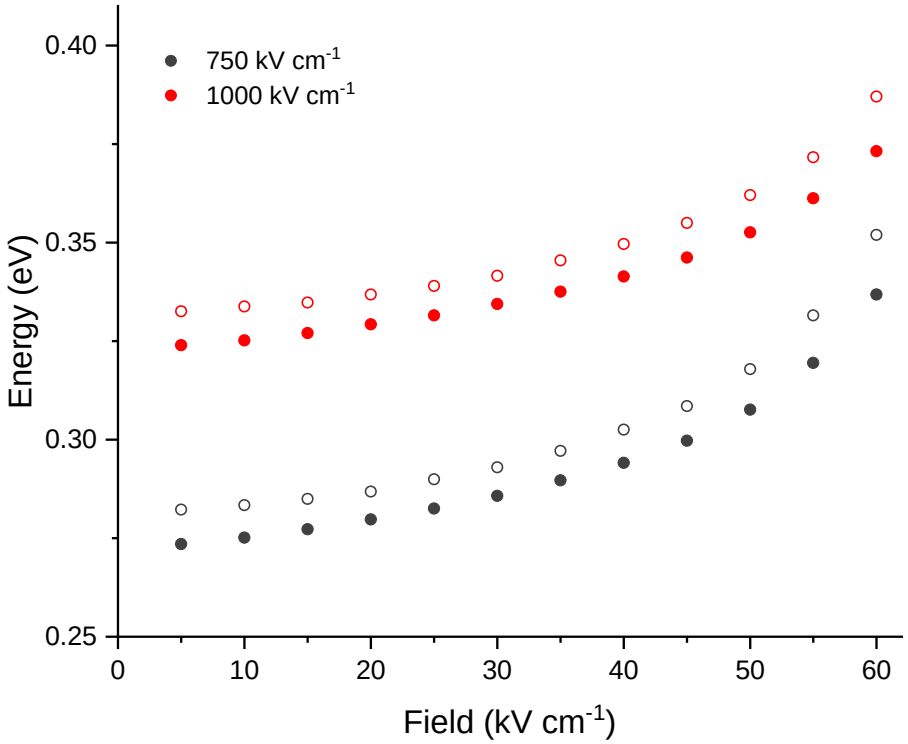


Figure 5.1.2: Average electron energy vs applied electric field for confining field strengths of 750 and 1000 kVcm^{-1}. Showing results for equilibrium, figure 4.1.3, (solid circles) and non-equilibrium (open circles) simulations for comparison.

5.1.2 Velocity

Steady state velocity characteristics are shown in figure 5.1.3, for the range of confining fields 250-1000 $kVcm^{-1}$, and are compared to the equilibrium results from figure 4.1.1. For all confining fields, the velocities from the non-equilibrium results are lower than the equilibrium results. At low applied electric fields the difference is small, but as the applied electric field increases, the non-equilibrium results become much lower than the equilibrium results, this effect of non-equilibrium phonons limiting the electron velocity is also seen in bulk [2, 4, 20, 21]. The lower velocity when non-equilibrium phonons are included in bulk is commonly attributed to the increase in average electron energy causing an earlier transition to the upper valleys with higher effective masses [34]. However, since the 2DEG simulated is considered to be in one valley only, the two sub-bands have the same effective mass and hence this cannot be the case for these results. Another effect of introducing non-equilibrium phonons is an increase in electron-phonon interactions, due to the increase in the POP scattering rates, which leads to an increased randomisation of the electron momentum, known as diffusive heating [34, 88]. Diffusive heating is also regularly associated with the decrease in electron velocity when non-equilibrium phonons are included and is the likely reason for the lower velocities in the above results.

In figure 5.1.4, the velocity results for both the equilibrium and non-equilibrium 2DEG simulations (confining field of 500 $kVcm^{-1}$, alloy disorder potential of 1.5 eV) are compared to experimental results from *Palacios* [33] and *Matulionis* (for differing voltage pulse lengths used, 3 ns and 100 ns) [13, 16]. The confining field has not yet been tuned to generate sheet densities to match the experimental data, this is performed in section 5.5.3. Figure 5.1.4 shows that both the equilibrium and non-equilibrium simulation results agree well with the results of *Palacios* in the very low field range. The velocity limiting caused by the introduction of the non-equilibrium phonons causes the

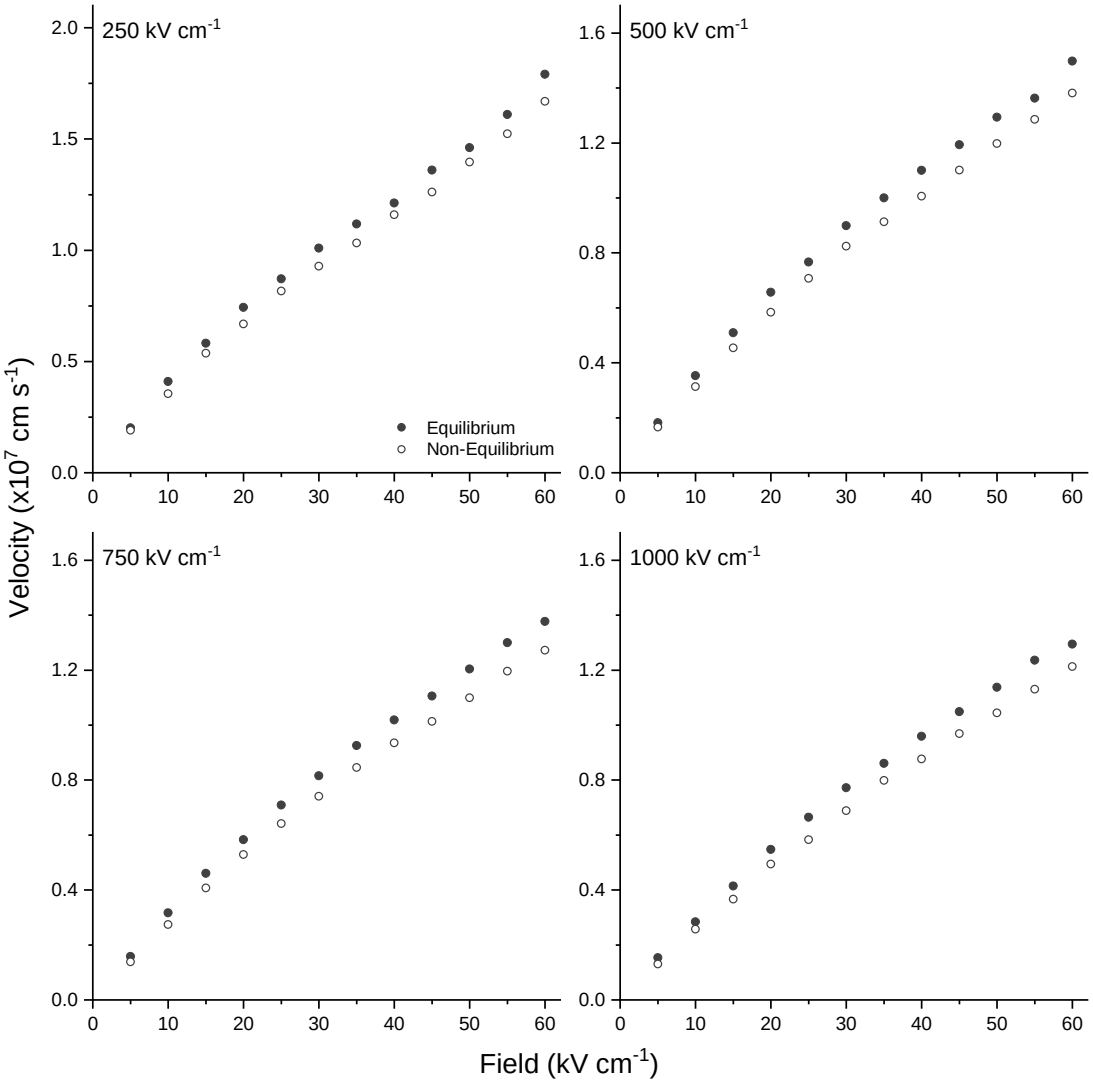


Figure 5.1.3: Average electron velocity vs applied electric field for confining fields ranging from 250-1000 kVcm^{-1}, comparing the equilibrium results, from figure 4.1.1, (solid circles) to non-equilibrium results (open circles).

non-equilibrium results to now match well at higher fields (40 kVcm^{-1} onwards), where the equilibrium results overshot the experimental results. Such close agreement between Monte Carlo simulation results and published experimental results have not been seen in other published Monte Carlo simulations [13, 16, 35], including *Palacios et al.'s* own simulation results where they attempted to generate their experimental results used here [33].

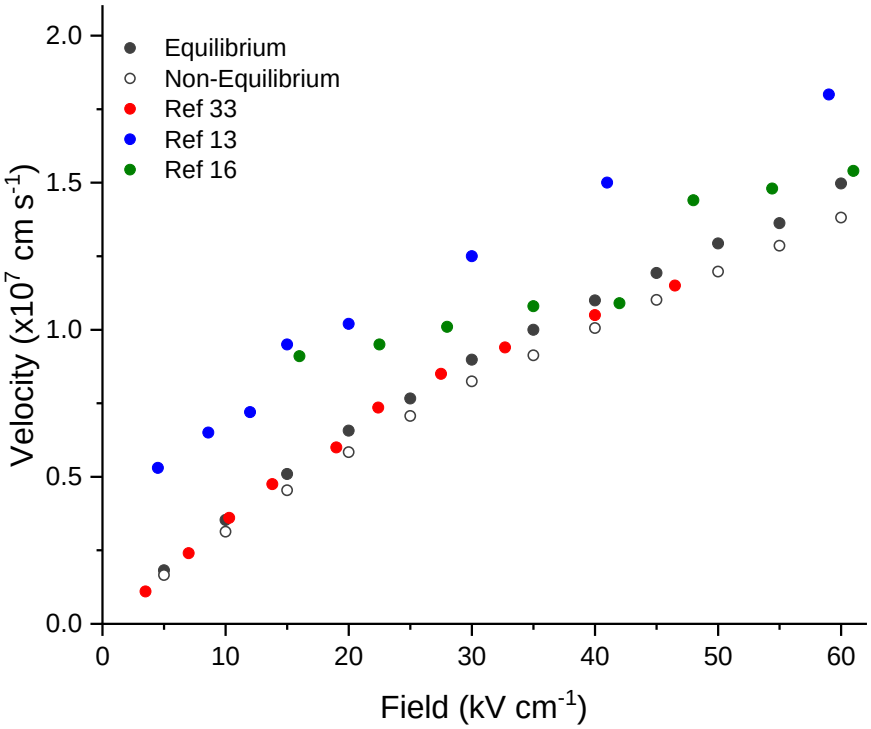


Figure 5.1.4: Average electron velocity vs applied electric field results from the equilibrium and non-equilibrium 2DEG simulations (confining field of 500 kV cm^{-1}, alloy disorder potential of 1.5 eV), compared to simulation results from Palacios [33], and experimental results from Matulionis for a 3 ns [13] and 100 ns voltage pulse [16].

5.2 Relaxation Times

The introduction of non-equilibrium phonons will also have an effect on the momentum and energy relaxation times. The increase in the electron momentum randomisation, which was previously seen to reduce the electron velocity, would be expected to lead to faster relaxation of the electron momentum. Previous analytical investigations [89] have also shown that one of the effects of introducing non-equilibrium phonons is to slow the rate of energy relaxation. The equations for energy, τ_e, and momentum, τ_m, relaxation time calculations are repeated here [85]:

$$\tau_e = \frac{\langle \mathrm{E} \rangle - \mathrm{E}_0}{eF\langle v \rangle}, \qquad 5.2.1$$

$$\tau_m = \frac{\langle p \rangle}{eF} \qquad 5.2.2$$

where $\langle\ \rangle$ represents the ensemble average, E is the electron energy, v is velocity, p is momentum, e is the electronic charge, F is the applied electric field and E_0 is the zero-field average electron energy. Since all electrons in the simulation have the same effective mass, the ensemble average momentum, $\langle p \rangle$, is simply calculated as $m^*\langle v \rangle$.

