The Physical Modelling of Heterojunction Field Effect Transistors

Robert Drury, BSc.

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
Department of Electronic and Electrical Engineering

September 1993

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others

Abstract

This project is primarily concerned with the incorporation of quantum effects into physical models for heterojunction field effect transistors. Several simulations have been developed including a one-dimensional Schottky-gate model which self-consistently solves the effective mass Schrödinger equation with Poisson's equation. This model employs a fast, accurate and robust solution algorithm based upon an expanded Newton scheme. This work is extended to two-dimensions, permitting charge transport and hence adding the current-continuity equation. All three equations are solved under non-equilibrium conditions. Finally a quasi-two-dimensional HFET model has been written, also including quantum mechanics which produces excellent agreement with measured characteristics

As a rigorous solution of the full two-dimensional Schrödinger equation and corresponding transport equation is very demanding and computationally expensive the problem has been simplified by assuming the electron wavefunction to take the form of Bloch, or travelling wave solutions is the directions parallel to the heterojunction. The component normal to the heterojunction interface is then solved by taking multiple one-dimensional solutions sampled at various positions throughout the device.

This new approach requires alternative solution algorithms to be developed since the conventional schemes are not applicable. This thesis reviews the physics behind semiconductor heterojunctions, discusses the solution schemes used in the models and presents results from the one-dimensional, two-dimensional and quasi-two-dimensional simulations.

Contents

The Physics of Heterojunctions (cont.)

List of Tables

List of Figures

Symbol Index

γ	Degeneracy factor
δ	Prefix for small increment
ε	Permittivitty
ε_0	Permittivitty of free space
ε_r	Relative permittivitty
λ	Electron energy level (eigenvalue)
λ_{th}	Thermal or deBroglie wavelength
μ	Electron mobility
ξ	Spatial component of the electron wavefunction
π	Pi (3.1415926)
ρ	Net charge density
τ_p	Electron momentum scattering relaxation time
τ_w	Electron energy scattering relaxation time
ϕ	Temporal component of the electron wavefunction
ϕ	Fermi potential
ϕ'	Normalised fermi potential
χ	Electron affinity
ψ	Electrostatic potential
Ψ	Full electron wavefunction (eigenvector)
ψ'	Normalised electrostatic potential
ω	Frequency
Ω	Volume

B	Magnetic field
C_{GC}	Gate to channel capacitance
C_{GD}	Gate to source capacitance
C_{GS}	Gate to drain capcacitance
D	Electron diffusivity
D	Electric displacement vector
E	Electric field
E_A	Acceptor ionisation energy
E_C	Conduction band edge
ΔE_C	Conduction band discontinuity
E_D	Donor ionisation energy
E_F	Fermi-level
E_G	Conduction band - valence band separation
E_H	Conduction band discontinuity due to the heterojunction
E_T	Trap ionisaton energy
E_V	Valance band edge
ΔE_V	Valence band discontinuity
f	Electron distribution function
f_0	Fermi function
f_T	Cut-off frequency
F	Helmholtz free energy
F	Force
$F1$	Poisson equation functional
$F2$	Transport equation functional
g_m	Transonductance

$g(E)$	Density of states
H	Perturbation operator
H	Magnetic field strength
h	Planks constant
$\hbar$	Reduced Planks constant
$h_{i,j}$ ($g_{i,j}$)	i,j^{th} mesh spacing in the x (y) direction
i	complex notation
I_D	Drain current
I_S	Source current
J	Electron current density
J	Jacobian matrix
k	Wavevector
k_B	Boltzmann's constant
m^*	Electron effective mass
m_0	Electron rest mass
n	Total electron density (n_{2D}+n_{3D})
n'	Normalised electron density
n_D	filled donor sites
n_{2D}	Two dimensional (quantum) electron density
n_{3D}	Three dimensional (bulk) electron density
n_s	Sheet electron density
N_A	Acceptor density
N_A^-	Ionised acceptor density
N_C	Density of states
$N_{C\,GaAs}$	Density of states in GaAs

N_D	Donor density
N_D^+	Ionised donor density
N_F	Minimum noise figure
N_T	Trap density
N_T^-	Ionised trap density
p	Momentum
q	Electron charge
Q	Heat flow to electron gas
R_D	Drain access resistance
R_S	Source access resistance
R_y^*	Rydburg constant
t	Time
T	Kinetic energy
T_0	Lattice temperature
T_e	Electron temperature
u	Group velocity
u_k	Bloch state periodic function
v	Velocity
V_{bi}	Built in Schottky bias
V_{th}	Thermal voltage (k_BT/q)
V_{xc}	Exchange-correlation energy
U	Total potential energy
U_E	External bias potential
U_L	Perfect crystal lattice potential
U_Q	Quantum correction energy

U_S	Thermal motion scattering potential
V_{xc}	Exchange-correlation potential
w	Energy per electron
W	Total energy
x	Aluminum mole fraction
$x_{i,j}$ ($y_{i,j}$)	i,j^{th} mesh position in the x (y) direction
Z	Gate finger length

List of Abbrieviations

2DEG	Two-Dimensional Electron Gas
BTE	Boltzman Transport Equation
HEMT	High Electron Mobility Transistor
HFET	Heterojunction Field Effect Transistor
IC	Integrated Circuit
IGFET	Insulated Gate Field Effect Transistor
LCAO	Linear Combination of Atomic Orbitals
LU	Lower-Upper (as in decomposition)
MBE	Molecular Beam Epitaxy
MESFET	MEtal Semiconductor Field Effect Transistor
MODFET	MOdulation Doped Field Effect Transistor
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
MOVPE	Metal-Organic Vapour Phase Epitaxy
PC	Personal Computer
SIS	Semiconductor-Insulator-Semiconductor
TEGFET	Two-dimensional Electron Gas Field Effect Transistor

Acknowledgements

I would like to thank my supervisors, Professor Chris Snowden and Dr. Bob Miles for their help and guidance throughout this work.

This project was funded by SERC, the Science and Engineering Reasearch Council and M/A-COM Corporate R&D, Lowell, USA.

Chapter 1

Introduction

1.1 A brief historical background

Over the past thirty years solid state technology has rapidly matured, moving from germanium and silicon, the original materials used to make semiconductor devices, to more exotic materials, in particular III-V compound semiconductors. One material in common use is gallium arsenide (GaAs) although there is a lot of interest in other compounds such as indium-gallium arsenide (InGaAs) and indium phosphide (InP). These materials are used because of their superior performance when compared with germanium and silicon, particularly due to their higher low field electron mobilities, which permit faster devices to be produced. Unfortunately much of the original silicon technology was not transferable to these new materials, for example the most popular silicon FET devices separate the gate from the rest of the device via a thin oxide layer, thus forming a Metal Oxide Semiconductor Field Effect Transistor (MOSFET). Although the first GaAs technologies were also aimed at producing enhancement mode insulated gate structures (IGFET) it was soon discovered that the surface states present at the GaAs/insulator interface were three or four orders of magnitude greater than that with silicon [Haigh, 1989]. Consequently the vast majority of the electrons induced by the gate were trapped and unable to participate in charge transport. The rectifying gate structure was therefore made using a Schottky contact [Mead, 1966], thus producing the MEtal Semiconductor Field Effect Transistor (MESFET).

Modern processing techniques, namely Molecular Beam Epitaxy (MBE) and Metal-Organic Vapour Phase Epitaxy (MOVPE), have made possible the fabrication of many novel structures, some pushing device technology into the quantum regime. One such device, the Heterojunction Field Effect Transistor (HFET) is produced by varying the material composition of the crystal from which the device is made. Consequently the electron wavefunctions can be engineered, significantly altering them from their classical form. These altered wavefunctions can then be used to improve device performance. In general these devices all follow a similar basic prescription whereby a doped wide band

gap material is grown, whilst maintaining a true crystalline structure, on top of a narrow band gap material. For example a common configuration consists of a bottom layer of epitaxial, unintentionally doped GaAs approximately 1 μm thick on top of which is grown a heavily doped AlGaAs layer. This upper layer, the larger band gap material, is typically doped at around 10^{24} donors m^{-3}, has an aluminum mole fraction between 0.2 and 0.4 and is less than 500 Å thick. The wafer is then processed in a similar fashion to the GaAs MESFET, with a Schottky gate evaporated onto the surface, used to modulate the current flowing between two ohmic contacts, the source and drain, illustrated in Figure 1,1.

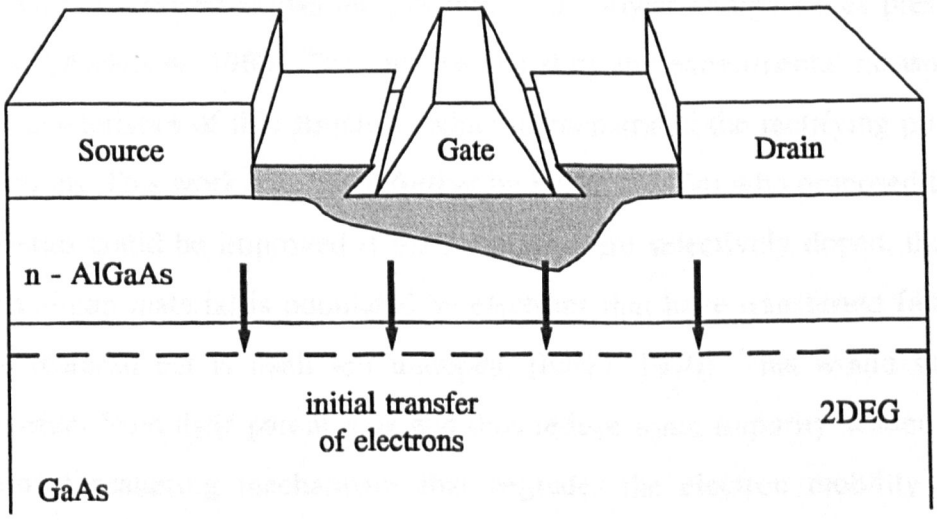


Figure 1,1

Cross-section of a typical AlGaAs/GaAs HFET

Although different technologies are used an analogy can be drawn between the MOSFET and the HFET, based upon the layered structure of both devices. The abrupt change in band gap across the interface present in both devices produces a thin sheet of carriers in the narrow band gap material, Si and GaAs respectively. These carriers are contained within a potential well of approximately atomic dimensions and thus are described not by classical mechanics, but rather quantum mechanics. However MOSFETs differ from HFETs in that they are not made from a single crystal, but rather

from the crystalline/amorphous structure formed from the silicon and SiO_2 layers. The crystalline nature of the HFET interface allows the electrons in the wide band gap material to transfer to the lower energy states present in the narrow band gap material[1]. Here the analogy with the MOSFET ends, as this transfer is not possible since any electrons present in the SiO_2 layer are bound locally (SiO_2 is an insulating material) and not permitted to travel into the silicon channel. This analogy is useful because the first quantum models, self-consistently solving Schrödinger's equation with Poisson's equation, were produced for MOS devices, and then later modified for use in heterostructure devices.

The accumulation of charge at the interface of a heterojunction was first demonstrated by Anderson, using a Ge/GaAs structure, where the electrons in the wider band gap material (GaAs) diffused across the junction to the lower energy states present in the germanium [Anderson, 1960]. This was validated by the experimental measurement of the I-V characteristics of this structure, which demonstrated the rectifying properties of these junctions. This work was taken further by Esaki and Tsu who proposed that device characteristics could be improved if the structure were selectively doped, that is if the narrow band gap material is populated by electrons that have transferred from a larger band gap material but is itself left undoped, [Esaki, 1970]. This would separate the charge carriers from their parent ions and thus reduce ionic impurity scattering, one of the dominant scattering mechanisms that degrades the electron mobility[1]. The first modulation doped structure was grown by Dingle *et al.* Here an AlGaAs/GaAs superlattice was grown, doping the wide band gap AlGaAs layers and leaving the GaAs undoped. The electron accumulation layers, labelled a **2 D**imensional **E**lectron **G**as (2DEG) exhibited higher mobilities than are obtainable through either a uniformly doped superlattice structure or "equivalently doped" bulk GaAs, [Dingle, 1978]. The first working devices were produced by Mimura *et al.*, who reported HFETs with mobilities that were "significantly higher than those of the GaAs MESFET" at both 77K and 300K [Mimura, 1980]. Mimura produced both HFETs and MESFETs having the same geometric dimensions and measured their characteristics. At low temperature (77K) a dramatic increase in the transconductance of the HFET over the MESFET was observed and Mimura consequently predicted that "the high speed performance of HFET should

---

1 Discussed more fully in Chapter 2, The Physics of Heterojunctions

be three times superior to that of the MESFET at 77K". Hiyamizu *et al.* continued developing the HFET, further separating the 2DEG electrons from their parent ions by incorporating an undoped AlGaAs spacer layer between the doped AlGaAs and undoped GaAs layers, [Hiyamizu, 1983]. They were rewarded with around an order of magnitude increase in the mobility of the carriers, at least when the device was operated below 77K. Unfortunately the spacer layer has a detrimental effect on the 2DEG density, which reduces with increasing spacer layer thickness. Consequently the thickness of this layer has to be optimised to produce the highest carrier mobility whilst retaining the largest number of free carriers.

Hence by 1983 the basic components of a modern HFET were in place, a MESFET-like terminal configuration, used to modulate the current flowing through a layered crystal structure containing a 2DEG, formed at the heterojunction interface, and possessing a high mobility due to it's separation from the ionic impurities. Each of the above workers labelled this device differently, and hence HFETs are also known by several other acronyms, each derived from a different part of the devices manufacture or operation. These include **MO**dulation **D**oped **F**ield **E**ffect **T**ransistor (MODFET), **H**igh **E**lectron **M**obility **T**ransistor (HEMT) and **T**wo-dimensional **E**lectron **G**as **F**ield **E**ffect **T**ransistor (TEGFET). HFET is the chosen acronym in this work because it is more general encompassing the whole spectrum of heterostructure devices.

1.2 HFET Performance

The HFET is one of the most promising devices over a range of criteria including outright performance and processing requirements. The following table is taken from Morkoç and Solomon's review [Morkoç, 1984], illustrating the ranking of HFETs versus other integrated circuit technologies. This table, giving the HFET the lowest score predicts this device to be one of the dominant devices in high speed IC applications, although it assumes all the parameters to have and equal weighting and fails to include economic factors.

Parameters	Josephson Junction	HFET	GaAs MESFET	Vertical FET	HBT	Silicon MOSFET	BJT
Speed	1	1	3	1	2	5	4
Power-delay product	1	1	2	2	4	1	4
Lithography requirements	2	2	3	5	1	4	1
Threshold control	3	4	4	4	1	2	1
Processing complexity	4	2	1	4	5	3	5
Materials problems	5	4	3	5	4	1	2
Score	16	14	16	21	17	16	17

Table 1.1

Relative ranking of IC technologies, after Morkoç. (1 = excellent, 5 = poor)

The high speed of these devices due to an increased average electron mobility has many benefits. The first, high frequency operation, is determined by the conventional equation, equation (1.1).

$$f_T = \frac{g_m}{2\pi (C_{GS}+C_{GD})} \tag{1.1}$$

Where g_m is the transconductance and C_{GS} and C_{GD} are the gate to source and gate to drain capacitances respectively.

The higher average mobilities increase the average velocity which together with the improved charge control increases the transconductance, g_m and therefore the cutoff frequency, f_T. This reaches a theoretical limit in excess of 300 GHz for 1 μm gate length GaAs devices [Haigh, 1989]. Conventional HFETs have shown good cutoff

frequencies, with Chao [Chao, 1983] publishing values in excess of 70 Ghz for a 0.25 μm gate AlGaAs/GaAs HFET. If the average velocity is increased further by using material other than GaAs, InGaAs for example, still higher cutoff frequencies are possible. Hikosaka *et al.* review the cutoff frequencies of various different types of HFET devices [Hikosaka, 1988], showing the increase in f_T with decreasing gate length and changing material composition, Figure 1,2. Other reported figures include cutoff frequencies in excess of 100 Ghz for a 0.1 μm gate length AlInAs/InGaAs pseudomorphic[2] HFET, [Wang, 1988].

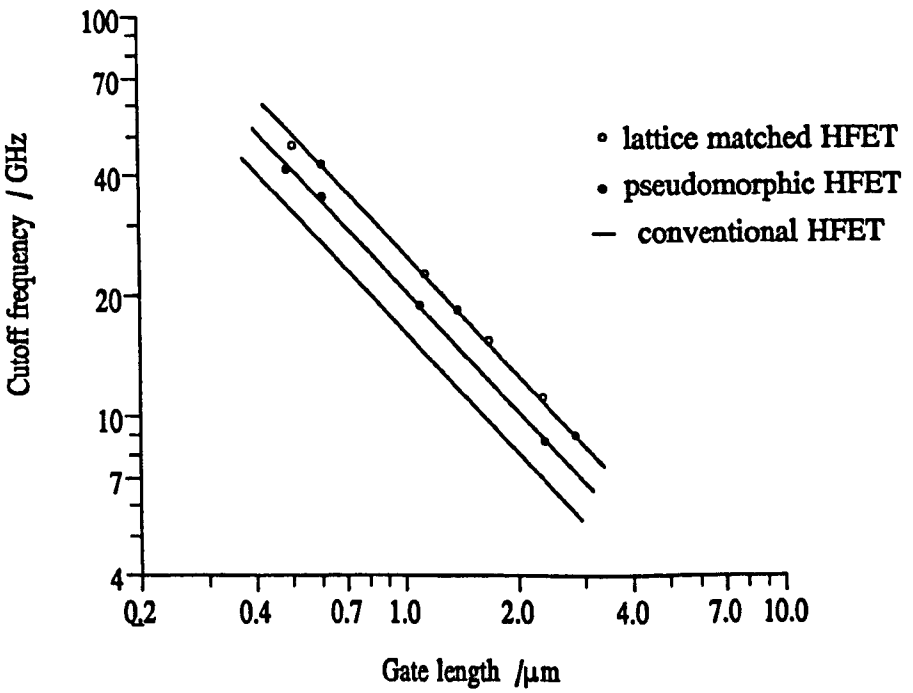


Figure 1,2

Variation of f_T with gate length and device type, after [Hikosaka, 1988]

For high frequency operation, particularly with regard to MMICs, one of the principle benefits of HFETs is their low noise [Ladbrooke, 1989]. Delagebeaudeuf *et al.* has shown evidence that the minimum noise equation for MESFETs, equation (1.2), is

---

[2] Both AlGaAs and GaAs materials have a similar lattice constant, the size of the crystal cell. However this is not always the case, with some HFETs being made form non-lattice matched materials, for example InGaAs. These devices accommodate a small amount of strain, typically only using a thin layer of the non-matching material, and are called "pseudomorphic" HFETs.

applicable to HFETs [Delagebeaudeuf, 1985].

$$N_{F\,\min} = 1 + 2.5\,\omega C_{GC}\left(\frac{R_S + R_D}{g_m}\right)^{0.5} \qquad (1.2)$$

where $N_{F\,min}$ is the minimum noise figure, ω the frequency, C_{GC} the gate to channel capacitance and R_S and R_D the source and drain access resistances respectively.

The increased transconductance of HFETs when coupled with their lower source and drain access resistances (the latter due to the higher low-field mobility[3]), reduce their noise figure with respect to MESFETs. This has been verified by Pospieszalski *et al.* who reported an HFET with a noise figure between 0.4 and 0.5 dB with an associated gain between 14.2 and 15.1 dB for a device operating at 8 GHz at room temperature [Pospieszalski, 1988].

1.3 Modelling of HFETs

It has become apparent, especially in recent years, that good semiconductor device design is aided by good physical models. The main aims of device models are to give a better understanding of device operation and to accurately predict device performance, both of which can then used to aid the design of better devices. The former, specifically for the device engineer, gives insight into the variation of both microscopic and macroscopic variables for different device layouts and configurations. For example, the location of the seed for avalanche breakdown can be monitored and factors such as the shape and depth of the gate recess altered to delay the onset of this effect. Accurate simulation is necessary such that the performance of a particular device can be tested and optimised before production and hence reduce both the time and expense involved in developing a new device. The 'performance' of a device is quite a subjective term,

[3] Most of the source-drain potential is dropped over the gate region, consequently the source and drain access regions are relatively low-field.

formed from absolute measures of operation such as gain, cutoff frequency and noise figure and other factors such as the susceptibility of these parameters to the variations in fabrication which necessarily accompany the manufacture processes. It is obviously desirable not only to have fast, high gain devices but also to obtain a high yield.

All physical device models, provided they are correct, fulfil the first aim automatically, as they provide a window into the device and enable the user to monitor the evolution of variables, including those that may be either difficult or even impossible to measure. The processes occurring deep within the device are immediately observable, and the effects of changes in the device layout and composition can readily be checked. This information can then be fed back, acting as a guide for device improvement and optimization, and may even be automated, running the device simulations as part of an optimisation package, homing in on parameters such as the 'best' doping profile and device layout.

Unfortunately the second aim is harder to fulfil since an accurate simulation requires both accurate physical equations, a suitable choice of domain and corresponding data. The choice of the physical equations is not trivial; too rigorous and the simulation will produce a challenging algorithm and require excessive computation time and too simple and the simulation may fail to model important device characteristics. Both extremes defeat the object of accurate device simulation making their respective schemes either too slow or too inaccurate to operate as a predictive tool. Even when appropriate device equations have been chosen further problems arise as they may require information that is not readily available. In the case of the HFET these unknown parameters include values for the energy band alignment across the heterojunction interface, accurate velocity - energy curves and even on a more mundane level accurate values for the trapping densities at the surface. Hence the device modeller is required to use his or her judgement at an early stage in the design of a semiconductor device model, both in the choice of device equations, modelled domain and in assigning values to the parameters that are either unknown or have uncertain values.

GaAs MESFETs have a whole array of physical models, both Monte Carlo and hydrodynamic, which have been adapted for the simulation of HFETs. The Monte Carlo

schemes have the advantage that they are quasi-exact, requiring few assumptions and are able to take account of most microscopic processes including non-stationary dynamic effects, velocity overshoot and ballistic transport. Also current continuity is guaranteed and noise properties are automatically obtained. However they are very computer intensive relying on long simulation times to improve the statistical accuracy of the results [Salmer, 1988]. Also it is unclear how they would manage a changing sub-band energy structure, and the effects this would have on the scattering probabilities[4].

The hydrodynamic schemes started with the analytic device models, progressing through the drift-diffusion approximations to the current 'state of the art' two-dimensional models which are based upon a full solution of the Boltzmann transport equation. The first of these, the analytic models, requires many simplifications to the physical device equations and the modelled domain to produce a tractable scheme. Several examples of these models are available in the literature including [Park, 1986], [Weiss, 1988] and [Hida, 1986]. The drift-diffusion equations using an exponential relationship for the discretization of the current continuity equation was first employed by Scharfetter and Gummel [Scharfetter, 1969] and an example of these equations being applied to heterostructure devices is given by [Mawby, 1988]. Finally the more comprehensive non-equilibrium transport models have been developed by Shawki, [Shawki, 1990a] and [Shawki, 1990b].

However the new devices made possible by the advanced fabrication methods, with features close to the atomic scale, require new models that take account of the novel effects that occur in the quantum regime. Conventional models, even the highly sophisticated non-equilibrium transport hydrodynamic schemes are based in the classical world, using approximations for the electron wavefunction derived at the classical limit. These schemes start to conflict with fundamental concepts in physics once the scale over which events occur approaches the deBroglie wavelength, (1.3).

---

[4] Monte Carlo models follow the time evolution of the particles, allowing them to be accelerated by the electric fields present, scattering the particles from one another according to a predetermined scattering probability. If this probability varies as a function of the sub-band structure a comprehensive Mote Carlo model would be much more complicated and several orders of magnitude slower, being required to calculate the new scattering probabilities at each new time step.

$$\lambda_{th} = \frac{h}{m^* v} \tag{1.3}$$

Boltzmann's transport equation is deterministic, with both position and momentum specified at the same time. The Heisenburg uncertainty principle states that these quantities cannot be resolved simultaneously, rather these quantities can only be evaluated to within a certain accuracy, given by equation (1.4).

$$\Delta x . \Delta p = h \tag{1.4}$$

where Δx and Δp are the uncertainties in the values of position and momentum respectively and h is Plank's constant.

Thus the BTE is fundamentally in conflict with the Heisenburg uncertainty principle at these dimensions[5]. Consequently the simplifying assumptions used to derive the semi-classical device equations are no longer valid and parameters such as the electron wavefunction which were assumed to have a travelling wave form (the classical limit) now have to be calculated explicitly. In typical HFET devices this results in an electron sub-band energy structure, with the allowed energy states in the conduction band discretized, and accompanied by a corresponding change in the density of states. These changes have two principle effects, the first is to change the electron density, which now spreads out, blurring the sharp profiles across the heterojunction discontinuity obtained when using classical schemes. The second is to alter the transport properties of the electrons. The modelling of HFET's therefore requires a different transport equation that can accommodate this feature. Several alternative theories have been developed including the density matrix formalism, Feynman path integral and Wigner distribution function [Wang, 1989]. Several authors have attempted to include quantum physics in a physical device model, with perhaps the most successful, from the physics point of view being that of Yablik *et al.*, who self-consistently solves the three-dimensional time-

[5] The Boltzmann transport equation is always in conflict with the Heisenburg uncertainty principle but when the scale lengths are large with respect to the deBroglie wavelength the relative error arising from the uncertainty principle is small.

dependent Schrödinger equation with Poisson's equation, [Yablik, 1989]. The principle disadvantages with this model is "the numerical problems encountered in specifying the boundary conditions at the contacts of the device" and "the inclusion of dissipative effects into the Schrödinger equation". Also this method is very expensive in terms of computer resources. The paper only reports results for a very small device possessing a 26 nm gate and 120 nm total width, which has been simulated for only 0.01 ps. The paper predicts a 1 ps simulation to take approximately 70 hours on a CRAY X-MP computer. This is acceptable, although not desirable, for research purposes but a device engineer requires a model that runs at least two orders of magnitude faster.

Zhou and Ferry, [Zhou, 1993] also model small devices, although this is probably to emphasize the differences between the quantum and classical solutions. Their model incorporates quantum mechanics by including a 'quantum pressure' term in the electron energy, equation (1.5)

$$w = \frac{1}{2}m^* v^2 + \frac{3}{2}\gamma k_B T_e + U_Q \qquad (1.5)$$

where v is the average velocity, T_e the electron temperature, γ the degeneracy term due to Fermi-Dirac statistics and U_Q the quantum pressure term. This additional term is calculated via a moments solution of the Wigner distribution function and is corresponds to retaining the phase information contained within quantum mechanics from the potential energy. The conventional hydrodynamic equations are then solved with this correction. Whilst this method is better than the classical hydrodynamic schemes alone, and more tractable in terms of computer resources than the model proposed by Yablik, it fails to account for the phase information contained in the kinetic energy terms and also neglects any change in the density of states, and the consequent discretization of the 2DEG particle's motion.

A different approach is taken by T. Wang and C.H. Hsieh, [Wang, 1990], and S.H. Hg et al., [Ng, 1991], both of which choose to solve the one-dimensional Schrödinger equation normal to the heterojunction interface, and use classical transport mechanisms to evaluate current parallel to the heterojunction. These models are much simpler in that

they are almost quasi two-dimensional in nature, only permitting current flow parallel to the heterojunction. Whilst this is correct for the 2DEG, and excellent quasi two-dimensional MESFET models have been produced based upon this assumption, it is questionable whether the inclusion of quantum effects which may be a second order correction is worth the effort in a first order model.

1.4 Aims and scope of current research

This thesis reviews some of the physics and numerical techniques required to produce a complete self-consistent two-dimensional model of a Heterojunction Field Effect Transistor. Subsequently several models have been produced both in one- and two-dimensions all including quantum mechanics. The one dimensional models are essentially used to simulate the epitaxial layer structure of these devices, starting with a simple Schottky-gate model and progressing to include some transport effects. The two-dimensional FET models were produced with particular attention is focused upon the introduction of quantum mechanics and the effects this causes. In an effort to minimise the computational demands, and thus produce a model useful for device design, the equations were simplified, whilst allowing transport in both the x- and y-directions. This model, reported in Chapter 5, is superior to previous quantum models, as it accounts for the restricted degrees of freedom of the 2DEG whilst allowing a thorough two-dimensional analysis of current transport around the gate. The inclusion of the important quantum effects whilst maintaining a reasonable simulation speed has lead to a new model. This solves the Schrödinger over a two-dimensional grid assuming that the electron wavefunction parallel to the heterojunction is described by 'free electron' or 'Bloch state' wavefunctions. This assumption is justifiable since no events occur in this direction of a scale comparable to the deBroglie wavelength, and hence there is nothing to precipitate the quantisation of electron motion in this direction. Further this assumption allows the one-dimensional Schrödinger equation to be solved in slices across the device, under non-equilibrium conditions, and provides a natural and straight forward method for the partitioning of the electrons into their two- and three-

dimensional components. Current flow is permitted in two-dimensions[6], thus a proper analysis of substrate injection and the transfer of electrons participating in parallel MESFET conduction is possible. This scheme also solves Poisson's equation with the classical transport equations.

Chapter 2 details the physics underlying the heterojunction in equilibrium and non-equilibrium conditions. The effect of the heterojunction upon the unbiased 'bulk' parameters is discussed, together with a review of the methods used to determine the conduction and valence band offsets. The equations governing the heterojunction are presented together with their origins, and quantum mechanics introduced. The effects of the discrete nature of the sub-band structure at the interface is discussed, concluding with a description of the transport equations in this novel structure.

Chapter 3 reviews the general numerical techniques used in the subsequent FET models, and includes sections on the discretization and scaling of the device equations. The choice of independent variable is discussed, and the evaluation of all the dependent parameters described. This chapter finishes with a review of the solution schemes available for non-linear systems, and presents reasons for the choice of the particular method used in the FET models.

The simulation of heterojunction devices starts with a one-dimensional substrate model, reported in Chapter 4 and follows the work of [Stern, 1984], [Vinter, 1983] and [Yoshida, 1986]. Here the model assumes no current flow which is adequate for reverse biased Schottky gates. This was then extended to include current flow, requiring Schrödinger's equation to be solved under non-equilibrium conditions. Both one-dimensional models are fast, robust and accurate, based upon a modified Newton-Raphson iterative solution scheme. Particular attention is drawn to the form of the partial derivatives of the quantum electron density, required by the solution scheme, and which turn out to be surprisingly straight forward. This chapter concludes with a section on the results of this model, simulating several different structures, and a brief discussion on how the one-dimensional model can be extended to produce rough guides for the design

---

[6] Only the bulk three-dimensional electrons are allowed to participate in motion across the heterojunction. The electrons associated with the 2DEG are bound in this direction and hence motion in this direction is forbidden.

of transistors.

The full two-dimensional model is described in Chapter 5, outlining the assumptions used in the scheme. This chapter continues by describing the modelled domain, the associated boundary conditions and the discretization schemes required for the current continuity equation. Results from classical simulations on GaAs MESFETs are followed by examples of HFET devices, illustrating the variation of the electron wavefunction across the device and the effect of quantum mechanics.

Chapter 6 describes an alternative model, reviewing the quasi two-dimensional (Q2D) scheme, including quantum mechanics. An improvement on the conventional Q2D model is presented, that is more self-consistent with the device physics. This produces results that are in excellent agreement with the measured characteristics of a pulse doped pseudomorphic HFET which are illustrated in the final section of this chapter.

Chapter 7 draws conclusions from all the preceding chapters and outlines further work on both the full two-dimensional and quasi-two-dimensional simulation schemes. The outline for a new model is presented, an intermediate scheme between the full two-dimensional and quasi-two-dimensional models that should be one or two orders of magnitude faster than the present full two-dimensional model but still retain two-dimensional current flow. Finally the use of this work is outlined and the future possibilities regarding applications discussed.

Chapter 2

Physics of Heterojunctions

2.1 Introduction

A heterojunction is defined within the context of this work to be a junction between two different materials, possessing different bulk parameters, whilst maintaining a single crystalline structure. This chapter discusses the effects of placing two different materials in such intimate contact under both equilibrium and non-equilibrium conditions. It outlines the equations that describe the junction, introducing the quantum mechanics used in the subsequent device models and the novel effects this causes.

2.2 Heterojunction formation

At a junction between two different semiconductors a transition region will form across the interface, perturbing the equilibrium electron distribution of each bulk material, and possessing different properties from both of the original semiconductors. In order to visualize the formation of a heterojunction we can consider the two solids in isolation, and the effects of bringing them closer together, until they are in crystalline contact. At large separations the materials will have no effect upon one another, and thus all parameters will have their bulk, equilibrium values, illustrate in Figure 2.1. As the semiconductors are brought closer together the wavefunctions associated with the conduction band electrons[1] in both materials start to overlap and consequently these electrons start to spend time on each side of the heterojunction. When the two semiconductors are finally brought together forming a single crystalline structure the potential barrier caused by the vacuum between the two solids ceases to exist and the wavefunctions can extend fully into both materials. The conduction band electrons now have the freedom to travel throughout the whole structure, and will distribute themselves

[1] Note that only the conduction band electrons participate in chemical reactions, since the valence band electrons are too tightly bound to the nucleus.

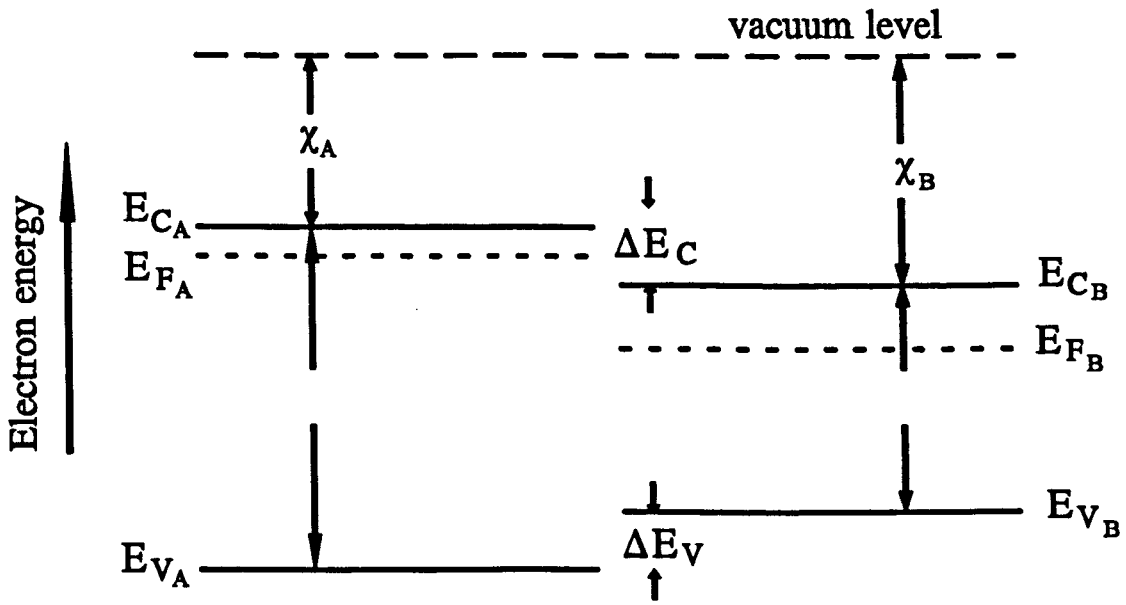


Figure 2.1

Energy band diagrams for two different semiconductors in isolation

in the lowest energy states possible. In the modulation doped structures typically used in n-type HFET's, this results in an initial migration of electrons across the heterojunction from the wide to narrow band gap material. The resulting local space-charge region creates an electric field which in turn opposes further diffusion. Equilibrium occurs when the drift and diffusion currents cancel to produce no net transport. Figure 2.4 shows a typical band diagram for such a junction under equilibrium conditions.

There are limitations to the choice of materials that can be used to make heterojunctions, one property that needs to be matched is the lattice constant, ie. the size of the crystal cell. Figure 2.2 shows a scatter plot of the lattice constant and band gap for some of the common materials used in heterostructure devices. Most heterostructure devices are made from lattice matched materials, AlGaAs/GaAs, InP/InGaAs, AlGaSb/GaSb, InAlAs/InP etc., though in some cases a small strain is accommodated producing a 'pseudomorphic' device. A good example of the materials commonly used in pseudomorphic devices is given by the AlGaAs/InGaAs/GaAs system. From Figure 2.2 it is evident that even small mole fractions of Indium in the InGaAs layer will increase

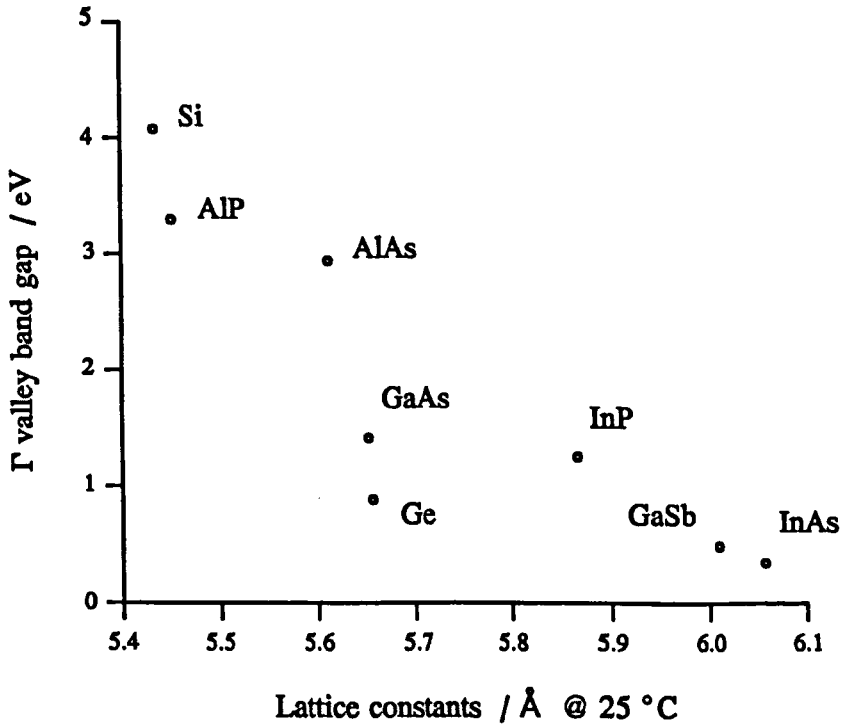


Figure 2.2

Energy band versus lattice constant for several semiconductor materials and compounds

the mismatch between the more usual AlGaAs GaAs interface. This device is fabricated to take advantage of the smaller band gap of the InGaAs compared with GaAs, producing a larger conduction band discontinuity at the AlGaAs/InGaAs interface and consequently increasing the effects of the heterojunction.

2.3 Bulk parameters at the interface

Physical parameters such as permittivity and effective mass are functions of the local energy band structure and together with the band gap are usually assumed to maintain their bulk values up to the interface, producing a discontinuity in these quantities at the heterojunction. These properties are only strictly defined for the bulk, and although to a first order approximation the above assumptions are valid, it is physically more correct to model the changing energy band structure throughout the transition region, and derive

the dependent parameters explicitly. This is not a trivial task, becoming heavily involved with quantum mechanics. A reasonable simplifying assumption would be to model the effects of the interface, assuming that they decrease logarithmically with distance, essentially returning the energy band structure to that of the bulk material within a few lattice constants of the heterojunction. This type of decay means that parameters such as effective mass only vary significantly from their bulk values over a few nanometres, and thus is very sharp indeed. In the models discussed in chapters 4, 5 and 6 these parameters are further simplified by using a linear approximation for their variation with position, which occurs as a natural consequence of the discretization scheme. The typical grid over which the device equations are solved, particularly in the two-dimensional model, will have a large internodal separation with respect to atomic dimensions, usually greater than 1-2 nm. Consequently the transition region for these parameters is contained within one or two nodal spacings. The discrete nature of the solution scheme linearly interpolates the value of any parameter positioned in between the grid points and hence linearly approximates the values of the parameters within the transition region. One may argue that the internodal separations should be small enough to resolve the changes in energy band structure, but on this scale, apart from the practical problems associated with implementing a model with far more mesh points, it becomes questionable whether other simplifying assumptions used to derive the device equations are still appropriate. With spatial discretization on this scale it may therefore be necessary to perform a full quantum analysis of the device.

The values for the bulk parameters for GaAs and AlGaAs used in the models described in chapters four and five are all taken from [Adachi, 1985], and have the following form.

$$E_G = 1.424 + 1.25x + 0.143x^2 \tag{2.1}$$

$$m^* = 0.067 + 0.083x \tag{2.2}$$

$$\epsilon_r = 13.18 - 3.12x \qquad (2.3)$$

where x is the aluminium mole fraction in the $Al_x Ga_{1-x} As$.

2.4 Conduction and valence band discontinuities

Whilst the bulk properties such as effective mass and relative permittivity are able to be approximated by a straight-forward linear interpolation, there is no simple derivation for the alignment of the energy bands. Consequently two of the most important parameters, the conduction and valence band edge discontinuities, have to be determined in a more rigorous fashion. These discontinuities, when coupled with the band bending at the interface, can form a two-dimensional potential well parallel to the heterojunction having dimensions of a few nanometres. This narrow well is responsible for many of the novel applications of heterojunction devices. The evaluation of the correct conduction and valence band edge discontinuities forms one of the most critical parameters in the analysis and modelling of heterojunction structures, with small errors amplified due to the exponential nature with which the electron density varies with energy. The difference in band gaps may be distributed between both the conduction and valence bands, and whilst it is apparent that the sum of the conduction and valence band discontinuities must equal the total change in band gap the exact values of these parameters is not yet known. The ratio of offsets may depend upon many variables including the quality of the materials and the interface, strain, doping profiles, temperature and in the AlGaAs/GaAs system the aluminium mole fraction. Considerable effort has been devoted to both calculating and/or measuring the exact values of these discontinuities, but to date the band offsets have only been evaluated to within an average accuracy of around ± 150 meV. A good review of the methods used to obtain the conduction and valence band edge discontinuities is given by Morkoç [Morkoç, 1991], and is briefly summarised below.

2.4.1 Theoretical models for the conduction band offset

One of the simplest theoretical models, proposed by Anderson [Anderson, 1962], sets the vacuum level constant across the heterojunction in equilibrium. This has the consequence that the conduction band edge discontinuity is equal to the difference in the electron affinities of the two semiconductors, denoted A and B, leading to equation (2.4).

$$\Delta Ec = \chi_B - \chi_A \tag{2.4}$$

thus the valence band discontinuity is given by equation (2.5)

$$\Delta E_V = E_{G_A} - E_{G_B} - \Delta E_C \tag{2.5}$$

As the value of electron affinity depends upon a given surface the conduction band discontinuity in this model depends upon the orientation of the crystal. Note also that the electron affinity is quite difficult to measure, depending strongly upon external variables, and thus this approach simply replaces one difficult problem with another!

Other simple theoretical models include those based upon a different constant parameter such as intrinsic level or conduction band edge, whilst more complicated versions generally include quantum mechanics. One such quantum mechanical approach is the 'Tight Binding Model' proposed by Harrison, where the valence band alignment is given from a method based upon linear combination of atomic orbitals (LCAO), [Harrison, 1977]. Another is based upon zeroing the charge associated with the Metal-Induced Gap States (MIGS), [Tersof, 1984], which are energy levels introduced into the energy gap of the semiconductor when the heterojunction forms.

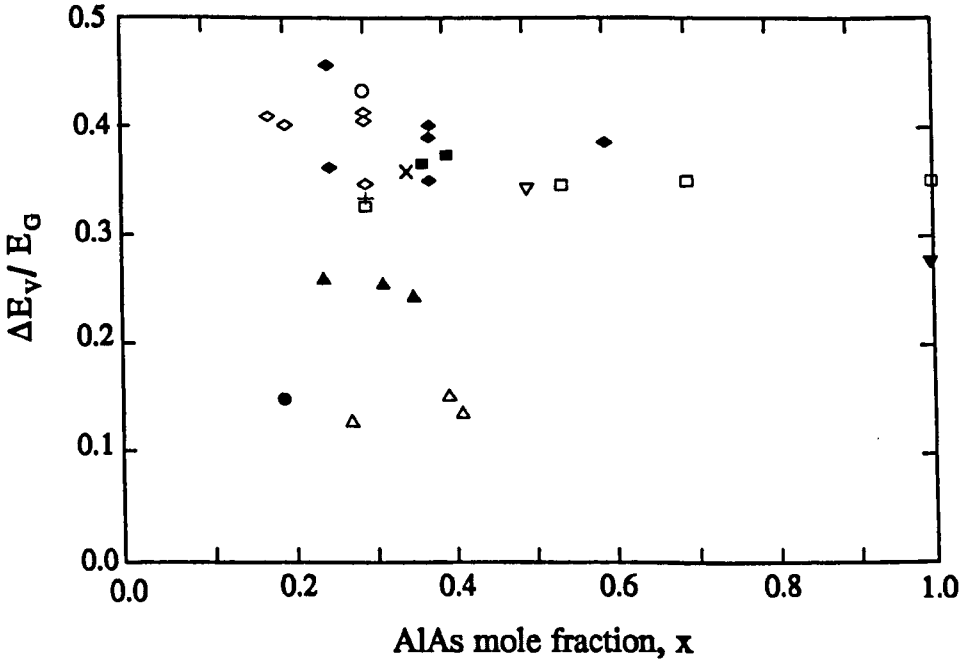


Figure 2.3

Ratio of valence band offset to total band gap [Wang, 1986]

Key: (•) Dingle 1974; (△) Gossard 1892; (○) Miller 1984; (▲) Dawson 1985; (+) Kröemer 1980; (×) Wu 1980; (■) Hickmott 1985; (◆) Batey 1985; (◊) Watenabe 1984; (□) Arnold 1984; (▽) Wang 1984; (▾) Wang 1985

2.4.2 Experimental techniques for finding the conduction band offset

Experimental techniques for measuring the conduction band offset are just as varied as the theoretical approach, including thermionic emission (I-V), Capacitance-Voltage (C-V) and an array of optical emission and absorption experiments. Thermionic measurements have been performed on a GaAs/AlGaAs/GaAs structure, forming a SIS capacitor, [Hickmot, 1985]. Here the conduction band discontinuity produces a potential barrier, ie. the 'insulator' in the SIS arrangement. The low voltage I-V data is then analyzed assuming the current is dominated by thermionic emission and the built in potential barrier, ΔE_C, extracted. This scheme works well provided the AlGaAs layer

is thick enough and not too heavily doped[2] to prevent a significant tunnelling current. Capacitance-Voltage measurements of a Schottky-barrier structure were used by Kröemer. Here the junction is reverse-biased and the sheet electron density obtained. This is then used with Gauss' law to determine the conduction band offset, [Kröemer, 1980]. This has problems resolving the electron sheet density on a scale below the Debye length due to the charge screening effect. Finally one of the optical methods uses absorption spectra of the approximate square wells produced by very thin GaAs-InGaAs-GaAs heterostructures [Dingle, 1974]. The discrete energy levels of the bound quantum states within the 2DEG produce associated peaks in the spectra from which the conduction band discontinuity can be extracted. Both of these latter methods require good models of the heterojunction structure to evaluate their respective data and deduce the energy band alignment.

