UNIVERSITÁ DEGLI STUDI DI UDINE

Dottorato di Ricerca in Ingegneria Industriale e dell'Informazione

Dipartimento Politecnico di Ingegneria e Architettura

Oves BADAMI

Numerical models, simulation tools, analysis and design of nanoscale MOS transistors

Advisor:

Prof. David ESSENI

External reviewers:

Prof. Elena GNANI

Prof. Hans KOSINA

Ph.D. Thesis

Ciclo XXX

Abstract

Mind boggling scaling of CMOS technologies for last five decades is an unparalleled engineering feat. This has been the source of the spectacular rise of the microelectronic industry. The Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET) has been scaled to reduce the cost of fabrication by packing more transistors onto the same area. The reduced dimensions also result in a reduction of the time taken by an electron to move from source to drain and thus improving the performance of the device. In addition to the scaling of physical dimensions, a number of modifications were made to the device structure that have played a vital role in order to keep Moore's law alive. The two major changes were: (a) adopting high-κ gate oxide when scaling of the then existing SiO_2 was no longer feasible; (b) adopting multigate architectures to maintain the electrostatic integrity and improve the drive current of the devices at extremely small channel lengths. These technology innovations have made the oxide-semiconductor interface more important than ever before. In fact, it has already been shown with simulations that the scattering due imperfections at the oxide-semiconductor interface is today the dominant cause of mobility degradation in nanoscale MOS transistors.

With the tremendous increase in the cost and complexity of the fabrication, the modeling and simulations provide an alternative way to evaluate the performance of the transistors before going into the fab. Of course, this demands that the models used in the simulations have a predictive ability namely they must be able to model the device operation on the basis of a sound physics content, rather than by multiplying the number of fitting parameters.

As a part of this work, we first extended the formulation of a nonlinear model for surface roughness scattering in planar transistors to account for tensorial carrier screening, which is important in multigate devices. The model was then incorporated in an existing comprehensive Multisubband Monte Carlo simulator. This allowed us to analyze a wide range of planar structures from bulk devices to heterostructure devices. Motivated by the good results from the nonlinear model for planar devices, we developed a nonlinear model for surface roughness scattering for the 3D devices having fairly arbitrary cross-section shape, so as to analyze transistors without any approximation to their geometrical structure. This nonlinear surface roughness scattering model was then incorporated along with other important scattering mechanisms in a simulation framework largely extended and improved starting from the existing solver for Schrödinger-Poisson and 1D Boltzmann Transport Equation to analyze 3D FETs with arbitrary cross-section shape and biasing scheme. This simulation framework was then used extensively to calculate mobility values for silicon and InAs based multigate FETs having different cross-section shapes and areas. The simulator was then extended to account for the impact of self-heating and series resistance on the device characteristics, as they are known to significantly degrade the performance of the transistor. In order to simulate a complete transfer characteristic with all the scattering mechanisms, we developed a novel iteration scheme and a method to solve the Schrödinger equation

adaptively to minimize the time needed in the simulation. The mobility analysis was then extended to a thorough comparison of different performance metrics. We have also benchmarked vertically stacked lateral nanowires (stacked NWs) against traditional FinFET, and simulations showed that surface roughness has a much stronger impact on the stacked NWs.

It is a well-known fact that the dispersion relationship (electronic band structure) for confined devices is very different from their bulk crystals and it depends strongly on the confinement. Moreover the cross-section shape and growth direction of the nanowire can also have a significant impact on the dispersion relationship. In this thesis, we have also developed a semi-empirical tight-binding simulator for bulk and nanowires with arbitrary cross-section shape. In particular, we integrated the solver with a general purpose mesh generator so as to calculate the band structures for a fairly arbitrary device features and crystallographic orientations.

Contents

Abstract **iii**

1 Introduction **1**
1.1 Journey of the Field Effect Transistor . . . 1
1.2 Modeling and simulations of semiconductor devices . . . 2
1.2.1 Importance of oxide-semiconductor interface . . . 6
1.2.2 Importance of electronic band structure . . . 7
1.3 Thesis organization . . . 8

2 Surface roughness scattering in planar transistors **11**
2.1 Introduction . . . 11
2.2 Linear and nonlinear model for surface roughness scattering . . . 12
2.2.1 Linear model . . . 14
2.2.2 Nonlinear model . . . 15
2.2.3 Additional comments . . . 17
2.3 Screening . . . 18
2.3.1 Physical interpretation and dielectric function . . . 18
2.4 Application to linear model . . . 19
2.5 Application to nonlinear model . . . 20
2.6 Multisubband Monte Carlo simulation framework . . . 21
2.6.1 Basics of Monte Carlo . . . 22
Step1: Calculation of free flight duration . . . 22
Step2: Identification of scattering mechanism and final state . . . 23
2.6.2 Additional scattering mechanisms . . . 24
Acoustic and optical non-polar phonon scattering . . . 24
Polar optical phonon scattering . . . 25
Alloy scattering . . . 25
Coulomb scattering . . . 26
2.6.3 Parallelization . . . 26
2.7 Validation of the nonlinear surface roughness scattering implementation . . . 27

3 Mobility analysis of planar transistors **29**
3.1 Introduction . . . 29
3.2 Comparison with experimental mobilities: Bulk devices . . . 29
3.3 Comparison with experimental mobilities: Thin body devices . . . 31

3.4 Impact of the energy barrier between buffer and active layer 33

4 Surface roughness scattering in 3D transistors 37
4.1 Introduction 37
4.2 Nonlinear model for surface roughness scattering 38
4.2.1 Coordinate system 38
4.2.2 Calculation of the matrix elements 38
4.2.3 Calculation of the ensemble averaged squared matrix elements 40
4.2.4 Simplifications to ensemble averaged squared matrix elements 42
4.3 Validation of approximations 44
4.3.1 Number of modes 44
4.3.2 Simplified ensemble averaged squared matrix elements 45
4.4 Screening 46
4.5 Some more comments on use of Fourier series 47

5 Mobility analysis in 3D transistors 49
5.1 Introduction 49
5.2 Simulator description 49
5.2.1 Schrödinger-Poisson solver 50
5.2.2 Boltzmann transport equation 51
Mobility calculations 53
5.3 Comparison with experimental data 54
5.3.1 Gate-all-around architecture 54
5.3.2 Back gate biased transistor 55
5.4 Mobility analysis for different cross-section shapes and materials 57

6 Analysis of different architectures for 3D Devices transistors 61
6.1 Introduction 61
6.2 Simulation frame work 62
6.2.1 Series resistance 63
6.2.2 Self-heating methodology 63
6.2.3 Optimization schemes 65
Schrödinger solver 65
Optimized iteration scheme 66
6.3 Performance analysis 67
6.3.1 Device architecture and material analysis 69
6.3.2 Impact of surface roughness on I_{on} 72
6.3.3 Impact of self-heating on I_{on} 75

7 Electronic band structure calculation: Tight binding method 79
7.1 Introduction 79
7.2 Bulk 80
7.2.1 Theory 80

7.2.2 Validation 83
7.3 Comments on the basis set 84
7.4 Confined Structure: Nanowires 86
7.4.1 Arbitrary Orientation 86
7.4.2 Theory 87
Arbitrary cross-section 88
7.4.3 Validation 89

8 Conclusion and future work **93**
8.1 Conclusion 93
8.2 Future Work 94

A Fine print in matrix element derivation **95**

List of Figures

1.1 Illustration of different methodologies to simulate the transport a transistor. It also shows the optimum transport formulation for simulating the present and near future transistors. 4

2.1 Sketch of a planar device indicating the coordinate system. 13
2.2 Sketch of the cross-section of a planar device illustrating the concept of screening. 18
2.3 Comparison of the mobility versus Si thickness for Si double gate SOI MOSFETs using different screening formulations and without screening, and either using the linear or the nonlinear surface roughness scattering models. For DG structures there is a significant inconsistency in mobility values using either the tensorial or the scalar screening, as already pointed out in [24]. The surface roughness parameters, Δ_{rms} = 0.66 nm and Δ_{rms} = 0.21 nm has been used with linear and non-linear model respectively. The correlation length, Λ = 1.4 nm is used for both models. 27

3.1 Plot of comparison between the simulated and experimental mobility versus effective electric field for a bulk unstrained silicon MOSFETs for different temperatures. The surface roughness parameters used are Δ_{rms}=0.21 nm and Λ=1.4 nm with the nonlinear model which are very well in the experimental range. 30
3.2 Comparison between the experimental and simulated mobilities (with linear and nonlinear model for surface roughness scattering). $\Delta_{rms}= 0.4$ nm and $\Lambda= 1.5$ nm were needed with the nonlinear surface roughness model. The data for the linear model has been taken from [46] ($\Delta_{rms}= 1.3$ nm and $\Lambda= 1.5$ nm). The channel doping of 2×10^{17} cm^{-3} was considered in the simulations. Other scattering parameters are same as in [46]. Experimental data are taken from [66, 67]. 30
3.3 (a) Schematic of the fabricated (and simulated) structure. (b)Comparison between the experimental [70] and simulated mobilities. We emphasize that the Δ_{rms}=0.55 nm used for simulations is same as experimentally measured. . . . 31
3.4 (a) Schematic of the fabricated (and simulated) structure. (b) Surface roughness scattering was also considered at the active-buffer layer interface with the parameters listed in Table.3.1. We also assumed interface charge density of 2×10^{12} cm^{-2} to improve the match between the experiments and simulations. 32

3.5 Comparison between the experimental and simulated values of mobility for different values of well thichkness,T_w from Fig.3.3 and Fig.3.4. The degradation of the mobility with well thickness is much lower than the T_w^6 trend. . . . 33
3.6 Comparison between the experimental and simulated mobilities for bulk-like structure (shown on left). The small value of Δ_{rms} needed to match the experimental mobilities suggest an interface quality comparable with Si-SiO_2. . 33
3.7 Shown is the impact of the energy barrier between the active and buffer layers on mobility at fixed inversion density, N_{inv}=6×10^{12} cm^{-2}. The SR parameters are the same as in Fig.3.4. The structure and the geometrical dimensions used in the simulation are shown on the right. 34
3.8 Wavefunction corresponding to the lowest subband for the III-V heterostructure shown in the figure 3.7. The value of the wavefunction at the active-buffer layer is much larger than that at the oxide-buffer layer. The surface roughness scattering and material parameters are listed in Table3.1 and Table3.2 respectively. 35

4.1 (a) sketch of the cross-section of a MuGFET where x is the transport direction and $\mathcal{I}_0$ denotes the curve describing the semiconductor-oxide interface. s and η are respectively the abscissa along $\mathcal{I}_0$ and normal to $\mathcal{I}_0$, and $\mathcal{I}_\eta$ is the curve at a distance η from $\mathcal{I}_0$. D_0 and D_η are the perimeters of $\mathcal{I}_0$ i.e. the perimeter of the semiconductor region) and of $\mathcal{I}_\eta$. (b) sketch of the cross-section of a MuGFET perturbed by a surface roughness stochastic process $\Delta(s,x)$. The dashed area is the perturbed region $P_R\left[\Delta(s,x)\right]$ where $\Delta(s,x)$ is non null; the grey region is the semiconductor. 39
4.2 Numerical recipe for calculation of the $\left\langle \left| \mathcal{M}_n^{n'}(q_x) \right|^2 \right\rangle$. 43
4.3 Form factors as calculated from Eq.4.20 for D_0=10 nm (a) and D_0=40 nm (b). The $q_{gg'}$=$2\pi(g'-g)/D_0$ values are obtained by setting g=0 and g'=$0,2,4,6$. The values of the form factor where ($q_{ll'}=q_{gg'}$) is dominant one (Approx1) and with increase in D_0 Form factor becomes more representative of dirac delta function (Approx2). 43
4.4 Plot of the lowest subband wavefunction in a triangular GAA nanowire (cross-section area = 40 nm^2) in the two valleys described by (a)$m_{yy}=0.916m_0$ and $m_{zz}=0.190m_0$ (b)$m_{yy}=0.190m_0$ and $m_{zz}=0.916m_0$ for different modes. . . 45
4.5 $\langle|\mathcal{M}_n^{n'}(q_x)|^2\rangle$ for intra-subband transitions in the lowest subband (i.e. n=n'=0), for Si$-SiO_2$ FETs and different cross-sections. The oxide thickness is 1 nm and the electron density is $N_{inv}\approx10^{13}$ cm^2. In all plots: (a) Tri-Gate rectangular FinFET; (b) triangular GAA nanowire FETs. Area is 40 nm^2 for both the cross-sections. Transport direction is $[100]$. The insets shows colormaps of the electron concentration profiles in both the cross-section emphasizing the dissimilarity between the two architectures. 45
4.6 Plot of mobility versus inversion charge density for (a) FinFET and (b)GAA triangular cross-section at a fixed area = 40 nm^2. 46

5.1 Illustration of the discretized (x,E_n) space. Filled symbols are the actual unknowns while the unknowns from the neighboring section are open symbols. The mapping the unknowns from the adjacent sections are illustrated with the arrows from the sections $x_i \pm 1$ to x_i. Dashed oval clubs the points having the same occupation function value in the uniform transport as they have the same kinetic energy. The discretization in the energy and transport direction are linked by $\Delta E = eF\Delta x$. 54
5.2 Mobility versus inversion charge density for a circular (R=7.5 nm) gate-all-around nanowire. A good match with the experiments is obtained at higher inversion density with Δ_{rms}=0.17 mn and Λ=1.4 nm. The results show that impact of screening is modest. The material parameters in the simulations are listed in Table.5.1 and 5.2. Experimental data is taken from [109]. 56
5.3 (a) Simulation domain for the back-gated nanowire. Homogeneous Neumann boundary conditions are imposed in the external domain with dashed lines, whereas Dirichlet boundary conditions are imposed at the bottom of the SiO_2 film (solid line); (b) Squared magnitude of the lowest subband wavefunction at $N_{inv}\approx 1\times 10^7$ cm^{-1} for back-gated nanowire. (c) Simulation domain for the same nanowire with a gate-all-around biasing condition. and (d) squared magnitude of the lowest subband. 56
5.4 Linear fit to the experimental values capacitance versus the radius of nanowire. 57
5.5 Comparison between the experimental and simulated mobility values for a backgated InAs nanowire (radius=7.5 nm) FET. When the nanowire is simulated with the gate all around biasing scheme results in overestimation of the mobility values. 58
5.6 (a) Impact of linear inversion charge density on mobility for gate all around triangle, square, circle and trigated FinFET devices for fixed area, A$\approx$40 nm^2. (b) Impact of the cross-section area on mobility for different cross-section shapes at a fixed inversion charge density (normalization of the inversion charge density is done with the gate perimeter). All the device have similar mobility values across the range of area and inversion density considered. . . 58
5.7 Plot of mobility versus area for different cross-section shapes with InAs channel for (a) $N_{inv} = 2 \times 10^{12} cm^{-2}$ and (b)$N_{inv} = 5 \times 10^{12} cm^{-2}$. Linear to areal charge density conversion is done with the gate perimeter. At lower N_{inv} the mobility follows a trend of A^3 at small cross-section areas however at higher N_{inv} mobility oscillates depending on the position of the subbands with respect to fermi level Fig.5.8. We have also included mobility curves for silicon for comparison. 59

5.8 (a) Plot of the mobility versus area for FinFET at $N_{inv} = 5 \times 10^{12}$ cm^{-2}. (b) Subband energies with reference to the Fermi level (0 eV) for the lowest subbands in the Γ and L valley. (c) plot of $f(E)\,[1 - f(E)]$ assuming equilibrium occupation function. (d) Total scattering rate for the second subband in device with $A = 20$ nm^2. As the subbands move into the window having non negligible value of $f(E)\,[1 - f(E)]$ there is sharp increase in the intersuband scattering which reduces the mobility. 60

6.1 Standard program flow for one bias point. It must be noted in this methodology the scattering calculations are included in the simulations from the first iteration. 65

6.2 Illustrates the lowest subband profile by solving Schrödinger equation (empty symbols) in all the slices, with simple (solid line) and modified (dashed line) optimized and for circular gate-all-around nanowire. 66

6.3 Optimized program flow which allows to progressively turn-on the scattering mechanism unlike the standard program flow is shown in Fig.6.1. With the new iteration scheme, the computationally heaviest scattering mechanism needs to be included only in the last few iterations. 68

6.4 Schematic of the different architectures simulated in this work. (a) FinFET with W:H=1:2, (b) stacked NWs with H=W, (c) Circular NW and (d) Square NW. All the architectures have same footprint because of same width (W=5 nm) channel length (Lg=14 nm) and source/drain length (Ls=Ld=25 nm). Transport is oriented along $\langle 110 \rangle$ and $\langle 100 \rangle$ for respectively sSi and $In_{0.53}Ga_{0.47}As$. Source and drain doping were taken to be 10^{20} cm^{-3} and $5{\times}10^{19}$ cm^{-3} for sSi and $In_{0.53}Ga_{0.47}As$ based devices, respectively. The oxide is HfO_2 for all the architectures with T_{ox}=2.8 nm (EOT=0.5 nm). The footprint is defined as W×(Ls+Lg+Ld). 69

6.5 Simulated I_{DS} versus V_{GS} for strained silicon and $In_{0.53}Ga_{0.47}As$ for stacked nanowires and FinFET. (a) The current is normalized by the gate perimeter while in (b) the absolute value of currents are plotted. All the devices have the same footprint. Surface roughness parameters used in the simulations are Δ_{rms}=0.21 nm and Λ=1.4 nm for both the silicon and $In_{0.53}Ga_{0.47}As$. 70

6.6 Plot of FO3 delay versus the parasitic capacitance. The simulation results show that with the increase in C_{par} the delay degradation for FinFET and stacked NWs is lower than for other GAA devices. This advantage of FinFET and stacked NWs for large C_{par} values stems from their large drive current per unit footprint. 71

6.7 Impact of $\Delta_{rms}$of surface roughness on the on-current for (a) strained silicon and (b) $In_{0.53}Ga_{0.47}As$ based FinFET and stacked nanowires. The current degrades super-linearly with an increase in Δ_{rms}. The results highlight that the impact of surface roughness scattering is architecture dependent. Moreover, at larger $\Delta_{rms}$the $In_{0.53}Ga_{0.47}As$ based transistor provide larger current than strained silicon-based FETs. The larger ballistic and phonon limited current is larger in silicon than $In_{0.53}Ga_{0.47}As$ because of larger N_{sh} which more than compensates for the lower electron velocity. 73
6.8 Schematic illustrating the coordinate system in an arbitrary shaped cross-section. 73
6.9 Plot of $M_{nn}(s)$ with n=0 along 's' for (a) strained silicon and (b) $In_{0.53}Ga_{0.47}As$based architectures. The range of integration has been done in the range $-3\Delta_{rms} \leq \eta \leq 3\Delta_{rms}$. Larger values of $M_{nn}(s)$ for stacked nanowires as compared to FinFET clarify the stronger degradation observed in stacked nanowires with an increase in surface roughness. 74
6.10 The impact of surface roughness on the on-current by with and without intersubband scattering due to surface roughness for (a) FinFET and (b) stacked nanowires. The simulations suggest that intersubband scattering due to surface roughness is an important reason for degradation of the on-current. . . . 74
6.11 Density of states at virtual source of FinFET at V_{GS}=0.7 V (lowest subband energy is taken as 0 eV). The close spacing of the subbands for sSi as compared to $In_{0.53}Ga_{0.47}As$ result in larger number of intersubband transitions in the former. The transition from lower subband to higher subband is more favorable because of the larger DoS in the later. 75
6.12 Effective temperature profile for FinFET architecture for different values of the thermal conductivity. 75
6.13 Impact of thermal conductivity of semiconductor on the on-current for circular gate-all-around and FinFET. The simulation results highlight that different architectures show different degradation for the same value of thermal conductivity. Vertical arrows show experimental thermal conductivity for different nanowire width, W [117]. 76
6.14 Plot of electron velocity and linear electron density at virtual source for (a) circular gate-all-around nanowire and (b) FinFET. In both the cases even though the charge concentration increases slightly with the reduction in the thermal conductivity the reduction in the velocity is much more and hence it results in the degradation of the on-current. The reduction in V_x with decreasing κ_{sc} (that is with increasing temperature - see Fig.6.12), is due to an increased phonon scattering and it is contrary to the ballistic behavior because the ballistic injection velocity (thermal velocity), increases with the increase in temperature [36]. 77

7.1 Illustration of the projection of the 3D crystal lattice on a plane. In the diamond crystal structure both the atoms are of the same type. The connections C1-A2, C1-A3, C1-A4 are related to C1-A1 connection through symmetry relations [147]. Similar considerations are also valid for anion atom. 83
7.2 Comparison of the band structures extracted from [149] and calculated in this work for silicon along [100] and [110] directions. In all there are 19 fitting parameters corresponding to various hopping and onsite terms for group IV semiconductors which are taken from [149]. 84
7.3 Same as 7.2 except for germanium. 84
7.4 Comparison of the parabolic effective masses approximation and lowest conduction band obtained from the tight binding method for (a) silicon at X point and (b) germanium at Γ and L points. The values of the effective masses used in the parabolic E-k relationship have been taken from [36]. 85
7.5 Shown is a projection of a unit cell for a circular nanowire with a radius of 0.3 nm with [100] orientation. Different symbols ('cross' and 'circle') represent 2 different types of atoms in a zinc-blende crystal structure. The different bonds between 3^{rd} atom and its nearest neighbors are also illustrated. 88
7.6 Illustration of a continuous arbitrary cross-section (solid line) along with discretized cross-section (dashed line). Control point is any point that lies inside the cross-section while a test point can lie any where in the plane. 89
7.7 Comparison between the E-k relationship for germanium nanowire with diameter of 1.6 nm whose transport is oriented along [100], [110] and [111] crystallographic direction. It also shows the 2D projection of atoms (circle and cross) in the unit cell that is used in the simulations. The reference data is taken from BandStructure Tool ([155]) 90
7.8 Comparison between the E-k relationship for silicon nanowire with diameter of 1.6 nm along whose transport is oriented along [100], [110] and [111] crystallographic direction. It also shows the 2D projection of atoms (circle and cross) in the unit cell that is used in the simulations. The reference data is taken from BandStructure Tool ([155]) 91

List of Tables

2.1 The values of Δ_{rms} and Λ reported in experiments for (100) Si MOSFETs [22]. 11

3.1 Surface Roughness parameters used in simulations. 31
3.2 Material parameters for the semiconductor and dielectrics employed in the simulations. 32

5.1 Bulk longitudinal and transverse effective mass, non-parabolicity coefficient, α and electron affinity, χ for InAs. The energy difference between the L and Γ conduction band minima in InAs is taken to be 0.716 eV [69]. 54
5.2 Effective mass, $m\,(\Gamma)$, electron affinity, χ, and relative permittivity, κ, for different oxides used in the analysis of the experimental. 55

Chapter 1

Introduction

1.1 Journey of the Field Effect Transistor

The rich history of the field effect transistor, the workhorse of the semiconductor industry, started with a patent filed by Julius Edgar Lilienfeld in 1925. The patent provided essentially the working principle of the field effect transistor. The use of the solid-state device was proposed as a replacement to the thermionic tubes that were used. This substitution was important because of myriad problems like reliability, size and power consumption associated with the use of vacuum tubes. The real breakthrough for field effect transistor didn't come until 1960 when Atalla and Kahng first demonstrated the working Field Effect Transistor [1]. The reason for such delay was the presence of surface states which blocked the gate field and thus hampered the proper functioning of the field effect transistor.

With the invention of the integrated circuit by Jack Kilby and Robert Noyce and the successful fabrication of metal-oxide-field effect transistor, the era of complementary Metal-Oxide-Semiconductor Field Effect Transistor began. In 1965, Gordon Moore with the help of just four data points predicted that the number of components on a chip would double every year [2]. The scaling however has slowed down and the packing density is now doubled every two years. This is known as the Moore's law. Since 1965 it has become the guiding principle for the semiconductor industry.

Complying with Moore's law provided a double advantage to the semiconductor industry. With the reduction of the feature sizes the companies were able to pack more transistors per unit chip area that reduced the fabrication cost, and secondly it made the transistors work faster (electrons had to travel smaller distance). Even though the Moore's law proposed that the devices must be scaled, the question of how it can be done was not answered. It was realized that direct scaling of the channel length degrades the device characteristics and hence Dennard *et al.* developed a set of general guidelines related to device design in order to correctly scale the device dimensions [3]. These laws also quantified the improvement in the device performance that can be achieved with the scaling. In order to reap these double advantage the semiconductor industry started on the journey of scaling of the device dimensions, which has resulted in the reduction of channel length from few micrometers to a few nanometers.

Soon it was realized that simply following the prescription laid down by Dennard *et al.* was not possible because reducing the dimension of the gate oxide would lead to increase in

the tunneling through the gate dielectric. This would adversely impact the performance of the device and it would increase the off-state power dissipation. The scaling of the gate oxide was extremely crucial in maintaining the electrostatic integrity of the device. After year 2000 the SiO_2 gate oxide was replaced by hafnium-based oxide which had a higher permittivity (high-κ material). This essentially allowed to maintain the electrostatic integrity without reducing the physical gate oxide thickness.

The use of high-κ materials for gate oxide prolonged the life of planar architecture by a few more years until it became apparent that it wasn't possible to continue the scaling in the bulk planar devices because of the detrimental effects of the close proximity between the source and channel (commonly knows as short channel effects). The semiconductor industry rose up to the challenge by adopting new 3D architecture [4]. The use of 3D architecture wasn't surprising because it was already recognized that the 3D architectures have better immunity to short channel effects [5]. These architectures have more than one gate thus improving the gate control. This has allowed continuing to scale channel to length less than 14 nm.

The amazing progress and innovations in the fabrication technology have provided an unusual device architecture to continue the scaling in the sub 7 nm range (channel length less than 12 nm). In this architecture, lateral gate all around nanosheets are vertically stacked and thus they deliver better dynamic and electrostatic performance as compared to FinFETs with very small modification to the fabrication process [6].

The evolutionary methods in the past have prolonged the ability to scale the transistors and have allowed the industry to continue to reap different benefits from it. Evidently, the scaling of the physical dimensions will hit a roadblock as off-state current increases (due to source to drain tunneling or poor electrostatics). Hence, it will be necessary to abandon the evolutionary devices and adopt revolutionary ideas like the use of 2D materials or devices based on concepts other than electron transport.

1.2 Modeling and simulations of semiconductor devices

As we have seen earlier, the tremendous progress in the semiconductor fabrication technology has allowed the manufacturer to continue packing more and more transistors per unit area on the chip and keep Moore's law alive. This has increased the fabrication complexity because of use of high-κ materials and use of 3D architectures. In order to fabricate these devices, the cost of the tools required to manufacture has gone up as well which has a negative impact on the revenue of the semiconductor companies. Moreover, fabrication of devices for evaluating their performance would naturally mean that the fabrication setup must be optimized. This is a very cumbersome and tedious task involving a variation of multiple parameters through trial and error and based on earlier experiences. In addition, after fabrication of devices, we can only observe limited quantities thus restricting the insight that we can achieve in the device performance. Hence we may not be able to isolate the real mechanisms that cause the degradation in the performance.

This is exactly where the modeling and simulations of transistors come into the picture. Modeling and simulations enable us to study the device behavior using physics and maths. The main advantage of the simulations is that they allow us to isolate different mechanisms and study their impact on the system which is not possible in experiments. The device simulations provide us with an opportunity to evaluate different architectures and materials without going through lots of very expensive fabrication processes and without the wastage of precious natural resources. It also allows us to 'characterize' the device in great depths by allowing us to visualize different variables like potential profile and electron density, which can provide us with unique insights into the device behavior thus allowing us to fine-tune the fabrication process and improve the device performance. Thus the simulations have become an integral part of the cycle for development of new technology nodes for many technology generations.

The second important advantage of using simulations is that they have a predictive power which can be used in analyzing possible materials and architectures for future transistors. In other words, the simulation tools essentially use the computers made from existing technology and analyze different options for future devices so as to select the best possible device configuration. For the simulator to have predictive power, first of all, the number of fitting parameters must be kept to a minimum. The reason for this is that scaling has brought to the forefront effects that were considered to have a negligible impact on the device performance in the devices of yesteryears. An example of such an effect is that in earlier transistors the electron density profile was considered to have a maximum at the oxide-semiconductor interface, while actually it is slightly away (1-2 nm) from the interface. It is easy to see that such a simplification is justified when the gate oxide is much thicker than such offset, while it is completely incorrect to neglect the electron density profile when the gate oxide thickness is comparable to the offset. Thus in order to be able to analyze transistors for future technology nodes that constantly bring to forefront different physical phenomena, the simulators must improve by enriching their physics content rather than adding new fitting parameters to maintain the predictive power of the simulators. But at the same time, we must also bear in mind that including physics related to phenomena that do not significantly impact the performance of the devices would increase the computation time with no improvement in accuracy. It must always be borne in mind that the simulations are in any case approximations to the real fabricated devices.

Simulation is essentially solving a set of (discretized) mathematical equations over a certain region, which has been discretized using standard meshing techniques. These mathematical equations capture the operation of the device. The computational load and accuracy of the simulations vary depending upon the equations to be solved. A number of different methodologies have been developed and applied in analyzing the devices to capture the transport in the transistors. These methodologies range from simple particle conservation based methodology (Drift diffusion formulation) to full quantum transport formulations like Non Equilibrium Green's function (NEGF) or Wigner equation, which can account for the wave nature of the electron. Different transport formulations that can be used to model the carrier transport through a transistor are shown in Fig.1.1.

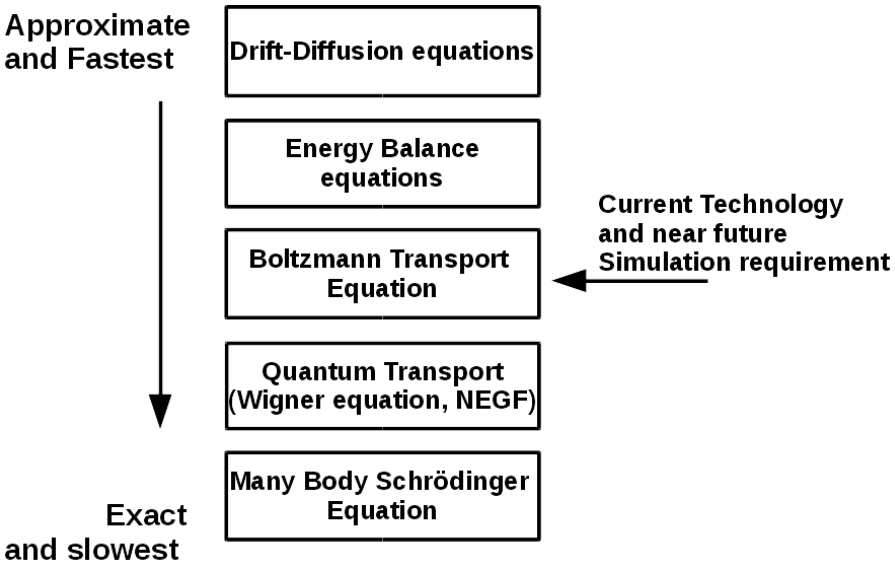


FIGURE 1.1: Illustration of different methodologies to simulate the transport a transistor. It also shows the optimum transport formulation for simulating the present and near future transistors.

The different transport formulations can be classified into mainly two categories: semi-classical and quantum transport. We believe that the Boltzmann Transport Equation (BTE) is a very effective formulation for modeling the electron transport in a nanoscale transistor provided the impact on the device performance of quantum effects along the transport (e.g. tunneling) are small. For a 3D electron gas in a bulk semiconductor the BTE is a 7-dimensional partial differential equation in the phase space. Straightforward usage of the BTE would, of course, be computationally prohibitive and, moreover, would result in lots of information that is not needed in modeling of long channel devices where low field parameters like mobility can be very accurate in benchmarking the performance of the devices. Thus to model the transport the TCAD simulators used only two moments of the Boltzmann Transport equation, which results in the drift-diffusion equations. Drift-Diffusion is the simplest and fastest of all the methods and it simply enforces the carrier continuity in the real space. As the channel length of the devices were constantly scaled to improve the performance, the transport became increasingly non-local and brought velocity overshoot to the forefront. The drift-diffusion model was able to reproduce the terminal characteristics of the devices but needed unphysical parameters [7]. This indicated that the physical effects involved in the device operation were beyond the simple carrier continuity and drift-diffusion was no longer valid. Thus the drift-diffusion approach became questionable. Notwithstanding these considerations, the drift-diffusion is still used in many simulation analysis by calibrating the parameters to either experiments or results from a more accurate simulation methodology. The reason for the use of the drift-diffusion till today in benchmarking of the performance analysis lies in its simplicity. The drift-diffusion formulation needs lower computational resources and the simulation time is much lower than other methods. Thus in the industry, where time is of essence, drift-diffusion method is still used.

In order to retain a computational burden comparable to the drift-diffusion model and also improve upon its predictive power, a number of extensions to the drift-diffusion were

proposed (like the addition of carrier temperature based on local energy balance) and the number of moments of the Boltzmann Transport equation was also increased [7, 8]. However, with all the improvements in the model a number of critical physical issues like closure to the moments, spurious velocity overshoot persisted [7]. In addition to the physical issues, inclusion of the higher moments in the simulations suffered from numerical issues like having a stable set of discretization and boundary condition as well. Also, the convergence with the addition of higher moments was not as good as with drift-diffusion [7]. And with continued scaling of the transistors, the failure of the moment based method became obvious and the researchers focused their attention to more accurate transport formalisms.

The benefit of the scaling to the modeling and simulations came in the form of improvement in the computational facilities. This allowed researchers to include more detailed physics-based models in the simulations. With the failure of the moment based methods, lot of effort was invested in the development and improvement of more sophisticated transport methodologies [9, 10, 11]. Boltzmann Transport Equation is a semi-classical method that describes the evolution of a carrier ensemble in the phase space. This is done using a Newtonian mechanics and carrier scatterings are treated quantum mechanically through the Fermi's Golden rule, so the overall approach is labeled as semi-classical. BTE has been solved primarily with the Monte Carlo method due to the relatively lower memory requirements but off late even deterministic methods are becoming increasingly popular, because they overcome the inherent statistical noise in the Monte Carlo method and because the availability of better computers have made deterministic methods computational affordable. [12]. Moreover since the modern day devices are confined in at least one direction, this also reduces the dimensionality of the BTE which goes a long way in reducing the computational load.

The scaling of the devices has reduced the channel length so significantly that it is literally possible to count the atoms in the channel. This makes one wonder if the electrons can really be treated as particles or if there is a need to employ quantum transport formulations. Quantum transport methods must be employed especially for the cases where the quantum effects are dominant like in the case of Tunnel FETs. Among the different quantum transport methods, Non-equilibrium Green's function (NEGF) is the most popular quantum transport formalism and it is based on the calculation of the Green's function of the Schrödinger equation from which all the quantities of interest like electron concentration, current etc can be calculated. A thorough review of this method can be found in [13, 14]. Another quantum transport formulation is based on the Wigner function which represents a quasi-probability distribution function (as it can be negative) [15]. Even though the Wigner equation based transport formalism has existed for a very long time its application to semiconductor devices has not received a lot of attention because deterministic simulations would be computationally very expensive and a stable Monte Carlo method was not possible due to the negative sign problem, as discussed in detail in [16]. The main attraction of using Wigner equation based transport formalism is its similarity with the Boltzmann Transport Equation. This allows us to include the different scattering mechanisms through the scattering operator as in the Boltzmann Transport Equation [17]. However, the application of the quantum transport

formalism has been limited because it requires large computational resources.