5.2.1 Momentum Relaxation

The momentum relaxation times when non-equilibrium phonons are included are calculated for each applied field strength using equation 5.2.2, for a confining field of 500 kVcm^{-1}. The results are shown in figure 5.2.1, along with the equilibrium results for comparison. The non-equilibrium results follow the same trend as the equilibrium results, the maximum relaxation time occurs at the lowest applied electric field, the relaxation time then steadily decreases as the applied electric field increases. As discussed in the previous chapter, this behaviour matches results presented by *Bulutay* [85], who plots the momentum relaxation time versus electron energy for bulk GaN. For all applied electric fields, the relaxation times when non-equilibrium phonons are included are lower than the equilibrium results. This is consistent with the theory that an increase in electron-phonon interactions due to the introduction of non-equilibrium phonons causes an increase in electron momentum randomisation, which results in a reduced velocity and faster momentum relaxation. However, this effect is small, as seen in the steady state velocity outputs in figure 5.1.3, and by the minimal difference of just 2-3 fs in the momentum relaxation times.

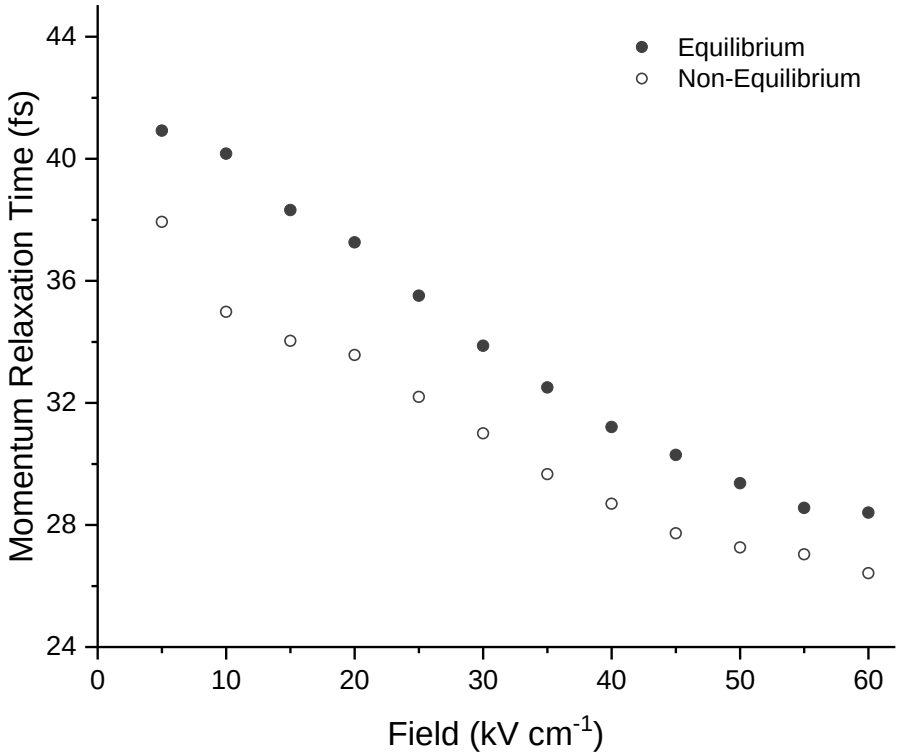


Figure 5.2.1: Momentum relaxation time vs applied electric field for a confining field of 500 kVcm^{-1}, comparing equilibrium (solid circles) and non-equilibrium (open circles) results.

5.2.2 Energy Relaxation

The energy relaxation times when non-equilibrium phonons are introduced are calculated for each applied field strength using equation 5.2.1, again with a confining field of 500 kVcm^{-1}. The results are shown in figure 5.2.2, and are compared to the equilibrium results from chapter 4. For the whole field range the non-equilibrium relaxation times are slightly higher than those from the equilibrium simulation. This is consistent with the analytical investigations that show the introduction of non-equilibrium phonons slow the energy relaxation rate [89] and also consistent with the steady state energy results shown in figure 5.1.1 where the non-equilibrium results are slightly greater.

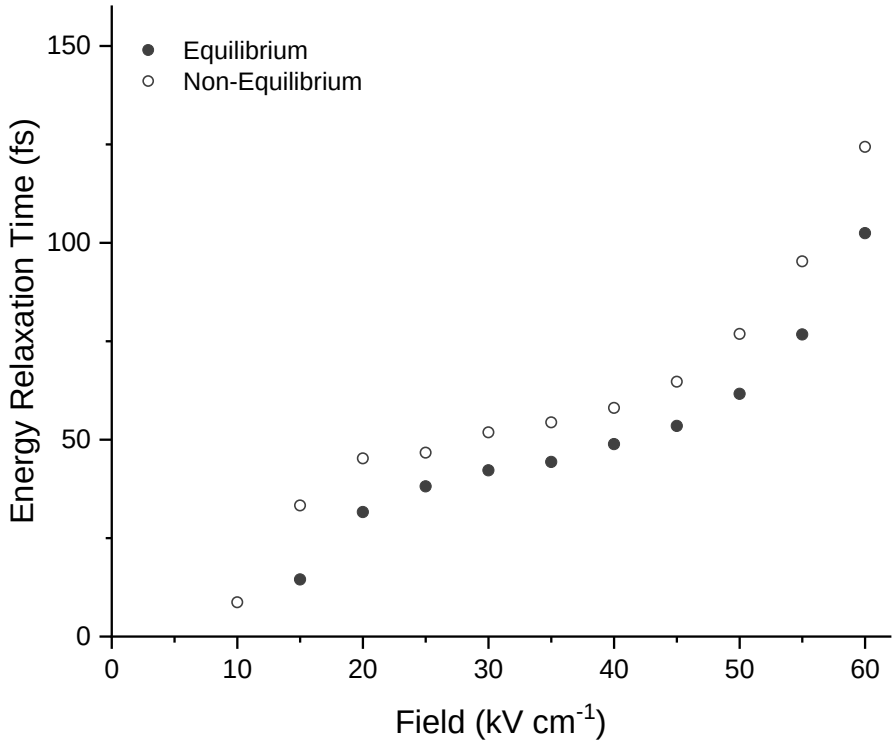


Figure 5.2.2: Energy relaxation time vs applied electric field for a confining field of 500 kVcm^{-1}, comparing equilibrium (solid circles) and non-equilibrium (open circles) results.

5.3 Low Field Mobility

As discussed in the equilibrium results in the previous chapter, electron mobility is an important characteristic in semiconductor devices and has been studied extensively, both experimentally and theoretically. Electron mobility, μ, is defined by

$$v_d = \mu F \qquad 5.3.1$$

Where v_d is the electron drift velocity and F is the applied electric field. The electron drift velocity is proportional to the applied field at low applied fields, allowing the electron mobility to be calculated as the gradient of the low-field region of the steady state velocity results shown in figure 5.1.3. The simulation, with non-equilibrium phonons included, is run with smaller field steps of 1 kVcm^{-1} for a range of confining fields. The results are shown in figure 5.3.1, where linear analysis has been performed on the first ten field steps to calculate the electron mobility for each confining field. The

comparison between the equilibrium and non-equilibrium results are shown in table 5.2. In section 5.1.2, it is shown that the steady state velocity is lower when non-equilibrium phonons are introduced, due to the increased randomisation of the electron momentum. As this effect increases with applied electric field, meaning lower velocities for higher electric fields, the non-equilibrium electron mobility, calculated as the gradient of the velocity-field results, should be smaller.

In table 5.2, it is shown that the non-equilibrium mobility results are noticeably lower for all confining field strengths. For a confining field of 250 $kVcm^{-1}$, the non-equilibrium mobility result is 9.5% lower. For a confining field of 750 $kVcm^{-1}$, the biggest difference occurs, with the result 15.2% lower. For confining fields of 500 and 1000 $kVcm^{-1}$, the non-equilibrium results both lie in this range and have similar differences, at 12.6% and 12.2% respectively. In the previous chapter, the equilibrium results were compared to experimental results by *Bajaj et al* [14], who measured the Hall mobility in an AlGaN/GaN HEMT with a sheet density of $7.8x10^{11}$ cm^{-2}, which can be converted to a 450 $kVcm^{-1}$ confining field (see Appendix A). The closest confining field used is 500 $kVcm^{-1}$, and the equilibrium result (362 $cm^2V^{-1}s^{-1}$) was found to be lower than the experimental value (445 $cm^2V^{-1}s^{-1}$). Introducing non-equilibrium phonons leads to a slightly lower mobility, hence the result of 316.1 $cm^2V^{-1}s^{-1}$ is still significantly lower than the *Bajaj et al.* [14] results. The simulation results have not been tuned to fit any experimental data in any way. The alloy scattering parameters could be used to provide a better fit, as in chapter 4, and is again considered in section 5.5.

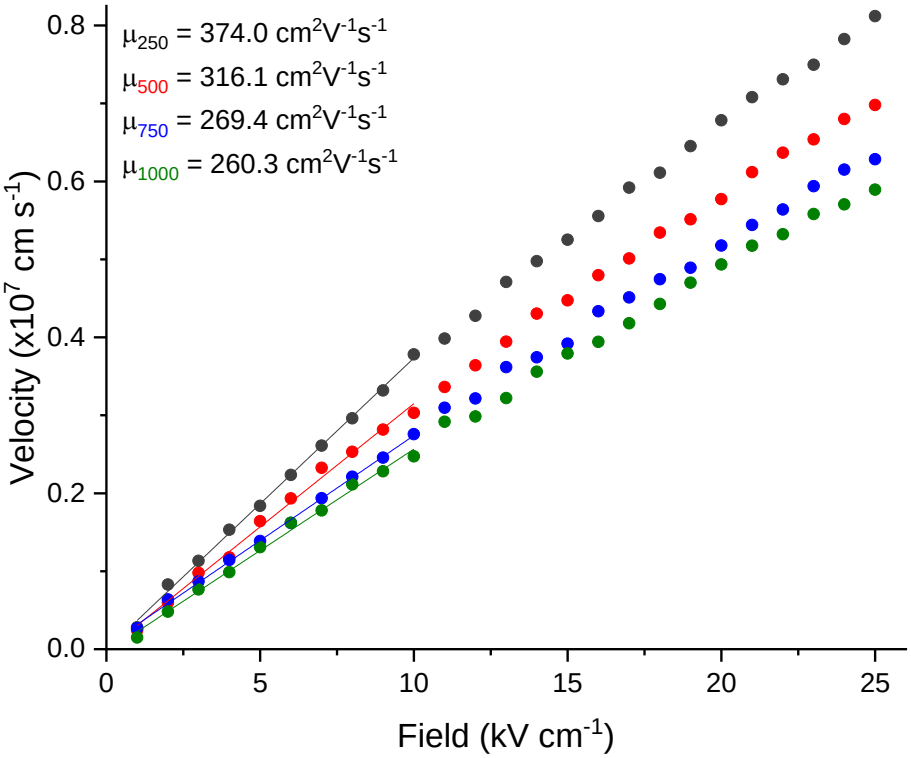


Figure 5.3.1: Linear fit analysis, performed within OriginPro, of the non-equilibrium low-field steady state velocity-field results, generating the electron mobility for each confining field strength. Colour and number in subscript corresponds to the colour of the data plot and confining field strength in $kVcm^{-1}$.

Confining Field ($kVcm^{-1}$)	Equivalent Sheet Density ($x10^{12}\ cm^{-2}$)	Equilibrium Mobility ($cm^2V^{-1}s^{-1}$)	Non-Equilibrium Mobility ($cm^2V^{-1}s^{-1}$)
250	0.434	413.4	374.0
500	0.868	361.9	316.1
750	1.30	317.9	269.4
1000	1.74	296.5	260.3

Table 5.2: Comparison between equilibrium and non-equilibrium results for electron mobility for each confining field strength, and equivalent sheet density (sheet density-electric field conversion shown in Appendix A).

5.4 Transient

With the introduction of non-equilibrium phonons having an effect on the steady state characteristics, specifically the increased randomisation of electron momentum causing the velocity to be lower, this is also likely to have an effect on how the electron properties change over time. Electron properties are output at the end of each time step and here the

transient velocity and energy characteristics when non-equilibrium phonons are introduced are investigated. The results are compared to the equilibrium results for the same applied fields as in chapter 4 (10, 50, and 70 $kVcm^{-1}$), for a confining field of 500 $kVcm^{-1}$, and are also compared to the equilibrium bulk results.