Unfortunately many of the reported measurements of the conduction band and valence band discontinuities are only performed on one sample, thus there is little information on the variation of this important parameter with variables such a doping concentration, temperature etc. Figure 2.3 illustrates the wide variation in the published data on this variable. Taking the average of this data the valence band offset to band gap ratio in the AlGaAs/GaAs system is 0.33 with a standard deviation of 0.09. This compares with 0.23 from Anderson's rule, just outside these error limits.

2.5 Thermodynamic equilibrium

When the interface is at thermodynamic equilibrium the system can be described by the equations of state [Finn, 1986]. At a constant temperature, volume and number of particles equilibrium will occur when the Helmholtz free energy, F, is minimised, equation (2.6)

---

[2] A highly doped AlGaAs layer will induce a lot of band bending near the heterojunction, again leading to a narrow potential barrier.

$$\delta F = 0 \tag{2.6}$$

where δ denotes a small increment

Consider two materials, denoted A and B, with electrons allowed to flow across the interface until equilibrium is achieved. Equation (2.6) can be rewritten in terms of partial derivatives, equation (2.7).

$$\delta F = \left.\frac{\partial F}{\partial N}\right|_A \delta N_A + \left.\frac{\partial F}{\partial N}\right|_B \delta N_B = 0 \tag{2.7}$$

where δN_A and δN_B are the number of electrons flowing out of regions A and B respectively. Since the total number of electrons is conserved the number flowing out of region A equals the number flowing into region B, hence

$$\delta N_A = -\delta N_B \tag{2.8}$$

Substituting equation (2.8) into equation (2.7) it is evident by inspection that at thermal equilibrium the partial derivatives of F with respect to N on either side of the heterojunction must be equal, ie.

$$\left.\frac{\partial F}{\partial N}\right|_A = \left.\frac{\partial F}{\partial N}\right|_B \tag{2.9}$$

Electrons are added or removed from the system at the electro-chemical energy (the Fermi-level), thus the rate of change of energy with respect to number of electrons is this quantity, equation (2.10)

$$\frac{\partial F}{\partial N} = E_F \tag{2.10}$$

Consequently at thermal equilibrium the Fermi-level is constant across the heterojunction. This result can also be derived by considering electron currents derived from the Boltzmann Transport Equation, *see Appendix B*. Expressing the electron density as a function of energy it can be shown that in the isothermal case[3].

$$J = \mu n \nabla E_F \tag{2.11}$$

where J is the net electron current, μ the mobility and n the electron density. At equilibrium $J = 0$ and since μ and n are both non-zero

$$\nabla E_F = 0 \tag{2.12}$$

In equilibrium both approaches give a constant Fermi-level throughout the whole device, and hence this parameter can be used as good reference point. With no current flowing and the band discontinuities evaluated the energy band profile is obtainable through equation (2.13) and equation (2.14).

$$E_C = E_H - q.\psi \tag{2.13}$$

$$E_V = E_C - E_G \tag{2.14}$$

where E_H is the potential associated with the conduction band discontinuity and is therefore dependent upon the material type and thus position. The electrostatic potential, ψ, is given from the self-consistent solution of Poisson's equation with the charge density, equation (2.15).

---

[3] In equilibrium the carriers are all at the same temperature, that of the lattice, T_0

$$\nabla.(\epsilon\nabla\psi) + q.\rho = 0 \qquad (2.15)$$

For a typical n-type heterojunction this produces an energy band diagram similar to that shown below in Figure 2.4 clearly illustrating the two-dimensional potential well parallel to the interface. In this example both materials are doped at 10^{23} donors m^{-3}. When the narrow band gap material is left undoped the conduction band edge on this side is raised such that the Fermi-level tends to the centre of the band gap of this material. In typical HFET structures the wide band gap material is usually thin and terminates in either a Schottky gate or a free surface. In both cases the conduction band edge on this boundary is also raised above its equilibrium value, effectively reverse biased. In the case of the Schottky barrier this is due to the 'built-in' voltage and on the free surfaces due to the electrostatic field set up by filled traps. Hence the potential barriers seen by the electrons are raised at both extremes, increasing their confinement within the potential well.

Under equilibrium conditions the Fermi-level is determined from equation (2.12) and ψ found from Poisson's equation. However when the device is biased, and therefore placed under non-equilibrium conditions, there is no simple solution for the Fermi-level. Also the net electron momentum is now non-zero and the electron energy may be other than that of the lattice[4]. To model the device under these conditions Poisson's equation has to be supplemented by other device equations, one for each independent variable. These usually take the form of conservation equations, including the particle (current continuity), momentum and energy conservation equations.

2.6 Origin of the independent device equations

Semiconductors are modelled by solving the set of physical equations with their associated variables. However not all equations are independent, with some redundancy, and hence one is required to choose the 'independent variables', and relate all other

---

[4] Charges moving in electrostatic fields will gain or lose energy as work is done by or on the electric field

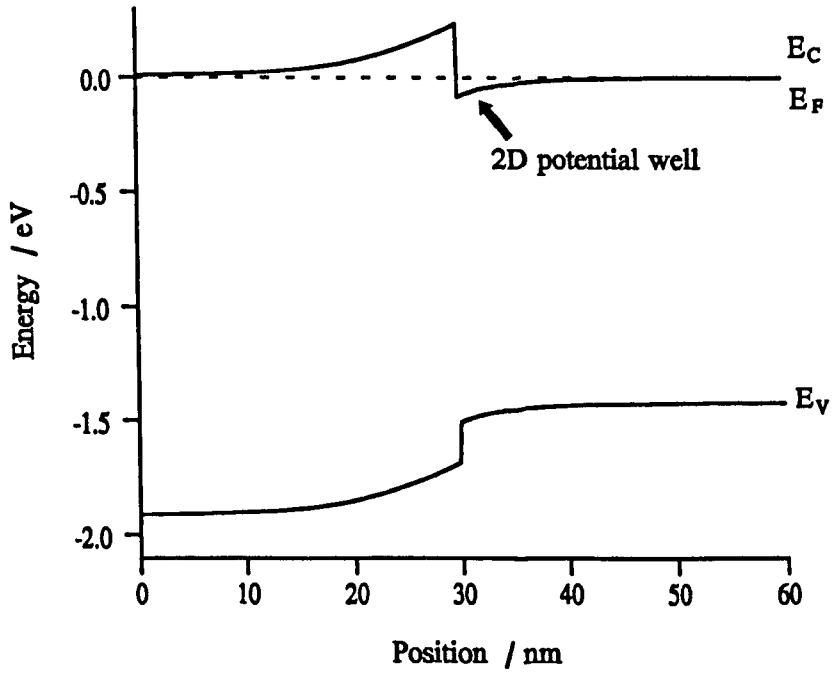


Figure 2.4

Energy band diagram of an $Al_{0.3}Ga_{0.7}As/GaAs$ heterojunction

variables in these terms. The choice is relatively arbitrary and made to simplify the solution of the whole set of equations. The choice of independent variable used in the FET models described in chapters 4,5 and 6 is discussed at the start of Chapter 3, and the following equations are labelled accordingly.

The basic physical device modelling equations, Poisson's and current continuity, can be derived from Maxwell's electromagnetic equations, (2.16) to (2.19), see [Selberherr, 1984].

$$\nabla . D = \rho \tag{2.16}$$

$$\nabla . B = 0 \tag{2.17}$$

$$\nabla \times E = -\frac{\partial B}{\partial t} \tag{2.18}$$

$$\nabla \times H = J + \frac{\partial D}{\partial t} \tag{2.19}$$

At low frequency[5] and in the steady state limit the partial derivatives with respect to time can be neglected, hence Maxwell's third equation becomes

$$\nabla \times E = 0 \tag{2.20}$$

From differential vector calculus it can be shown that if the curl of a vector is zero it may be expressed as a gradient field.

$$\nabla \times E = 0 \quad \rightarrow \quad E = -\nabla \psi \tag{2.21}$$

Substituting equation (2.21) into Maxwell's first equation we obtain Poisson's equation, the first device modelling equation.

$$\nabla . (\varepsilon \nabla \psi) + \rho = 0 \tag{2.22}$$

where ε is the total permittivity, $\varepsilon_0 \varepsilon_r$

The second principle device equation, which expresses current continuity, is obtained by taking the divergence of Maxwell's fourth equation, again at the low frequency limit

---

[5] The term 'low' here means sub MHz. A rough guide is to set the displacement current density less than 1% of the drift and diffusion current densities, at approximately 10^6 Am^{-2}. The displacement current is given by the product of permittivity and the partial derivative of electrostatic field with respect to time. Given the above limit and typical fields of the order 10^8 Vm^{-1} the minimum increment in time is 10^{-6} s, corresponding to a frequency 1 Mhz.

$$\nabla.(\nabla \times H) = \nabla.J = 0 \tag{2.23}$$

The third and fourth device equations, momentum and energy conservation are usually derived from the Boltzmann Transport Equation (BTE), although this is strictly incorrect at the heterojunction[6].

Expressing the probability of a particle having a given momentum at a certain position at a particular time as a distribution function, the electron current, momentum and energy conservation equations may be derived formally through Liouville's theorem [Selberherr, 1984]. This states that the derivative of the distribution function with respect to time over a particle trajectory is zero, equation (2.24).

$$\frac{df(x,k,t)}{dt} = 0 \tag{2.24}$$

Expanding equation (2.24) explicitly in terms of x, k and t gives the Boltzmann Transport Equation.

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial k}{\partial t}.\frac{\partial f}{\partial k} + \frac{\partial x}{\partial t}.\frac{\partial f}{\partial x} = 0 \tag{2.25}$$

By substituting Force for the partial derivative of k with respect to time and velocity for the partial derivative of x with respect to time one obtains the BTE in its more conventional form, equation (2.26).

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{F}{\hbar}.\frac{\partial f}{\partial k} + v.\frac{\partial f}{\partial x} = 0 \tag{2.26}$$

---

[6] This has already been mentioned in the introduction chapter and is expanded further in the non-equilibrium section later in this chapter

From classical mechanics the distribution function $f(x,k,t)$ is given by the Fermi-function (electrons are Fermions), and taking the first moment of the BTE this distribution can be shown to give the classical current equation, *see Appendix B*.

$$J = \mu n \nabla E_F \tag{2.27}$$

Evaluating the second and third moments of the BTE give the momentum (2.28) and energy conservation (2.29) equations, [Bløtekjær, 1970].

$$\frac{\partial p}{\partial t} + \nabla.(vp) = qnE - \nabla(nk_BT) + \left(\frac{\partial p}{\partial t}\right)_C \tag{2.28}$$

$$\frac{\partial W}{\partial t} + \nabla.(vW) = qnv.E - \nabla.(vnk_BT) - \nabla.Q + \left(\frac{\partial W}{\partial t}\right)_C \tag{2.29}$$

where

$$\left(\frac{\partial p}{\partial t}\right)_C = -\frac{m^* nv}{\tau_p(w)} \tag{2.30}$$

$$\left(\frac{\partial W}{\partial t}\right) = -\frac{n(w-1.5k_BT)}{\tau_{w_{ii}}(w_i)} - \frac{nw}{\tau_{w_{ij}}(w_i)} + \frac{nw}{\tau_{w_{ji}}(w_j)} \tag{2.31}$$

Both these equations have fairly straight forward phenomenological explanations based upon Gauss' law, which states that the rate of change of a quantity within a certain volume equals the net rate at which that quantity is flowing across the boundaries of that volume minus the rate at which the quantity is being destroyed within that volume. For example the terms in the energy conservation equation can be described, in order, as the rate of change of energy plus the flux of energy across an arbitrary boundary equals the Joule heating minus the energy lost to electron pressure minus the energy lost to heat

flow to the electron gas minus the energy lost to the lattice within that boundary.

2.7 Origin of the dependent device equations

The principle device variables, electrostatic potential, Fermi-level, momentum and electron energy are supplemented by dependent parameters such as electron and ionised donor density and electron mobility. For example the net charge density, ρ, used in Poisson's equation, equation (2.22), is given by the sum of the charge of the unbound electrons[7] and holes, ionised donors, acceptors and traps. Some of these variables are less important, for example if the material is predominantly n-type then the concentration of holes and ionised acceptors is low, and also if traps are neglected the net charge density reduces to the sum of ionised donors and electrons. These extra variables need to be expressed in terms of the independent variables, giving the following 'dependent' equations.

2.7.1 Ionised donors

The ionised donors are created by dopants such as silicon losing an electron to the conduction band. The empty donor sites then form free states in the band gap of the semiconductor, Figure 2.5, which are positively charged when vacant, and neutral when occupied. The probability of the donor sites being occupied is given from Fermi-Dirac statistics, hence the number of non-ionised donors is given by equation (2.32)

$$n_D = \frac{N_D}{1 + \alpha e^{(E_D - E_F)/k_B T}} \tag{2.32}$$

where E_D is the donor ionisation energy and α is usually taken as 0.5. Here α is

---

[7] The valence band electron's charge is cancelled by the atomic nucleus.

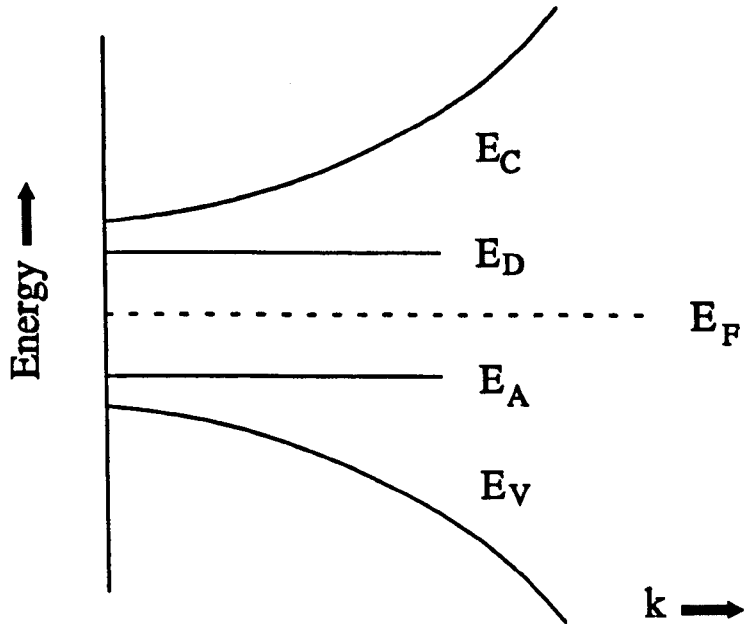


Figure 2.5

Total density of states of a doped semiconductor, showing the conduction band, valence band and the states created by donor and acceptor doping.

introduced because each site at the donor energy can only accept one electron, as opposed to the conduction band where each site can accept two, one electron in each spin state. N_D^+, the ionised donor density is given by N_D - n_D: [Peirret, 1987]

$$N_D^+ = N_D - n_D = \frac{N_D}{1 + 2e^{(E_F - E_D)/k_B T}} \tag{2.33}$$

From this equation it is apparent that the ionised donor density approaches its maximum when the Fermi-level is far below the conduction band edge. E_D, the donor ionisation energy varies as a function of the aluminum mole fraction. Equation (2.34) is used for the donor ionisation energy - conduction band minimum separation (measured in mev), [Adachi, 1985].

$$E_D = 5.2 + 7.9x + 7.1x^2 \tag{2.34}$$

2.7.2 Electron density

It is evident from the conduction band profile shown in Figure 2.4, that in highly doped devices the band bending at the heterojunction coupled with the sharp change in conduction band edge produces an approximately triangular potential well that can approach atomic dimensions. When events occur that are smaller than or equal to the size of the electron's wavepacket, the de-Broglie or thermal wavelength, classical approximations are no longer valid, and quantum effects have to be taken explicitly into account. From quantum theory the electron density is given by the square of the electron wavefunction, equation (2.35).

$$n(x,t) = \langle \Psi^*(r,t)\Psi(r,t) \rangle \tag{2.35}$$

where the angle brackets denote an average over the entire ensemble of electrons and the electron wavefunction, $\Psi(r,t)$, is given from the solution of Schrödinger's equation, equation (2.36).

$$i\hbar\frac{\partial}{\partial t}\Psi_0(r,t) = T(r,t)\Psi_0(r,t) + U(r,t)\Psi_0(r,t) \tag{2.36}$$

Here $T(r,t)$ and $U(r,t)$ are the kinetic and total potential energy operators respectively. The kinetic energy is simply given from the classical description, equation (2.37)

$$E_K = \frac{p^2}{2m_o} \tag{2.37}$$

where the momentum is usually converted into the $-i\hbar\nabla$ operator [Eisberg, 1985].

Potential energy on the other hand is rather more complicated, being formed from the sum of both microscopic and macroscopic components. The latter corresponds to macroscopic electric fields resulting from any externally applied biases and any bulk charge displacement, U_E in Figure 2.6. The microscopic term however, is more

complicated, being formed from the local electric fields derived from the lattice ions. This term can be further separated into two components, the first U_L corresponding to the potential of a perfectly periodic lattice, with the ions in their equilibrium positions, and the second U_S, a random scattering potential introduced to include the lattice ions motion.

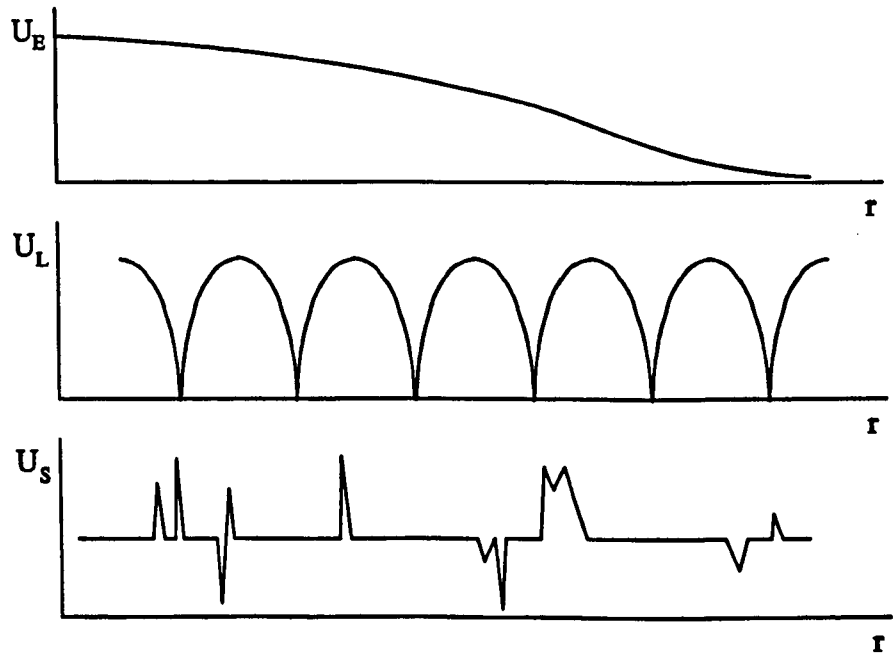


Figure 2.6

The three components of the potential energy of an electron, after [Datta, 1989]

If the temporal and spatial components of wavefunction are separated, equation (2.38), the two sides of Schrödinger's equation can be evaluated separately.

$$\Psi(r,t) = \Phi(t)\,\xi(r) \tag{2.38}$$

Which when substituted into the electron density equation (2.35) gives equation (2.39)

$$n = \sum_m \sum_n \left\langle \Phi_m^*(t)\,\Phi_n(t) \right\rangle \xi_m^*(r)\,\xi_n(r) \tag{2.39}$$

Solving the temporal components of the Schrödinger equation involves the left hand side of equation (2.36), producing a first order differential equation. Assuming the total energy to be constant with time gives equation (2.40)

$$\Phi_m(t) = e^{-iE_m(k)t/\hbar} \tag{2.40}$$

In device modelling the solutions are typically sought over a long time scale compared with the quantum fluctuations which has the result that the summation over the entire ensemble of electrons, under equilibrium conditions, produces the familiar Fermi function, equation (2.41), [Datta, 1989].

$$\left\langle \Phi_m^*(t)\, \Phi_n(t) \right\rangle = \begin{cases} f_0(E_m) & , \quad m = n \\ 0 & , \quad m \neq n \end{cases} \tag{2.41}$$

Substituting this function into equation (2.39) the electron density equation becomes

$$n = \sum_m f_0(E_m)\, \xi_m^*(r)\, \xi_m(r) \tag{2.42}$$

Consequently a similarity with the classical analysis can be drawn [Dekker, 1958], with the summation over m replacing the integral over energy and the $\xi_m^*(r)\xi_m(r)$ product corresponding to the density of states.

The periodic nature of the potential modifies the free particle wavefunction, producing Bloch state spatial wavefunctions.

$$\xi(r) = u_k(r)\, e^{ik.r} \tag{2.43}$$

where $u_k(r)$ is a function of the crystal having the same periodicity as the lattice. When

these wavefunctions are used in the Schrödinger equation it can be manipulated, with the lattice potential, and hence the $u_k(r)$ part of the wavefunction taken into the kinetic energy term, producing the effective mass Schrödinger equation, equation (2.44).

$$-\frac{\hbar^2}{2m^*}\nabla^2\xi_m + (U_E(r) - \lambda_m)\xi_m = 0 \qquad (2.44)$$

where m^* is the effective mass. Here the energy dispersion curves for the perfectly periodic lattice potential, Figure 2.7, are analyzed, and the effective mass deduced.

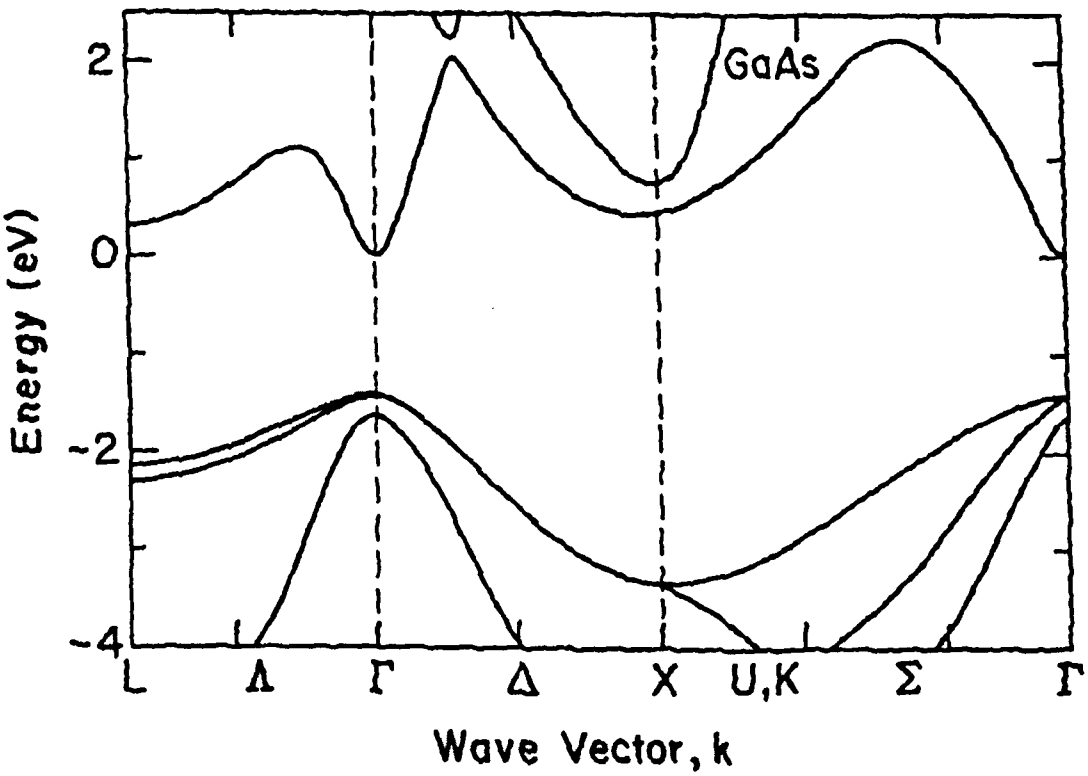


Figure 2.7

Energy dispersion curves for GaAs, [Morkoç, 1991]

Fortunately the energy band diagram is nearly parabolic at its minima, producing a constant effective mass with respect to electron energy. Thus the true wavefunction is formed from a product of the periodic wavefunction associated with the lattice and an

'envelope' function derived from the solution of the effective mass Schrödinger equation.

2.7.3 Electron density at the classical limit

At the classical limit, with flat energy bands, the electron envelope wavefunctions take the form of travelling waves, equation (2.45).

$$\xi_m(r) = \frac{e^{ik.r}}{\sqrt{\Omega}} \tag{2.45}$$

where Ω is the total volume of the domain under consideration.

The sum over k-states in the electron density equation (2.42) can be converted to an integral, producing a density of states that is proportional to the square root of the electron's energy. Thus the total number of electrons is given by a Fermi-integral, equation (2.46).

$$n = N_C \frac{2}{\sqrt{\pi}} \mathcal{F}_{\frac{1}{2}}\left(\frac{E_F - E_C}{k_B T}\right) \tag{2.46}$$

Here the conduction band edge - Fermi-level separation has been substituted for electron energy. It is instructive to note that both the classical and quantum mechanical treatments produce the same result, in accordance with the correspondence principle.

2.7.4 Electron density at the quantum limit

If one considers a simple one-dimensional structure comprising of two layers of different materials, and there is a change in the electron's potential energy over a scale

comparable to the deBroglie wavelength at the interface, then the electrons wavefunctions can no longer described by the solutions obtained at the classical limit. Here the wavefunction is again split into it's temporal and spatial components, with the former again producing the Fermi-function. In this case however, the spatial wavefunctions can be split further, into components parallel and perpendicular to the interface. The parallel wavefunctions are unbound, and hence can be written as free electron wavefunctions, but the normal components (in the y-direction) now have to be calculated explicitly from the effective mass Schrödinger equation solved across the heterojunction.

$$-\frac{\hbar^2}{2}\frac{\partial}{\partial y}\left(\frac{1}{m^*}\frac{\partial \xi_k(\psi,\phi)}{\partial y}\right) + \left(U_E - \lambda_k(\psi,\phi)\right)\xi_k(\psi,\phi) = 0 \qquad (2.47)$$

where U_E, the macroscopic potential energy is comprised of several terms including the electrostatic potential, the energy associated with the heterojunction conduction band discontinuity and the electron-electron exchange correlation energy. This case is evaluated fully in *Appendix A*, and leads to a relatively straight forward derivation of the electron density, equation (2.48).

$$n_{2D}(y) = N_{C_{2D}} \sum_k |\xi_k(y)|^2 \ln\left(1 + e^{(E_F(y)-\lambda_k)/k_B T}\right) \qquad (2.48)$$

where $\xi_k(y)$ and λ_k are the eigenvectors and eigenvalues obtained from the effective mass Schrödinger equation, [Datta, 1989], [Stern, 1972] and [Vinter, 1984].

Comparing the quantum electron density, equation (2.48), with that at the classical limit, equation (2.46), it is apparent that one of the principle effects of introducing quantum mechanics is to modify the density of states term, which changes from a parabolic curve, $g(E) \propto E^{1/2}$, to a spatially dependent term proportional to the square of the electron wavefunction, Figure 2.8.

All the above arguments can be applied in the general case, solving the full Schrödinger equation. However it is not clear, due to the changing band structure, whether using the effective mass Schrödinger equation is valid across a heterojunction. The original effective mass Schrödinger equation was derived for the bulk case, with a single semiconductor material, and hence the effective mass was outside the spatial partial derivative. Morrow and Brownstein [Morrow, 1993] however have shown that the effective mass equation can be applied to heterostructures provided the kinetic energy operator has the form of equation (2.49).

$$T = \frac{1}{2}\left(m^{\alpha} p m^{\beta} p m^{\alpha}\right) \tag{2.49}$$

where p is momentum and m is the effective mass and where α and β are given by equation (2.50).

$$2\alpha + \beta = -1 \tag{2.50}$$

The kinetic energy operator used in equation (2.44) satisfies the above criterion, and can be derived from more phenomelogical arguments, namely keeping the electron envelope wavefunction and its first derivative continuous at the interface. This operator is also Hermitian, and it can be shown that if quantum transport were explicitly included then current continuity would be guaranteed. Unfortunately a full solution of Schrödinger's equation under non-equilibrium conditions is not trivial, requiring the boundary conditions to be defined at the contacts, [Yablik, 1989]. Thus a simplified scheme is used which is described more fully in the Chapter 5.

The case described above creates a sub-band structure, splitting the conventional conduction band in the 2DEG potential well. This creates conceptual problems when describing the effective mass in such systems.

2.7.5 Derivation of the effective mass

The effective mass parallel to the heterojunction can be calculated from the energy dispersion curve in this direction. In the classical limit, near the conduction band minimum, the energy dispersion curve is approximately parabolic with wave vector, given by equation (2.51), [Chambers, 1990].

$$E = E_{C0} + \alpha\left(k_x^2 + k_y^2 + k_z^2\right) \tag{2.51}$$

where α is a constant. The effective mass is given by equation (2.52).

$$m_{xx}^{*} = \frac{1}{\hbar^2}\frac{\partial^2 k_x}{\partial E^2} \tag{2.52}$$

Thus, differentiating equation (2.52) twice with respect to k_x and inverting one obtains a constant value, irrespective of the value of k_y and k_z. When Schrödinger's equation is solved the density of states in the y-direction becomes quantised. This forms a step function, which intersects the classical isotropic energy dispersion curve at discrete values of k_y. This produces a set of energy dispersion curves, one for each sub-band, which are described by the same parabola as the classical conduction band minima, simply shifted up the energy axis, Figure 2.8. One can conclude that the effective mass of each of the sub-bands parallel to the heterojunction is the same as the classical conduction band effective mass.

However, the density of states formed in the y-direction of this system causes conceptual problems when the effective mass is considered. The usual description of effective mass, given in equation (2.52) predicts either zero or an infinite mass[8] depending on the electrons energy. This is clearly inconsistent with the constant value used in Schrödinger's equation used to derive the sub-band structure. On the other hand an

---

[8] The discretization that occurs in this direction limits the wavevectors to certain values, forming delta functions in energy. Hence the derivative of E with respect to k_y is zero.

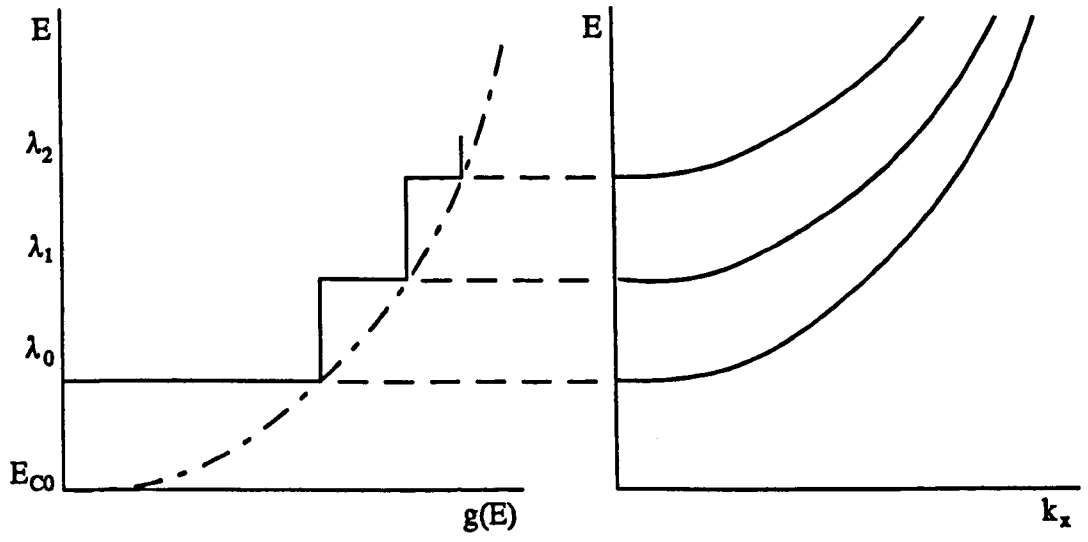


Figure 2.8

Two-dimensional density of states, and the corresponding sub-band dispersion curves in the k_x and k_z directions.

infinite effective mass will produce the quantised bi-dimensional motion for this structure that many researchers predict. One must conclude that the definition of effective mass in such systems is not as straight forward as equation (2.52). For bulk systems the effective mass was originally derived from the periodic nature of the crystal, and hence my only explanation is that the electron has the normal effective mass at certain energies, defined from the Schrödinger equation, and is effectively infinite elsewhere. This interpretation is used in the subsequent device models, where the Schrödinger equation is solved using the normal classical effective mass but the current equations modified. If the separation between consecutive bound state energies is larger than the thermal energy then charge transport is restricted to two dimensions, those parallel to the heterojunction. This effect is not discussed in most of the literature concerning the self-consistent solution of the effective mass Schrödinger equation, with all the afore mentioned papers on this subject using the straight forward classical value for this parameter.

2.7.6 Quasi-continuous conduction band edge

The potential well associated with most HFET structures is roughly triangular in shape, gradually curving to the equilibrium conduction band edge positions. As this well widens the separation between consecutive sub-bands decreases and the two-dimensional density of states approaches the classical parabolic curve. In an effort to again reduce the computational demands only the bound states up to this point are calculated with the contribution to the total electron density beyond this point approximated by the classical Fermi-integral, extending from a 'quasi-continuous conduction band edge to infinity, *see Chapter 3*. The quasi-conduction band edge is defined from the intersection of the real conduction band edge and the energy at which the separation between consecutive bound state drops below the thermal energy. The quasi-continuous conduction band is illustrated as the hatched region in Figure 2.9.

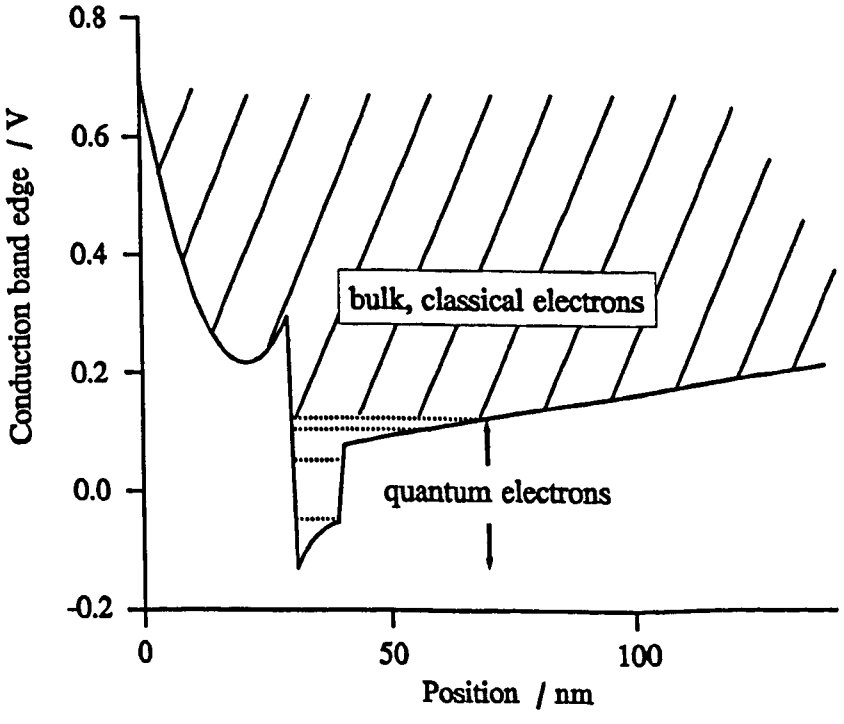


Figure 2.9

Separation of the quantum and classical electron density

2.8 Non-equilibrium conditions

When the device or structure is biased the system is perturbed away from equilibrium. Currents will flow in an attempt to bring the system back to equilibrium once again. Classically current can be described by equations of the form in equation (2.53).

$$J = qnv \tag{2.53}$$

where n is the electron density and v the mean velocity.

The electron density in these conditions can be approximated by the equilibrium distributions, although strictly they should be derived from non-equilibrium statistics, typically a displaced Maxwellian distribution, and hence the equations for the electron density derived in the earlier sections in this chapter can be used. The mean velocity causes more problems, being a complex function of time and energy. The conventional hydrodynamic approach is to obtain v from an approximate solution of Boltzmann's transport equation. As discussed in the introduction, the Boltzmann transport equation is inappropriate at these dimensions, and a rigorous analysis would involve solving a transport scheme based firmly in the quantum regime. Several theories have been developed, discussed more fully in [Wang, 1990]. These schemes are non-trivial, not fully developed and do not follow the same strategy. The desire to produce a modelling scheme usable by device engineers, and thus one that is comparatively computationally inexpensive whilst maintaining accuracy has lead to some simplifications. A similar scheme to the classical description is used for the heterojunction system, but here the equation for mobility is non-isotropic, differing from that used in the classical case.

When the dynamics of an electron's motion within the 2DEG are considered it is apparent that when the electron moves in a direction perpendicular to the interface it must exchange energy with the system, gaining or losing energy as work is done by or on the electric field. This change in energy is transferred to the kinetic energy of the particle's motion in this direction. From the density of states diagram, Figure 2.8, Figure 2.9, it is evident that in the direction normal to the heterojunction there are no vacant energy states available for the particle to move into and hence this

event is forbidden. Consequently the particle is not allowed to move[9] in this direction corresponding to 'zero' mobility. The two-dimensional well that forms at a heterojunction is roughly triangular in shape, and as the width of the well increases with energy the separation between consecutive bound states decreases. Once this separation is smaller than $k_B T$, the thermal energy, the electrons are again permitted to move in this direction as they now have enough energy to transfer to the vacant states. The correspondence principle is again confirmed, with the quantum mechanical solution tending to that described in the classical limit when the scale length of the potential energy increases beyond the debroglie length.

However, in the direction parallel to the heterojunction there is no such sub-band structure and hence electrons travelling in this direction are allowed to scatter into new states having a different energy, and thus are allowed to move. The mobility in this direction therefore possesses a more conventional shape.

In summary, when the scale length of the electron potential energy is less than the debroglie length one can conclude that the electron mobility is no longer isotropic, but is separated into two components, one perpendicular and one parallel to the heterojunction. The normal component of mobility is then zero for electrons associated with sub-band energy levels whose separation from neighbouring bound state energy levels is less than $k_B T$, the thermal energy, and the parallel component given from the two-dimensional mobility discussed in the following section. Once the bound state energy separation is less than the thermal energy full three-dimensional motion is regained and the mobility in all directions is given by the conventional three-dimensional mobility equation. The total electron density in such systems is therefore separated into two components; 'quantised' electrons, restricted to motion parallel to the heterojunction, and 'bulk' electrons possessing all three degrees of freedom. Current normal to the heterojunction is calculated from the conventional current equation, with the exception that only the three-dimensional bulk, electron density is used. Current parallel to the interface includes all the electrons, both two- and three-dimensional, however the mobility associated with each type is different.

---

[9] This explains why the valence band electrons do not participate in charge transport as their energy bands are essentially full and thus there are very few vacant states for these electrons to scatter into.

2.9 Conventional solutions of the Boltzmann Transport Equation

A full solution of the Boltzmann transport equation is non-trivial, even when at the classical limit, where various simplifying assumptions can be made. Although some work was done on the solution of a full non-equilibrium transport scheme it was found that the solution of the energy conservation equation in a scheme including Schrödinger's equation was very difficult. This model was not robust, and failed to converge for any drain-source bias in excess of 1 volt. Consequently the problem was simplified further, implementing the drift-diffusion approximation to describe charge transport. Although not satisfactory for an accurate device model with sub-micron features, the drift-diffusion approximation allowed the solution of Schrödinger's equation within a two-dimensional device model to be investigated and it was hoped that when the problems associated with this scheme were resolved the model could be extended to include non-equilibrium transport effects.

In the drift-diffusion approximation carrier heating is neglected, and assuming that the carriers are at the same temperature as the lattice, then the electron velocity can be expressed as the product of mobility and the gradient of the Fermi-level, equation (2.54), *see Appendix B*

$$v = \frac{\mu \nabla E_F}{q} = \mu \nabla \phi \qquad (2.54)$$

where ϕ is the Fermi-potential and μ, the electron mobility, is a function of electric field.

When the separation of the carriers is performed, splitting them into their two- and three-dimensional components, the currents in the x- and y-directions take on different forms, equations (2.55) and (2.56) respectively.

$$J_x = q\left(\mu_{2D} n_{2D} + \mu_{3D} n_{3D}\right) \frac{\partial \phi}{\partial x} \tag{2.55}$$

$$J_y = q\,\mu_{3D} n_{3D} \frac{\partial \phi}{\partial y} \tag{2.56}$$

2.10 Electron mobility

The problem thus reduces to finding μ, the electron mobility in terms of ϕ, ψ and $\nabla\psi$, the electrostatic and Fermi potentials and the electric field respectively. This term is formed from a combination of various scattering mechanisms, Figure 2.10, including

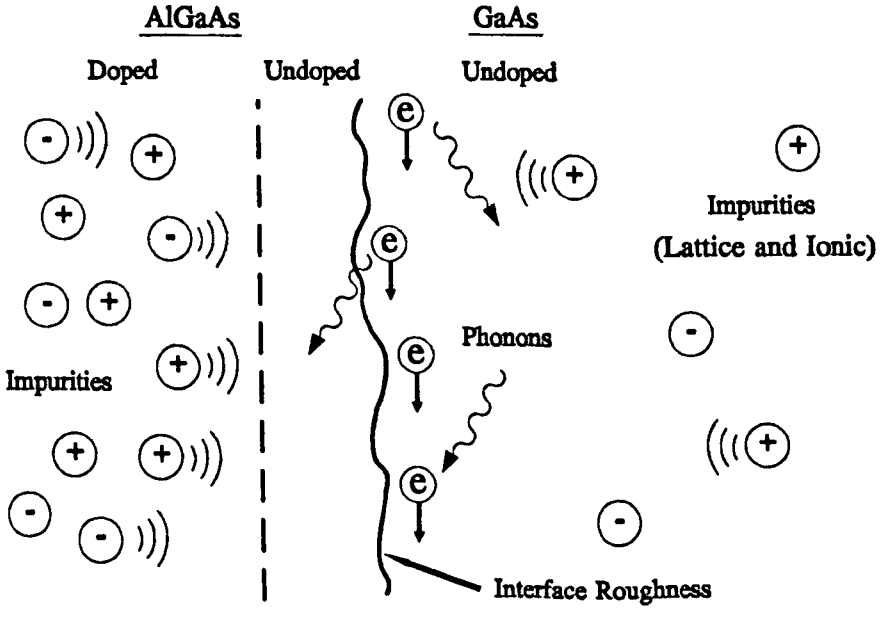


Figure 2.10

Figure of the various scattering mechanisms encountered by the electrons

ionic impurity and phonon absorption and emission scattering. An increase in the number of ionised donors increases the probability of the electrons being scattered by this mechanism, and thus reduces the overall mobility. This effect is reduced within a

heterostructure device by taking advantage of their ability to spatially transfer electrons from one region to another. Thus by modulating the doping profile, with a high doping density in the wide band gap material the electrons can be positioned in a region of low ionised donor density. Thus the electrons are physically separated from their parent ions, creating a 2DEG with a high carrier density in a region of high mobility. This effect is only prominent in the low field characteristics of the electron velocity curve, as other scattering mechanisms, including polar optical [Morkoç], become dominant in high field transport, illustrated in the mobility curves shown in Chapter 3.

2.11 Two-dimensional electron mobility

Electrons confined within the 2DEG are subject to a different set of scattering events to those in the bulk material, and hence one would expect a different mobility. Since these electrons are restricted to two-dimensional motion they are subject to any scattering event that involves transport normal to the heterojunction interface. The two-dimensional mobility is further complicated by the sub-band structure present in the potential well. In the bulk material the 'negative differential resistance' displayed by GaAs and InGaAs is attributed to the transfer of electrons from a low energy band, Γ, the higher energy and lower mobility L and X bands. This transfer can also occur between the electron sub-bands of the 2DEG system, but only if there is a low electron density, where the upper levels are relatively unpopulated and hence permit vacant states for the electrons to scatter into. This effect is reported by Störmer *et al.* where a Schottky gate system was used to vary the population of the 1st and 2nd sub-bands [Störmer, 1981]. A decrease in the electron mobility was demonstrated once the second sub-band began to be populated corresponding to the onset of inter sub-band scattering. However if the 2DEG density rises such that the upper sub-bands fill up and this scattering process is reduced. This reduction in the inter sub-band scattering is not expected to occur in the bulk GaAs since the separation of the Γ, L and X bands are at least an order of magnitude greater than that of the energy levels in the sub-band structure, and hence would require exceedingly high electron densities to significantly fill the upper bands in this system.

Although there have been several calculations of the low field two-dimensional mobility and its variation with carrier density and lattice temperature [Drummond, 1981], [Hirakawa, 1986] and [Luo, 1988] there has been little work on the field dependence of this variable. Most workers use the temperature dependence of the low field mobility, fitting the 'temperature' with an equivalent electron temperature. This assumes the electron density to be described by an equilibrium-type distribution function, either Boltzmann or Fermi-Dirac. One author and his coworkers [Lei, 1986a] and [Lei, 1986b] have investigated the mobility/field relationship using a non-Boltzmann balance equation. Here the first two sub-bands are approximated by the Fang-Howard-Stern variational method and used to calculate an electron mobility that includes an inter sub-band scattering term.