During my PhD program, we have used Boltzmann Transport Equation to model the electron transport in nanoscale FETs. We have selected a semi-classical methodology rather than full-quantum formulation and consistently limited the analysis to a channel length down to 14 nm. At such channel lengths, the quantum effects along the transport can be neglected as shown by a recent study benchmarking the performance of InAs and strained silicon-based nanowire FETs [18]. In this study, it was demonstrated that for InAs and strained silicon-based FETs the source to drain tunneling is negligible. The result is significant as InAs has a very low effective mass (m=0.023m_0) and energy gap, and even with such a low effective mass the source to drain tunneling contributes less than 10% to the off-current. Another crucial information that we can retrieve from the literature is that even at small channel lengths the scattering tends to dominate the transport resulting in small ballistic ratios, defined here as the ratio of the on-current with scattering to on-current without scattering [19]. Thus the transport methodology must be able to take into account different scattering mechanisms with ease. Another important practical consideration while simulating the devices is that the computational burden must be reasonable. Boltzmann Transport equation seems a reasonable trade-off even in this latter respect.

1.2.1 Importance of oxide-semiconductor interface

The MOSFETs have been the backbone of the semiconductor industry. Their operation relies primarily on the conduction of electrons from source to drain and the conduction takes place along the interface between the semiconductor and oxide. Thus it is obvious to see that this interface would be of paramount importance in working of the transistor. The scaling of the MOSFET since it was first demonstrated by Atalla and Kahng has brought two principal changes: namely the use of more than one gate, and the fact that the interface along which the conduction of electron takes place is today frequently formed by materials other than silicon and silicon dioxide. Use of multiple gates is essential to keep the detrimental short channel effects at bay. The use of high-κ gate oxide was needed to increase the gate capacitance without reducing the physical thickness of the gate oxide. Thus because of the first measure the number of the interfaces along which the conduction of electron takes place increases, while the second measure results in the interface between materials other than Si-SiO_2. Thus both the cases point out that the importance of the oxide-semiconductor interface would increase further.

The oxide-semiconductor interface being the most important region of the FET, even small imperfections can have a sizable impact on the performance of the devices. The interface affects the performance either due to the formation of trap states or due to roughness at the surface. The atoms at the interface can have unsaturated bonds which lead to formation of trap states. The impact of trap states on the performance of the modern day devices is studied in [20] and [21]. The interface is not smooth at the atomic scale and the impact of this roughness is the main topic of study in this thesis. These small variations break the periodicity of the crystal and thus result in a random potential which can lead to scattering

of electrons as they travel from source to drain. The surface roughness is a random process and the scattering due to the roughness is proportional to the nonlinear transformation of the surface roughness. Traditionally in the modeling of surface roughness scattering a number of approximations are made which convert the nonlinear transformation to linear transformation of a random process. Models based on these approximations result in overestimation of the surface roughness necessary to reproduce experiments, such as mobility measurements. A new model was proposed by our group that retained the nonlinear transformation and was able to reproduce the experimental mobilities [22]. But in that study the applicability of the model still had several limitations, like the use of scalar screening, and the formulation for planar structures only. It was necessary to overcome them to increase the range of applicability of the model.

1.2.2 Importance of electronic band structure

Electronic band structure (or simply band structure) is perhaps one of the most important relationship for any material as it enables us to extract a wide spectrum of data, starting from whether a material is a semiconductor or not, based on its energy gap. Among many parameters that can be calculated from the band structure, we point out that simply taking a derivative of the relationship with respect to wavevector gives us the velocity, which is one of the two necessary terms to calculate the current (other being the electron density). Thus by analyzing the band structure of different materials, we can look for the ones that give us higher velocity.

The dispersion relationship is not only necessary to describe the dynamics of a carrier transport but also vital from the electrostatics calculations point of view. From the band structure we can calculate the density of states, which indicates the maximum number of carriers that there can be hosted in the material per unit volume per unit energy. Generally it has been observed that materials that have high density of states have lower group velocity and vice versa. Hence simply opting for materials having large group velocity may not necessarily be the best option. Thus it is necessary to evaluate the band structure as accurately as possible.

Different methods have been developed to calculate the band structure that range from complete calculations from first principles (generally called as *ab-initio* methods) to a simple analytical description near the band edges (parabolic dispersion relations). Depending upon the method that is used, the computation time and load can vary over a wide range. In fact for most of the realistic devices the use of *ab-initio* methods would take prohibitively large time to simulate and simple parabolic relationship is incorrect. Thus a trade off is necessary and for this semi-empirical methods have been developed. In these class of methods, there are a number of parameters (with physical significance) which are calibrated such that they can reproduce the band structure data from experiments or from other more accurate methods.

1.3 Thesis organization

In the light of above discussion we set out to develop and improve physics based models related to the transport in nanoscale transistors and analyze different viable architectures for future CMOS technology nodes. This thesis is organized as follows:

In Chapter 2 we extend the formulation of the nonlinear model for surface roughness scattering to account for the tensorial screening and briefly discuss the Multi Subband Monte Carlo simulation framework, in which the nonlinear model was implemented [11, 23]. The extension of the nonlinear model to account for tensorial screening is necessary because the scalar screening is applicable only to single gate devices, while we have already discussed that the semiconductor industry has already adopted the multigate architecture [24]. The model had been originally developed in standalone form and the simulator was applicable in the circumstances where the surface roughness is dominant scattering mechanism. This limited the application of the nonlinear model to a very small set experimental data sets. Hence it was necessary to incorporate the nonlinear model in a comprehensive simulation framework to analyze a wide range of devices and architectures in conjunction with other relevant scattering mechanisms.

In Chapter 3 we first calibrate the surface roughness parameters in the nonlinear model using the bulk experimental mobility data sets for silicon and III/IV compounds. Here we reinforce the view that the nonlinear model is able to reproduce the experimental mobility sets for silicon with the experimentally measured values of the surface roughness parameters and it can also reproduce the mobility in III-V based FETs with more realistic values of the surface roughness parameters. We further analyze more complicated device architectures and demonstrate that, even for quantum well structures, the nonlinear model can reproduce the mobility data sets with measured values of the parameters related to surface morphology. We finally conclude this chapter by demonstrating that a buffer layer used to improve the mobility in heterostructure can be a double-edged sword.

In Chapter 4 we develop the nonlinear model for surface roughness scattering for 3D devices having a fairly arbitrary cross-section shape. This is necessary as we have already discussed in the previous section the migration to gate-all-around architecture is inevitable. We also validate various approximations made in the course of development of the model.

In Chapter 5 we develop a complete simulation framework using the existing Boltzmann Transport Equation and Schrödinger-Poisson solvers. First of all we analyze the experimental mobility data in gate-all-around and backgated device architecture and extract credible values of the surface roughness parameters. We then discuss the impact of cross-section shape, area, and material on the mobility and show that, even though different cross-section shape is relatively unimportant for silicon-based FETs, it has a significant impact for InAs based devices.

In Chapter 6 we discuss the extension of the simulator in the previous chapter to account for relevant effects like series resistance and self-heating, which are known to cause significant degradation in the performance of transistors. In order to be able to perform a comprehensive analysis, we have developed and implemented a number of methodologies

to streamline the simulator. We extend the simple mobility analysis to a thorough comparison of different analog and digital performance metrics for devices based on silicon or $In_{0.53}Ga_{0.47}As$ with different architectures. The analysis highlights the important fact that surface roughness and self-heating cause unequal amount of degradation in different architectures.

In Chapter 7 we discuss the tight binding formulation for both the bulk and its extension to nanowires and validate the implementation using the data available in the literature. Since the cross-section shapes of the practical structures do not have an analytical expression we have also developed a method by which we can interface the tight binding simulator with a general purpose mesh generator which is discussed as well.

In Chapter 8 we summarize important outcomes of the thesis, possible extensions and improvements to the present work. There is always room for improvement!

Chapter 2

Surface roughness scattering in planar transistors

2.1 Introduction

In order to achieve larger I_{on} and keep the detrimental effects of scaling of the physical dimensions under control, ultra-thin-body SOI and multigate FETs will be continued to be employed in future CMOS technology nodes and they may even be replaced with the gate-all-around nanowires [25]. In order to improve the gate control and also keep the tunneling current through gate oxide in-check, high-κ materials (like HfO_2, Al_2O_3, Zr_2O_3) are being used instead of SiO_2 [26]. To further improve the I_{on} silicon has been replaced and strained-silicon and other semiconductor materials (like Ge, InGaAs) that have larger mobility are being actively studied [27, 28, 29]. These modifications to the FETs, namely an interface between semiconductor and oxide other than familiar Si-SiO_2, and the employment of new device architectures having a larger number of surfaces as compared to bulk-planar devices suggest that surface related effects will continue to play an increasingly important role in future CMOS technologies. Thus the impact of surface morphology on the transport must be modeled with physical accuracy, that is with models that can reproduce the experimental mobility data with experimentally measured parameters related to surface structure (root mean squared values of surface roughness and correlation length, Λ) with AFM or TEM techniques

TABLE 2.1: The values of Δ_{rms} and Λ reported in experiments for (100) Si MOSFETs [22].

Δ_{rms} and Λ from AFM or TEM measurements for the Si-SiO_2 interface			
	Δ_{rms}[nm]	**Λ[nm]**	**Spectrum**
S.M.Goodnick *et al.* [30]	0.14-0.20	0.6-2.5	Gaussian
	0.14-0.20	0.7 - 3.7	Exponential
T.Yamanaka *et al.* [31]	0.21	—	—
A.Pirovano *et al.* [32]	0.18	1.8	Gaussian
O.Bonno *et al.* [33]	0.18	$\simeq$4.1	Exponential

Historically the surface roughness scattering in FETs has been modeled by Prange-Nee and generalized Prange-Nee model, here referred to as the linear model [34, 35]. The root mean square values of surface roughness (Δ_{rms}) extracted from simulations using linear model has been consistently larger than the experimentally measured values with TEM/AFM technique [30, 36, 37, 38, 39, 40, 41, 11, 35, 42, 43, 44, 45]. Also, the values of Δ_{rms} extracted from simulations for the case of III-V MOSFETs are extremely large (greater than 1 nm) [46, 47]. These discrepancies provided the motivation for the development of the nonlinear model [22] for modeling of surface roughness scattering. The nonlinear model in the momentum relaxation time framework along with scalar screening provided a very good match with the universal mobility curves for silicon with the value of Δ_{rms} that is consistent with the experimentally measured values at low temperature [22].

In this chapter, we will briefly review the linear surface roughness scattering models followed by a discussion on the nonlinear surface roughness scattering model. We will also discuss the extension of the nonlinear model to account for cross-correlation terms which will be necessary to calculate the tensorial screening. We will then describe the screening formulations and their applications to the nonlinear model. Then we introduce the Multi-subband Montecarlo (MSMC) framework in which the nonlinear model was implemented and the necessity to implement a parallelization scheme. Finally, we validate the implementation of the nonlinear model in the MSMC framework.

2.2 Linear and nonlinear model for surface roughness scattering

An accurate calculation of the matrix elements is extremely important in the semiclassical framework as they are then used to compute the scattering rates via the first-order Born approximation (which is equivalent to the Fermi's Golden Rule). According to the Fermi's Golden Rule, the scattering rates, $S_{n,n'}(\mathbf{k},\mathbf{k}')$ produced by the static scattering potential, $U_{sc}(\mathbf{R})$, and corresponding to the transition from a state with wavevector $\mathbf{k}$ in the n^{th} subband to $\mathbf{k}'$ in the m^{th} subband is given by [36]

$$S_{n,n'}(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}|\mathcal{M}_{nn'}(\mathbf{k},\mathbf{k}')|^2\delta(E_n(\mathbf{k}) - E_n(\mathbf{k}')) \tag{2.1}$$

where $\mathcal{M}_{nn'}(\mathbf{k},\mathbf{k}')$ is the matrix element is given by

$$\mathcal{M}_{nn'}(\mathbf{k},\mathbf{k}') = \int_{\Omega} \Phi^{\dagger}_{n\mathbf{k}}(\mathbf{R})U_{sc}(\mathbf{R})\Phi_{n'\mathbf{k}'}(\mathbf{R})d\mathbf{R} \tag{2.2}$$

where Ω is the normalization volume and $\Phi_{n\mathbf{k}}(\mathbf{R})$ is the wavefunction corresponding to the unperturbed system.

To proceed forward we make an ansatz related to the form of the envelope wavefuction. We assume that the wavefunction can be written as plane wave ($e^{i\mathbf{k}\cdot\mathbf{r}}$) in the transport plane ($\mathbf{r}\equiv(x,y)$) and $\xi_n(z)$ in the confinement direction, thus the $\Phi_{n\mathbf{k}'}(\mathbf{R})$ is given by

$$\Phi_{n\mathbf{k}}(\mathbf{R}) = \frac{1}{\sqrt{A}}\xi_n(z)e^{i\mathbf{k}\cdot\mathbf{r}} \tag{2.3}$$

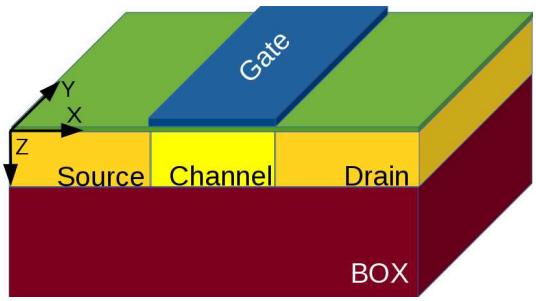


FIGURE 2.1: Sketch of a planar device indicating the coordinate system.

where A is the normalization area and $\xi_n(z)$ can be calculated by solving the Schrödinger equation in the cross-section

$$\left[\frac{-\hbar^2}{2}\frac{\partial}{\partial z}\frac{1}{m_z}\frac{\partial}{\partial z}+U(z)\right]\xi_n(z)=\varepsilon_n\xi_n(z) \tag{2.4}$$

where m_z is the effective mass in the confinement direction, $U(z)$ is the total potential energy and ε_n is the parabolic eigen energy.

In the particular case of surface roughness scattering in a planar device sketched in Fig.2.1, the $\mathcal{M}_{nn'}(\mathbf{k},\mathbf{k}')$ can be written

$$\mathcal{M}_{nn'}(\mathbf{q})=\frac{1}{A}\int_A M_{nn'}[\Delta(\mathbf{r})]e^{-i\mathbf{q}\cdot\mathbf{r}}d\mathbf{r} \tag{2.5}$$

where $\mathbf{q}=\mathbf{k}'-\mathbf{k}$, is the wave-vector modification produced by the scattering, A is the normalization area and $M_{nn'}[\Delta(\mathbf{r})]$ is defined

$$M_{nn'}[\Delta(\mathbf{r})]=\int_z \xi_n(z)(\hat{H}_{p,\mathbf{r}}-\hat{H}_0)\xi_{n'}(z)dz \tag{2.6}$$

where $\hat{H}_0$ and $\hat{H}_{p,\mathbf{r}}$ are the unperturbed and perturbed Hamiltonians. It is important to note that this separation is not rigorously correct unless the confinement direction coincides with the principal ellipsoidal axes (i.e. there is no coupling between the 'z' and **r** directions in the inverse effective mass tensor in the device coordinate system) [48]. Also, the additional term due to different electron masses in the oxide and semiconductor has been neglected as compared to the potential energy barrier [22]. The Hamiltonian under the parabolic effective mass approximation can be written as [36]

$$\hat{H}_{p,\mathbf{r}z}=\frac{-\hbar^2}{2}\frac{\partial}{\partial z}\frac{1}{m_z}\frac{\partial}{\partial z}+U(z) \tag{2.7}$$

The m_z for the ultra thin body system is given by

$$m_z=m_{ox,T}\Theta(-z+\Delta_F(\mathbf{r}))+m_{ox,B}\Theta(z-T_w-\Delta_B(\mathbf{r}))+m_{sct}[\Theta(z-\Delta_F(\mathbf{r}))-\Theta(-z+T_w+\Delta_B(\mathbf{r})] \tag{2.8}$$

where $m_{ox,T}$, $m_{ox,B}$ and m_{sct} are the electron's effective mass in the top and bottom oxide and semiconductor region, $\Theta(x)$ is the Heaviside step function, T_w is the thickness of

the semiconductor film and $\Delta_{F/B}(\mathbf{r})$ is the random process describing the front and back oxide-semiconductor interface in x-y plane. The total potential energy is the sum of the electrostatic potential energy calculated from the Poisson equation and the difference in the electron affinities between the semiconductor and oxide and it can be written as

$$U(z) = -e\phi(z) + \Phi_B[\Theta(-z + \Delta_F(\mathbf{r})) + \Theta(z - T_w - \Delta_B(\mathbf{r}))] \tag{2.9}$$

where $\phi(z)$ is the electrostatic potential, e is the elementary electron charge and Φ_B is the potential barrier between the semiconductor and oxide. The unperturbed Hamiltonian, $\hat{H}_{0,z}$, described by definition the case where there is no surface roughness, i.e. $\Delta_{F/B}(\mathbf{r})$ = 0.

2.2.1 Linear model

Prange-Nee and generalized Prange-Nee models have been extensively used in the literature to model the surface roughness scattering since they were first published [34, 40, 42, 49, 50, 51]. In order to calculate the perturbed Hamiltonian (assuming that $\Delta_B(\mathbf{r}) = 0$) we introduce a coordinate transformation given by

$$z' = \frac{T_w}{T_w - \Delta_F(\mathbf{r})}(z - \Delta_F(\mathbf{r})) \tag{2.10}$$

Then by expressing the potential energy, $U(z)$, in terms of z' and linearzing for small values $\Delta(\mathbf{r})$ values and keeping terms only up to first order in $\Delta(\mathbf{r})$

$$\phi(z) = \phi\left(z' + \frac{T_w - z'}{T_w}\Delta(\mathbf{r})\right) = \phi(z') + \left(\Delta_F(\mathbf{r}) - z'\frac{\Delta_F(\mathbf{r})}{T_w}\right)\frac{d\phi(z')}{dz'} \tag{2.11}$$

With some straightforward but tedious algebraic manipulations we can write the perturbed Hamiltonian as [36]

$$\hat{H}_{p,\mathbf{r}z} = \hat{H}_{0,z'} + \frac{\Delta_F(\mathbf{r})}{T_w}\left(\hbar^2\frac{d}{dz'}\frac{1}{m_z}\frac{d}{dz'} - ez'\frac{d\phi(z')}{dz'}\right) - e\Delta_F(\mathbf{r})\frac{d\phi(z')}{dz'} \tag{2.12}$$

In order to evaluate the matrix element we linearize the wavefunction (similar to the potential energy) about the perturbed interface which results in

$$\xi_n(z) = \xi_n\left(z' + \frac{T_w - z'}{T_w}\Delta(\mathbf{r})\right) = \phi(z') + \left(\Delta_F(\mathbf{r}) - z'\frac{\Delta_F(\mathbf{r})}{T_w}\right)\frac{d\xi_n(z')}{dz'} \tag{2.13}$$

By keeping only the first order terms in $\Delta_F(\mathbf{r})$ we arrive at [36]

$$\begin{aligned} M_{nn'}[\Delta_F(\mathbf{r})] = -\Delta_F(\mathbf{r})\Bigg[\int_0^{T_w}\xi_n(z)e\frac{d\phi}{dz}\xi_{n'}(z)dz - (\varepsilon_n - \varepsilon_{n'})\int_0^{T_w}\xi_n(z)\frac{d\xi_{n'}(z)}{dz}dz\Bigg] \\ - \frac{\Delta_F(\mathbf{r})}{T_w}\int_0^{T_w}\xi_n(z)\left(\frac{\hbar^2}{m_z}\frac{d^2}{dz^2} - ze\frac{d\phi}{dz}\right)\xi_{n'}(z)dz - \frac{\Delta_F(\mathbf{r})}{T_w}(\varepsilon_n - \varepsilon_{n'})\int_0^{T_w}\xi_n(z)z\frac{d\xi_{n'}(z)}{dz}dz \end{aligned} \tag{2.14}$$

When Φ_B is very large and in fact when it tends to infinity we can neglect the wavefunction penetration in the oxide and Eq.2.14 reduces to the the Prange-Nee model, where $M_{nn'}[\Delta_F(\mathbf{r})]$ is given by

$$M_{nn'}[\Delta_F(\mathbf{r})] = \Delta_F(\mathbf{r}) \left[\frac{\hbar^2}{2m_z} \frac{d\xi_n(0)}{dz} \frac{d\xi_{n'}(0)}{dz} \right] \tag{2.15}$$

The rationale behind the linearization of the $\xi_n(z)$ and $\phi(z)$ around the interface position is to be able to write the $M_{nn'}[\Delta_F(\mathbf{r})]$ as product of non-statistical part, $K^F_{nn'}$ (terms in the square bracket) and $\Delta_F(\mathbf{r})$ itself, which is exactly what has been achieved in Eq.2.14 and 2.15. Hence, these models are referred here as the linear models. This allows us to write the matrix elements as simply proportional to the Fourier transform of the $\Delta_F(\mathbf{r})$. Substituting Eq.2.14 (or Eq.2.15) in Eq.2.5

$$\mathcal{M}^F_{nn'}(q) = K^F_{nn'} \left[\frac{1}{A} \int_A \Delta_F(\mathbf{r}) e^{-i\mathbf{q}\cdot\mathbf{r}} d\mathbf{r} \right] \tag{2.16}$$

Thus the ensemble averaged squared matrix element is

$$\left\langle |\mathcal{M}^F_{nn'}(q)|^2 \right\rangle = |K^F_{nn'}|^2 \frac{S_F(\mathbf{q})}{A} \tag{2.17}$$

where the spectrum of the surface roughness is defined as

$$S_F(\mathbf{q}) = \frac{1}{A} \left| \int_A \Delta_F(\mathbf{r}) e^{-i\mathbf{q}\cdot\mathbf{r}} d\mathbf{r} \right|^2 \tag{2.18}$$

Since we have assumed that the random process is completely characterized by its first (mean) and second (correlation function) moments, we write the spectrum in terms of the autocorrelation function of the $\Delta_F(\mathbf{r})$ using the Wiener–Khinchin theorem, which states that power spectral density is the Fourier transform of the correlation function, as

$$S_F(\mathbf{q}) = \int_A C_F(\mathbf{r}) e^{-i\mathbf{q}\cdot\mathbf{r}} d\mathbf{r} \tag{2.19}$$

where $C_F(\mathbf{r})$ is the autocorrelation function of the surface roughness. In the literature, $C_F(\mathbf{r})$ is generally modeled with an either exponential [30] or Gaussian form [50] (with the exponential form being in better agreement with the experiments, which is as expected from theory of memoryless random process [52]). Assuming that $\Delta_F(\mathbf{r})$ and $\Delta_B(\mathbf{r})$ are uncorrelated then the ensemble averaged squared matrix elements at the different interfaces are simply additive and hence for a planar UTB-SOI we have an overall matrix element

$$\left\langle |\mathcal{M}_{nn'}(q)|^2 \right\rangle = \left\langle |\mathcal{M}^F_{nn'}(q)|^2 \right\rangle + \left\langle |\mathcal{M}^B_{nn'}(q)|^2 \right\rangle \tag{2.20}$$

2.2.2 Nonlinear model

The linear model depends primarily on the assumption that $\Delta(\mathbf{r})$ is small, that justifies the linear expansion of the wavefunction and electrostatic potential. Even though the $\Delta(\mathbf{r})$

may be small (though with the linear model Δ_{rms} can be even greater 1 nm [46]), change in wavefunction is significant around the interface and this makes the $M_{nn'}[\Delta(\mathbf{r})]$ extremely nonlinear [22]. Thus in order calculate $M_{nn'}[\Delta(\mathbf{r})]$ we must evaluate the integration in Eq.2.6 numerically. The contribution to the integral comes from the kinetic and potential energy terms. As discussed at length in [22] the contribution to the integral from the kinetic energy term is typically negligible for most of the silicon-oxide systems. However, for III-Voxide system it becomes instead unphysically large because the parabolic effective mass approximation is inadequate to calculate the kinetic energy contributions when m_{ox}/m_{sct} is large. Hence in this work the kinetic energy contributions are neglected. This allows us to write the Eq.2.6 at the front interface as

$$M_{nn'}[\Delta_F(\mathbf{r})] \approx \int_0^{\Delta_F(\mathbf{r})} \xi^{\dagger}_{n'}(z)\Phi_B\xi_n(z)dz \quad . \tag{2.21}$$

Eq.2.21 shows that the $M_{nn'}[\Delta_F(\mathbf{r})]$ is not simply proportional to the $\Delta_F(\mathbf{r})$, rather it is nonlinear transformation of the random process, $\Delta_F(\mathbf{r})$. Hence the ensemble averaged squared matrix elements is not simply proportional to the spectrum of the surface roughness, rather we must calculate the correlation function of $M_{nn'}[\Delta_F(\mathbf{r})]$. If we assume that the $\Delta_F(\mathbf{r})$ is a zero mean stationary random process and that it has Gaussian first and second distribution, then the autocorrelation function of $M_{nn'}[\Delta_F(\mathbf{r})]$ can be written as [53]

$$C^{nn'}_{nn'}(\mathbf{r}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} M_{nn'}(\Delta_{F1})M_{nn'}(\Delta_{F2})f(\Delta_{F1},\Delta_{F2},\mathbf{r})d\Delta_{F1}d\Delta_{F2} \tag{2.22}$$

where $f(\Delta_{F1},\Delta_{F2},\mathbf{r})$ is the joint probability distribution function which is taken to be gaussian and it is given by

$$f(\Delta_{F1},\Delta_{F2},\mathbf{r}) = \frac{1}{2\pi\, C_\Delta(0)\sqrt{1-C^2_{\Delta,N}(\mathbf{r})}}\exp\left[-\frac{\Delta^2_{F1}+\Delta^2_{F2}-2C_{\Delta,N}(\mathbf{r})\Delta_{F1}\Delta_{F2}}{2C_\Delta(0)(1-C^2_{\Delta,N}(\mathbf{r}))}\right] \tag{2.23}$$

where $C_\Delta(\mathbf{r})$ is the autocorrelation function of $\Delta(\mathbf{r})$ and $C_{\Delta,N}(\mathbf{r}) = C_\Delta(\mathbf{r})/C_\Delta(0)$. We now assume that the surface roughness has an exponential autocorrelation function

$$C_\Delta(r) = \Delta^2_{rms}e^{-r\sqrt{2}/\Lambda} \tag{2.24}$$

where Δ_{rms} is the root mean squared value and Λ is the correlation length (approximately the distance between 2 peaks). Then using the Wiener-Khinchin theorem, we can calculate the ensemble averaged squared matrix elements as the Fourier transform of the correlation function

$$\left\langle|\mathcal{M}^F_{nn'}(q)|^2\right\rangle = \frac{1}{A}\int_0^{+\infty} C^{nn'}_{nn'}(r)e^{-i\mathbf{q}\cdot\mathbf{r}}dr \tag{2.25}$$

Converting in to the polar coordinates ($\mathbf{r} \equiv (r, \theta)$ and $\mathbf{q} \equiv (q, \phi)$) as $C_\Delta(r)$ and consequently $C^{nn'}_{nn'}(\mathbf{r})$ depends only on $r = |\mathbf{r}|$, leads to

$$\begin{aligned} \left\langle |\mathcal{M}^F_{nn'}(q)|^2 \right\rangle &= \frac{1}{A} \int_0^{+\infty} r dr C^{nn'}_{nn'}(r) \int_{-\pi}^{\pi} e^{-iqrcos(\theta)} d\theta \\ &= \frac{2\pi}{A} \int_0^{+\infty} r dr C^{nn'}_{nn'}(r) J_0(qr) \end{aligned} \tag{2.26}$$

where $J_0(x)$ is the zeroth order Bessel function of the first kind. Note that, because of the radial symmetry of the correlation function Fourier transform is same as zeroth order Hankel transform [54]. The total ensemble averaged squared matrix element stemming from scattering at the front and back interface can be written as

$$\left\langle |\mathcal{M}_{nn'}(q)|^2 \right\rangle = \left\langle |\mathcal{M}^F_{nn'}(q)|^2 \right\rangle + \left\langle |\mathcal{M}^B_{nn'}(q)|^2 \right\rangle \tag{2.27}$$

In summary, we first calculate the $M_{nn'}[\Delta(\mathbf{r})]$ numerically, which is followed by the evaluation of the autocorrelation function using the Eq.2.22. Then using Eq.2.26 we can calculate the ensemble averaged squared matrix elements.

2.2.3 Additional comments

The contribution from different interfaces to the scattering rates in SOI structure are considered to be additive (i.e. the front and back interfaces are uncorrelated). This allows us to write the matrix element as a sum of the matrix element from individual surfaces. Thus the total ensemble averaged squared matrix elements can be written as a sum of the individual contributions from different interfaces. This is simple to see for linear models but for nonlinear model an important hint is to realize that the matrix elements calculated in Eq.2.21 are space invariant transformation of $\Delta(\mathbf{r})$ [22]. Further more to the best of our knowledge till day no experimental study has shown the correlations between different interfaces in FET. Validating this simplifying assumption can be a topic of future study.

The ensemble averaged squared matrix elements for linear and nonlinear models derived in the previous subsection considered only the impact of the asperities of the surface on the modulation of the subband energies through the perturbation in the Hamiltonian (modification to the potential energy term). However, a second-order impact of the surface roughness is the modifications of the electron concentration profile through the variation in the electrostatic potential, which has been neglected and can be a topic for future study. The contribution of the second term to the ensemble averaged squared matrix elements can be expected to be relatively unimportant because mobility studies neglecting it [40] and those accounting for it [49] arrive at similar values of Δ_{rms} and Λ.

2.3 Screening

2.3.1 Physical interpretation and dielectric function

Screening is a physical phenomenon of smearing out the scattering potential by mobile charge carriers in response to the scattering potential. Consider a device with a rough interface (sketched as shown in Fig.2.2), which also illustrates the direction of the gate electric field. In the presence of a strong gate field, electrons accumulate near the surface. These accumulated electrons would then create an electrostatic (screening) field counter to the gate field that would prevent the other electrons (screened) that are traveling from left to right from coming much closer to the interface. Thus the screened electrons would see a smaller scattering potential (in this example surface roughness). This reduction of the scattering potential will make it easy (as compared to the case where the screening field is not there) for the electrons to travel and hence improve the mobility. Thus neglecting screening would lead to overestimation of the scattering. An important aspect of screening is that it is expected to play a role only at larger gate biases (in the on-state of the device).

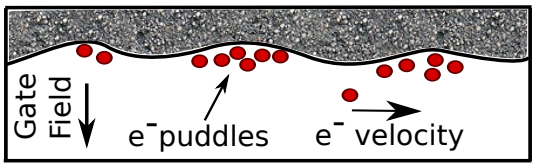


FIGURE 2.2: Sketch of the cross-section of a planar device illustrating the concept of screening.

To put things in a mathematical framework, it is necessary to calculate the dielectric constant (a scalar quantity for a 3D system) which scales the applied electric field (or scattering potential) inside the system. However, a confined system responds by reorienting along confinement (intersubband components) and free direction (intrasubband components) [52]. Thus instead of a scalar dielectric constant for a 3D system, we get a dielectric matrix. This problem was solved by Lindhard using the first-order perturbation theory and random phase approximation. The Lindhard dielectric function, $\varepsilon_{nn'}^{mm'}(\mathbf{q})$, is given by [36, 52, 55]

$$\varepsilon_{nn'}^{mm'}(\mathbf{q}) = \delta_{m,n}\delta_{m',n'} - \frac{e^2}{q(\varepsilon_{sct}+\varepsilon_{ox})}\Pi_{n,n'}(\mathbf{q})F_{mm'}^{nn'}(\mathbf{q}) \tag{2.28}$$

The second term is the contribution of the free charge carriers to the dielectric function. The polarization factor, $\Pi_{n,n'}(\mathbf{q})$, is given by

$$\Pi_{n,n'}(\mathbf{q}) = \frac{1}{A}\sum_{\mathbf{k}}\frac{f_{n'}(\mathbf{k}+\mathbf{q}) - f_n(\mathbf{k})}{E_{n'}(\mathbf{k}+\mathbf{q}) - E_n(\mathbf{k})} \tag{2.29}$$

and formfactor in the Eq.2.28 is given by

$$F_{mm'}^{nn'}(q) = \int dz \xi_m(z)\xi_{m'}^{\dagger}(z)\int dz_0 \xi_n(z_0)\xi_{n'}^{\dagger}(z_0)\phi_{pcN}(q,z,z_0) \tag{2.30}$$

where $\phi_{pcN}(q, z, z_0)$ is the Green's function and it satisfies

$$\left[\frac{\partial^2}{\partial z^2} + q^2\right] \phi_{pcN}(q, z, z_0) = -\frac{e}{\varepsilon}\delta(z - z_0) \tag{2.31}$$

Since the dielectric function, by definition, relates (scales) the total electric potential, screened, (sum of external and induced due to reorientation of the charge in the material) and external potential, unscreened, we can write the matrix elements due to scattering potential as

$$\mathcal{M}_{nn'}(q) = \sum_{mm'} \varepsilon_{nn'}^{mm'}(q)\mathcal{M}_{mm'}^{scr}(q) \tag{2.32}$$

In the long wavelength limit ($q \to 0$) and assuming a bulk MOS system, $\phi_{pcN}(q, z, z_0)$ is a constant for both bulk ($\phi_{pcN}(q \to 0, z, z_0) \approx 1$) and thin body devices ($\phi_{pcN}(q \to 0, z, z_0) \approx \frac{\epsilon_{sct}+\epsilon_{ox}}{2\epsilon_{ox}}$) [24]. By using the orthogonality of the wavefunctions, the form factor can be simplified as

$$F_{mm'}^{nn'}(q \to 0) \approx \delta_{mm'}\delta_{nn'} \tag{2.33}$$

Because of the Dirac delta function, the intersubband transitions are always unscreened. Further more the eq.2.33 allows us to write

$$F_{mm}^{nn}(q \to 0) \approx F_{mm}^{mm}(q \to 0) \tag{2.34}$$

Using eq.2.32 and 2.34 along with 2.28 results in a symmetric dielectric matrix which can be analytically inverted [36] and thus the screened matrix elements can be written as

$$\mathcal{M}_{nn}^{scr}(q) = \frac{\mathcal{M}_{nn}(q) + \alpha(q) + \sum\limits_{n\neq m} \Pi_{nn}(q)F_{mm}^{nn}(q)\left[\mathcal{M}_{mm}(q) - \mathcal{M}_{nn}(q)\right]}{\varepsilon_S} \tag{2.35}$$

We further assume that $\mathcal{M}_{nn'}(q) > [\mathcal{M}_{mm}(q) - \mathcal{M}_{nn}(q)]$ thus simplifying screening calculations to

$$\mathcal{M}_{nn}^{scr}(q) = \frac{\mathcal{M}_{nn}(q)}{\varepsilon_S} \tag{2.36}$$

where

$$\varepsilon_S(q) = det(\boldsymbol{\varepsilon}) \tag{2.37}$$

where $\varepsilon_S(q)$ is the scalar dielectric function.