5.4.1 Velocity

Figure 5.4.1 shows how the velocity evolves over time for the three applied electric fields and for each of the simulations, equilibrium, non-equilibrium (2DEG) and bulk (equilibrium). For all field strengths, the saturation velocity in the non-equilibrium results are lower than the equilibrium results, and this difference increases with an increase in the applied field. This is expected from the steady state velocity results in figure 5.1.3, where the non-equilibrium velocities are lower than the equilibrium results, and this is enhanced as the applied electric field increases. For all applied electric fields, the initial rapid increase in the velocities are similar, and the peak values reached are very similar. This is followed by a slightly prolonged decrease in velocity to the lower saturation values in the non-equilibrium results, creating a slightly larger novel overshoot. For all fields, the peak and saturation velocities for the non-equilibrium results are lower than in bulk.

5.4.2 Energy

The energy evolution over time for the three applied electric fields are shown in figure 5.4.2, comparing the equilibrium and non-equilibrium results. For all applied electric fields, the non-equilibrium results saturate at a higher energy, this effect becomes more noticeable as the applied electric field increases. For 50 and 70 $kVcm^{-1}$, both results begin with the same trend before the non-equilibrium results saturate slightly later and at higher energies. The higher saturation energies are consistent with the results shown in figure 5.1.1, where the non-equilibrium energies are slightly greater than the equilibrium results.

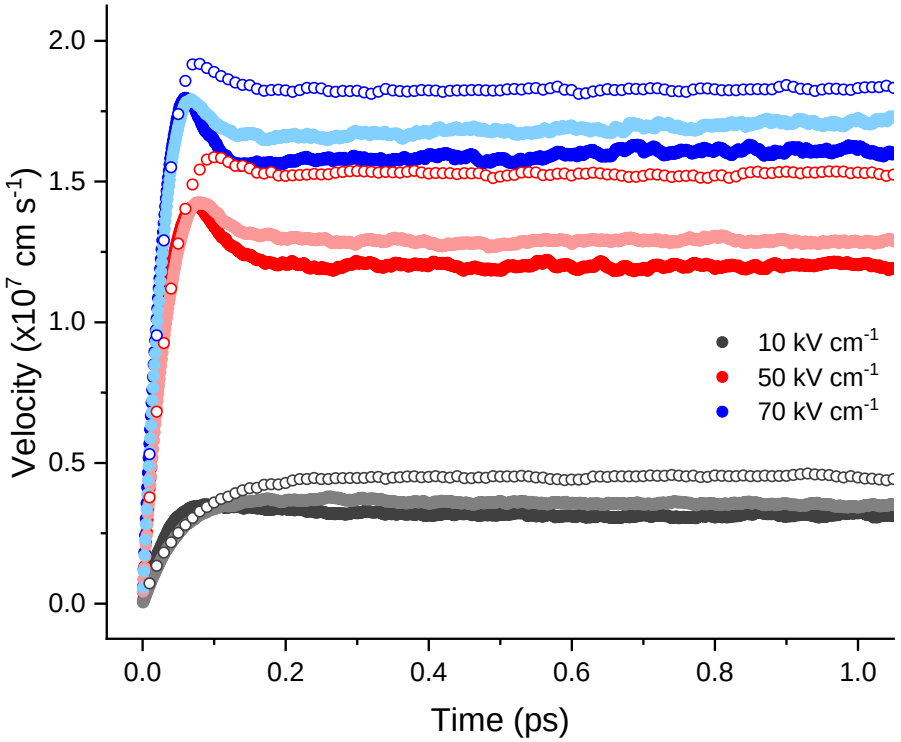


Figure 5.4.1: Transient velocity characteristics for a low applied electric field, mid-electric field before the energy runaway and just after the energy runaway (in equilibrium). Comparing the equilibrium 2DEG (pale solid circles), non-equilibrium 2DEG (solid circles) and equilibrium bulk EMC (open circles) results.

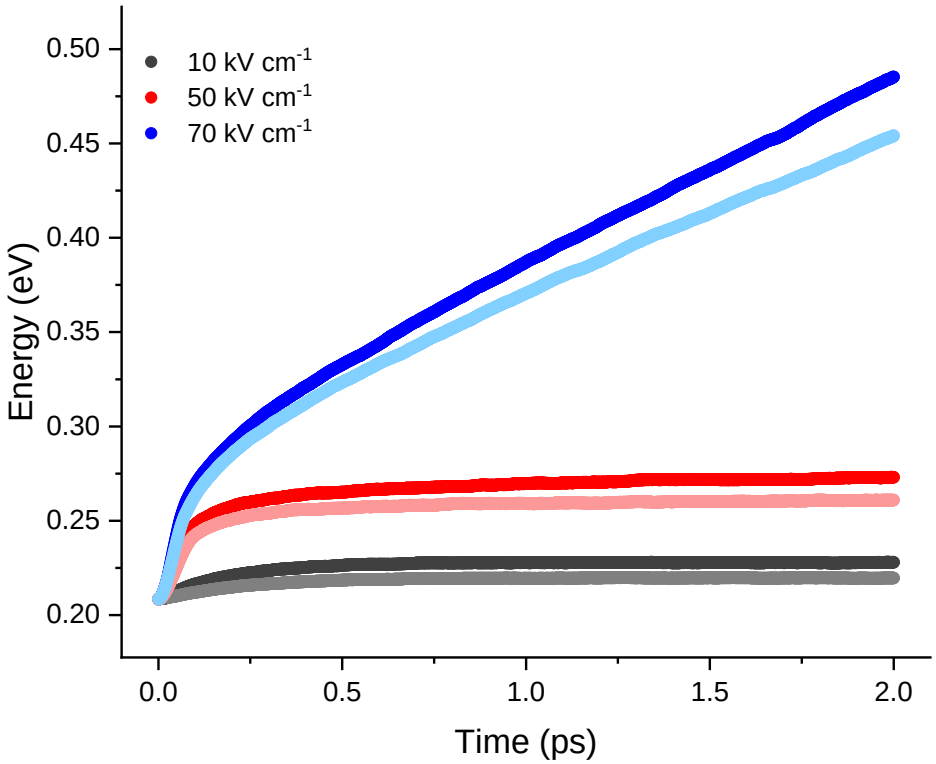


Figure 5.4.2: Transient energy characteristics for a low applied electric field, mid-electric field before the energy runaway and just after the energy runaway (in equilibrium). Comparing the non-equilibrium (solid circles) and equilibrium (pale solid circles) results.

5.5 Effect of the Alloy Disorder Potential

As in the previous chapter, the alloy disorder potential is lowered, and completely removed, to lessen the alloy scattering rates and the effect on the steady state velocity and low field mobility is investigated. The alloy disorder potential is tuned due to the different compositions and materials used in the growth of the experimental structures. *Palacios* [33] used a 0.7 μm iron doped buffer layer, 1.8 μm unintentionally doped GaN and 29 nm $Al_{0.35}Ga_{0.65}N$ barrier, with a 0.6 nm AlN interlayer between the GaN buffer and AlGaN barrier. *Matulionis* [53] used a 1 μm magnesium doped GaN buffer layer, an undoped 25 nm $Al_{0.15}Ga_{0.85}N$ layer and was protected by a 33 nm Si_3N_4 layer. Both structures were grown on sapphire substrates.

5.5.1 Steady State Velocity

The equilibrium simulation produced velocities that were low when compared to experimental results, and with the introduction of non-equilibrium phonons lowering the velocity this difference is enhanced (see figure 5.1.3, comparing velocity results to experimental results). The non-equilibrium simulation is repeated with alloy disorder potentials ranging from 1.0-1.5 eV, in 0.1 eV steps (for a confining field strength of 500 $kVcm^{-1}$) and the results are shown in figure 5.5.1. The same effect as in section 4.5.1 is seen, with a lower alloy disorder potential, hence a lower alloy scattering rate, producing a small increase in the velocity results at low applied electric fields and the increase in velocity becomes more prominent as the field increases.

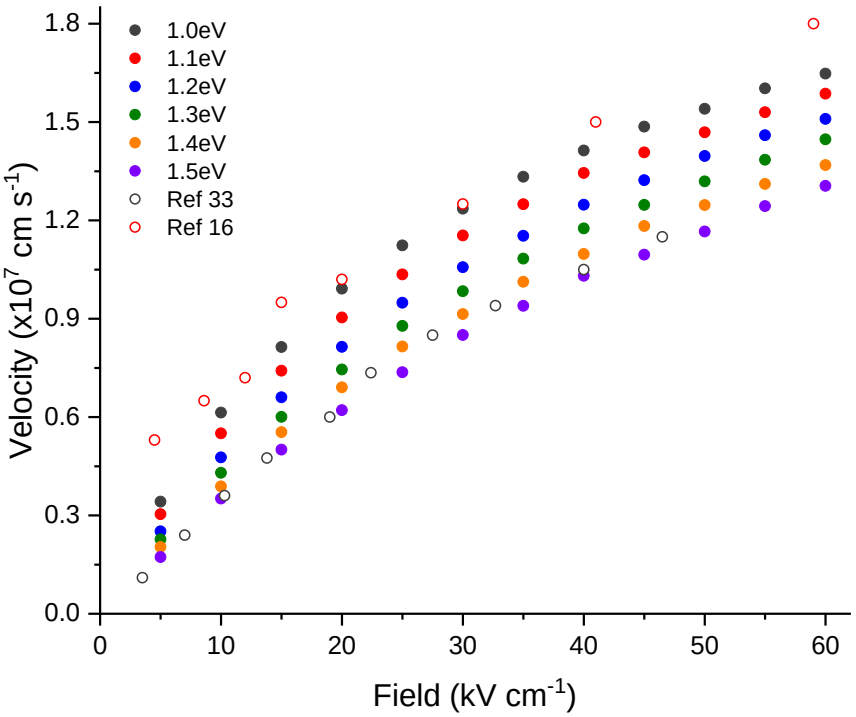


Figure 5.5.1: Steady state velocity results for varying alloy disorder potential, compared to experimental results from Palacios [33] and Matulionis [16].

5.5.2 Low Field Mobility

The field step is changed to 1 kVcm^{-1} to generate low field velocity results, to allow for the calculation of the low field mobility for varying alloy disorder potentials. Linear analysis is performed on the first 10 field steps and the results are shown in figure 5.5.2. As expected from the velocity results in section 5.5.1, a lower alloy disorder potential produces a greater mobility result due to the increased velocities. The increased mobility results, however, are lower than the equivalent results from the equilibrium case in section 4.5.2. For an alloy disorder potential of 1.0eV, the equilibrium result is 622.5 cm^{2}V^{-1}s^{-1} whereas with non-equilibrium phonons introduced the result is 505.1 cm^{2}V^{-1}s^{-1}. This is expected, since the steady state velocity results have shown that the increase in velocity is not as pronounced in the non-equilibrium case, due to the increase in electron-phonon interactions lowering the velocity. The increase in mobility is investigated further with further reductions in the alloy disorder potential. Alloy scattering is completely removed

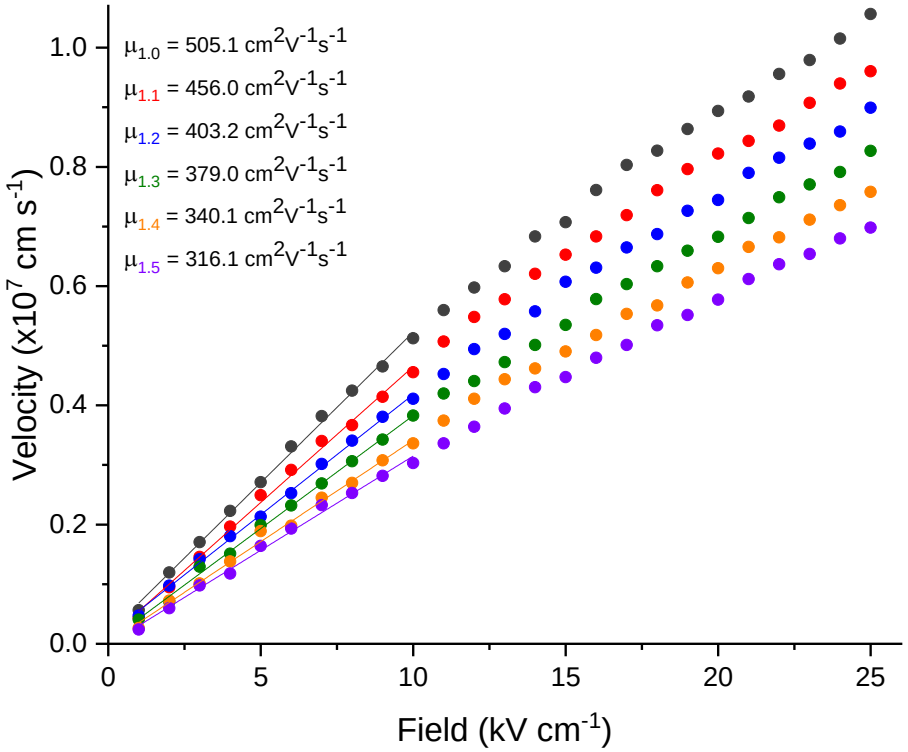


Figure 5.5.2: Linear fit analysis, performed within the OriginPro software, of the low-field steady state velocity results, generating the electron mobility for each alloy disorder potential. Colour and number in subscript corresponds to the colour of the data plot and alloy disorder potential in eV.