The case of an HFET is more complicated still since the number of sub-bands and their energy levels and wavefunctions vary along the channel. Thus the field dependent two-dimensional mobility curves would also vary along the channel. One possible solution to this problem would be to model the mobility with two parameters, field and the electron density. This would produce a different mobility under the gate from the rest of the device due to the reduced carrier density in this region. Perhaps the only rigorous solution would be to calculate the electron mobility in situ, as part of the FET model itself. In this scheme the 'true' electron wavefunctions, energy levels and populations could be used to produce a local mobility. This scheme would necessarily be very computer intensive, probably increasing the run time by one or two orders of magnitude. This would be unsuitable for CAD purposes and consequently with present computational limits the 'two-parameter' electron mobility is probably a good compromise between speed and accuracy.

Chapter 3

Numerical Modelling of Heterojunction Field Effect Transistors

3.1 Introduction

The numerical modelling of semiconductor devices requires the physical equations outlined in Chapter 2 to be expressed in terms transferable to computer code. This chapter describes the general numerical schemes used to discretize, scale, evaluate and solve these equations. The primary device equations are obtained from Maxwell's equations and the Boltzmann transport equation:

$$\nabla.(\epsilon \nabla\psi) + q\left(N_D^+ - n\right) = 0 \tag{3.1}$$

$$\frac{\partial n}{\partial t} + \frac{1}{q}\nabla.J = 0 \tag{3.2}$$

$$\frac{\partial p}{\partial t} + \nabla.(vp) = qnE - \nabla(nk_BT) + \left(\frac{\partial p}{\partial t}\right)_C \tag{3.3}$$

$$\frac{\partial W}{\partial t} + \nabla.(vW) = qnv.E - \nabla.(vnk_BT) - \nabla.Q + \left(\frac{\partial W}{\partial t}\right)_C \tag{3.4}$$

The electron density is formed from the sum of the two- and three-dimensional components

$$n = n_{2D} + n_{3D} \tag{3.5}$$

where

$$n_{2D}(y) = N_{C_{2D}} \sum_k |\xi_k(y)|^2 \ln(1 + e^{(E_F(y) - \lambda_k)/k_B T}) \tag{3.6}$$

$$-\frac{\hbar^2}{2} \frac{\partial}{\partial y}\left(\frac{1}{m^*} \frac{\partial \xi_k(\psi,\phi)}{\partial y}\right) + (U_E - \lambda_k(\psi,\phi))\,\xi_k(\psi,\phi) = 0 \tag{3.7}$$

and

$$n_{3D} = N_{C_{3D}} \mathcal{F}_{\frac{1}{2}}\left(\frac{E_F - E_C'}{k_B T}\right) \tag{3.8}$$

Finally the other dependent variables are given by equations (3.9) and (3.10).

$$N_D^+ = N_D - n_D = \frac{N_D}{1 + 2e^{(E_F - E_D)/k_B T}} \tag{3.9}$$

$$\mu = \mu(w,t) \tag{3.10}$$

Although these equations comprise a full description of carrier dynamics in an HFET they are not trivial to solve, even if the quantum treatment of the electron density is neglected. For this reason the full hydrodynamic treatment of carrier dynamics is omitted and the older and simpler drift-diffusion approximation used instead. This

approach is strictly only valid for long gate length devices, exactly the opposite to what is generally produced, but greatly reduces the complexity of the numerical schemes, halving the number of independent variables. Once the effects of explicitly including quantum mechanics into a simple device models have been investigated, and satisfactory numerical schemes produced that are both accurate and robust they may be included in a more comprehensive device simulations involving the momentum and energy conservation equations. The drift-diffusion approximation neglects the energy transport equation, or at least omits the Joule-heating term, setting the carrier temperature to that of the lattice. The second term in the momentum conservation equation is also assumed to be small with respect to the right hand side, and thus the electron velocity can be expressed in terms of the Fermi-level, equation (3.11)

$$v = \mu \frac{\nabla E_F}{q} \tag{3.11}$$

With the momentum and energy conservation equations solved the problem is then reduced to solving Poisson's and the current continuity equations self-consistently.

3.2 Choice of independent variables

Most conventional drift-diffusion models are based upon a Boltzmann statistics description for the electron density, rewriting the equation for velocity, equation (3.11), in terms of field and the gradient of electron density[1], equation (3.12)

[1] Hence the name: drift (field) and diffusion (∇n)

$$v = \mu E - D\frac{\nabla n}{n} \tag{3.12}$$

where D is the electron diffusivity. The Einstein relationship [Selberherr, 1984] is usually assumed and D is written in terms of μ, the electron mobility, equation (3.13).

$$D = \frac{\mu k_B T}{q} \tag{3.13}$$

This permits the simulation designers to choose ψ and n, the electrostatic potential and electron density respectively, to be independent variables. This is the preferred set because although the overall self-consistent solution is non-linear, each separate device equation is linear in at least one variable and hence easily soluble requiring only one matrix inversion. For example conventional algorithms usually follow a simple Gummel scheme solving first for potential using Poisson's equation, assuming the charge density to be fixed, and then for electron density using the current continuity equation assuming the mobility to be fixed. However this does have disadvantages, particularly in the solution of the current continuity equation. Probably the simplest and most common scheme uses the semi-implicit time discretization, equation (3.14).

$$\frac{n^{k+1} - n^k}{\Delta t} = \frac{\nabla . J^{k+1} + \nabla . J^k}{2} \tag{3.14}$$

Even in this method, which calculates the current at the $k+1^{th}$ time step, the time increment has to be kept small to guarantee stability. The maximum permissible increment to maintain stability is given by equation (3.15), [Selberherr, 1984]

$$\Delta t < \min\left(\frac{2\lambda^2}{\mu n}\right) \tag{3.15}$$

Difficulties also occur in the discretization of this equation. The partial derivative is commonly evaluated by taking the difference of the currents at the half-node positions. The electron density however varies exponentially with position, and hence a simple algebraic average is not appropriate to find the half-node values of this variable. The linear differencing is replaced by one that varies exponentially with position, the Scharfetter-Gummel differencing scheme [Scharfetter, 1969].

Unfortunately when quantum mechanics proper are introduced the problem in this form is no longer tractable for several reasons. The first is that the electron density is no longer a simple exponential function of energy and hence the Scharfetter-Gummel scheme is not applicable. Secondly the solution of Schrödinger's equation requires the Fermi-level to be determined. Whilst it may not be impossible to derive E_F from the independent variables ψ and n, it does require the inversion of equation (3.6). This approach is complex, lengthy and does not produce an elegant solution[2], consequently an alternative approach is used, changing the independent variables.

If the electron density is replaced by the Fermi-potential as an independent variable the scheme immediately becomes more tractable. Working with the set $\{\psi, \phi\}$ instead of $\{\psi, n\}$ has the advantage that all dependent variables, ionised donor density, electron density and mobility[3], are simple functions of ψ and ϕ. This has the added bonuses that the electron density can never be less that zero, and also because both independent variables are linear functions of position, and do not vary exponentially like the electron density, their internodal values are calculable from a

---

[2] One method proposed for effectively inverting the two-dimensional electron density equation uses an iterative scheme, re-solving the electron density equation for different Fermi-levels until the output matches the true electron density. This scheme is inelegant and time consuming since it requires Schrödinger's equation to be solved many times, once for each new guess at the Fermi-level.

[3] In the drift-diffusion problem the electron mobility is derived as a simple function of electric field.

simple linear combination of the values at the nodes, equation (3.16)

$$f|_{i+\alpha} = (1-\alpha) f_i + \alpha f_{i+1} \qquad : \quad 0 < \alpha < 1 \tag{3.16}$$

One of the main disadvantages of this choice of variables is that the principle device equations, Poisson's and current continuity, are no longer linear in the independent variables, in fact quite the opposite, as they now depend exponentially upon both ψ and ϕ. This means that the simple solution scheme used in conventional models is not applicable, at least not without an inner iterative loop within the Gummel scheme to solve each equation. An alternative solution scheme is therefore necessary, with the standard for non-linear systems, Newton-Raphson, chosen for this work [Kurata, 1982], [Press, 1989]. This solution method is more complicated than the Gummel scheme, requiring the evaluation of the Jacobian, a matrix formed from the partial derivatives of each equation with respect to each variable, but has the advantage of quadratic convergence as opposed to the logarithmic convergence of Gummel's scheme. A further disadvantage, because the $\{\psi, \phi\}$ set are a relatively unpopular choice for independent variables, it that there is very little literature published on the solution of the device equations in this form.

3.3 Discretization

The principle device variables, ψ and ϕ, are solved self-consistently by sampling their values at different positions within the device domain and linking them to one another via the principal device equations, Poisson's and current continuity. Sampling the variables at set positions requires the spatial derivatives in all the device equations to be discretized, which can be performed by several schemes, the most common being finite difference, finite element and their variations. In general all discretization methods can be used to solve a problem, but in practice some schemes

are more applicable to certain problems. For example, in models with irregular boundaries the finite element discretization is usual as this scheme can easily adapt to non-rectangular shapes. The discretization chosen in this work follows the well tested central-difference scheme which is conceptually simple and thus avoids any unnecessary numerical complications. All finite difference schemes can be derived from a Taylor expansion of the variables as follows. Consider a uniform grid with an arbitrary variable, a, that is a function of position. We have, by expanding $a(x+h)$

$$a(x+h) = a(x) + h\frac{\partial a(x)}{\partial x}\bigg|_x + \frac{h^2}{2!}\frac{\partial^2 a}{\partial x^2}\bigg|_x + O(h^3) \qquad (3.17)$$

where $O(h^3)$ is the sum of all components of third order and greater. Combining with a similar expression for $a(x-h)$ one can derive the central difference equations, which to second order accuracy are given by equations (3.18) and (3.19)

$$\frac{\partial a}{\partial x}\bigg|_x \sim \frac{a(x+h) - a(x-h)}{2.h} \qquad (3.18)$$

$$\frac{\partial^2 a(x)}{\partial x^2}\bigg|_x \sim \frac{a(x+h) - 2.a(x) + a(x-h)}{h^2} \qquad (3.19)$$

When discretized variables are used the expansions for $a(x+h)$ are laid over a mesh of grid points. The terms $a(x)$, $a(x+h)$ and $a(x-h)$ are replaced by a_i, a_{i+1} and a_{i-1} respectively, where h is the internodal separation and i the nodal index.

3.4 Mesh refinement

In order to model a device accurately the discretization error that occurs due to the

finite number of samples must be minimised, which in turn means that the distance between consecutive nodes must be small. In regions where there is little change in the variables the discretization error is small and hence it is wasteful in both computational time and memory to have a high node density in these areas; the gain in accuracy is minimal compared with the expense of solving the extra equations. A more efficient scheme will cluster the grid nodes in the regions where rapid changes occur in the variables, typically the electrostatic potential, and reduce the node density elsewhere. Unfortunately a non-uniform grid has incomplete cancellation of the second order terms in the Taylor expansions of the variables and thus the derivatives will not quite be second order in accuracy, but this is compensated for by the reduction in the sampling error in these regions. Further the central finite-difference equations have to be modified to work on non-uniform grids, taking account of the varying internodal separation, and thus producing equations (3.20) and (3.21), [Snowden, 1988].

$$\left.\frac{\partial a}{\partial x}\right|_x = \frac{(a_{i+1} - a_{i-1})}{(h_i + h_{i-1})} \tag{3.20}$$

$$\left.\frac{\partial^2 a}{\partial x^2}\right|_x = \frac{2a_{i+1}}{h_i.(h_i + h_i - 1)} - \frac{2a_i}{h_i.h_{i-1}} + \frac{2a_{i-1}}{h_{i-1}.(h_i + h_{i-1})} \tag{3.21}$$

A scheme is employed to fit the nodes, increasing their density in regions where rapid changes occur. Conventional mesh refinement schemes add and/or delete lines of nodes at certain positions, fitting the internodal separation to a variable. To accommodate the changing number of nodes the models store the device variables in either oversized arrays, larger than the average number of nodes, or incorporate a dynamic memory allocation scheme[4]. The former is wasteful in computer memory,

[4] This method stores each node as a separate entity possessing several attributes, including the values of the device parameters at that point, the internodal positions and pointers to the neighbouring nodes. Thus the addition of a new node only requires the allocation of some free memory and the alteration of the pointers and internodal separations of the adjacent nodes.

and thus is undesirable especially if the program is to run on a personal computer. The latter is more efficient in memory at the expense of generally more complicated code.

In the two-dimensional case however, the discretization, coupled with the solution scheme produces a large sparse set of linear equations. It is computationally very advantageous for the matrix inversion routine to keep the same pattern of zero/non-zero elements during the simulation. This in turn means that the number of nodes and their arrangement in the mesh may not be changed and hence the conventional method of adding and removing lines of nodes is not appropriate. A new method of mesh refinement has been developed where the nodes are rearranged in space, maintaining the same neighbours, but creating high node densities in regions of rapid change. This scheme first interpolates the values of the independent variables ψ and ϕ using a standard cubic spline routine. A fitting curve for the internodal separations is then extracted from one of the device parameters, typically the electric field, which is then used to find the new nodal positions using the scheme illustrated in Figure 3.1.

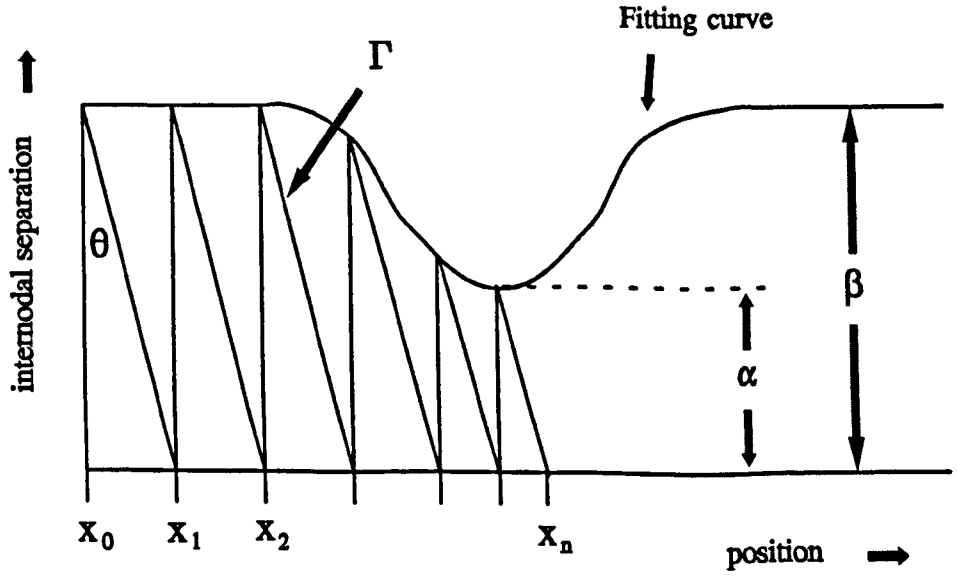


Figure 3.1

Internodal spacing fitting scheme

Here the fitting curve is shown at the top of Figure 3.1. An angle, θ, is chosen and the nodal positions are produced at the intersection of the line marked Γ with the x-axis. The angle θ is then varied, first using a bounding algorithm and then by bisection until the number of steps taken to reach the end of the modelled domain matches the number of nodes. Note the ratio α:β is the ratio of the minimum to maximum mesh size. Once the new mesh has been produced the independent variables are then fitted back to the new nodal positions using the original splined data.

This method produces a non-uniform mesh with nodes clustered around the edges of the contacts, at the surface and at each of the heterojunction interfaces. Figure 3.3 shows a typical non-uniform grid produced by this scheme for a single heterojunction FET.

It was found that the solution of the device equations depends critically upon the aspect ratio of the grid cells, especially in regions of high field. Although the equations were solved, the residual error was less than the convergence limit, the simulation produced results that were definitely non-physical.

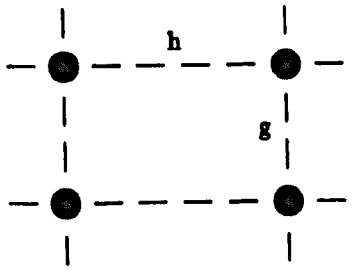


Figure 3.2

Mesh aspect ratio

A simple single heterojunction FET was modelled using a simple drift-diffusion current equation and classical statistics. This device showed a large droop in current as the bias was increased into the saturated region of device operation. There were no terms in the equations to produce such a response, and hence the solution algorithm was investigated. It was found that the simulation was using a grid with a very poor aspect ratio, greater than 20:1, and the solution depended upon the mesh. When the aspect ratio, h:g in was reduced to less than 5:1 the solutions became independent of the mesh producing true physical characteristics, with the current

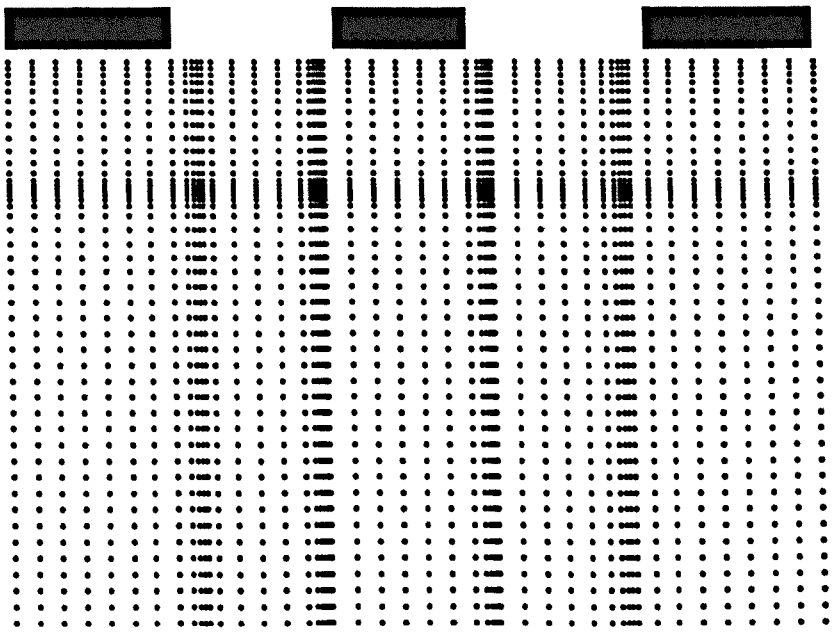


Figure 3.3

A typical non-uniform grid consisting of 60x45 nodes

slowly increasing with bias. This restriction on the aspect ratio is particularly significant in heterostructure devices, were highly doped materials are common. The high doping produces a model domain with a large aspect ratio (active region depth : source-drain separation), and thus requires many more nodes in the x-direction than the y-direction to maintain a cell aspect ratio of less than 5:1. This causes problems since to accurately model the heterojunction a large number of nodes are required in the y-direction. If an even larger number of nodes are required in the x-direction to maintain a good aspect ratio the overall scheme will become very large indeed! In the two-dimensional simulations discussed in Chapter 5 a non-uniform mesh is employed that is only non-uniform in the x-direction. This results in a much lower probability of a poor aspect ratio in a critical region, ie. in the potential well at the drain edge of the gate. With this mesh refinement the grid has the same nodal pattern in the x-direction shown in Figure 3.3, but a uniform internodal separation in the y-direction.

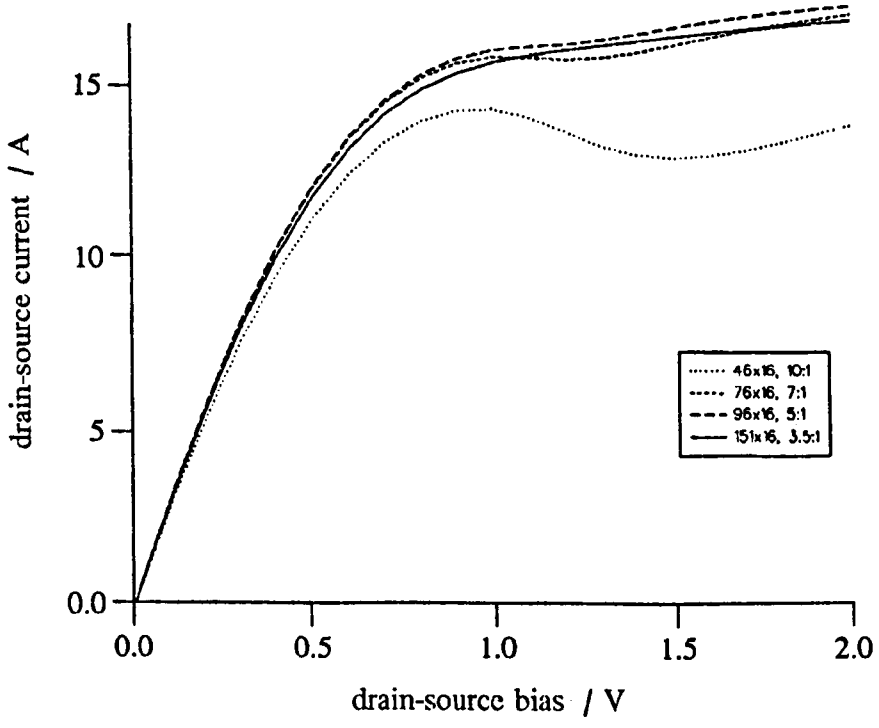


Figure 3.4

I_{DS}-V_{DS} curves for several different mesh configurations

3.5 Scaling of the variables

The conventional argument used for scaling the device variables is based upon changing their absolute values to be of around the same order of magnitude, preferably unity, aiming to reduce the possibility of round-off error. With workstations and modern compiler's providing double and quadruple precision variables, possessing a range of $\pm 10^{300}$ or more the problem of round-off error is not so prominent. In practice little or no change is observed when scaling has been incorporated into the models discussed in chapters 4,5 and 6. However scaling is still employed in the program code to simplify the equations, reducing the number of terms, and consequently omitting a lot of factors at the beginning of the code. This is more significant when the solution process is considered, which in both the one- and two-dimensional models requires the derivative of the device equations with respect to each of the variables. Fewer terms in the equations produces shorter more

compact code with less opportunity arising for programming errors. The variables are scaled according to a modified DeMari scheme, [DeMari, 1968], but where n_i, the intrinsic electron density in the DeMari scheme is replaced by $N_{C\ GaAs}$, the density of states in GaAs, Table 3.1. It is worth noting that this scaling scheme produces dimensionless scaled variables.

Variable	Description	Scaling factor	Value (at 300K)
ψ, ϕ	Potentials	$\frac{q}{k_B T}$	38.67
n, p, N_D^+, N_A^-, N_T^-	Charge densities	$\frac{1}{N_{C\ GaAs}}$	2.30×10^{-24}
x, y, h, g	Positions and internodal separations	$\left(\frac{k_B T \varepsilon_0}{q^2 N_{C\ GaAs}}\right)^{\frac{1}{2}}$	5.52×10^{8}

Table 3.1

Table of the variable scaling factors

3.6 Evaluation of the dependent variables

Some of the dependent variables are not straight forward to obtain, for example the two-dimensional electron density requires the solution of Schrödinger's equation, the three-dimensional electron density has no algebraic solution, and the electron mobility is obtained from Monte-Carlo data has to be characterised. This next section deals with the evaluation of these variables.

3.6.1 Two-dimensional electron density

The electron density is modelled by splitting it into two components, n_{2D} and n_{3D}, as discussed in Chapter 2.

$$n_{tot} = n_{2D} + n_{3D} \tag{3.25}$$

The two-dimensional electron density is obtained from the solution of the one-dimensional effective mass Schrödinger equation, (3.26), *see Appendix A*.

$$-\frac{\hbar^2}{2}\frac{\partial}{\partial y}\left(\frac{1}{m_e^*}\frac{\partial \xi_k(\psi,\phi)}{\partial y}\right) + \left(U_E - \lambda_k(\psi,\phi)\right)\xi_k(\psi,\phi) = 0 \tag{3.26}$$

where U_E is the total potential the electron experiences. Here the Schrödinger equation is solved in each vertical slice, perpendicular to the heterojunction, throughout the whole of the simulation domain. Upon each iteration of the global solution scheme there will be N separate solutions[5] of Schrödinger's equation. Schrödinger's equation is more complicated to solve than most partial differential equations, principally because there is no unique solution. Indeed in the modelling of the heterojunction we desire the first k solutions, until the separation between consecutive bound state energies is less than the thermal energy. The partial derivatives that form the first term in Schrödinger's equation are discretized using central finite differences, in a similar fashion to Poisson's equation, converting the whole equation to a matrix form. Many different techniques have been developed for solving the eigensystems of general matrices, for example Jacobi transformations, Householder reduction and QR/QL algorithms, to name but a few [Press, 1989], but since Schrödinger's equation is only required in one-dimension and its corresponding discretized matrix is tridiagonal we can take advantage of this special form.

---

[5] Where N is the number of nodes in the x-direction

Considering a general eigenvalue problem, equation (3.27)

$$A.x = \lambda x \tag{3.27}$$

it is apparent that if the zero vector is omitted eigenvalues occur when

$$\det |A - \lambda I| = 0 \tag{3.28}$$

where I is the identity matrix.

Thus we can expand out the determinant into an N^{th} degree polynomial and search for the roots. In general this is not a good computational method for finding eigenvalues, [Press, 1989], but in the case of a tridiagonal matrix the characteristic polynomial associated with the determinant is straight forward to evaluate. The roots are therefore found using the bisection method operating upon a Sturm sequence [Wilkinson, 1971]. This counts the number of changes of sign of the characteristic polynomial, and hence the number of roots. Here the upper and lower limits are initially set using Gershgorin's theorem, [Stöer, 1980], and then updated from the previously calculated eigenvalue. It is apparent that unless the system is degenerate the $k+1^{th}$ eigenvalue must be larger that the k^{th}, and thus the lower limit for the bisection algorithm can be set to this value. Unfortunately this does not produce a radical improvement in the speed of the eigenvalue solver as it only reduces the search region by a small amount[6], but since this information is available and there is no penalty in applying it, the lower limit is updated. Once an eigenvalue is computed it is checked for degeneracy, separation from the previous eigenvalue and position with respect to the conduction band edge at the top and bottom surfaces. The first check is to prevent problems in the calculation of the partial derivative of the two-dimensional electron density with respect to the electrostatic potential. This term

[6] Note: a 50% reduction in the search region will only reduce the number of bisections by one.

includes a denominator of the form λ_k - λ_{k-1}, which if the eigenvalues were degenerate would be identically zero, and thus the partial derivative infinite which is clearly non-physical. The second term checks for the quasi-continuous conduction band minimum, which occurs when the separation is less than the thermal energy and terminates the solution of Schrödinger's equation, assuming the electron density beyond this point is approximated by the Fermi-integral. Finally the boundary check is to ensure the solutions of Schrödinger's equation are true bound states, otherwise a quantum mechanical current is possible which is not included in the model. This is therefore accommodated within the scheme by truncating the solution of further 'bound' states and setting the quasi-continuous conduction band minima to this energy as opposed to the energy separation criterion.

If the eigenvalue passes all these tests it is taken as being a true bound state and used for the calculation of the 2DEG density. The eigenvalue is substituted back into Schrödinger's equation and the corresponding eigenfunction solved for using LU decomposition. Again the matrix is tridiagonal allowing a special LU decomposition algorithm to be used.

Part of the total potential, U_E, used in Schrödinger's equation, though not particularly significant is the exchange-correlation energy. This term is associated with 'many-body effects' that accompany the change from a single particle description. The exchange-correlation energy is purely a quantum mechanical effect, and can be approximated by the analytic equation below, equation (3.29) described by [Stern, 1984].

$$V_{xc}(x) = -\left[1 + 0.7734\,\beta \log_e\left(1 + \beta^{-1}\right)\right]\left(\frac{2}{\pi\,\alpha\,r_s}\right)Ry^* \quad (3.29)$$

where

$$\alpha = \left(\frac{4}{9\pi}\right)^{\frac{1}{3}} \quad , \quad \beta = \beta(x) = \frac{r_s(x)}{21} \tag{3.30}$$

$$r_s(x) = \left(\frac{4}{3}\pi a^{*3} n(x)\right)^{-\frac{1}{3}} \quad , \quad a^* = \frac{4\pi \epsilon_0 \epsilon_r \hbar^2}{m^*(x) q^2} \tag{3.31}$$

and Ry^* is the Rydberg constant.

This term is included for completeness as it does not effect the speed and/or robustness of the model even thought it does not significantly alter the solution. Note this term is also relatively inexpensive to calculate in terms of computational time and memory.

Attempts have been made to approximate the solution of the one-dimensional effective mass Schrödinger equation assuming a triangular potential well [Delagebeaudeuf, 1982] and [Widiger, 1985]. This shape has known solutions, taking the form of Airy functions, but is highly dependent upon the choice of empirical fitting parameter, producing worse results than simple Boltzmann statistics if chosen incorrectly [Yoshida, 1986]. Since this scheme is highly empirical, and the solution of Schrödinger's equation is not overwhelmingly more difficult, it was decided to include a fully self-consistent solution of Schrödinger's equation in the device models.

Self-consistent solutions of Schrödinger's equation over semiconductor interfaces were initially performed for the SiO_2/Si junction in silicon MOSFETs [Stern, 1968] and [Stern, 1972]. Here the abrupt change in band gap again produces a roughly triangular potential well parallel to the heterojunction, but in this case, because it cannot penetrate into the insulator, the wavefunction is assumed to tend to zero at the interface instead of at the free surface. There are numerous publications on the self-consistent solution of the one-dimensional effective mass Schrödinger equation over AlGaAs/GaAs heterojunctions, including papers by [Vinter, 1983], [Stern, 1984] and [Shey, 1988]. A good review of the classical Boltzman and Fermi-Dirac statistic

solutions, the triangular well approximation and the full self-consistent solution of Schrödinger equation is given by [Yoshida, 1986].

3.6.2 Three-dimensional electron density

The three-dimensional component of the total electron density is approximated using the Fermi-integral, equation (3.32)

$$n_{3D} = N_{C\,3D}\,\mathcal{F}_{1/2}\left(\frac{E_F - E_C'}{k_B T}\right) \tag{3.32}$$

but integrating from the quasi-continuous conduction band minima as opposed to the classical conduction band edge. The quasi-continuous conduction band edge is formed from the merging of the 'discrete' quantised sub-band structure, which occurs when the separation between consecutive bound state energies is less than the thermal energy. This profile is modelled by taking the intersection of the classical conduction band edge, derived from the electrostatic potential, and the first bound state energy satisfying the above criterion. This is illustrated in Figure 2.9 which shows the classical conduction band edge, the sub-band structure and the quasi-continuous conduction band profile. Also shown is the separation of the total electron density into the two- and three-dimensional components corresponding to 2DEG and bulk electrons.

The Fermi integral is not analytically soluble, and is therefore modelled using a curve fit approximation. Several curve fits have been proposed by different authors, the one chosen in this work, equation (3.33), is described by D. Bednarczyk and J. Bednarczyk [Bednarczyk, 1978] having a relative error of less than ± 0.5% over the whole range of arguments.

$$\mathcal{F}_{1/2}(\eta) = \frac{\frac{\sqrt{\pi}}{2}}{c(\eta) + e^{-\eta}} \tag{3.33}$$

where $c(\eta)$ is given by equation (3.34).

$$c(\eta) = \frac{\frac{3}{4}\sqrt{\pi}}{\left(50 + \eta^4 + 33.6\eta.(1 - 0.68e^{-0.17(\eta+1)^2})\right)^{3/8}} \tag{3.34}$$

This curve is displayed in Figure 3.5 in conjunction with its asymptotic expansions.

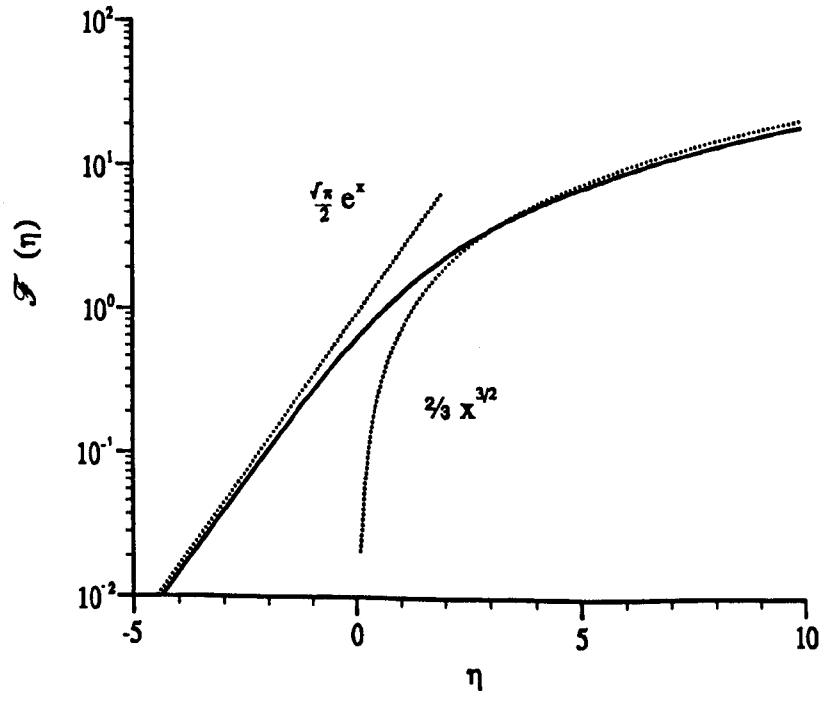


Figure 3.5

Curve fit Fermi-integral of order ½ and its asymptotic expansions

In both the one- and two-dimensional models, bulk electrons only occur when the

quasi-continuous conduction band edge is well above the Fermi-level, thus this term could be approximated using Boltzman statistics instead. However, the overall solution method requires an initial starting point which is provided by solving the equations using the classical approximations, namely the with the electron density described by the Fermi-integral. Thus the function for the Fermi-integral is coded and since there is no penalty in incorporating it into the final code it is used in the simulations.

3.6.3 Ionised Donor density

The ionised donors are also modelled using Fermi-Dirac statistics as described in chapter two. The modelling of this variable is therefore very straight forward, involving only the evaluation of one exponential function and its derivative with respect to energy. The only complication with the evaluation of this variable arises in ensuring that the exponential term does not overflow. This is simply done by capturing values outside the overflow limit. In most cases this is not required since the normal range of energies expected is well within the range of arguments, ie. the conduction band - Fermi-level separation is unlikely to be greater than 7 eV.

3.6.4 Traps

Ionised traps are treated in the same fashion as donors, with the exception that they are negative when filled and neutral when empty. Traps are used on the upper surface of the modelled domain to produce the surface potential. The surface potential is of a similar value as the Schottky 'built-in' bias, approximately 0.5 V, and provides the necessary band bending at this boundary to guarantee the confinement of the electron wavefunction. Hence this parameter allows the boundary conditions for the wavefunction to be defined, $\xi_k \mid_{surf} = 0$. The depth and density of the surface states

is fairly arbitrary, chosen such that they penetrate approximately 10-20 nm and produce a surface potential of the correct value.

3.6.5 Mobility

For simplicity the mobility is modelled using the conventional three-dimensional mobility equations described by [Snowden, 1984].

$$\mu = \frac{300\mu_0}{T_0}\left(\frac{1 + \dfrac{8.5\times10^4 E^3}{\mu_0 E_0^4\left(1 - 5.3\times10^{-4} T_0\right)}}{\left(1 + (E/E_0)^4\right)}\right) \tag{3.35}$$

where

$$\mu_0 = \frac{0.8}{1 + \sqrt{\dfrac{N_D^+}{10^{23}}}} \tag{3.36}$$

and

$$E_0 = 4\times10^5 \tag{3.37}$$

This produces the set of curves shown in Figure 3.6, illustrating the variation of the three-dimensional electron velocity with both field and ionised donor density.

In the previous chapter the two-dimensional mobility was discussed but for simplicity is omitted in the FET models produced as part of this work. This term will be important but to date there is little published data on the form of this parameter and suitable fitting curves. Indeed there may be no unique curve for this parameter as it

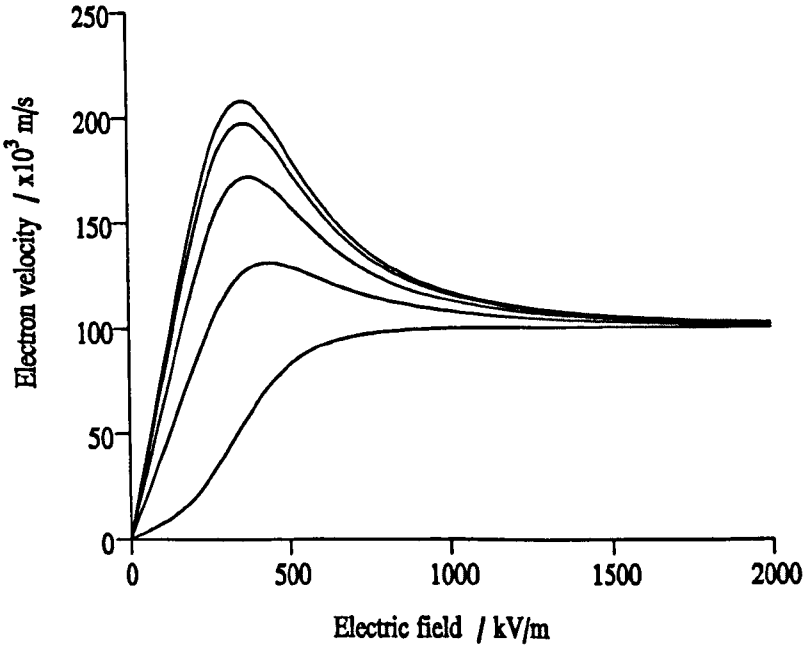


Figure 3.6

Electron velocity versus electric field

N_D+ = 10^{20}, 10^{21}, 10^{22}, 10^{23} and 10^{25} from the top down

is likely to be a function of the electron wavefunction profile and hence will be strongly dependent upon the FET structure and operating bias. It may be the case that to produce a fully self-consistent model the electron mobility will have to be evaluated during the solution process, changing from one iteration to the next, which will have the consequence that the simulation will become more computationally intensive and probably less robust.

3.7 Solution schemes

The discretized device equations using the chosen variables form a set of coupled, highly non-linear equations. The standard solution method for non-linear equations is the Newton-Raphson scheme (abbreviated to Newton scheme). This scheme

approximates the solution by a taking a series of tangents to the curve $f(x)$, and given the correct conditions is able to rapidly home in on the root of a non-linear equation. Preconditions for good convergence include a well behaved function containing no discontinuities in either the curve or its first derivative and a good initial guess. If these criteria are satisfied the scheme possesses quadratic convergence properties typically requiring less than 10-15 iterations to reduce the error below the numerical noise level for these device equations.

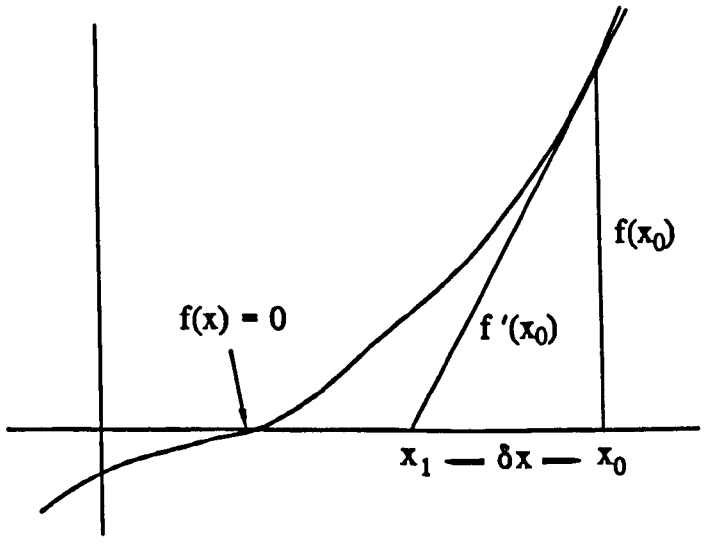


Figure 3.7

One step in the Newton scheme

The basic Newton scheme can be described with the aid of Figure 3.7. Given an initial guess, x_o, an improved solution is found, equation (3.38), at the intersection of the x-axis and the tangent to the curve $f(x_o)$. Once the new guess for the solution is obtained the procedure can be repeated until the convergence criteria are satisfied.

$$x_1 = x_0 - \frac{f(x_0)}{\dot{f}(x_0)} \tag{3.38}$$

There are three common criteria used for terminating the iteration sequence, the first is simply to stop after N steps. This is useful for setting an upper limit on the number of iterations, ie. if the solution scheme takes more than N steps it probably is not going to solve at all. The second method checks for the magnitude of the step taken by the solution vector, equation (3.39).

$$|x_{k+1} - x_k| \leq \epsilon \tag{3.39}$$

Again this method is probably best used as a check for the non-convergence of the scheme. If the solution vector is very small and the error in the residual large then again the equations are probably not going to converge in a reasonable number of iterations. The third and best method is to check for the error in the residual, equation (3.40). This method is infallible, since it is independent of the path taken to the root. The safest scheme will employ this last method as the test for convergence, and the other two methods as checks for non-convergence.

$$|f(x_k)| \leq \alpha \tag{3.40}$$

As stated earlier Newton's scheme requires the curve $f(x)$ and its first derivative $f'(x)$ to be continuous. It is also desirable for the curve to be smooth, with no local maxima, minima or inversion points between the current guess and the root. It is evident from Figure 3.7 why this is the case; if the curve has a region where the gradient is near zero, the solution increment would tend to $\pm\infty$, causing the scheme to diverge. Also it is evident why in general a good guess for the root is required, as this will minimise the region over which the function has to be well behaved.

When more than one independent variable is present, Newton's method has to be extended to solve several equations at the same time (a unique solution requires at least as many independent equations as variables). This is made possible by rearranging equation (3.38), and substituting δx for $x_1 - x_o$.

$$J^k_{(i,j)} . \delta x_i^k = -f(x_i^k) \tag{3.41}$$

where J, the Jacobian denotes the matrix formed by the partial derivatives of the equations with each of the independent variables and i the index referring to different independent variables. The solution vector is calculated from a set of linear equations, which is then added to the original value to form the updated guess for the root ie. $x_1 = x_o + \delta x$. A three variable case is illustrated in Figure 3.8.

$$J(x) \cdot \delta x = -f(x)$$

$$\begin{vmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} \\ \frac{\partial f_3}{\partial x_1} & \frac{\partial f_3}{\partial x_2} & \frac{\partial f_3}{\partial x_3} \end{vmatrix} \begin{vmatrix} \delta x_1 \\ \delta x_2 \\ \delta x_3 \end{vmatrix} = \begin{vmatrix} -f_1 \\ -f_2 \\ -f_3 \end{vmatrix}$$

Figure 3.8

Three variable Jacobian formed from partial derivatives

After discretization the device equations produce many independent variables, one for each of the main device variables, ψ and ϕ, at each of the grid nodes. Note that in the text that follows each independent variable at each of the nodes will be

labelled a minor variable, and each global device variable a major variable. A full 2D model having a typical mesh of 100x50 nodes will have at least 5000 independent minor variables for each major variable, all of which require their own equation for which a root has to be found. The large number of variables causes many practical computational problems associated with machine size and speed. Each major variable, dependent as well as independent in the above case requires 40 KBytes, and if a full Jacobian matrix were used it would require up to several GBytes of memory. Fortunately these minor variables are not all explicitly linked, each connected to no more than Nx5 other variables[7]. Consequently the majority of the partial derivatives that occur between these minor variables are identically zero and the Jacobian matrix has few non-zero values (less than 0.1% for the above example). This makes the matrix inversion procedure much less computationally intensive and consequently much quicker. The Jacobian matrix thus forms a large,

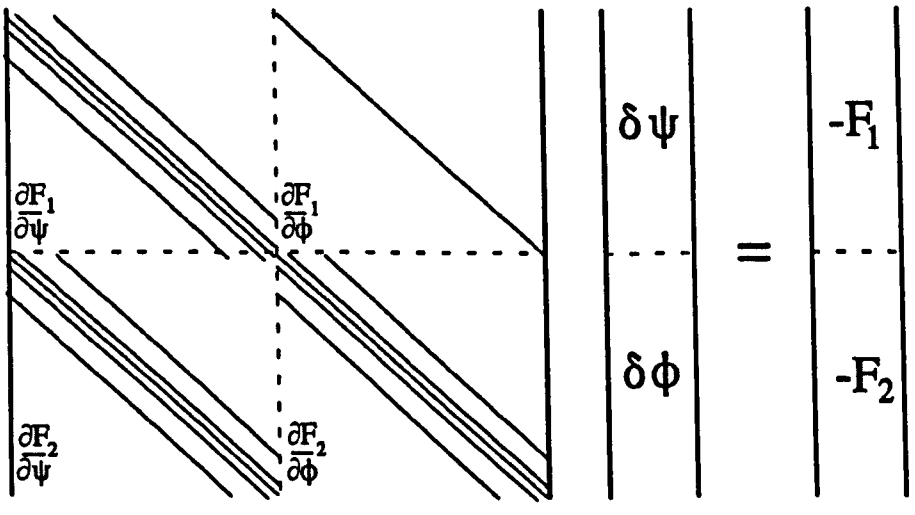


Figure 3.9

Figure illustrating the banded nature of the two-dimensional model Jacobian

sparse set of linear equations, which in this case, can be divided into four sub-matrices, with each sub-matrix in turn being pentadiagonal due to the grad and div.grad operators. Many numerical techniques, including both direct and iterative

---

[7] Where N denotes the number of independent major variables

methods, have been used to solve this set of linear equations, briefly reviewed in the next section.

3.8 Matrix inversion routines

The iterative schemes used include Gauss-Siedel and its variations, where the set is linearised by stepping through the equations, solving for one independent variable whilst holding the others constant, equation (3.42).

$$\begin{aligned} \alpha^{k+1} &: f_1(\alpha^k, \beta^k, \gamma^k) \\ \beta^{k+1} &: f_2(\alpha^{k+1}, \beta^k, \gamma^k) \\ \gamma^{k+1} &: f_3(\alpha^{k+1}, \beta^{k+1}, \gamma^k) \end{aligned}$$

(3.42) Example of a Gummel routine for three independent variables

The Gauss-Siedel scheme requires the equations to be well conditioned, ie. diagonally dominant, to work well. It is evident that if too great a dependence on the values of neighbouring nodes is present then the system is prone to oscillation which in turn means poor convergence if at all. This effect can be reduced by relaxing the solution increment, with each updated solution the sum of new and old components, equation (3.43), [Cole, 1993].

$$x^{k+1} = \alpha x^{k+1}_{calculated} + (1-\alpha)x^k \qquad (3.43)$$

Although the Gauss-Siedel scheme worked well for the device under low Drain-Source biases it had difficulty converging once the saturated current regime was entered. One must conclude that whilst the device equations are relatively well conditioned in the ohmic region of device operation they become ill conditioned once

the saturated region is entered. This is confirmed by a calculation of the L-infinity condition number of the Jacobian, which increases with the drain-source bias, indicating less well-conditioned matrices. This causes problems for convergence even when the equations are re-arranged to provide an excellent approximation for the solution.