2.4 Application to linear model

The application of screening to the linear model has been extensively discussed in the literature [36, 24, 42]. Here we briefly summarize the key steps. The matrix elements in the linear model for surface roughness scattering formulation can be written as

$$\mathcal{M}_{nn'}(q) = K_{nn'}\frac{\Delta(\mathbf{q})}{A} \tag{2.38}$$

Using eq.2.38 and 2.32 we can write

$$\mathcal{M}^{scr}_{nn'}(q) = K^{scr}_{nn'}\frac{\Delta(\mathbf{q})}{A} \tag{2.39}$$

where $K^{scr}_{nn'}$ is given by

$$K^{scr}_{nn'} = \sum_{mm'}\left(\varepsilon^{nn'}_{mm'}\right)^{-1} K_{mm'} \tag{2.40}$$

Squaring and taking ensemble average we get

$$\langle|\mathcal{M}^{scr}_{nn'}(q)|\rangle^2 = |K^{scr}_{nn'}|^2\frac{S_F(\mathbf{q})}{A} \tag{2.41}$$

The application of the scalar screening is simply as written in eq.2.36

2.5 Application to nonlinear model

The procedure to apply tensorial screening to the nonlinear model is more involved as compared to the linear model because the matrix elements, $\mathcal{M}_{nn'}(q)$, cannot be simply expressed as a product of the non-statistical part and $\Delta(\mathbf{r})$ and it is necessary that we calculate the correlation function of the $M_{nn'}[\Delta(\mathbf{r})]$ which directly leads to the ensemble averaged squared matrix elements as discussed in the earlier sections. To this end, we discuss the formulation in detail below.

If we denote ε^{-1} by $\mathbf{L}$ then Eq.2.32 can be formally solved for the screened matrix elements as

$$\mathcal{M}^{scr}_{nn'}(q) = \sum_{mm'} L^{mm'}_{nn'}(q)\mathcal{M}_{mm'}(q) \tag{2.42}$$

Multiplying eq.2.42 with it complex conjugate and then taking ensemble average we get

$$\left\langle|\mathcal{M}^{scr}_{nn'}(q)|^2\right\rangle = \sum_{\substack{(mm')\\(pp')}} L^{mm'}_{nn'}(q)L^{pp'}_{nn'}(q)^{\dagger}\left\langle\mathcal{M}_{mm'}(q)\mathcal{M}^{\dagger}_{pp'}(q)\right\rangle \tag{2.43}$$

where $\left\langle\mathcal{M}_{mm'}(q)\mathcal{M}^{\dagger}_{pp'}(q)\right\rangle$ is the spectrum of the cross-correlation between the unscreened matrix elements $M_{mm'}[\Delta(r)]$ and $M_{pp'}[\Delta(r)]$.

From an implementation point of view, it is very beneficial to be able to write the above equation in the matrix form. To this end we recall that the relationship between the screened and unscreened matrix elements in the matrix form can be written as

$$\boldsymbol{\mathcal{M}} = \boldsymbol{\varepsilon}\boldsymbol{\mathcal{M}}_{scr} \tag{2.44}$$

where $\boldsymbol{\varepsilon}$ is the dielectric matrix, $\mathcal{M}$ and $\mathcal{M}_{scr}$ are the row vectors of unscreened and screened matrix elements respectively (indexed by the subband pairs) and $\boldsymbol{\varepsilon}$ is a dielectric matrix. Eq.2.44 can be rewritten by noting the dielectric matrix is a non-singular matrix

$$\boldsymbol{\mathcal{M}}_{scr} = \boldsymbol{\varepsilon}^{-1}\boldsymbol{\mathcal{M}} \tag{2.45}$$

Taking hermitian conjugate of the Eq.2.45, post-multiplying it with Eq.2.45 and taking ensemble average we get

$$\langle \boldsymbol{\mathcal{M}}_{scr} \boldsymbol{\mathcal{M}}_{scr}^{\dagger} \rangle = \boldsymbol{\varepsilon}^{-1} \langle \boldsymbol{\mathcal{M}} \boldsymbol{\mathcal{M}}^{\dagger} \rangle \left(\boldsymbol{\varepsilon}^{-1} \right)^{\dagger} \tag{2.46}$$

The ensemble averaging process filters out the dielectric matrices because they don't have statistical dependence. The $\langle \boldsymbol{\mathcal{M}}_{scr} \boldsymbol{\mathcal{M}}_{scr}^{\dagger} \rangle$ is the full matrix of all the possible combinations of the screened matrix elements between different subbands and valleys but only diagonal elements have physical consequence.

Returning back to the discussion, by definition, the cross-correlation between 2 random process, $X(t)$, $Y(t)$ is given by [56]

$$C_X^Y = \boldsymbol{E}(X(t_1), Y(t_2)) \tag{2.47}$$

where $\boldsymbol{E}(...)$ is the expectation operator and, if we know the joint probability distribution function, then we can write it as

$$C_X^Y = \int X(t_1), Y(t_2) f(t_1, t_2) dt_1 dt_2 \tag{2.48}$$

where $f(t_1, t_2)$ is the joint probability distribution function. These definitions suggest that the autocorrelation function is a special case of the cross-correlation where the random processes are the same. This observation allows us to extend the definition of the autocorrelation function 2.22 to calculate the cross-correlation function as

$$C_{mm'}^{pp'}(\mathbf{r}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} M_{mm'}(\Delta_1) M_{pp'}(\Delta_2) f(\Delta_1, \Delta_2, \mathbf{r}) d\Delta_1 d\Delta_2 \tag{2.49}$$

where $f(\Delta_1, \Delta_2, \mathbf{r})$ is the joint probability distribution function which is same as stated in Eq.2.23.

For the scalar screening the screened matrix elements can be simply written just like in the case of linear model as

$$\left\langle |\mathcal{M}_{nn}^{scr}(q)|^2 \right\rangle = \frac{\left\langle |\mathcal{M}_{nn}(q)|^2 \right\rangle}{\varepsilon_S^2(q)} \tag{2.50}$$

where $\varepsilon_S(q)$ is the scalar dielectric function defined in Eq.2.37.

2.6 Multisubband Monte Carlo simulation framework

The Multi Subband Monte Carlo (MSMC) simulator [11] has been a developed over last many years in the group. Over the years it has been refined by including many relevant scattering mechanisms and different physical phenomena like screening [24] and traps [20, 21]. In addition, the computation has also been streamlined by using parallelization of scattering matrix elements [23] and use of sophisticated discretization method (pseudospectral method) to reduce the computation time needed for Schrödinger solver [57]. The simulator has been benchmarked across other independent simulators [58]. These features allow

us to analyze a wide spectrum of experimental data for planar devices across a range of temperatures.

2.6.1 Basics of Monte Carlo

This section will briefly overview important concepts in the Monte Carlo method for solving the Boltzmann Transport Equation; interested readers may refer to [36, 59, 60] for details. Monte Carlo technique has been used over last many decades to solve the Boltzmann Transport equation (BTE) by many authors [38, 11, 12, 61], to compute the occupation function, $f_n(\mathbf{R}, \mathbf{K})$ and simulate the devices which are driven far from equilibrium conditions. For solving the BTE, the motion of electrons (considered as particles) is simulated for a random amount of time (free flight time), and then instantaneously terminated by a randomly selected scattering event. The free flight duration and scattering events are selected according to the probability of the scattering mechanisms considered in the simulations [59]. After the scattering has taken place a suitable final state (energy and momentum) is selected according to the scattering mechanism with which the free flight was terminated. This procedure, when repeated a number of times (law of large numbers), allows us to estimate different physical quantities of interest like average velocity, distribution function etc. At the heart of the Monte Carlo simulator there are two steps: namely calculation of the free flight duration and identifying the scattering mechanism and state after scattering. The details of these two steps are summarized below.

Step1: Calculation of free flight duration

In order to calculate the free flight duration we let, $S_{tot}(\mathbf{K})$, denote the total scattering rate out of the state **K** given by

$$S_{tot}(\mathbf{K}) = \sum_{m} \sum_{\mathbf{K}'} S_m(\mathbf{K}, \mathbf{K}') \tag{2.51}$$

where we have summed over the possible scattering mechanisms, m, and over all the final states, **K'**. Also, by definition, $S_{tot}(\mathbf{K}(t))dt$ is nothing but the probability that an electron will scatter in time interval 'dt'. This allows us to write the probability that an electron hasn't scattered over free flight duration, t_{FF}, as $exp(-\int_0^{t_{FF}} S_{tot}K(t')dt')$. Thus the probability that an electron will scatter in time 'dt' after 't_{FF}' is given by

$$P(t_{FF}) = exp\left(-\int_0^{t_{FF}} S_{tot}K(t')dt'\right) \times S_{tot}(K(t_{FF}) = r \tag{2.52}$$

where 'r' is a random number having a uniform distribution function between 0 and 1.

It is extremely difficult to solve the above integral equation for t_{FF} and the simplest method is to use the concept of self scattering [36, 59]. This results in t_{FF}, as

$$t_{FF} = -\frac{ln(r)}{\Gamma} \tag{2.53}$$

where Γ is upper bound of the total scattering rate. At the end of t_{FF}, it is necessary to update the position of the particle in the phase space. This is done by using laws of classical mechanics adapted to Bloch waves (by defining momentum as $\hbar\mathbf{K}$), that is by solving equations

$$\frac{d(\hbar\mathbf{K})}{dt} = -\nabla_{\mathbf{R}} E(\mathbf{R}, \mathbf{K}) \tag{2.54}$$

$$\frac{d(\mathbf{R})}{dt} = \frac{1}{\hbar}\nabla_{\mathbf{K}} E(\mathbf{R}, \mathbf{K}) \tag{2.55}$$

where E is the total energy of the electron. Eq.2.55 is the definition of group velocity of the wave-packet. The trajectory of electrons in the phase space can be tracked with the Eq.2.54 and 2.55.

Step2: Identification of scattering mechanism and final state

After the termination of the free flight, the next step is to determine the scattering mechanism that posed to end the free flight. This is obtained by calculating the relative contribution of different scattering mechanisms as

$$P_m(\mathbf{K}) = \frac{1}{S_{tot}(\mathbf{K})} \sum_{\mathbf{K}'} S_m(\mathbf{K}, \mathbf{K}') \tag{2.56}$$

where $S_{tot}(\mathbf{K})$ is the total scattering rate, $S_m(\mathbf{K}, \mathbf{K}')$ is the scattering rate from $\mathbf{K}$ to $\mathbf{K}'$ due to m^{th} scattering mechanism and $P_m(\mathbf{K})$ is by definition the probability of the m^{th} mechanism interrupting the free flight. Then by generating a random number (uniformly distributed between 0 and 1), '$r1$', we can select the scattering mechanism j as

$$\sum_{i=1}^{j-1} P_i(\mathbf{K}) \leq r1 < \sum_{i=1}^{j} P_i(\mathbf{K}) \tag{2.57}$$

The probability that the electron scatters from $\mathbf{K}$ to $\mathbf{K}'$ due to j^{th} scattering mechanism, $P_j(\mathbf{K}, \mathbf{K}')$, is given by

$$P_j(\mathbf{K}, \mathbf{K}') = \frac{S_j(\mathbf{K}, \mathbf{K}')}{\sum_{\mathbf{K}'} S_j(\mathbf{K}, \mathbf{K}')} \tag{2.58}$$

Then by generating random number, $r2$, we can select the final state, 'n', as

$$\sum_{i=1}^{n-1} P_j(\mathbf{K}, \mathbf{K}') \leq r2 < \sum_{i=1}^{n} P_j(\mathbf{K}) \tag{2.59}$$

The degeneracy of the electron gas can play a vital role in carrier transport especially in III-V materials which have low density of states because degeneracy effectively forbids the transition to the final state if that is occupied, consistently with Pauli's exclusion principle. MSMC accounts for carrier degeneracy using the methodology discussed in [62]. To account for the Pauli's exclusion principle a random number (uniformly distributed between 0 and 1), r3, is generated. The carrier scattering to the final state is then rejected if the occupation function of the final state, f(**R**,**K**) is greater than r3 [36].

2.6.2 Additional scattering mechanisms

In order to analyze a wide range of devices comprising of heterostructures and different materials, we have used a number of different scattering mechanisms in the MSMC simulations. The scattering rates (from (n,**k**) to (n′,**k**′)) that are needed in the Monte Carlo simulations are calculated from the squared matrix elements for different scattering mechanisms (Eqs.2.61, 2.63, 2.65, 2.67, 2.68) according to the Fermi's Golden rule

$$S_{n,n'}(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}|\mathcal{M}_{nn'}(\mathbf{k},\mathbf{k}')|^2\delta(E_n(\mathbf{k}) - E_{n'}(\mathbf{k}') \pm \hbar\omega) \tag{2.60}$$

where '+' corresponds to increase in the energy of the final state (for example due to absorption of phonon) and '$-$' corresponds to reduction in the energy of the final state (for example due to emission of phonon) and for elastic processes (for example surface roughness scattering) there is no net exchange of energy ($\hbar\omega$=0 eV).

Here we will briefly discuss the most matrix element calculations of the relevant ones for this work in addition to the surface roughness scattering discussed in detail in section 2.2.

Acoustic and optical non-polar phonon scattering

The non-zero temperature of the crystal lattice leads to vibration of the atoms about their mean positions. These vibrations of the atoms breaks the periodicity of the lattice potential and thus cause electron scattering. From the dispersion relationship of the phonons they can be classified into acoustic (vibrations of the atom that are in-phase with each other in the long wavelength limit) and optical phonons (vibrations of the atom that are out-of-phase with each other in the long wavelength limit). The acoustic phonons are modeled in the elastic approximation, which is generally a good approximation at room and higher temperatures. The unscreened matrix elements for absorption and emission process are given by [36]

$$|M_{nn'}^{(ab)/(em)}(\mathbf{k},\mathbf{k}')|^2 = \delta_{\mathbf{k}',\mathbf{k}\pm\mathbf{q}}\frac{K_B T D_{ac}^2}{2\rho A v_s^2}F_{nn'} \tag{2.61}$$

where '+' corresponds to absorption and '$-$' corresponds to emission of a phonon, D_{ac} is the acoustic deformation potential, v_s is the sound velocity in the crystal and the form factor, $F_{nn'}$, is given by

$$F_{nn'} = \int_z |\xi_n(z)|^2|\xi_n'(z)|^2 dz. \tag{2.62}$$

Optical phonons in MSMC are assumed to be dispersionless and and thus have a a constant energy ($\hbar\omega_0$). The matrix element (for absorption,'$-$', and emission process,'+') for the optical phonon is given by [36]

$$|M_{nn'}^{(em)/(ab)}(\mathbf{k},\mathbf{k}')|^2 = \delta_{\mathbf{k}',\mathbf{k}+\mathbf{q}}\frac{\hbar D_{op}^2}{2\omega_0\rho A}F_{nn'}\left(n_{op} + \frac{1}{2} \pm \frac{1}{2}\right) \tag{2.63}$$

where D_{op} is the optical deformation potential, n_{op} are the number of the phonons corresponding to energy of $\hbar\omega_0$ and can be calculated using the Bose-Einstein statistics as

$$n_{op} = \frac{1}{exp(\frac{\hbar\omega_0}{K_B T}) - 1} \tag{2.64}$$

In the MSMC we consider bulk phonon modes and any impact of the quantization on the phonons is neglected.

Polar optical phonon scattering

In polar semiconductors, the bonds between the neighboring atoms have an ionic tendency (because of the different electronegativity of the atoms) in addition to covalent nature. This ionic nature of the bond leads to the formation of an electric dipole. When these dipoles oscillate because of the finite temperature, they give rise to a scattering potential. This scattering potential due to vibration of the dipoles can lead to significantly higher scattering rates than nonpolar phonons. The polar optical phonons are modeled by Frölich interactions [63, 52]. For bulk polar optical phonons the squared matrix element corresponding to the emission ('+') and absorption ('$-$') process is given by [36]

$$|M_{nn'}(q)|^2 = \frac{e^2\hbar\omega_{ph}}{4Aq}\left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0}\right)\left(n_{ph} + \frac{1}{2} \pm \frac{1}{2}\right) I_{nn'} \tag{2.65}$$

where $\hbar\omega_{ph}$ is the phonon energy, ϵ_∞ is the high frequency dielectric constant ϵ_0 is the static dielectric constant, n_{ph} is the number of phonons given by Bose-Einstein statistics and $I_{nn'}$ is the form factor given by

$$I_{nn'} = \int_z dz \int_{z'} dz' \xi_n(z)\xi^\dagger_{n'}(z)\xi_n(z)\xi^\dagger_{n'}(z')\xi_n(z')e^{-q|z-z'|} \tag{2.66}$$

The computational effort needed to evaluate the form factors for the POP phonon is much more than those of the nonpolar phonons because of its dependence on 'q'. This dependence on 'q' makes the scattering anisotropic.

Alloy scattering

The presence of foreign atoms randomly placed in an otherwise homogeneous crystal breaks the periodicity of the crystal structure and hence the periodicity of lattice potential. These fluctuations in the periodic lattice potential lead to scattering which is referred to as alloy scattering. The unscreened squared matrix element is given by [36]

$$|M_{nn'}|^2 = \frac{\Omega_C}{A}\Delta U^2 x(1-x)\int dz |\xi_n(z)|^2 |\xi_{n'}(z)|^2 \tag{2.67}$$

where Ω_C is the volume of the unit cell, x is the molar fraction of one of the materials and ΔU is the alloy scattering potential.

Coulomb scattering

The presence of charges which are randomly placed in a device lead to scattering potential. In the MSMC simulator the Coulomb scattering model comprises the effects of finite thickness gate stack with high-κ and interfacial layer. Let T_{HK} and T_{ITL} correspond to the thicknesses of high-κ dielectric and interfacial layer respectively. Then the squared unscreened matrix element is given by [36]

$$\begin{aligned}|M_{nn'}(\mathbf{q})|^2 =& \frac{1}{A}\int_{-T_{HK}-T_{ITL}}^{-T_{ITL}} |M^0_{nn'}(q,z')|^2 N_{HK}(z')dz' + \\ & \frac{1}{A}\int_{-T_{ITL}}^{0} |M^0_{nn'}(q,z')|^2 N_{ITL}(z')dz' + \\ & \frac{1}{A}\int_{0}^{y_{max}} |M^0_{nn'}(q,z')|^2 N_{semi}(z')dz' + \\ & \frac{1}{A}|M^0_{nn'}(q,-T_{ITL})|^2 N_{HK/ITL} + \frac{1}{A}|M^0_{nn'}(q,0)|^2 N_{ITL/semi}\end{aligned} \tag{2.68}$$

where $N_{HK}(z')$ is the volumetric charge density in high-κ oxide, $N_{ITL}(z')$ is the volumetric charge density in interfacial oxide, $N_{semi}(z')$ is the volumetric charge density in semiconductor, $N_{HK/ITL}$ is surface charge density at the interface between the high-κ and interfacial layer and $N_{ITL/semi}$ is surface charge density at the interface between the interfacial oxide and the semiconductor layer and $M^0_{nn'}(q,z')$ is given by

$$M^0_{nn'}(q,z') = \int_z \xi_{n'}(z)\xi_n(z)\left(\frac{e}{2q\epsilon_S}e^{-q|z-z'|} + Ce^{-qz}\right)dz \tag{2.69}$$

where C is the constant determined by the boundary condition.

2.6.3 Parallelization

The price of using accurate models for surface roughness scattering and screening formulations must be paid in terms of simulation time. The nonlinear model for calculation of single unscreened ensemble averaged squared matrix element needs four numerical integrations. In scalar screening we have to calculate only the autocorrelation function. Thus the number of matrix elements, N_{ME}, to be calculated with scalar screening are

$$N_{ME} = \sum_{i=1}^{N_{val}} N^2_{SB/val} \tag{2.70}$$

where $N_{SB/val}$ are the number of subbands per valley. However if the tensorial screening is used, then we have to evaluate all the cross-correlation terms as well, hence the N_{ME} is give by

$$N_{ME} = \left(\sum_{i=1}^{N_{val}} N^2_{SB/val}\right)^2 \tag{2.71}$$

Thus the tremendous increase in the computational load makes parallelization a necessity to keep the simulation time under practical limits.

2.7 Validation of the nonlinear surface roughness scattering implementation

Fig.2.3 compares the mobility in double gate silicon MOSFET in on-state (inversion charge density, N_{inv}=$2 \times 10^{13} cm^{-2}$) where surface roughness is expected to be dominant scattering mechanism. The mobility values are calculated with screening (scalar and tensorial) and without screening. The mobility values calculated using the linear and nonlinear model closely track each other (with different values of Δ_{rms}), thus validating the implementation of the nonlinear model. We also emphasize that as the well thickness reduces the mobility values with scalar and tensorial screening differ both qualitatively and quantitatively. This is because of the violation of assumptions in the formulations of scalar screening in double gate structures as discussed in [24]. The unscreened mobility values are significantly lower than the screened ones because the screening potential is overestimated as discussed in section 2.3.

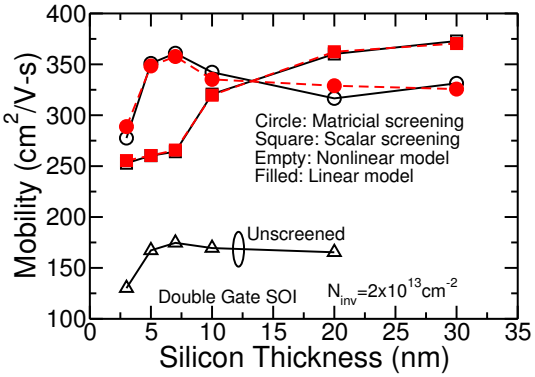


FIGURE 2.3: Comparison of the mobility versus Si thickness for Si double gate SOI MOSFETs using different screening formulations and without screening, and either using the linear or the nonlinear surface roughness scattering models. For DG structures there is a significant inconsistency in mobility values using either the tensorial or the scalar screening, as already pointed out in [24]. The surface roughness parameters, Δ_{rms} = 0.66 nm and Δ_{rms} = 0.21 nm has been used with linear and non-linear model respectively. The correlation length, Λ = 1.4 nm is used for both models.

Chapter 3

Mobility analysis of planar transistors

3.1 Introduction

In this chapter, we analyze a number of experimental mobilities in planar (bulk and ultra-thin-body) devices of different materials, which may be architectures of choice in sub 14 nm nodes [25]. In order to simulate the FETs, we have used the Monte Carlo based simulation suite discussed in the Chapter 2 using the nonlinear surface roughness scattering model and other relevant scattering mechanisms. We also account for screening, which can strongly influence the mobility values. We have used the experimental mobility values because they can be used as a proxy for determining FET performance, but using lower computational resources.

With the progress (read complexity) in fabrication technology and the rising cost of performing experimental research, it is necessary to analyze different architectures and materials with reasonable (in the range of experimental) values of physical and morphological parameters using simulation tools. In addition, simulations provide us with an opportunity to isolate different aspects of the transport in a controlled environment, and help identify the most important causes of degradations in performances.

3.2 Comparison with experimental mobilities: Bulk devices

Fig.3.1 shows the calibration of the surface roughness model against the experimental universal mobility curves for silicon and for a wide range of temperatures [64]. The simulations included intra-valley acoustic (in the elastic approximation), inelastic inter-valley optical phonons and surface roughness. The values of the Δ_{rms} and Λ were set to 0.21 nm and 1.4 nm respectively. These values of the surface roughness parameters are very well in the experimentally measured range (different experimental values are reported in the previous chapter). However, if the linear model is used the value of Δ_{rms} needed are about 2 to 3 times larger than the experimental values, as seen in the previous chapter. Fig.3.1 also compares the scalar and tensorial screening. As expected for bulk devices scalar screening is an excellent approximation for tensorial screening, and hence they both are in good agreement with the experimental data [24, 65].

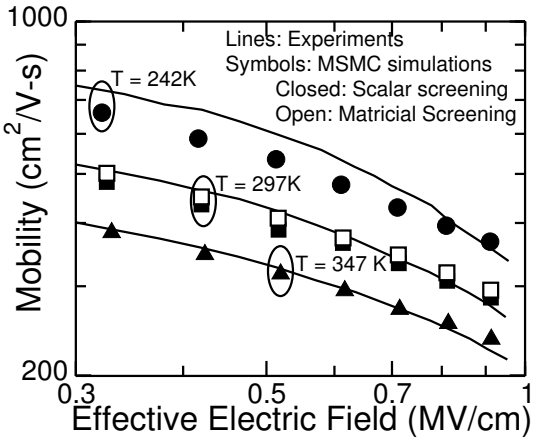


FIGURE 3.1: Plot of comparison between the simulated and experimental mobility versus effective electric field for a bulk unstrained silicon MOSFETs for different temperatures. The surface roughness parameters used are Δ_{rms}=0.21 nm and Λ=1.4 nm with the nonlinear model which are very well in the experimental range.

Fig.3.2 compares the experimental and simulated mobility for bulk III-V device. In order to account for the relevant scattering mechanisms in the MOSFET having $In_{0.53}Ga_{0.47}As$-HfO_2 system, we have taken alloy, elastic acoustic phonons, local and remote polar optical phonons and surface roughness scattering. In order to calibrate the linear model for surface roughness scattering the value of Δ_{rms} needed was 1.3 nm, which is a very large number [46]. However with the nonlinear model the Δ_{rms} required to reproduce experiments was 0.4 nm, a significantly smaller and reasonable value and larger than the well established Si-SiO_2 interface.

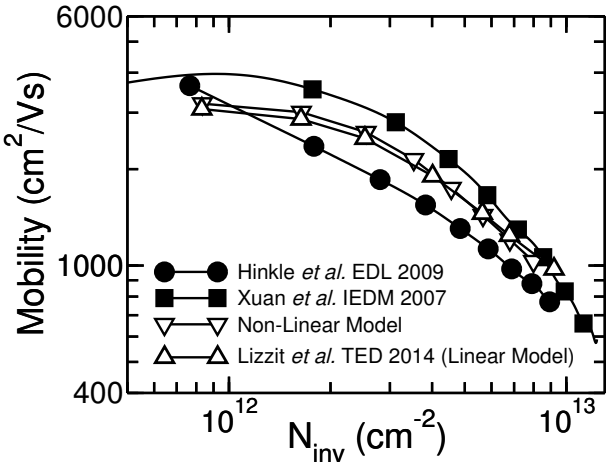


FIGURE 3.2: Comparison between the experimental and simulated mobilities (with linear and nonlinear model for surface roughness scattering). Δ_{rms}= 0.4 nm and Λ= 1.5 nm were needed with the nonlinear surface roughness model. The data for the linear model has been taken from [46] (Δ_{rms}= 1.3 nm and Λ= 1.5 nm). The channel doping of 2×10^{17} cm^{-3} was considered in the simulations. Other scattering parameters are same as in [46]. Experimental data are taken from [66, 67].

TABLE 3.1: Surface Roughness parameters used in simulations.

Structure	Semiconductor-Oxide Interface (Δ_{rms}; Λ)	Active-Buffer Interface(Δ_{rms}; Λ)
Bulk Si (Fig.3.1)	0.21 nm; 1.4 nm	—
Bulk-like III-V (Fig.3.6)	0.2 nm; 1 nm	0.0
III-V Quantum well(Fig.3.3)	0.55 nm; 1 nm	—
III-V Heterostructure (Fig.3.4)	0.55 nm; 1 nm	0.275 nm; 1 nm

3.3 Comparison with experimental mobilities: Thin body devices

We begin this section by comparing the experimental mobilities for an $In_{0.7}Ga_{0.3}As$ channel sandwiched between 2 dielectrics as shown in Fig.3.3. This structure corresponds to a deep uniform potential well. The material parameters are listed in Tab.3.2. The effective mass was calculated using $\mathbf{k}\cdot\mathbf{p}$ method [68] with the parameters in [69]. The comparison between the experimental [70] and simulated mobilities are shown in Fig.3.3 for different well thickness, T_w. An important feature of the comparison between the mobilities is that the value of Δ_{rms}=0.55 nm used in the simulation and experimentally measured are same. As can be seen, we obtain a good match for T_w=10 nm. However, for T_w=5 nm the simulated mobilities are larger than the experimental mobilities, which are quite small (≈70 cm^2/V-s) and can be reproduced only by setting Δ_{rms} to 0.65 nm [65].

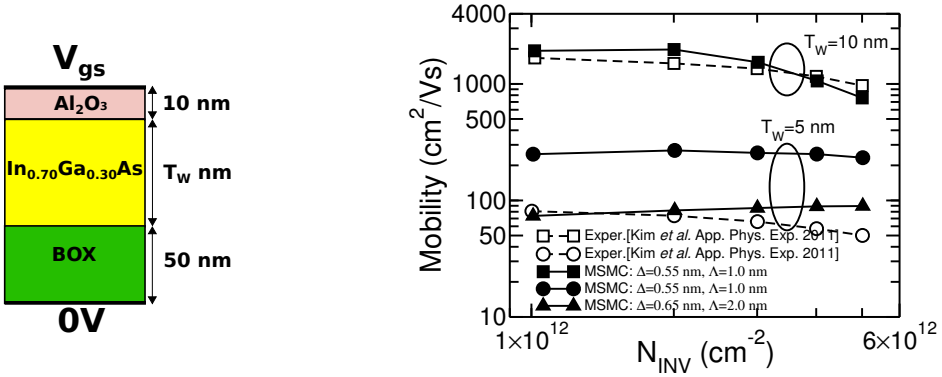


FIGURE 3.3: (a) Schematic of the fabricated (and simulated) structure. (b)Comparison between the experimental [70] and simulated mobilities. We emphasize that the Δ_{rms}=0.55 nm used for simulations is same as experimentally measured.

The next experimental mobility set had a heterostructure structure. In this case a channel of $In_{0.7}Ga_{0.3}As$ was placed between the $In_{0.3}Ga_{0.7}As$ buffer layers (Fig.3.4) with front and buried oxide of Al_2O_3 [75]. In order to accurately account for surface roughness scattering the nonlinear model was extended to account for scattering at the buffer-active layer interface and we assume that different interfaces are uncorrelated. Even though the buffer-active layers are expected to be smooth, the wavefunctions at these interfaces are much larger than at the oxide-semiconductor interfaces (see Fig.3.8). The material parameters are listed in

TABLE 3.2: Material parameters for the semiconductor and dielectrics employed in the simulations.

Material	Electron Affinity (eV)	Electron mass (m_0)	Relative permittivity (ε_0)
InAs	4.9	0.026	15.15
$In_{0.7}Ga_{0.3}As$	4.64	0.0365	14.30
$In_{0.3}Ga_{0.7}As$	4.32	0.0529	13.42
GaAs	4.07	0.067	12.9
Al_2O_3	1 eV[71]	0.23[72]	11
ZrO_2	1.64[73]	0.12 [74]	24

Tab.3.2. The surface roughness parameters Δ_{rms} and Λ used in the calibration of the nonlinear model with the experimental mobilities are listed in Tab.3.1. They are same as in Fig.3.3, which are the same as the experimentally measured in [70]. Fig3.4(b) illustrates the comparison between the experimental and simulated mobilities for different active layer thickness. The simulation results show that we have a good match for T_{act}=5 nm and T_{act}=3 nm but the mobilities for T_{act}=1 nm is overestimated.

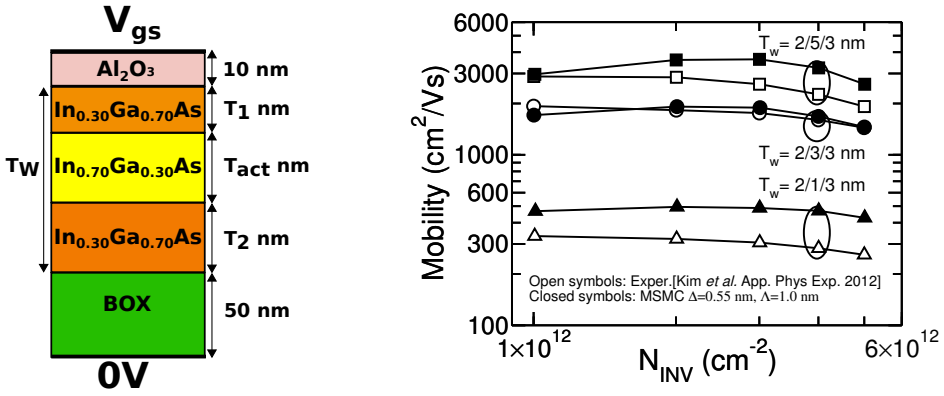


FIGURE 3.4: (a) Schematic of the fabricated (and simulated) structure. (b) Surface roughness scattering was also considered at the active-buffer layer interface with the parameters listed in Table.3.1. We also assumed interface charge density of 2×10^{12} cm^{-2} to improve the match between the experiments and simulations.

In Fig.3.5 we plot the experimental and simulated mobilities versus the well thickness at N_{inv}=5×10^{12} cm^{-2} for devices analyzed in Fig.3.3 and Fig.3.4. On the log-log plot, we observe that the dependence of experimental mobility versus T_w for planar III-V devices is significantly weaker than T_w^6, which has been observed in the experiments and simulations for silicon devices at low temperatures [76]. This is because of the reduced semiconductor-oxide barrier as compared to Si-SiO_2 and thus reducing the impact of surface roughness.

We further analyzed a more recent experimental mobility data set published in [77]. The structure consists of a 10 nm thick InAs channel on a lattice matched wide bandgap AlAsSb buffer layer (sketched in Fig.3.6(a)). The material parameters for the ZrO_2 (gate oxide) are listed in Tab3.2. Fig.3.6(b) shows the experimental and simulated mobilities. The surface

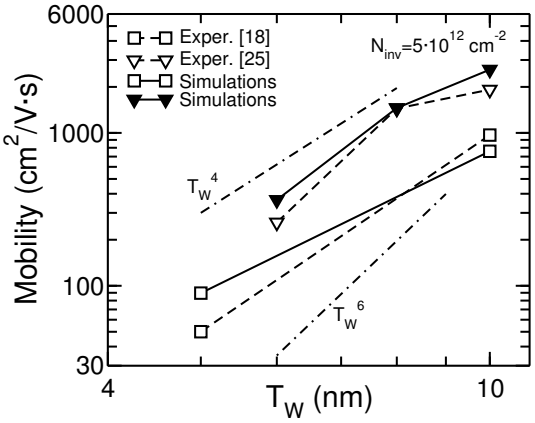
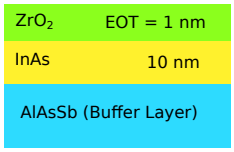


FIGURE 3.5: Comparison between the experimental and simulated values of mobility for different values of well thichkness,T_w from Fig.3.3 and Fig.3.4. The degradation of the mobility with well thickness is much lower than the T_w^6 trend.

roughness parameters Δ_{rms}=0.2 nm and Λ=1 nm were used in the simulations suggest excellent interface quality fully comparable to silicon based devices. We have compared the mobility up to N_{inv}=5×10^{12} cm^{-2}. To improve the quality of match with the experiments we have considered a small fixed interface charge density (N_s=2.8 × 10^{12} cm^{-2}).