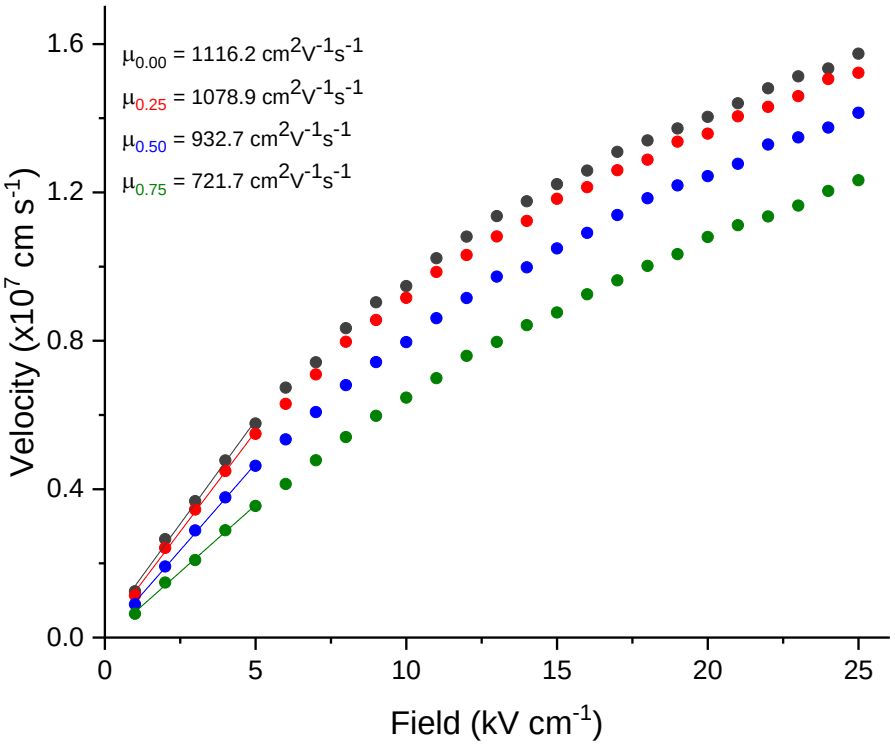


Figure 5.5.3: Linear fit analysis, performed within the OriginPro software, of the low-field steady state velocity results, generating the electron mobility for further reduced alloy disorder potential. Colour and number in subscript corresponds to the colour of the data plot and alloy disorder potential in eV.

from the simulation by setting the alloy disorder potential to 0 eV to generate a maximum mobility result. The simulation is repeated with alloy disorder potentials of 0.25, 0.50 and 0.75 eV and the results are shown in figure 5.5.3. With lower alloy disorder potentials, the mobility increase is enhanced, giving a rapid increase in velocity at low fields. Linear analysis is performed on the first 5 field steps, before the velocity begins to saturate, to ensure the low field mobility equation, equation 5.3.1, remains valid. Similar to the results in figure 5.5.2, the increased mobility results in the non-equilibrium simulation are lower than the equilibrium results. The maximum mobility obtained, for an alloy disorder potential of 0 eV, in section 4.5.2 is 2094.4 $cm^2V^{-1}s^{-1}$, compared to the maximum non-equilibrium result of 1116.2 $cm^2V^{-1}s^{-1}$.

5.5.3 Experimental Results

The velocity results in figure 5.1.4 are compared to experimental results from *Palacios* et al. [33], however, the confining field strength of 500 $kVcm^{-1}$, which corresponds to an electron sheet density of $0.868x10^{12}$ cm^{-2}, is significantly lower than the reported value of $1.46x10^{13}$ cm^{-2}. Converting a sheet density of $1.46x10^{13}$ cm^{-2} to a confining field strength (see Appendix A) generates a value of 8400 $kVcm^{-1}$, which corresponds to an effective well width of 2 nm. The non-equilibrium simulation is run with a confining field strength of 8400 $kVcm^{-1}$, and the alloy disorder potential is varied to find the value which produces the greatest match to the experimental velocity results. It is found that an alloy disorder potential of 0.9 eV (the equilibrium result from section 4.5.3) produces velocity results that are much lower than the experimental results. The alloy disorder potential must be lowered further to best match the *Palacios et al.* velocities [33]. Further reductions to the alloy disorder potential were tested until the velocities were found to best match the experimental data. It is found that an alloy disorder potential of 0.6 eV gives the best match up to an applied electric field of 35 $kVcm^{-1}$. After this the velocity

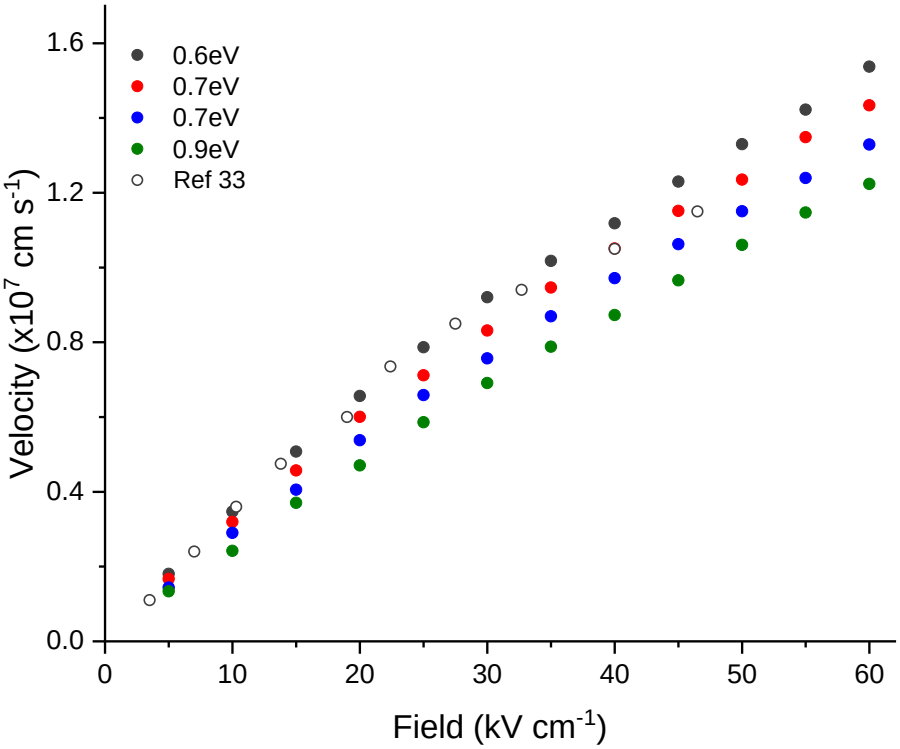


Figure 5.5.4: Steady state velocity results for a confining field strength of 8400 kVcm^{-1} (corresponding to an electron sheet density of 1.46x10^{13} cm^{-2}) and alloy disorder potentials of 0.6, 0.7, 0.8 and 0.9 eV, compared to experimental results from Palacios [33].

does not saturate as significantly as the experimental results, meaning the fit over the whole field range is not as accurate as when a lower sheet density is assumed, i.e. the results shown in figure 5.1.3 for a confining field of 500 kVcm^{-1}. The alloy disorder scattering varies with aluminium content in the barrier [90], and has a range in the literature from 1.0-1.5eV. The lower alloy disorder potential required here to match experiment is likely due to the absence of interface roughness scattering. The need for a lower alloy disorder potential to match the experiment results is consistent with results in sections 5.5.1 and 5.5.2, where the non-equilibrium velocity results are lower than the equilibrium results due to the increased electron-phonon interactions, so it is expected that a lower alloy disorder potential would be needed to reproduce the same velocity results. Although a reduction in the alloy scattering parameters produces a better match, there are scattering mechanisms not included in the simulation, such as interface roughness scattering, which may have an effect on the electron velocity and thus the

accuracy in comparison to experimental results. The very narrow well width may result in a single sub-band being present, which would result in a slightly increased POP scattering rate and hence slightly reduced velocity.

5.6 Phonon Behaviour

The inclusion of the phonon occupancy table, to allow for the introduction of non-equilibrium phonons into the simulation, allows the distribution of the phonons to be monitored throughout the simulation. The phonon distributions can be output at several time intervals to investigate the growth of the phonon distribution over time. The average phonon occupancy can also be calculated and output over time.

5.6.1 Phonon Distributions

At the early stages of the simulation, the phonon distribution will be around thermal, as not many phonons will have been emitted. Phonon occupation, N_q, can be calculated for a given temperature using:

$$|N_q| = \frac{1}{e^{\frac{\mathrm{E}_q}{k_B T}} - 1} \qquad 5.6.1$$

where E_q is the phonon energy, k_B is the Boltzmann constant and T is the temperature. Using the phonon energy of 0.0912 eV from table 4.1, and a temperature of 300 K, the phonon occupation at thermal for GaN is calculated as 0.03. Figure 5.6.1 shows the average phonon occupation over time for a range of applied electric fields, from 10 to 60 kVcm^{-1}, for a confining field of 500 kVcm^{-1}. At the start of the simulation, for each applied field, the average occupation is around the thermal value. For low applied electric fields, 10 and 25 kVcm^{-1}, the average occupation remains around thermal throughout the whole simulation time, increasing slightly. As the applied electric field increases, the average phonon occupation increases further. This is expected, as an increase in applied

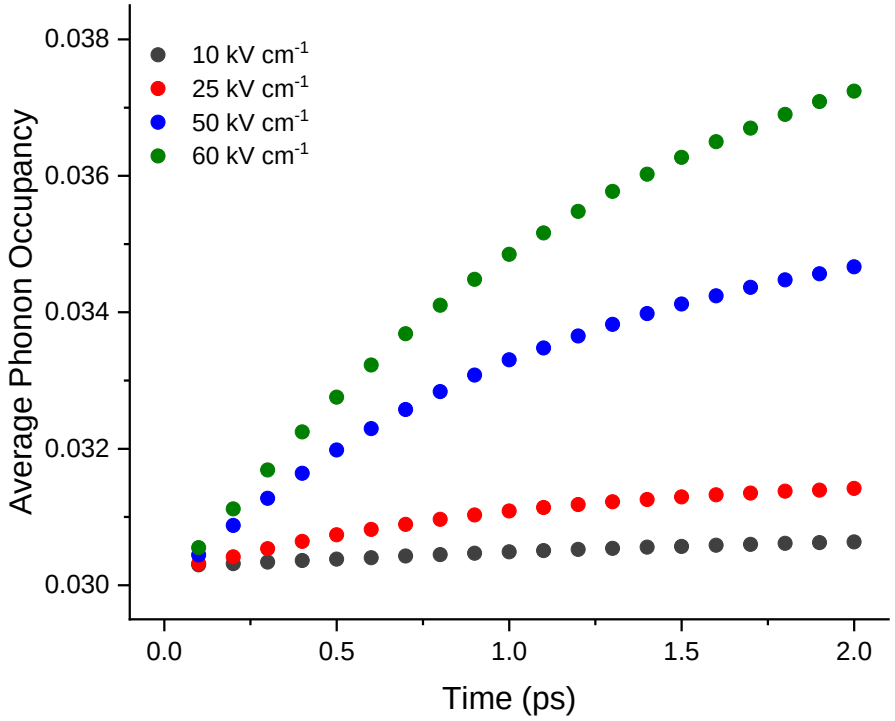


Figure 5.6.1: Average phonon occupancy over time for a range of applied electric fields.

electric field leads to an increase in electron energy, which results in more phonon interactions. Figure 5.6.2 shows the phonon distributions over the full q-space range for applied electric fields of 25 and 50 kVcm^{-1} at the beginning (0.1 ps), middle (1 ps) and end (2 ps) of the simulation run time. It is evident that the phonons are confined to a small area in q-space, this is due to momentum conservation rules in the plane. Also, for intra-band POP scattering via emission, the electron must have enough in-plane energy to emit a phonon (i.e. the in-plane energy must be equal to or greater than the phonon energy, 0.0912 eV), and this minimum energy relates to a minimum phonon wavevector.