The usual matrix equation to be solved has the form of equation (3.44).

$$J(x^k).\delta x = -f(x^k) \tag{3.44}$$

but by adding $J(x).x$ to each side, and noting that $x^k + \delta x = x^{k+1}$ this becomes:

$$J(x^k).x^{k+1} = J(x^k).x^k - f(x^k) \tag{3.45}$$

If the relative change in x is not large then $x^{k+1} \sim x^k$, and the scheme should be close to solution. This method solves directly for the next guess at the root of the equations, rather than the increment between guesses used in the conventional Newton scheme. This method was implemented, and even though it displayed an initial improvement, the ill-conditioning of the Jacobian again precluded convergence once the saturated current region of device operation was entered.

An alternative Gummel scheme was also developed which avoided the necessity of the inversion of a large matrix. This scheme solves Poisson's and the current continuity equations at each node using a local Newton scheme, producing a matrix with a rank of only two or three, depending on the number of independent global variables, ψ, ϕ, T_e etc. The corresponding inversion of this local Jacobian is much easier when compared to the inversion of the full mesh matrix. The grid nodes are then stepped through sequentially, resolving each mesh point until no further change takes place. Although this scheme avoided the computational expense connected with inverting a large sparse matrix it was unusable because each updated mesh point had

to be considerably relaxed to avoid oscillation and hence the overall scheme became very slow to converge.

Another method, that was semi-iterative in nature, divides the mesh into vertical or horizontal lines, sub-dividing the Jacobian into a set of tridiagonal matrices. The inversion of tridiagonal matrices using direct methods is fairly straight forward, with a solution to the whole Jacobian, the set of tridiagonal matrices, found by iterating each line solution within a relaxed Gummel loop. This scheme is more attractive when the tridiagonal matrices are solved using an LU decomposition algorithm, which can be stored and used to find the solution of the tridiagonal matrix upon subsequent iterations[8]. Again this scheme worked well at low biases but took progressively longer to converge as the drain-source bias is increased, until eventually it becomes unfeasible for device simulation, requiring millions of iterations.

Several direct methods were also tried including routines based upon Gaussian elimination and LU Decomposition. These were initially written by the author but more efficient routines were found in packaged routines, namely 'LINPAK' and 'SPARSE'. The latter of these programs is currently being used written by K.S. Kundert and A. Sangiovanni-Vincentelli [Kundert, 1988]. This routine works by LU decomposition and typically solves the matrix equation to within a normalised error of less than 10^{-17}, close to machine accuracy for these equations. Direct methods have the advantage that they produce a solution even when the matrix is ill-conditioned, although they can suffer from round-off error, where initially small errors in the solution at one end of the matrix are amplified, becoming large by the time the solution at the other end of the matrix is calculated. This effect can be reduced by reordering the matrix elements, an operation performed automatically within 'SPARSE'. If the matrix retains the same zero/non-zero pattern the reordering procedure need only be invoked once, at the beginning of the simulation, and the same reordering scheme used for all subsequent matrices. This is why the new mesh

---

[8] This procedure only requires the multiplication of the inverted matrix with the old solution, hence in the case of a tridiagonal matrix is performed in just 2xN operations, where N is the rank of the matrix

refinement algorithm was developed, mentioned in section 3.4, which retains the same nodal pattern.

If problems still occur a more accurate solution can be found using an iterative scheme with an approximate solution produced by SPARSE used as an initial starting point. This has not been necessary within the solution schemes incorporated into the full two-dimensional HFET model discussed in Chapters 5.

Chapter 4

One-Dimensional Models

4.1 Introduction

This chapter presents the modifications and additions to the numerical schemes described in the last chapter that are required to solve the device equations for the one-dimensional models. When Schottky barriers are modelled, due to their very low leakage current, the electron transport can be neglected, greatly simplifying the problem. This was the first model to include quantum mechanics developed in this study and although similar schemes are reported frequently in the literature an elegant, robust and fast scheme is described for the solution of the device equations for this structure. The conduction band, sub-band structure, wavefunctions and carrier profiles are displayed for several different heterostructures, clearly demonstrating the formation of one or more 2DEGs. Current is then introduced in the one-dimensional case as a prelude to the two-dimensional model, and the two schemes compared.

4.2 Schottky Gate / Heterojunction charge control model

The central problem with physical device modelling lies in accurately, robustly and efficiently solving the set of non-linear partial differential equations that describe the device operation. One of the simplest cases is demonstrated by a charge control model including a Schottky gate as one of the contacts. In this case, if the gate dimensions are assumed to be large with respect to the events occurring at the heterojunction, one can neglect the gate fringing effects and formulate the problem in one-dimension. The model can further be simplified when the rectifying properties of the Schottky junction are considered. The gate, which is usually operated in reverse bias, has an exceedingly low leakage current and thus to a good approximation this structure can be modelled neglecting electron transport. This has the consequence that the current continuity (particle conservation), momentum conservation and energy conservation equations are

all automatically solved. The device is effectively in equilibrium, and hence the Fermi-potential is flat. Since potentials are arbitrary variables, the Fermi-potential may be set at the reference point for all positions, equation (4.1).

$$\phi = 0 \tag{4.1}$$

Also the net momentum is zero, equation (4.2).

$$p = 0 \tag{4.2}$$

Finally, since there are no driving terms present in the energy conservation equation the average energy is the thermal energy of the carriers and set equal to that of the lattice, equation (4.3).

$$w = \frac{3}{2} k_B T_0 \tag{4.3}$$

This leaves ψ, the electrostatic potential, to be found via a self-consistent solution of Poisson's equation with the charge density. From the previous chapter the charge density is chosen to be a dependent variable, formed from the difference of ionised donors and electron density, both of which are functions of the conduction band edge - Fermi-level separation.

4.2.1 Boundary conditions

In all models the device equations derived in chapter two only apply within the simulation domain. On the surface they are complimented by the equations describing the boundary conditions. In one-dimension the boundary conditions are relatively simple as only the two extreme nodes are involved. In the Schottky - gate model ϕ, p and w

are solved hence only the boundary conditions for ψ are required. The Schottky gate contact is modelled simply by pinning the conduction band edge at the applied gate bias plus the Schottky 'built-in' potential, equation (4.4).

$$E_C = V_G + V_{bi} \tag{4.4}$$

As V_H, the conduction band discontinuity is known ψ can be obtained from equation (4.5).

$$\psi = V_H - \frac{E_C}{q} \tag{4.5}$$

There are two options available for the substrate boundary condition. Both are derived by stipulating that this boundary is in equilibrium, and thus should be unaffected when the gate bias is altered. The first option, method A, is to set the conduction band - Fermi-level separation constant, such that there is no net charge present at this point, equation (4.6).

$$\rho = N_D^+ - n = 0 \tag{4.6}$$

This equation is still highly non-linear due to the exponential terms present in both the ionised donor and electron densities. The second option, method B, is also true in equilibrium and sets the gradient of the conduction band edge to zero, implying no further change beyond this point.

$$\nabla E_C . n = 0 \tag{4.7}$$

where n is the unit vector normal to the boundary.

Both methods are valid, and if the modelled region extends far enough into the substrate they will both solve for the same value of conduction band edge. The first method is

used in this work because if the modelled region is truncated too early it will produce an error that slightly raises the conduction band edge above the true value, whereas the second method will tend to pull the conduction band edge down, Figure 4.1

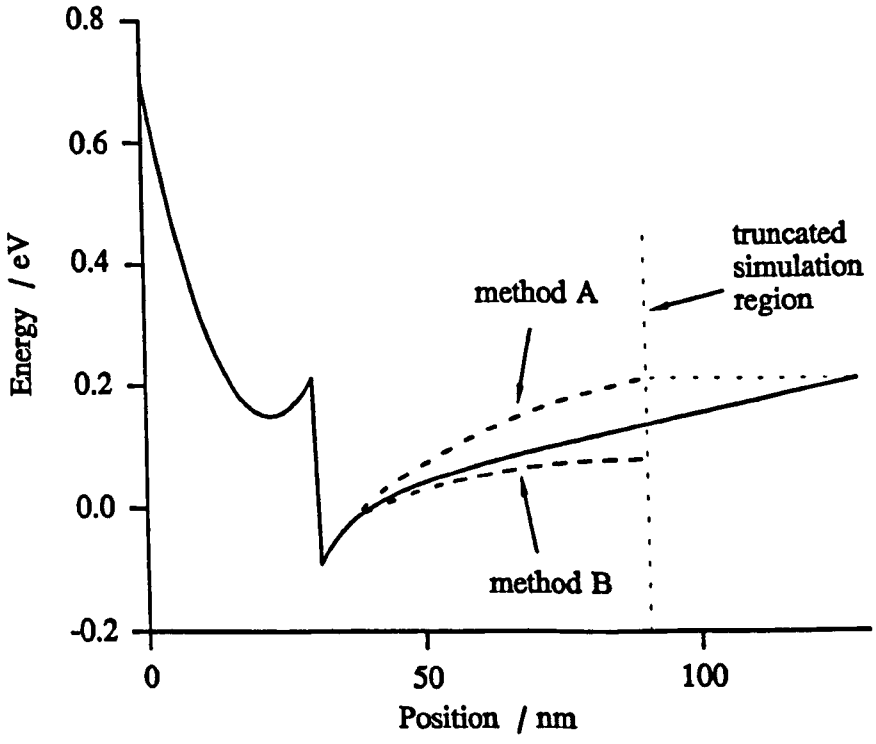


Figure 4.1

Conduction band edge under different boundary conditions

A raised conduction band edge, although strictly incorrect, will help contain the electron wavefunctions associated with the two-dimensional electrons, and consequently makes the model more robust. Thus in the one-dimensional Schottky gate - substrate model both boundaries are modelled using Dirichlet boundary conditions, even though the value of the substrate conduction band edge is not explicitly known.

4.2.2 Solution scheme

Once the boundary equations have been derived one has an equation for ψ at each of the nodes; Dirichlet's equation at the first, setting ψ to the Schottky gate potential,

Poisson's for all internal nodes and Dirichlet again for the substrate end, setting ψ such that the net charge is zero. This system is non-linear and consequently is solved using the Newton scheme described in Chapter 3. This involves the calculation of two components, the Jacobian and the residual. The evaluation of the latter is straight forward, again using the methods described in Chapter 3, whereas the Jacobian which is formed from the partial derivative of Poisson's and the Dirichlet equations with respect to ψ, and is discussed below.

4.2.3 Derivation of the 1D model Jacobian

The Jacobian is a matrix formed from the partial derivatives of the equations with each of the independent variables. In the Schottky gate model this translates to the matrix formed by the partial derivatives of Dirichlet's and Poisson's equations with respect to ψ. The leading diagonal elements of the Jacobian correspond to the partial derivatives of the equations with respect to the value of ψ at the same position, while the off-diagonal elements are the partial derivatives with respect to the neighbouring nodes.

Writing Poisson's equation and expanding ρ, the total charge density, we have equation (4.8).

$$\frac{\partial \epsilon}{\partial y}.\frac{\partial \psi}{\partial y} + \epsilon.\frac{\partial^2 \psi}{\partial y^2} + q(N_D^+(\psi) - n_{3D}(\psi) - n_{2D}(\psi)) = 0 \qquad (4.8)$$

When the central difference scheme is implemented the grad and div.grad elements, together with the ϵ and grad ϵ terms can be expressed in terms an operator matrix, P^{op}, which works upon ψ by multiplication, equation (4.9).

$$\left(\frac{\partial \epsilon}{\partial y}.\frac{\partial}{\partial y} + \epsilon.\frac{\partial^2}{\partial y^2}\right)^{op}.\psi \equiv P^{op}.\psi \tag{4.9}$$

In one-dimension P^{op} forms a tridiagonal matrix, with the elements constant throughout the simulation and independent of the value of ψ. Thus differentiating these terms with respect to ψ simply returns the same operator elements, equation (4.10).

$$\frac{\partial(P^{op}.\psi)}{\partial \psi} = P^{op} \tag{4.10}$$

This 'operator' notation is exploited in the program code, with the elements calculated at the beginning, stored, and then used in both the formulation of the Jacobian matrix and the evaluation of the residual error in the equations. The remaining terms in Poisson equation to be differentiated are the charge densities which are simply functions of the local potential, and hence are independent of the potentials at other positions in the device. This has the consequence that they contribute no off-diagonal elements to the Jacobian. Thus the derivative of Poisson's equation, labelled F1, with respect to ψ can be written as equation (4.11) and is illustrated in Figure 4.2.

$$\frac{\partial F1}{\partial \psi} = P^{op} + \frac{\partial N_D^+}{\partial \psi} - \frac{\partial n_{3D}}{\partial \psi} - \frac{\partial n_{2D}}{\partial \psi} \tag{4.11}$$

where the partial derivatives of the charges are derived in the next section.

The Dirichlet equations describing the boundary conditions have no off-diagonal elements, but take different forms. At the Schottky contact the boundary is given by equation (4.2). Differentiating this with respect to ψ returns a constant, 1. The substrate boundary on the other hand is more complicated, given by equation (4.6), and its derivative is formed from the difference of the donor and electron density partial derivatives at that point. It is the exponential functions in these terms that make the system highly non-linear.

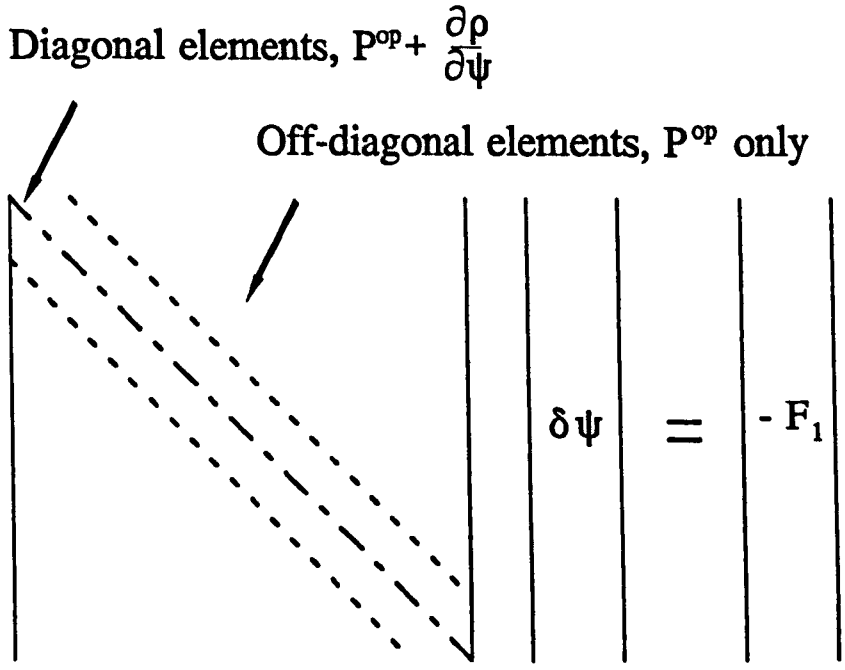


Figure 4.2

The Jacobian matrix showing the origin of the terms

4.2.4 Calculation of the partial derivatives of the charges

All the bulk charge terms, N_D^+ and n_{3D} (N_T^-, p and N_A^- if included) are simply functions of the local difference in conduction band edge and Fermi-level. This translates to the chosen variables as a sum, equation (4.12), where both ϕ and ψ are of the same sign.

$$-(E_C - E_F) = -q.(\psi - V_H + \phi) \qquad (4.12)$$

Thus the partial derivative with respect to ψ is the same as that with respect to ϕ. Although these partial derivatives are not required in the 1D Schottky model they are used in more complicated simulations where current flow is included.

4.2.5 Ionised donor density

The partial derivatives of N_D^+ are straight forward to calculate, equation (4.13) as this is a simple expression involving only one exponential dependent upon the (ψ+ϕ) sum. The only modification of this equation, when implemented on a computer, is to check for overflow of the argument. This check could possibly be omitted since if the (ψ+ϕ) sum is too large to be evaluated the whole scheme is probably in error and failing.

$$\frac{\partial N_D^+(\psi,\phi)}{\partial\psi} = \frac{-\frac{q}{k_B T} N_D}{\left(1 + e^{q.(\psi+\phi-V_h)/(k_B T)}\right)^2} \qquad (4.13)$$

4.2.6 Bulk electron density

The bulk, three-dimensional electron density, n_{3D} is given by an expression that is not analytically soluble, $\mathcal{F}_{1/2}(\eta)$, the Fourier integral, equation (4.14).

$$n_{3D} = N_C \sqrt{\frac{\pi}{2}}\, \mathcal{F}_{1/2}\left(\frac{E_C - E_F}{k_B T}\right) \qquad (4.14)$$

This term is evaluated using a curve fit equation described in Chapter 3, and thus although not elegant the derivative is obtainable by differentiating this curve fit equation. Note that even if there were an exact method for evaluating the derivative of the Fermi-integral the Jacobian requires the derivative of the equations used in the model, which in this case is the curve fit equation.

4.2.7 Two-dimensional electron density

As shown in Appendix A the two-dimensional electron density is given by an equation of the form, equation (4.15).

$$n_{2D}(\psi,\phi) = N_{C_{2D}} \sum_k |\xi_k(\psi)|^2 \log_e\left(1 + e^{\frac{q}{kT}(\phi - \lambda_k(\psi))}\right) \qquad (4.15)$$

The derivative of n_{2D} with respect to ϕ is relatively straight forward since neither ξ_k nor λ_k depend upon ϕ.

$$\frac{\partial n_{2D}(\psi,\phi)}{\partial \phi} = N_{c_{2D}} \sum_k |\xi_k(\psi)|^2 \frac{1}{1 + ex_k(\psi,\phi)} \cdot \frac{\partial ex_k(\psi,\phi)}{\partial \phi} \qquad (4.16)$$

where

$$ex_k(\psi,\phi) = e^{\frac{q}{k_B T}(\phi - \lambda_k(\psi))} \qquad (4.17)$$

$$\frac{\partial ex_k(\psi,\phi)}{\partial \phi} = \frac{q}{k_B T} ex_k(\psi,\phi) \qquad (4.18)$$

The derivative with respect to ψ is not so apparent since both ξ_k and λ_k depend upon this variable and one needs to resort to perturbation theory to obtain this parameter. Differentiating n_{2D} with respect to ψ and explicitly expressing the derivatives of ξ_k and λ_k gives equation (4.19).

$$\frac{\partial n_{2D}}{\partial \psi} = N_{c_{2D}} \sum_k \left\{ \log_e(1+ex_k(\psi,\phi))\, \xi_k^*(\psi) \frac{\partial \xi_k(\psi)}{\partial \psi} + \frac{|\xi_k(\psi)|^2}{1+ex_k(\psi,\phi)} \frac{\partial ex_k(\psi,\phi)}{\partial \psi} \right\} \tag{4.19}$$

where

$$\frac{\partial ex_k(\psi,\phi)}{\partial \psi} = -\frac{q}{k_B T} ex_k(\psi,\phi) \frac{\partial \lambda_k(\psi)}{\partial \psi} \tag{4.20}$$

The partial derivatives of ξ_k and λ_k with respect to ψ are calculated as follows. From the first principles of calculus one can define a partial differential as

$$\frac{\partial a(\psi)}{\partial \psi} = \frac{a(\psi+H) - a(\psi)}{H}\bigg|_{H\to 0} \tag{4.21}$$

Expanding the wavefunction and energy levels in terms of H, a small perturbation in energy, and truncating to first order gives equations (4.22) and (4.23).

$$\xi_k(\psi+H) \approx \xi_k(\psi) + \delta\xi_k(H) \tag{4.22}$$

$$\lambda_k(\psi+H) \approx \lambda_k(\psi) + \delta\lambda_k(H) \tag{4.23}$$

hence upon substitution one obtains

$$\frac{\partial \xi_k}{\partial E} \approx \frac{\delta\xi_k(H)}{H}\bigg|_{H\to 0} \tag{4.24}$$

$$\frac{\partial \lambda_k}{\partial E} \approx \frac{\delta\lambda_k(H)}{H}\bigg|_{H\to 0} \tag{4.25}$$

The charge on the electron is negative, thus energy is proportional to -ψ, and the partial derivatives of ξ_k and λ_k with respect to ψ are

$$\frac{\partial \xi_k}{\partial \psi} = -\frac{\partial \xi_k}{\partial E} \tag{4.26}$$

$$\frac{\partial \lambda_k}{\partial \psi} = -\frac{\partial \lambda_k}{\partial E} \tag{4.27}$$

Applying perturbation theory to equations (4.22) and (4.23) one obtains

$$\delta \lambda_k(H) = \int \xi_k^* H \xi_k \tag{4.28}$$

$$\delta \xi_k(H) = \sum_l \alpha_l(H) \xi_l(\psi) \tag{4.29}$$

where

$$\alpha_{l \neq k}(H) = \frac{\int \xi_k^* H \xi_l}{(\lambda_l - \lambda_k)} \tag{4.30}$$

and α_k is obtained by normalising $\xi_k^* \xi_k$.

$$\alpha_k(H) = \left(1 - \sum_{l \neq k} \alpha_l^2\right)^{\frac{1}{2}} - 1 \tag{4.31}$$

As H, the energy perturbation operator, has no off-diagonal elements it may be taken outside the integral, giving[1]

[1] The elements of ξ_k form an orthonormal set

$$\delta\lambda_k(H) = H\int \xi_k^* \xi_k = H \tag{4.32}$$

$$\alpha_{l\neq k} = \int \xi_l^* H \xi_k = H\int \xi_l^* \xi_k = 0 \tag{4.33}$$

and substituting $\alpha_{l\neq k}$ into equation (4.31).

$$\alpha_k = 0 \tag{4.34}$$

Substituting equations (4.32), (4.33) and (4.34) into (4.24) and (4.25) one obtains the partial derivatives of ξ_k and λ_k with respect to ψ.

$$\frac{\partial\lambda_k}{\partial\psi} = -1 \tag{4.35}$$

$$\frac{\partial\xi_k}{\partial\psi} = 0 \tag{4.36}$$

Finally using this result in equation (4.19) one has

$$\frac{\partial n_{2D}}{\partial\psi} = N_{c_{2D}} \sum_k |\xi_k|^2 \frac{1}{1+e^{\frac{q}{kT}(\phi-\lambda_k)}} \frac{q}{k_B T} e^{\frac{q}{k_B T}(\phi-\lambda_k)} \tag{4.37}$$

and by inspection it is evident that

$$\frac{\partial n_{2D}}{\partial\psi} = \frac{\partial n_{2D}}{\partial\phi} \tag{4.38}$$

Upon reflection this result is not wholly surprising as the electron potential energy is not

a function of the absolute value of the conduction band edge, but measured with respect to a reference, the Fermi-level. Thus an equivalent change in electron potential energy is possible by moving either the conduction band edge or the Fermi-level in the opposite directions.

$$\frac{\partial n_{2D}}{\partial E_C} = -\frac{\partial n_{2D}}{\partial E_F} \tag{4.39}$$

and since

$$E_C = q.(V_H - \psi) \quad and \quad E_F = q.\phi \tag{4.40}$$

the partial derivatives of n_{2D} with respect to ψ and ϕ are identical.

4.3 Solution of the Jacobian and Hasselgrove relaxation

The Schottky gate charge control model is one-dimensional and thus the potential at each node is only explicitly linked to that of its two nearest neighbours. As discussed in the previous section this creates a tridiagonal Jacobian matrix. Tridiagonal matrices are special in that they can easily be inverted requiring only 2N operations[2]. This is a consequence of their special format which does not produce any fill-in's (ie. non-zero values in positions where there previously were none), and thus also allows them to be stored very efficiently as three vectors, one for each of the diagonal lines. One of the most efficient solution algorithms for tridiagonal matrices is based upon the LU decomposition algorithm [Press, 1989] and is used in this simulation.

The updated 'guess' for the solution vector is produced from the sum of the current value and the increment calculated from the Jacobian matrix equation. In some cases the solution scheme is more robust if relaxed, with the new solution vector formed by only

---

[2] Matrix inversion is usually an N^3 operation, where N is the rank of the matrix.

adding a fraction, α, of the solution increment, equation (4.41).

$$x^{k+1} = x^k + \alpha\,\delta x \tag{4.41}$$

where α lies between 0 and 1.

From experience this system of equations generally does not require relaxation and thus if the scheme was permanently relaxed it would be significantly slowed down. Note that with no relaxation an ideal Newton's scheme has quadratic convergence properties, when moderate relaxation factors (0.6-0.9) are used this is reduced to super-linear convergence, and at high relaxation factors (<0.3) can even become logarithmic, asymptotically approaching the solution. Therefore a good compromise between speed and stability is given by the Hasselgrove scheme, [Powell, 1970], which checks for convergence upon each iteration and only introduces the relaxation factor if this criterion fails. A record is made of the original values, the increment added and the new variable checked for convergence. If the new error is greater than that produced by the original variable this step is repeated with a relaxed increment. The increment can be successively reduced until convergence is obtained. One obvious problem with this scheme is that the solution vector can be reduced to such an extent that the new 'guess' for the solution vector is almost the same as the original, and hence no real change has occurred, and therefore the next iteration will produce a similar result. Consequently the model places a lower limit on the relaxation factor, which is typically not permitted to drop below 0.1, even if this initially causes the overall error to increase. A typical solution flow chart is illustrated in Figure 4.3.

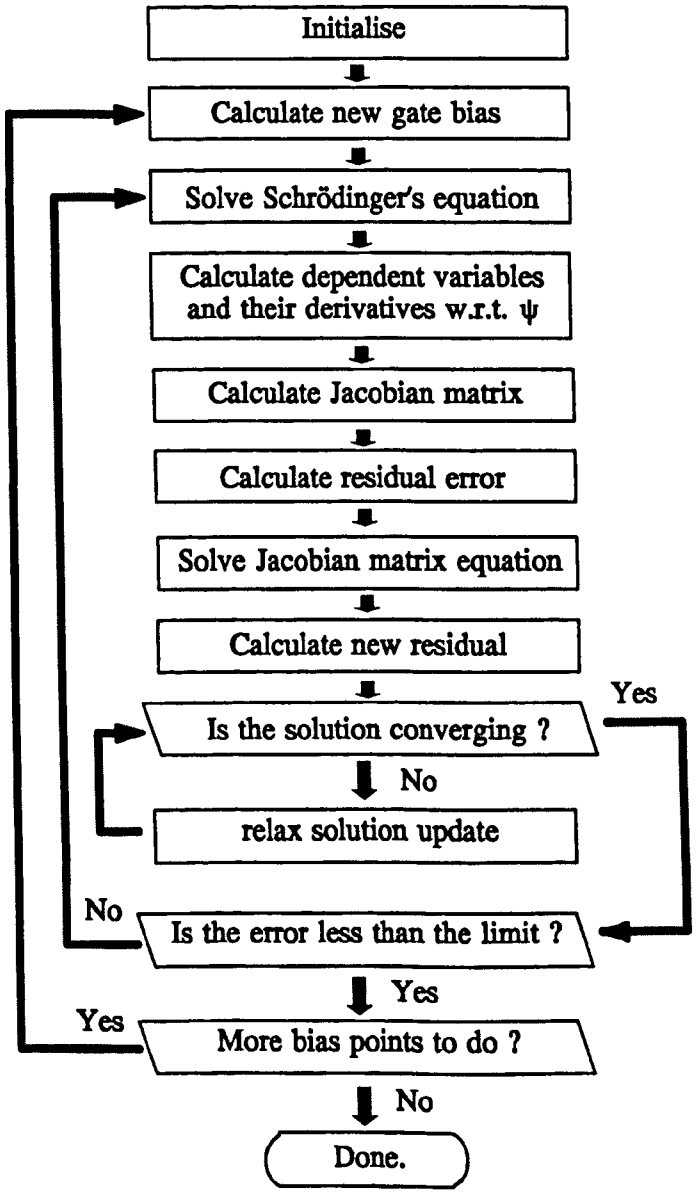


Figure 4.3

Flow chart of the Newton solution process

Figure 4.4 shows the typical convergence properties obtained for the Newton scheme applied to these equations, note the logarithmic scale on the y-axis. It is also worth noting that this curve corresponds to the first solution point, hence the initial guess for the solution, whilst not poor, is certainly not excellent. Subsequent bias points, provided the increment in bias is not too great, have a much better initial starting point using the previous bias' solution.

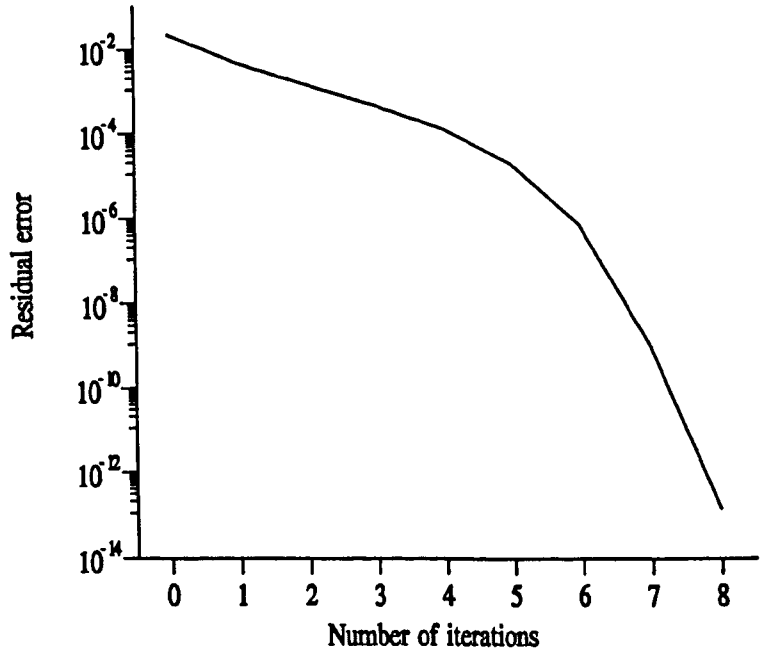


Figure 4.4

Plot of the residual error versus iteration for a typical heterojunction structure.

4.4 Results of the one-dimensional Schottky gate model

This model has been implemented on several one-dimensional structures, including a simple single AlGaAs/GaAs heterojunction, a pseudomorphic AlGaAs/InGaAs/GaAs heterojunction, a multi-channel AlGaAs/GaAs/AlGaAs/GaAs structure and a homogeneous delta-doped MESFET structure. All simulations produced quantised bound states and the conduction band edge and sub-band structure, bound state wavefunctions and total electron density profile are shown below.

The single heterojunction structure is formed from a 30nm $Al_{0.3}Ga_{0.7}As$ layer, doped at $1.5x10^{24}$ m^{-3} on top of an undoped GaAs substrate 100nm deep. The pseudomorphic device is similar except here a thin (10nm) undoped $In_{0.2}Ga_{0.8}As$ placed between the AlGaAs and GaAs layers. The multi-channel structure is formed from four layers, 30nm of $Al_{0.3}Ga_{0.7}As$, 20nm GaAs, 20nm of $Al_{0.3}Ga_{0.7}As$, followed by 100nm GaAs. Both

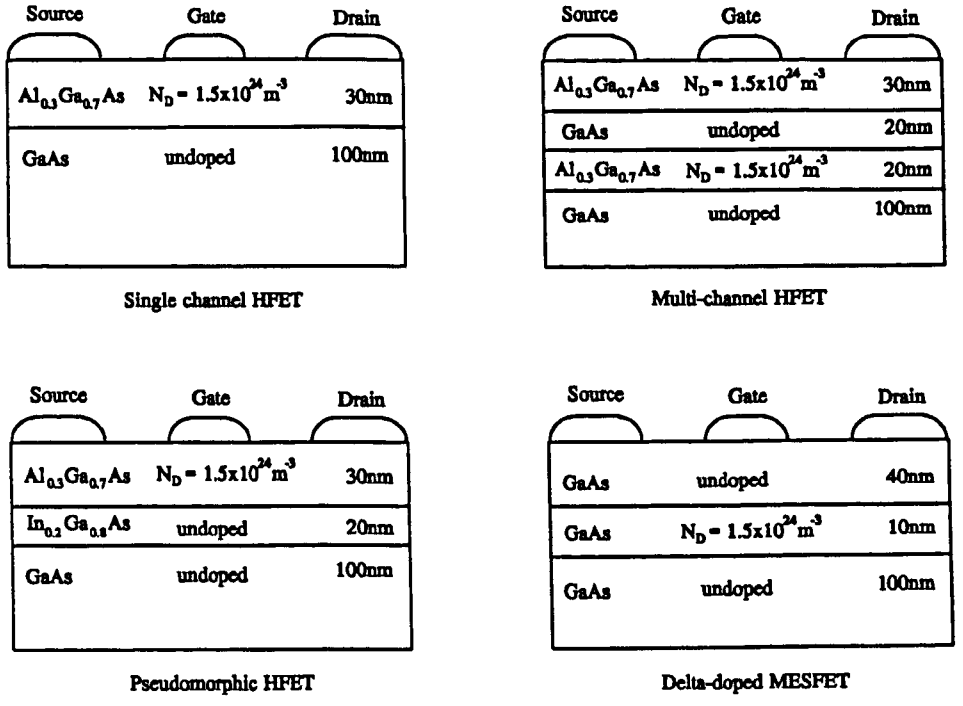


Figure 4.5

Cross-sections of the devices modelled with the 1D Schottky gate model

AlGaAs layers are doped at $1.5x10^{24}$ m^{-3} and both GaAs layers left undoped. Finally the delta-doped structure is formed from a narrow layer of highly doped material, again $1.5x10^{24}$ m^{-3}, placed 40nm from the Schottky gate and extending a further 10nm into the substrate.

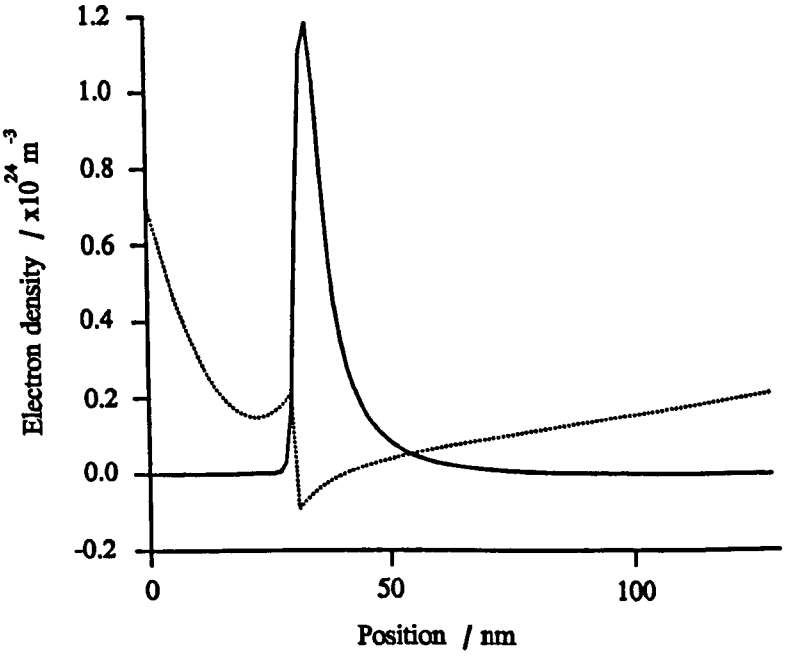


Figure 4.6

Single heterojunction electron density

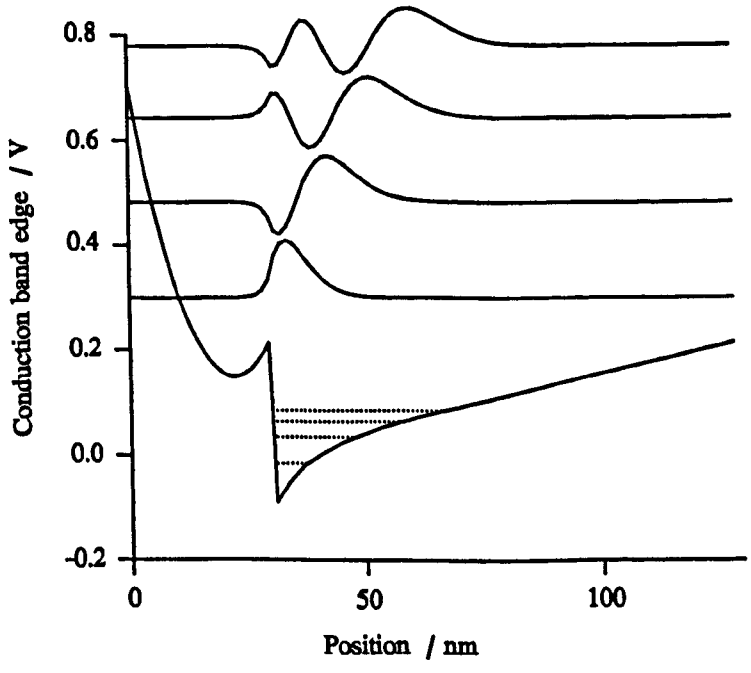


Figure 4.7

Single heterojunction conduction band profile, with the first 4 bound states

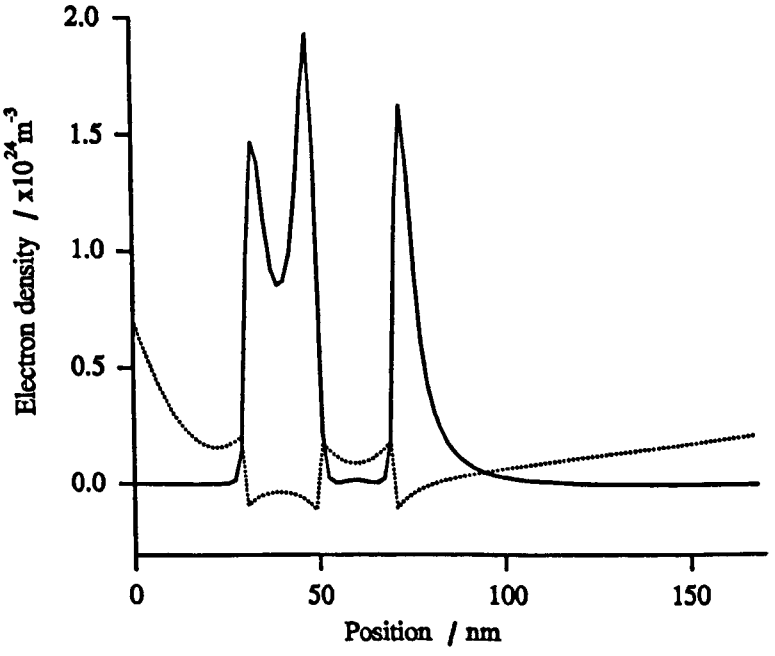


Figure 4.8

Multichannel HFET electron profile

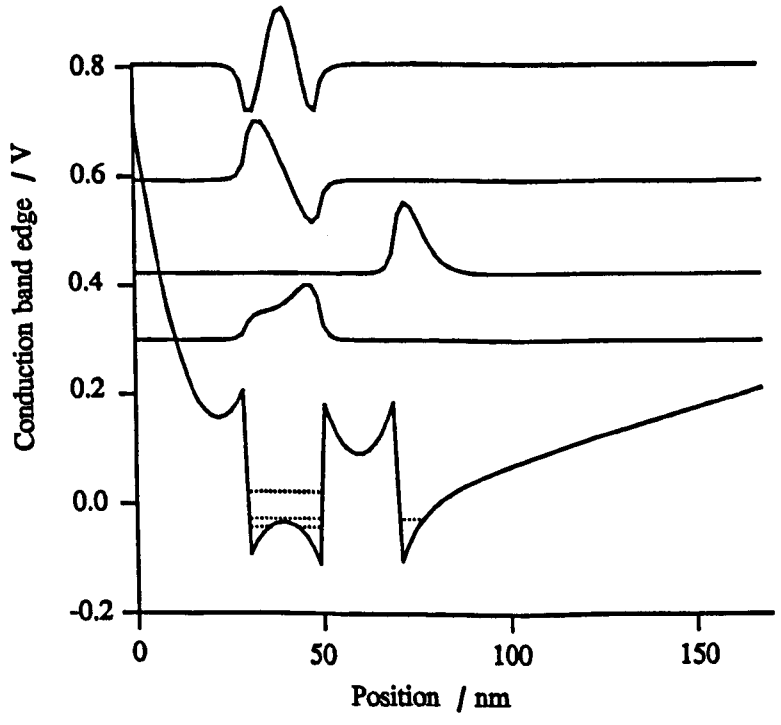


Figure 4.9

Multichannel HFET conduction band edge, with the first 4 bound states

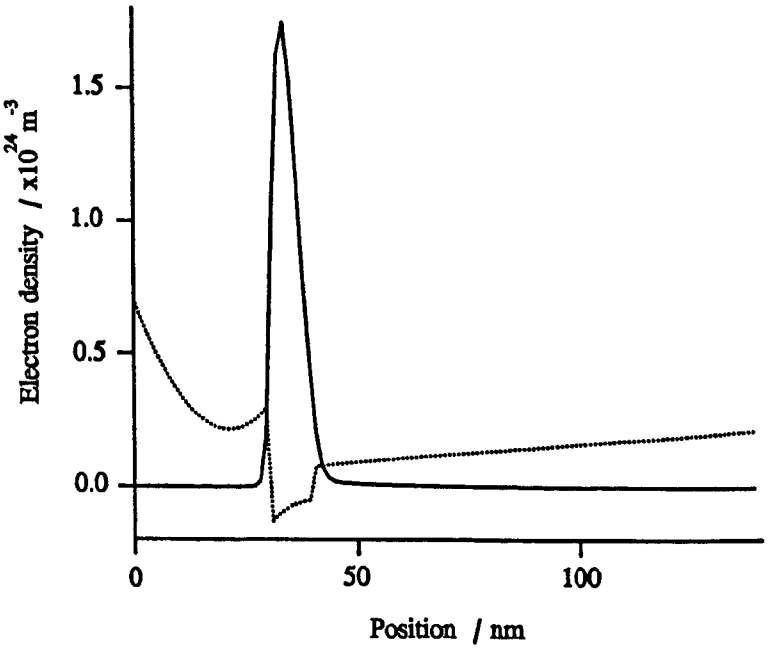


Figure 4.10

Pseudomorphic HFET electron profile

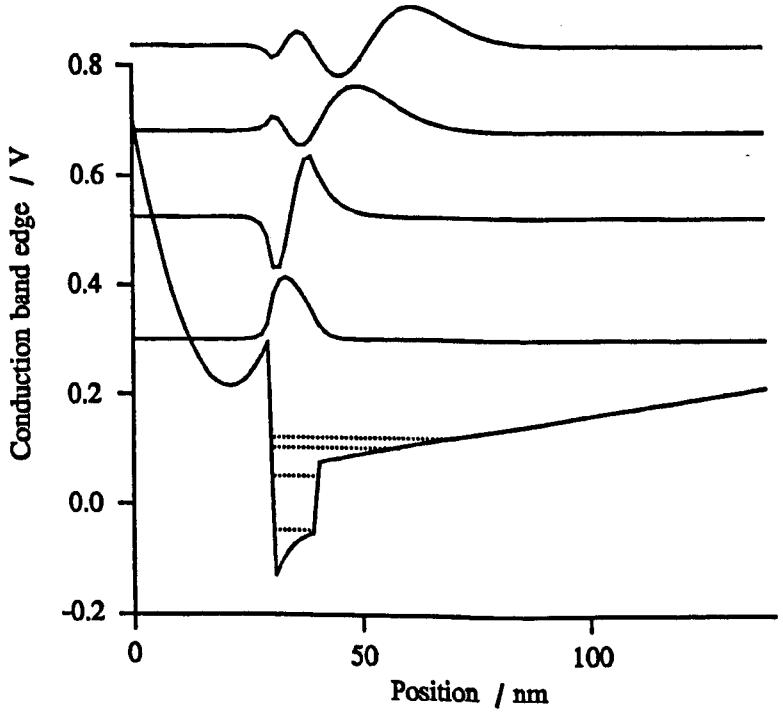


Figure 4.11

Pseudomorphic HFET conduction band profile, with the first four bound states

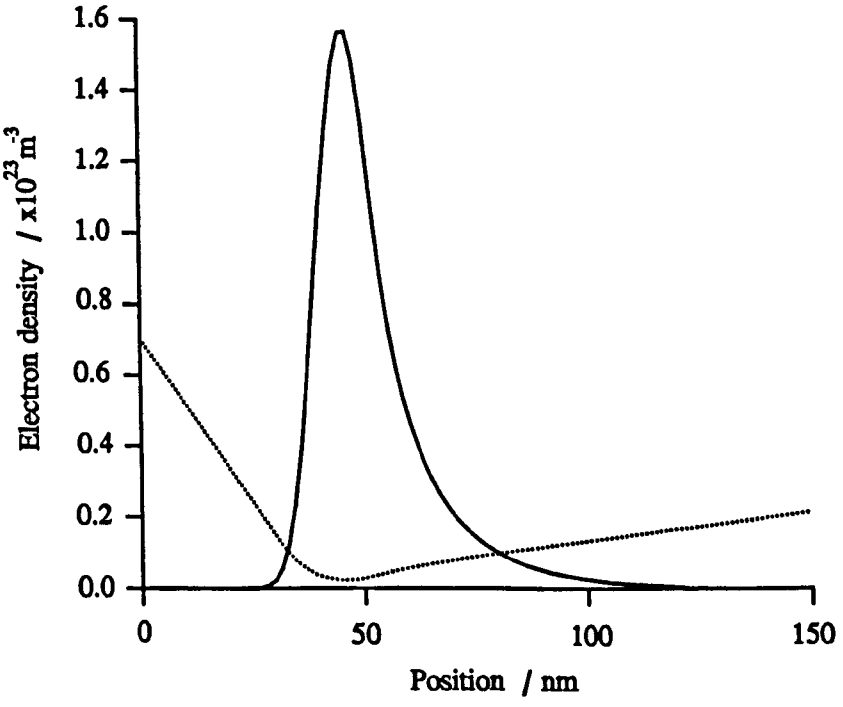


Figure 4.12

Delta-doped MESFET electron profile

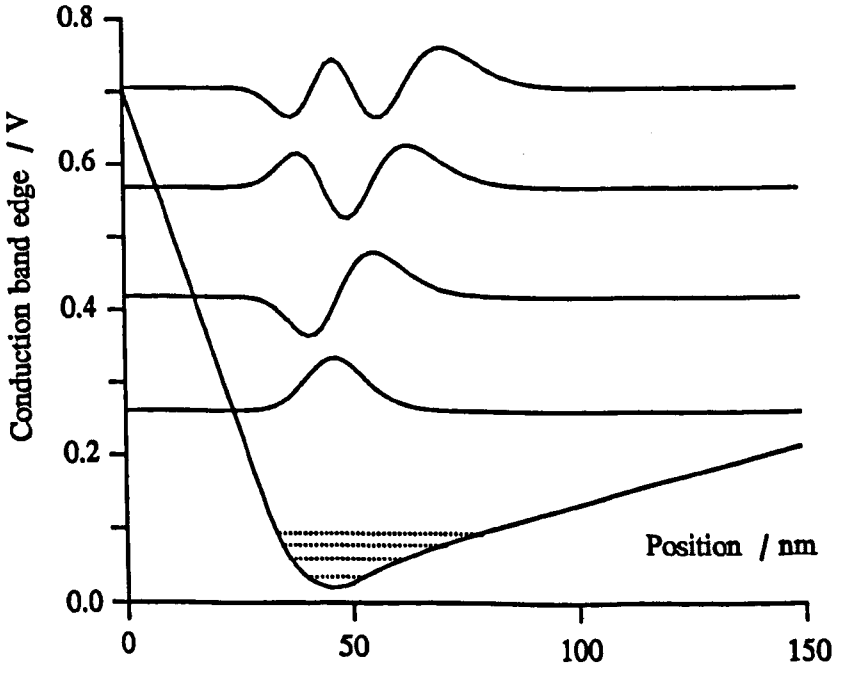


Figure 4.13

Delta-doped MESFET conduction band profile, with the first 4 bound states

These solutions were compared to the classical results and errors of up to 10 - 15% were found in the sheet electron density. This confirms the work done by Yoshida who compared schemes including classical Boltzmann, classical Fermi-integral, full quantum mechanics, and the triangular well approximation, [Yoshida, 1986].