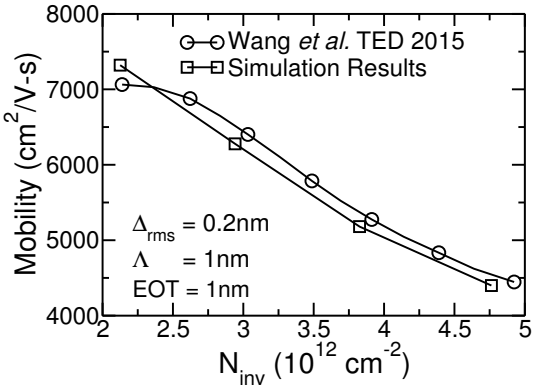


FIGURE 3.6: Comparison between the experimental and simulated mobilities for bulk-like structure (shown on left). The small value of Δ_{rms} needed to match the experimental mobilities suggest an interface quality comparable with Si-SiO_2.

In the above three analyses of the experiments, Fig.3.4, Fig.3.3 and Fig3.6, we restricted the comparison up to N_{inv}=5×10^{12} cm^{-2}. Because at higher N_{inv} there can be electrons that get trapped in interface states and thus contributing to the charge neutrality but not to the transport. Thus at higher N_{inv} there can be fictitious roll-off in mobility versus N_{inv} curves.

3.4 Impact of the energy barrier between buffer and active layer

The importance of buffer layers has increased because of their ability to suppress the detrimental effects of surface roughness, remote phonons and fixed charges at the semiconductor-oxide interface [28]. This is achieved by pushing the wavefunction away from the oxide. It is interesting to notice, however, that the influence of the energy barrier at the active-buffer interface is at least twofold, and it is studied here considering an InAs active layer. Fig.3.7

shows impact active-buffer barrier height on mobility with and without surface roughness at active-buffer layer interface. When the roughness of the active-buffer layer interface is neglected the mobility continues to increase with the increase in the barrier height, while the mobility versus buffer-active layer barrier flattens out if the corresponding surface roughness is accounted for.

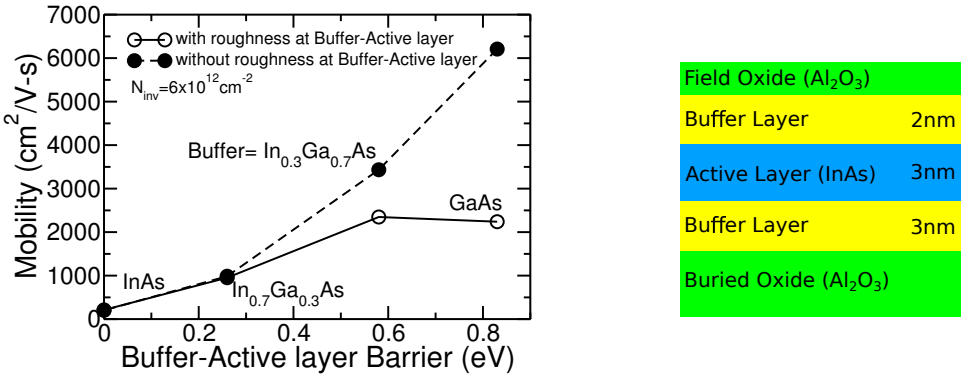


FIGURE 3.7: Shown is the impact of the energy barrier between the active and buffer layers on mobility at fixed inversion density, N_{inv}=6 $\times$ 10^{12} cm^{-2}. The SR parameters are the same as in Fig.3.4. The structure and the geometrical dimensions used in the simulation are shown on the right.

At small energy barriers, the dominant effect is that, by increasing the barrier, the electron wavefunction is pushed farther from the semiconductor-oxide interface (as can be seen in Fig.3.8), which reduces the surface roughness scattering, thus improving the mobility at given N_{inv}, as illustrated in Fig.3.7. But when the barrier at the active-buffer interface is increased further, the surface roughness at the interface becomes significant and degrades the mobility improvement. In these simulations, all the parameters were kept constant except the material parameters of the buffer layer which are selected according to the buffer material case by case. While the results in Fig.3.7 are focused on the SR related effects due to the change in the energy barrier between active and buffer layer, we acknowledge that the physical picture of such heterostructure transistors is more complex than described in our simulations, because changing the molar fraction in the buffer layer also affects its lattice constant and thus the lattice mismatch with the InAs active layer. The possible strain induced by such a lattice mismatch depends on the device architecture and the fabrication process, and it is not accounted for in our simulations.

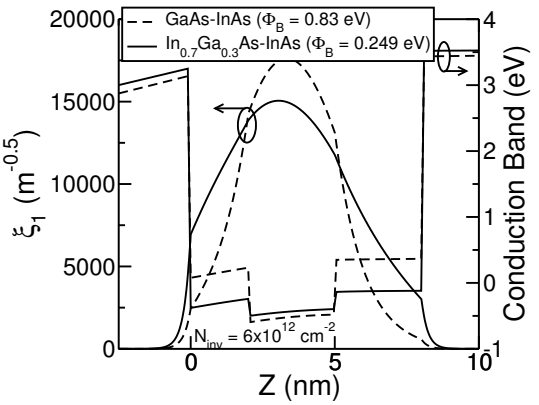


FIGURE 3.8: Wavefunction corresponding to the lowest subband for the III-V heterostructure shown in the figure 3.7. The value of the wavefunction at the active-buffer layer is much larger than that at the oxide-buffer layer. The surface roughness scattering and material parameters are listed in Table3.1 and Table3.2 respectively.

Chapter 4

Surface roughness scattering in 3D transistors

4.1 Introduction

The constant drive to continue scaling has pushed the critical physical dimensions of the MOS transistor in few nanometers. In order to be able to switch off the transistor at such small dimensions, it is necessary to give a boost to the gate control to counteract the short channel effects. In order to enhance the gate control, multiple gate devices are expected to replace all the planar architectures beyond 7 nm node [25, 78, 79, 80, 81, 82, 83]. However, this transformation of the device architecture increases the surface to volume ratio which makes the relative importance of the surface related effects in the operation of FETs even more critical [84, 85, 86, 87, 88, 89].

The first thorough surface roughness analysis of 3D architecture was performed on cylindrical nanowires with a simplifying assumption in terms mass anisotropy and was dependent on the symmetry considerations of the circular geometry [42]. However, such simplifying assumptions are not justified when dealing with real-life devices. Moreover it used a linear model to account for surface roughness scattering. Even the more recently developed surface roughness scattering models that account for arbitrary cross-section shape and mass anisotropy are based on the perturbative treatment of the impact of surface aspersions on the wavefunction and potential [90, 91, 92]. These models assume that matrix elements are linearly proportional the surface roughness, while it has been demonstrated that they are strongly nonlinear [22]. These linear models tend to overestimate the Δ_{rms} parameters as discussed in chapter 2. In this chapter, we will discuss in detail the derivation of the nonlinear model for devices having fairly arbitrary and smooth cross-sections [93, 94]. We will also develop simplified versions of the ensemble averaged squared matrix element to reduce the computational burden. We will further justify the use of simplifying assumption with thorough numerical analysis and verification.

4.2 Nonlinear model for surface roughness scattering

The growing importance of the surface roughness and the need for models that can reproduce the experimentally measured mobility with credible values of the parameters related to surface morphology (correlation length, Λ, and root mean squared values of surface roughness, Δ_{rms}) provided the motivation for this work. The good results obtained in this respect from the nonlinear surface roughness model for planar devices encouraged us to develop the nonlinear model for surface roughness scattering for 3D devices having a fairly arbitrary cross-section. In the development of the model we will assume that different interfaces are uncorrelated and also neglect the impact of surface roughness on the potential profile as discussed in chapter 2. Here will discuss in detail its derivation [93, 94].

4.2.1 Coordinate system

In order to accurately model the surface roughness the first step is to describe the oxide-semiconductor interface, assuming that the device has the same nominal cross-section along the transport direction. To this end, let us consider the arbitrary, smooth cross-section of the device shown in the Fig.4.1. The transport is taken to be along 'x' and cross-section is in the **r**≡(y,z) plane. The nominal interface between oxide and semiconductor is indicated by $\mathcal{I}_0$ and contours at a distance η in the direction normal to the $\mathcal{I}_0$, $\hat{n}$, are labeled by $\mathcal{I}_\eta$. A natural coordinate system to model the contour ($\mathcal{I}_0$) is a curvilinear system which defines the position along the contour, s_0, and normal to the contour, η. More generally for $\mathcal{I}_\eta$ we define the coordinate system by (s_η, η). The perturbations to the nominal oxide-semiconductor interface (surface roughness), $\Delta(x, y, z)$, is shown by the dashed region in Fig.4.1(b). These perturbations are a function of the interface position which is conveniently expressed in the coordinate $\mathbf{r}_I \equiv$(s,x). The $\mathrm{P_R}\left[\Delta(s, x)\right]$ is the small region surrounding nominal interface ($\eta_{min} \leq \eta \leq \eta_{max}$) where the perturbation is non-null. It is this region that contributes to the surface roughness scattering.

4.2.2 Calculation of the matrix elements

In order to calculate the matrix elements we first make an ansatz that the electron envelope wavefunction for 1D electron gas can be written as

$$\Psi_{n,k_x}(\mathbf{r}, x) = \xi_n(\mathbf{r})\frac{e^{i\,k_x\,x}}{\sqrt{L_x}} \tag{4.1}$$

where L_x is the normalization length along the transport direction, $\xi_n(\mathbf{r})$ is the component in the confinement plane and a plane wave ($e^{i\,k_x\,x}$) along the transport direction. The envelope wavefunction in the confinement plane can be calculated by solving the Schrödinger equation

$$\left[-\frac{\hbar^2}{2}\nabla\cdot(W_{yz}\nabla) + U(\mathbf{r})\right]\xi_n(\mathbf{r}) = \varepsilon_n^{(p)}\xi_n(\mathbf{r}) \tag{4.2}$$

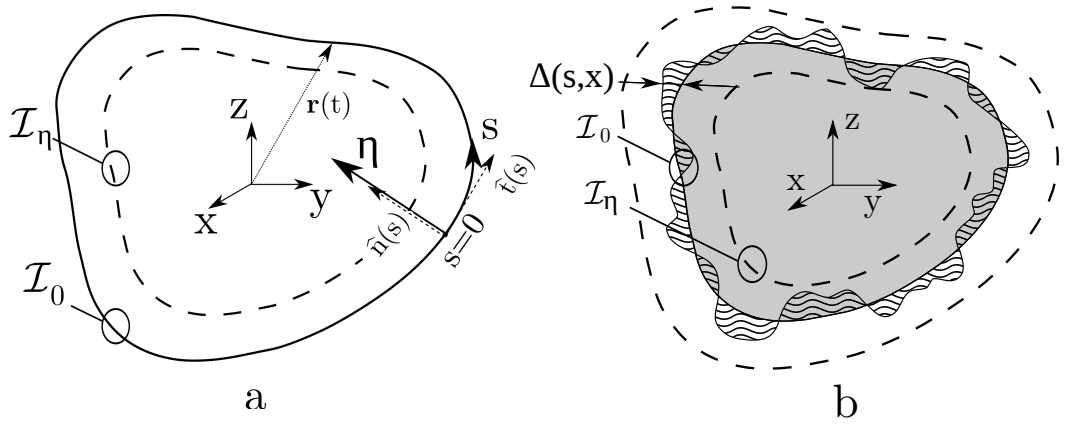


FIGURE 4.1: (a) sketch of the cross-section of a MuGFET where x is the transport direction and $\mathcal{I}_0$ denotes the curve describing the semiconductor-oxide interface. s and η are respectively the abscissa along $\mathcal{I}_0$ and normal to $\mathcal{I}_0$, and $\mathcal{I}_\eta$ is the curve at a distance η from $\mathcal{I}_0$. D_0 and D_η are the perimeters of $\mathcal{I}_0$ i.e. the perimeter of the semiconductor region) and of $\mathcal{I}_\eta$. (b) sketch of the cross-section of a MuGFET perturbed by a surface roughness stochastic process $\Delta(s,x)$. The dashed area is the perturbed region $\mathrm{P_R}\left[\Delta(s,x)\right]$ where $\Delta(s,x)$ is non null; the grey region is the semiconductor.

where $U(\mathbf{r})$ is the total potential energy, $\varepsilon_n^{(p)}$ is the parabolic eigen energy and W_{yz} is the 2D inverse effective mass tensor in the device coordinate system.

By definition, the matrix element for surface roughness scattering is given by

$$\mathcal{M}_n^{n'}(q_x) = \frac{1}{L_x}\int_{L_x} M_{nn'}[\Delta(s,x)]e^{-i\,q_x x}\,dx \tag{4.3}$$

where q_x is difference between the initial and final wavevector, k_x, $M_{nn'}[\Delta(s,x)]$. We here consider only the potential energy contribution, because the kinetic energy term can lead to unphysical matrix elements in effective mass formulation of the Hamiltonian [22], and thus we can write matrix element as

$$M_{nn'}[\Delta] = \iint\limits_{\mathrm{P_R}[\Delta(s,x)]} \xi_{n'}^{\dagger}(\mathbf{r})\,\Phi_B\xi_n(\mathbf{r})\,d\mathbf{r} \tag{4.4}$$

where Φ_B is the potential barrier between the oxide and semiconductor and $\xi_n(\mathbf{r})$ is the wavefunction corresponding to the n^{th} subband. The integration in the eq.4.4 is in the Cartesian coordinate system. However, as we have discussed earlier, the most convenient coordinate system for the surface roughness (i.e. the perturbation to the contour, $\mathcal{I}_0$) is the curvilinear coordinate system defined by $'s_0'$ in the cross-section. Hence it is necessary to change the coordinate system from (x,y,z) to (x,s_0,η).

To this end, we expand the wavefunction along the closed curves, $\mathcal{I}_\eta$, in Fourier series, in fact because any function on a closed curve is always periodic. Thus the wavefunction can be written as

$$\xi_n(s_\eta,\eta) = \sum_l \phi_{n,l}(\eta)\,e^{i2\pi l s_\eta/D_\eta} \qquad\qquad l = 0,\pm1,\pm2,\ldots \tag{4.5}$$

l is the mode, with $\phi_{n,l}(\eta)$ defined as

$$\phi_{n,l}(\eta) = \frac{1}{D_\eta} \int_{\mathcal{I}_\eta} \xi_n(\mathbf{r})\, e^{-i2\pi l s_\eta / D_\eta} d\mathbf{r}\,. \tag{4.6}$$

Using the results derived in appendix A (specifically eq.A.9 and eq.A.11) and substituting eq.4.5 in eq.4.4 we obtain

$$M_{nn'}[\Delta] = \int_{D_0} M_{nl}^{n'l'}\left[\Delta(s,x)\right]\, e^{-i\, q_{ll'} s} ds \tag{4.7}$$

where $M_{nl}^{n'l'}\left[\Delta(s,x)\right]$ is defined as

$$M_{nl}^{n'l'}\left[\Delta(s,x)\right] = \int_0^{\Delta(s,x)} D_\eta\, \phi^\dagger_{n'l'}(\eta)\, \Phi_B\, \phi_{nl}(\eta)\, d\eta \tag{4.8}$$

and $q_{ll'}$=$2\pi(l'-l)/D_0$. Physically speaking, $q_{ll'}$ can be considered as change is the wavevector associated with the modes. The form of $M_{nl}^{n'l'}\left[\Delta(s,x)\right]$ as derived in the eq.4.8 is same as in the planar device except that here it must be evaluated for each pair of modes (l,l'). According to the fig.4.1, for $\Delta(s) > 0$ the perturbation enters into the semiconductor and for $\Delta(s) < 0$ it enters into the oxide region.

An important remark on using Fourier series is that it separates wavefunction into component normal ($\phi_{n,l}(\eta)$ a constant for each $\mathcal{I}_\eta$) to contour $\mathcal{I}_\eta$ and component ($e^{-i2\pi l s_\eta / D_\eta}$) along the contour $\mathcal{I}_\eta$. This makes the nonlinear transformation of random process described by Δ(s,x) (eq.4.8) a time (here time is equivalent 's') invariant, because it doesn't depend explicitly on 's', which is important in the calculation of the ensemble averaged squared matrix element, $\left\langle \left| \mathcal{M}_n^{n'}(q_x) \right|^2 \right\rangle$.

4.2.3 Calculation of the ensemble averaged squared matrix elements

Using the eqs.4.3, 4.7 and 4.8 allow us to write the unscreened matrix element as

$$\mathcal{M}_n^{n'}(q_x) = \frac{1}{L_x\, D_0} \sum_{l,l'} \int_{L_x} \left[\int_{D_0} M_{nl}^{n'l'}\left[\Delta(s,x)\right]\, e^{-i\, q_{ll'} s} ds \right] e^{-i\, q_x x} dx \tag{4.9}$$

Taking complex conjugate of eq.4.9, multiplying it with itself and taking ensemble average we get ensemble averaged squared matrix element as

$$\left\langle \left| \mathcal{M}_n^{n'}(q_x) \right|^2 \right\rangle = \frac{1}{(L_x D_0)^2} \sum_{\substack{l,l' \\ g,g'}} \iint_{L_x} dx\, dx' \iint_{D_0} ds\, ds' e^{-i\, q_{ll'} s} e^{+i\, q_{gg'} s'} C_{nlg}^{n'l'g'}(\boldsymbol{\tau}) e^{-i\, q_x (x-x')} \tag{4.10}$$

where $\boldsymbol{\tau}=(s-s', x-x')$ and the correlation function, $C_{nlg}^{n'l'g'}(\boldsymbol{\tau})$, is defined as

$$C_{nlg}^{n'l'g'}(\boldsymbol{\tau}) = \left\langle M_{nl}^{n'l'}\left[\Delta(s,x)\right]\left(M_{ng}^{n'g'}\left[\Delta(s',x')\right]\right)^{\dagger}\right\rangle \tag{4.11}$$

We again emphasize that in eq.4.10 we could write the correlation function as a function of difference between (s,x) and (s',x') because $M_{nl}^{n'l'}\left[\Delta(s,x)\right]$ is a time invariant transformation of the random process, $\Delta(s,x)$. The form of the correlation function is same as in the case of planar structures except here we have a set of modes (l,g) and (l',g'). Using the Wiener–Khinchin theorem we can express the correlation function in terms of its power spectrum, $S_{nlg}^{n'l'g'}(q_s,q_x')$, as

$$C_{nlg}^{n'l'g'}(\boldsymbol{\tau}) = \frac{1}{(2\pi)^2}\int\limits_{q_s} dq_s \int\limits_{q_{x'}} dq_x'\, S_{nlg}^{n'l'g'}(q_s,q_x')e^{i\,q_s(s-s')}e^{i\,q_x'(x-x')} \tag{4.12}$$

where we have assumed that the L_x and D_0 are large enough to allow us to treat the q_x and q_s as continuous variables. These are reasonable assumptions because the length of the device is generally much larger than all the relevant length scales and the correlation length, Λ, is generally between 1 to 2 nm, which is much smaller than the perimeter of the cross-section of the realistic devices [25]. Putting eqs.4.10 and 4.12 together we get the final form for the ensemble averaged squared matrix element as

$$\begin{aligned}\left\langle\left|\mathcal{M}_n^{n'}(q_x)\right|^2\right\rangle = \frac{1}{\left(2\pi L_x D_0\right)^2}\sum_{\substack{l,l'\\g,g'}}\int\limits_{q_s} dq_s \int\limits_{q_x'} dq_x'\, S_{nlg}^{n'l'g'}(q_s,q_x')\iint\limits_{L_x} dx\,dx' e^{i\,(q_x'-q_x)\,(x-x')}\times\\ \times\int\limits_{D_0} ds\, e^{-i(q_{ll'}-q_s)s}\int\limits_{D_0} ds'\, e^{+i\left(q_{gg'}-q_s\right)s'}\end{aligned} \tag{4.13}$$

However the procedure is not yet complete as we have not specified the form of the power spectrum. As in the case of planar architectures, the correlation function for stationary random process having Gaussian first and second moment is given by [53]

$$\begin{aligned}C_{nlg}^{n'l'g'}(\boldsymbol{\tau}) = \frac{1}{2\pi\, C_\Delta(0)\sqrt{1-C_{\Delta,N}^2(\boldsymbol{\tau})}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} M_{nl}^{n'l'}[\Delta_1]\,\left(M_{ng}^{n'g'}[\Delta_2]\right)^{\dagger}\times\\ \times\exp\left[-\frac{\Delta_1^2+\Delta_2^2-2C_{\Delta,N}(\boldsymbol{\tau})\Delta_1\Delta_2}{2C_\Delta(0)(1-C_{\Delta,N}^2(\boldsymbol{\tau}))}\right]d\Delta_1\,d\Delta_2\end{aligned} \tag{4.14}$$

where $C_\Delta(\boldsymbol{\tau})$ is the auto-correlation function of the roughness process itself, $\Delta(s,x)$, and $C_{\Delta,N}(\boldsymbol{\tau}) = C_\Delta(\boldsymbol{\tau})/C_\Delta(0)$. The correlation function for the surface roughness, $\Delta(s,x)$ is given by assuming exponential form [30]

$$C_\Delta(\tau) = \Delta_{rms}^2 e^{-\tau\sqrt{2}/\Lambda} \tag{4.15}$$

As the correlation function depends only on the magnitude of $\boldsymbol{\tau}$, its Fourier transform (equivalent to Hankel transform) can be written as

$$S_{nlg}^{n'l'g'}(q) = 2\pi \int_0^\infty \tau\, C_{nlg}^{n'l'g'}(\tau) J_0(q\tau) d\tau \tag{4.16}$$

where $J_0(x)$ is the zeroth order Bessel function of the first kind.

In the limit L_x tends to infinity the integration over (x, x') can be written as

$$\lim_{L_x\to\infty} \iint_{L_x} e^{i\,(q_x'-q_x)\,(x-x')} dx\, dx' = 2\pi\, L_x \delta(q_x' - q_x) \tag{4.17}$$

and the integration along the contour $\mathcal{I}_0$ reduces to

$$\frac{1}{D_0}\int_{D_0} e^{i\,q\,s} ds = \frac{sin(0.5\,D_0\,q)}{0.5\,D_0\,q} \equiv sinc\left(\frac{qD_0}{2}\right). \tag{4.18}$$

Using the eq.4.17 and 4.18 along with 4.13 we get

$$\left\langle \left| \mathcal{M}_n^{n'}(q_x) \right|^2 \right\rangle = \frac{1}{L_x D_0} \sum_{\substack{l,l' \\ g,g'}} \int_{q_s} dq_s\, F\left(q_{ll'}, q_{gg'}, q_s\right) S_{nlg}^{n'l'g'}(q_s, q_x) \tag{4.19}$$

where the form factor $F\left(q_{ll'}, q_{gg'}, q_s\right)$ if given by

$$F\left(q_{ll'}, q_{gg'}, q_s\right) = \frac{D_0}{2\pi} \text{sinc}\left[\frac{(q_{ll'} - q_s)\, D_0}{2}\right] \text{sinc}\left[\frac{(q_{gg'} - q_s)\, D_0}{2}\right]. \tag{4.20}$$

The flowchart summarizing the key equations for the calculating the ensemble averaged squared matrix element is shown in Fig.4.2

4.2.4 Simplifications to ensemble averaged squared matrix elements

Even though the final expression for the ensemble averaged squared matrix does not look computationally ominous, it is in fact much more computationally heavy then its planar counterpart. The main culprits for the computational burden are the evaluations of the correlation function (which has a double integral) and power spectrum must be evaluated for all the pairs of pairs of modes (l, g) and (l', g'). This hints us that if we can reduce the number of modes then it would drastically reduce the number of times the correlation function and its power spectrum is calculated. In addition, what ever simplification are made, they must be independent of the cross-section shape. These two criteria for simplifications are satisfied by the form factor (4.20).

We first of all notice that the $F\left(q_{ll'}, q_{gg'}, q_s\right)$ is essentially a product of two "$sinc$" functions and we know that a "$sinc$" function, by definition, is peaked where its argument is zero. Thus we can expect that the value of the form factor at $(q_{ll'} = q_s)$ and $\left(q_{gg'} = q_s\right)$ be much larger than for other arguments. Thus the form factor for $q_{ll'}$=$q_{gg'}$ (i.e. for $(l-l')$=$(g-g')$)

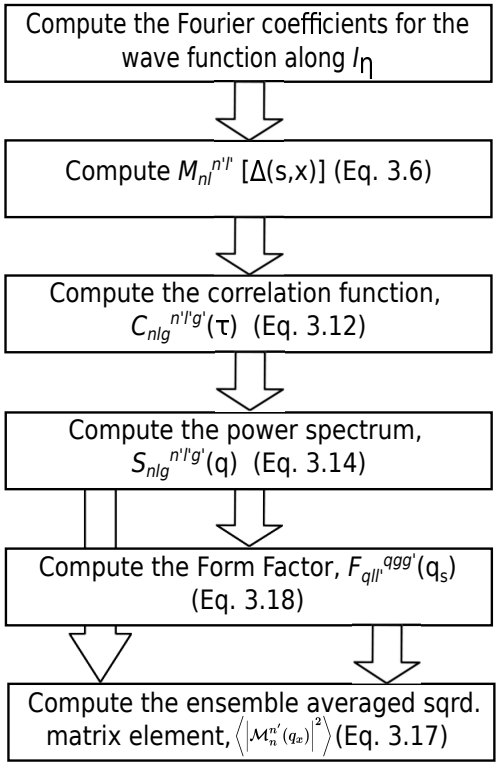


FIGURE 4.2: Numerical recipe for calculation of the $\left\langle \left|\mathcal{M}_n^{n'}(q_x)\right|^2 \right\rangle$.

is the dominant one (see 4.3) as compared to other form factors. This is confirmed by the calculations in fig.4.3. This allows us to restrict the summation in eq.4.19 over those modes that satisfy $(l - l') = (g - g')$. Thus the eq.4.19 can be written as

$$\left\langle \left|\mathcal{M}_n^{n'}(q_x)\right|^2 \right\rangle \approx \frac{1}{L_x D_0} \sum_{(l-l')=(g-g')} \int_{q_s} dq_s \, F\left(q_{ll'}, q_{ll'}, q_s\right) S_{nlg}^{n'l'g'}(q_s, q_x). \tag{4.21}$$

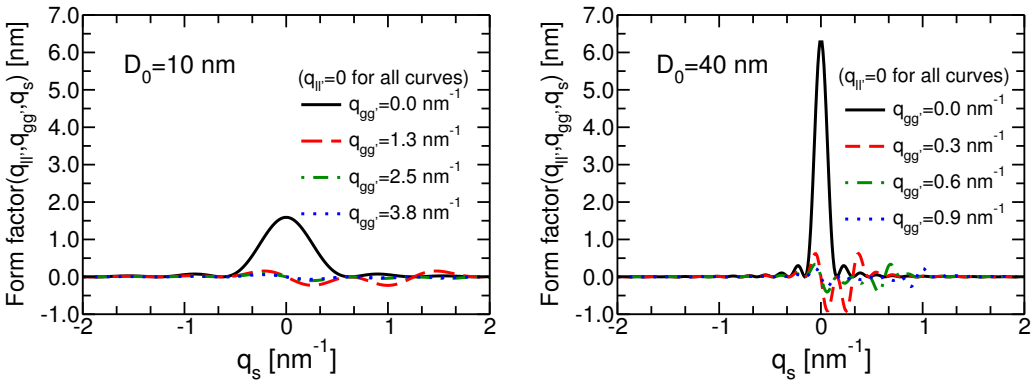


FIGURE 4.3: Form factors as calculated from Eq.4.20 for D_0=10 nm (a) and D_0=40 nm (b). The $q_{gg'}$=$2\pi(g' - g)/D_0$ values are obtained by setting g=0 and g'=$0, 2, 4, 6$. The values of the form factor where $(q_{ll'} = q_{gg'})$ is dominant one (Approx1) and with increase in D_0 Form factor becomes more representative of dirac delta function (Approx2).

An important point that is highlighted in Fig.4.3(b) is that for large values of the perimeter, D_0, the "$sinc$" becomes extremely peaked. In the limit for D_0 tending to infinity, $F(q_{ll'}, q_{ll'}, q_s) \approx \delta(q_{ll'} - q_s)$, thus allowing us to write the eq.4.21 as

$$\left\langle \left| \mathcal{M}_n^{n'}(q_x) \right|^2 \right\rangle = \frac{1}{L_x D_0} \sum_{(l'-l)=(g'-g)} S_{nlg}^{n'l'g'}(q_{ll'}, q_x). \tag{4.22}$$

As it can be seen the use of Dirac delta function reduces the integral over the 'q_s'. The eq.4.21 and eq.4.22 provide remarkable simplification over the complete formulation eq.4.19 and they reduce the computational burden up to 10 times. The quality of the approximations made in deriving the eq.4.21 and eq.4.22 degrade continuously with the reduction of the D_0 and hence depending upon the accuracy needed these approximations may or may not be used. However in our simulations, (see section. 4.3.2) show that these approximations give accurate results for $D_0 \approx 27$ nm. And consequently we can say that these are excellent approximations for all practical D_0 values.

4.3 Validation of approximations

An important aspect of the ensemble averaged squared matrix element is that the wavefunctions are expanded in Fourier series. The number of modes needed to represent the wavefunction is important from the computational standpoint, in fact if we have "l" modes then the correlation function and power spectrum must be evaluated l^4 times. We will then validate the approximations that led us to the simplified formulations in eq.4.21 and 4.22.

4.3.1 Number of modes

Because the wave functions are expressed in Fourier modes, as discussed earlier (see eq.4.6), an upper bound must be set for the number of modes to be considered. To get a measure of the number of modes needed, we first plot wavefunctions along the interface, $\mathcal{I}_0$, for different number of modes in Fig.4.4. As we can see, as the number of modes increases the match between the actual wavefunction (numerically calculated from the Schrödinger equation) and its series expansion improves. In particular, with $l = \pm 5$ even the minute features of the wavefunctions are accurately captured.

Fig.4.5 plots the intrasubband $\left\langle \left| \mathcal{M}_0^0(q_x) \right|^2 \right\rangle$ for trigated FinFET and GAA triangular architectures for different number of modes. As the number of modes increases the $\left\langle \left| \mathcal{M}_n^{n'}(q_x) \right|^2 \right\rangle$ saturates very quickly. Fig.4.5 highlights the fact that even for significantly different structures (as seen from the electron concentration profiles), modes up to $|l|=5$ are sufficient to obtain an accurate evaluation of $\left\langle \left| \mathcal{M}_n^{n'}(q_x) \right|^2 \right\rangle$.

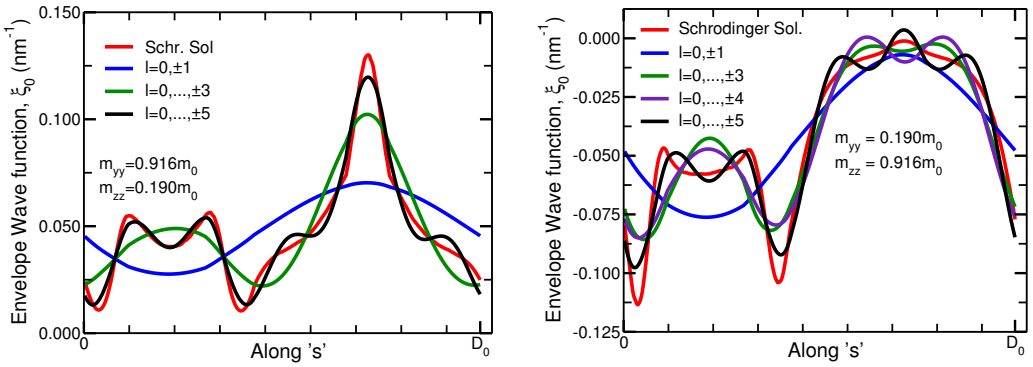


FIGURE 4.4: Plot of the lowest subband wavefunction in a triangular GAA nanowire (cross-section area = 40 nm^2) in the two valleys described by (a)$m_{yy} = 0.916m_0$ and $m_{zz} = 0.190m_0$ (b)$m_{yy} = 0.190m_0$ and $m_{zz} = 0.916m_0$ for different modes.

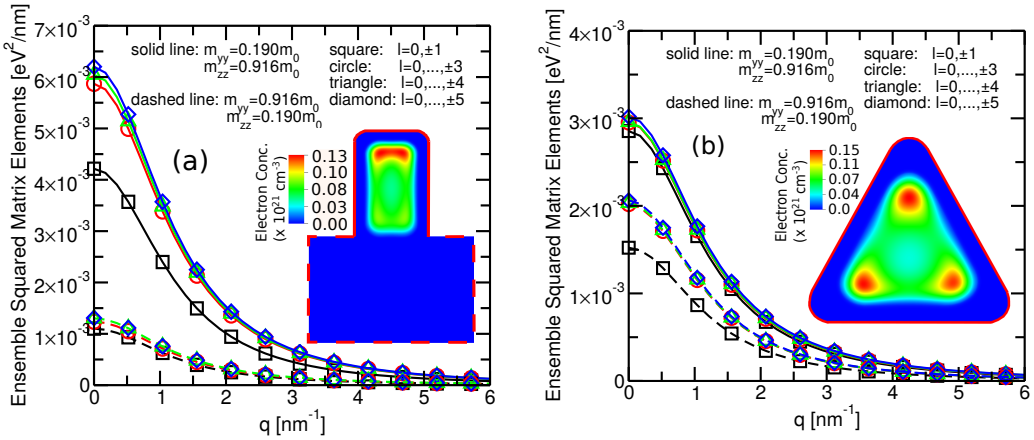


FIGURE 4.5: $\langle|\mathcal{M}_n^{n'}(q_x)|^2\rangle$ for intra-subband transitions in the lowest subband (i.e. $n{=}n'{=}0$), for Si$-$SiO$_2$ FETs and different cross-sections. The oxide thickness is 1 nm and the electron density is N$_{inv}\approx 10^{13}$ cm^2. In all plots: (a) Tri-Gate rectangular FinFET; (b) triangular GAA nanowire FETs. Area is 40 nm^2 for both the cross-sections. Transport direction is [100]. The insets shows colormaps of the electron concentration profiles in both the cross-section emphasizing the dissimilarity between the two architectures.