Figure 5.6.3 shows the same phonon distributions from figure 5.6.2, with a reduced scale on both axis to more clearly show the distribution in the small area of q-space that is affected. The minimum phonon wavevector associated with the minimum energy required for POP emission scattering is shown in figure 5.6.3 as $q_{\hbar\omega}$. Also in figure 5.6.3, a minimum wavevector is shown based on the ensemble average in-plane energy for the given snapshot. The average in-plane energy for each snapshot (0.1 ps, 1 ps and 2 ps) is

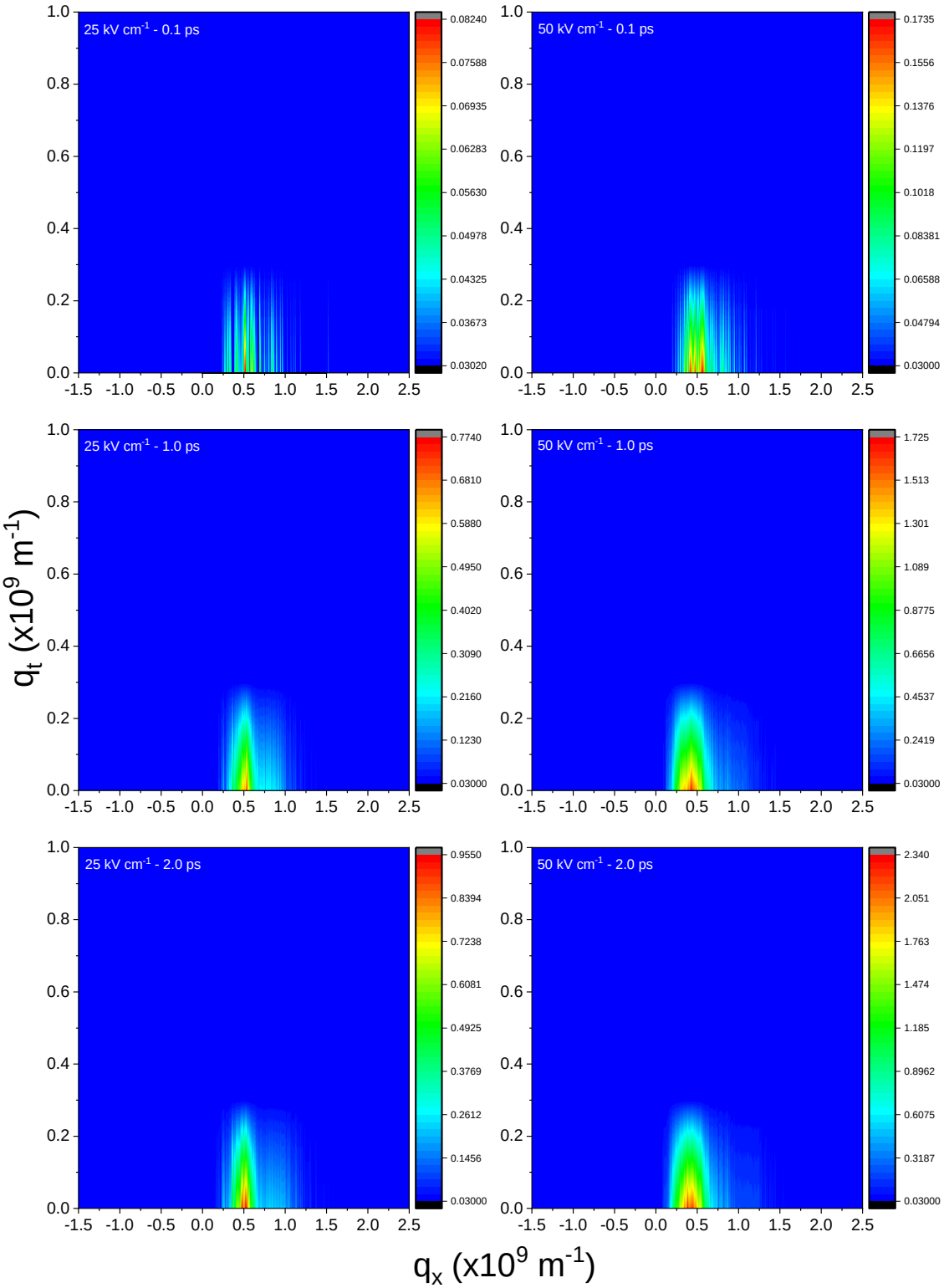


Figure 5.6.2: Phonon distributions as a function of phonon wavevector for applied fields of 25 and 50 kVcm^{-1} after 0.1, 1 and 2 ps. q_x is the component of the phonon wavevector in the direction of the applied electric field (x-direction) and q_t is the y-z component. Confining field strength of 500 kVcm^{-1}.

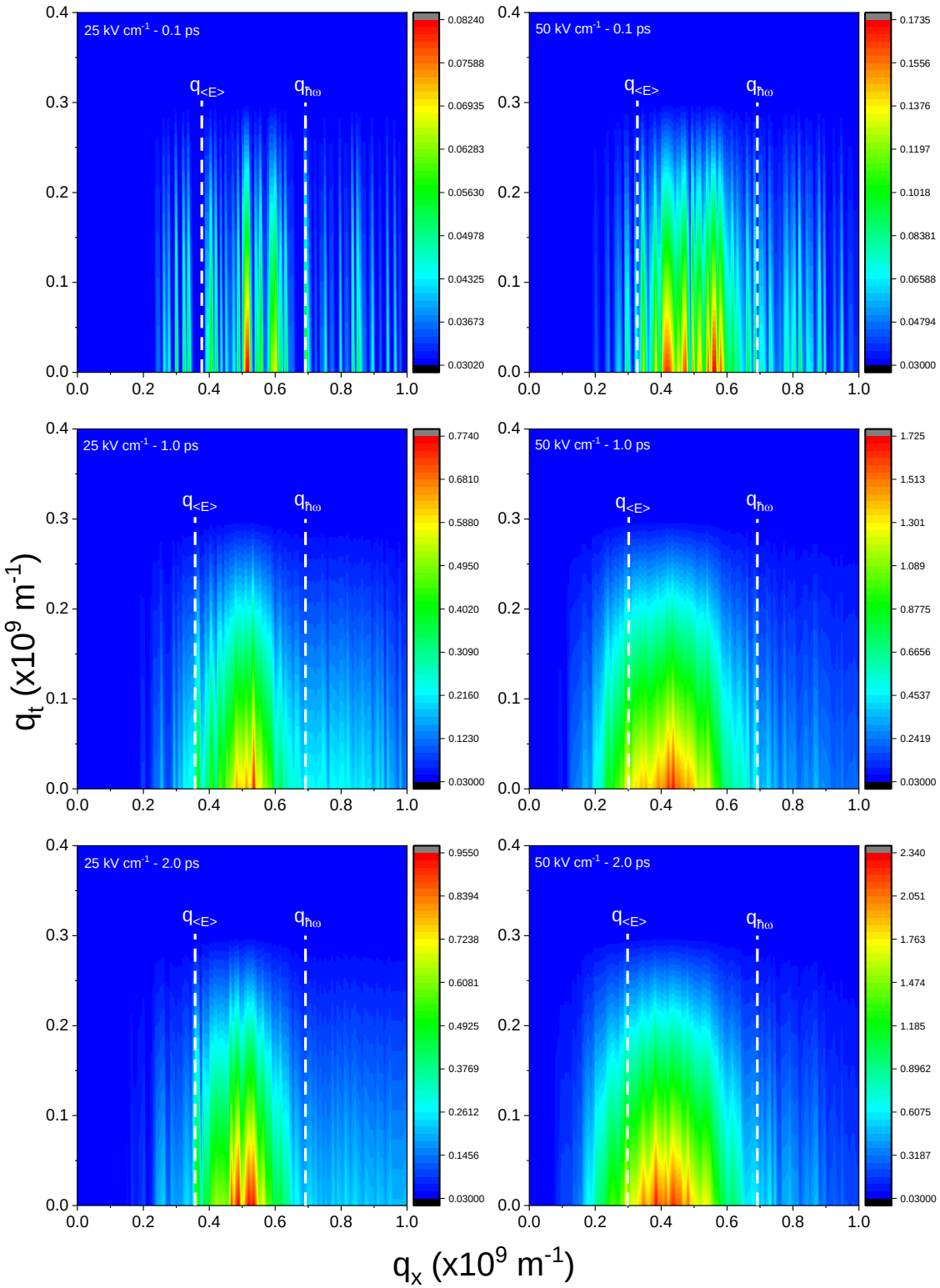


Figure 5.6.3: Phonon distributions as a function of phonon wavevector for an applied field of 25 and 50 kVcm^{-1} after 0.1, 1 and 2 ps, for a confining field strength of 500 kVcm^{-1}. $q_{\hbar\omega}$ represents the minimum wavevector calculated from the minimum in-plane energy required to emit a phonon, $q_{\langle E\rangle}$ represents the minimum wavevector calculated based on the average in-plane energy for the given snapshot.

taken from the transient outputs from section 5.4.2. The phonon energy is then added to this average in-plane energy to calculate the energy an electron must have to emit a phonon and end up with the average energy. The minimum phonon wavevector associated with this energy, $q_{<E>,}$ is then calculated. For all phonon distributions shown, the hotspot lies between these two wavevectors, along the q_x axis. For an applied electric field of 25 $kVcm^{-1}$, the hotspot is around $0.5x10^9$ m^{-1} for all time steps. For an applied electric field of 50 $kVcm^{-1}$, when more electron-phonon interactions occur and the average distribution increases more significantly (as shown in figure 5.6.1), the hotspot starts around $0.5x10^9$ m^{-1} but spreads out to between 0.3 and $0.5x10^9$ m^{-1} in the final time step.

Figure 5.6.4 shows the minimum phonon wavevector against in-plane energy and the calculated minimum wavevectors at the phonon energy, $q_{\hbar\omega}$, and ensemble average electron energy (at 0.1 ps for 50 $kVcm^{-1}$), $q_{<E>}$. This shows that the minimum wavevector decreases as electron energy increases, explaining the relative positions of the wavevectors in figure 5.6.3. Although $q_{<E>}$ is calculated based on the average energy, in reality the distribution of electron energies will spread below this energy and above it. For electrons with energies greater than the ensemble average, the minimum wavevector is less than $q_{<E>}$ and for electrons with energy lower than the average the minimum wavevector is greater than $q_{<E>}$. This explains why the distributions in figure 5.6.3 spread below $q_{<E>}$, and don't lie purely between the two wavevectors, $q_{\hbar\omega}$ and $q_{<E>}$. In two-dimensional systems, electron-phonon scattering events favour forward scattering, while in one-dimensional systems there is only forward scattering [91]. With a very narrow well, the interactions may effectively act as in the one-dimensional case. This preference for forward scattering is seen in figures 5.6.2 and 5.6.3, where the distribution changes in the positive q_x direction, and close to the q_x axis (i.e. $q_t = 0$ and the phonon wavevector is purely in the x-direction).