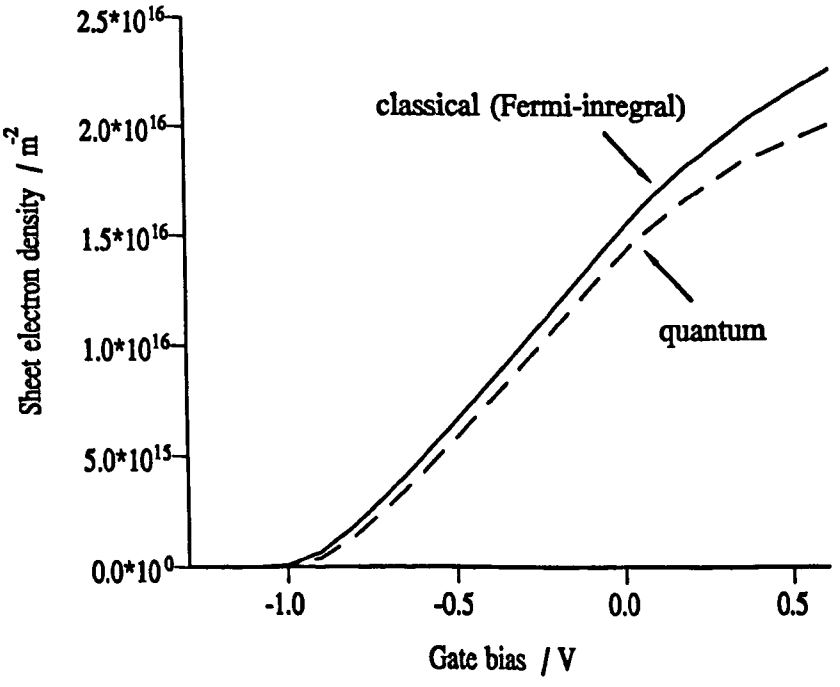


Figure 4.14

Classical and quantum sheet electron densities for a pseudomorphic HFET

4.5 Rapid evaluation of device characteristics

This model, although simple, omitting many of the more complicated effects due to charge transport, is still useful in device design, quickly giving an indication of the amount of charge control of the gate. Also, although to be treated with caution, it can be utilised to give an indication of device operation. For example if the gate fringing effects are neglected, corresponding to a large gate width, the gate capacitance can be modelled. Further if the electron velocity under the gate is assumed to be constant with gate bias, again more likely to be true for long gate length, an average velocity can be used and the transconductance of the device obtained.

Transconductance is defined by equation (4.42)

$$g_m = \left.\frac{\partial I_{DS}}{\partial V_{GS}}\right|_{V_{DS}} \tag{4.42}$$

Assuming a constant 'average' electron velocity I_{DS} can be expressed by equation (4.43)

$$I_{DS} = q Z n_s \bar{v} \tag{4.43}$$

where Z is the device length (the length of the gate fingers) and n_s is the sheet electron density, given from the integral of electron density over the modelled region. Thus as q, Z, and $\bar{v}$ are constant with gate bias they can be taken out of the partial derivative giving equation (4.44)

$$g_m = q Z \bar{v} \left.\frac{\partial n_s}{\partial V_{GS}}\right|_{V_{DS}} \tag{4.44}$$

This simple model can be extended further to obtain characteristics such as the cutoff-frequency and minimum noise figure. These parameters are calculated from an

equivalent circuit model [Ladbrooke, 1989], and take the following form:

$$f_T = \frac{g_m}{2\pi (C_{GC} + C_{GD})} \tag{4.45}$$

and

$$NF_{min} = 1 + K_F \omega C_{GC} \left(\frac{R_S + R_G}{g_m} \right)^{1/2} \tag{4.46}$$

These equations have been incorporated into the one-dimensional Schottky gate model and the following plots produced showing the variation of these values with parameters such as doping density. The simulation was chosen to model a proposed structure for a power device from M/A-COM inc. The layout shown in Figure 4.15 incorporates two pulse doped layers that supply the channel with electrons. A section under the gate has been modelled with the one-dimensional simulator using the layer structure shown in Table 4.1. Here the last layer is the superlattice buffer and hence is formed from undoped alternating layers of AlGaAs and GaAs. The resulting conduction band profile and electron density is shown in Figure 4.16, together with the channel capacitance, Figure 4.17, Figure 4.18, transconductance, Figure 4.18, and minimum noise figure, Figure 4.19.

These parameters were all taken at zero gate bias and using an average velocity under the gate of 5×10^4 ms^{-1}. The figures illustrate that there is no point in raising the doping density above 4×10^{24} donors m^{-3}, as this is the value at which the transconductance plateaus. Indeed if the doping density is increased the model predicts that the transconductance will start to oscillate with donor concentration. This is due to the two pulse doped layers bending the conduction band enough to create two potential wells, and consequently two 2DEGs. The alternating transconductance is due to the two 2DEGs and their interplay.

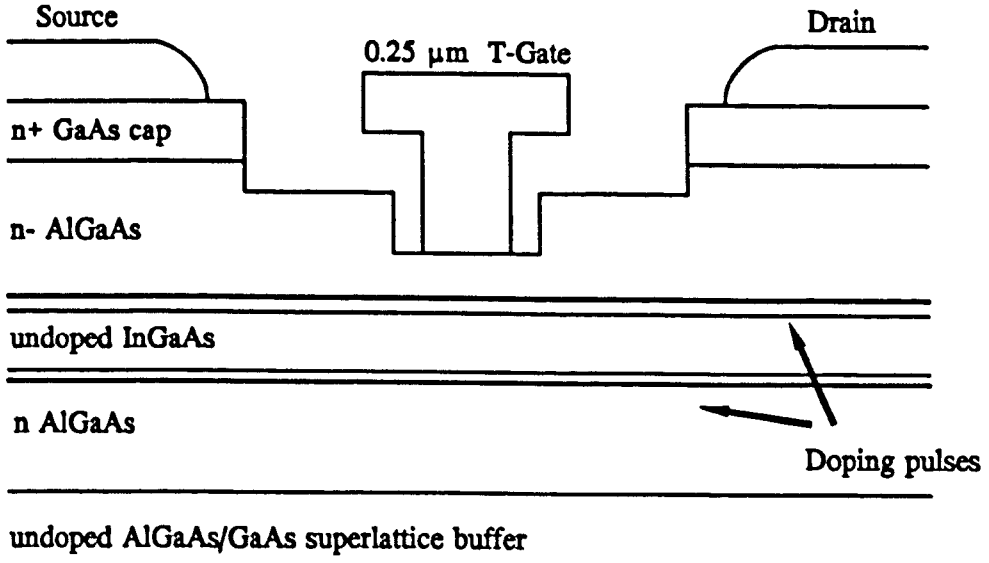


Figure 4.15

Cross-section of the proposed M/A-COM power HFET

Layer N°.	Width / nm	% Al	% In	N_D / m^{-3}
1	50	24	0	$2x10^{20}$
2	1	24	0	$1x10^{25}$
3	1	24	0	$2x10^{20}$
4	12	0	20	$1x10^{20}$
5	1	24	0	$2x10^{22}$
6	1	24	0	$1x10^{25}$
7	50	24	0	$2x10^{22}$
8	100	24 / 0	0	$1x10^{20}$

Table 4.1

Table of the composition of the layers in the proposed M/A-COM power HFET

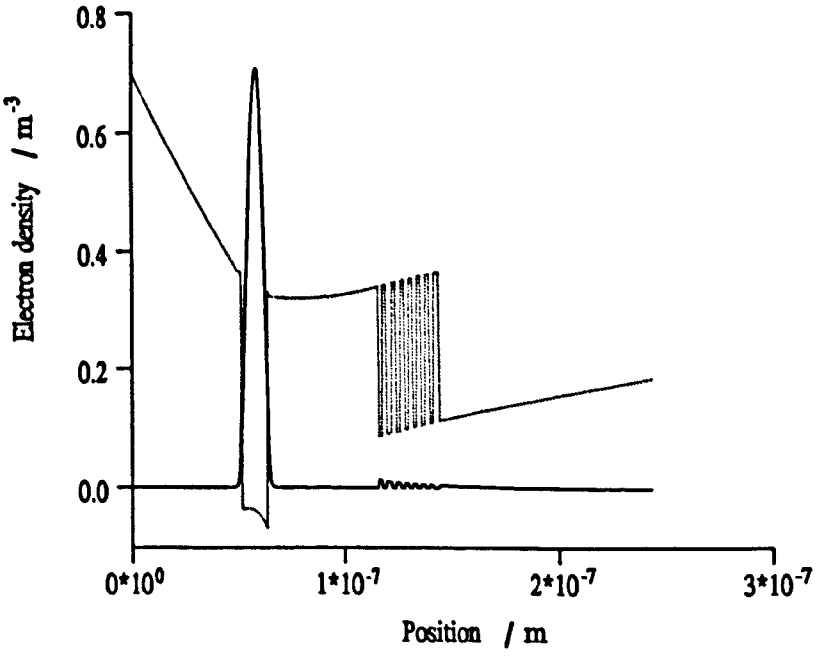


Figure 4.16

Conduction band edge and electron density for the proposed M/A-COM power HFET

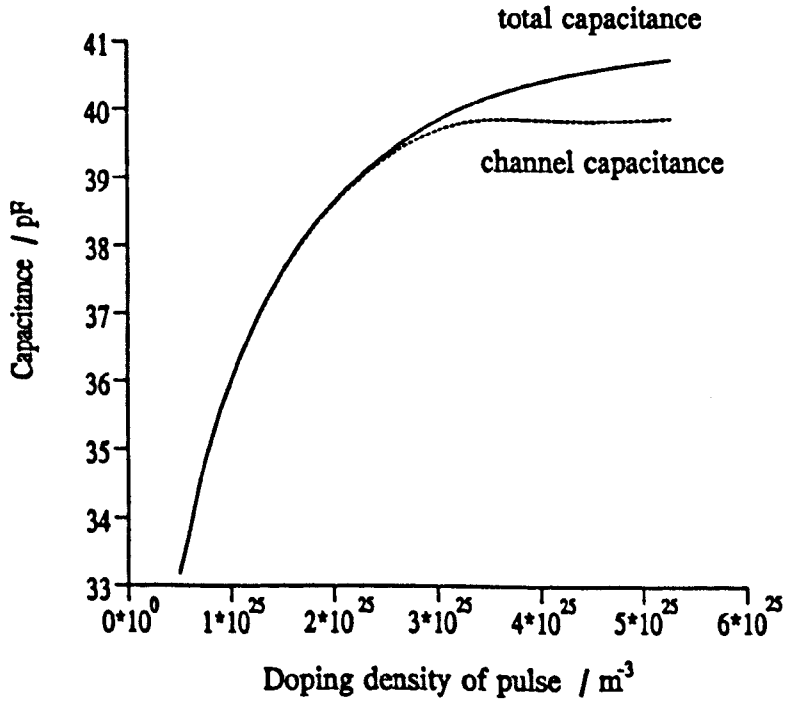


Figure 4.17

Total and channel capacitances versus pulse doping density

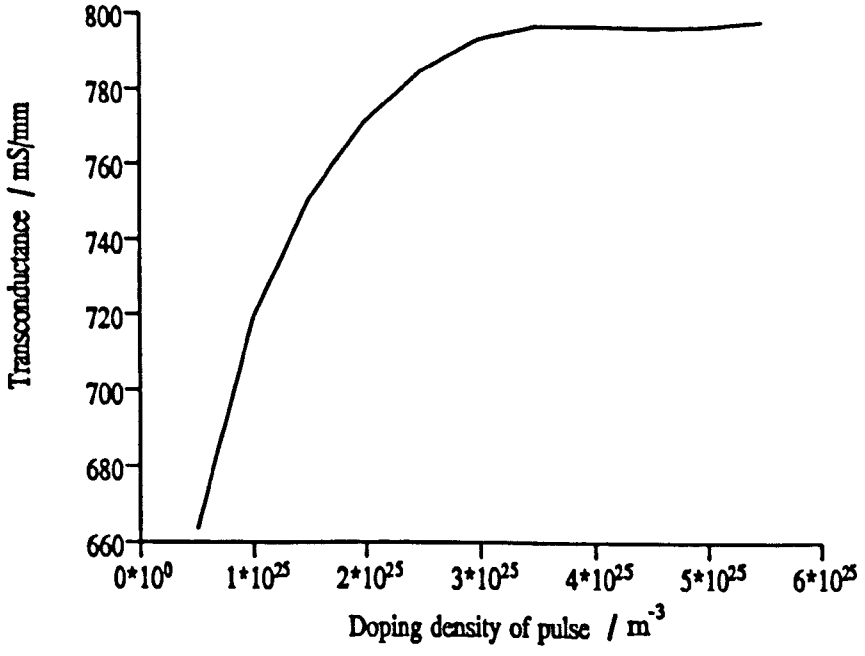


Figure 4.18

Transconductance versus pulse doping density

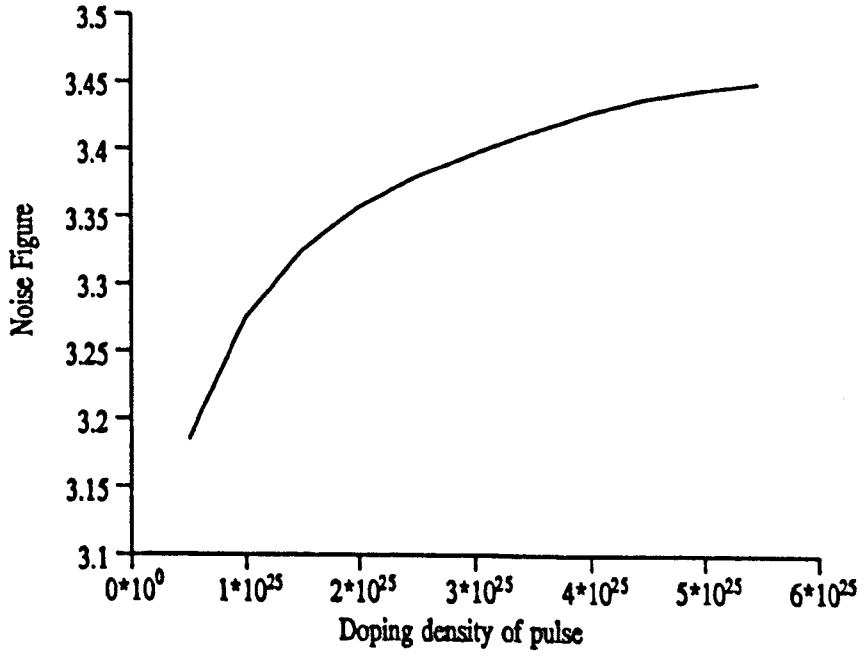


Figure 4.19

Noise Figure versus pulse doping density

@ 10 GHz

4.6 One dimensional current model

Numerous problems presented themselves in the development of the full two-dimensional model, many linked with the new form of the current continuity equations used in this simulation. In an effort to reduce the number of variables, and hence the complexity of this scheme, charge transport was added to the one-dimensional model in order to study its effects, and solve the associated problems in a simpler system before returning to the full FET model. This was done and its results discussed more fully in the next chapter, but had the consequence that the one-dimensional model including current was developed which was able to validate the flat Fermi-level assumption in the Schottky gate model. Figure 4.20 illustrates a typical solution, with the Fermi-level in the charge transport model pinned at the gate bias. Both schemes were implemented on the same structure and the total electron density calculated for various bias points. Figure 4.21 shows that these curves only deviate once the intrinsic reverse bias, the Schottky 'built-in' bias is removed by external potentials. Thus in normal operating circumstances, with the gate reverse-biased the simple Schottky gate model is accurate.

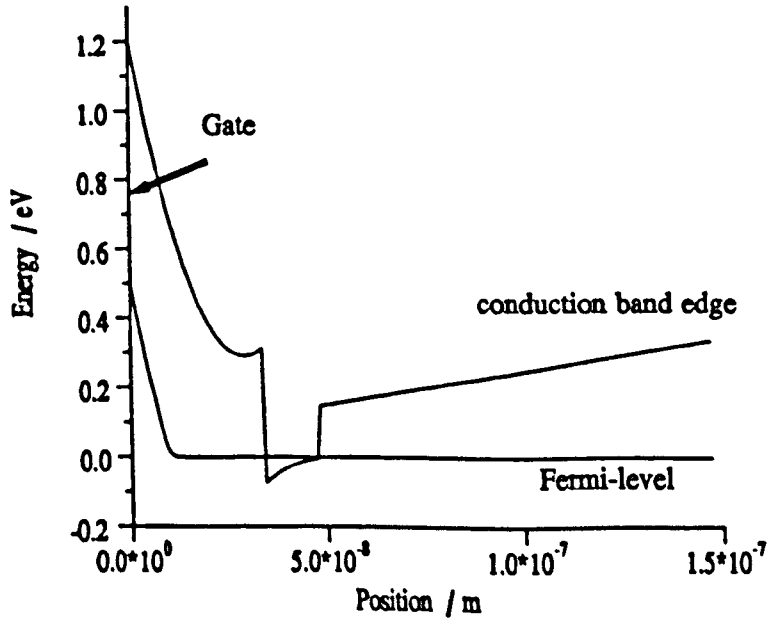


Figure 4.20

Fermi-level and conduction band edge of a pseudomorphic HFET

The simulation includes current flow, with the gate reverse biased at -0.5 V

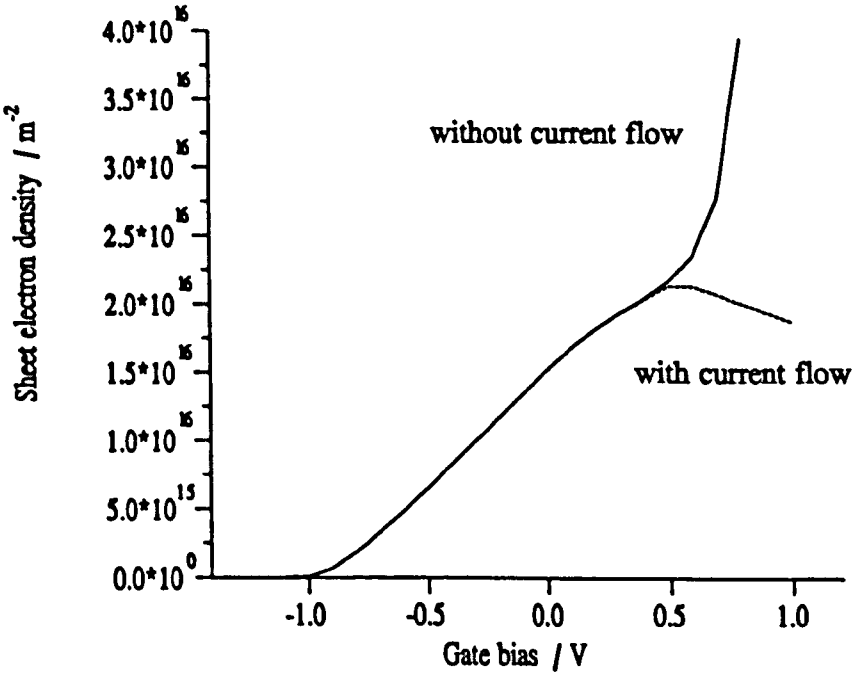


Figure 4.21

Sheet electron density versus gate bias for one-dimensional current and Schottky gate models

It is worth noting that only a drift-diffusion current was included, with a much greater current expected if tunnelling was allowed. A greater current implies a larger gradient in the Fermi-level, which would be produced by reducing the sharp drop present at the gate distributing it throughout the rest of the device. Even so under the gate this gradient would still be small compared with the electric field and thus the simple model excluding charge transport is still valid, although probably over a slightly smaller range.

Chapter 5

Two-Dimensional HFET Model

5.1 Introduction

This chapter describes the equations and numerical schemes used in the two-dimensional models. It describes how the quantum mechanical equations are solved over two-dimensions to obtain the two-dimensional electron density, and how this effects the transport within the device. Also a review of the charge transport equations is given and the numerical schemes required to solve them are discussed. The boundary conditions used in the two-dimensional models are presented and the methods used to obtain convergence when the scheme becomes very ill-conditioned. Finally this chapter concludes with some results and discusses their relevance to the future of HFET modelling.

5.2 Quantum mechanics in two-dimensions

The rigorous solution of Schrödinger's equation in two-dimensions is non-trivial, and certainly is not as straight forward as the solution in one-dimension. The principal problem associated with the solution over two-dimensions is related to the form of the quantum transport equation, (5.1).

$$J_{quantum} = -\frac{iq\hbar}{2m^*}\left\langle (\nabla\xi)^*\xi - \xi^*(\nabla\xi)\right\rangle \qquad (5.1)$$

For the net current to be real the value of the term within the angle brackets must be complex; $i^2 = -1$. This is satisfied if the wavefunction, ξ, has solutions in the form of travelling waves, equation (5.2).

$$\xi(k) = \sum_{k} \alpha_k e^{ik.r} \qquad (5.2)$$

Taking the gradient of the ξ with respect to r would then return $ik\xi$, and hence the term within the angle brackets would indeed be complex. The first problem associated with this solution is that a wavefunction formed from travelling waves is itself complex, and hence the solution of the Schrödinger's equation would require much more computational effort. This would complicate the code and although not straight forward would in itself not produce an intractable problem. However the second problem is much harder to solve. This relates to the boundary condition of a travelling wave. For the bound state solution produced by the one-dimensional scheme the boundary conditions for the wavefunctions are straight forward, the wavefunctions tending to zero at both extremes. However when current is present this condition is certainly not applicable at the terminals, and possibly not applicable on the artificial internal boundaries perpendicular to the heterojunction. There is no clear method for resolving this problem and although there would be the additional benefit that a rigorous solution of Schrödinger's equation automatically guarantees current continuity the full two-dimensional equations have to be simplified to produce a tractable problem.

A simpler scheme has been devised whereby Schrödinger's equation is solved over two-dimensions by taking multiple solutions of the one-dimensional equation in slices across the heterojunction interface, Figure 5.1, and assuming the electron wavefunctions parallel to the heterojunction have the form of travelling wave solutions. This assumption can be justified by considering the size of the 'events' the electron encounters. Perpendicular to the interface the conduction band edge changes over approximately 10 nm, just small enough to produce bound states with energies that are separated by more than $k_B T$, the thermal energy. In the direction parallel to the interface however, no events on such a small scale occur. The gate, the smallest feature in this direction, is at least an order of magnitude larger than the debroglie wavelength, the minimum size required to produce quantised motion. Consequently the electron wavefunctions in these directions can be assumed to be formed from Bloch states and the two-dimensional electron density derived using the equations from Chapter 2 and *Appendix A*.

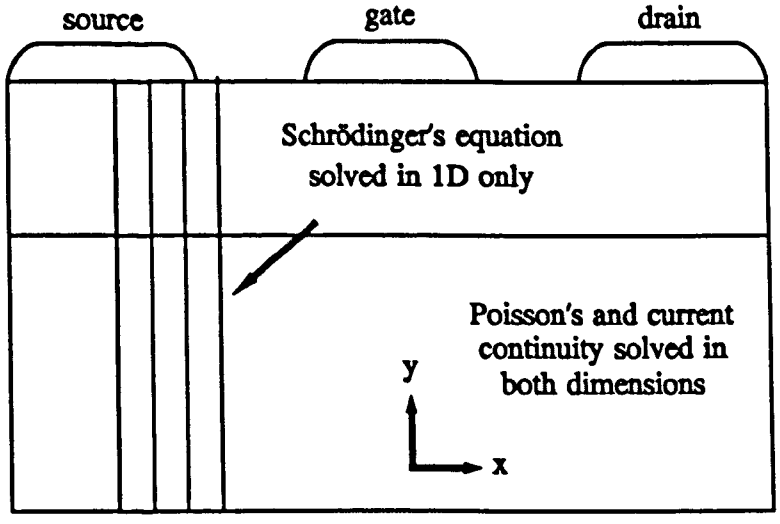


Figure 5.1

Diagram of the one-dimensional slices over which Schrödinger's equation is solved

The quantum solver stops once the separation between consecutive bound states drops below $k_B T$, and the contribution to the total electron density beyond this point is calculated from a Fermi-integral as described in section 2.7.6. This leads to a natural partitioning of the charge and its respective degrees of freedom, also it is worth noting that although there is current flow in the direction normal to the heterojunction interface it is not inconsistent with the chosen solution of Schrödinger's equation, even though the wavefunctions in this direction are not complex. The only electrons participating in transport in the y-direction are derived from the classical Fermi-integral with the electrons associated with the Schrödinger equation solution restricted motion parallel to the heterojunction.

5.3 Two-Dimensional Boundary Conditions

In the full 2D model there are essentially four different types of boundary, ohmic and Schottky contacts, the free surface and the artificial internal boundaries, illustrated in

Figure 5.2.

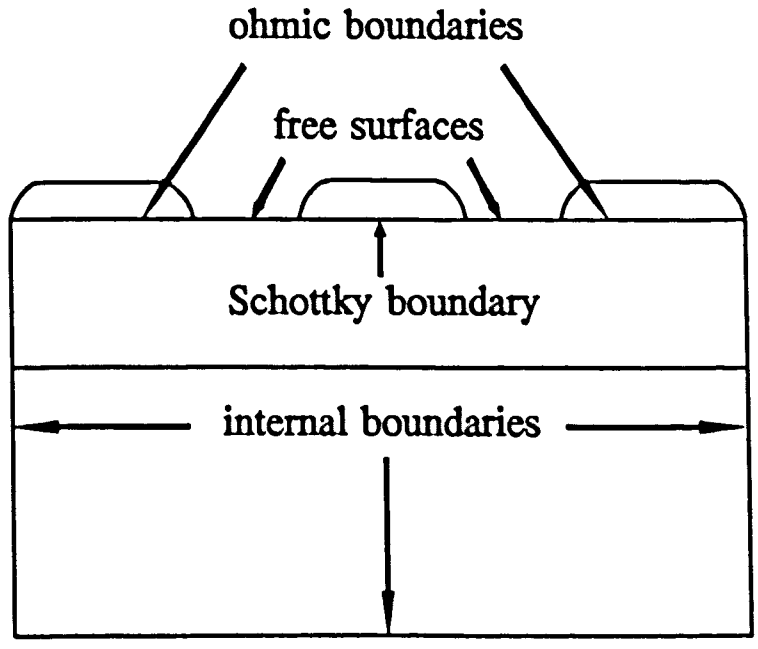


Figure 5.2

Boundaries present in the two-dimensional model

Schottky barriers place the semiconductor in intimate contact with a high conductivity material, the gate metal, and hence it is reasonable to assume, due to the low resistance of the gate metal that the Fermi-level on the Schottky surface is fixed constant at the gate bias. Across the junction there exists a discontinuity in the conduction band edge, principally determined by the surface states [Sze, 1981], which pin it on the semiconductor side to the Schottky 'built in' bias, V_{bi}, above the Fermi-level, Figure 5.3.

Ohmic boundaries on GaAs form similar energy band profiles to the Schottky gate, but in this case a highly doped region is created in the semiconductor next to the contact resulting in a spike in the conduction band edge that is thin enough for a considerable tunnelling current to occur. Thus they form low resistance non-rectifying contacts. Again one can expect the Fermi-levels to align either side of the junction, at a constant potential due to the very low resistance of the contact metals.

As discussed in the section 5.2 problems occur with the solution of Schrödinger's equation at this contact due to the large currents flowing in the direction perpendicular

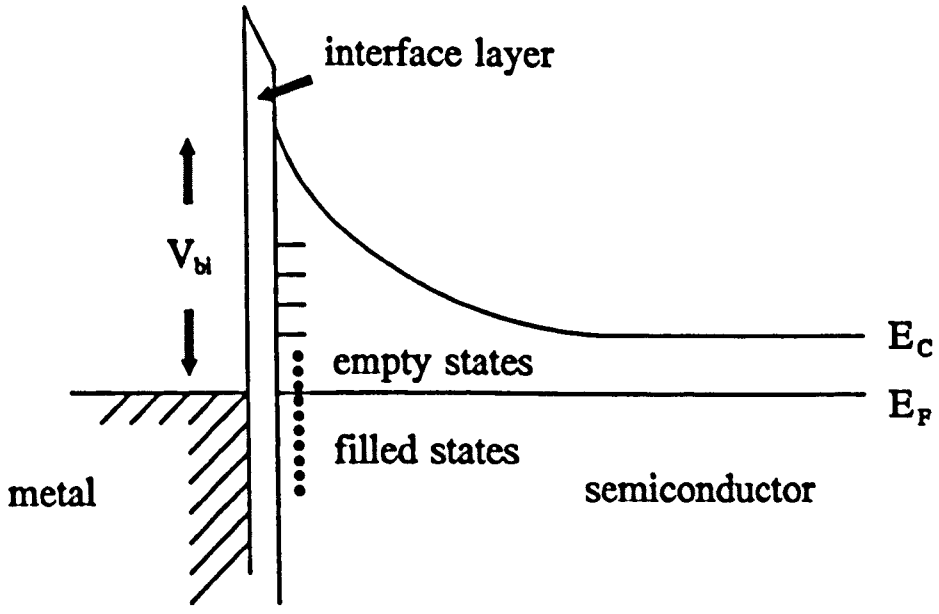


Figure 5.3

Schottky gate energy band diagram showing the empty and filled surface states

to the heterojunction. In these regions the true solution of Schrödinger's equation will produce Bloch state wavefunctions. Consequently the approximations used to separate the full two-dimensional solution of Schrödinger's equation into a quasi-two-dimensional form are no longer applicable. The model initially simulated HFETs by avoiding this inconsistency, separating the modelled region into three sections. The first and last sections, under the source and drain contacts respectively, were modelled employing the classical approximation for the electron wavefunctions and hence the electron density was evaluated solely via the Fermi-integral from the conventional conduction band edge. The middle section, under the gate, possesses much less current flow across the heterojunction and hence was modelled by explicitly calculating the electron wavefunctions from the quasi-2D solutions of Schrödinger's equation. This scheme was inherently ill-conditioned, with large discontinuities in the electron density occurring at the junctions between the two types of region. This ill-conditioning was reduced by truncating the modelled region, between the Schottky and ohmic contacts, and setting the Fermi-level on edges of the modelled region constant at the source and drain biases. The potential lost over the source and drain access resistances is then added to the applied bias after the device equations have been solved. This scheme is justified by the relatively low access resistances possessed by FETs and the large pad areas that are

typically employed. The low resistance means that very little potential is dropped between the ohmic contacts and the artificial boundary, and the large pad area means that the potential across the slice will be approximately constant. This assumption can be readily checked when the equipotential lines produced by conventional two-dimensional models are considered, and it is observed that in these regions the equipotential lines are approximately parallel and perpendicular to the interface[1]. This produces a model domain similar to that shown in Figure 5.4.

At the free surface one can assume the material above the semiconductor is non-conducting, hence there is no current crossing this boundary, restricting the current flow to within the semiconductor material. As current is characterised by equation (5.3).

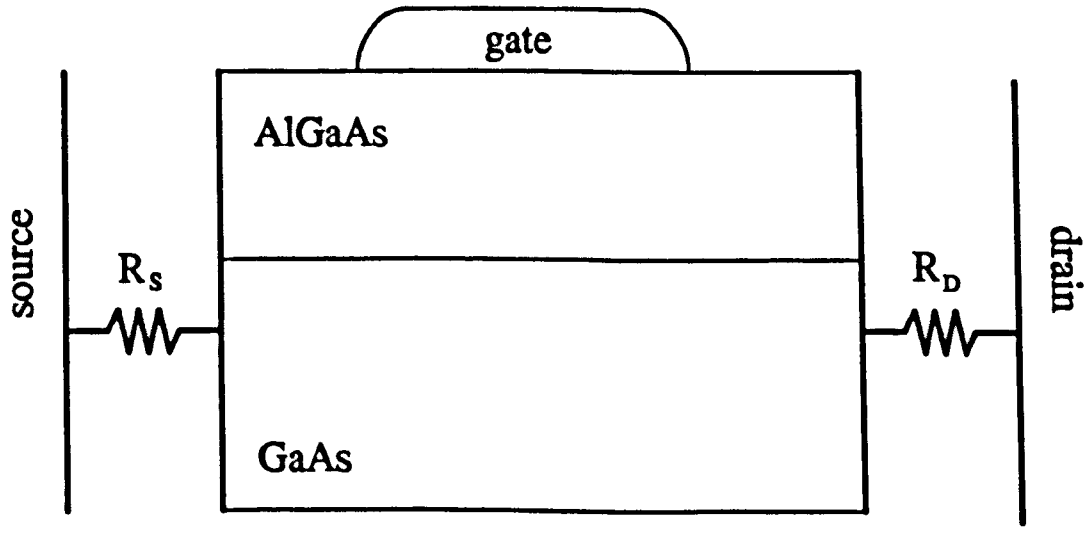


Figure 5.4

Diagram of modelled domain and the connections to the source and drain via the lumped access resistances

[1] It is also worth noting that this approximation is employed in quasi-two-dimensional models which have been shown to produce very accurate results [ref]

$$J = \mu q n \nabla \phi \tag{5.3}$$

it is apparent that $\nabla\phi.n = 0$, where n is the unit vector normal to the surface. The conduction band edge at this boundary is solved via Poisson's equation, assuming that the normal component of electrostatic field above the surface is zero. Note this surface will have a relatively high trapping density due to surface oxidation/contamination and crystalline defects.

Finally the artificial internal boundary in the substrate is chosen arbitrarily such that there is no change in the device characteristics if it is moved. It is obviously unnecessary and wasteful to model the 'whole' device, extending to regions that have no effect on the transistor operation. It is therefore necessary to truncate the model dimensions, only including the important regions. One suitable boundary condition assumes these areas are far enough from the active regions to be in equilibrium and thus to forbid current flow across these boundaries. If current were permitted to cross these boundaries then they are not in equilibrium and it is probable that moving them would effect the simulated characteristics. If this boundary condition distorts the current flow within the device then the internal boundaries have been poorly chosen and should be moved further from the active region. Thus, as on the free surface, the Fermi-level is set such that $\nabla\phi.n = 0$ and the conduction band edge by solving Poisson's equation with the normal component of field is set to zero on the boundary.

Before the truncated device domain was implemented a method was developed to reduce the singularities that occurred at the quantum/classical interfaces. This scheme blended the two solutions for the electron density, smoothly changing from classical to quantum and back to classical, Figure 5.5. Although this improved the performance of the model, making it more robust, the scheme was still ill-conditioned and failed to converge for most biases. This method of blending was also used on the contacts, merging the Neuman and Dirichlet boundary conditions. Conventional device models switch abruptly between the two surface types producing sharp singularities in the solutions at these points. It is difficult to believe that such abrupt change physically exist, especially when the device manufacture process is considered, diffusing alloys into the semiconductor

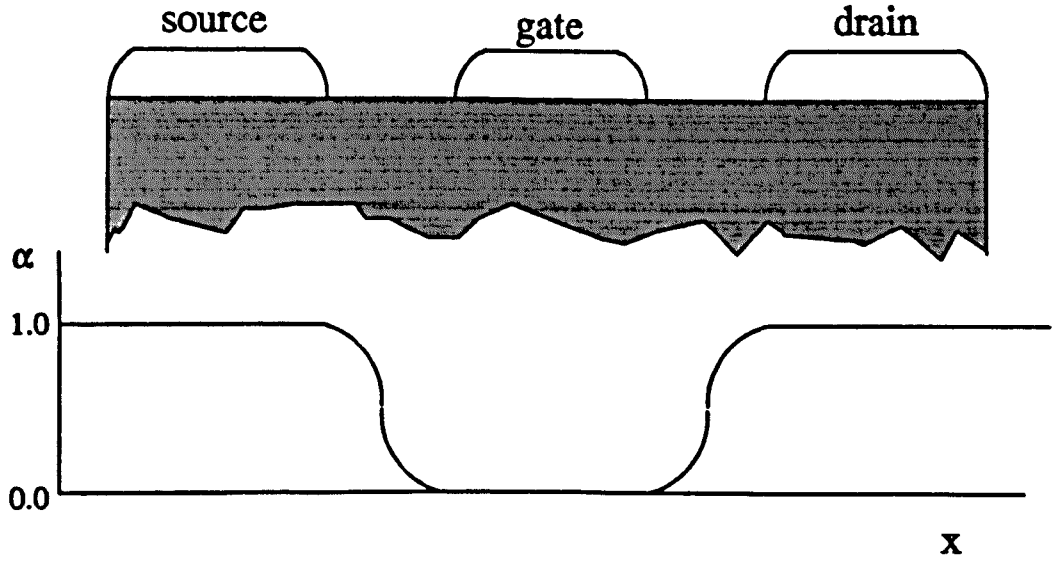


Figure 5.5

Variation of the fitting parameter α with lateral position

material. Thus a more physical description smooths the change in boundary conditions, allowing some nodes to have a fraction of one boundary condition and the remainder in the other boundary condition. The smoothing is achieved by assigning a value between 0 and 1 to a variable representing the contribution of Dirichlet boundary conditions. Thus a value of 0 implies 100% Neuman, 1 100% Dirichlet and otherwise a blend of the two types, equation (5.4).

$$\text{Boundary condition} = \alpha\ \text{Dirichlet} + (1-\alpha)\ \text{Neuman} \qquad (5.4)$$

This also allows a very general and elegant description of the surface, and if the internal equations are substituted for 'Neuman' in the above equation then the Dirichlet biasing at the ohmic contacts can be extended into to device, below the surface.

This scheme makes the solution process more robust by decreasing the size of any singularities that may occur at the contact edges, hence it is applied for numerical reasons, although as the above argument illustrates it should produce a more physical

simulation.

The two-dimensional FET model solves a similar set of equations to the one-dimensional Schottky gate model. Poisson's equation is expanded to include the partial derivatives with respect to x, thus producing a Jacobian sub-matrix with a pentadiagonal structure as opposed to the tridiagonal format produced in the one-dimensional scheme. Schrödinger's equation is solved at each vertical line of nodes using the same techniques deployed in the one-dimensional model, and finally the current continuity equation using Newton's method applied to the discretization described in the next section.

5.5 Discretization of the charge transport equations

As derived in chapter two the steady-state charge transport in the absence of holes is given by equation (5.5).

$$\nabla . J = 0 \tag{5.5}$$

where J, the electron current is given by equation (5.6).

$$\nabla . J = \nabla . (\mu q n \nabla \phi) = 0 \tag{5.6}$$

Two schemes were originally considered for the discretization of this equation. The first explicitly expanded the equation in terms of its grad and div.grad operators, equation (5.7).

$$\nabla . J = \mu n \nabla^2 \phi + \nabla(\mu n) . \nabla \phi = 0 \tag{5.7}$$

Here q, the electronic charge is dropped from both sides. The μn term varies exponentially with position and consequently the $\nabla(\mu n)$ tern cannot be formed from a simple linear discretization. This term was evaluated by again expanding it, but this time in terms of the independent parameters, ϕ and ψ, giving equation (5.8).

$$\nabla(\mu n) = \frac{\partial(\mu n)}{\partial\psi}.\nabla\psi + \frac{\partial(\mu n)}{\partial\phi}.\nabla\phi \qquad (5.8)$$

This expansion produces grad terms operating upon linear variables, thus the simple linear finite differences can be used without undue errors. Thus there are no 'Bernoulli-like' terms present that are required in more conventional schemes, where n is chosen as an independent variable. One disadvantage of this method is that when the Newton scheme is used to solve the equations it requires the dependent variables μ and n to be differentiated twice with respect to the independent parameters ψ and ϕ, once for equation (5.8) and once when this equation is differentiated to be used in the Jacobian. This discretization method was implemented, and although it worked well for low biases and has the advantage that the current continuity equation is simply a linear function of the grad and div.grad operators, as discussed in section 4.2.2, it was found to be poor when modelling the saturated regime of the device characteristics. The discretization was found not to conserve current, which was later attributed to a poor approximate of the derivative of μ[2]. In the drift-diffusion model μ is a function of three variables, ψ, ϕ and $|E|$, the electrostatic and Fermi potentials and the absolute value of the electrostatic field respectively. The partial derivative of μn with respect to position, used in equation (5.7) should therefore take the form of equation (5.9).

$$\nabla(\mu n) = \frac{\partial(\mu n)}{\partial\psi}\nabla\psi + \frac{\partial(\mu n)}{\partial\phi}\nabla\phi + \frac{\partial(\mu n)}{\partial|E|}\nabla|E| \qquad (5.9)$$

This last term is itself a function of ψ ($E = -\nabla\psi$) was not taken explicitly into account

---

[2] Newton's scheme itself can operated even if inaccurate guesses for the elements of the Jacobian matrix are used but when the function itself includes a partial derivative, this term must be evaluated accurately to produce an accurate overall solution.

in the original scheme, and became very important in regions were the electrostatic field varied rapidly. This effected the current continuity, producing the spikes on the current plot shown in Figure 5.6, which should be constant between the source and drain contacts. This method should be more easily applied to non-equilibrium simulations where the electron mobility, μ, is a function of ϕ, ψ and w, the Fermi- and electrostatic potentials and electron energy only. Thus the derivative of mobility with respect to each of these parameters is straight forward to calculate.

A better scheme investigated, which is still used in present models, is derived from a

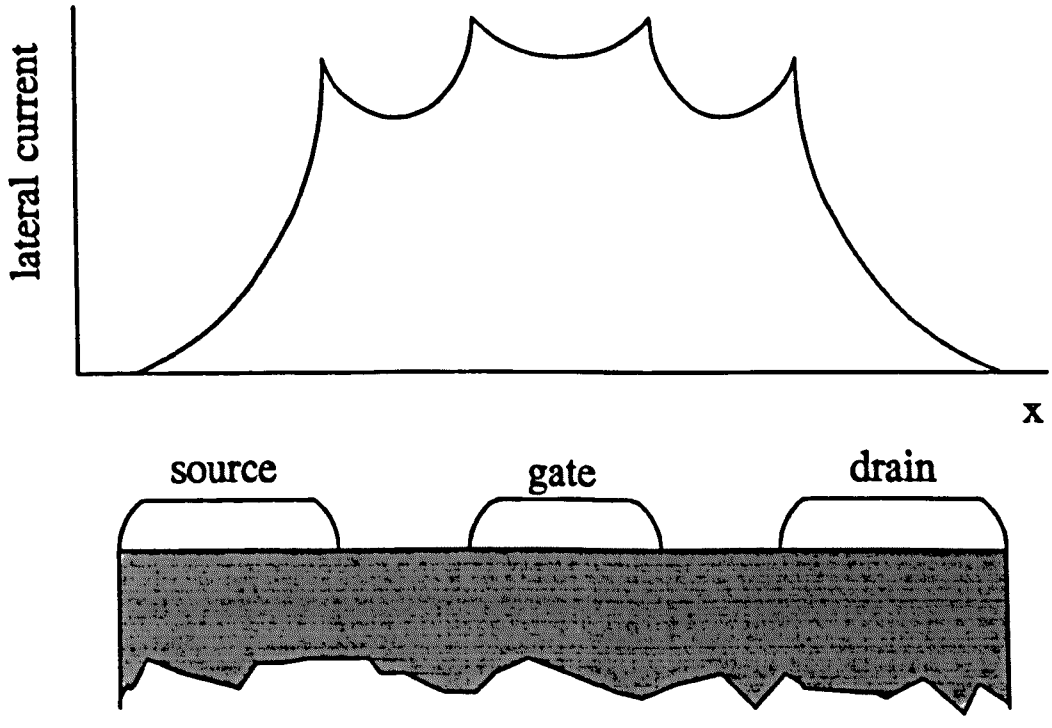


Figure 5.6

The total electron current density in the x-direction

form of Gauss' law and discretizes the current continuity equation at its highest level, equation (5.10).

$$\nabla.J = \frac{J_{x\,(i+1/2,j)} - J_{x\,(i-1/2,j)}}{\frac{h_{(i,j)} + h_{(i-1,j)}}{2}} + \frac{J_{y\,(i,j+i/2)} - J_{y\,(i,j-1/2)}}{\frac{g_{(i,j)} + g_{(i,j-1)}}{2}} \qquad (5.10)$$

where h and g are the internodal separations in the x- and y-directions respectively. If one considers a rectangular Gaussian surface drawn at the half-nodes, Figure 5.7, and assumes the current across each boundary line to be constant then equation (5.10) can easily be derived by integrating the current density normal to the surface.