4.3.2 Simplified ensemble averaged squared matrix elements

Fig.4.6 shows the mobility (the formulation for mobility calculations for 3D devices will be discussed in the following chapter) versus inversion charge density for FinFET and triangular GAA nanowires using the exact formulation (eq.4.19) and the approximate formulations eq.4.21 and 4.22. We see that both the approximate forms of the $\left\langle \left|\mathcal{M}_n^{n'}(q_x)\right|^2 \right\rangle$ give excellent match with the results obtained from exact formulation for the wide range of inversion density. This isn't surprising because, as we have already noted, approximations to the form factors are independent of the cross-section shapes and the oxide-semiconductor interface is long enough to sustain the simplifications.

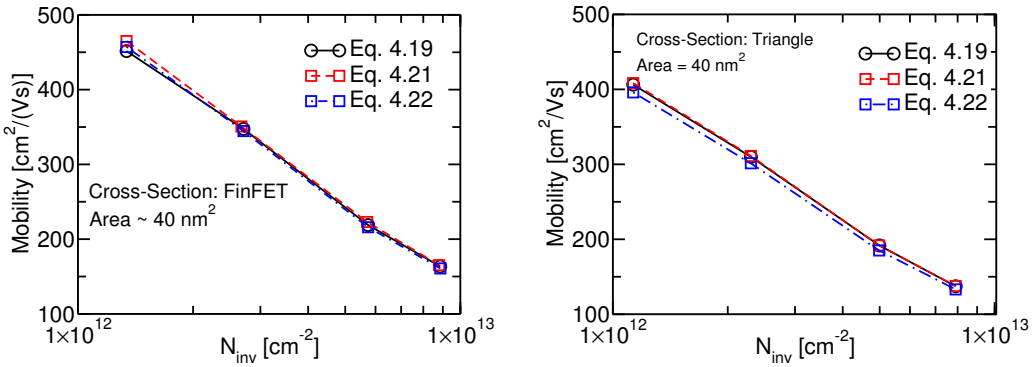


FIGURE 4.6: Plot of mobility versus inversion charge density for (a) FinFET and (b)GAA triangular cross-section at a fixed area = 40 nm^2.

4.4 Screening

The screening formulation developed for the planar device is general and it can be easily adapted to the surface roughness model for 3D transistors developed in this chapter. We recall the definition of the screened matrix element, $\mathcal{M}^{scr}_{nn'}(q_x)$

$$\mathcal{M}^{scr}_{nn'}(q_x) = \sum_{mm'} L^{nn'}_{mm'}(q_x)\mathcal{M}^{uns}_{mm'}(q_x). \tag{4.23}$$

where $\mathbf{L}$ is the inverse of the dielectric matrix, q_x is the difference between the initial and final wavevectors, (n, n') are the initial and final subband indices, $\mathcal{M}^{scr}_{nn'}(q_x)$ and $\mathcal{M}^{uns}_{nn'}(q_x)$ are the screened and unscreened matrix elements respectively. By multiplying with the complex conjugate and taking ensemble average we get

$$\begin{aligned}\left\langle \left|\mathcal{M}^{scr}_{n,n'}(q_x)\right|^2 \right\rangle = \sum_{m,m'} \left|L^{n,n'}_{m,m'}(q_x)\right|^2 \left\langle \left|\mathcal{M}^{uns}_{m,m'}(q_x)\right|^2 \right\rangle + \\ + \sum_{(m,m')\neq(p,p')} L^{n,n'}_{m,m'}(q_x) L^{n,n'}_{p,p'}(q_x)^{\dagger} \left\langle \mathcal{M}^{uns}_{m,m'}(q_x) \left(\mathcal{M}^{uns}_{p,p'}(q_x)\right)^{\dagger} \right\rangle\end{aligned} \tag{4.24}$$

which is formally identical to the corresponding derivation for planar FETs.

As we know that the matrix elements are written in terms of Fourier components (eq.4.9), we define the correlation function as

$$\begin{aligned}C^{m'p',l'g'}_{mp,lg}(\boldsymbol{\tau}) = \frac{1}{2\pi\, C_\Delta(0)\sqrt{1-C^2_{\Delta,N}(\boldsymbol{\tau})}} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} M^{m'l'}_{ml}[\Delta_1] \left(M^{p'g'}_{pg}[\Delta_2]\right)^{\dagger} \times \\ \times \exp\left[-\frac{\Delta_1^2+\Delta_2^2-2C_{\Delta,N}(\boldsymbol{\tau})\Delta_1\Delta_2}{2C_\Delta(0)(1-C^2_{\Delta,N}(\boldsymbol{\tau}))}\right] d\Delta_1\, d\Delta_2\end{aligned} \tag{4.25}$$

where (m, m', p, p') are subband indices and (l, l', g, g') are Fourier modes of $(\xi_m, \xi_{m'}, \xi_p, \xi_{p'})$ respectively and other symbols have same meaning as in unscreened case. Note that eq.4.25, unlike eq.4.14, also allows for different subband pairs (mm' and ll') as well. After calculating

the $C_{mp,lg}^{m'p',l'g'}(\boldsymbol{\tau})$ we can proceed as earlier to calculate the final ensemble averaged squared matrix element.

4.5 Some more comments on use of Fourier series

The starting point of the derivation of the matrix element is expansion of the wavefunction in Fourier components along the contours represented by $\mathcal{I}_\eta$. As expected the Fourier coefficients can be complex as well but the ensemble averaged squared matrix elements must be real valued. Here we discuss how despite of the complex starting point the end result is guaranteed to be real.

We start by noting that the envelope wavefunction from the Schrödinger equation is real. This forces the Fourier coefficients, $\phi_{n,l}(\eta)$, to satisfy

$$\phi_{n,l}(\eta) = [\phi_{n,-l}(\eta)]^\dagger \tag{4.26}$$

The complex nature of the Fourier coefficients, $\phi_{n,l}(\eta)$, forces the matrix elements defined in eq.4.8 to be also complex, but because of the property in 4.26 we can write

$$\begin{aligned} M_{n,-l}^{n',-l'}\left[\Delta(s,x)\right] &= \int_0^{\Delta(s,x)} D_\eta\, \phi_{n',-l'}^\dagger(\eta)\, \Phi_B\, \phi_{n,-l}(\eta)\, d\eta \\ &= \int_0^{\Delta(s,x)} D_\eta\, \phi_{n',l'}(\eta)\, \Phi_B\, \phi_{n,l}^\dagger(\eta)\, d\eta \\ &= \left[M_{n,l}^{n',l'}\left[\Delta(s,x)\right]\right]^\dagger \end{aligned} \tag{4.27}$$

This allows us to extend the reasoning to correlation function as

$$\begin{aligned} C_{n,-l,-g}^{n',-l',-g'}(\boldsymbol{\tau}) &= \left\langle M_{n,-l}^{n',-l'}\left[\Delta(s,x)\right]\left(M_{n,-g}^{n',-g'}\left[\Delta(s',x')\right]\right)^\dagger\right\rangle \\ &= \left\langle \left(M_{n,l}^{n',l'}\left[\Delta(s,x)\right]\right)^\dagger M_{n,g}^{n',g'}\left[\Delta(s',x')\right]\right\rangle \\ &= \left[C_{n,l,g}^{n',l',g'}(\boldsymbol{\tau})\right]^\dagger \end{aligned} \tag{4.28}$$

So that for the power spectrum we have

$$S_{n,-l,-g}^{n',-l',-g'}(q) = \left[S_{n,l,g}^{n',l',g'}(q)\right]^\dagger \tag{4.29}$$

Finally, the evenness of the $sinc(x)$ function along with eq.4.29 coupled with the summation over all the modes (l, l', g, g') forces the $\left\langle \left|\mathcal{M}_n^{n'}(q_x)\right|^2 \right\rangle$ to be real valued. This is particularly simple to see in the eq.4.22.

Chapter 5

Mobility analysis in 3D transistors

5.1 Introduction

At sub 7 nm nodes, multi-gate architectures will be preferred over their planar architectures because of their ability to keep the short channel effects under control and deliver better performance. With the increasing cost and complexity of fabrication it is necessary to perform a thorough analysis of different architectures and materials using the simulations. Moreover the tremendous improvements in the fabrication technology also provide us with ability to fabricate complex architectures which increases the architecture-space. Thus simply analyzing all the architectures with detailed analysis of analog and digital performance parameters would be counter productive. Also the simulation framework must be general enough to account for different architectures without any major simplification to the geometry, biasing schemes and approximations to materials parameters like isotropic mass approximations [42].

In this chapter, we will discuss in detail the complete simulation frame work that we developed to analyze the 3D device architectures (having fairly arbitrary cross-section) comprising of the Schrödinger, Poisson and Boltzmann Transport Equation solvers in which we have included the surface roughness model that we have developed in the previous chapter 4 and other relevant scattering mechanisms. We then analyze two experimental mobility data sets and extract reasonable values of the surface roughness parameters. We finally discus the dependence of mobility on cross-section shape and area in silicon and InAs based devices and highlight the importance of electrostatics in mobility.

5.2 Simulator description

In this section we will discuss the simulation frame work that will be used to analyze devices having fairly arbitrary cross-sections. We will first describe the Schrödinger-Poisson solver, which can account for arbitrary geometry and complete anisotropy in the effective masses and dielectric permittivity. We will then discuss the 1D Boltzmann Transport Equation solver and modifications in the discretization scheme for mobility calculations.

5.2.1 Schrödinger-Poisson solver

In order to accurately model the electrostatics and account for quantization effects, we solve the parabolic effective mass Schrödinger equation in the cross-section (y-z plane) given by

$$\left[-\frac{\hbar^2}{2}\nabla\cdot(W_{yz}\nabla)+U(\mathbf{r})\right]\xi_n(\mathbf{r})=\varepsilon_n^{(p)}\xi_n(\mathbf{r}) \tag{5.1}$$

where $\varepsilon_n^{(p)}$ is the n^{th} parabolic eigen energy, $\xi_n(\mathbf{r})$ is the envelope wavefunction in the cross-section of the device, $U(\mathbf{r})$ is the total potential energy and W_{yz} is the 2D inverse effective mass tensor in the device coordinate system which is evaluated by appropriate coordinate transformation (change of basis) of the effective masses in the ellipsoidal to device coordinate system [36]. For an arbitrary orientation (when the ellipsoidal, crystallographic and device coordinate axes are not aligned) W_{yz} is a full 2×2 tensor. The complete envelope wavefunction is taken to be of the form

$$\Psi_{n,k_x}(\mathbf{r},x)=\xi_n(\mathbf{r})\frac{e^{i\,k_x\,x}}{\sqrt{L_x}} \tag{5.2}$$

where we have assumed periodic boundary conditions along the transport direction, x and L_x is the normalization length while k_x is the discrete wavevector. We again emphasize here that we have neglected the phase term that arises due to the coupling between the confinement and transport direction (non zero values of w_{xy} and w_{xz}) [48, 95], because it is expected to have limited impact on the transport [22].

The parabolic eigen energies obtained from the solution of eq.5.1 are then corrected to account for non-parabolicity in the E-k relationship, which are very important in III-V semiconductors even though they are modest for silicon. The parameters for the non-parabolicity coefficient and effective masses have been benchmarked against $\mathbf{k}\cdot\mathbf{p}$ and empirical pseudopotential method [21]. For 1D electron gas, the nonparabolic eigen energies can be calculated as [42]

$$E_n=\langle U_n\rangle+\frac{\sqrt{1+4\alpha\left[\frac{\hbar^2}{2}\frac{k_x^2}{m_x}+\varepsilon_n^{(p)}-\langle U_n\rangle\right]}-1}{2\alpha} \tag{5.3}$$

where α is nonparabolicity coefficient, $\langle U_n\rangle$ is the expectation value of the total potential energy ($\iint\limits_{y\,z}|\xi_n(y,z)|^2\,U(y,z)\,dy\,dz$), m_x is the mass along the transport direction which is given by [95]

$$m_x=\frac{det(W_{yz})}{\omega_{t1}*\omega_{t2}*\omega_l} \tag{5.4}$$

where ω_{t1}, ω_{t2} and ω_l are the inverse effective masses in the ellipsoidal coordinate system along the principal transverse and longitudinal axes.

At the interface between two materials it is necessary that there is continuity of the wavefunction ($\xi_n(y,z)$) and of the flux of the wave function $\mathbf{W}_{yz}\nabla\xi_n$ [96]. The Schrödinger equation solver discussed above accounts for multi-valley, anisotropic masses and wavefunction penetration into the oxide.

The solution of the Schrödinger equation depends on the potential energy (the electrostatic part), which is calculated from the Poisson equation which can be written as

$$\nabla \cdot (\varepsilon(\mathbf{r})\nabla)\,\phi(\mathbf{r}) = q^2(N_D(\mathbf{r}) - N_A(\mathbf{r}) - n(\mathbf{r})) \tag{5.5}$$

where $\phi(\mathbf{r})$ is the electrostatic potential energy, $\boldsymbol{\varepsilon(\mathbf{r})}$ is the position dependent permittivity, $n(\mathbf{r})$ is the electron concentration, $N_D(\mathbf{r}) - N_A(\mathbf{r})$ is the net doping in the material. The concentration $n(\mathbf{r})$ depends on the Schrödinger equation and on the occupation of states through either Fermi-Dirac statistics or transport equation. This interdependence must be taken into account by solving the Schrödinger, Poisson and transport equation or Fermi-Dirac statistics in a self-consistent fashion. Because of the similarity between the differential operator of the Schrödinger and Poisson equation they can be written into similar abstract problems, which can then be solved using standard libraries like MKL or LAPACK [97].

We have used Discrete Geometric Approach (DGA) to solve the Schrödinger-Poisson problem because of its ability to account for arbitrary cross-section with ease. The main advantage of the DGA over other discretization schemes is that for a general polyhedral mesh, analytical closed form expression can be derived for the elements of the stiffness matrix, which is a very important advantage over other discretization methods [97]. In addition to this, the time independent Schrödinger equation can be written as a standard eigenvalue problem, unlike the finite element method where it gets transformed into a generalized eigenvalue problem which is more computationally challenging [98]. The boundary conditions are naturally satisfied by the DGA formulation [97].

5.2.2 Boltzmann transport equation

Boltzmann transport equation (BTE) is used to model the transport in the devices. Steady state BTE for a 1D system is a 2D equation for the nth subband given by [36]

$$v_{g,n}\frac{\partial f_n(x,k_x)}{\partial x} - \left(\frac{1}{\hbar}\frac{\partial E_{Tn}}{\partial x}\right)\frac{\partial f_n(x,k_x)}{\partial k_x} = S_n^{in}(k_x) - S_n^{out}(k_x) \tag{5.6}$$

where $v_{g,n}(=\frac{1}{\hbar}\frac{\partial E_{Tn}}{\partial k_x})$ is the group velocity for the electron, $f_n(x,k_x)$ is the occupation function for the electrons, E_{Tn} is the potential energy of the electron, sum of the kinetic energy and subband energy. $S_n^{in}(k_x)$ and $S_n^{out}(k_x)$ are the in-scattering and out-scattering rates respectively. Eq.5.6 is solved for the occupation function, which can then be used for calculating the different observable quantities like electron concentration and the current. The scattering rates are given by

$$S_n^{in}(k_x) = [1 - f_n(x,k_x)]\frac{L_x}{2\pi}\sum_{n'}\left[S_{nn'}(k_x,k_x')f_{n'}(x,k_x')dk_x'\right] \tag{5.7}$$

$$S_n^{out}(k_x) = f_n(x,k_x)\frac{L_x}{2\pi}\sum_{n'}\left[S_{n'n}(k_x',k_x)[1 - f_{n'}(x,k_x')]dk_x'\right] \tag{5.8}$$

where $S_{nn'}(k_x, k'_x)$ is the scattering rate from a state (n, k_x) to (n', k'_x), L_x is length of the device along the transport, and $[1-f_n(k_x,x)]$ and $[1-f'_n(k'_x,x)]$ enforce the Pauli's exclusion principle and make the BTE a nonlinear equation. A very convenient transformation of the integral from the k-space to the energy can be introduced by using the nonparabolic E-k relationship (eq.5.3).

Eq.5.6 can be simplified by performing a change of variables from $f_n(x,k_x)$ to $f_n(x,E_{Tn})$. According to the E-k relationship defined in eq.5.3, two values of k_x (positive and negative k_x corresponding to the forward and backward moving electrons) correspond to the same value of energy (because of squaring of k_x). Hence simply performing the change of variables from $f_n(x,k_x)$ to $f_n(x,E_{Tn})$ would be incomplete and we must add another index that would account for positive or negative k_x (physically speaking it would show the direction of movement of electrons). Thus in order to completely describe the occupation function we introduce f_n^+ and f_n^- occupations as

$$f_n(x,k_x) = \begin{cases} f_n^+(x,E_{Tn}) & \text{for } k_x\text{>0} \\ f_n^-(x,E_{Tn}) & \text{for } k_x\text{<0} \end{cases} \tag{5.9}$$

Using the occupation function, $f_n^{\pm}(x,E_{Tn})$ and the following identities

$$\frac{\partial f_n(x,k_x)}{\partial x} = \frac{\partial f_n^{\pm}(x,E_{Tn})}{\partial x} + \frac{\partial E_{Tn}}{\partial x}\frac{\partial f_n^{\pm}(x,E_{Tn})}{\partial E_{Tn}} \tag{5.10}$$

$$\frac{\partial f_n(x,k_x)}{\partial k_x} = \frac{\partial E_{Tn}}{\partial k_x}\frac{\partial f_n^{\pm}(x,E_{Tn})}{\partial E_{Tn}} \tag{5.11}$$

one can recast eq.5.6 into two first order ordinary differential equations [99, 100, 101]

$$v_{g,n}^+(x,E_n)\frac{df_n^+(x,E_{Tn})}{dx} = S_n^{in,+} - S_n^{out,+} \tag{5.12a}$$

$$v_{g,n}^-(x,E_n)\frac{df_n^-(x,E_{Tn})}{dx} = S_n^{in,-} - S_n^{out,-} \tag{5.12b}$$

where $v_{g,n}^+(x,E_{Tn})$, $v_{g,n}^-(x,E_{Tn})$ correspond to the group velocity of forward ($k_x > 0$) and backward ($k_x < 0$) moving electrons, $f_n^+(x,E_n)$, $f_n^-(x,E_{Tn})$ are the occupation functions of the electrons having positive and negative k_x respectively and the total energy, E_{Tn}, play the role of a parameter in eq.5.12.

Once occupation functions are known, macroscopic quantities like the volumetric electron concentration, $n_{inv}(\mathbf{r})$ and the current, I, in a cross-section can be calculated as

$$n_{inv}(\mathbf{r}) = \frac{1}{\pi}\sum_{\nu,n}\mu_\nu\left|\xi_{\nu,n}(\mathbf{r})\right|^2\left(\int_{\varepsilon_{\nu,n}}^{+\infty}\frac{f_{\nu,n}^-(E_{\nu,n})}{\hbar\left|v_{g,\nu,n}^+(E_{\nu,n})\right|}dE_{\nu,n} + \int_{\varepsilon_{\nu,n}}^{+\infty}\frac{f_{\nu,n}^+(E_{\nu,n})}{\hbar\left|v_{g,\nu,n}^-(E_{\nu,n})\right|}dE_{\nu,n}\right) \tag{5.13}$$

$$I = \frac{e}{\hbar\pi}\sum_{\nu,n}\mu_\nu\left(\int_{\varepsilon_{\nu,n}}^{+\infty}f_{\nu,n}^+(E_{\nu,n})dE_{\nu,n} - \int_{\varepsilon_{\nu,n}}^{+\infty}f_{\nu,n}^-(E_{\nu,n})dE_{\nu,n}\right) \tag{5.14}$$

where μ_ν denotes the multiplicity of a given valley ν and the nonparabolic subband energy $\varepsilon_{\nu,n}$ is obtained by using Eq.5.3 and setting $k_x = 0$. Further details related to BTE solver can

be found in [101]

In the above discussion, we have implicitly assumed that energy broadening due to collisions is and to the coupling to the source-drain reservoirs negligible. Since BTE is essentially a semi-classical transport equation we have assumed that both the position and momentum of the electron can be specified simultaneously with arbitrary precision. Furthermore by writing the right hand side of the BTE as in eq.5.7 and eq.5.8 we have assumed that the scattering process is instantaneous in time and completely localized in real space [52].

Mobility calculations

In order to extract the correct values of the mobility, the length of the device must be much longer than the relaxation lengths. Mobility by definition are near equilibrium performance parameters and hence the simulation conditions must be such that there is small uniform electric field along the transport direction. Thus in order to mimic these conditions we solve the Schrödinger and Poisson equations in the device cross-section assuming equilibrium. In order to discretize the BTE, we further assume a uniform mesh along the transport and the discretization in energy is taken as $\Delta E = eF\Delta x$. The BTE in the device cross-section is discretized using an upwind/downwind scheme which allows us to write the LHS of the BTE as

$$\left.\frac{df^+}{dx}\right|_{x_s} = \frac{f^+_{s,n,j} - f^+_{s-1,n,j}}{x_s - x_{s-1}} \tag{5.15a}$$

$$\left.\frac{df^-}{dx}\right|_{x_s} = \frac{f^-_{s+1,n,j} - f^-_{s,n,j}}{x_{s+1} - x_s}. \tag{5.15b}$$

where $f^+_{s,n,j}$ is the occupation function f^+_n at a given position s and energy j in the discretized domain (see Fig.5.1). Since mobility is a characteristic of uniform transport, the occupation functions having same kinetic energy have the same occupation, which allows us to map the occupation functions outside the simulation domain on to the same cross-section as shown in the Fig.5.1. This, in turn leads to discretized equations

$$\left.\frac{df^+}{dx}\right|_{x_s} = \frac{f^+_{s,n,j} - f^+_{s,n,j-1}}{\Delta x} \tag{5.16a}$$

$$\left.\frac{df^-}{dx}\right|_{x_s} = \frac{f^-_{s,n,j+1} - f^-_{s,n,j}}{\Delta x}. \tag{5.16b}$$

With this approach, we are effectively simulating an infinitely long device, which is a requirement of the mobility calculation. Straightforward simulation of the very long device would be much more computationally expensive. Mobility is then calculated from its definition as the ratio of average velocity to electric field. As the BTE is a nonlinear equation as discussed earlier, it is solved with the Newton-Raphson method with Fermi-Dirac occupation function as an initial guess.

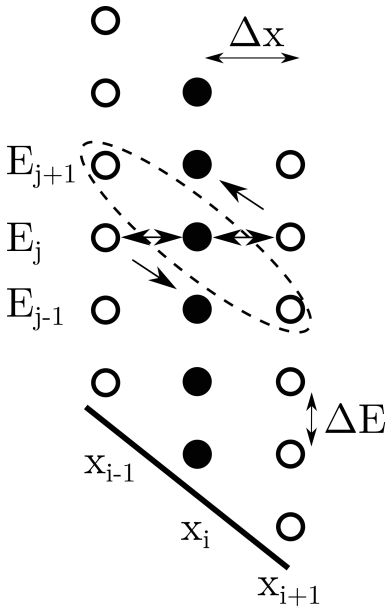


FIGURE 5.1: Illustration of the discretized (x,E_n) space. Filled symbols are the actual unknowns while the unknowns from the neighboring section are open symbols. The mapping the unknowns from the adjacent sections are illustrated with the arrows from the sections $x_i \pm 1$ to x_i. Dashed oval clubs the points having the same occupation function value in the uniform transport as they have the same kinetic energy. The discretization in the energy and transport direction are linked by $\Delta E = eF\Delta x$.

5.3 Comparison with experimental data

In this section, we will analyze two experimental data sets and examine the surface roughness parameters necessary to reproduce them. In the simulations we have considered non polar acoustic and optical phonons [36] and polar optical phonons [102] in addition to surface roughness scattering (as discussed in the previous chapter). The scattering due to different mechanisms are accounted for without any simplifying assumption of momentum relaxation time [42, 47] or linearization of the BTE [51, 103]. The phonon scattering parameters are same as listed in [36]. The material parameters used in the simulations are listed in Table.5.1 and Table.5.2.

Material	Valley	**effective mass**		α	χ
		m_l [m_0]	m_t [m_0]	[eV^{-1}]	[eV]
InAs	Γ[46]	0.026	0.026	2.5	4.9
	L[104]	1.565	0.124	0.45	

TABLE 5.1: Bulk longitudinal and transverse effective mass, non-parabolicity coefficient, α and electron affinity, χ for InAs. The energy difference between the L and Γ conduction band minima in InAs is taken to be 0.716 eV [69].

5.3.1 Gate-all-around architecture

The first experimental data set that we consider in this study was published in [109]. The structure is a gate all around InAs cylindrical nanowire FET with gate oxide of Zr_2O_3. The nanowire has a diameter of 15 nm. Fig.5.2 plots the simulated and experimental mobility

SiO_2			ZrO_2			In_2O_3		
$m(\Gamma)$ [105] $[m_0]$	χ [105] $[eV]$	κ [106] $[\varepsilon_0]$	$m(\Gamma)$ [107] $[m_0]$	χ [107] $[eV]$	κ [106] $[\varepsilon_0]$	$m(\Gamma)$ [47] $[m_0]$	χ [108] $[eV]$	κ [47] $[\varepsilon_0]$
0.5	0.95	3.9	0.30	3.0	24	0.30	3.5	15

TABLE 5.2: Effective mass, $m(\Gamma)$, electron affinity, χ, and relative permittivity, κ, for different oxides used in the analysis of the experimental.

versus inversion charge density (linear electron density is normalized by the circumference of the nanowire). Impact of different scattering mechanisms (acoustic, polar and non polar optical scattering, surface roughness scattering and Coulomb scattering due to fixed oxide charges) on the mobility is also shown in the Fig.5.2. The phonon limited mobility obtained from the simulations is much larger than the experimental values (bulk phonon limited mobility at room temperature for InAs is about 20,000 cm^2/V-s [110]). However, when the surface roughness scattering with $\Delta_{rms} = 0.17$ nm and $\Lambda = 1.4$ nm is included in the calculations we obtain a good match with the experiments at high inversion charge density. These surface roughness values are consistent with the ones for the bulk-like planar InAs case as shown in chapter 2. However, at low inversion density, there is a significant difference between the experimental and simulated mobilities. In order to further study this we included the Coulomb scattering due to fixed charges (N_{fix}) in the simulations with $N_{fix} = 4 \times 10^{12}$ cm^{-2} and $N_{fix} = 8 \times 10^{12}$ cm^{-2}. We observe that by increasing the fixed charge density the match between the simulations and experiments improves, thus suggesting that the Coulomb scattering is important in this range of N_{inv}. Since such a large N_{fix} would induce a significant shift in the threshold voltage which would be inconsistent with the experiments, this route was not pursued further [111].

In addition, fig.5.2 shows that carrier screening plays a modest role in the simulated InAs GAA MOSFET. This can be explained by noting that, in a degenerate 1D electron gas, intra-subband transitions at energies close to the Fermi level result in a very large exchanged wavevector $q=(k'_x - k_x)$, which reduces drastically the effect of screening [36, 52]. Inter-subband transitions may have a significantly smaller q, but inter-subband screening is weaker [36, 52].

5.3.2 Back gate biased transistor

We now compare the mobility for a backgated InAs nanowire FET as measured in [112]. The structure used in simulations consists of thick back oxide and InAs is covered by 2.2 nm of native oxide (taken to be In_2O_3). We also consider air surrounding the nanowire to mimic the experimental conditions and to account for parasitics (Fig.5.3(a)). The asymmetric backgate biasing scheme results in lack of radial symmetry in the potential profile and consequently there is no radial symmetry in the wavefunction. Because of the potential profile the wavefunction offsets towards the BOX. This is shown in Fig.5.3(b), which plots the $|\xi_n(\mathbf{r})|^2$ for the lowest subband. This is completely different for the gate all around biasing scheme which has a radial symmetry, whose structure and $|\xi_n(\mathbf{r})|^2$ for the lowest subband is shown in Fig.5.3(c) and (d), respectively. It is imperative that we account for the lack of

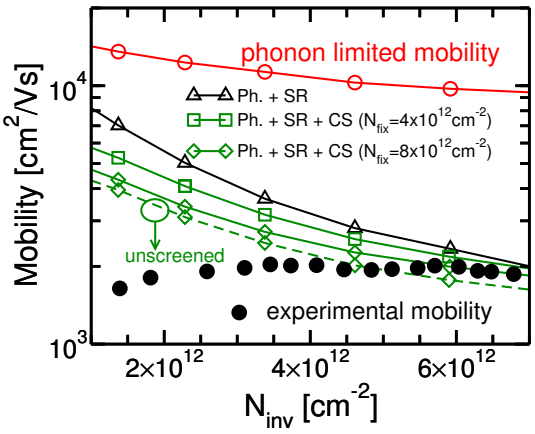


FIGURE 5.2: Mobility versus inversion charge density for a circular (R=7.5 nm) gate-all-around nanowire. A good match with the experiments is obtained at higher inversion density with Δ_{rms}=0.17 mn and Λ=1.4 nm. The results show that impact of screening is modest. The material parameters in the simulations are listed in Table.5.1 and 5.2. Experimental data is taken from [109].

symmetry in the wavefunctions as wavefunctions lie at the heart of all the scattering calculations.

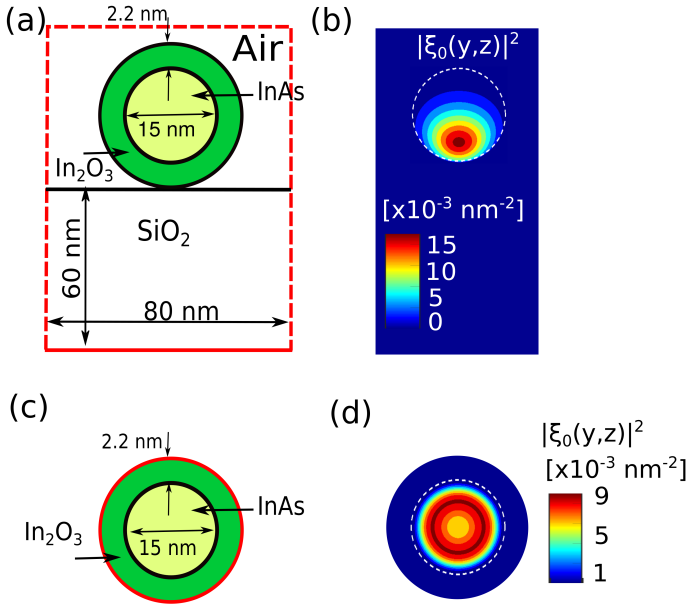


FIGURE 5.3: (a) Simulation domain for the back-gated nanowire. Homogeneous Neumann boundary conditions are imposed in the external domain with dashed lines, whereas Dirichlet boundary conditions are imposed at the bottom of the SiO_2 film (solid line); (b) Squared magnitude of the lowest subband wavefunction at N_{inv}≈1×10^7 cm^{-1} for back-gated nanowire. (c) Simulation domain for the same nanowire with a gate-all-around biasing condition. and (d) squared magnitude of the lowest subband.

The mobility data in [112] is given as a function of over drive voltage (see supporting information in [112]) which brings in the uncertainty associated with the value of the threshold voltage. To put the experimental data and simulation results on equal footing we converted $(V_{GS} - V_T)$ into inversion charge density by extracting the capacitance for radius=7.5 nm. In the extraction procedure, the capacitance data as function of nanowire radius (Fig.4 of [112])

was fitted with a straight line and then was extrapolated to $R = 7.5\ nm$ (shown in Fig.5.4), which results in $C_G \approx 43aF/\mu m$.

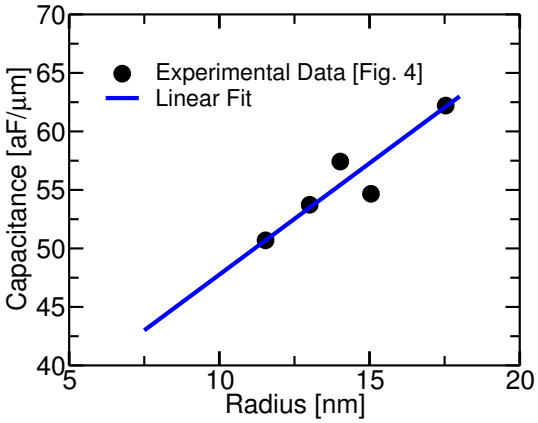


FIGURE 5.4: Linear fit to the experimental values capacitance versus the radius of nanowire.

Fig.5.5 plots the experimental and simulated mobility versus the linear electron density. A good agreement with the experimental mobility is achieved with using the surface roughness parameters as Δ_{rms}=0.55 nm and Λ=2.0 nm. It is interesting to note that Δ_{rms}=0.55 nm has been experimentally extracted for III-V quantum well MOSFET [70]. If a linear model with GAA biasing scheme is used to match the same experimental data, this results in Δ_{rms}=1.2 nm, which is a very large value. The ability of the nonlinear model to reproduce the experimental data with credible values of the surface roughness parameters as compared to linear model encourages us to question the fundamental assumption of linearization of the wavefunction (and electrostatic potential) in the linear surface roughness models. Furthermore we simulated the same nanowire but with the gate-all-around biasing (Fig.5.3(c)) with the same values of the surface roughness parameters. The mobility for such a configuration was much larger than in the backgated configuration. This can be explained by the fact that in the backgated configuration the wavefunction near oxide is larger than in the GAA configuration (Fig.5.3(b) and (d)) which results in higher scattering rates. This observation reiterates the importance of the model and simulation methodology to account for biasing schemes and arbitrary cross-section.

5.4 Mobility analysis for different cross-section shapes and materials

Fig.5.6 shows the dependence of mobility on linear inversion charge density and cross-section area. Results correspond to a fixed N_{inv} for different cross-section shapes for relaxed silicon based device (transport orientation is $\langle 100 \rangle$). In these analyses we have considered acoustic and optical phonons with parameters from [12] and surface roughness scattering with Δ_{rms}=0.21 nm and Λ=1.4 nm, which is same as needed to match the universal mobility curves in Si bulk MOSFETs [113]. The mobility simulations suggest that circular cross-section has a modest advantage over other architectures. In all the architectures, the mobility

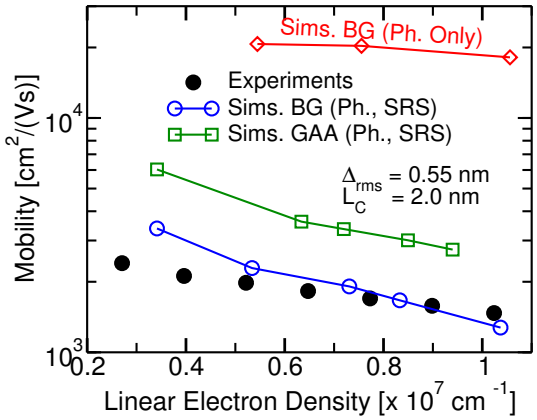


FIGURE 5.5: Comparison between the experimental and simulated mobility values for a backgated InAs nanowire (radius=7.5 nm) FET. When the nanowire is simulated with the gate all around biasing scheme results in overestimation of the mobility values.

reduces monotonically with the reduction in the area due to increase in the surface roughness scattering. Fig.5.6(b) also shows the phonon limited mobility which is much larger than the total mobility thus suggesting that surface roughness is the dominant scattering mechanism.