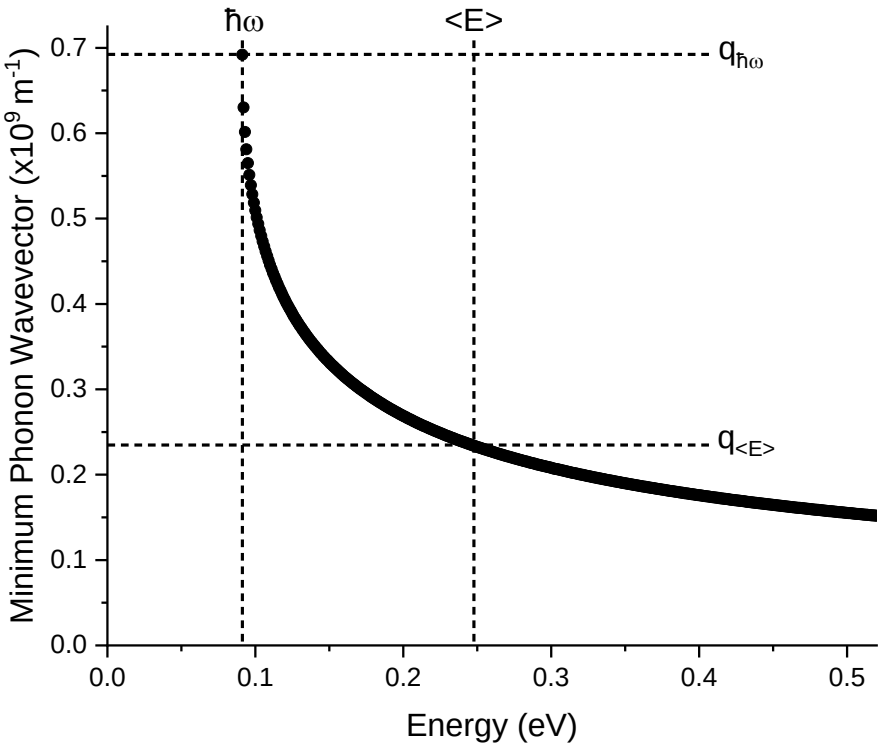


Figure 5.6.4: Minimum phonon wavevector against electron energy. Showing the phonon energy, ħω, the ensemble average energy (taken at 0.1 ps for an applied electric field of 50 kVcm^{-1}), <E>, and the associated minimum phonon wavevectors for each energy.

The phonon occupancy increases over time as more phonons are emitted and reabsorption takes place. This characteristic is enhanced for higher applied electric fields (maximum phonon occupancy of 0.955 for 25 kVcm^{-1} and 2.34 for 50 kVcm^{-1}). Figure 5.6.5 shows the phonon distribution at the end of the simulation time (2 ps) for an applied electric field of 10 kVcm^{-1}. *Ramonas et al.* [35], who also include 3D non-equilibrium phonons in a 2DEG simulation, show a phonon distribution for a field of 10 kVcm^{-1} (for intra-band scattering assuming the q_z component of the phonon wavevector = 0). The phonon distribution reaches a maximum value of 0.14 (which rearranging equation 5.6.1 equates to a temperature of 513 K), whereas figure 5.6.5 shows a maximum of 0.244 (which equates to a temperature of 660 K). However, the distribution in figure 5.6.5 includes all scattering processes (including inter-band) which may explain the slight increase in values. Another difference between the simulations is that *Ramonas* [35] treat acoustic scattering as an inelastic process. Acoustic scattering at room temperature is regularly

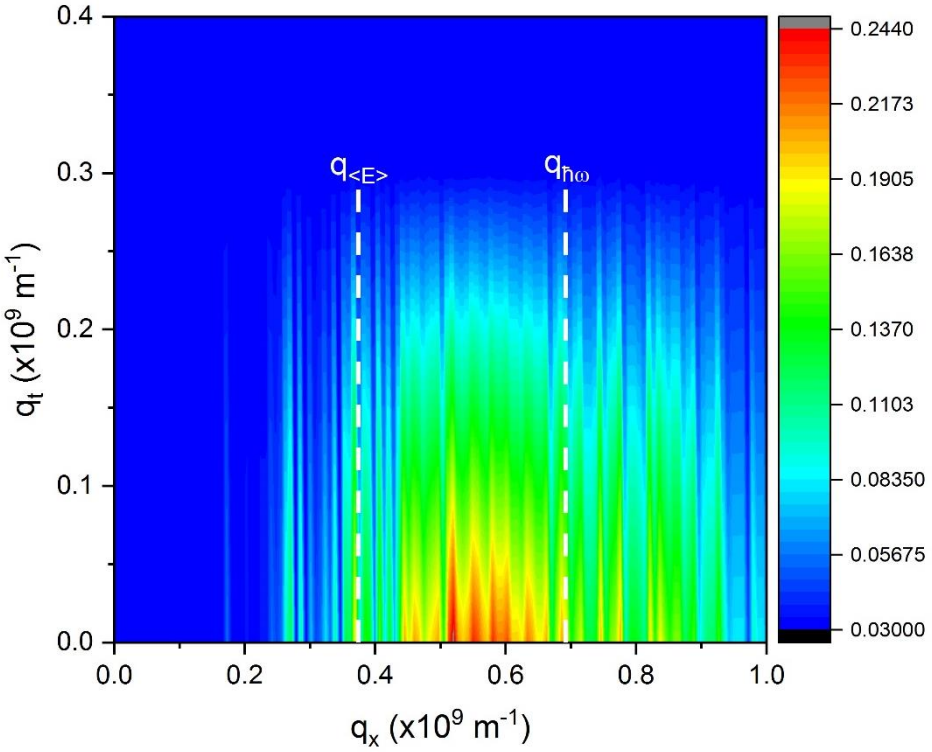


Figure 5.6.5: Phonon distribution as a function of phonon wavevector for an applied field of 10 kVcm^{-1} after 2 ps, for a confining field strength of 500 kVcm^{-1}. $q_{\hbar\omega}$ represents the minimum wavevector calculated from the minimum in-plane energy required to emit a phonon, $q_{\langle E\rangle}$ represents the minimum wavevector calculated based on the average in-plane energy for the given snapshot.

assumed to be an elastic process because the acoustic phonon energies are small. However, this means that electrons are not able to dissipate energy until they are accelerated to energies above the phonon energy. At low fields, where the average electron energy is low, this can lead to inaccurate calculations of energy relaxation times. *Ramonas'* simulation is focused on low electric fields and hence they include inelastic acoustic scattering. Treating acoustic scattering as an inelastic process is another possible way for an electron to dissipate energy and would lessen the effect of an energy runaway, observed in the steady state results in section 5.1.1.

5.6.2 Polar Optical Phonon Scattering Rates

From the expressions for the POP scattering rates given in section 2.3.3.3, it is shown that the scattering rates are dependent on the phonon occupation. As the phonon distribution

evolves over time, the scattering rates must be updated accordingly, as described in section 2.4. Due to the small area of q-space where the phonon distribution increases, shown and explained in section 5.6.1, the average phonon occupation does not increase significantly at low applied electric fields, and hence the scattering rates would not differ greatly. However, as the applied electric field increases causing the phonon distribution to increase further in the confined area, the average phonon occupation does appear to increase and a difference in the scattering rates over time can be seen.

Figure 5.6.6 shows the intra-band POP scattering via absorption rates, in the lowest sub-band, at the beginning (0.1 ps), middle (1 ps) and end (2 ps) of the simulation run time, for applied electric fields of 10, 25, 50 and 60 $kVcm^{-1}$. Intra-band absorption is chosen as it most clearly shows the increase in scattering rates. It is evident that for an applied electric field of 10 $kVcm^{-1}$, the scattering rate remains consistent for all energy values throughout the simulation. For an applied electric field of 25 $kVcm^{-1}$, there are very small increases in the scattering rate at low energies over time. For higher applied electric fields, 50 and 60 $kVcm^{-1}$, the increase in scattering rate over time is more noticeable. The increase is more prominent at low energies, but is still significant at high energies.

There are minor bumps that arise from the numerical integration used in the recalculation of the phonon occupation crossing bin boundaries. The phonon bin size was varied and the final number was chosen to minimise the numerical effects, which are very small near the start of the simulation but continue to grow as the simulation continues and become noticeable towards the end of the simulation time. This section shows the increase in POP scattering rates over time and how this is enhanced for larger applied electric fields, which would lead to an increase in electron-phonon interactions. This supports the theory of diffusive heating having an increased effect on the steady state electron velocity and energy as the applied electric field increases, as discussed in section 5.1.

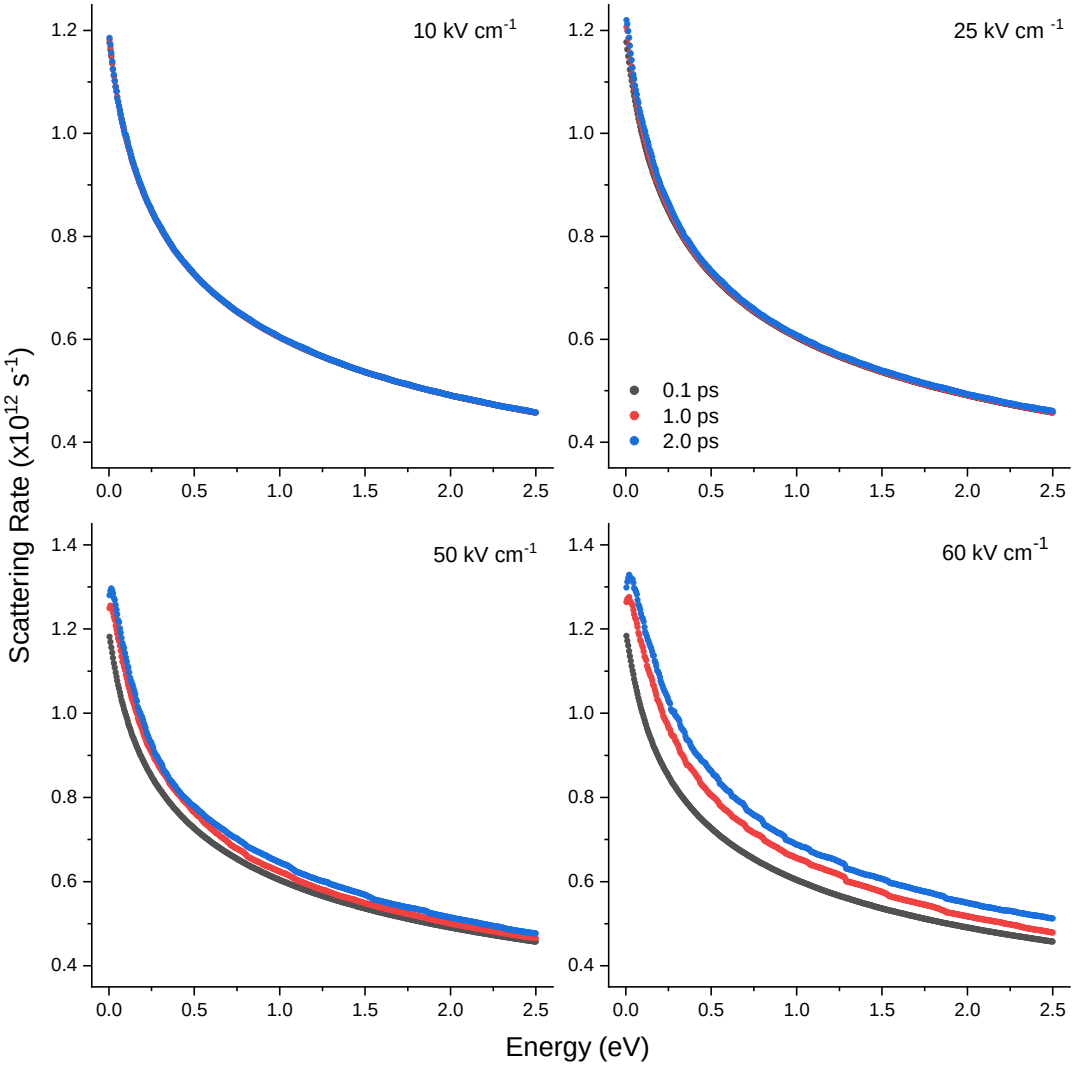


Figure 5.6.6: Intra-band POP via absorption scattering rates as a function of electron energy for an applied field of 10, 25, 50 and 60 kVcm^{-1} after 0.1, 1 and 2 ps.

5.7 Summary

In this chapter, the effect of non-equilibrium phonons on the transport characteristics of a two-dimensional electron gas created at a Gallium Nitride/Aluminium-Gallium Nitride interface were presented. The chapter began by comparing the steady state velocity results from the non-equilibrium simulation to those of the equilibrium simulation and with published experimental results [13, 16, 33]. It was shown that introducing non-equilibrium phonons causes a reduction in the electron velocity. This was caused by diffusive heating (the increase in the randomisation of electron momentum due to an increase in electron-phonon interactions) and was found to become more significant as

the applied electric field increased. Momentum and energy relaxation times from the non-equilibrium 2DEG simulation were then compared to the equilibrium results as well as published data. The momentum relaxation times were found to be lower when non-equilibrium phonons are introduced. This is consistent with the theory that an increase in electron-phonon interactions, leading to an increased randomisation of the electron momentum, produces lower velocities and faster momentum relaxation. The energy relaxation times were slightly increased, consistent with analytic investigations [89] that state introducing non-equilibrium phonons leads to slower energy relaxation rates.