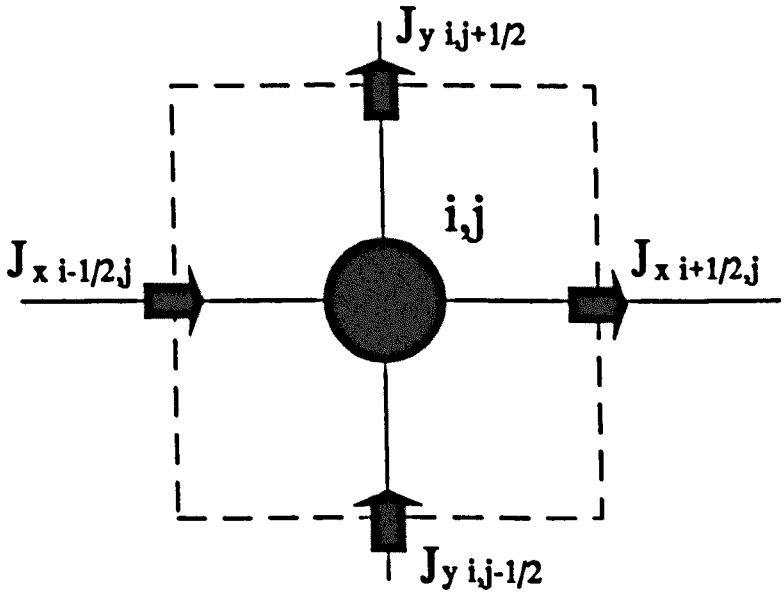


Figure 5.7

Gaussian surface draw around node (i,j) for the solution of the current continuity equation

This method requires the dependent variables μ and n to be evaluated at the half-node positions.

Whilst mobility is a linear variable of position electron density is not, consequently an arithmetic average of this variable is not adequate.

$$n_{i+1/2} = \frac{(n_{i+1} + n_i)}{2} \tag{5.11}$$

The electron density, at least in the classical limit, is an approximately exponential

function of ϕ and ψ, which do vary linearly with position thus a geometric average is potentially applicable though is not strictly correct for anything other than Boltzmann statistics.

$$n_{i+1/2} = (n_{i+1}.n_i)^{1/2} \tag{5.12}$$

Here, if the electron density is expressed in exponential form, the geometric average can be shown to evaluate it at the half node by taking the algebraic or half node values of the independent variable

$$\left(e^{t_i}.e^{t_{i+1}}\right)^{1/2} = e^{(t_i+t_{i+1})/2} \tag{5.13}$$

Since the Fermi-integral and the 2DEG electron density in the classical and quantum regimes respectively are not strictly exponential functions of the electron potential energy the FET model evaluates the mobility and electron density as functions of the half-node potential energy, equation (5.14).

$$n_{i+1/2}(\psi,\phi) = n\left(\frac{\psi_{i+1}+\psi_i}{2}, \frac{\phi_{i+1}+\phi_i}{2}\right) \tag{5.14}$$

As stated earlier these independent parameters ψ and ϕ vary linearly in between the nodes and consequently an arithmetic average is applicable to find the half-node values of these variables. This scheme requires three times the computer memory, storing the values of the electron density at the node, and at the half-nodes in both the x- and y-directions, but is guaranteed to produce a more accurate value for the half-node parameters than either of the two simpler schemes. The partial derivatives of the half-node variables are also returned from the same function, but here they are with respect to the half-node electron potential energy. To describe these derivatives in terms of the independent variables, ψ and ϕ at the node, this half-node potential energy has

to be further differentiated, equation (5.15), (5.16).

$$\frac{\partial n_{i+1/2}}{\partial U_{E_i}} = \frac{\partial n_{i+1/2}}{\partial U_{E_{i+1/2}}} \cdot \frac{\partial U_{E_{i+1/2}}}{\partial U_{E_i}} \qquad \textbf{(5.15)}$$

Since $U_{E\,i+½} = 0.5*(U_{E\,i} + U_{E\,i+1})$ the derivative of $n_{i+½}$ with respect to each independent variable at the node is simply half that of the derivative at the half-node, equation (5.16).

$$\frac{\partial n_{i+1/2}}{\partial U_{E_i}} = 0.5 . \frac{\partial n_{i+1/2}}{\partial U_{E_{i+1/2}}} \qquad \textbf{(5.16)}$$

5.6 Results of a GaAs MESFET simulation

This current discretization scheme was first implemented on the GaAs MESFET structure illustrated in Figure 5.8, omitting the quantum solutions. This was to ensure that the equations solved to the expected values on a simple structure before attempting the simulation of a more complex heterojunction device. Since this simulation omits the solution of Schrödinger's equation the more conventional simulation boundaries were chosen, extending the modelled region underneath the source and drain contacts. This scheme produced the following results, illustrated in Figures 5.9 to 5.12.

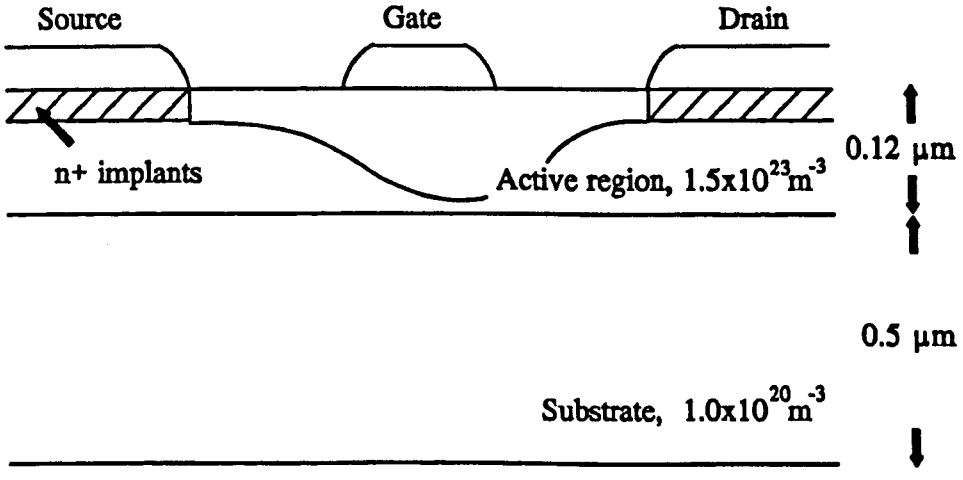


Figure 5.8

0.5 μm gate GaAs MESFET structure

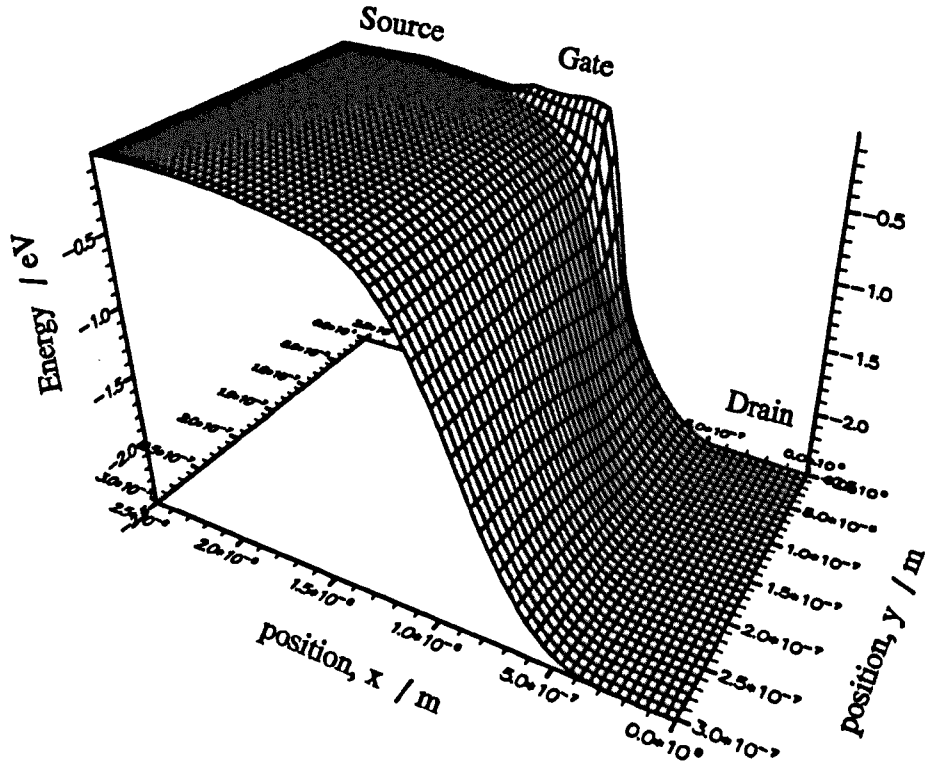


Figure 5.9

Fermi-level for a 0.5 μm gate MESFET

The above Figure displays the Fermi-level for a GaAs MESFET biased at 2.5 V_{DS} and 0 V_{GS}. This parameter is pinned at the source, gate and drain potentials along the top edge of the diagram, with all other surfaces defined using Neuman boundary conditions, preventing current from leaving the simulated domain. It is instructive to note the equipotential lines in the active channel under the gate are nearly parallel and normal to the 'free-surface'. This feature is the underlying principle behind the quasi-two-dimensional FET models described further in Chapter 6.

Immediately under the gate the Fermi-level is pulled up to the gate bias, with the subsequent gradient being the driving force behind the gate conduction current.

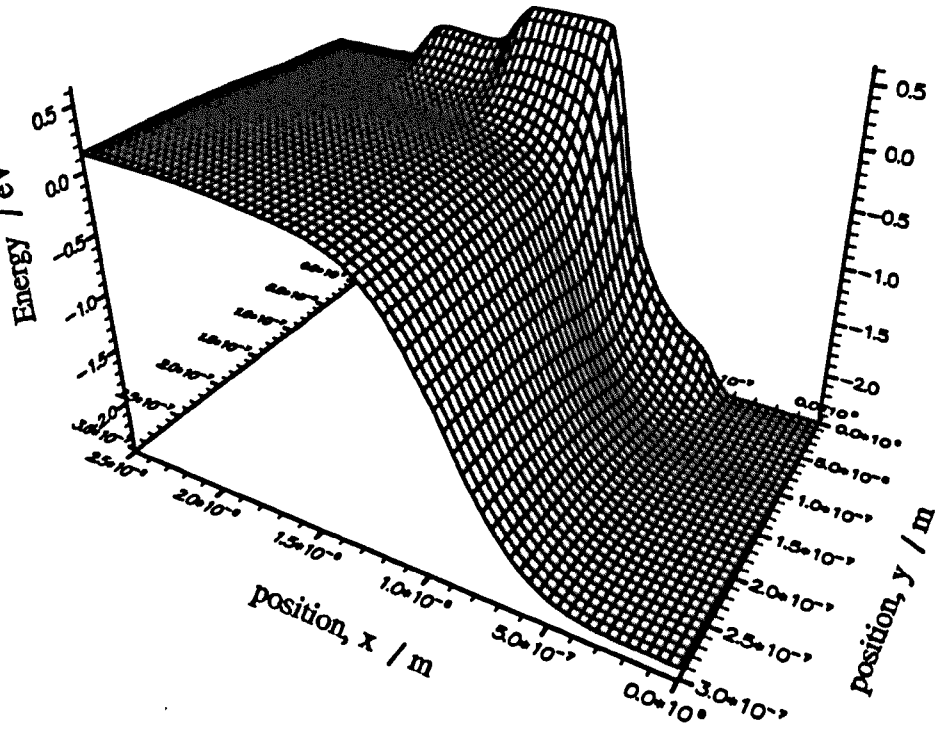


Figure 5.10

Conduction band edge for a 0.5 μm gate MESFET

This figure shows the equivalent conduction band edge to the Fermi-level on the previous page. Here the conduction band edge is raised above the equilibrium value at the surface due to the Schottky built-in bias and the surface traps. The substrate is simply low-doped, and hence is only raised slightly above that of the active channel. If traps were to be included in this region the substrate would become semi-insulating and the conduction band edge would be raised further and hence produce a much greater confining potential.

The conduction band at the source and drain contacts is degenerate, lying below the Fermi-level due to the high doping present immediately underneath these contacts.

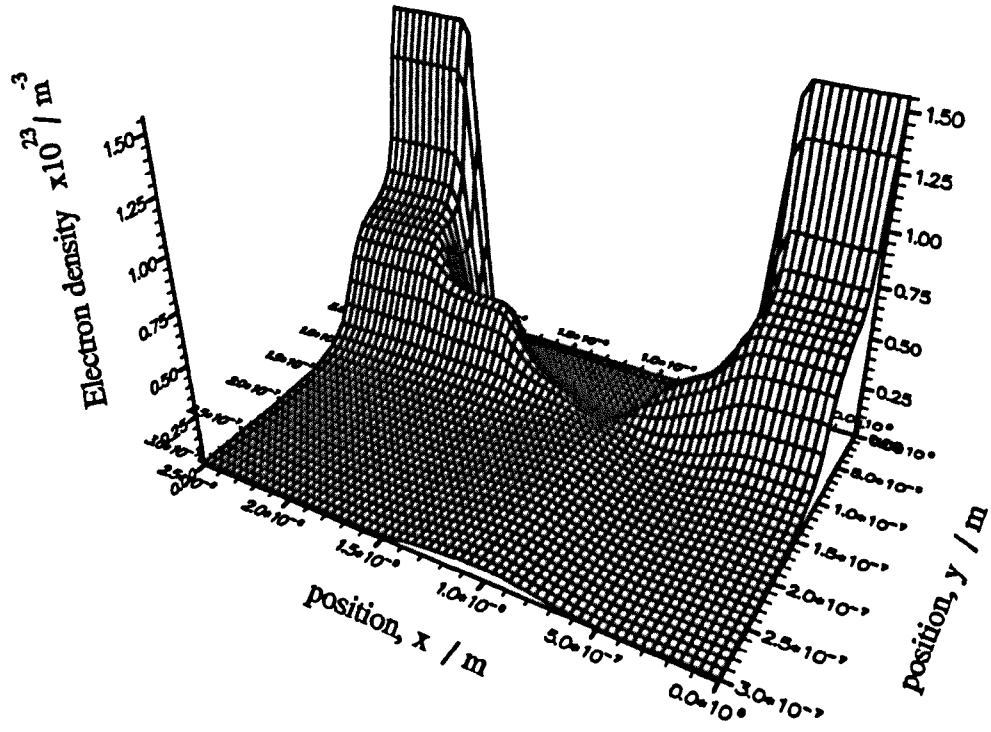


Figure 5.11

Electron density for a 0.5 μm gate MESFET

This figure again illustrates the MESFET biased at 2.5 V_{DS}, 0 V_{GS} but here the electron profile is displayed. The high doping under the ohmic contacts produces the high electron densities at these positions, which rapidly reduced to the ambient active channel level and then lower still as the substrate is entered. The surface traps and 'built-in' gate bias deplete the free electron density on the top surface, which extends approximately 0.1 μm into the active channel. Under the gate, especially at the drain edge, the source-drain bias tends to reverse bias the gate further extending the depletion region almost to pinch off. At this point the effect of substrate injection is apparent, increasing the electron density in the substrate by several orders of magnitude.

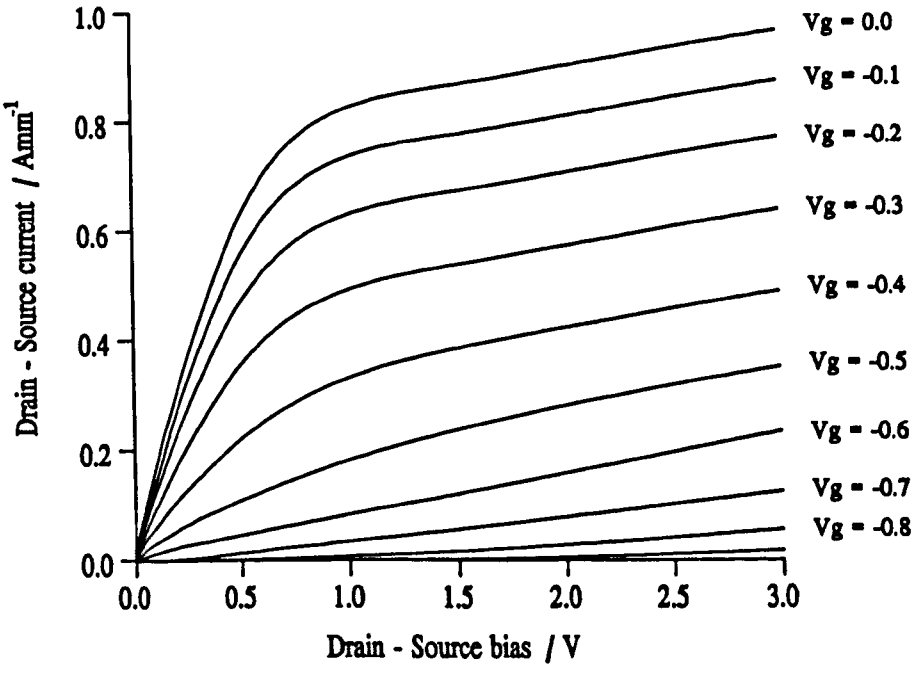


Figure 5.12

I_{DS}-V_{DS} characteristics for a 0.5 μm gate MESFET

Finally the I_{DS}-V_{DS} characteristics for this device are shown in Figure 5.12. These curves appear correct, showing a typical transconductance, pinch-off and output conductance. This model has no problems with convergence, with the source-drain bias taken greater than 10 Volts and the source-gate bias varied over several volts. The overall solution scheme is very fast solving the above curves, requiring 200 individual solutions in approximately 1 hour 20 minutes, which is at least an order of magnitude faster than the more conventional time-dependent Gummel based algorithms.

Overall these results compare very well with those expected for a GaAs MESFET, which can be determined from the results presented by [Barton, 1988] and [Lee, 1993] in their respective theses.

5.7 Results of an AlGaAs/GaAs simulation

The scheme incorporating quantum mechanics was then used to simulate the single channel conventional AlGaAs/GaAs HFET structure illustrated below. At present the simulation does not include a gate recess model, and hence only planar structures were modelled.

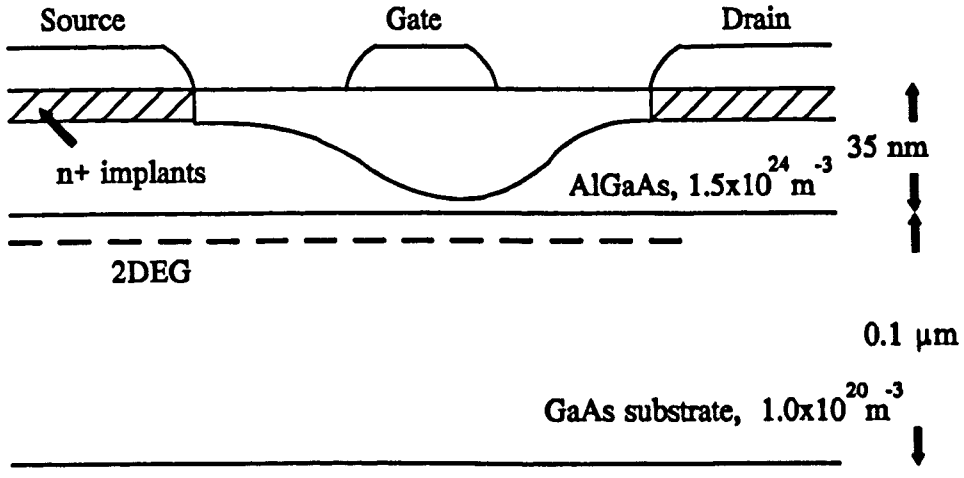


Figure 5.13

0.5 μm gate single channel HFET structure

This simulation had many problems associated with numerical stability and convergence, the results of which are apparent in the I_{DS}-V_{DS} curves, Figure 5.20, where the quantum simulation is truncated above 1.5 Volts. This instability is linked to the inclusion of quantum mechanics, and the resulting discretization that occurs. When small changes in potential energy can alter the number of bound states a particular potential profile can support it is not surprising the scheme has difficulties converging. This effect is particularly prominent once the drain-source bias is sufficient to 'saturate' the output current. Here the potential well at the drain edge of the gate is pulled out and hence cannot support a 2DEG. Either side of this point the potential well gradually returns,

supporting one and then more bound states. At these positions the number of 'quantum' solutions is particularly prone to oscillation, with the residual error in the global device equations correspondingly unstable. This can be attributed to a functional that has either local maxima, minima or an inversion point between the current estimate for the solution and the true value of the root. Note also that there may even be multiple roots.

The situation can be improved by reducing the increment in the drain-source bias, and hence decreasing the probability that a maxima/minima or inversion point exists between the current estimate and the true root. This was incorporated into the two-dimensional FET model, which incorporated an algorithm that reduced the drain-source increment if a satisfactory solution is not found within N iterations, where N was typically between 50 and 100. This allowed the scheme to converge for slightly higher biases, increasing the range by approximately 10%. This modification has a limited application since beyond a certain bias the increments became prohibitively small, producing a FET model that was impractical for device design.

Alternative solution schemes were also tried, in an effort to reduce the effect of the poor solution curve. These incorporated Gummel's algorithm at different levels. The first method, already mentioned in Chapter 3, reduces the global solution scheme to a local one, whereby a solution is found for each node, stepping through the device mesh in sequence. This corresponds to using Newton's algorithm on a local scale and linking consecutive local solutions via Gummel's scheme. The second method solved each global equation, Poisson's and current continuity, independently using Newton's algorithm and then successively iterated these solutions within Gummel's scheme until a self-consistent solution was obtained. Both these alternatives worked for low biases, although not as quickly or robustly as the original solution scheme, and both failed at higher biases for the same reason. Gummel's algorithm requires that the equations being solved to be reasonably well behaved, otherwise excessive relaxation factors have to be incorporated to prevent oscillation. The Poisson and current-continuity equations, when solved self-consistently with Schrödinger's equation become very ill-conditioned, especially at higher drain-source biases. Hence both the above schemes had to be relaxed to such an extent that they again were impractical for device simulation.

One further improvement that increased the numerical stability of the scheme was to simply decrease the inter-nodal spacing. The finer meshes produced in this case were able to extend the maximum drain-source bias by a further 10-20%. This improvement is then limited by the machine size and speed.

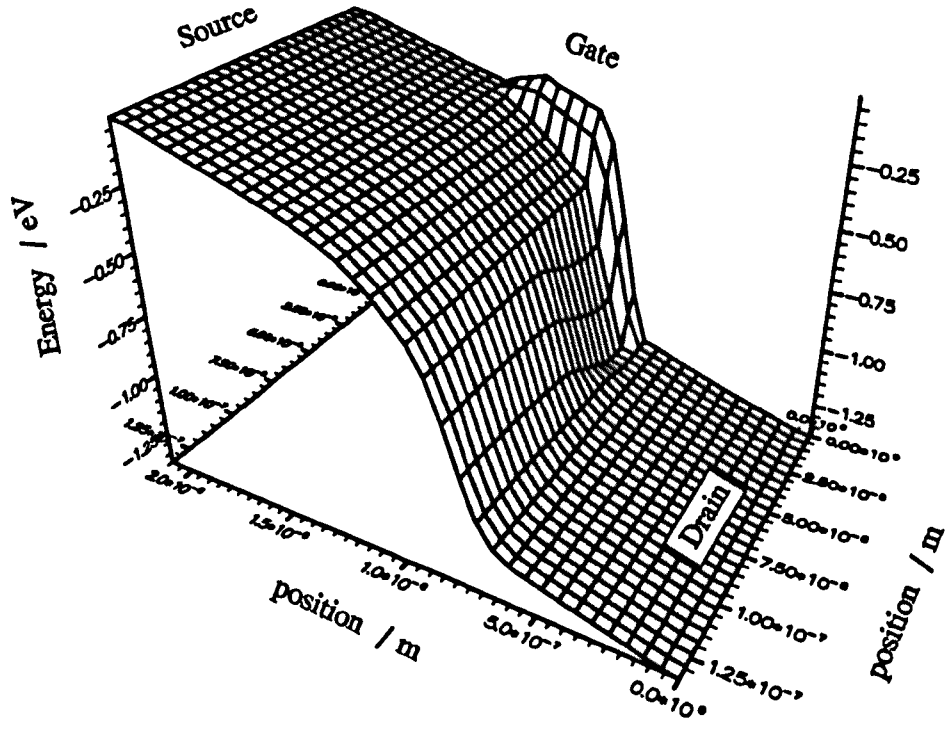


Figure 5.14

Fermi-level for a single-channel HFET

V_{GS} = 0.0, V_{DS} = 1.7

This figure illustrates the Fermi-level for the 'quantum' solution, biased at 1.3 V_{DS}, 0 V_{GS}, just beyond the 'knee' and entering into the saturated current region on the I_{DS}-V_{DS} curves. This profile is very similar to the MESFET curve, at least in shape, with the biases being pinned at the source, gate and drain contacts.

The HFET Fermi-level would differ from the MESFET curves if the source and drain contacts were modelled in a more conventional way, placing them on the top edge of the modelled domain. In this case, because of the significant current in the *y*-direction,

the Fermi-level would be split into two sections. The top layer, in contact with the top surface would retain its present shape, but the lower section, below the heterojunction would be altered. Here, because of the spike in the conduction band edge there is a corresponding low electron density, and hence a large access resistance to the 2DEG. To maintain a significant current through this region the driving forces (the gradient of the Fermi-level) must compensate for the low carrier density. Thus a significant proportion of the drain-source bias would be dropped over this region, and hence the 2DEG would experience a lower driving force. This effect will be reduced because of the large source and drain pad area, although it may still exist at the gate edges of these contacts. To model the whole ohmic pad area would be prohibitive due to the extra computational demands, and hence the scheme used in the present model was adopted whereby the regions around the ohmic contacts are modelled via lumped access resistances.

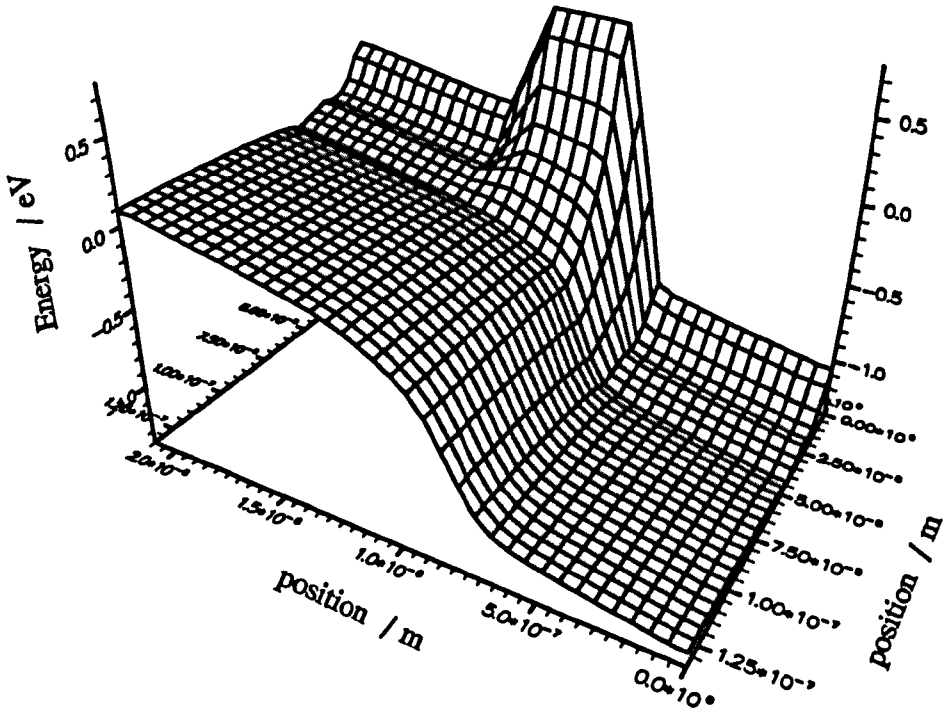


Figure 5.15

Conduction band profile for a single-channel HFET

$V_{GS} = 0.0$, $V_{DS} = 1.7$

Figure 5.15 shows the conduction band profile for the single channel HFET. Here the modelled domain is truncated to within 0.8 μm either side of the gate, with the true source and drain contacts connected via 1.7 Ω lumped access resistances. The local source-drain bias (omitting the lumped access resistances) is 1.3 Volts. The conduction band profile clearly illustrates the abrupt change across the heterojunction, although the spline routine used to produce the graphics has tended to smear out this discontinuity. It is also visible that the two-dimensional potential well is pulled up at the drain edge of the gate, and this corresponds to a reduction in the number of bound states. This effect increases as the device is biased, and at the point shown the well will not support any 'bound' states, and hence all the electrons at this point are derived from classical mechanics and free to travel in both the x- and y-directions.

Again the gate and surface pull the conduction band edge up, as in the case of the MESFET, guaranteeing the electrons wavefunctions are bound at this surface.

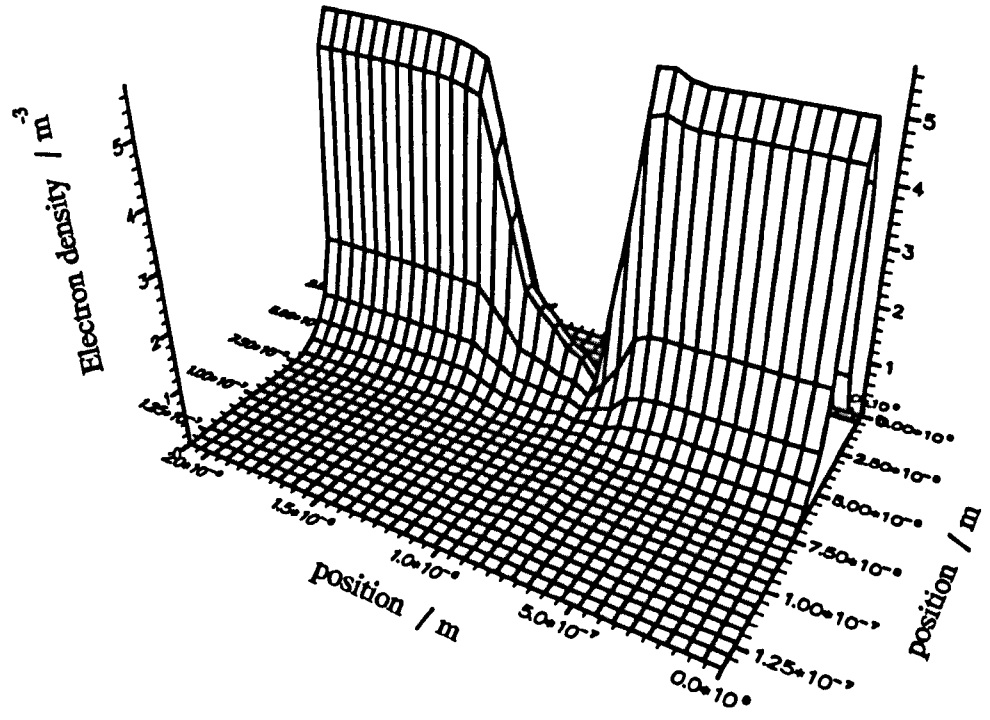


Figure 5.16

Total electron density of a single-channel HFET

$V_{DS} = 0.0$, $V_{DS} = 1.7$

Figure 5.16 shows the total electron profile for the device at the same bias point as the conduction band profile on the previous page. The 'spike' of electrons in the 2DEG is clearly visible at the heterojunction interface. In the quantum case the 2DEG extends into the AlGaAs region than with classical mechanics, penetrating the high potential of this region.

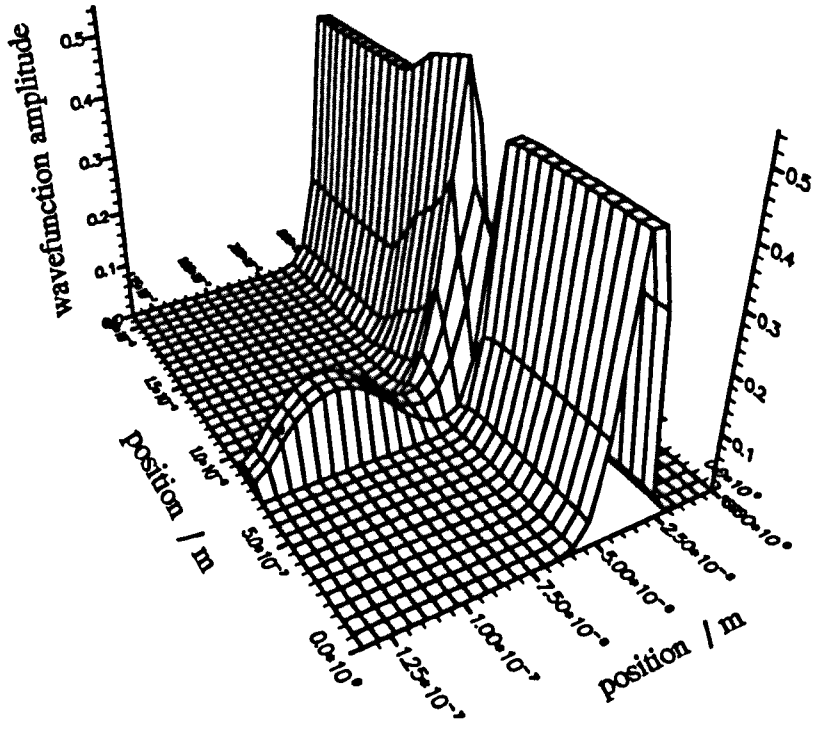


Figure 5.17

First bound state wavefunction for a single-channel HFET

The y-component of the first three bound states for the single channel HFET are shown in Figure 5.17, Figure 5.18 and Figure 5.19. Apart from the drain edge of the gate these take on the form illustrated in Chapter 4, increasing the number of peaks with bound states. The figures also show the wavefunction solutions at the drain edge of the gate, even though in the model this information is omitted and the electron density calculated from the classical Fermi-integral. At this point, as discussed earlier the potential well is drawn out to such an extent that it will not support any 'bound' states, and consequently the wavefunctions illustrated at this point are incorrectly bound at the substrate boundary, where in reality they take the form of travelling waves.

The other two components of the wavefunctions (in the x- and z-directions) take the form of Bloch state travelling waves over the whole modelled domain.

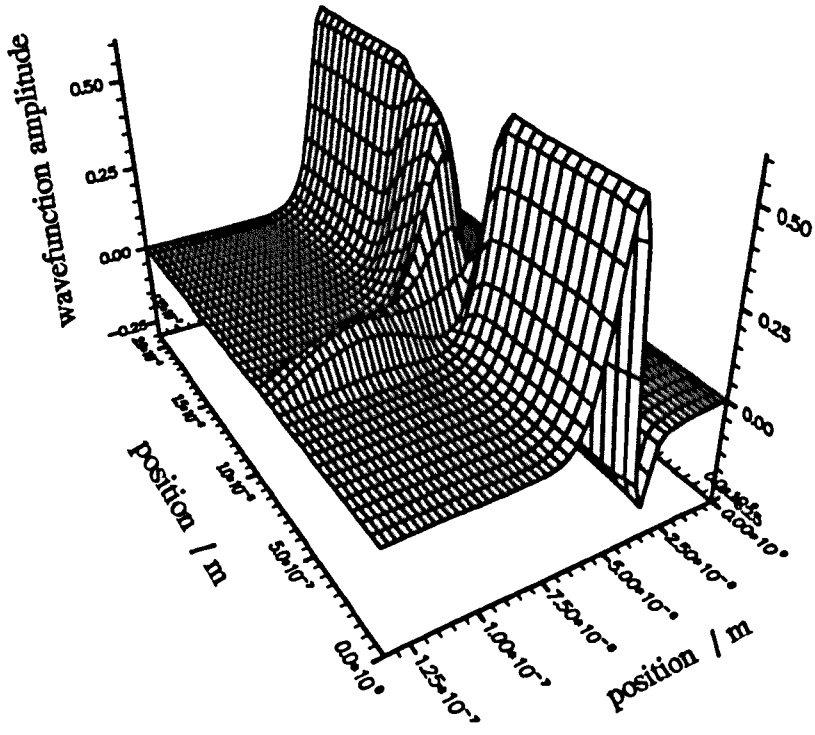


Figure 5.18

Second bound state wavefunction for a single-channel HFET

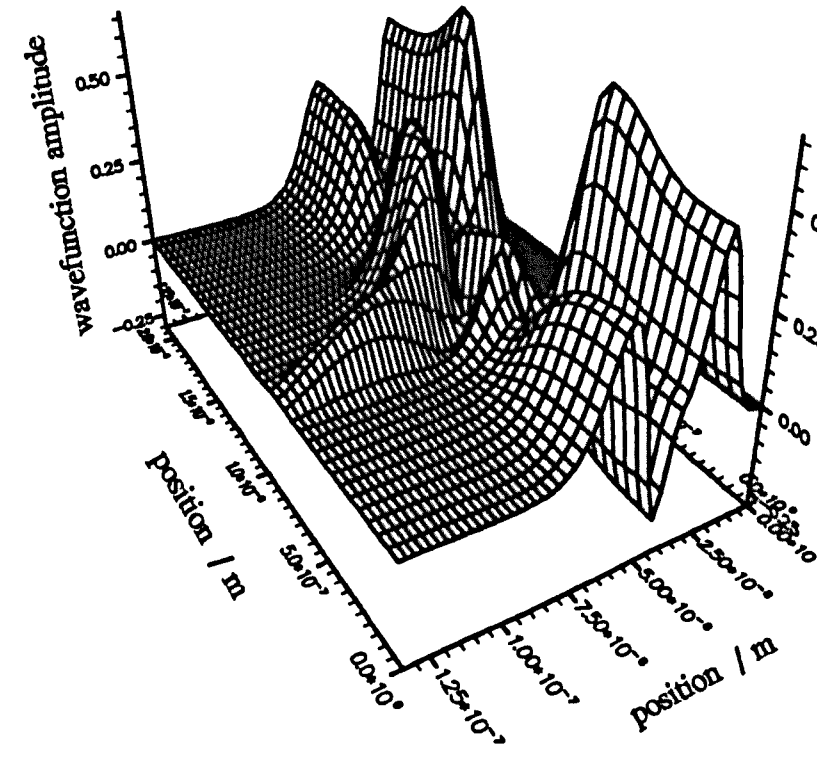


Figure 5.19

Third bound state wavefunction for a single-channel HFET

Figure 5.20 shows the I_{DS}-V_{DS} characteristics for the HFET device shown in Figure 5.13. The quantum and classical solutions are compared over a range of drain-source and gate biases. Generally the curves are similar, one would not expect too great a change as current classical schemes produce reasonably accurate simulations. Here the quantum current approximately 20% lower than the classical case. This drop in current is attributed to the lower free electron density produced in the quantum simulation, and not the restriction in the degrees of freedom allowed for the 2DEG electrons.

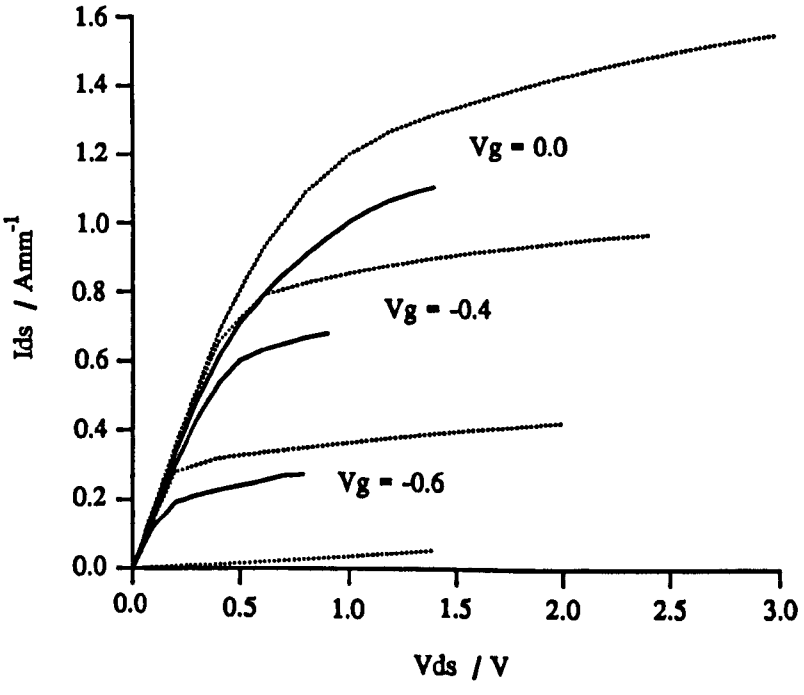


Figure 5.20

I_{DS}-V_{DS} Characteristics for a single channel HFET

The same device was also modelled changing the maximum number of bound states and the degrees of freedom of the 2DEG. Figure 5.21 illustrates these results. The maximum number of bound states was unset for most simulations, with the typical number calculated around 3 or 4. The 2DEG was allowed to participate in transport in the *y*-direction in two cases, labelled "quantum, 2D motion, 1 bound state" and "quantum, 2D motion". The former of these was restricted to one bound state only, with all the rest of the electrons evaluated by classical statistics. When the 2DEG was allowed to move in the *y*-direction the current increased slightly due to the increased freedom within the

system, but even when there was only one bound state, and hence the incorporation of quantum mechanics was minimal, the current didn't approach the classical solution.

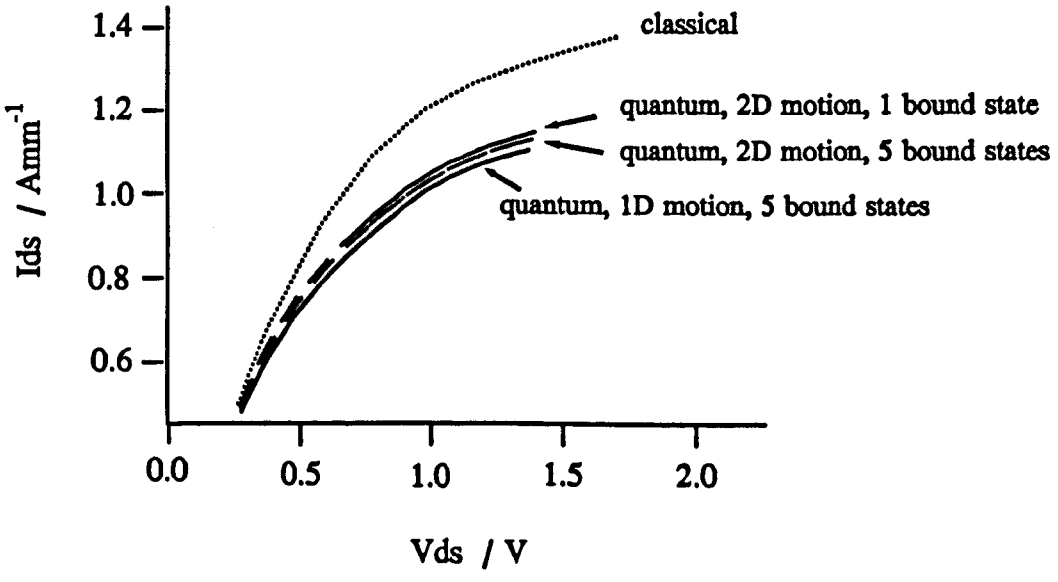


Figure 5.21

Comparison of the classical and quantum I_{DS}-V_{DS} curves

5.8 Results of a pulse-doped pseudomorphic HFET simulation

The full two-dimensional simulation was then run on a pulse-doped pseudomorphic HFET structure, similar to that described in section 4.5. Here the second doping pulse was omitted to simplify the structure, which took the form illustrated in Figure 5.22. This simulation again produces a Fermi-level with a similar profile to the MESFET and the single-channel HFET, Figure 5.23, but here the potential drop at the drain edge of the gate is slightly more abrupt. This is due to the reduced substrate injection in this device which maintains a low carrier density out of the InGaAs layer. Consequently any

significant current in these regions requires a large driving force, and hence a large gradient in the Fermi-level.