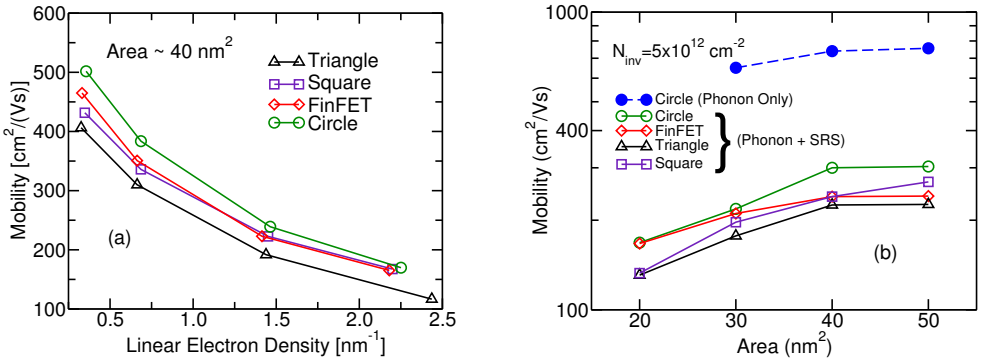


FIGURE 5.6: (a) Impact of linear inversion charge density on mobility for gate all around triangle, square, circle and trigated FinFET devices for fixed area, A≈40 nm^2. (b) Impact of the cross-section area on mobility for different cross-section shapes at a fixed inversion charge density (normalization of the inversion charge density is done with the gate perimeter). All the device have similar mobility values across the range of area and inversion density considered.

We further analyzed the devices with having the same architecture as in Fig.5.6, but with InAs channel, HfO_2 gate oxide and $\langle 100 \rangle$ as the transport orientation. We have used polar optical, non polar acoustic and optical phonons with scattering parameters from [36], and surface roughness roughness scattering. The surface roughness parameters used are same as in case of silicon to have a fair comparison between different materials and we have also extracted similar values for the latest III-V mobility data as shown earlier (in Chapter3).

Fig.5.7(a) plots the mobility for different architectures at N_{inv}=2×10^{12} cm^{-2}. Fig.5.7(a) also shows that for cross-section area less than 20 nm^2 mobility decreases with a trend of A^3 which is consistent with the Prange-Nee picture in the quantum well FETs [40]. Such a trend has also been shown in the GAA circular nanowire ([42]). This reduction in the mobility with the cross-section is because with the reduction in the area the wavefunction gets increasingly

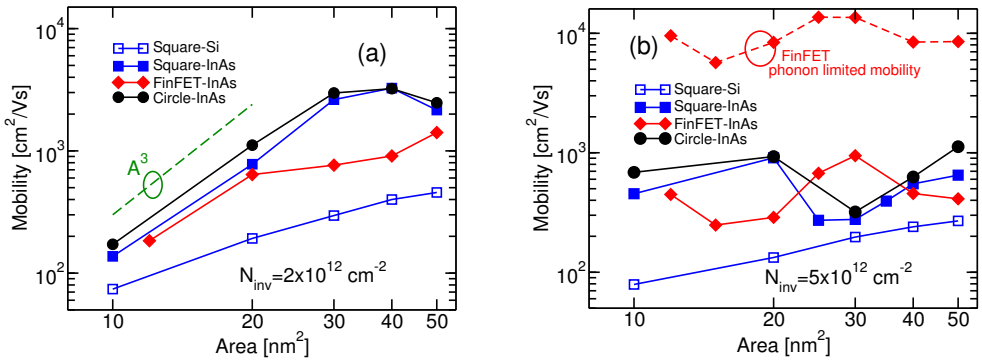


FIGURE 5.7: Plot of mobility versus area for different cross-section shapes with InAs channel for (a) $N_{inv} = 2 \times 10^{12} cm^{-2}$ and (b)$N_{inv} = 5 \times 10^{12} cm^{-2}$. Linear to areal charge density conversion is done with the gate perimeter. At lower N_{inv} the mobility follows a trend of A^3 at small cross-section areas however at higher N_{inv} mobility oscillates depending on the position of the subbands with respect to fermi level Fig.5.8. We have also included mobility curves for silicon for comparison.

squeezed towards the oxide and thus resulting in an increase in interaction between oxide-semiconductor interface and electrons thus resulting in an increase in the surface roughness scattering.

Fig.5.7(b) shows the mobility for different cross-sections at fixed N_{inv}=5×10^{12} cm^{-2}. Here we see that unlike the case of N_{inv}=2×10^{12} cm^{-2} there is no clear trend and different cross-sections tend to behave quite differently in terms of mobility values. These mobility fluctuations are characteristic of a typical 1D system (density of states has a peak at subband minima). These mobility results are analyzed further in Fig.5.8 using case of FinFET. We notice that as the cross-section area reduces the mobility first increases and then reduces until it is minimum at A=15 nm^2, as shown in Fig.5.8(a). As we move from A=50 nm^2 to A=30 nm^2 the third and the fourth subbands which lie just below the Fermi level (Energy=0 eV) move higher in energy and eventually cross-over the Fermi level. When the subbands come near the Fermi level, there is a sharp increase in the scattering which decays away as the subbands move away from the Fermi level. In this region, the subbands from the L-valley are significantly above the Fermi level and are expected to play a negligible role in transport. Thus the mobility increases and is maximum at A=30 nm^2 where there are no subbands near the Fermi level, as the mobility is essentially governed by the electrons around Fermi level (due the product of matrix element and $f(E)(1-f(E))$). As we continue to reduce the area further, 2nd subband in the Γ-valley moves closer to Fermi level. For A=15 nm^2 it is necessary to include the L-valley into the simulations as they modify the subband energy level because of redistribution of the electron concentration in the subbands. This modification in the energy level is particularly important for the A= 15 nm^2 as the second subband in the Γ valley is extremely close to the Fermi level and hence the L-valleys are taken into account for A=15 nm^2 only. Further as we move from A=15 nm^2 to A=12 nm^2 the second subband moves above the Fermi level (thus reducing the product of scattering rate and $f(E)(1-f(E))$), so that the mobility increases. Similar modulation of mobility due to the position of subbands is observed for other architectures as well. We reiterate that the L-valleys may not play a direct role in scattering but by affecting the subband energy levels

if they lie in close proximity to the Fermi level.

Such mobility fluctuations have been observed in experiments in p-type silicon nanowires at low temperatures [114] and mobility simulations of InAs nanowires [47] as well.

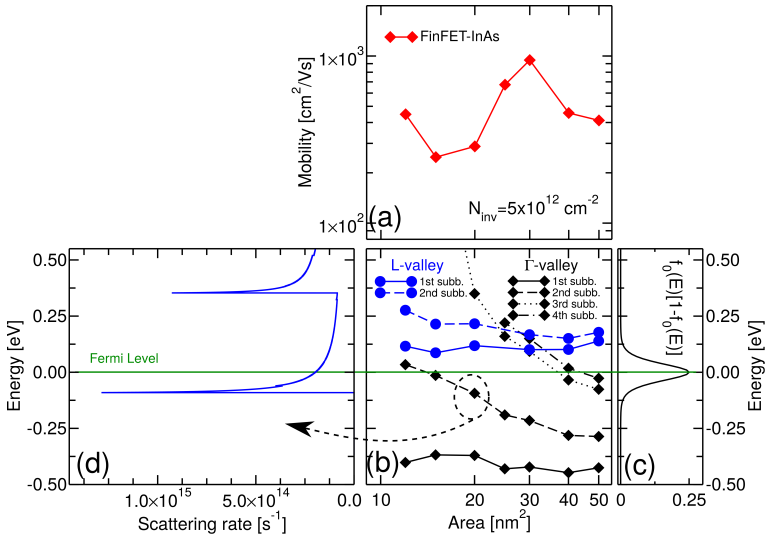


FIGURE 5.8: (a) Plot of the mobility versus area for FinFET at $N_{inv} = 5 \times 10^{12}$ cm^{-2}. (b) Subband energies with reference to the Fermi level (0 eV) for the lowest subbands in the Γ and L valley. (c) plot of $f(E)\,[1 - f(E)]$ assuming equilibrium occupation function. (d) Total scattering rate for the second subband in device with $A = 20$ nm^2. As the subbands move into the window having non negligible value of $f(E)\,[1 - f(E)]$ there is sharp increase in the intersuband scattering which reduces the mobility.

Chapter 6

Analysis of different architectures for 3D Devices transistors

6.1 Introduction

The remarkable progress in the fabrication technology allows us to fabricate complex device architectures to continue to improve device performance with each technological generation for FETs. However, the cost and complexity of the fabrication necessitate that a good preliminary analysis of the device architectures is made using simulations which are relatively cheap and less time-consuming to weed out architectures that do not perform up to the mark. The computational tools that must be used to analyze the devices should not rely on the moment based transport models (drift-diffusion or hydrodynamic model) because of their low field assumptions are not justified in the far from equilibrium transport regime enforced in nanoscale FETs [59]. Moreover, use of a large number of fitting parameters in these models mask away the physical insights in the transport phenomenon. In addition, simple ballistic or even phonon limited analysis using accurate transport methods like Boltzmann transport equation can be misleading in the analysis of the advantages of an architecture over another.

When devices are in on-state the electrons are far away from their subband edge, especially near the channel-drain region. The electrons dissipate the excess energy through the inelastic phonon scattering. This electron-phonon interaction can increase the lattice temperature (known as self-heating) in a transistor. A thorough analysis of the impact of the self-heating for the planar devices has shown that the self-heating causes a non-negligible degradation in the device performance [115]. Consequently we can expect that the degradation must be even more significant for the 3D devices because of the reduced available paths for heat to escape the nanowire. This expectation has been corroborated by simulation studies for circular gate-all-around nanowire FET [116]. Furthermore, it is a well-known fact that the reduced thickness of semiconductor has a strong negative impact on the thermal conductivity of the nanowire thus increasing the self-heating [117, 118, 119]. This impact is compounded by a strong dependence of thermal conductivity on the surface roughness as well [118, 119]. Thus while analyzing different architectures for their relative performances, self-heating must also be taken into account as different architectures can have different self-heating impact and it can even be a performance bottleneck.

Even though accurate transport methodologies have existed for a very long time their usage and application for real devices has been quite limited even because of their large computational burden [120, 121, 122]. Despite the progress in the availability of computational resources, the use of accurate models (transport or scattering or electrostatic related) has been limited and till today calibrated drift-diffusion based methodologies or ballistic analysis have been mainly used to benchmark different 3D architectures and materials. The use of these approximated methods is expected and partly justified because of the extremely large computational time involved in using sophisticated transport methodologies and accurate scattering models [123]. Thus in order to be able to use the sophisticated formulations, there is a need to develop algorithms or iteration schemes able to reduce the computational burden.

In this chapter, we will extend the simulation framework that we had discussed in the previous chapter and also discuss the inclusion of the series resistance and self-heating effects in the simulator. We will then introduce the optimization schemes that we have developed and implemented in the simulator to reduce the computational load. Finally, we will perform a comprehensive performance analysis of different architectures and of the impact of self-heating on the device performance.

6.2 Simulation frame work

For complete device analysis the 3D device is divided into a series 2D sections and the spacing between the 2D sections determines the mesh along the transport direction. In each section, the parabolic band effective mass Schrödinger equation is solved in all the 2D sections and the Poisson equation is solved in the entire 3D device, unlike the mobility calculations in which only one section is analyzed with periodic boundary conditions. Our Schrödinger-Poisson solver uses a Discrete Geometric Approach (DGA) for discretization of the equations having an improved numerical efficiency compared to finite element method [97]. DGA allows us to account for arbitrary cross-section shapes and have a standard eigenvalue problem unlike Finite Element Method [98]. The parabolic eigen energies are then corrected to account for the non-parabolicity in the E-k relationship [42]. The resultant subband energy (both parabolic and nonparabolic) profile along the transport direction, the wavefunctions and potential profile are then passed to the Boltzmann transport equation solver. The Boltzmann transport equation is solved subject to the boundary condition that the incoming electrons have Fermi-Dirac distributions. Explicit equations and their discretization are as discussed in the Chapter 5. These equations are solved in a Gummel iterative way until a self-consistent solution is reached.

In the simulations we have included the nonlinear surface roughness scattering model that was developed in Chapter 4. In addition to this we have also included nonpolar acoustic and optical phonons and for III-V materials polar optical phonons [36, 102]. The use of nonlinear surface roughness scattering model allows us to use realistic values of the surface morphology parameters as we have demonstrated with mobility analysis in chapter 5. Scattering rates are included in the Boltzmann solver without any simplifications (e.g.

relaxation time approximation). We have neglected the screening in these simulations to reduce the computational burden, because we have verified that screening plays a relatively modest role in the 1D systems of our interest [94].

6.2.1 Series resistance

ITRS predicts that series resistance can degrade the on-current by up to 50% in high-performance logic and more than 50% in low power devices [25]. Moreover, the impact of the series resistance is bias dependent and thus it is difficult to correct in a simple way. In this work, the external series resistance ($R_{ext} = R_S + R_D$) is considered as a lumped resistor connected to the source and drain terminals. The impact of the series resistance is accounted with a self-consistent method by correcting the drain bias as

$$V_{DS} = V_{DD} - I^n_{DS} R_{ext} \tag{6.1}$$

where V_{DD} is the applied bias, V_{DS} is drop across the nanowire and I^n_{DS} is the drain current in the n^{th} iteration. Because of non-monotonic convergence it is necessary to damp the potential drop across the R_{ext} to avoid large oscillations. Hence we have implemented an under-relaxation scheme for calculating the I^n_{DS} [124]

$$I^n_{DS} = \alpha I^n_{DS} + (1 - \alpha) I^{n-1}_{DS} \tag{6.2}$$

where α is the relaxation coefficient between 0 and 1. The parameter α must be chosen with trial and error. It is easy to see that after the simulation has converged $I^n_{DS} \approx I^{n-1}_{DS}$ and hence the drop across the external resistance is correctly accounted without any artifact from the relaxation scheme. Because of the drop across the contact resistance, the Fermi level at the drain floats up while at the source it moves down and thus there is a reduction in the voltage drop across the nanowire with respect to the externally applied V_{DS}. In addition to the reduction in the drain bias, the barrier between the source and channel also increases, thus inducing a reduction of the current flowing through device.

6.2.2 Self-heating methodology

It has been shown in the literature that the self-heating can cause significant degradation in the performance of the device in terms of I_{on} degradation [116]. A number of methodologies exist in the literature to account for self-heating effects like solving coupled Boltzmann Transport equations for phonons and electrons as well as a fully coupled quantum electron phonon transport etc [115, 116, 119, 125].

In this work we have used a simple yet physically transparent methodology to account for an increase in temperature due to inelastic scattering as described in [119]. A very important advantage of this methodology is that it is computationally very efficient as it requires only the solution of a linear 1D differential equation. An additional advantage of this formalism is that it provides a seamless integration with the overall simulation framework developed so far. In this approach, 1D Fourier heat transport equation is solved along the

transport direction to calculate the lattice temperature profile, $T(x)$

$$-A\kappa_{sc}\frac{d^2}{dx^2}T(x)+\frac{\kappa_{BOX}}{T_{BOX}}\left[T(x)-T_0\right]W=H_{1D}(x) \tag{6.3}$$

where A is the semiconductor cross-section area, κ_{sc} is the thermal conductivity of the semiconductor, κ_{BOX} is the thermal conductivity of the buried oxide (BOX), T_{BOX} is the thickness of the BOX, T_0 is the temperature at the bottom of the BOX (taken to be 300 K), W is the width of the device and H_{1D} is the heat generated in the nanowire because of the inelastic phonon scattering. In this work, we have assumed that the thermal conductivity is independent of the position and temperature. The heat conduction through the buried oxide is accounted for with a simple phenomenological heat transfer rate, namely the second term on the left-hand side of the Eq.6.3. The heat transfer through the thin gate oxide is neglected because of the reduced thermal conductivity as compared to bulk oxides [126]. Eq.6.3 is solved subject to Dirichlet boundary conditions at the source and drain contacts and it has also a heat sink at the bottom of the device (below the BOX).

In order to calculate the $H_{1D}(x)$, we first define a spectral current from section i to $i\pm1$ (the current is considered to be positive when the carriers are moving along positive x), $\mathcal{I}_{i\rightarrow i\pm1}(E)$, as

$$I_{i\rightarrow i\pm1}=\int_{-\infty}^{+\infty}dE\ \ \mathcal{I}_{i\rightarrow i\pm1}(E) \tag{6.4}$$

where $I_{i\rightarrow i\pm1}$ is the total current moving from section i to $i\pm1$. We recall that the spectral power of an electron moving from section $i\pm1$ to i can be written as (similar to the power law in electrodynamics)

$$P_{i\rightarrow i\pm1}(E)=\frac{E}{-e}\mathcal{I}_{i\rightarrow i\pm1}(E) \tag{6.5}$$

where e is the elementary electron charge and E is the total energy. However when electrons can exchange energy with the crystal lattice (through inelastic phonon scattering), then the amount of energy transfered to the system is just the difference between the incoming and outgoing electron energy. Thus the amount of heat generated from the imbalance in the spectral current in the section i can be written as

$$H_{1D}(x_i)dx=\int_{-\infty}^{+\infty}dE\left[P_{i\rightarrow i+1}(E)+P_{i\rightarrow i-1}(E)\right]=\int_{-\infty}^{+\infty}dE\frac{E}{-e}\left[\mathcal{I}_{i\rightarrow i+1}(E)+\mathcal{I}_{i\rightarrow i-1}(E)\right] \tag{6.6}$$

Eq.6.6 is intrinsically linked to the transport (scattering) mechanism, hence it provides an explicit way to calculate the heat generation. In the absence of inelastic scattering the energy of an electron is conserved when the electron moves through the device, consequently the spectral current is conserved.

$$\mathcal{I}_{i\rightarrow i+1}(E)+\mathcal{I}_{i\rightarrow i-1}(E)=0 \tag{6.7}$$

Eq.6.3 is solved self-consistently with the Schrödinger, Poisson and Boltzmann transport

equation. In the self-consistent scheme, the temperature enters into the phonon scattering calculation and thus enters also the BTE. The complete simulation flow for one bias point is shown in Fig.6.1

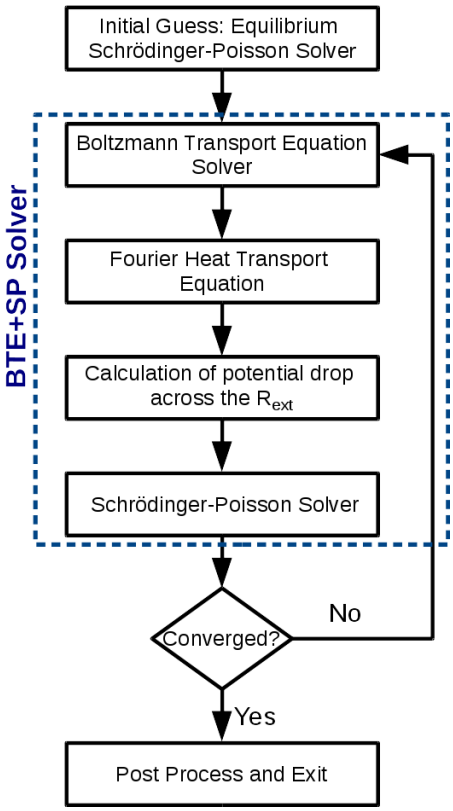


FIGURE 6.1: Standard program flow for one bias point. It must be noted in this methodology the scattering calculations are included in the simulations from the first iteration.

6.2.3 Optimization schemes

In order to be able to analyze the device with practical sizes and relevance, we have developed several optimization schemes. We will discuss them in detail below.

Schrödinger solver

The mesh spacing along the transport direction is dictated by the convergence of the Boltzmann Transport equation. This implies that the Schrödinger equation too must be solved in all the slices and thus its solution becomes a major computational challenge in the electrostatics module [127]. Simply using a coarse mesh for the Schrödinger solver isn't a correct solution because it would result in the incorrect location of top-of-the-barrier which is important for transport (especially in the case of ballistic transport). We deal with this problem by solving the Schrödinger equation in a number of sections that is dynamically adjusted depending on the variation of the potential profile. The variation of the potential is measured by the absolute value of the maximum change in the potential between the two slices. The parabolic and non-parabolic eigen energies between the two slices in which the

Schrödinger equation is solved are calculated by a linear interpolation (Schrödinger equation is always solved in the first and last slice) while the wavefunction is kept the same because wavefunctions are expected not to change significantly with the small variation in the potential. Direct application of this approach can still lead to incorrect top-of-the-barrier (see Fig.6.2) as the variation of the potential near the top-of-the-barrier is small and so the resulting mesh in the top-of-the-barrier region can be quite coarse. Thus we modify this approach by solving the Schrödinger equation in all the slices in the first iteration and then use the adaptive scheme except for a small region around the top-of-the-barrier. It is important to note that the top-of-the-barrier is determined in the beginning of each iteration as the top-of-the-barrier can change from one iteration to another.

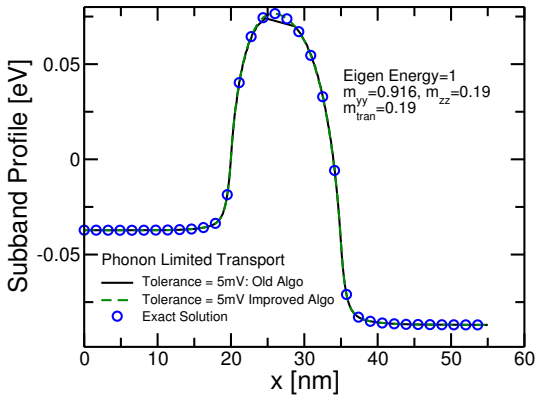


FIGURE 6.2: Illustrates the lowest subband profile by solving Schrödinger equation (empty symbols) in all the slices, with simple (solid line) and modified (dashed line) optimized and for circular gate-all-around nanowire.

This scheme is adaptive in the sense that in the source and drain regions the potential profile is essentially flat and thus the difference between the two successive slices in which Schrödinger has to be solved is much larger as compared to the channel region where the potential changes much faster. This optimization scheme besides reducing the number of times during a complete simulation when the Schrödinger equation is solved, it also reduces the number of times the ensemble averaged squared matrix elements and form factor in polar optical phonon are calculated, which are among the computationally heaviest process in the entire simulation.

Optimized iteration scheme

Even with the adaptive solution of the Schrödinger equation the computational burden remains extremely high, due to the nonlinear surface roughness scattering model and form factor calculations for polar optical phonons in III-V based devices. In addition to these scattering related calculations, another important time-consuming part is the setting up of the scattering rate matrices which couples a given state of an electron (specified by eigenenergy and wavevector) to other states. These procedures lie at the very heart of the simulation methodology and any approximations to these formulations may affect the accuracy of the simulations. Thus a further reduction of the simulation time calls for an overhaul of the iteration scheme.

Fig.6.3 shows the novel iteration scheme which brings the simulation time down to practical limits. This procedure progressively refines the initial guess by switching on increasingly heavier scattering mechanism. Here we start with an initial guess from equilibrium Schrödinger-Poisson solver. This crude initial guess is then further refined by performing only the ballistic simulations until the error reduces below a tolerance criterion ("StartPhError"). This is followed by further improving the potential profile by turning on the phonon scattering (which is computationally less heavy than surface roughness scattering). When the error is reduced below a tolerance level ("StartSRError") then surface roughness scattering is activated, which is the heaviest is included in the simulation. The simulation is terminated when the error reduces below the actual error BTE-SP tolerance criterion. In this iteration scheme, the computationally heaviest scattering schemes are turned-on only in the last few iterations.

The starting points of phonon and surface roughness scattering must not be set to BTE-SP tolerance criterion. This would be counterproductive as the simulation would first converge under the ballistic case and then it would turn on the phonon scattering and again converge till the BTE-SP tolerance criterion and then it would turn on the all the scattering mechanism and then it would finally converge. The values of StartPhError and StartSRError must be chosen on the basis of trial and error and it can even depend on biasing conditions as surface roughness scattering is expected to be strong in the on-state. In our experience, setting StartPhError to be an order of magnitude more than BTE-SP tolerance criterion and StartSRError to be 5 times BTE-SP tolerance criterion is good choice.

6.3 Performance analysis

In this section we will analyze different architectures having channel length L_G=14 nm. Device width (W) and the gate oxide thickness (T_{ox}) were adjusted such that the subthreshold swing for FinFET was about 75 mV/dec as specified by the ITRS. Here the FinFET was chosen as the base case as it is the current industry standard. The schematics of different architectures considered in this study are shown in Fig.6.4. The dimensions of the other architectures (Stacked nanowires, circular GAA, and square GAA nanowire) are chosen such that they have the same footprint. Silicon-based FETs were assumed to be strained with a tensile stress of 2 GPa. However the $In_{0.53}Ga_{0.47}As$ based devices were assumed to be relaxed as the strain has been shown to be ineffective in improving the performance in both simulations and experiments [128, 129]. The transport for silicon-based devices was taken to be along $\langle 110 \rangle$ while transport in $In_{0.53}Ga_{0.47}As$ based FETs was assumed to be along $\langle 100 \rangle$. At such small channel lengths, the use of Boltzmann Transport equation may appear questionable because of its inability to account for source to drain tunneling, however it has been demonstrated in a recent study employing full quantum transport for InAs and strained silicon based nanowire FETs that contribution of source to drain tunneling at L_G=14 nm is less than 10% [18].

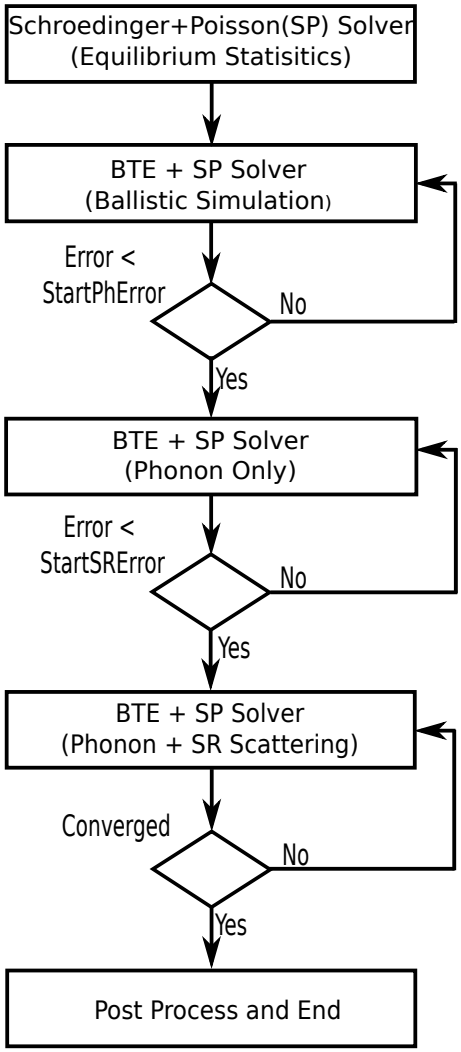


FIGURE 6.3: Optimized program flow which allows to progressively turn-on the scattering mechanism unlike the standard program flow is shown in Fig.6.1. With the new iteration scheme, the computationally heaviest scattering mechanism needs to be included only in the last few iterations.

The normalized off-state current was assumed to be 100 nA/μm consistent with the ITRS forecast for 14 nm channel length device. The drain bias was taken to be 0.7 V. The simulations also account for external series resistance, $R_{ext} = R_S + R_D = 202\,\Omega\mu m$. The conversion from $\Omega\mu m$ to Ω was performed by dividing it by the gate perimeter. The source/drain length (L_S/L_D) was taken to be 25 nm for both silicon and $In_{0.53}Ga_{0.47}As$ based devices to have a fair comparison. The source/drain doping was taken to be 10^{20} cm^{-3} and 5×10^{19} cm^{-3} for silicon and $In_{0.53}Ga_{0.47}As$ based devices. An HfO_2 gate oxide was considered with the physical gate thickness of 2.8 nm (EOT=0.5 nm). As different architectures have the same footprint as defined in Fig.6.4, this also allows us to compare the absolute values of the current as well. The scattering parameters are taken from [36, 12]

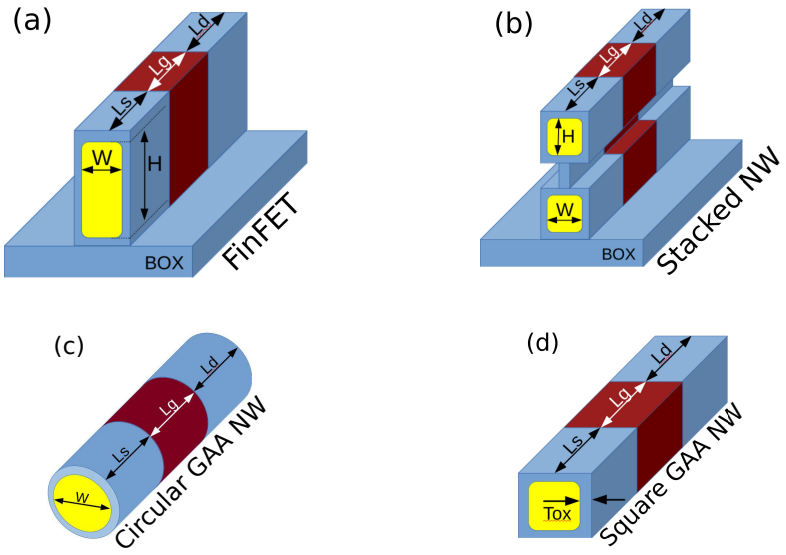


FIGURE 6.4: Schematic of the different architectures simulated in this work. (a) FinFET with W:H=1:2, (b) stacked NWs with H=W, (c) Circular NW and (d) Square NW. All the architectures have same footprint because of same width (W=5 nm) channel length (Lg=14 nm) and source/drain length (Ls=Ld=25 nm). Transport is oriented along $\langle 110 \rangle$ and $\langle 100 \rangle$ for respectively sSi and $In_{0.53}Ga_{0.47}As$. Source and drain doping were taken to be 10^{20} cm^{-3} and 5×10^{19} cm^{-3} for sSi and $In_{0.53}Ga_{0.47}As$ based devices, respectively. The oxide is HfO_2 for all the architectures with T_{ox}=2.8 nm (EOT=0.5 nm). The footprint is defined as W×(Ls+Lg+Ld).

6.3.1 Device architecture and material analysis

Fig.6.5 shows the transfer characteristics for FinFET and stacked nanowire architectures for strained silicon and $In_{0.53}Ga_{0.47}As$ based devices with normalized current (normalization is done by the gate perimeter) and absolute current. The surface roughness parameters, Δ_{rms}=0.21 nm and Λ=1.4 nm, that are used in these simulations correspond to near ideal interface [65, 77, 94]. Fig.6.5(a) shows that the absolute current for stacked nanowires is more than FinFET for both the strained silicon and $In_{0.53}Ga_{0.47}As$. However, the slight advantage for stacked nanowire is not observed in terms of normalized current. This slightly larger current that is observed in the case of stacked nanowires compared to FinFET is due to larger gate capacitance in the former. Another important observation from the Fig.6.5 is that the strain silicon-based devices have better I_{on} than the $In_{0.53}Ga_{0.47}As$ based devices.

Architecture	SS (mV/dec)	I_{on} (mA/μm)	T_{sw} (ps)	E_{sw} (fJ/μm)	N_{sh} (10^7/cm)	V_x (10^6cm/s)	gm (S/cm)	gds (S/cm)	BR (%)
FinFET	78.99	1.088	0.621	0.473	1.57	10.77	17.28	1.417	70.8
StackedNW	75.77	1.127	0.709	0.561	1.81	10.70	21.55	0.696	63.8
Circle	64.34	1.106	0.580	0.450	1.09	10.08	21.64	0.942	57.1
Square	65.89	1.002	0.561	0.393	1.24	10.09	20.88	1.472	56.3

TABLE 6.1: Calculated figures of merit mainly related to digital qualifications for strained silicon based FETs. Subthreshold swing, SS, is average value calculated between 100 nA/μm and 5 μA/μm. I_{on} is calculated at V_{GS}=V_{DS}=V_{DD}. Q_{on} is calculated at V_{DS}=0, V_{GS}=V_{DD} and Q_{off} at V_{DS}=V_{DD}, V_{GS}=0 V. Intrinsic delay, T_{sw}=(Q_{on} − Q_{off})/I_{on}. Switching energy, E_{sw}=V_{DD}(Q_{on} − Q_{off}). N_{sh} and V_x=I_{on}/(qN_{sh}) are the electron concentration and velocity at virtual source. BR is the ratio of I_{on} to ballistic on current. gm and gds are calculated at V_{DS} = V_{GS} = V_{DD}/2.

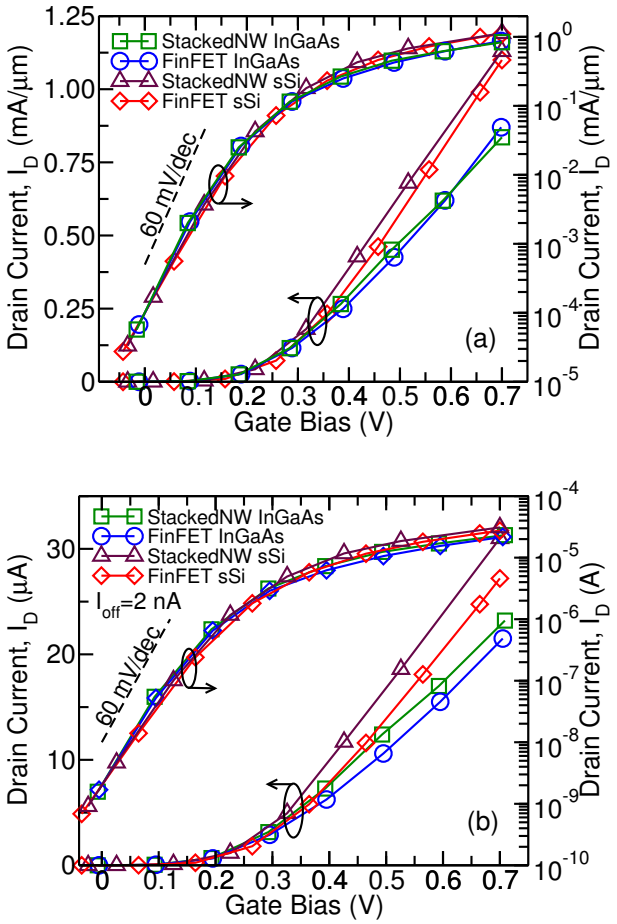


FIGURE 6.5: Simulated I_{DS} versus V_{GS} for strained silicon and $In_{0.53}Ga_{0.47}As$ for stacked nanowires and FinFET. (a) The current is normalized by the gate perimeter while in (b) the absolute value of currents are plotted. All the devices have the same footprint. Surface roughness parameters used in the simulations are Δ_{rms}=0.21 nm and Λ=1.4 nm for both the silicon and $In_{0.53}Ga_{0.47}As$.