The chapter then investigated the effect on the low field mobility and it was found that the low field mobility values, for a range of confining field strengths, when non-equilibrium phonons are introduced reduced significantly. The lowest decrease was 9.5% for an applied electric field of 250 $kVcm^{-1}$, with the greatest reduction being 15.2% for an applied electric field of 750 $kVcm^{-1}$. This was consistent with the steady state velocity results, where the non-equilibrium results were lower, and this was enhanced as the applied electric field increased. Next, the transient characteristics were investigated. It was shown that the initial velocity evolution over time were similar for both the equilibrium and non-equilibrium results, however, the non-equilibrium results saturated to lower velocities, consistent with the steady state results showing lower velocities when non-equilibrium phonons are introduced. The difference in saturation velocities increased as the applied electric field increased, consistent with the steady state results and with the increase in electron-phonon interactions for larger applied electric fields. The transient velocities for both the equilibrium and non-equilibrium simulations showed signs of a minor novel overshoot, before the electrons had enough energy to begin emitting phonons. This lack of any significant overshoot is consistent with experimental results [18, 19].

Next, the effect of the alloy disorder potential was studied, it was found to have the same effect as in the equilibrium results from the previous chapter. Lowering the alloy disorder potential, and hence the alloy scattering rate, produces an increase in steady state velocities. This effect becomes more prominent as the applied electric field increases, and hence generates a steeper velocity-field curve which led to higher low-field mobility results. However, due to the increased electron-phonon interactions, the non-equilibrium velocities are lower than in the equilibrium case, so these increased mobilities were smaller than the equilibrium results. Alloy scattering was totally removed from the simulation, by setting the alloy disorder potential to 0 eV, to calculate a maximum mobility value for the non-equilibrium simulation. The maximum was found to be 1116.2 $cm^2V^{-1}s^{-1}$, which is significantly lower than the equivalent result in the equilibrium simulation of 2094.4 $cm^2V^{-1}s^{-1}$. The value of 1116.2 $cm^2V^{-1}s^{-1}$ is well below the experimental values of *Matulionis* (1500 $cm^2V^{-1}s^{-1}$) and *Palacios* (1670 $cm^2V^{-1}s^{-1}$), however these results were for different sheet densities. It was found that, for a confining field strength of 8400 kV, to generate an electron sheet density of $1.46x10^{13}$ cm^{-2} to match that reported by *Palacios* [33], an alloy disorder potential of 0.6 eV is needed to produce steady state velocities that best match the published results, much lower than the 0.9 eV needed in the equilibrium case.

The chapter ended by investigating the phonon behaviour. The evolution of the phonon distribution over time was shown to be confined to a small area of q-space due to momentum conservation within the plane. The hotspot was found to lie between 0.3-$0.5x10^9$ m^{-1} on the q_x axis ($q_t = 0$). The average phonon distribution was found to undergo minimal increase for low applied electric fields, however this increase became more significant at higher applied electric fields. The peak phonon occupancy equates to an equivalent temperature reaching ~3000 K, which compares favourably with the experimental results of ~3600 K [30]. The effect on the POP scattering rates was shown

to be minimal for low applied electric fields (10 and 25 $kVcm^{-1}$) where the average phonon distribution increases slightly. The increase in scattering rates became more prominent for higher applied electric fields (50 and 60 $kVcm^{-1}$). The increase in scattering rates at higher applied electric fields would lead to diffusive heating having a greater effect, and this explained why the difference in the equilibrium and non-equilibrium steady state velocity results is small for low applied electric fields but increased and became more prominent as the applied electric field increased. The mobility results, varying the alloy scattering parameters, cover the spread of the experimentally determined values. While it is clear that non-equilibrium phonons are important in GaN based devices, the details of the individual device structures, doping, composition, etc. are also important in determining the peak velocities and mobilities.

Chapter 6

Conclusions & Future Work

In this work an Ensemble Monte Carlo (EMC) algorithm, simulating electron transport in bulk III-V semiconductor materials, has been successfully ported from the CPU architecture to the massively parallel GPU architecture. A series of optimisations, including architectural changes, a memory strategy and general physics simulation changes, were performed leading to significant reductions in the simulation run time. Of the optimisations, increasing the time step (usually determined by the frequency of transient output) and changing from double to single precision floating point numbers (at the expense of precision) can be implemented within the CPU algorithm to obtain reductions in the simulation run time. Refactoring of the algorithm for the particular, highly parallel GPU architecture saw further reductions in run times. The most significant reductions were obtained from maximising the time steps, ensuring all electrons encountered at least one scattering event and reducing the number of inactive threads for a given iteration, from replacing mathematical functions with the corresponding CUDA optimised functions and from utilising local memory to have each thread create its own local version of the electron it had been assigned to simulate. The optimised code running on the CPU ran in 56.5 s, while the GPU version ran in 18.0 s, which is approximately 30% of the CPU run time, proving that it is possible to gain significant performance increases in semiconductor EMC algorithms by utilising GPUs.

Also in this work, electron transport within the two-dimensional electron gas (2DEG) created at an Aluminium Gallium Nitride/Gallium Nitride (AlGaN/GaN) heterojunction was explored. New scattering rates and a new scattering routine were introduced to the bulk EMC, based on the triangular well approximation, assuming the two lowest sub-bands. Steady state velocity, momentum and energy relaxation times and electron

mobility results were compared to experimental and other simulation results. The results were found to match remarkably well with experiment (given a significant difference in electron sheet density). Steady state velocity results show no peak or signs of negative differential resistance (NDR), consistent with experimental data [16, 33] and unlike bulk GaN [2, 20]. Transient velocity results show no sign of a significant overshoot, also unlike bulk GaN [4, 21] and again consistent with experimental results.

The effect of introducing non-equilibrium phonons to the 2DEG EMC were also investigated. Non-equilibrium phonons are shown to reduce the steady state velocity due to diffusive heating [86, 88], caused by an increase in electron-phonon interactions. Momentum relaxation times decrease slightly compared to the equilibrium 2DEG results, consistent with the theory of diffusive heating causing an increase in randomisation of the electron momentum and faster relaxation times. Energy relaxation times slightly increase, consistent with previous analytic results [89]. The non-equilibrium phonon effects are found to be small, especially at low applied electric fields. It is shown that the phonon distribution only changes in a small area of q-space, due to the electron confinement, and thus the phonon occupancy does not significantly increase and the polar optical phonon scattering rates show very little change at low electric fields. The phonon occupancy grows more significantly for higher applied electric fields, hence the change in the scattering rate becomes more noticeable as the applied electric field increases, and the increase in electron-phonon interactions is more prominent at high electric fields. This effect was seen in the comparison of the steady state velocity results, which were only slightly reduced at low applied electric fields and were lessened much further at higher applied electric fields.

The work in this thesis shows that electron confinement should be included in simulations to accurately represent GaN based HEMTs and to reproduce experimental results, and also provides a base for possible future work in two-dimensional simulations of device

structures. Simple devices containing an anode, cathode and active region have previously been modelled [20]. Whilst GaN Gunn diodes operating in one spatial dimension have been simulated [92], along with a proof of concept for two-dimensional devices, these device models could be expanded such that more complex devices can be simulated, for example, switching from vertical devices to model planar devices. The device model calculates the electric field based on the current charge density, which could be regularly fed to the 2DEG EMC. A full device model would require a significant amount of computation power, and it would be possible to improve the performance of such simulations by utilising GPGPU for highly parallel sections of the algorithm.

Appendix A

Carrier Sheet Density to Applied Electric Field Conversion

Experimental results regularly quote the electron sheet density of their 2DEG, and the minimisation parameter, b, is often given in terms of the sheet density, n_{sh}, as [72]:

$$b = \left(\frac{77\pi m^* e^2}{4\varepsilon\hbar^2}\, n_{sh}\right)^{\frac{1}{3}} \qquad \text{A.1.1}$$

where m^* is the effective mass, e is the electronic charge, ε is the dielectric constant and $\hbar$ is the reduced Planck constant. In equation 2.3.3, repeated here, the minimisation parameter is given in terms of the applied confining electric field, F_z, as [71]:

$$b = \left(\frac{14\, m^*}{\hbar^2} e\, F_z\right)^{\frac{1}{3}} \qquad \text{A.1.2}$$

where the symbols have the same meaning. Equations A.1.1 and A.1.2 both define the minimisation parameter and thus can be seen as equal. Setting A.1.1 equal to A.1.2 and solving in terms of the confining electric field strength yields

$$F_z = \frac{11\, \pi\, e\, n_{sh}}{8\, \varepsilon} \qquad \text{A.1.3}$$

and hence using equation A.1.3 it is possible to calculate the corresponding field strength for a given sheet density, and vice versa.

References

[1] J. Lee, J. Kim and H. Jeon, Current Applied Physics, vol. 9, 663, 2008.

[2] S. Yamakawa, R. Akis, N. Faralli, M. Saraniti and S. M. Goodnick, Journal of Physics: Condensed Matter, vol. 21, 174206, 2009.

[3] S. Mingiacchi, P. Lugli, A. Bonfiglio, G. Conte, M. Eickhoff, O. Ambacher, A. Rizzi, A. Passaseo, P. Visconti and R. Cingolani, Physica Status, Solidi (a), vol. 190, 281, 2002.

[4] B. E. Foutz, S. K. O'Leary, M. S. Shur, and L. F. Eastman, Journal of Applied Physics, vol. 85, 7727, 1999.

[5] B. Aslan, L. F. Eastman and Q. Diduck, International Journal of High Speed Electronics and Systems, vol. 19, 1, 2009.

[6] R. Gaska, Q. Chen, J. Yang, A. Osinsky, M. Asif Kahn and M. Shur, IEEE Electron Device Letters, vol. 18, 492, 1997.

[7] K. Krishnamurthy, J. Martin, B. Landberg, R. Vetury and M. J. Poulton, IEEE MTT-S International Microwave Symposium Digest, pp. 303-306, 2008.

[8] B. Kim, D. Derickson and C. Sun, Asia-Pacific Microwave Conference (IEEE), pp. 1-4, 2007.

[9] K. Joshin and T. Kikkawa, IEEE Radio and Wireless Symposium, pp. 65-68, 2008.

[10] A. Ashok, D. Vasileska, O. L. Hartin and S. M. Goodnick, IEEE Transactions on Electronic Devices, vol. 57, 562, 2010.

[11] C. Xie and A. Pavio, MILCOM 2007 – IEEE Military Communications Conference, pp. 1-4, 2007.

[12] X. Ding, Y. Zhou, J. Cheng, CES Transactions on Electrical Machines and Systems, vol. 1, 54, 2019.

[13] L. Ardaravicius, A. Matulionis, J. Liberis, O. Kiprijanovic, M. Ramonas, L. F. Eastman, J. R. Shealy and A. Vertiatchikh, Applied Physics Letters, vol. 83, 4038, 2003.

[14] S. Bajaj, O. F. Shoron, P. S. Park, S. Krishnamoorthy, F. Akyol, T-H. Hung, S. Reza, E. M. Chumbes, J. Khurgin and S. Rajan, Applied Physics Letters, vol. 107, 153504, 2015.

[15] O. Kiprijanovic, A. Matulionis, J. Liberis and L. Ardaravicius, Lithuanian Journal of Physics, vol. 45, 447, 2005.

[16] L. Ardaravicius, M. Ramonas and J. Liberis, O. Kiprijanovic, A. Matulionis, J. Xie, M. Wu, J. H. Leach and H. Morkoc, Journal of Applied Physics, vol. 106, 073708, 2009.

[17] A. Matulionis, J. Liberis, L. Ardaravicius, M. Ramonas, I. Matulioniene and J. Smart, Semiconductor Science and Technology, vol. 17, I-9, 2002.

[18] T-H. Yu and K. F. Brennan, Journal of Applied Physics, vol. 91, 3730, 2002.