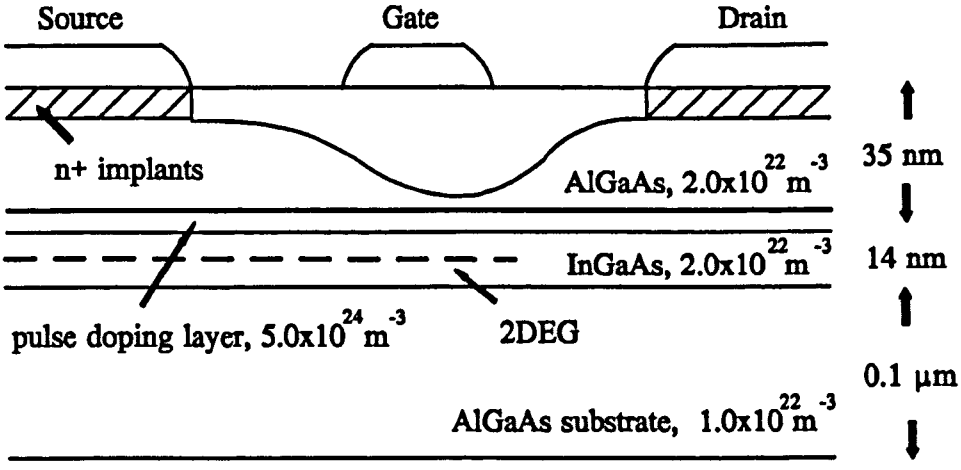


Figure 5.22

0.5 μm gate pulse-doped pseudomorphic HFET structure

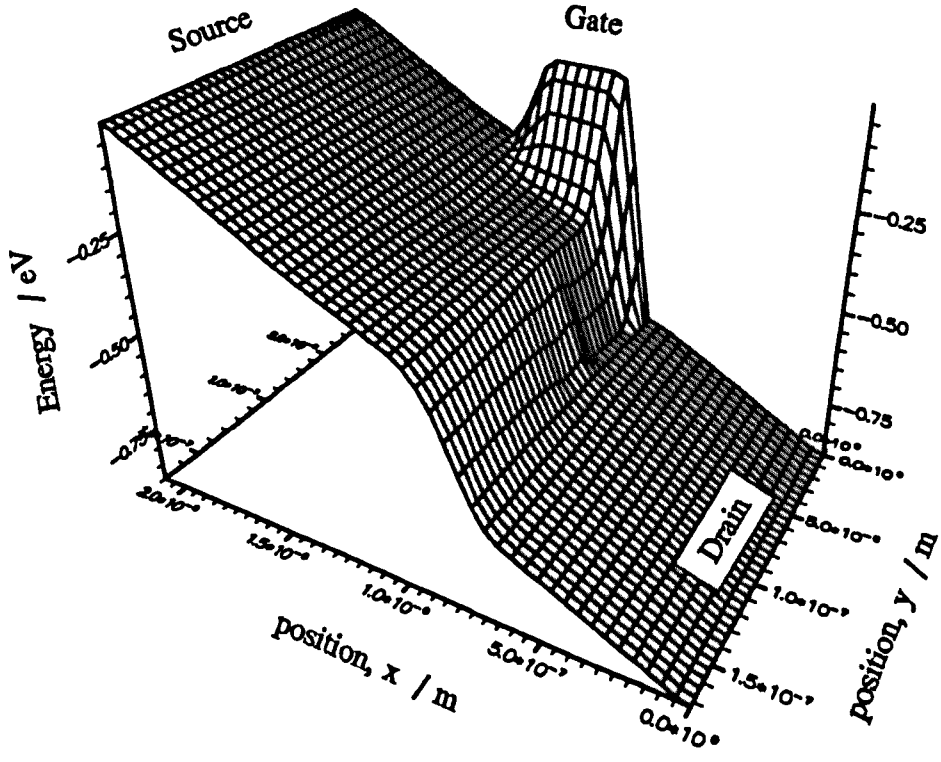


Figure 5.23

Fermi-level for a pulse-doped pseudomorphic HFET

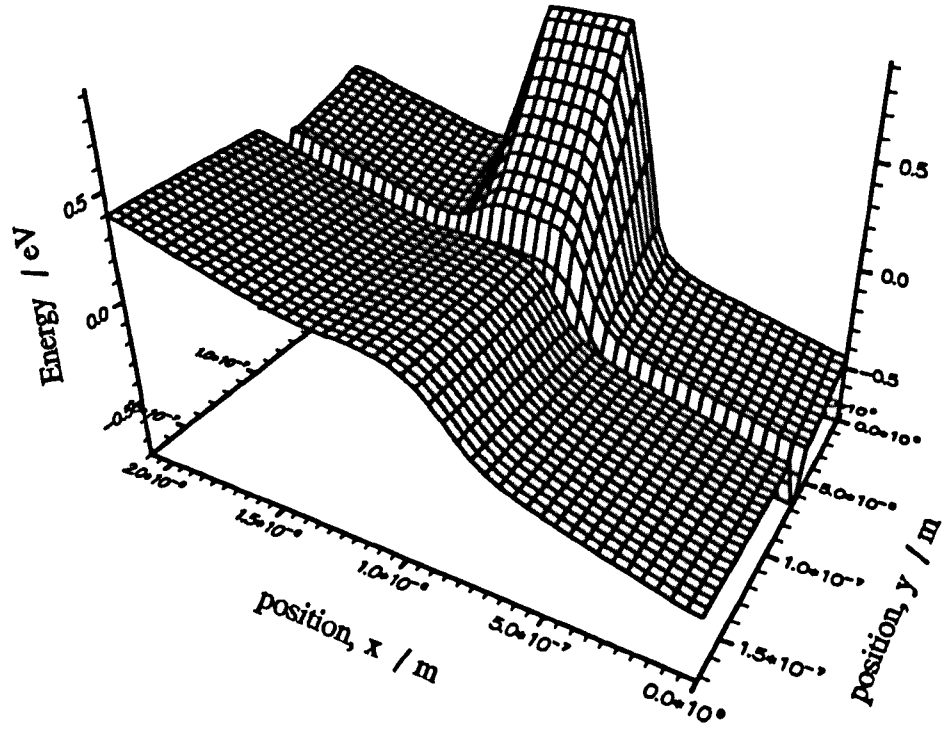


Figure 5.24

Conduction band profile for a pulse-doped pseudomorphic HFET

This figure illustrates the conduction band edge for this device. Here the sharp changes in the conduction band edge are more visible due to the larger conduction band discontinuity of the AlGaAs/InGaAs system as opposed to AlGaAs/GaAs. The low doped upper AlGaAs layer is apparent under the gate, where the potential induced by the Schottky built-in bias decreases linearly to the channel as opposed to both the MESFET and single-channel HFET profiles which decrease quadratically[3].

The electron density and ionised charge density profiles are shown in Figure 5.25 and Figure 5.26. The ionised charge density is essentially zero for most of the device, becoming negative at the surface due to the surface traps, and positive in the pulse doped region.

---

[3] The charge under the gate in these devices is roughly constant at the donor density. With values in excess of 1.0×10^{23} m^{-3} the solution of Poisson's equation gives a quadratic profile, whereas lower than this and the donor density is effectively zero, which then produces a linear profile.

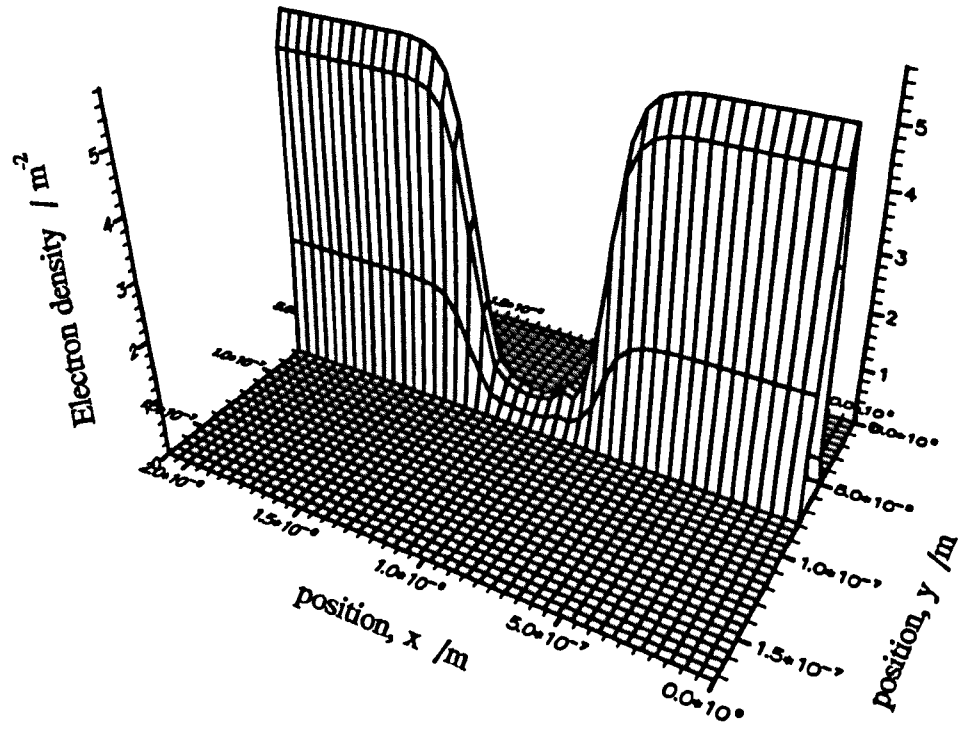


Figure 5.25

Electron density for a pulse-doped pseudomorphic HFET

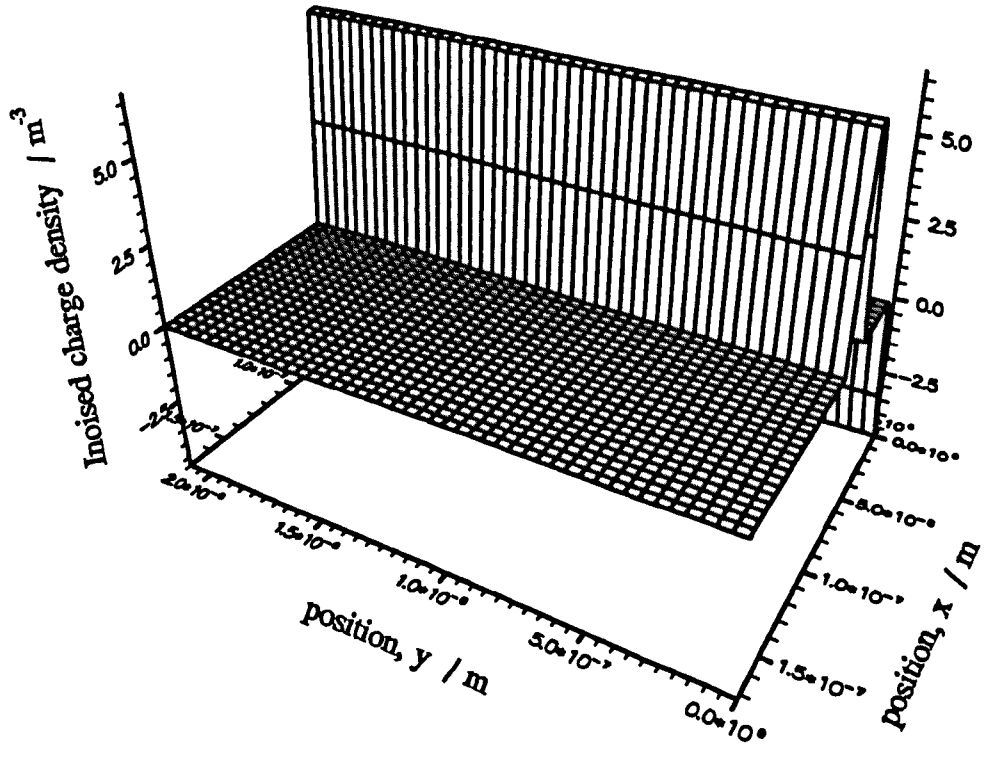


Figure 5.26

Ionised charge density for a pulse-doped pseudomorphic HFET

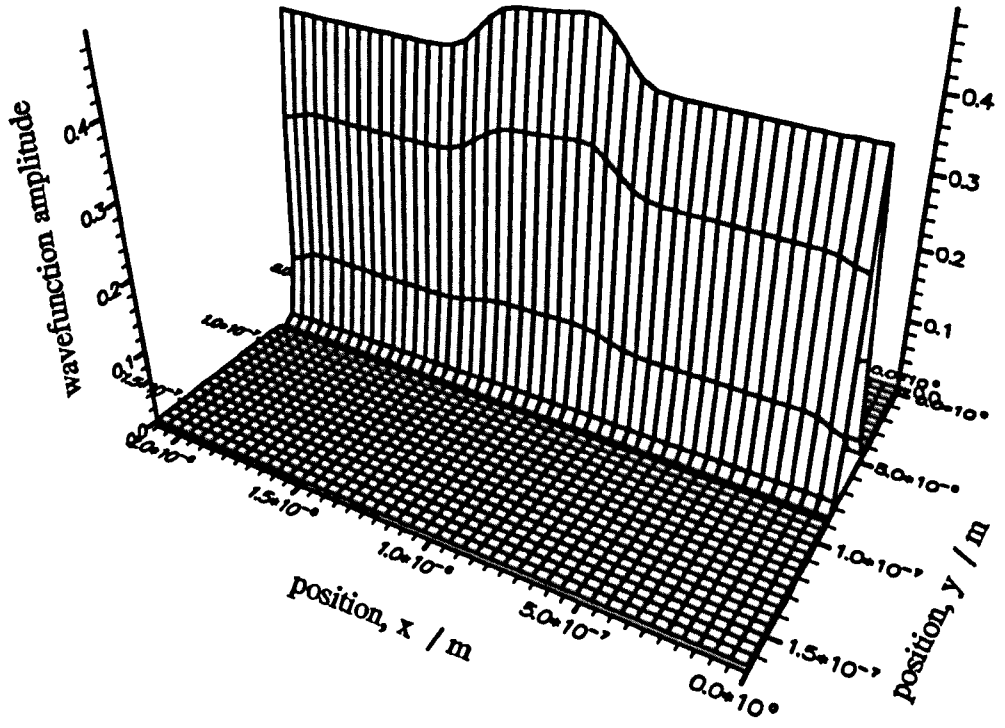


Figure 5.27

First bound state wavefunction for a pulse-doped pseudomorphic HFET

Figure 5.27, Figure 5.28 and Figure 5.29 show the first three bound state wavefunctions for this device. Again only the y-component is displayed with the x- and z-components taking the form of travelling waves. In this device, due to the deeper potential well the 2DEG is supported, even under the drain-edge of the gate. Here the first two wavefunctions have true solutions, producing the 2DEG, with the third neglected as this in no longer bound by the well.

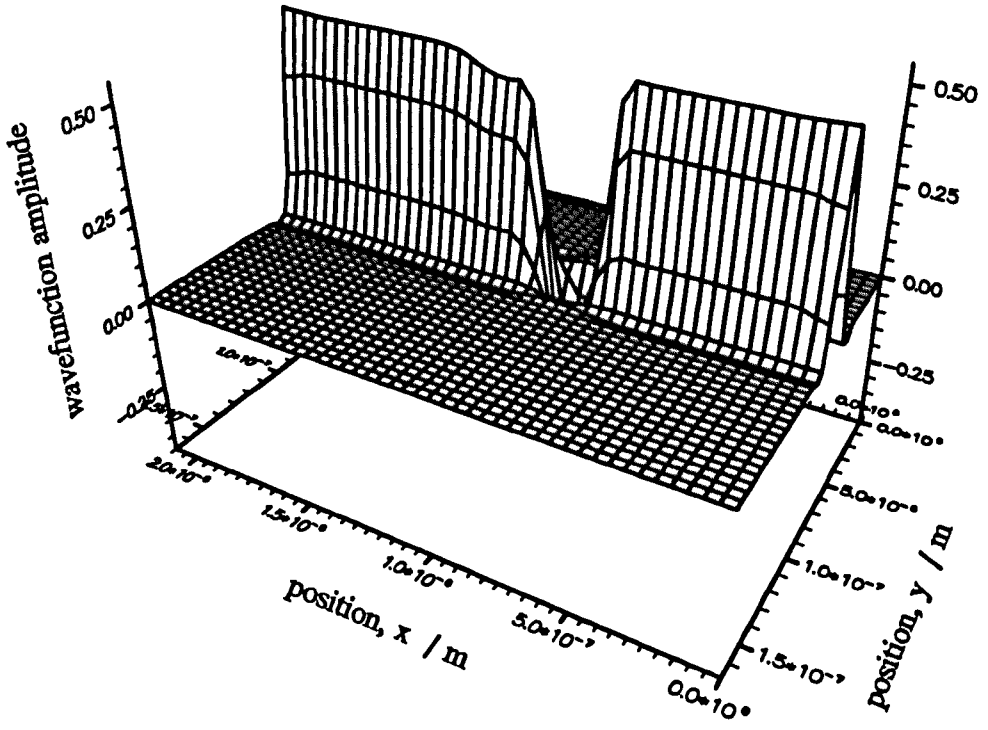


Figure 5.28

Second bound state wavefunction for a pulse-doped pseudomorphic HFET

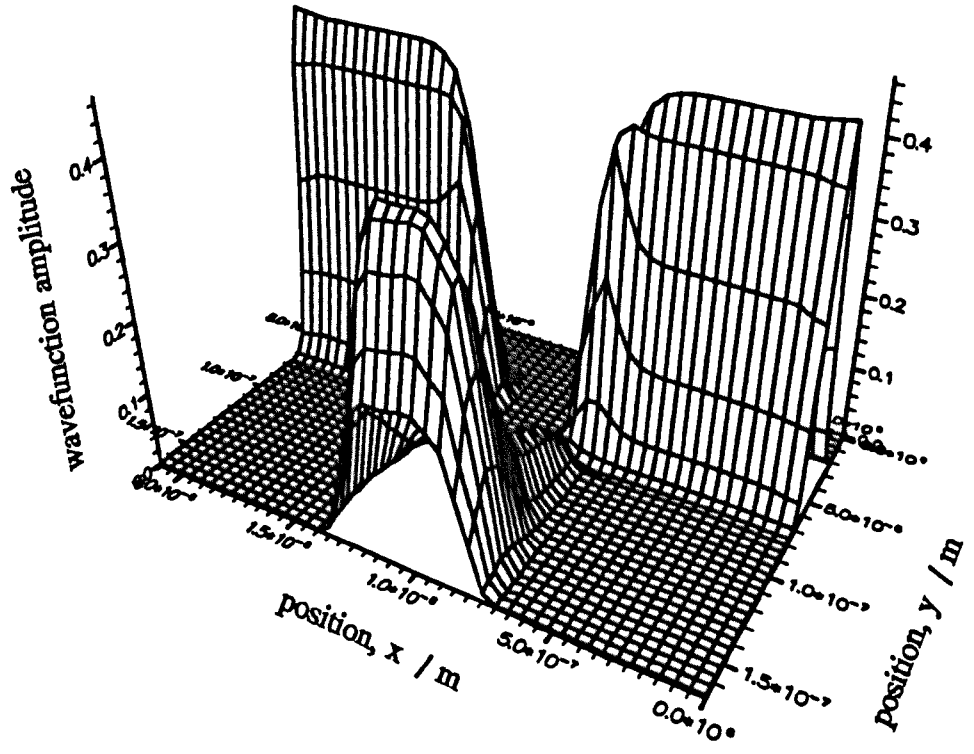


Figure 5.29

Third bound state wavefunction for a pulse-doped pseudomorphic HFET

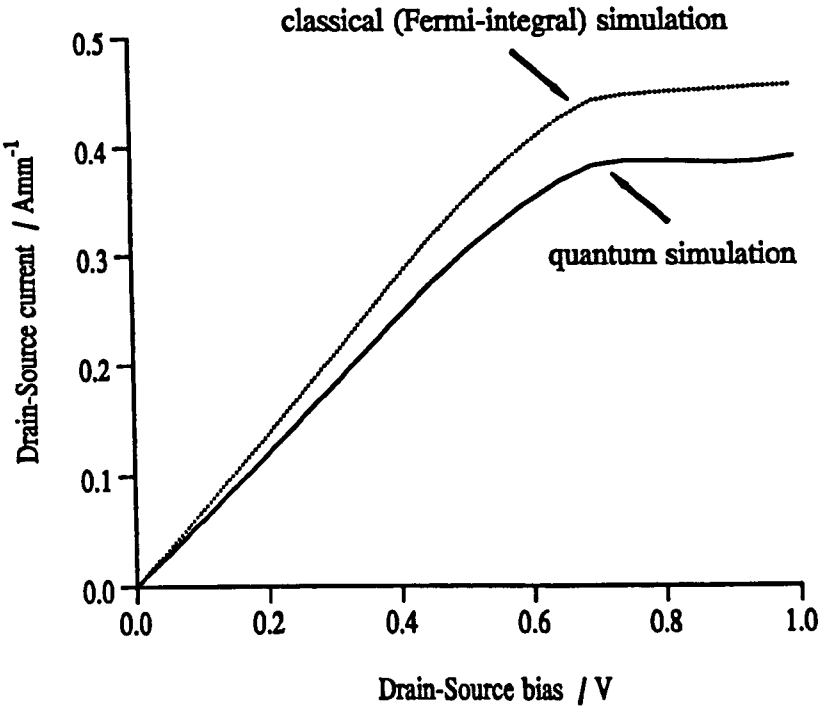


Figure 5.30

I_{DS}-V_{DS} curves for quantum and classical simulations of a pulse doped pseudomorphic HFET

Figure 5.30 illustrates the I_{DS}-V_{DS} characteristics for this device, comparing the quantum and classical solutions. Again the quantum solution reduces the output current by approximately 20%, tracking the classical solution fairly closely.

5.9 Discussion

The results presented in the previous two sections illustrate the significant change in output characteristics that occur when quantum mechanics is included to explicitly calculate the electron wavefunction across a heterojunction interface. These results generally imply a 20% reduction in the output current throughout all applied biases. The

importance of the correct determination of the electron wavefunctions are demonstrated, producing significantly different results even without modifying the transport equations. If a more rigorous analysis were performed, with the 2DEG mobility calculated as part of the simulation process one might expect further changes, only increasing the requirement for a full quantum analysis of the device.

The low-doped AlGaAs layer in the pulse-doped pseudomorphic device allows the Fermi-level under the gate to vary more freely than in the case of the single-channel HFET. Consequently more of the drain-source potential is dropped between the source and the gate in this device creating a smaller electric field at the drain edge of the gate. Also, due to the deeper potential well, the pseudomorphic device retains true bound states in this region, and therefore it is in this device that the largest effects of a rigorous determination of the two-dimensional mobility are expected to be observed. The electron density in the single-channel structure is nearly 'pinched-off' with no corresponding shape produced in the pseudomorphic device. This is probable due to the greater potential dropped at the gate in the single-channel structure, raising the potential well further above the Fermi-level. If a greater drain-source bias were applied to the pseudomorphic device a similar electron profile could be expected in this structure also.

The 'discrete' nature of the quantum mechanical solutions produces many numerical problems, with the maximum drain-source bias attained in the region of 2.0 - 2.5 Volts. Reducing both the mesh spacing and the bias increment between solutions helps to stabilize the solution procedure, perhaps adding a further 25% to the maximum attainable drain-source voltage. One further modification that could possibly increase the stability of the solution process would be to make the scheme time dependent. This would slow down the steady-state solution process by at least an order of magnitude but with the wavefunctions allowed to evolve along a slowly changing path extra stability may be introduced.

Chapter 6
Quasi Two-Dimensional Models

6.1 Introduction

If the regions around the contacts are omitted and substrate injection is low then the Fermi-level can be considered approximately one-dimensional, independent of the depth into the device. Hence the full two-dimensional problem can be reduced to a 'Quasi Two Dimensional' (Q2D) scheme whereby the two-dimensional Schrödinger and Poisson's equations are solved self-consistently with a one-dimensional transport equation. This can be used to greatly simplify the device equations, rewriting them in a 'current-driven' form, and consequently radically improving the solution times. Current driven solvers, as their name implies, require the imposition of a drain-source current and calculate the corresponding drain-source voltage as opposed to both the Schottky-gate and full two-dimensional FET models discussed in Chapters 4 and 5, which set the biases on the contacts and calculate the resulting current. Q2D device models are discussed fully in [Panoja, 1990] and [Rodrigues, 1991] and briefly summarised below.

6.2 Quasi-two-dimensional Models

Consider a slice in the y-direction, away from the contacts, through an HFET device, illustrated in Figure 6.1. The current normal to this surface is given by the surface integral of the current density in this direction, equation (6.1).

$$I = \oint J.n\,dydz \tag{6.1}$$

where n is the unit vector in the x-direction.

If the fringing effects at the ends of the gate fingers are neglected the current density can be considered independent of position in the z-direction, thus integrating in this

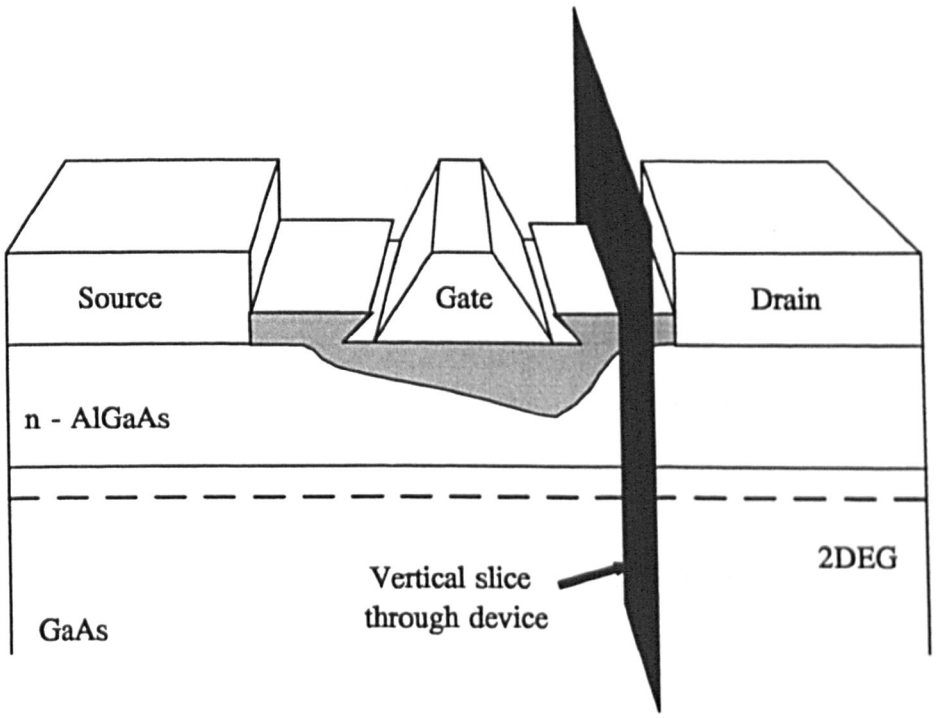


Figure 6.1

Slice through an HFET for use in the quasi-two-dimensional simulation

direction simply multiplies the other terms by Z, the gate finger width. Expanding the current density in terms of electron density and velocity gives equation (6.2).

$$I = qZ\int nv.dy \tag{6.2}$$

If the Fermi-level is independent of y then the gradient of the Fermi-level is also independent of y. Expanding v, the electron velocity in terms of the driving force gives equation (6.3).

$$v = \mu \frac{\partial \phi}{\partial x} \tag{6.3}$$

The low field electron mobility does vary with position corresponding to the local ionised donor concentration but since the majority of the current occurs in the 2DEG and hence the narrow band gap material this quantity can be taken as a constant and equal the the narrow band gap value. Thus the velocity term is approximately

independent of position in the y-direction and can be taken outside the integral. Consequently the equation for current becomes (6.4).

$$I = qZ\mu\nabla\phi\int n.dy \tag{6.4}$$

If the sheet electron concentration can determined the gradient for the Fermi-level can be calculated from the current equation, which when integrated over the length of the device gives the source-drain potential. The problem then reduces to finding a suitable approximation for the sheet electron charge density. Conventional Q2D HFET simulations do this by approximately solving Poisson's equation (6.5).

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{q}{\varepsilon_0\varepsilon_r}\left(N_D^* - n\right) = 0 \tag{6.5}$$

By integrating over the slice the electron density can be substituted by terms from the current-continuity equation. Further replacing the partial derivatives of ψ with respect to position with the electric field and neglecting the diffusion component of the current one has equation (6.6).

$$\int\left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} - \frac{q}{\varepsilon}N_D^*\right).dy - \frac{dI}{\varepsilon Z\mu E_x} = 0 \tag{6.6}$$

where d is the total depth into the device.

Multiplying through the equation by the field in the x-direction and expanding the first partial derivative using a backward-differencing scheme one has a quadratic equation in $E_x|_i$. Here the partial derivative with respect to y and the ionised donor term are usually grouped together and solved assuming the field in the x-direction is zero. This function is performed at the begining of the simulation, using the Schottky-gate model described in Chapter 4, and produces a look-up table varying as a function of the surface

conduction band edge - channel potential[1] separation.

$$\frac{d}{h}E_{x_i}^2 - \frac{d}{h}E_{x_i}.E_{x_{i-1}} + E_{x_i}.\int\left(\frac{\partial E_y}{\partial y} - \frac{q}{\varepsilon}N_D^+\right).dy + \frac{dI}{\varepsilon Z\mu} = 0 \qquad (6.7)$$

A typical plot of sheet-electron density versus surface potential is illustrated in Figure 6.2.

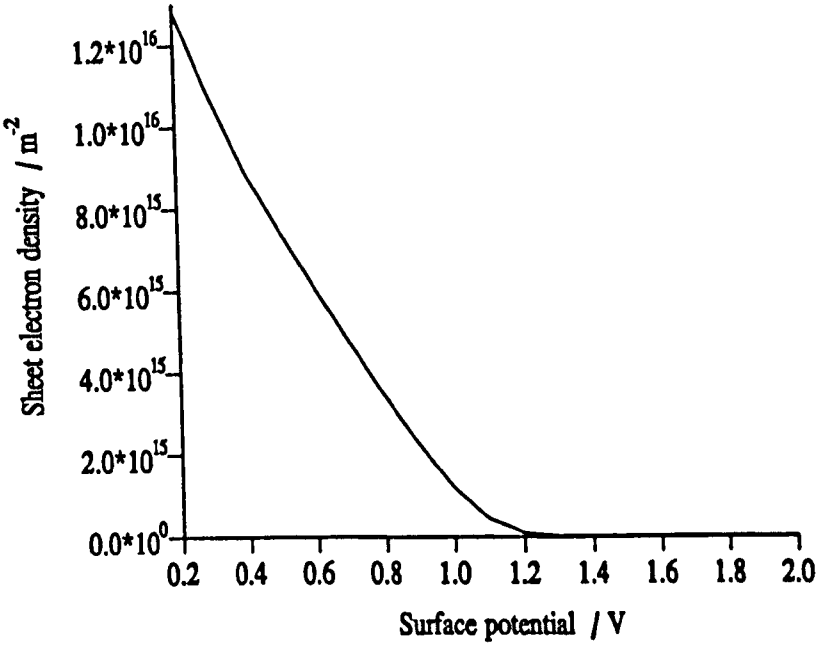


Figure 6.2

Variation of the sheet electron density with surface potential

6.3 Two-dimensional look-up table

The 'zero-field' assumption used to derive this relationship is quite accurate when the partial derivative of E_x with respect to x tend to minus infinity, however as this term

1 The "channel-potential " corresponds to the Fermi-potential in the absence of a diffusion current.

tends to plus infinity problems occur, as this equation would produce a negative electron density. This has been investigated with the one-dimensional Schottky-gate model by adding a term to the one-dimensional Poisson's equation to represent the partial derivative with respect to x. This term is constant accross the modelled domain which is self-consistent with the Q2D approximation. This simulation produced the following results, varying *dEx/dx* at a constant gate bias. The simple approximation is shown to fail when the partial derivative is large and positive as it would continue the gradient from the left of the figure into the 'negative electron' region, illustrated in Figure 6.3 by the dotted line.

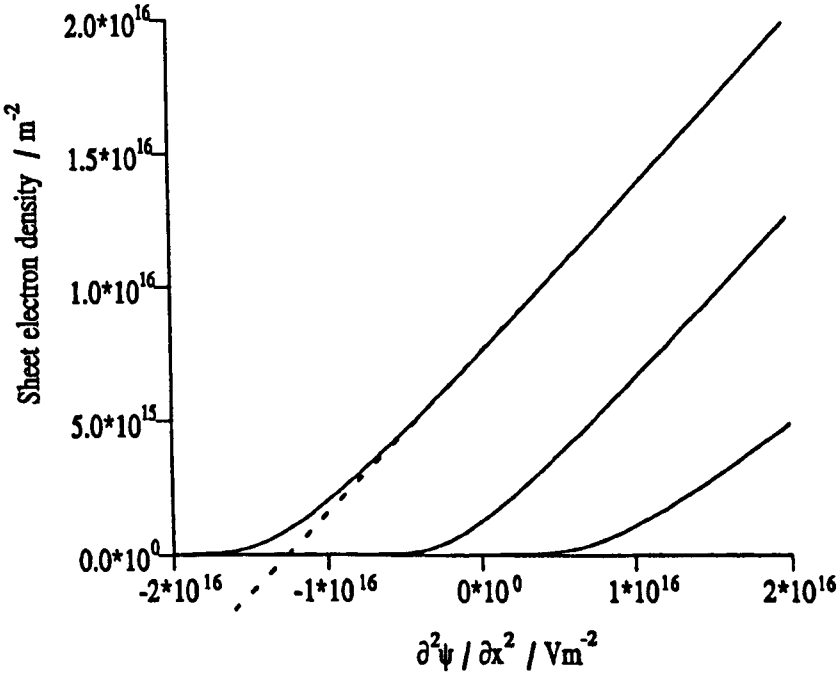


Figure 6.3

Variation of sheet electron density with the gradient of electric field

To incorporate these effects, which are important because they effect the most significant region of the FET, the drain edge of the gate, a model has been written whereby a two-dimensional look-up table is produced at the beginning of a simulation, giving the values of the electron density for varying surface potentials and rate of change of the field in the x-direction. However, the two-dimensional look-up table is expensive in both time and computer memory. It would be preferable to approximate

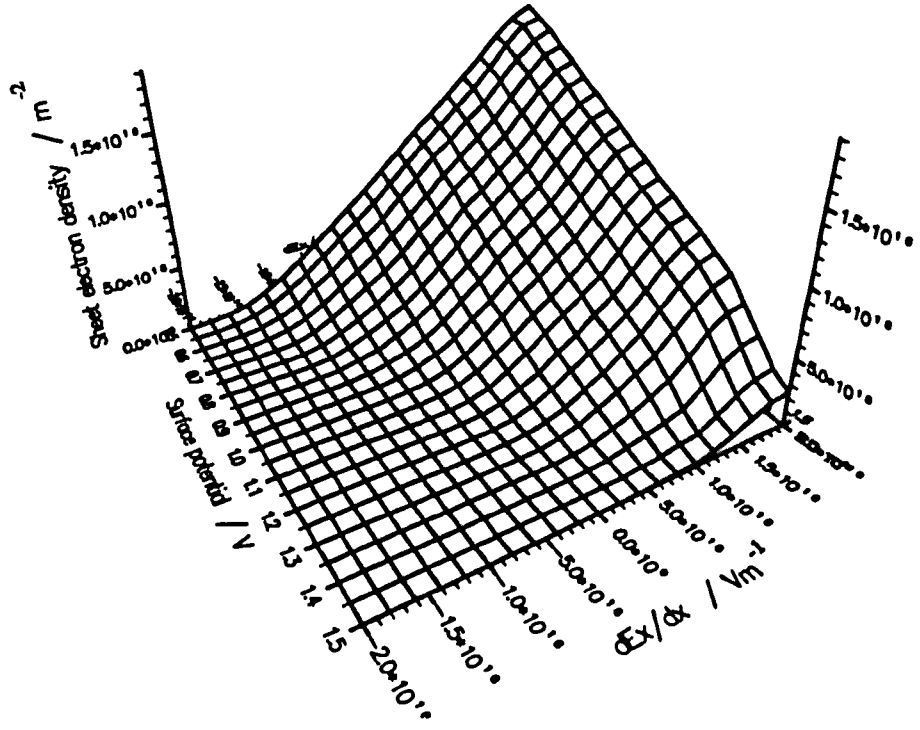


Figure 6.4

Two-dimensional variation of the sheet electron density

this shape using curve-fits and fewer calculations. One possible scheme would be to approximate the curve of sheet electron density with respect to dEx by one or two linear curves and an exponential tail. This curve could be obtained from 5-6 solution points at a constant surface potential and different dEx. This curve is easily invertable and hence is would be trivial to translate it on the dEx axis such that is corresponds to the full two-dimesnional curve. Other alternatives include simple series expansions, although care would have to be taken to ensure the curves had the right shapes at the extremities.

Apart from the large increase in speed, Q2D models offer other advantages, including a relatively simple formulation of the energy conservation equations and subsequently non-equilibrium transport. The energy conservation equations can all be formulated in a backward difference form and solved as the scheme progresses from one node to the next, equation (6.8).

Also other important physical parameters such as the gate recess are relatively easily introduced, with different sections of the device simply requiring different look-up

$$T_{e_{i+1}} = T_{e_i} + h.\left(\frac{2q}{5k_B}E_i - \frac{3(T_{e_i} - T_0)}{5v_i\tau_i}\right) \qquad (6.8)$$

tables. One region of concern may be the walls of the recess, which if obliquely angled would require many different look-up tables and hence become expensive in computational effort, although at it may be possible to simply interpolate between to different look-up tables for this region. Further problems may be associated with the memory requirements for extra look-up tables and the computational effort required to produce them, especially if the program is designed to run on a PC.

6.4 Quasi-two-dimensional model results

This model has been used to simulate a pulse doped pseudomorphic device produced by M/A-COM inc. The two-dimensional look-up table was produced with approximately 400 samples and used in conjunction with the above equations. It is evident that this scheme corresponds fairly well to the device characteristics, remarkably well for such a simple scheme. There was no fitting performed on the device parameters. The discrepancies between the measured and calculated I-V characteristics can be explained, for example the higher saturation currents would be reduced if DX centre trapping were included, which would decrease the number of free electrons, secondly the higher output conductance could be achieved if a substrate injection model were incorporated. One possible method of performing the latter would be to extend the one-dimensional substrate solver to solve for 'hot-electrons', and thus produce a three-dimensional look-up table. The problems with size and computational effort are considerable, especially when it is desirable to produce a fast simulations suitable for CAD purposes, but these problems could be greatly reduced if the curves are fitted to polynomial expansions or other series.

The modelled device is shown in Figure 6.5, with five gate fingers of 50 μm, giving a total gate length of 0.25 mm. Figure 6.6 illustrates a typical band diagram produced by the model, in this case taken at a drain-source bias of approximately 0.65 V. Here the

gate recess is simply modelled by decreasing the surface potential in the regions under the GaAs cap layers. Figure 6.7, Figure 6.8 and Figure 6.9 show the corresponding electron temperature, sheet electron density and electric field respectively.

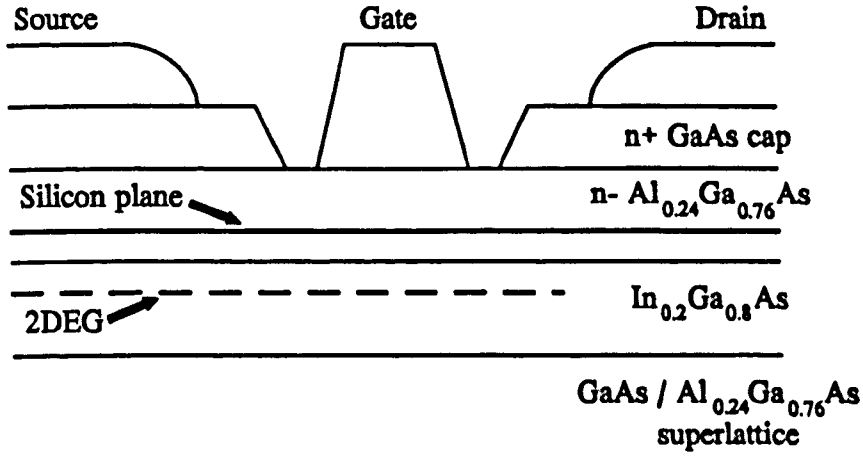


Figure 6.5

Pulse doped pseudomorphic HFET

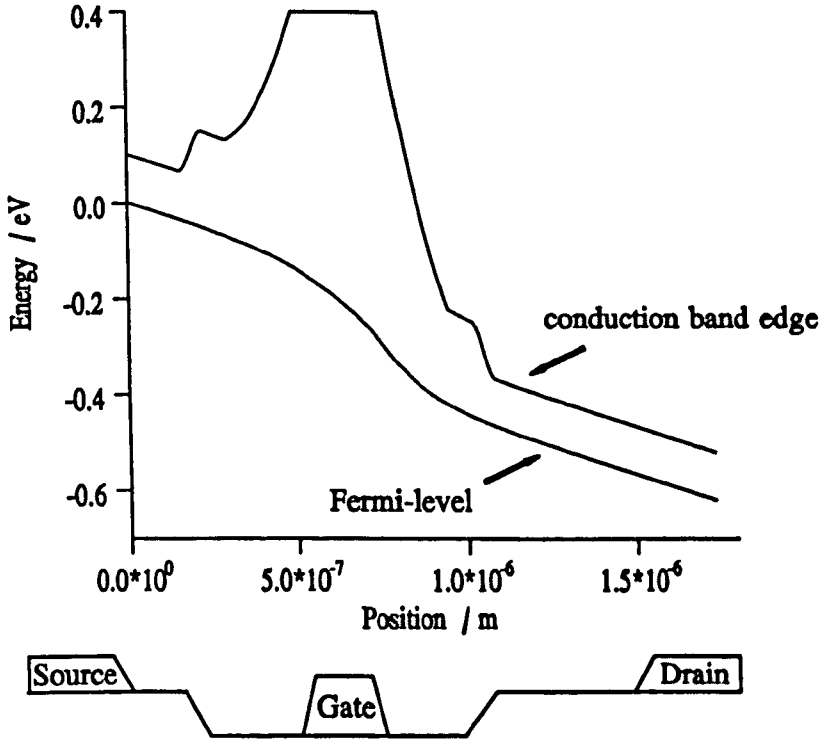


Figure 6.6

Conduction band edge and Fermi-level at the surface

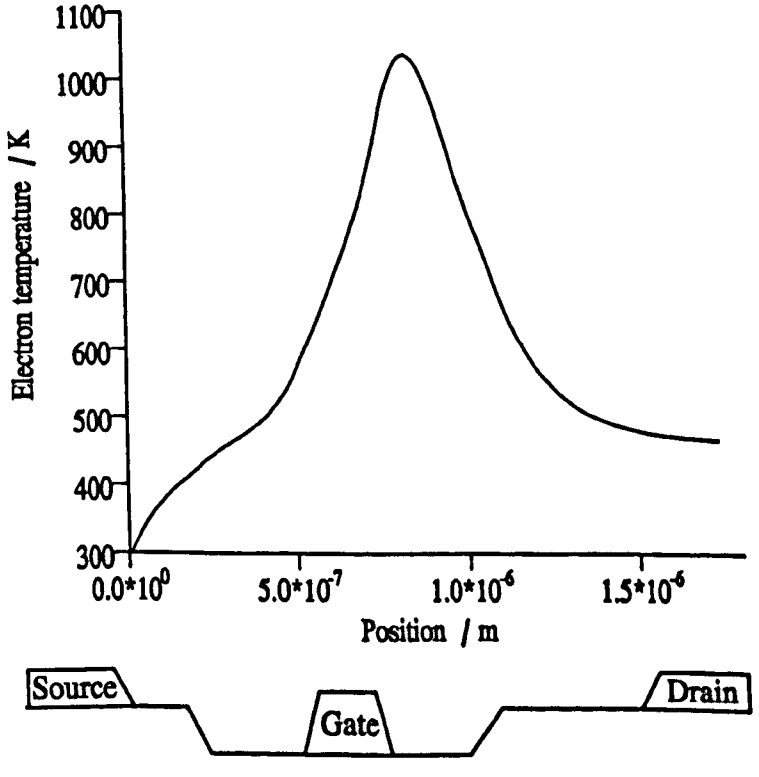
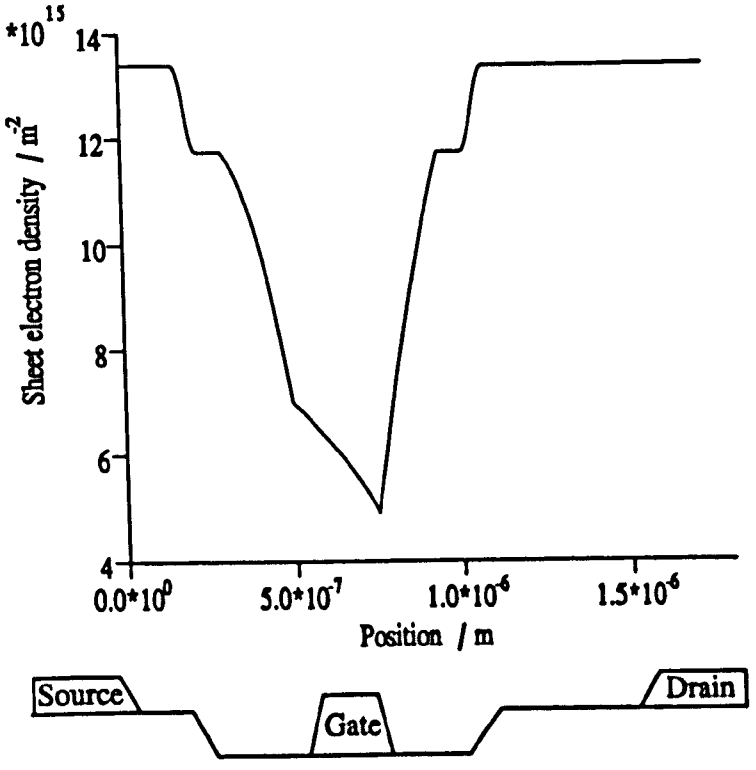


Figure 6.7

Electron temperature



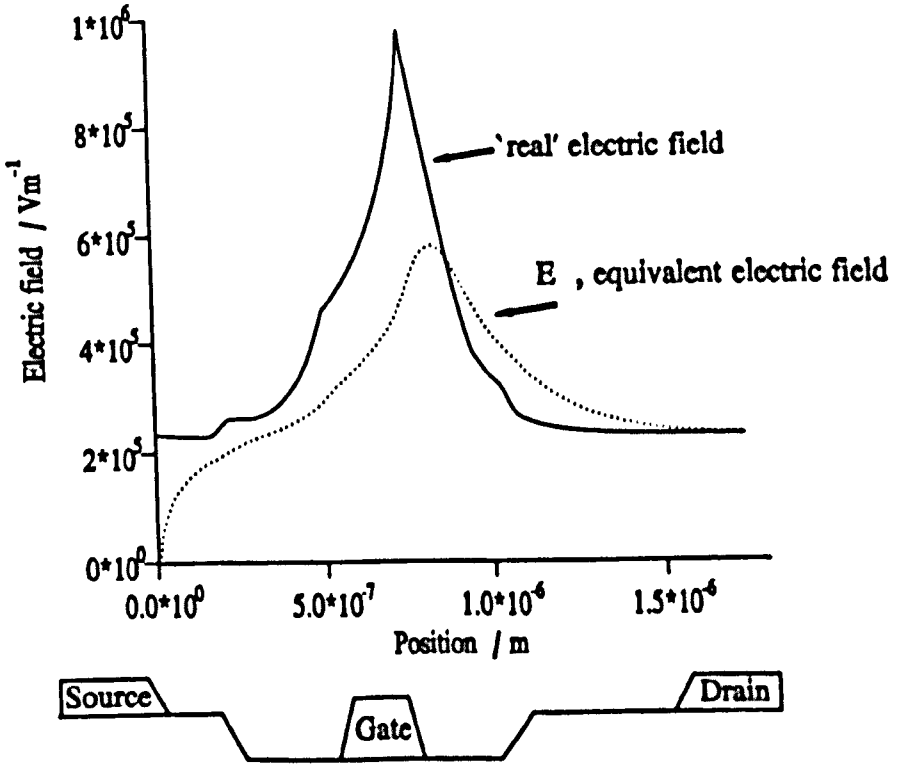


Figure 6.9

Electric field

The model was solved for a range of drain-source currents and gate biases, producing the I_{DS}-V_{DS} characteristics displayed in Figure 6.10. These results compare very well with the measured characteristics of this device, despite the simple nature of this model. This would imply that the majority of the fundamental device physics has been included and the model can be 'fine-tuned' with some additional parameters such as the inclusion of D-X trapping centres and a substrate injection model discussed at the beginning of this section.