Architecture	SS (mV/dec)	I_{on} (mA/μm)	T_{sw} (ps)	E_{sw} (fJ/μm)	N_{sh} (10^7/cm)	V_x (10^6cm/s)	gm (S/cm)	gds (S/cm)	BR (%)
FinFET	73.05	0.820	0.405	0.233	0.58	22.37	13.44	1.091	59.2
StackedNW	74	0.745	0.587	0.301	0.73	17.38	13.53	0.525	48.2
Circle	61.77	0.878	0.226	0.139	0.353	24.46	15.17	0.539	54.7

TABLE 6.2: Same as Tab.6.1 except for $In_{0.53}Ga_{0.47}As$ based FETs.

Tab.6.1 and Tab.6.2 report important analog and digital figures of merits for strained silicon and $In_{0.53}Ga_{0.47}As$ respectively, and for different architectures considered in this study. It is observed that the stacked nanowire has larger electron concentration at the virtual source, N_{sh}, as compared to the FinFET for both the materials, thus suggesting the larger absolute current in stacked nanowires is due to larger gate capacitance than FinFET. The larger gate capacitance in case of stacked nanowires also results in larger switching energy, E_{sw}, and switching time, T_{sw}, as compared to FinFET. Moreover a higher I_{on} is observed in strained silicon-based devices as compared to $In_{0.53}Ga_{0.47}As$ based devices, even though the latter has higher velocity because the silicon based device has a much larger density of states in the silicon resulting in larger electron concentration that more than compensates for smaller velocity. The subthreshold swing for the circular and square GAA nanowires are

smaller than FinFET and stacked nanowires because of their better gate control. Small ballistic ratios for all the architectures and materials highlight that the transport even in such scaled devices is far from ballistic regime. Thus the need to accurately account for all the relevant scattering mechanisms is emphasized.

Another important aspect of scaling is that there is an increase in the relative proportion of the parasitics (for example due to increase in the length of interconnect) [130]. It is thus necessary to account for parasitics as well while evaluating different architectures on the basis of their switching delays. This is accounted by first calculating the increase in the charge as the device turns on as

$$Q_{int} = Q_{on} - Q_{off} \tag{6.8}$$

where Q_{on} and Q_{off} are the total charge respectively in the on-state (V_{DS}=0 and V_{GS}=V_{DD}) and off-state (V_{DS}=V_{DD} and V_{GS}=0). Q_{on} and Q_{off} were calculated by integrating the electron concentration over the entire device. Thus allowing us to have a ball-park number for internal capacitance as

$$C_{int} = Q_{int}/V_{DD} \tag{6.9}$$

Further fringe capacitance, C_{frn}, was taken to be 1.2 times the intrinsic capacitance in-line with ITRS 2013. An estimate for charge associated with the fringe capacitance can calculated as

$$Q_{frn} = C_{frn}V_{DD} \tag{6.10}$$

Thus the total amount of charge that an inverter has to drive assuming a fan out of 3 is

$$Q_{FO3} = 3 \times 2 \times (Q_{int} + Q_{frn}) \tag{6.11}$$

Thus the total delay, D_T, is given by

$$D_T = \frac{Q_{int} + Q_{frn} + Q_{FO3} + Q_{par}}{I_{on}} = T_{sw} + \frac{Q_{frn} + Q_{FO3} + Q_{par}}{I_{on}} \tag{6.12}$$

where the charge due to the parasitic capacitance, Q_{par}, is given by $V_{DD} \times C_{par}$

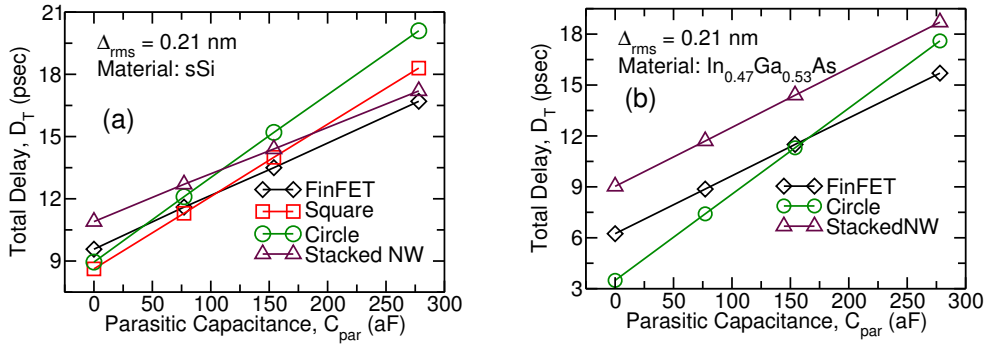


FIGURE 6.6: Plot of FO3 delay versus the parasitic capacitance. The simulation results show that with the increase in C_{par} the delay degradation for FinFET and stacked NWs is lower than for other GAA devices. This advantage of FinFET and stacked NWs for large C_{par} values stems from their large drive current per unit footprint.

Fig.6.6(a) plots FO3 delay for strained silicon, D_T, versus the parasitic capacitance for different architectures. It can be seen that in the absence of the parasitic capacitance (C_{par}=0 aF) the square and circle GAA have a smaller delay as compared to the FinFET and stacked nanowires. This is because of the larger gate capacitance as already highlighted in Tab.6.1 and Tab.6.2 by larger electron concentration at the virtual source. However, with an increase in the parasitic capacitance the delay of the square and circle GAA nanowires increases much more than the FinFET and stacked nanowires. This is because of the larger value of the absolute current (or current per unit footprint as all the architectures occupy the same area) in the FinFET and stacked nanowire as compared to the square and circle GAA FET. Thus suggesting that, despite poorer electrostatics of FinFET and stacked nanowires indicated by larger subthreshold swing in Tab.6.1 and Tab.6.2 perform better than circular and square GAA nanowires when FO3 delay is considered with parasitic capacitance. Fig.6.6(b) plots FO3 delay, D_T, versus the parasitic capacitance for different architectures for $In_{0.53}Ga_{0.47}As$. As the absolute values of the on-current for $In_{0.53}Ga_{0.47}As$ FinFET and stacked nanowires are very similar, they result in almost similar dependence on parasitic capacitance. However, the comparison of the D_T for FinFET and circular GAA nanowire is qualitatively similar to as observed in Fig.6.6(a).

6.3.2 Impact of surface roughness on I_{on}

The analysis considered in the previous section assumed a very high-quality interface by setting Δ_{rms}=0.21 nm. Such small value of the Δ_{rms} is on the lower side of the experiments even for the matured Si-SiO_2 interfaces. In III-V FET, such small value of the Δ_{rms} has been extracted only for the very recent mobility experiment [65, 94].

Fig.6.7 shows the impact of surface roughness, on the on-current, I_{on} for silicon and $In_{0.53}Ga_{0.47}As$ based FinFET and stacked nanowires. We first notice that the ballistic and phonon limited currents (Δ_{rms}=0 nm) for strained silicon-based devices are larger than $In_{0.53}Ga_{0.47}As$ because of the larger density of states, as discussed earlier. At small Δ_{rms} values the current is larger for the stacked nanowires compared to the FinFET (see also Fig.6.5(a)), but when the SR increases its impact is more pronounced in stacked nanowires, which eventually leads to a cross-over between the I_{on} in stacked nanowires and in FinFETs for large Δ_{rms}. This behavior is observed in both silicon and $In_{0.53}Ga_{0.47}As$ based devices.

To interpret this result we define

$$M_{nn}(s) = \int_{\eta_{min}}^{\eta_{max}} \xi_n(s,\eta)^{\dagger} \Phi_B \xi_n(s,\eta) \tag{6.13}$$

which lies at the heart of surface roughness calculations as discussed in chapter 4, where $\xi_n(s,\eta)$ is the wavefunction corresponding of the n^{th} subband, Φ_B is the barrier between oxide and semiconductor, s is the abscissa along the oxide-semiconductor interface and η is the coordinate in the direction normal to the interface (as illustrated in the Fig.6.8).

In the Fig.6.9 we illustrate, plots of $M_{nn}(s)$ versus the normalized perimeter for FinFET and stacked nanowires for strained silicon and $In_{0.53}Ga_{0.47}As$. The matrix element defined

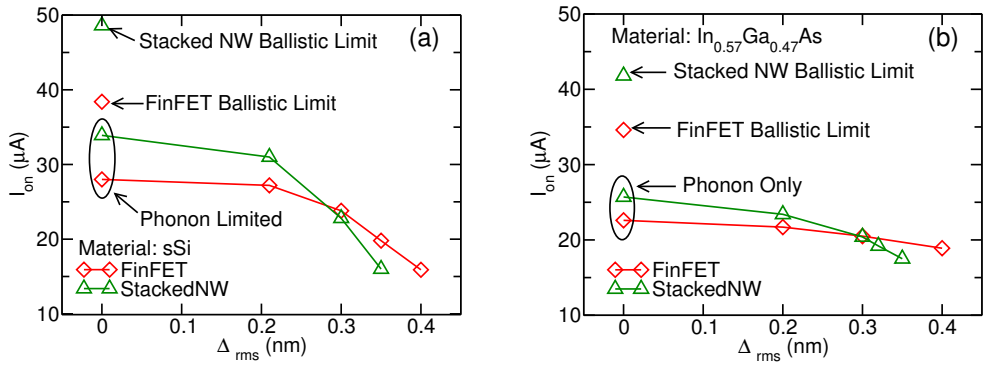


FIGURE 6.7: Impact of $\Delta_{rms}$of surface roughness on the on-current for (a) strained silicon and (b) $In_{0.53}Ga_{0.47}As$ based FinFET and stacked nanowires. The current degrades super-linearly with an increase in Δ_{rms}. The results highlight that the impact of surface roughness scattering is architecture dependent. Moreover, at larger $\Delta_{rms}$the $In_{0.53}Ga_{0.47}As$ based transistor provide larger current than strained silicon-based FETs. The larger ballistic and phonon limited current is larger in silicon than $In_{0.53}Ga_{0.47}As$ because of larger N_{sh} which more than compensates for the lower electron velocity.

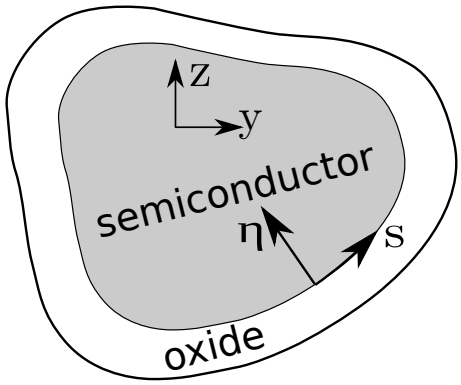


FIGURE 6.8: Schematic illustrating the coordinate system in an arbitrary shaped cross-section.

in eq.6.13 is larger in stacked nanowires than in FinFET for both the strained silicon and $In_{0.53}Ga_{0.47}As$. As the wavefunction penetration in the oxide is considered in this work and the cross-section area of individual nanowire in the stacked nanowire architecture is significantly smaller than that of FinFET leads to larger values of the wavefunction at the oxide-semiconductor interface (assuming the same oxide-semiconductor potential barrier in both the cases). This results in a larger surface roughness scattering in stacked nanowire as compared to the FinFET. Thus with increase in the surface roughness the degradation of the on-current is much more pronounced in the stacked nanowire. Thus if the surface roughness is neglected in the transport calculations it can lead to incorrect qualitative (in terms of judging the best possible architecture) and quantitative predictions, because impact of surface roughness is different for different architectures.

Another important point arises from the Fig.6.7 when comparing on-current from different materials. It shows that even though at small Δ_{rms} silicon-based FETs provide higher on-current than the $In_{0.53}Ga_{0.47}As$ based devices, as the surface roughness increases the degradation in the silicon-based FETs is significantly more than the $In_{0.53}Ga_{0.47}As$ based FETs. The reason for this is that the subbands in strained silicon are much more closely

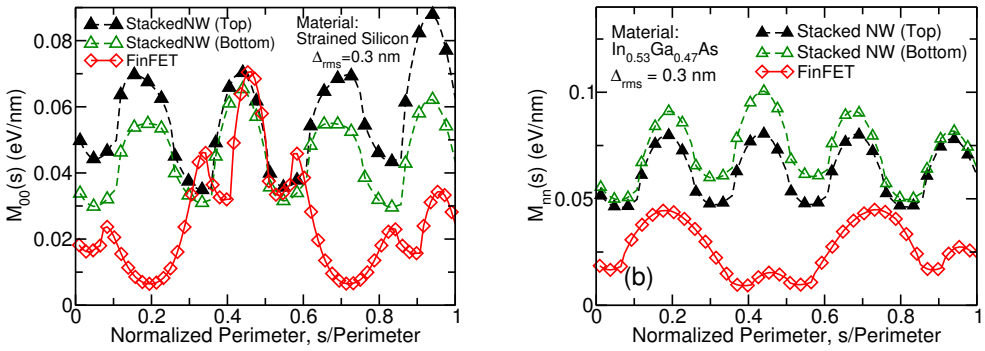


FIGURE 6.9: Plot of $M_{nn}(s)$ with n=0 along 's' for (a) strained silicon and (b) $In_{0.53}Ga_{0.47}As$based architectures. The range of integration has been done in the range $-3\Delta_{rms} \leq \eta \leq 3\Delta_{rms}$. Larger values of $M_{nn}(s)$ for stacked nanowires as compared to FinFET clarify the stronger degradation observed in stacked nanowires with an increase in surface roughness.

packed than $In_{0.53}Ga_{0.47}As$. This results in a stronger intersubband scattering which degrades the on-current. To highlight the importance of the inter-subband scattering we plot the impact of surface roughness by turning off the intersubband scattering due to surface roughness for FinFET (Fig.6.10(a)) and stacked nanowire (Fig.6.10(b)). It can be seen that when the intersubband scattering is turned-off the amount on-current degradation significantly reduces. Fig.6.11 plots the density of states (DoS) at the virtual source in strained silicon and $In_{0.53}Ga_{0.47}As$ based FinFET. It can be seen that the subbands are much more closely spaced in the case of silicon due to their larger confinement mass unlike the case of $In_{0.53}Ga_{0.47}As$ which has a very small effective mass. Closely spaced subbands allow a larger number of intersubband transitions as illustrated in Fig.6.11. Since surface roughness is an elastic scattering mechanism, a transition from a lower subband to a higher subband is accompanied by a reduction in the kinetic energy (as the potential energy increases) and thus a degradation of the electron velocity. This results in degradation of the on-current. The transition for the higher subband to lower subband is also possible (which would have an exactly opposite impact), but such transitions have a lower rate because the rate is proportional to the density of states that is larger in the higher subband because of the 1D nature of the electron gas (DoS $\alpha(E - \varepsilon_n)^{-0.5}$).

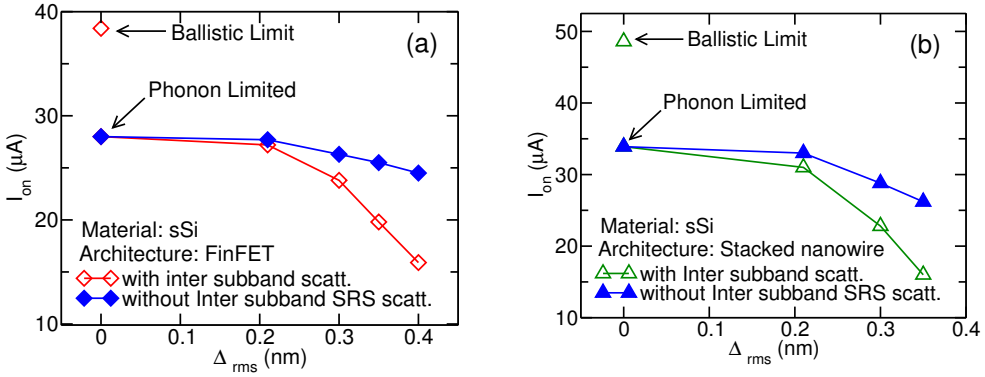


FIGURE 6.10: The impact of surface roughness on the on-current by with and without intersubband scattering due to surface roughness for (a) FinFET and (b) stacked nanowires. The simulations suggest that intersubband scattering due to surface roughness is an important reason for degradation of the on-current.

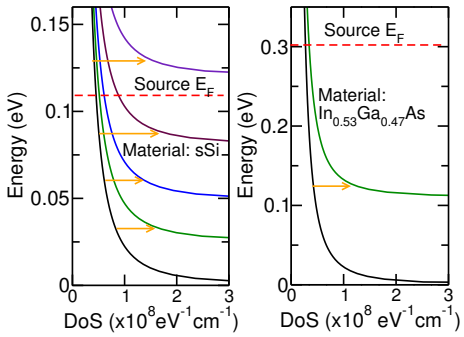


FIGURE 6.11: Density of states at virtual source of FinFET at V_{GS}=0.7 V (lowest subband energy is taken as 0 eV). The close spacing of the subbands for sSi as compared to $In_{0.53}Ga_{0.47}As$ result in larger number of intersubband transitions in the former. The transition from lower subband to higher subband is more favorable because of the larger DoS in the later.

6.3.3 Impact of self-heating on I_{on}

The thermal conductivity has a very strong dependence on the semiconductor thickness; with a reduction in the dimensions of the nanowire the thermal conductivity degrades from its bulk value of approximately 148 W/mK to few units in W/mK [117, 118, 119, 131]. The thermal conductivity also depends on surface roughness and fabrication method [118, 132]. Furthermore, the theoretical estimates of the thermal conductivity tend to overestimate the experimentally measured values because of a number of approximations made in the calculations, as discussed in [118, 119]. Because of this uncertainty in the value of the thermal conductivity, we have used it as a parameter and varied it over the range of practical interest to get an estimate of the reduction in the on-current due to self-heating.

Fig.6.12 shows the effective temperature along the transport direction for FinFET. The temperature at the source and drain boundaries was set to T=300 K using the Dirichlet boundary condition. Smaller values of the thermal conductivity lead to inefficient extraction of the heat generated due to inelastic scattering and consequently the temperature in the nanowire increases.

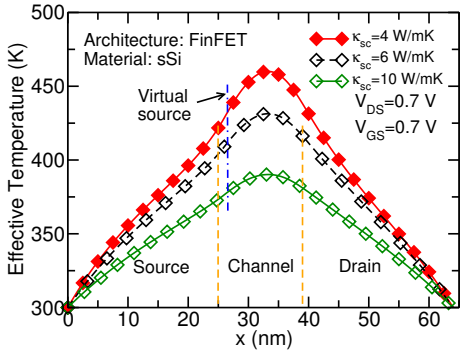


FIGURE 6.12: Effective temperature profile for FinFET architecture for different values of the thermal conductivity.

Fig.6.13 plots the normalized on-current versus the thermal conductivity for FinFET and circular GAA nanowires. We have also plotted the on-current obtained when self-heating is

neglected. Fig.6.13 shows that as the thermal conductivity reduces degradation in the on-current increases, because the reduction in the thermal conductivity leads to increase in the temperature in the nanowire (as shown in Fig.6.12). The increase in temperature leads to increase in the phonon scattering (due to Bose-Einstein statistics) which causes the reduction in the on-current. On comparing the impact of self-heating on different architectures, it can be easily seen that the degradation due to self-heating in circular nanowire much more as compared to FinFET. This can be explained by noticing that even though the two architectures have the same footprint, the cross-section area of the FinFET is more than the circular GAA nanowire which eases the extraction of the heat generated.

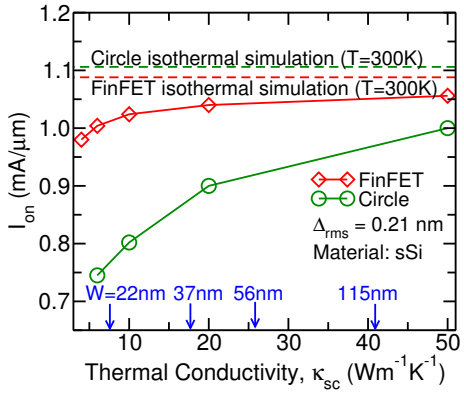


FIGURE 6.13: Impact of thermal conductivity of semiconductor on the on-current for circular gate-all-around and FinFET. The simulation results highlight that different architectures show different degradation for the same value of thermal conductivity. Vertical arrows show experimental thermal conductivity for different nanowire width, W [117].

To explore the cause of degradation further we plot in Fig.6.14 the electron velocity and electron concentration at the virtual source for different thermal conductivities for circular GAA nanowire and FinFET. It can be seen that in both cases with the reduction in the thermal conductivity electron concentration and electron velocity at the virtual source follow opposite trends. However, the reduction in the velocity due to self-heating is much more than the increase in the electron concentration, so that an overall reduction in the on-current is observed.

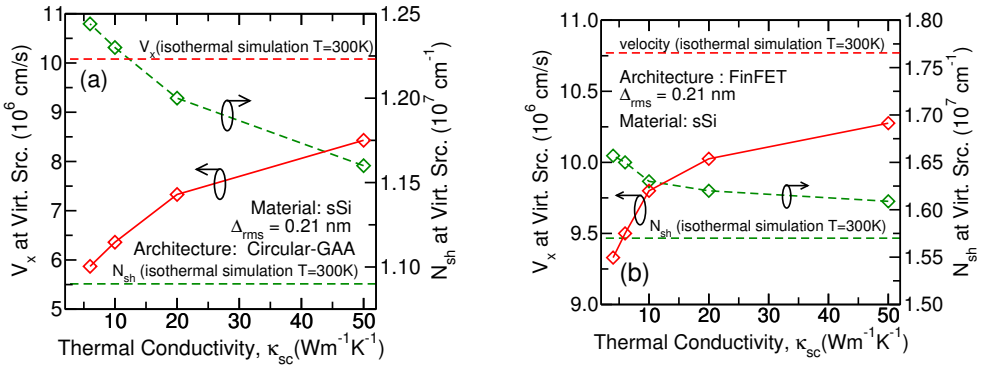


FIGURE 6.14: Plot of electron velocity and linear electron density at virtual source for (a) circular gate-all-around nanowire and (b) FinFET. In both the cases even though the charge concentration increases slightly with the reduction in the thermal conductivity the reduction in the velocity is much more and hence it results in the degradation of the on-current. The reduction in V_x with decreasing κ_{sc} (that is with increasing temperature - see Fig.6.12), is due to an increased phonon scattering and it is contrary to the ballistic behavior because the ballistic injection velocity (thermal velocity), increases with the increase in temperature [36].

Chapter 7

Electronic band structure calculation: Tight binding method

7.1 Introduction

The electronic band structure (or band structure) of a crystal or a nanostructure lists the allowed energy levels as a function of the electron wavevector. Band structures are fundamental to the study of materials for electronic states from which a plethora of useful parameters like effective mass, energy band gap, density of states, electrical resistivity can be calculated. These parameters are crucial in studying several material properties which can in turn be very useful while evaluating the materials from different applications.

A number of methods exist in the literature to calculate the band structures and among them, $\mathbf{k}\cdot\mathbf{p}$, pseudopotential and tight binding method are the popular ones. The simplest among them is the $\mathbf{k}\cdot\mathbf{p}$ method, which is based on the expansion of Schrödinger equation around a point in $\mathbf{k}$-space generally taken to be the Γ point. $\mathbf{k}\cdot\mathbf{p}$ is inherently based on perturbation theory and it is applicable only around the point about which the Schrödinger equation is expanded [36]. $\mathbf{k}\cdot\mathbf{p}$ method is suitable for studying valence bands for both direct and indirect semiconductors and conduction band for direct semiconductors as it has minimum at the Γ point. In [133] a 30 band $\mathbf{k}\cdot\mathbf{p}$ method has been developed to reproduce the energy bands in the first Brillouin zone. In pseudopotential method, the potential energy due to nuclei and all the other electrons is approximated by a form of a potential profile (hence the name pseudopotential) where the rapid variations of the nuclei potential energy near the atomic core is canceled out thanks to Phillips-Kleinman cancellation theorem. However, the application of the pseudopotential method to confined structures has been limited because they rely on the plane wave basis which increases the computational burden especially for transport [134, 135, 136, 137, 138, 139].

Recently tight binding (or linear combination of atomic orbitals or Bloch's) method has virtually become a method of choice because of its simplicity and intuitive formulation [9, 140, 141, 142]. As the name suggests the linear combination of atomic orbitals (tight binding method) essentially uses atomic orbitals to expand the electron wavefunction and thus solve the Schrödinger equation. The tight binding method uses a smaller number of basis functions as compared to other methods that rely on plane wave basis. Thus the computational effort is much smaller and hence allows us to analyze systems having larger unit cells

(effectively larger systems). Another reason for widespread use of the tight binding method is its ability to resolve the material at an atomic level which allows us to take the heterostructures and strain effects naturally into account. However, the tight binding formulation has a number of integrals involving the basis functions that are treated as fitting parameters, which can be obtained by matching the band structure obtained from tight binding to more sophisticated methods; then these parameters are used to analyze different situations [143]. This approximation of treating integrals as fitting parameters places tight binding in a sweet spot, where the use of *ab-initio* methods is impractical but the system has strong quantum effects that make use of simpler approximations (like parabolic effective mass) questionable.

In this chapter, we will first discuss the theory of the tight binding method for bulk materials and validate the implementation using the results available in the literature. We will then extend the tight binding theory to confined systems (nanowires) and discuss a methodology to account for arbitrary cross-sections. We will also validate the implementation with the results available in the literature. It is not the purpose of this chapter to provide in an in-depth literature survey of the tight binding method.

7.2 Bulk

In this section, we will discuss the theory for the tight binding method as applied to the bulk crystals and also validate the implementation for different materials.

7.2.1 Theory

In order to describe a bulk crystal the single electron Hamiltonian can be written as

$$H = -\frac{\hbar^2}{2m}\nabla^2 + \sum_{i=1}^{N} U(\mathbf{r} - \mathbf{R}_i) \tag{7.1}$$

where $U(\mathbf{r} - \mathbf{R}_i)$ is the potential energy felt by the electron due to an atom located at $\mathbf{R}_i$ approximated as a sphere and N is the total number of atoms in the crystal. Since the crystal is periodic along x, y and z direction, the wavefunction for the electron must satisfy the Bloch theorem [144]

$$\Psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} \sum_i e^{i\mathbf{k}\cdot\mathbf{R}_i} \psi_{\mathbf{k}}(\mathbf{r} - \mathbf{R}_i) \tag{7.2}$$

where $\mathbf{k}$ is the three-dimensional wavevector and $\psi_{\mathbf{k}}(\mathbf{r} - \mathbf{R}_i)$ is some function to be determined. Expanding the $\psi_{\mathbf{k}}(\mathbf{r} - \mathbf{R}_i)$ in terms of Löwdin atomic orbitals (generally in tight binding method only the orbitals whose energy are in the vicinity of the energy range of interest are considered [143]), the wavefunction can be written as

$$\Psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} \sum_{i,j} e^{i\mathbf{k}\cdot\mathbf{R}_i} C_j(\mathbf{k}) \phi_j(\mathbf{r} - \mathbf{R}_i) \tag{7.3}$$

where $C_j^i(\mathbf{k})$ is the coefficient (or weight) of the j^{th} orbital located locate at $\mathbf{R}_i$ (to be determined) and $\phi_j(\mathbf{r} - \mathbf{R}_i)$ is the Löwdin orbital centered at $\mathbf{R}_i$. Functions, $\phi_j(\mathbf{r} - \mathbf{R}_i)$ are defined

as the eigen functions of the problem

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + U(\mathbf{r}-\mathbf{R}_i)\right]\phi_j(\mathbf{r}-\mathbf{R}_i) = E_{jj}\phi_j(\mathbf{r}-\mathbf{R}_i) \tag{7.4}$$

where E_{jj} is the energy of the j^{th} orbital. As we have written $\psi_{\mathbf{k}}(\mathbf{r}-\mathbf{R}_i)$ in terms of atomic orbitals (Löwdin orbitals) in eq.7.3, tight binding method is also sometimes referred to Linear Combination of Atomic Orbitals (LCAO). Substituting eq.7.3 in eq.7.1 allows us to write the Schrödinger equation as

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + \sum_{n=1}^{N} U(\mathbf{r}-\mathbf{R}_n)\right]\sum_{i,j} e^{i\mathbf{k}\cdot\mathbf{R}_i} C_j^i(\mathbf{k})\phi_j(\mathbf{r}-\mathbf{R}_i) = E_{\mathbf{k}}\sum_{i,j} e^{i\mathbf{k}\cdot\mathbf{R}_i} C_j^i(\mathbf{k})\phi_j(\mathbf{r}-\mathbf{R}_i) \tag{7.5}$$

Multiplying the eq.7.5 with $e^{-i\mathbf{k}\cdot\mathbf{R}_{i'}}\phi_{j'}^{\dagger}(\mathbf{r}-\mathbf{R}_{i'})$ and integrating over the crystal volume we get

$$\begin{aligned}&\sum_{i,j}\int d\mathbf{r} e^{-i\mathbf{k}\cdot\mathbf{R}_{i'}}\phi_{j'}^{\dagger}(\mathbf{r}-\mathbf{R}_{i'})\left[-\frac{\hbar^2}{2m}\nabla^2 + \sum_{n=1}^{N} U(\mathbf{r}-\mathbf{R}_n)\right] C_j^i(\mathbf{k})\phi_j(\mathbf{r}-\mathbf{R}_i)e^{i\mathbf{k}\cdot\mathbf{R}_i} \\ &= E_{\mathbf{k}}\sum_{i,j} C_j^i(\mathbf{k})\int d\mathbf{r}\phi_j(\mathbf{r}-\mathbf{R}_i)\phi_{j'}^{\dagger}(\mathbf{r}-\mathbf{R}_{i'})e^{i\mathbf{k}\cdot(\mathbf{R}_i-\mathbf{R}_{i'})}\end{aligned} \tag{7.6}$$

The left hand side can be rearranged as

$$\begin{aligned}&\sum_{i,j}\int d\mathbf{r} e^{-i\mathbf{k}\cdot\mathbf{R}_{i'}}\phi_{j'}^{\dagger}(\mathbf{r}-\mathbf{R}_{i'})\left[-\frac{\hbar^2}{2m}\nabla^2 + \sum_{n=1}^{N} U(\mathbf{r}-\mathbf{R}_n)\right] C_j^i(\mathbf{k})\phi_j(\mathbf{r}-\mathbf{R}_i)e^{i\mathbf{k}\cdot\mathbf{R}_i} \\ &= \sum_{i,j} C_j^i(\mathbf{k})e^{i\mathbf{k}\cdot(\mathbf{R}_i-\mathbf{R}_{i'})}\left[\int d\mathbf{r}\phi_{j'}^{\dagger}(\mathbf{r}-\mathbf{R}_{i'})\left(-\frac{\hbar^2}{2m}\nabla^2 + \sum_{n=1}^{N} U(\mathbf{r}-\mathbf{R}_n)\right)\phi_j(\mathbf{r}-\mathbf{R}_i)\right]\end{aligned} \tag{7.7}$$

Up to this point, we have not made any approximation except for writing $\psi_{\mathbf{k}}(\mathbf{r}-\mathbf{R}_i)$ using a limited number of atomic orbitals (i.e. we have used only the atomic orbitals that lie in the energy range of interest). Depending on the configuration of i, i' and n the term in the square bracket can belong to one of the following categories

- Case I: All the functions in the integrations lie on the same atomic site. This corresponds to $i = i' = n$. In this case using eq.7.4 and orthogonality of orbital wavefunction the integration in the square bracket in eq.7.7 can be evaluated as

$$\int d\mathbf{r}\left[\phi_{j'}^{\dagger}(\mathbf{r}-\mathbf{R}_i)\left(-\frac{\hbar^2}{2m}\nabla^2 + U(\mathbf{r}-\mathbf{R}_i)\right)\phi_j(\mathbf{r}-\mathbf{R}_i)\right] = E_{jj}\delta_{j,j'}\delta(\mathbf{R}_i-\mathbf{R}_{i'}) \tag{7.8}$$

- Case II: The functions lie on the nearest neighboring atoms (i.e. the atoms are connected by bond vectors). These are referred to as 2 center integrals. In this case, the

integration in the square bracket is treated as a fitting prameter and can be written as

$$\int d\mathbf{r}\left[\phi_{j'}^{\dagger}(\mathbf{r}-\mathbf{R}_{i'})\left(-\frac{\hbar^2}{2m}\nabla^2+U(\mathbf{r}-\mathbf{R}_i)\right)\phi_j(\mathbf{r}-\mathbf{R}_i)\right]=H_{ii'}^{jj'} \tag{7.9}$$

where $H_{ii'}^{jj'}$ is a parameter that can be used to fit the band structure data available, obtained either from experiments or from other numerical methods like empirical pseudopotential method or DFT calculations.

- Case III: All other combinations

$$\int d\mathbf{r}\left[\phi_{j'}^{\dagger}(\mathbf{r}-\mathbf{R}_{i})\left(-\frac{\hbar^2}{2m}\nabla^2+U(\mathbf{r}-\mathbf{R}_i)\right)\phi_j(\mathbf{r}-\mathbf{R}_i)\right]=0 \tag{7.10}$$

Since these are $\phi_j(\mathbf{r}-\mathbf{R}_i)$ are constructed from the atomic orbitals they are localized so that overlap between the $\phi_j(\mathbf{r}-\mathbf{R}_i)$ and $\phi_j(\mathbf{r}-\mathbf{R}_{i'})$ can be approximately neglected. Thus the right-hand side of eq.7.6 simplifies to

$$E_{\mathbf{k}}\sum_{i,j}C_j^i(\mathbf{k})\int d\mathbf{r}\phi_j(\mathbf{r}-\mathbf{R}_i)\phi_{j'}^{\dagger}(\mathbf{r}-\mathbf{R}_{i'})e^{i\mathbf{k}\cdot(\mathbf{R}_i-\mathbf{R}_{i'})}=E_{\mathbf{k}}C_j^i(\mathbf{k})\delta_{j,j'}\delta_{i,i'} \tag{7.11}$$

In the above discussion, we have assumed that the overlap between the wavefunctions $\phi_j(\mathbf{r}-\mathbf{R}_i)$ and $\phi_j(\mathbf{r}-\mathbf{R}_{i'})$ is non-zero only when they are nearest neighbors. This is questionable because the orthogonalized Löwdin orbitals have contributions from the neighboring atoms and hence they do not decay as fast as the atomic orbital. Thus a cut-off distance must be set and in general, in literature, it is taken to be nearest neighbors. We have also assumed that the three center integrals ($n \neq i \neq i'$) are zero which may not be so but in the spirit of simplifying the problem and limiting the number of fitting parameters, we will embrace these approximations with no further discussion [143].