[19] D. Liu, D. Lin, G. Huang, Z. Li and K. Guo, Superlattices and Microstructures, vol. 112, 57, 2017.

[20] D. R. Naylor, "Development of Monte-Carlo Simulations for III-V Semiconductors employing an analytic band-structure", Ph.D. Thesis, University of Hull, 2012.

[21] R. P. Joshi, S. Viswanadha, P. Shah and R. D. del Rosario, Journal of Applied Physics, vol. 93, 4836, 2003.

[22] D. R. Anderson, "Phonon-Limited Electron Transport in Gallium Nitride and Gallium Nitride-Based Heterostructures", Ph.D. Thesis, University of York, 2002.

[23] S. Ghosh, S. M. Dinara, A. Bag, A. Chakraborty, P. Mukhapadhyay, S. Kabi and D. Biswas, Physics of Semiconductor Devices, pp. 269-272, 2014.

[24] G. Atmaca, P. Narin, E. Kutlu, T. V. Malin, V. G. Mansurov, K. S. Zhuravlev, S. B. Lisesivdin and E. Ozbay, IEEE Transactions on Electron Devices, vol. 65, 950, 2018.

[25] S. Krishnamurthy, M. van Schilfgaarde, A. Sher, A.-B. Chen, Applied Physics Letters, vol. 71, 1999, 1997.

[26] M. Piccardo, L. Martinelli, J. Iveland, N. Young, S. P. DenBaars, S. Nakamura, J. S. Speck, C. Weisbuch, J. Peretti, Phys. Rev. B, vol. 89, 235124, 2014.

[27] S. K. O'Leary, B. E. Foutz, M. S. Shur and L. F. Eastman, Journal of Materials Science: Materials in Electronics, vol. 17, 87, 2006.

[28] E. Ahmadi, S. Keller and U. K. Mishra, Journal of Applied Physics, vol. 120, 115302, 2016.

[29] M. Tapajna, N. Killat, V. Palankovski, D. Gregusova, K. Cico, J-F. Carlin, N. Grandjean, M. Kuball and J. Kumzmik, IEEE Transactions on Electron Devices, vol. 61, 2793, 2014.

[30] T. Brazzini, M. A. Casbon, H. Sun, M. J. Uren, J. Lees, P. J. Tasker, H. Jung, H. Blanck and M. Kuball, Microelectronics Reliability, vol. 55, 2493, 2015.

[31] J. W. Pomeroy, M. J. Uren, B. Lambert and M. Kuball, Microelectronics Reliability, vol. 55, 2505, 2015.

[32] W. S. Tan, M. J. Uren, P. W. Fry, P. A. Houston, R. S. Balmer and T. Martin, Solid-State Electronics, vol. 50, 511, 2006.

[33] T. Palacios, S. Rajan, A. Chakraborty, S. Keikman, S. Keller, S. P. DenBaars and U. K. Mishra, IEEE: Transactions on Electronic Devices, vol. 92, 2117, 2005.

[34] A. Dyson, D. R. Naylor and B. K. Ridley, IEEE Transactions on Electron Devices, vol. 62, 3613, 2015.

[35] M. Ramonas, A. Matulionis and L. Rota, Semiconductor Science and Technology, vol. 18, 118, 2003.

[36] E. Tea, H. Hamzeh and F. Aniel, Journal of Applied Physics, vol. 110, 113108, 2011.

[37] A. Stephen, G. M. Dunn, C. H. Oxley, J. Glover, M. Montes Bajo, D. R. S. Cumming, A. Khalid, M. Kuball, Journal of Applied Physics, vol. 114, 043717, 2013.

[38] O. Ambacher, B. Foutz, J. Smart, J. R. Shealy, N. G. Weimann, K. Chu, M. Murphy, A. J. Sierakowski, W. J. Schaff and L. F. Eastman, Journal of Applied Physics, vol. 87, 334, 2000.

[39] O. Ambacher, J. Majweski, C. Miskys, A. Link, M. Hermann, M. Eickhoff, M. Stutzmann, F. Fernardini, V. Fiorentini and V. Tilak, Journal of Physics: Condensed Matter, vol. 14, 3399, 2002.

[40] M. Gonschorek, J. F. Carlin, E. Feltin, M. A. Py, N. Grandjean, V. Darakchieva, B. Monemar, M. Lorenz and G. Ramm, Journal of Applied Physics, vol. 103, 093714, 2008.

[41] G. Koley and M. G. Spencer, Applied Physics Letters, vol. 86, 042107, 2005.

[42] H. Xiao-Guang, Z. De-Gang and J. De-Sheng, Chinese Physics B, vol. 24, 067301, 2015.

[43] F. Bernardini, V. Fiorentini and D. Vanderbilt, Physical Review B, vol. 56, R10024, 1997.

[44] O. Ambacher, J. Smart, J. R. Shealy, N. G. Weimann, K. Chu, M. Murphy, W. J. Schaff, L. F. Eastman, R. Dimitrov, L. Wittmer, M. Stutzmann, W. Rieger and J. Hilsenbeck, Journal of Applied Physics, vol. 85, 3222, 1999.

[45] A. Bykhovski, B. L. Gelmont and M. S. Shur, Journal of Applied Physics, vol. 81, 6332, 1997.

[46] P. M. Asbeck, E. T. Yu, S. S. Lau, G. J. Sullivan, J. Van Hove and J. M. Redwing, Electronics Letters, vol. 33, 1230, 1997.

[47] E. T. Yu, G. J. Sullivan, P. M. Asbeck, C. D. Wang, D. Qiao and S. S. Lau, Applied Physics Letters, vol. 71, 2794, 1997.

[48] M. B. Nardelli, K. Rapcewicz and H. Bernholc, Applied Physics Letters, vol. 71, 3135, 1997.

[49] T. Takeuchi, H. Takeichi, S. Sota, H. Sakai, H. Amano and I. Akasaki, Japanese Journal of Applied Physics, vol. 36 part 2, L177, 1997.

[50] B. K. Ridley, B. E. Foutz and L. F. Eastman, Physical Review B, vol. 61, 16862, 2000.

[51] S. Das Sarma and B. A. Mason, Annals of Physics, vol. 163, 78, 1985.

[52] T. Bradley, “Inside Kepler” [Online], Available: http://www.nvidia.co.uk/content/PDF/isc-2012/2-ISC12-Inside-Kepler-Architecture.pdf

[53] B. Molnar, G. Tolnai and D. Legrady, Annals of Nuclear Energy, vol. 132, 46, 2019.

[54] J. Wei and F. E. Kruis, Journal of Computational Physics, vol. 249, 67, 2013.

[55] Y. Nejahi, M. S. Barhaghi, J. Mick, B. Jackman, K. Rishaidat, Y. Li, L. Schwiebert and J. Potoff, SoftwareX, vol. 9, 20, 2019.

[56] D. Vasileska, S. M. Goodnick, "Computational Electronics, Synthesis Lectures on Computational Electromagnets", Morgan & Claypool Publishers, 2006.

[57] K. Tomizawa, "Numerical Simulation of Submicron Semiconductor Devices", Artech House, 1993.

[58] M. Lundstrom, "Fundamentals of carrier transport", 2nd ed., Cambridge University Press, 2000.

[59] C. Kittel, "Introduction to Solid State Physics", 8th ed., John Wiley & Sons Inc., 2000.

[60] R. D. Kronig and W. G. Penney, Proceedings of the Royal Society of London, vol. 130, pp. 499-513, 1931.

[61] C. Bulutay, B. Ridley and N. Zakhleniuk, Physical Review B, vol. 62, 15754, 2000.

[62] B. Guo, H. Guo, S. Zhang and D. Song, Physica B: Condensed Matter, vol. 405, 4925, 2010.

[63] J. Wu, W. Walukiewicz and E. Haller, Physical Review B, vol. 65, 233210, 2002.

[64] D. K. Ferry, "Semiconductor Transport", 1st ed., Taylor & Francis, 2000.

[65] B. K. Ridley, "Quantum Processes in Semiconductors", 4th ed., Oxford University Press, 1999.

[66] B. K. Ridley, “Electrons and Phonons in Semiconductor Multilayers”, 2nd ed., Cambridge University Press, 2009.

[67] C. Jacoboni, “Theory of Electron Transport in Semiconductors – A Pathway from Elementary Physics to Nonequilibrium Green Functions”, 1st ed., Springer, 2010.

[68] K. Yokoyama and K. Hess, Physical Review B, vol. 33, 5595, 1986.

[69] T. Ando, A. B. Fowler and F. Stern, Reviews of Modern Physics, vol. 54, 437, 1982.

[70] G. Bastard, “Wave Mechanics Applied to Semiconductor Heterostructures”, Halsted Press, 1988.

[71] E. Yamaguchi, Journal of Applied Physics, vol. 56, 1722, 1984.

[72] K. S. Yoon, G. B. Stringfellow and R. J. Huber, Journal of Applied Physics, vol. 62, 1931, 1987.

[73] P. J. Price, Annals of Physics, vol. 133, 217, 1981.

[74] P. Lugli, C. Jacoboni, L. Reggiani and P. Kocevar, Applied Physics Letters, vol. 50, 1251, 1987.

[75] J. H. Leach, C. Y. Zhu, M. Wu, X. Ni, X. Li, J. Xie, U. Ozgur, H. Morkoc, J. Liberis, E. Sermusksnis, A. Matulionis, H. Cheng, C. Kurdak, Applied Physics Letters, vol. 95, 223504, 2009.

[76] NVIDIA, “CUDA Toolkit Documentation” [Online], Available: docs.nvidia.com/cuda/

[77] G. M. Amdahl, AFIPS ’67 (Spring): Proceedings of the April 18-20, 1967, Spring Joint Computer Conference, pp. 483-485, 1967.

[78] W. Knap, S. Contreras, H. Alouse, C. Skierbiszewski, J. Camassel, M. Dyakonov, J. L. Robert, J. Yang, Q. Chen, M. Asif Khan, M. L. Sadowski, S. Huant, F. H. Yang, M. Golran, J. Leotin and M. S. Shur, Applied Physics Letters, vol. 70, 2123, 1997.

[79] E. Bellotti and F. Bertazzi, "Nitride Semiconductor Devices", edited by J. Piprek, Wiley-VCH, 2007.

[80] V. Bougrov, M. E. Levinshtein, S. L. Rumyantsev and A. Zubrilov, "Properties of Advanced Semiconductor Materials", John Wiley & Sons, Inc., 2001.

[81] T. P. Chow and Ghezzo, "III-Nitride, SiC, and Diamond Materials for Electronic Devices", vol. 423, edited by D. LK. Gaskill, C. D. Brandt and R. J. Nemanich, Material Research Society Symposium Proceedings, 1996.

[82] M. Semenenko, O. Yilmazoglu, H. L. Hartnagel and D. Pavlidis, Journal of Applied Physics, vol. 109, 023703, 2011.

[83] T. Hoshino and N. Mori, Japanese Journal of Applied Physics, vol. 87, 04FG06, 2018.

[84] S. Wang, Y. Dou, H. Liu, Z. Lin and H. Zhang, Journal of Electronic Materials, vol. 47, 1560, 2018.

[85] C. Bulutay, B. K. Ridley and N. A. Zahleniuk, Physical Review B, vol. 68, 115205, 2003.

[86] D. J. Suntrup III, G. Gupta, H. Li, S. Keller, U. K. Mishra, Applied Physics Letters, vol. 105, 263506, 2014.

[87] K. T. Tsen, D. K. Ferry, A. Botchkarev, B. Sverdlov, A. Salvador, H. Morkoc, Applied Physics Letters, vol. 72, 2131, 1998.

[88] R. Mickevicius and A. Reklaitis, Journal of Physics: Condensed Matter, vol. 1, 9401, 1989.

[89] B. K. Ridley, Semiconductor Science and Technology, vol. 4, 1142, 1999.

[90] D. Yu. Protasov, T. V. Malin, A. V. Tikhonov, A. F. Tsatsulnikov and K. S. Zhuravlev, Semiconductors, vol. 47, 33, 2013.

[91] J.-Z. Zhang, A. Dyson and B. K. Ridley, Physical Review B, vol. 84, 155310, 2011.

[92] N. Appleyard, "Monte-Carlo Simulations of Gunn Diodes and Hot-Phonon Effects in Bulk Semiconductors", Ph.D. Thesis, University of Hull, 2016.