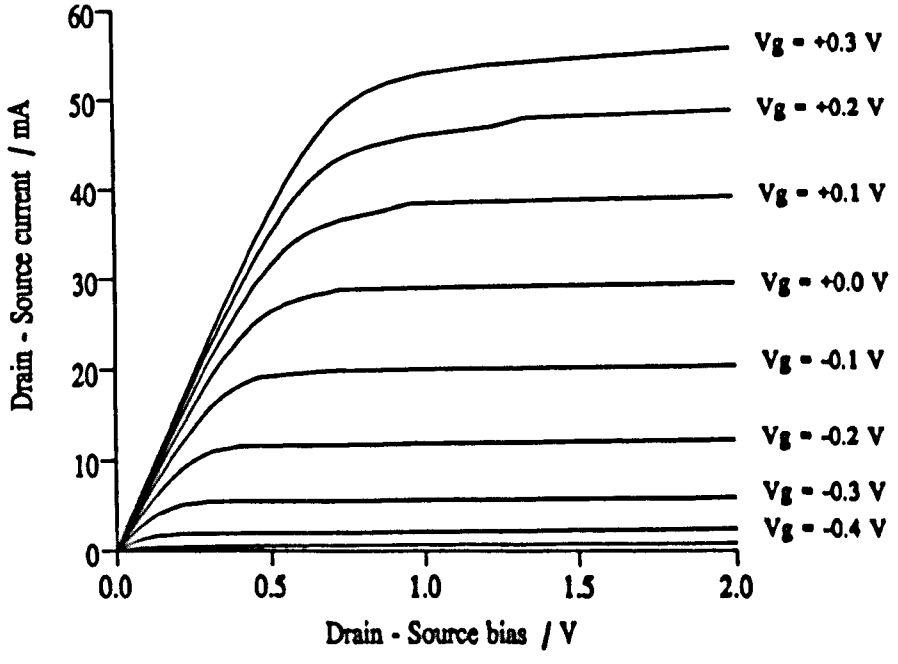


Figure 6.10

Calculated results for the M/A-COM pulse-doped pseudomorphic HFET

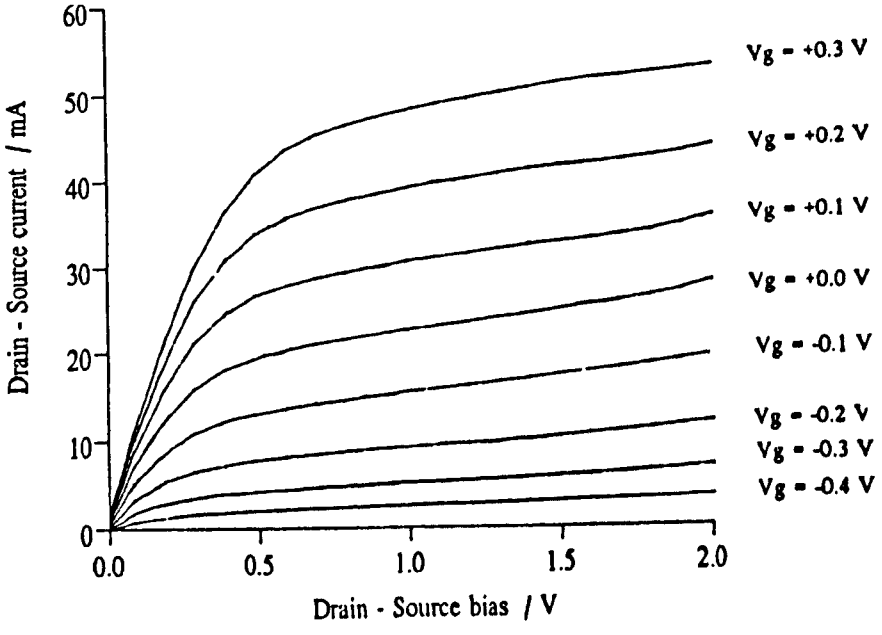


Figure 6.11

Measured results for the M/A-COM pulse-doped pseudomorphic HFET

Chapter 7
Conclusions and Future Work

7.1 Conclusions

Modern field effect transistors with their ultra-small features and complex processing technologies are entering into the quantum world. The physical intuition that accompanied the design of larger devices, namely Silicon and GaAs MESFETS, is not applicable to these smaller structures, producing apparently anomalous effects if quantum analysis is omitted. Since quantum mechanics is counter-intuitive and deeply mathematical these structures will be most reliably designed with the aid of a physical model, able to accurately predict device performance quickly and robustly. The intention of this study was to produce a simulation that filled this role, incorporating quantum mechanics into a conventional FET model, for use in the design and characterisation of HFETS.

This work has produced several new models including a one-dimensional Schottky-gate model, a one-dimensional diode model, a full two-dimensional HFET model and a quasi-two-dimensional HFET model, all of which solve Schrödinger's equation in order to calculate the 2DEG concentration. A clear compromise is required between the accuracy and speed in these simulations when used for CAD purposes, with a good scheme able to cover all important criterion affecting the operation of the device while being fast enough to enable interactive design.

The one-dimensional Schottky-gate model has been developed that solves Schrödinger's equation self-consistently with Poisson's equation. Here the current-continuity and momentum and energy conservation equations are all set at their respective equilibrium values due to the extremely low current densities. This equations are solved using an expanded Newton algorithm which produces an extremely efficient code that is both fast and robust. This model clearly demonstrates the different results produced when quantum mechanics is explicitly taken into account, illustrating the sub-band structure

of several different heterostructures. This model was extended to extract the gate-channel and gate-drain capacitances from which the transconductance, cut-off frequency and minimum noise figure can be approximated. Although straight-forward in nature, this scheme can give an indication of the general performance of a new device structure which is useful where no similar devices have previously been produced, and consequently no 'rules of thumb' exist to guide the designer. This scheme could be used to set limits on the device configuration from which a more accurate simulation could work.

The Schottky-gate model was extended further to include charge transport, producing the diode model. This scheme was used to validate the 'zero current' approximation of the Schottky-gate model, which was found to be accurate provided the surface was reverse biased. Once, V_{bi}, the Schottky built-in potential was removed by external biasing the two schemes diverged, with the simpler Schottky-gate model significantly overestimating the sheet electron density.

The two-dimensional HFET model self-consistently solves the Poisson, current-continuity and Schrödinger equations, again within the framework of a modified Newton algorithm. The solution of Schrödinger's equation in two dimensions is non-trivial, especially under non-equilibrium conditions. Consequently this task is performed assuming the events in the x- and z-directions are large enough to allow a classical description of the electron wavefunction in these directions. The wavefunction normal to the heterojunction interface is then determined by multiple one-dimensional solutions. This new model, solving the two-dimensional Schrödinger equation under non-equilibrium conditions, separates the total electron density into two components. The electrons associated with the 2DEG are found from quantised states where the inter sub-band separation is greater than the thermal energy of the electrons. Since transport normal to the heterojunction requires scattering to different energy levels, and the energy separation is greater than the energy of the electron, this event is forbidden and consequently the 2DEG is restricted to motion parallel to the interface. The potential well that creates the 2DEG widens as the energy increases, and consequently the bound state energy levels get closer together until the thermal energy of the electrons is sufficient to produce a 'quasi-continuous conduction band'. Electrons associated with the

higher energy levels, and hence the quasi-continuous conduction band are treated like normal 'bulk' carriers and have all three degrees of freedom. The calculation of the three-dimensional electron density is further simplified by approximating it with the Fermi-integral taken from the onset of the quasi-continuous conduction band edge to infinity. It is these carriers that permit a current to exist normal to the heterojunction allowing substrate injection to occur and the transfer of the electrons involved in parallel MESFET conduction. This novel sectioning of the electrons and their respective degrees of freedom reduces the output current, but only by 1-2%. This small change from an effect that appears to be far more significant can be attributed to the low current densities normal to the heterojunction, indeed the only position within the modelled domain to have a significant current in the y-direction is at the drain-edge of the gate. In the case of the single-channel HFET both the quantum and classical simulations produce a potential well has been pulled out to such an extent that it will not support any 'quantised' bound states. Consequently both schemes have a similar number of three-dimensional electrons able to move into the substrate. The pseudomorphic device however, is slightly different since the potential well in this case is deep enough to support the 2DEG at the drain-edge of the gate and hence in this device the quantum and classical treatments do produce very different electron densities able to travel into the substrate. Here the restricted degrees of freedom of the 2DEG does not alter the output currents profoundly due to the much smaller current crossing the interface in this device. The reduced substrate current is caused by the increased potential barrier presented by the substrate. If quantum transport, and consequently a tunnelling current, were explicitly included, the simulated domain could be extended under the source and drain contacts where one could expect the reduced mobility of the 2DEG to play a more important role due to the larger current densities normal to the interface.

The solution of Schrödinger's equation prevented the conventional solution techniques from being applied to this model, and consequently a scheme was developed where the device equations are solved using an expanded Newton's scheme, which although more complex to code generally possesses fast and efficient convergence properties. The equations are discretized over a non-uniform grid for which a new refining algorithm has been produced. The overall scheme solves each bias at least an order of magnitude faster than the more conventional time-dependent Gummel based algorithms, typically

2-3 minutes per bias point on a SUN IPX workstation. The discrete nature of the quantum solution produces a numerically unstable scheme that has problems converging for high drain-source biases. The instabilities are assumed to be caused by the profile of the solution curve which may possess local maxima, minima or inversion points in between the initial guess and the root of the equations. Further the solution may not even be unique. The scheme was successfully improved by taking smaller increments in bias between consecutive solution points and incorporating a relaxation algorithm to the solution process. Although this produces more bias points in any given voltage sweep it is not much more expensive in computational resources since each point generally requires less iterations to solve. One other improvement that worked for some device structures was to simply increase the node density and hence reduce the internodal separation. This is certainly more expensive since the matrix inversion procedure requires between N^2 and N^3 operations, where N is the number of nodes. Further this scheme was not reliable producing worse results for some device configurations.

This quantum HFET model has been used to model several devices including a MESFET, a simple single-channel AlGaAs/GaAs HFET and a pulse-doped pseudomorphic HFET. Whilst several important effects have been omitted, namely non-equilibrium transport and the two-dimensional electron mobility, both HFET simulations have demonstrated a significant change from the classical results. These different devices produce similar output characteristics, with the drain current reduced by approximately 20% by the quantum description. This is attributed to the lower free electron density produced by the quantum simulation, since the restriction on the degrees of freedom were found not to effect the steady-state results by more than a few percent. Further changes are expected if a rigorous derivation for the two-dimensional electron mobility were included, although the inclusion of non-equilibrium transport may well counter-act this since the heated electrons should gain enough energy to scatter to the higher sub-bands and restore all three degrees of freedom.

The results from the full-two-dimensional simulations, in particular the potential profiles encouraged work on a quasi-two-dimensional model. The Fermi-level was almost one-dimensional in all cases, perpendicular to the active channel. Thus the

quasi-two-dimensional approximation appears to apply to heterostructure devices as well as MESFETS. The conventional Q2D device equations were investigated and found to be inadequate for HFET simulation, especially at the drain-edge of the gate where large positive gradients in the electric field exist. A new scheme was therefore developed, moving the solution of the two-dimensional Poisson equation to the sheet electron density look-up table[1]. This produces a sheet electron density that varies as a function of two parameters, the surface potential and the gradient of the electric field. The electron density was then self-consistently solved with the current continuity and energy conservation equations. This model was applied to the structure of a pulse-doped pseudomorphic device produced by M/A-COM inc. from which the D.C. I_{DS}-V_{DS} characteristics were obtained and found to be in remarkable agreement with measured results. This was without the model parameters being fitted and hence is very encouraging for such a simple model. Although this new Q2D model needs to be verified on other different structures it may be the route for HFET CAD design.

7.2 Further development

The full two-dimensional model could be developed in five main areas, the first three are related to the charge transport equations, followed by the incorporation of a gate recess model, with the last targeted on the numerical stability. One of the first additions to the model would be that of non-equilibrium transport. This is not trivial, even in the simple single band approximation where at least one other variable would be added, increasing the rank of the Jacobian matrix by 50% and requiring more than twice the computational effort to produce. In the case of an HFET these equations are complicated further by the sub-band structure occurring in the two-dimensional potential well. Non-equilibrium effects occur when the electrons gain enough energy from the field to transfer to the higher energy L and X bands. If other bands are present, ie. the sub-bands at the heterojunction interface, the electrons can scatter to these levels as well. The equations governing this system become more complicated, not only due to the larger

[1] The conventional Q2D scheme splits the solution of Poisson's equation. The y-component is used to determine a one-parameter electron look-up table with the second 'x-component' solved with the transport equations in the channel simulator.

number of available bands but also because the bands are closer together and may move relative to one another. The electron sub-bands will change electron population and hence alter the number of vacant sites available. Work was done on the incorporation of the energy conservation equations into both the one- and two-dimensional models, but just using the classical two-band format. Although the one-dimensional model had limited success the two-dimensional FET model was found to be extremely numerically unstable with few drain-source biases in excess of 1 Volt successfully solved. This model incorporated the same macroscopic discretization scheme for the energy conservation equation as the current-continuity equation, evaluating the energy flux at the half-node positions. The electron temperature was assumed to vary linearly in between the nodes, which was perhaps the source of the numerical instability. The next step in the incorporation of the energy conservation equation could therefore be to use an exponential variation for the electron temperature similar to that used in more conventional models.

The second development for the two-dimensional FET model would be to include a more rigorous determination of the two-dimensional electron mobility. This parameter has already been discussed in Chapters 2, 3, and 5. In conclusion, the modelling of this parameter within present computational limits is probably best done using predetermined mobility surface varying as a function of two-parameters, the electric field and the electron density.

The third improvement in the transport equations, and most complicated, is the inclusion of a tunnelling current. In regions where there exists a significant current in the y-direction the band profiles will be effected by a tunnelling current through the potential spike at the heterojunction. To fully describe this phenomena quantum transport has to be included, with all the associated problems discussed in the introduction to Chapter 5, section 5.1. Hence this route is certainly non-trivial and probably too expensive in computational resources to be applied. Therefore a simpler approximation is required that doesn't require the evaluation of the 'free' electron wavefunctions. Further if tunnelling were introduced the regions directly beneath the source and drain ohmic contacts could be included in the modelled domain.

Most modern FET structures include gate recesses as part of the device layout. At present the two-dimensional FET model omits this important detail. A more practical FET model would therefore be produced if a gate-recess were included. There are several ways in which this could be done. One common scheme discretizes the modelled domain and simply neglects the nodes falling in the recess region. A better alternative is to include these nodes, and further extend the simulated gate vertically upwards to the real gates boundaries. The nodes between the gate and the free surface can then be given the attributes of air of the pacifying material. In this way the effects of the different gate shapes can be investigated.

Finally work could continue on the solution algorithm, making it more robust and able to simulate devices under higher biases. Whilst the I_{DS}-V_{DS} curves generated to date by the quantum solution follow those produced by the classical scheme this may not always be the case. One could argue that as the gate bias is increased and the potential well is pulled out that the effects of the quantum transport will become more pronounced as the effects of inter sub-band scattering becomes more prominent.

7.3 Full two-dimensional current driven model

Work on both the full two-dimensional and quasi-two-dimensional (Q2D) models has resulted in the outline of an intermediate scheme being developed, using some of the approximations used in the quasi-2D scheme, but allowing full two-dimensional treatment of charge transport. The proposed scheme uses the current driven sections of the Q2D model and some of the assumptions about the surface potential, but instead of using a sheet electron density expands the slice through the device and solves the transport equation in this direction also.

If one considers a slice through the device, perpendicular to the heterojunction interface, we can draw gaussian surfaces around each node. If the current density from the left is known the gradient of the Fermi-level to the left of this slice can be determined using a backward differencing scheme. Hence the Fermi-level on the slice can also be

calculated by integrating in the x-direction. This could be used to find a self-consistent solution of the conduction band edge and the charge density via Poisson's equation. With the charge density, field and Fermi-level determined the current flow in the y-direction can be calculated and hence by Gauss' law the current exiting each cell from the right. Thus current continuity from one slice to another is maintained.

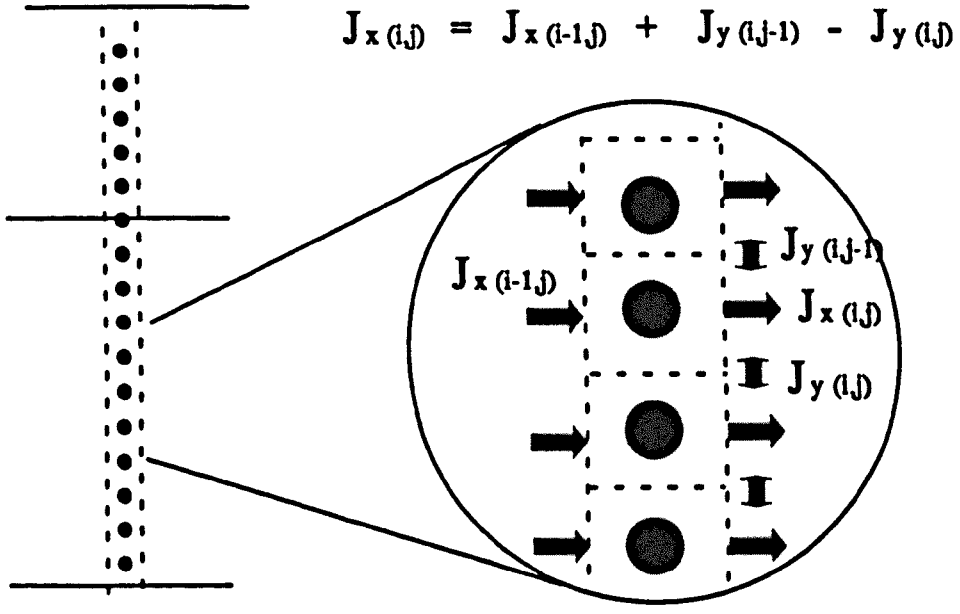


Figure 7.1

One-dimensional slice through FET used in current driven solver

This model would require additional information on the boundaries of each slice, but one could use a scheme similar to that used in the Q2D models, for example the conduction band edge / Fermi-level separation on the surface could be determined from the surface and gate potentials.

This scheme would perform a full two-dimensional simulation of semiconductor devices, whilst making a significant improvement in the speed of the simulation. In the full two-

dimensional scheme described in Chapter 5 each bias point requires a few minutes to solve. In this model over 90% of the simulation time is spent inverting the Jacobian matrix equation, a task that requires approximately N^3 operations, where N is the number of nodes. The one dimensional scheme would possess a tridiagonal matrix structure, and enable the whole device to be solved in 3xN operations, offering a significant increase in speed. This would be particularly useful where the features change over a small scale typical of heterojunctions and the simulation requires many nodes.

7.4 Applications

This work is useful, both in the results it has produced, and the possible future applications of the full two-dimensional and quasi-two-dimensional models. Although the full two-dimensional model omitted an accurate description of the electron transport, and consequently is not expected to produce reliable predictions on device behaviour, it is instructive to observe the different results produced when quantum mechanics is included in a FET model. These significant changes occur even with a bulk mobility used for the 2DEG. The techniques developed to solve Schrödinger's equation in two-dimensions, provide a better approximation for the bound state wavefunctions and their associated energy levels than has previously been available. This information could be used to explicitly calculate the scattering rates occurring in the 2DEG and hence determine the two-dimensional electron mobility directly. Unfortunately this scheme would probably be very computationally intensive, adding another layer of complexity.

The quasi-two-dimensional model is still probably the best tool for CAD since this scheme offers accuracy to within processing tolerances and impressive speed. Full parameter extraction including D.C., pulsed I-V and s-parameter is available within seconds. This allows the automation of device design to become a real proposition, with iterative procedures that home in on the best device configuration being able to produce useful results within minutes.

References

Adachi S., "GaAs, AlAs, and $Al_xGa_{1-x}As$: Material parameters for use in research and device applications", *J. Appl. Phys.*, Vol. 58, No. 3, pp. R1, August 1985

Anderson R.L. "Germanium-Gallium Arsenide Heterojunctions" *IBM J. Res. & Develop.*, Vol. 4, pp. 283-287, 1960

Anderson R.L., "Experiments on Ge-GaAs Heterojunctions", *Solid State Electronics*, Vol. 5, pp. 341, April 1962

Barton T.M., "Characterisation of the Physical Behavior of GaAs MESFETs", *Ph.D Thesis*, Dept. Electrical and Electronic Engineering, University of Leeds, 1988

Bednarczyk D. and Bednarczyk J., "The Approximation of the Fermi-Dirac Integral $\mathcal{F}_{1/2}(\eta)$", *Phys. Lett.*, vol. 64A, no. 4, pp. 409, January 1978

Blakemore J.S., "Approximations for Fermi-Dirac Integrals. Especially the Function $\mathcal{F}_{1/2}(\eta)$ Used to Describe Electron Density in a Semiconductor", *Solid State Electronics*, vol. 25, no. 11, pp.1067, 1982

Bløtekjær K., "Transport Equations for Electrons in Two-Valley Semiconductors", *IEEE*, Vol. ED-17, No. 1, pp. 38-47, January 1970

Chambers. R.G., "Electrons in Metals and Semiconductors", *Chapman and Hall* (Publishers), p. 88, ISBN 0-412-36840-4, 1990

Cole E.A.B., "Numerical Methods and their Application to Device Modelling", Compound Semiconductor Device Modelling, C.M. Snowden and R.E. Miles (eds.), *Springer-Verlag* (Publishers), ISBN 3 540 19827 X, 1993

Datta S., "Quantum Phenomena, part of the modular series on solid-state devices", *Addison-Wesley Publishing Co.*, ISBN 0 201 07956 9, 1989

Delagebeaudeuf D. and Linh N.T., "Metal-(n) AlGaAs-GaAs Two-Dimensional Electron Gas FET", *IEEE*, Vol. ED-29, No. 6, p. 955, 1982

Delagebeaudeuf D., Chevrier J., Laviron M. and Delascuse P. "A new relationship between the Fukui coefficient and optimal current value for low-noise operation of field-effect transistors", *IEEE*, Vol. EDL-6, No. 9, pp. 444-445, September 1985

Dekker A.J., "Solid State Physics", *Macmillan & Co. Ltd.* (Publisher), pp. 305 - 314, 1957

DeMari A., "An Accurate Numerical One-Dimensional Solution of the P-N Junction", *Solid State Electron*, Vol. 11, pp. 33-58, 1968

Dingle R., Störmer H.L., Gossard A.C. and Wiegmann W., "Electron mobilities in modulation-doped semiconductor heterojunction superlattices", *Appl. Phys. Lett.*, Vol. 33, No. 7, pp. 655-677, 1978

Dingle R., Wiegmann W. and Henry C.H., "Quantum States of Confined Carriers in Very Thin $Al_xGa_{1-x}As$-GaAs-$Al_xGa_{1-x}As$ Heterostructures", *Phys. Rev. Lett.*, Vol. 33, No. 14, pp.827, September 1974

Drummond T.J., Morkoç H., Hess K. and Cho A.Y., "Experimental and theoretical electron mobility of modulation doped $Al_x Ga_{1-x}$ As/GaAs heterostructures grown by molecular beam epitaxy", *J. Appl. Phys.*, Vol. 58, No. 8, pp. 5231-5234, August 1981

Duh H.H.G., Pospieszalski M.W., Kop W.F., Ho P., Jabra A.A., Chao P.C., Smith P., Lester L.F., Ballingall J.M. and Weinreb S., "Ultra-Low-Noise Cryogenic High-Electron-Mobility Transistors", *IEEE*, Vol, ED-35, No. 3, pp. 249-256, March 1988

Eisberg R. and Resnick R., "Quantum Physics of atoms, molecules, solids, nuclei and particles (2^{nd} Ed.)", *Wiley* (Publishers), p.145, ISBN 0-471-87373-X, 1974, 1985

Esaki L. and Tsu R., "Superlattice and Negative Differential Conductivity in Semiconductors", *IBM J. Res. & Develop.*, Vol. 14, p. 61, 1970

Ferry D.K. and J.R. Zhou, "Modelling of Quantum Effects in Ultrasmall HEMT Devices", *IEEE*, Vol. ED-40, No. 2, pp. 421, February 1993

Finn C.B.P., "Thermal Physics", *Routlidge & Kegan Paul*, London, (Publishers) ISBN 0-7102-0660-7, 1986

Haigh D. and Everard J. (Eds.), "GaAs technology and its impact on circuits and systems", *Peter Peregrins* (Publisher), ISBN 0 86341 197 8, pp. 392-405. 1989

Harrison W.A., "Elementary Theory of Heterojunctions", *J. Vac. Sci. Tech.*, Vol. 14, pp. 1016, July 1977

Hg S.H. , Khoie R. and Venkat R., "A Two-Dimensional Self-Consistent Numerical Model for High Electron Mobility Transistor", *IEEE*, Vol. ED-38, No. 4, pp. 852, April 1991

Hickmot T.W., Solomon P.M., Fischer R. and Morkoç H., "Negative charge, barrier heights and the conduction-band discontinuity in $Al_xGa_{1-x}As$ capacitors", *J. Appl. Phys.*, Vol. 57, No.8, pp. 2844, April 1985

Hida H. , Itoh T. and Ohata K., "A Novel 2DEGFET Model Based on the Parabolic Velocity-Field Curve Approximation", *IEEE*, Vol. ED-33, No. 10, pp. 1580-1586, October 1986

Hikosaka K., Sasa S., Harada N. and Kuroda S., "Current-Gain Cotoff Frequency Comparison of InGaAs HEMTs", *IEEE*, Vol. EDL-9, No. 5, pp. 241-243, May 1988

Hirakawa K. and Sakaki H., "Mobility of the two-dimensional electron gas at selectively doped *n*-type $Al_xGa_{1-x}As$/GaAs heterojunctions with controlled electron concentrations", *Phys. Rev. B*, Vol. 33, No. 12, pp. 8291-8302, June 1986

Hiyamizu S., Saito J., Nanbu K. and Ishikawa T., "Improved Electron Mobility Higher than 10^6 cm^2/Vs in Selectively Doped GaAs/N-AlGaAs Heterostructures Grown by MBE", *Jap. J. Appl. Phys.*, Vol. 22, No. 10, pp. L609-L611, October 1983

Kreyszig E., "Advanced Engineering Mathematics", Fifth Edition, pp. 476, *John Wiley & Sons*, 1983, ISBN 0-471-88941-5

Kröemer H., Chien W., Harris J.S. and Edwall D.D., "Measurement of isotype heterojunction barriers by C-V profiling", *Appl. Phys. Lett.*, Vol. 36, No. 4, pp. 295, February 1980

Kundert K.S. and Sangiovanni-Vincentelli A., "Sparse, A Sparse Linear Equation Solver, Version 1.3a, 1 April 1988", Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Calif. 94720

Kurata M., "Numerical Analysis for Semiconductor Devices", *Lexington Books* (Publishers), p. 143, ISBN 0-669-04043-6, 1982

Ladbrooke P.H., "MMIC Design: GaAs FETs and HEMTs", *Artech House, Inc.* (Publishers), ISBN 0 89006 314 1, p. 189, 1989

Lei X.L., Zhang J.Q. and Birman J.L., "Hot-electron transport in GaAs-AlGaAs heterojunctions", *Phys. Rev. B*, Vol. 33, No. 6, pp. 4382-4385, March 1986 (a)

Lei X.L., Horing H.J.M. and Zhang J.Q., "Balance-equation analysis of linear and non-linear electronic transport in quasi-two-dimensional quantum-well superlattices", Vol. 34, No. 2, pp. 1139-1147, July 1986 (b)

Li C.L., "Avalanche Breakdown in GaAs MESFETS", *Ph.D Thesis*, Dept. Electrical and Electronic Engineering, University of Leeds, 1993

Luo J.K., Ohno, H., Matsuzaki K. and Hasegawa H., "Low Field Transport Properties of Two-Dimensional Electron Gas *in Selectively* Doped N-AlGaAs/GaInAs Pseudomorphic Structures", *Jap. J. Appl. Phys.*, Vol. 27, No. 10, pp. 1931-1840, October 1988

Mawby P.A., "Characterisation and Fabrication of Heterojunction Bipolar Transistors", *PhD Thesis*, Dept. Electrical and Electronic Engineering, University of Leeds, April 1988

Mead C.A., "Schottky Barrier Gate Field Effect Transistor", *Proc. of the IEEE*, Vol. 54, No. 2, pp. 307-308, February 1966

Mimura T., Hiyamizu S., Fujii T. and Nanbu K., "A New Field-Effect Transistor with Selectively Doped GaAs/n-$Al_xGa_{1-x}As$ Heterojunctions", *Jap. J. Appl. Phys.*, Vol. 19, No. 5, pp. L225-L227, May 1980

Morkoç H. and Solomon P.M., "The HEMT: a superfast transistor", *IEEE Spectrum*, Vol. 21, pp. 28-35, February 1984

Morkoç H. , Unlu H. and Ji G., "Pprinciples and technology of MODFETs", Volume 1, *John Wiley & Sons* (Publishers), ISBN 0 471 92920 4, p.18, 1991

Morrow R.A. and Brownstein K.R., "Model effective mass Hamiltionians for abrupt heterojunctions and the associated wave-function-matching conditions", *Phys. Rev. B*, Vol. 30, No. 2, p.678, July 1984

Pantoja R.R., "Computer Aided Modelling and Characterisation of MESFET devices and Circuits", *Ph.D Thesis*, Dept. Electrical and Electronic Engineering, University of Ledds, 1990

Park K. and Kwack K.D., "A Model for the Current-Voltage Characteristics of MODFETs", *IEEE*, Vol. ED-33, No. 5, pp. 673-676, May 1986

Pierret R.F. "Volume VI, Modular series on solid state devices, Advanced semiconductor fundamentals", *Addison-Wesley Publishing Co.* (Publishers), p. 123, ISBN 0-201-05338-1, 1987

Powell M.J.D., "A Fortran Subroutine For Solving Systems of Non-linear Algebraic Equations; Numerical Methods for Non-linear Algebraic Equations", Rabinavitz P. Ed., *Gorden & Breach* (Publishers), 1970

Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., "Numerical Recipes. The Art of Scientific Computing", pp. 48, *Cambridge University Press*, 1986 and 1989, ISBN 0-521-35465-X

Rodrigues P.J.C., "Field Effect Transistor Physical Modelling for Non-Linear Subsystem Design", *Ph.D Thesis*, Dept. Electrical and Electronic Engineering, University of Leeds, 1991

Salmer G., Zimmermann J. and Fauquembergue R., "Modelling of MODFETs", *IEEE*, Vol. MTT-36, No. 7, pp. 1124-1140, July 1988

Selberherr S., "Analysis and Simulation of Semiconductor Devices", *Springer-Verlag* (Publishers), ISBN 3 211 81800 6, 1984

Sharfetter D.L. and Gummel H.K., "Large Signal Analysis of a Silicon Read Diode Oscillator", *IEEE*, Vol. ED-16, pp. 64-77, 1969

Shawki T., Salmer G. and El-Sayed O., "MODFET 2-D Hydrodynamic Energy Modeling: Optimisation of Subquarter-Micron-Gate Structures", *IEEE*, Vol. ED-37, No. 1, pp. 21-29, January 1990a

Shawki T., Salmer G. and El-Sayed O., "2-D Simulation of Degenerate Hot Electron Transport in MODFETs Including DX Center Trapping", *IEEE*, Vol . CAD-9, No. 11, pp. 1150-1163, November 1990b

Shey A.J. and Ku W.H., "On the Charge Control of the Two-Dimensional Electron Gas for Analytic Modelling of HEMTs", *IEEE*, Vol. EDL-9, No. 12, December 1988

Snowden C.M., "Semiconductor Device Modelling", *Peter Peregrinus Ltd.* (Publishers), ISBN 0 86341 130 4, 1988

Stern F., "", *Phys. Rev. Lett.*, Vol. 21, pp. 1697, 1968

Stern F., "Self-Consistent Results for n-Type Si Inversion Layers", *Phys. Rev. B*, Vol. 5, No. 12, pp. 4891, June 1972

Stern F., "Electron energy levels in GaAs-$Ga_{1-x}Al_xAs$ heterojunctions", *Phys. Rev. B*, Vol. 30, No. 2, pp. 840, July 1984

Stöer J. and Burlisch R., "Introduction to Numerical Analysis", *Springer Verlag* (Publishers), p. 385, 1980

Sze S.M., "Physics of Semiconductor Devices", *John Wiley & Sons* (Publishers) p. 250, ISBN 0-471-09837-X, 1981

Tersof J., "Theory of semiconductor heterojunctions: The role of quantum dipoles", *Physical Review B*, Vol. 30, No. 8, pp.4874, October 1984

Vinter B., "Subbands in Back-Gated Heterojunctions", *Appl. Phys. Lett.* Vol. 44, No. 3, pp. 307, February 1984

Wang C.T., "Introduction to Semiconductor Technology: GaAs related compounds", *John Wiley & Sons* (Publishers), ISBN 0 471 63119 1, pp. 555-577, 1989

Wang G.W., Chen K.W., Schaff W.J. and Eastman L.F., "A 0.1-μm Gate $Al_{0.5}In_{0.5}As/Ga_{0.5}In_{0.5}As$ MODFET Fabricated on GaAs Substrates", *IEEE*, Vol. ED-35, No. 7, pp. 818-823, July 1988

Wang T. and Hsieh C.H., "Numerical Analysis of Non-equilibrium Electron Transport in AlGaAs/InGaAs/GaAs Pseudomorphic MODFET's", *IEEE*, Vol. ED-37, No. 9, pp. 1930, September 1990

Wang W.I., "On the Band Offsets of AlGaAs/GaAs and Beyond", *Solid State Elec.*, Vol. 29, No. 2, pp. 133, 1986

Weiss M. and Pavlidis D., "The Influence of Device Physical Parameters on HEMT Large-Signal Characteristics", *IEEE*, Vol. MTT-36, No. 2, pp. 239-249, February 1988

Widiger D.J., Kizilyalli I.C. and Coleman J.J., "Two-Dimensional Transient Simulation of an Idealized High Electron Mobility Transistor", *IEEE*, Vol. ED-32, No. 6, p. 1092, 1985

Wilkinson J.H., "The Algebraic Eigenvalue Problem", *Springer Verlag* (Publishers), p. 300, 1971

Yablik M.C., Neofotistos G., Diff K., Guo H. and Gunton J.D., "Quantum Mechanical Simulation of Charge Transport in Very Small Semiconductor Structures", *IEEE*, Vol. ED-36, No. 6, pp. 1009, June 1989

Yoshida J., "Classical Versus Quantum Mechanical Calculation of the Electron Distribution of the n-AlGaAs/GaAs Heterointerface", *IEEE*, Vol. ED-33, No. 1, p. 154, 1986

Appendix A

Derivation of the electron charge densities

One of the fundamental principles of quantum mechanics is that when taken to the macroscopic limit the results derived should agree with those predicted by classical mechanics. This 'correspondence principle' is demonstrated for the bulk three-dimensional electron density in semiconductors.

In general the electron distribution is given as the integral over all space and time of the square of the electron wavefunction.

$$n = \langle \Psi^*(r,t)\, \Psi(r,t) \rangle \qquad \text{(A.1)}$$

where the angle brackets denote summation over the entire ensemble of electrons, and $\Psi(r,t)$ is given from a solution of the Schrödinger equation, equation (A.2).

$$-\frac{\hbar}{i}\frac{\partial \Psi(r,t)}{\partial t} = H\,\Psi(r,t) \qquad \text{(A.2)}$$

If H, the Hamiltonian or total energy operator, is independent of time then the total wavefunction may be split into temporal and spatial components, $\Phi(t)$ and $\psi(r)$ respectively, giving equation (A.3).

$$\Psi(r,t) = \sum_m \Phi_m(t)\, \xi_m(r) \qquad \text{(A.3)}$$

Upon substitution into equation (A.1) this gives

$$n = \sum_m \sum_n \left\langle \Phi_m^*(t)\, \Phi_n(t) \right\rangle \xi_m^*(r)\, \xi_n(r) \qquad \text{(A.4)}$$

Since $\Phi_m(t)$ is summed over a very large number of electrons, and involves time averaging over long periods compared with the time scale of the microscopic fluctuations ($\sim 10^{-12}$ seconds) it can be shown that

$$\left\langle \Phi_m^*(t)\, \Phi_n(t) \right\rangle = \begin{cases} f_o(E_m) : & m = n \\ 0 : & m \neq n \end{cases} \qquad \text{(A.5)}$$

where $f_o(E_m)$ is the Fermi-Dirac factor, equation (A.6).

$$f_o(E_m) = \frac{1}{e^{(E_m - E_F)/k_B T} + 1} \qquad \text{(A.6)}$$

thus the equation for electron density becomes

$$n(r) = \sum_m f_o(E_m) \left| \xi_m(r) \right|^2 \qquad \text{(A.7)}$$

The form of ξ_m, the spatial electron wavefunction, depends upon the scale of the events it encounters with respect to it's own wavelength. If the scale of the conduction band edge is large compared with the deBroglie (or thermal) length, equation (A.8).

$$\lambda_{th} = \frac{h}{m^* v} \qquad \text{(A.8)}$$

then the wavefunction can be approximated by Bloch states, equation (A.9).

$$\xi_k(r) = \frac{e^{ik.r}}{\Omega^{\frac{1}{2}}} \tag{A.9}$$

where Ω is the volume of the device.

When substituted into the Schrödinger equation the corresponding eigenstates are given by equation (A.10).

$$\lambda_k = E_{C0} + \frac{\hbar^2 k^2}{2m^*} \tag{A.10}$$

It is worth noting that here the electron density is independent of position.

The summation over k is converted to an integral over energy and the classical Fermi-integral to the power half is formed.

$$n = \frac{(2m^*)^{\frac{3}{2}}}{2\pi^2\hbar^3} \int_{E_{C0}}^{\infty} \frac{(E-E_{C0})^{\frac{1}{2}} dE}{e^{(E-E_F)/k_BT} + 1} \tag{A.11}$$

which can be rearranged to equation (A.12).

$$n = N_C \frac{2}{\sqrt{\pi}} \mathcal{F}_{\frac{1}{2}}\left(\frac{E_F-E_{C0}}{k_BT}\right) \tag{A.12}$$

where

$$N_C = \frac{1}{4\hbar^3}\left(\frac{2m^*k_BT}{\pi}\right)^{\frac{3}{2}} \tag{A.13}$$

In practical heterojunction FET's, typically doped in excess of 10^{24} donors m^{-3}, one cannot assume the electrons wavefunctions to be accurately approximated by Bloch states as the abrupt change in conduction band edge creates a potential well on the scale of the order 10-20 nm. This is clearly comparable or smaller than the deBroglie wavelength, which is ~30 nm in GaAs at 300K. In this case 'quantum sub-bands' are formed accross the heterojunction, and if the energy separation between consecutive energy levels is greater than the thermal energy, k_BT, then the electron's motion in this direction will become discretised. As the smallest event in the directions parallel to the heterojunction is produced by the gate, ~500 nm, one can assume the electron wavefunctions in these directions can still be approximated by the the classical Bloch states. Thus the spatial electron wavefunction is split further, into components parallel and normal to the heterojunction interface. In this case the electron wavefunction has the form of equation (A.14).

$$\xi_{m,k_x,k_z}(r) = \frac{e^{ik_x x} e^{ik_z z}}{\sqrt{A}} \xi_m(y) \qquad \text{(A.14)}$$

where A is the cross-sectional area.

Substituting this into the time-independent Schrödinger equation we obtain the corresponding eigenstates.

$$\lambda_{m k_x k_z}(r) = \lambda_m + \frac{\hbar^2}{2m^*}\left(k_x^2 + k_z^2\right) \qquad \text{(A.15)}$$

With the electron density, in this case, given by equation (A.16).

$$n(y) = \sum_m \sigma_m |\xi_m(y)|^2 \qquad \text{(A.16)}$$

where

$$\sigma_m = \sum_{k_x,k_z} \frac{f_o(\lambda_m,k_x,k_z)}{A} \qquad \text{(A.17)}$$

Again converting the summation over k_x and k_y to an integral over energy it can be shown that

$$\sigma_m = \frac{m^*}{\pi \hbar^2} \int_{E_m}^{\infty} \frac{dE}{e^{(E-E_F)/k_BT} + 1} \qquad \text{(A.18)}$$

hence

$$\sigma_m = \frac{m^* k_B T}{\pi \hbar^2} \ln\left(1 + e^{-(\lambda_m - E_F)/k_BT}\right) \qquad \text{(A.19)}$$

and finally

$$n(y) = \frac{m^* k_B T}{\pi \hbar^2} \sum_m |\xi_m(y)|^2 \ln\left(1 + e^{-(\lambda_m - E_F)/k_BT}\right) \qquad \text{(A.20)}$$

where both $\xi(m)$ and $\lambda(m)$ are given by a solution of the one-dimensional Schrödinger equation in the direction normal to the heterojunction.

Appendix B

Derivation of the classical charge transport equation

In steady state flow the electron distribution function, which is a function of space, momentum, will be different from that in thermal equilibrium in the absence of flow, f_o. For small changes from equilibrium the new distribution function can be expanded in terms of the original distribution via perturbation theory, giving, to a first approximation

$$f(\vec{x},\vec{k},t) = f_o(\vec{x},\vec{k}) + f'(\vec{x},\vec{k},t) \qquad \textbf{(B.1)}$$

The steady state is defined by Liouville's equation *(B.2)* which when explicitly expanded gives the Boltzman Transport Equation (BTE) equation *(B.3)*

$$\frac{df(\vec{x},\vec{k},t)}{dt} = 0 \qquad \textbf{(B.2)}$$

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial \vec{k}} \cdot \frac{\partial \vec{k}}{\partial t} + \frac{\partial f}{\partial \vec{x}} \cdot \frac{\partial \vec{x}}{\partial t} = 0 \qquad \textbf{(B.3)}$$

substituting F for $\frac{\partial \vec{k}}{\partial t}$ and u for $\frac{\partial \vec{x}}{\partial t}$ one obtains the BTE in its more conventional form

$$\frac{\partial f}{\partial t} = -\frac{\vec{F}}{\hbar}.\nabla_k f - \vec{u}.\nabla_x f \qquad \text{(B.4)}$$

For small perturbations the distribution function can be assumed to relax back to equilibrium in an exponential fashion, with a time constant τ.

$$f - f_o = (constant)\, e^{-t/\tau} \quad : \quad \frac{\partial f}{\partial t} = -\frac{f - f_o}{\tau} \qquad \text{(B.5)}$$

Substitution of *(B.5)* into *(B.4)* gives a simplified Boltzman Transport Equation, from which the first order perturbation correction can be calculated.

$$f'(\vec{x},\vec{k},t) = \tau\left(\frac{\vec{F}}{\hbar}.\nabla_k f_o + \vec{u}.\nabla_x f_o\right) \qquad \text{(B.6)}$$

From semiconductor statistics the equilibrium electron distrubution functions is obtained from Fermi-Dirac statistics.

$$f(\vec{x},\vec{k},t) = \frac{1}{e^{(E-E_F)/k_BT}+1} \qquad \text{(B.7)}$$

If $E - E_F > k_BT$ then the unity term in the denominator is negligible, and hence Boltzman statistics can be used

$$f(\vec{x},\vec{k},t) = e^{-(E-E_F)/k_BT} \qquad \text{(B.8)}$$

where E, the energy of the electrons, is the sum of their potential and kinetic

components

$$E(\vec{k},\vec{x}) = E_{co}(\vec{x}) - q.\psi(\vec{x}) + \frac{\hbar^2.\vec{k}.\vec{k}}{2m^*} \qquad \textbf{(B.9)}$$

When either of these distribution functions are substituted into *(B.6)* the partial derivatives of f with respect to x and k can be evaluated.

$$\nabla_k f_o = \left(\frac{\hbar^2.\vec{k}}{m^*}\right)\frac{\partial f_o}{\partial E} = \hbar\vec{u}\frac{\partial f_o}{\partial E} \qquad \textbf{(B.10)}$$

$$\nabla_x f_o = \left(-\vec{F} - \nabla_x E_F - \frac{(E-E_F)}{T}.\nabla_x T\right)\frac{\partial f_o}{\partial E} \qquad \textbf{(B.11)}$$

where

$$\frac{\partial f_o}{\partial E} = \frac{-f_o(1-f_o)}{k_B T} \qquad \textbf{(B.12)}$$

for F-D statistics and

$$\frac{-f_o}{k_B T} \qquad \textbf{(B.13)}$$

for Boltzman statistics.

Assuming the temperature to be constant throughout the device equation *(B.11)* can be simplified and the first order correction is calculated as

$$f' = \tau \vec{u} \frac{\partial f_o}{\partial E} \nabla_x E_F \qquad \textbf{(B.14)}$$

The electron current is the product of the group velocity and distribution function summed over all k-space

$$\vec{J} = \frac{-q}{4\pi^3} \int_{V_k} \vec{u} \left(f_o + f' \right) d\vec{k} \qquad \textbf{(B.15)}$$

Substituting *(B.14)* for f' gives

$$\vec{J} = \frac{-q}{4\pi^3} \int_{V_k} \vec{u} f_o - \tau f_o (1 - f_o) \frac{\vec{u}.\vec{u}}{k_B T} \nabla_x E_F \qquad \textbf{(B.16)}$$

which gives

$$\vec{J} = q \mu n \nabla_x \phi \qquad \textbf{(B.17)}$$

where $q\phi = E_F$ and $\mu = \frac{q\tau}{m^*}$

Publications

'*The Physical Modelling of Single and Multichannel Heterojunction FETs*', 6th Meeting of the International GaAs Simulation Group, III-V Device Simulation Workshop, Lille, October 1991

'*A New Full Two-Dimensional Quantum Model for Pseudomorphic HFETs*', 7th Meeting of the International GaAs Simulation Group, III-V Device Simulation Workshop, Harrogate, April 1993

'*A New Fast Full 2D Quantum Model for Heterojunction Field Effect Transistors*', International Workshop on Computational Electronics, Leeds, August 1993

'*Two-Dimensional Quantum Modelling of Millimetre-Wave Heterojunction Field Effect Transistors*', IEE colloquium of the Modelling of Millimetre Active Devices and Circuits, Leeds, September 1993 (Accepted but postponed)