Putting the left and right hand side of the eq.7.6 together and recognizing that in the nearest neighbour approximation $(\mathbf{R}_i-\mathbf{R}_{i'})$ is just the bond vector, $\mathbf{b}_{ii'}$, we get

$$E_{jj}C_j(\mathbf{k})+\sum_{j'}\sum_{i'=1}^{NN(i)}H_{ii'}^{jj'}e^{i\mathbf{k}\cdot\mathbf{b}_{ii'}}C_{j'}(\mathbf{k})=E_{\mathbf{k}}C_j(\mathbf{k}) \tag{7.12}$$

where $NN(i)$ is the list of nearest neighbors of the atom located at $\mathbf{R}_i$, E_{jj} is the onsite energy term corresponding to the j^{th} orbital energy, $H_{ii'}^{jj'}$ is energy corresponding to the interaction between the orbital j located at $\mathbf{R}_i$ and orbital j' located at $\mathbf{R}_{i'}$, $E_{\mathbf{k}}$ is the eigen energy and $C_j(\mathbf{k})$ is the eigen vector. Eq.7.12 is the final tight binding equation, which for each wavevector $\mathbf{k}$ is an eigen value problem.

In the Slater and Koster formulation [143], the overlap between the 2 orbitals j and j' located on atoms i and i' is divided further into its individual components of σ, π and δ bonds using appropriate rotations of the orbitals using direction cosines. Thus $H_{ii'}^{jj'}$ can be written as

$$H_{ii'}^{jj'}=\alpha_\sigma V_{ii',\sigma}^{jj'}+\alpha_\pi V_{ii',\pi}^{jj'}+\alpha_\delta V_{ii',\delta}^{jj'} \tag{7.13}$$

where $V^{jj'}_{ii',\sigma}$, $V^{jj'}_{ii',\pi}$ and $V^{jj'}_{ii',\delta}$ are the weights of σ, π and δ bonds, α_σ, α_π and α_δ are the direction cosines needed to break of the bond connecting the atoms located at $\mathbf{R}_i$ and $\mathbf{R}_{i'}$ into its σ, π and δ components. $V^{jj'}_{ii',\sigma}$, $V^{jj'}_{ii',\pi}$ and $V^{jj'}_{ii',\delta}$ are the material parameters and they are fitted to match the available band structure data [145, 146]. However this is not the only way and in some old literature the $H^{jj'}_{ii'}$ are directly fitted to the band structure data [147, 148].

In a zinc-blende or diamond crystal structure each atom is connected to 4 atoms (nearest neighbors) which is schematically shown in Fig.7.1

Because of the periodicity along the x, y and z directions (same as in section 7.2), the unit

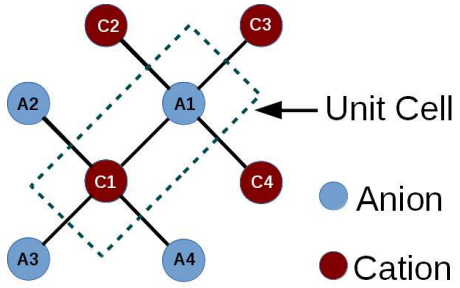


FIGURE 7.1: Illustration of the projection of the 3D crystal lattice on a plane. In the diamond crystal structure both the atoms are of the same type. The connections C1-A2, C1-A3, C1-A4 are related to C1-A1 connection through symmetry relations [147]. Similar considerations are also valid for anion atom.

cell (smallest repeating unit in the crystal) consists of 2 atoms. In the diamond crystal both the atoms are of same type (for example in Si crystal both are Si atoms) while in the case of zinc-blende crystal structure the 2 atoms are of different elements (for example in InAs crystal one atom is of In and another is of As). Thus the tight binding matrix can be written as

$$\begin{bmatrix} H_{aa} & H_{ac} \\ H_{ca} & H_{cc} \end{bmatrix} \begin{bmatrix} C^a \\ C^c \end{bmatrix} = E_{\mathbf{k}} \begin{bmatrix} C^a \\ C^c \end{bmatrix} \tag{7.14}$$

where $[H_{aa}]$ and $[H_{cc}]$ are the diagonal matrices corresponding to the onsite energies (orbital energies) corresponding to the anion and cation in the unit cell (for diamond lattice both are the same), $[H_{ac}]$ and $[H_{ca}]$ are the matrices corresponding to the interaction between the two atoms in the unit cell (because of the periodicity of the crystals all components of all the 4 neighboring atoms get clubbed up into the off-diagonal terms), $E_{\mathbf{k}}$ is the eigen energy. $[H_{aa}]$, $[H_{cc}]$, $[H_{ac}]$ and $[H_{ca}]$ are square matrices of dimension equal to number of orbitals taken into account. Because the Hamiltonian must be hermitian (as we are accounting only the real band structure) we have

$$H_{ac} = H^{\dagger}_{ca} \tag{7.15}$$

7.2.2 Validation

Having discussed the theory for bulk crystals we proceed to validate the implementation with the results available in the literature. Fig7.2 and Fig7.3 show the bulk band structure

for silicon and germanium respectively.

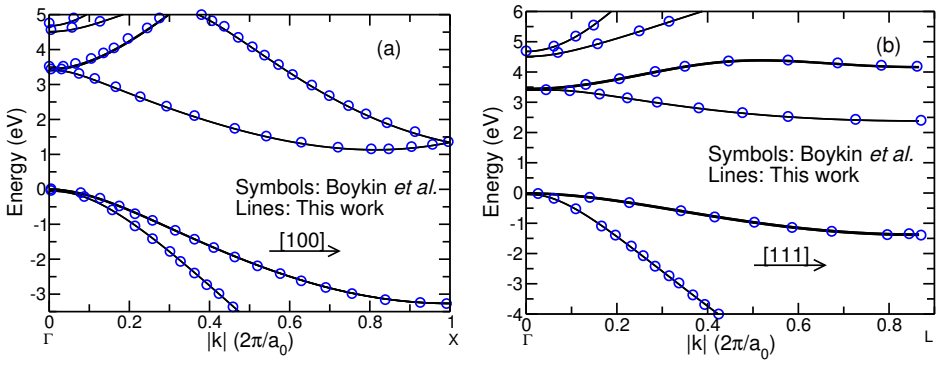


FIGURE 7.2: Comparison of the band structures extracted from [149] and calculated in this work for silicon along [100] and [110] directions. In all there are 19 fitting parameters corresponding to various hopping and onsite terms for group IV semiconductors which are taken from [149].

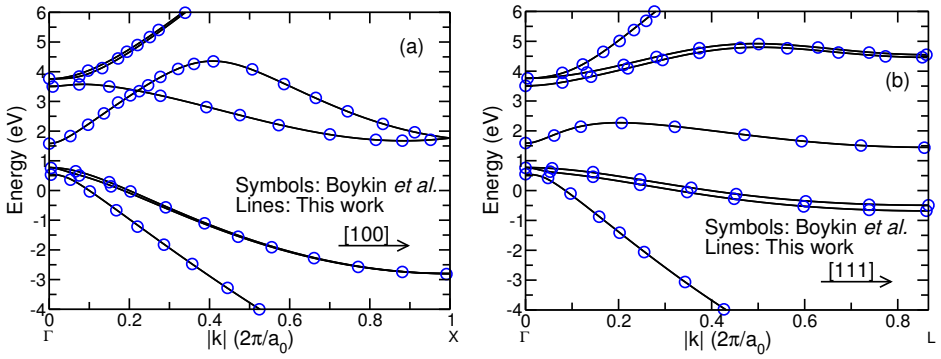


FIGURE 7.3: Same as 7.2 except for germanium.

The parameters used in the tight binding simulations were taken from [149]. The band structures were also extracted from the same reference so as to have a fair comparison. It can be seen that both band structures for silicon and germanium along [100] and [111] can be very accurately reproduced.

Another important feature that is extracted from the band structures is the effective mass (curvature of the bands). Fig7.4 shows the parabolic E-k relationship and lowest conduction band from the tight binding method in silicon and germanium. The parabolic E-k relationship with the textbook effective mass of $0.916m_0$ for silicon at $\approx 0.82\ \frac{2\pi}{a_0}$ along [100] and $0.037m_0$ and $1.588m_0$ at Γ and L point respectively for germanium match very well [36]. This further validates the implementation of the tight binding method as applied to the bulk crystals.

7.3 Comments on the basis set

The computational burden for calculating the band structure using tight binding is directly dependent on the number of orbitals used. Hence we need to justify the use of 10 orbitals (one s, three p, one s^* and five d orbitals) per atom (excluding spin). In order to obtain a qualitative similitude with the real bandstructure, we need at least 4 orbitals per atom

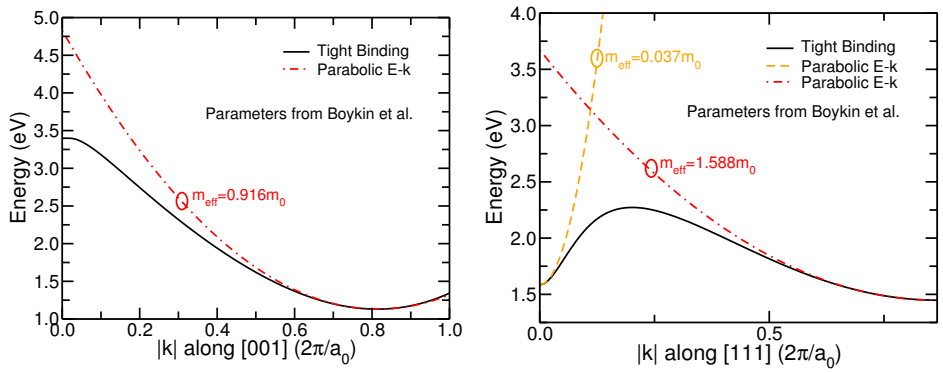


FIGURE 7.4: Comparison of the parabolic effective masses approximation and lowest conduction band obtained from the tight binding method for (a) silicon at X point and (b) germanium at Γ and L points. The values of the effective masses used in the parabolic E-k relationship have been taken from [36].

(one *s* and 3 *p* orbitals) [150] (excluding spin-orbit interaction). With this configuration (4 orbitals and nearest neighbor coupling) we can reproduce the valence bands satisfactorily, however it is not possible to correctly reproduce the bandgap in the indirect semiconductors having diamond or zinc-blende crystal structure because it neglects the contributions of the higher energy orbitals at the X and L points [151]. We can overcome this inadequacy by adding coupling beyond the nearest neighbors or by adding higher orbitals [150]. Even though adding beyond the nearest neighbor coupling can solve this issue without increasing the number of orbitals, it goes against the spirit of the tight binding method to reduce the number of fitting parameters. Also, the interaction between the atoms is expected to drop quite rapidly as the distance between the atoms increases. In addition the use of higher orbitals (for example *d* orbitals) can be naturally taken into account by simply increasing the size of the problem.

Vogl *et al.* in [148] added an excited state, *s**, to rectify the error in conduction band and retained the nearest neighbor coupling. This increased the size of the problem to be solved (5 orbitals per atom excluding spin). The addition of this higher excited orbital can reproduce the energy bandgap and position of the minima of the conduction band quite accurately as demonstrated by in [148] but the second conduction band is still incorrect. Most importantly, by restricting to this orbital set and nearest neighbor interaction it is not possible to reproduce the second derivative (effective mass) at the X point for silicon, which is an extremely important parameter for electronic applications, correctly [146]. This orbital set along with the nearest neighbor can be used for calculation of the optical properties of the material, where only the energy gaps are important [145].

Thus in order to correct the curvature of the first conduction band and the second conduction band, we must include higher orbitals (*d* orbitals) [145]. Moreover with the addition of *d* orbitals the density-of-states, deformation potentials is greatly improved [145]. Thus we need at least 10 orbitals per atom (20 orbitals with spin) to correctly reproduce the essential quantities needed for electronic applications.

7.4 Confined Structure: Nanowires

The dimension of the practical devices is always limited in at least one or two dimensions. Thus the periodic boundary conditions along all the three directions used in the previous section are not justified. And the confinement can drastically change the band structure of a material, for example, even though bulk silicon has an indirect band gap but strongly confined thin films or nanowires have direct band gap because of folding of the bands onto the Γ point. Thus there is a need to modify the derivations of the Hamiltonian.

7.4.1 Arbitrary Orientation

Crystal orientation is an important tool in improving the device performance by modifying the electron effective mass along both the transport and confinement direction [152, 153, 141]. In order to apply the tight binding method the first and foremost what we need is the position of the atoms with respect to each other in a unit cell. The position of the atoms can vary depending upon the crystallographic orientation of the nanowire.

In order to get the positions of the atoms in the crystal we recall that the semiconductors are generally diamond or zinc-blende crystal structures that is are face-centered crystal (FCC) with a motif. For a coordinate system where $x \equiv (1,0,0)$, $y \equiv (0,1,0)$ and $z \equiv (0,0,1)$ the basis vectors for FCC are

$$\begin{aligned} \boldsymbol{v_1^0} &= 0.5a_0[0,\ \ 1,\ \ 1]^T \\ \boldsymbol{v_2^0} &= 0.5a_0[1,\ \ 0,\ \ 1]^T \\ \boldsymbol{v_3^0} &= 0.5a_0[1,\ \ 1,\ \ 0]^T \end{aligned} \tag{7.16}$$

Using these basis vectors, we can obtain the position of atoms when the crystal is grown along [100] direction. The bond vectors connecting i^{th} atom to its 4 nearest neighbors in same coordinate system are given by

$$\begin{aligned} \boldsymbol{b_1^0} &= 0.25a_0[+1,\ \ +1,\ \ +1]^T \\ \boldsymbol{b_2^0} &= 0.25a_0[+1,\ \ -1,\ \ -1]^T \\ \boldsymbol{b_3^0} &= 0.25a_0[-1,\ \ +1,\ \ -1]^T \\ \boldsymbol{b_4^0} &= 0.25a_0[-1,\ \ -1,\ \ +1]^T \end{aligned} \tag{7.17}$$

However, when the nanowires are grown along any other direction (for example along [110] or [111]) it is necessary we rotate the basis and bond vectors. Let the unit vectors of the new coordinate system be given by $\hat{x} \equiv [a_x,\ \ b_x,\ \ c_x]$, $\hat{y} \equiv [a_y,\ \ b_y,\ \ c_y]$ and $\hat{z} \equiv [a_z,\ \ b_z,\ \ c_z]$ in the old coordinate system, where $\hat{x} \cdot \hat{y} = 0$ and $\hat{x} \times \hat{y} = \hat{z}$. We can perform this transformation using a rotation matrix, $\boldsymbol{R}$, given by

$$\boldsymbol{R}^{-1} = \begin{bmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{bmatrix} \tag{7.18}$$

Thus the new basis and bond vectors can be written as

$$\begin{aligned} \boldsymbol{v_i} &= R\boldsymbol{v_i^0} \qquad \text{i=1, 2 or 3} \\ \boldsymbol{b_i} &= R\boldsymbol{b_i^0} \qquad \text{i=1, 2, 3 or 4} \end{aligned} \tag{7.19}$$

7.4.2 Theory

The first and most obvious observation is that the use of the unit cell of two atoms that we have used in bulk crystal is no longer valid in a 1D system which is unconstrained along the transport direction. The unit cell consists of all the atoms that lie in the cross-section and are repeated along the transport direction, z. These atoms can be listed by using the projection of the basis vectors, $\boldsymbol{v_1}$, $\boldsymbol{v_2}$ and $\boldsymbol{v_3}$ and basis atom (diamond or zinc-blende is a face-centered crystal with a basis) on the $x-y$ plane. With these projected basis vectors, $\boldsymbol{v_{1t}}$, $\boldsymbol{v_{2t}}$ and $\boldsymbol{v_{3t}}$ we can generate the projection of the 3D crystal onto a 2D plane by generating primitive vectors as

$$a_i = n_{1i}\boldsymbol{v_{1t}} + n_{2i}\boldsymbol{v_{2t}} \tag{7.20}$$

Since the projection of the 3D basis vectors onto a 2D plane makes them linearly dependent and hence we can generate a projection of the crystal with the projection of any 2 basis vectors and n_{1i}, n_{2i} are integers. Varying n_{1i} and n_{2i} from $-\infty$ to $+\infty$ this creates an infinite mesh of atoms, therefore we must *reject* the atoms that lie outside the cross-section. The procedure for rejection is trivial for analytic curves, but can be challenging for arbitrary cross-section and it will be discussed later. Thus with a large number of atoms in the unit cell the size of the Hamiltonian increases. The Hamiltonian is a square matrix of rank $N_{orb} \times N_{atoms}$.

The Hamiltonian for the bulk case consists of onsite energy components and the hopping term (the interaction between the two atoms). It is easy to see that onsite terms would be the diagonal blocks of the overall Hamiltonian and the off-diagonal blocks would be four off-diagonal blocks corresponding to the neighboring atoms which in the bulk case were written in H_{ac} using symmetry. In order to calculate these off-diagonal blocks, we recall the off-diagonal block for the bulk Hamiltonian can be written as 7.12

$$H_{ac} = H_{i1}e^{i\mathbf{k}\cdot\boldsymbol{b_{i1}}} + H_{i2}e^{i\mathbf{k}\cdot\boldsymbol{b_{i2}}} + H_{i3}e^{i\mathbf{k}\cdot\boldsymbol{b_{i3}}} + H_{i4}e^{i\mathbf{k}\cdot\boldsymbol{b_{i4}}} \tag{7.21}$$

where H_{i1}, H_{i2}, H_{i3} and H_{i4} are the terms of the part of Hamiltonian along the bond vectors $\mathbf{b}_{i1}$, $\mathbf{b}_{i2}$, $\mathbf{b}_{i3}$ and $\mathbf{b}_{i4}$ respectively, $\mathbf{k} \equiv (k_x, k_y, k_z)$ and $\mathbf{b}_{ii'} \equiv (b^x_{ii'}, b^y_{ii'}, b^z_{ii'})$. Thus the part of Hamiltonian describing the interconnection between the atoms can be written as

$$H_{ac} = \left[H_{i1}e^{ik_z b_{i1z}}\right] e^{i\mathbf{k}_t\cdot\boldsymbol{b_{i1t}}} + \left[H_{i2}e^{ik_z b_{i2z}}\right] e^{i\mathbf{k}_t\cdot\boldsymbol{b_{i2t}}} + \left[H_{i3}e^{ik_z b_{i3z}}\right] e^{i\mathbf{k}_t\cdot\boldsymbol{b_{i3t}}} + \left[H_{i4}e^{ik_z b_{i4z}}\right] e^{i\mathbf{k}_t\cdot\boldsymbol{b_{i4t}}} \tag{7.22}$$

where $\mathbf{k}_t \equiv (k_x, k_y)$ and $\boldsymbol{b_{ii't}} \equiv (b_{ii'x}, b_{ii'y})$ are the projection of wavevectors and bond vectors on the $x-y$ plane. For a system confined along x and y, k_x and k_y have no physical significance (k_x and k_y are a consequence of periodicity). Performing an inverse Fourier transform of H_{ac} in eq.7.22 along x and y direction removes the exponential terms containing

k_x and k_y. $H_{ii'}e^{ik_z b_{ii'z}}$ is the coupling matrix between the atoms connected by the projection of the bond vector $\mathbf{b}_{ii't}$.

As an illustration, consider a toy cross-section in which the positions of atoms are as shown in Fig.7.5. For this structure the Hamiltonian is given by

$$H_{TB} = \begin{bmatrix} H_{11} & 0 & 0 & 0 & 0 & H_{i3} & H_{i1} & 0 & 0 \\ 0 & H_{22} & 0 & 0 & 0 & H_{i4} & 0 & H_{i1} & 0 \\ 0 & 0 & H_{33} & 0 & 0 & H_{i2} & H_{i4} & H_{i3} & H_{i1} \\ 0 & 0 & 0 & H_{44} & 0 & 0 & H_{i2} & 0 & H_{i3} \\ 0 & 0 & 0 & 0 & H_{55} & 0 & 0 & H_{i2} & H_{i4} \\ H_{i3}^{\dagger} & H_{i4}^{\dagger} & H_{i2}^{\dagger} & 0 & 0 & H_{66} & 0 & 0 & 0 \\ H_{i1}^{\dagger} & 0 & H_{i4}^{\dagger} & H_{i2}^{\dagger} & 0 & 0 & H_{77} & 0 & 0 \\ 0 & H_{i1}^{\dagger} & H_{i3}^{\dagger} & 0 & H_{i2}^{\dagger} & 0 & 0 & H_{88} & 0 \\ 0 & 0 & H_{i1}^{\dagger} & H_{i3}^{\dagger} & H_{i4}^{\dagger} & 0 & 0 & 0 & H_{99} \end{bmatrix} \tag{7.23}$$

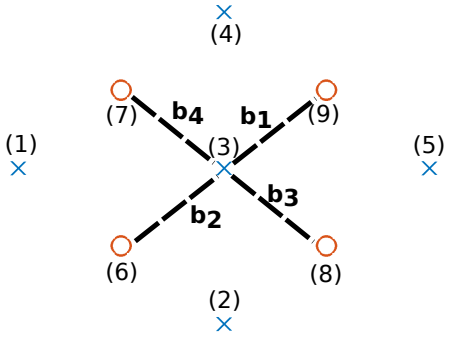


FIGURE 7.5: Shown is a projection of a unit cell for a circular nanowire with a radius of 0.3 nm with [100] orientation. Different symbols ('cross' and 'circle') represent 2 different types of atoms in a zinc-blende crystal structure. The different bonds between 3^{rd} atom and its nearest neighbors are also illustrated.

However solving the above Hamiltonian directly would result in unphysical states that lie in the band gap due to dangling bonds. The atoms at the surface (atom 1, 3, 4 and 5) must be passivated as discussed in [154]. In this method, the energy needed by an electron to occupy the unsatisfied bond is increased while the energy of the bonds that are connected are kept unchanged.

Arbitrary cross-section

Even though the formulation for the tight binding method discussed earlier is applicable to any cross-section, getting a list of the atoms that lie in the cross-section can be a challenge if the cross-section shape is arbitrary and a simple mathematical relationship(s) does not exist as is the case with most of the practical devices. This calls for developing a general methodology with which we can *reject* the atoms that lie outside a certain perimeter as it is straightforward to generate an infinite crystal lattice.

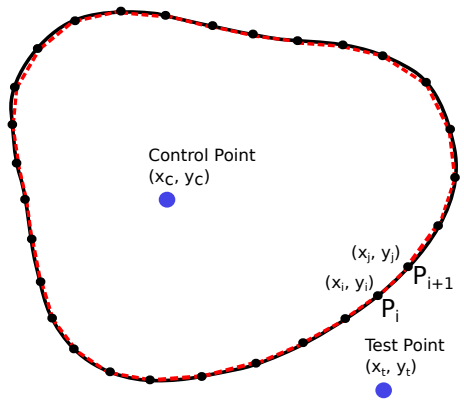


FIGURE 7.6: Illustration of a continuous arbitrary cross-section (solid line) along with discretized cross-section (dashed line). Control point is any point that lies inside the cross-section while a test point can lie any where in the plane.

Fig.7.6 shows an arbitrary cross-section shape which can be made with a general mesh generator tool. The mesh generator will give the list of ordered mesh points that lie on a curve (for example as shown by black bullets). Another ingredient needed is coordinate of any point (let's call it control point with coordinate (x_c, y_c)) that lies inside the curve. Now with the list of points, we can break the curve into line segments (as shown by the dashed lines joining the neighboring nodes). So in order to determine if any point in the $x-y$ plane lies inside the curve then they must lie on the same *side* of all the line segments. Since we know the coordinates of the nodes we can define a function, f, as

$$f = y - mx - c \tag{7.24}$$

where (x, y) are the coordinates of a point, where $m = \frac{y_j - y_i}{x_j - x_i}$ is the slope of the line segment P_iP_{i+1} and c is the intercept of the line segment P_iP_{i+1} on y-axis. The sign of the residual will determine whether a point lies either to the left or right side of the line segment depending on its orientation. Since we know the control point (x_c, y_c) we can easily calculate f and lets say that it is f_c (see Fig.7.6). Now for any test point (x_t, y_t) we can calculate the f and lets say that it is f_t. Then if the ratio of f_t and f_c is positive then control point and test point lie on the same side of the segment, P_iP_{i+1}. Performing the same calculations for all the line segments that describe the cross-section we can identify the atoms that lie inside the contour.

7.4.3 Validation

In order to validate the implementation we have simulated the band structure for silicon and germanium nanowires using the Band Structure Lab tool available on nanohub.org [155] and one we have developed. We have considered a cylindrical nanowire with a radius of 0.8 nm grown along [100], [110] and [111] direction. We see that we obtain excellent agreement between the two tools. Fig.7.7 and Fig.7.8 show the band structure for germanium and silicon having different crystallographic orientation. Fig.7.7 and Fig.7.8 also show the 2D projection of the unit cell used in the simulations

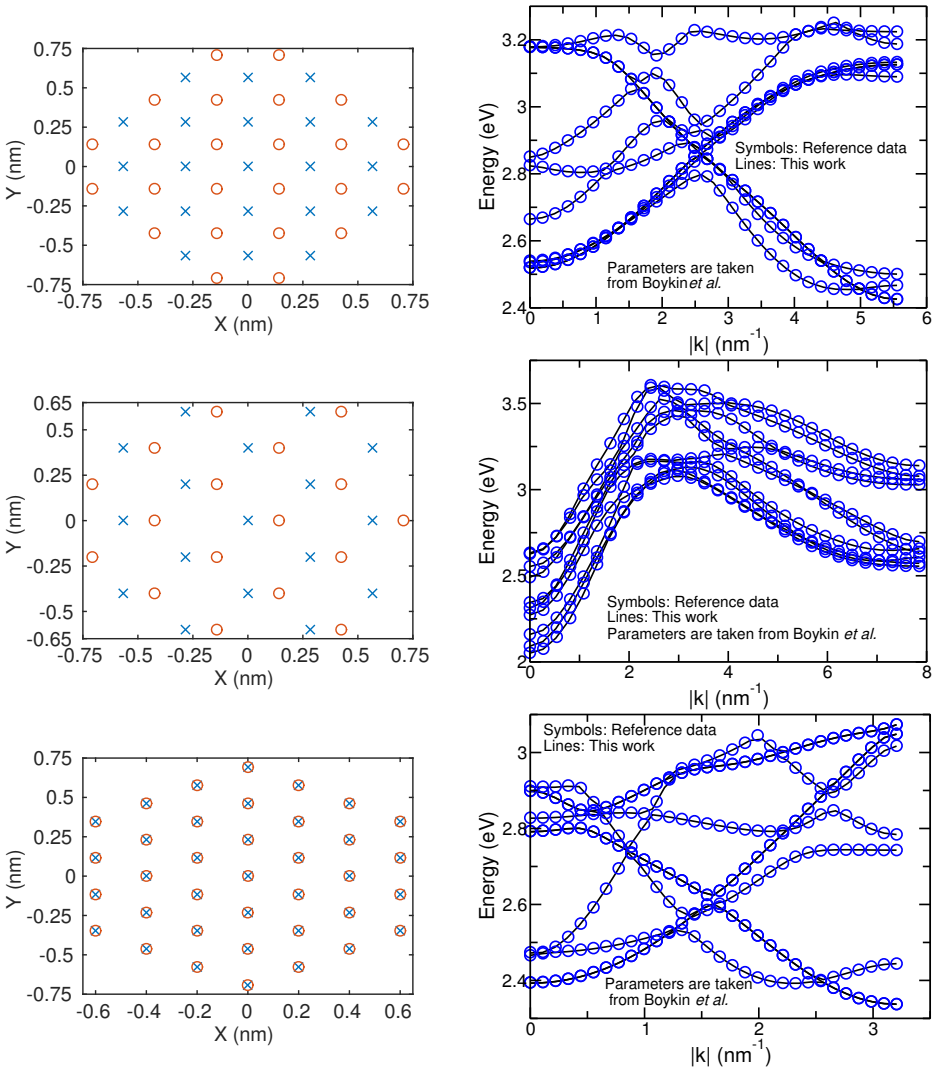


FIGURE 7.7: Comparison between the E-k relationship for germanium nanowire with diameter of 1.6 nm whose transport is oriented along [100], [110] and [111] crystallographic direction. It also shows the 2D projection of atoms (circle and cross) in the unit cell that is used in the simulations. The reference data is taken from BandStructure Tool ([155])

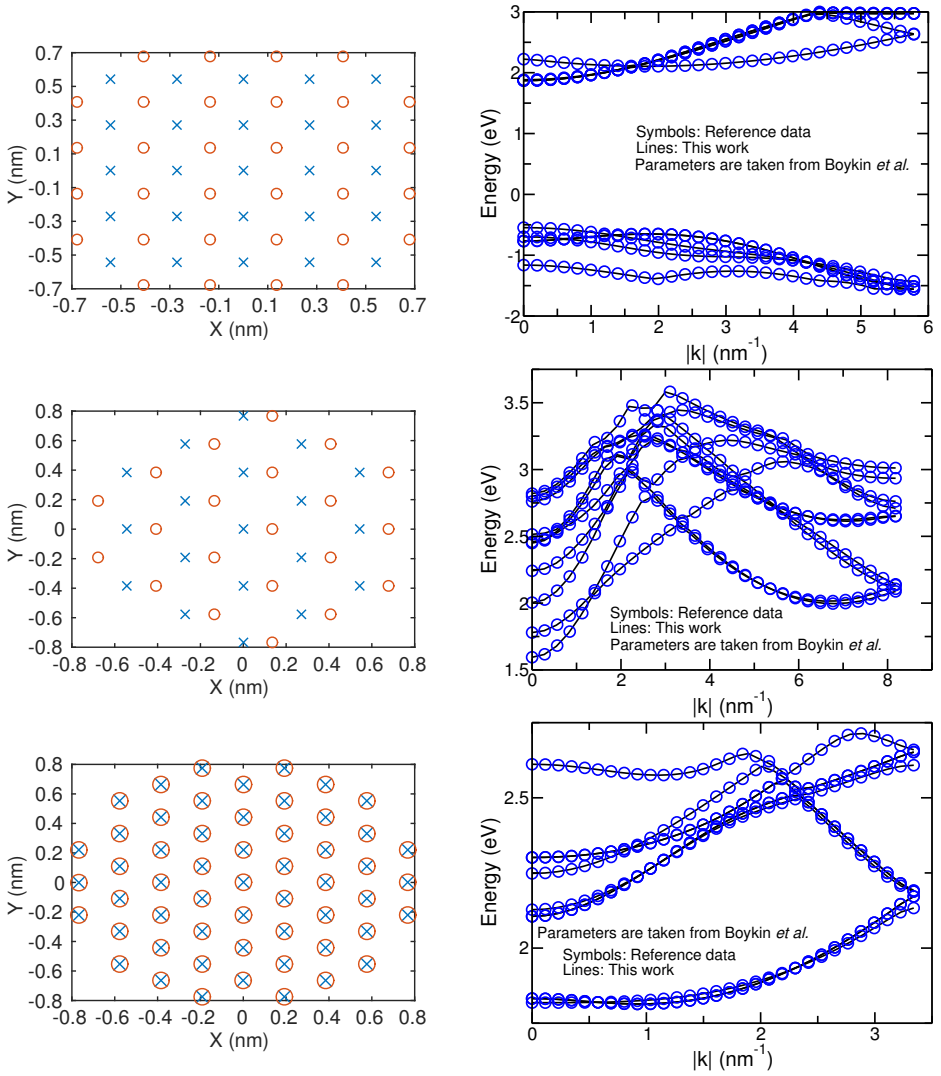


FIGURE 7.8: Comparison between the E-k relationship for silicon nanowire with diameter of 1.6 nm along whose transport is oriented along [100], [110] and [111] crystallographic direction. It also shows the 2D projection of atoms (circle and cross) in the unit cell that is used in the simulations. The reference data is taken from BandStructure Tool ([155])

Chapter 8

Conclusion and future work

8.1 Conclusion

In this thesis, we have worked extensively on the surface roughness scattering which is the dominant scattering mechanism in modern-day MOSFET. We have first extended the formulation of the nonlinear surface roughness scattering model for planar devices to account for tensorial screening. The extension was necessary so as to make the model applicable for multigate architectures. The nonlinear model was then implemented in a Multi-subband Monte Carlo framework comprising a wide range of scattering mechanism. The implementation of the nonlinear model was parallelized using openMP API to reduce the simulation time. Using the nonlinear model for surface roughness scattering, we were able to reproduce some experimental mobility data sets with values of the surface parameters in good agreement with physical measurements of the interface morphology for both bulk and quantum well transistors.

We have developed a nonlinear surface roughness scattering model for 3D architectures having fairly arbitrary cross-section shape and bias conditions. We have also developed a complete device simulation framework based on the deterministic solution of the Boltzmann transport equation for a 1D electron gas, in which new surface roughness model was implemented along with other relevant scattering mechanisms and physical effects, such as series resistance and self-heating. The simulator was then used for analysis and design of nanoscale MOSFETs. Our mobility analysis suggests that using nonlinear surface roughness scattering model allowed us to extract credible values of the surface roughness from the mobility data sets. Our analysis also highlighted that accurately accounting for biasing scheme is an important aspect while analyzing a device. Our simulation analysis suggests that performing ballistic or phonon limited analysis can give incorrect qualitative results because the impact of surface roughness depends upon the architecture: stacked nanowires are more sensitive to surface quality than their FinFET counterparts. The self-heating effects are also architecture dependent and can cause non-negligible degradation in the on-current, which is because of the reduction in the carrier velocity, even though the thermal velocity increases with increase in temperature.

8.2 Future Work

We present below a list of possible future developments related the work discussed in this thesis:

- **Nonlinear surface roughness scattering model:** The nonlinear surface roughness model that has been discussed in this thesis assumes that different interfaces in the FET are uncorrelated. This assumption may be questionable especially in modern day transistors where the semiconductor in the channel is just a few atoms thick [156]. Another relevant effect of the surface roughness is the redistribution of the electron profile due to modification in the electrostatic potential produced by of the surface roughness itself [157] which has been neglected in this thesis without a thorough justification. The nonlinear model can be modified in future to account for these effects.

- **Optimization Schemes:** The optimization scheme developed in this work can be improved further by reducing the number of times the surface roughness scattering elements are calculated. Further modifications to the iteration scheme may be developed to make the simulator more efficient.

- **Full band analysis:** The transport analysis performed in this work relied on effective mass approximation. The parabolic E-k relationship was corrected to account for the nonparabolic effects. Even though the use of nonparabolic correction improves the agreement with the full band E-k relationship, it may be questionable in the on-state where the carrier distribution is expected to be very far away form quasi-equilibrium distribution. Another consequence of the use of parabolic effective mass approximation is that it forces us to ignore the kinetic energy contributions to the surface roughness scattering calculations. These may be significant for III-V FETs and at the heterostructure interfaces between the buffer and active layer.

Appendix A

Fine print in matrix element derivation

To convert the integral from the Cartesian coordinate (4.4) to curvilinear coordinate system we introduce a parameterization of the oxide-semiconductor interface, $\mathcal{I}_0$ as

$$\mathbf{r}(t) = f(t)\hat{i}_y + g(t)\hat{i}_z \tag{A.1}$$

where t is the parameter, $\hat{i}_y$ and $\hat{i}_z$ are the unit vectors along y and z directions. This allows us to define an abscissa $s(t)$ along the $\mathcal{I}_0$ which is the length of the curve as [158]

$$s\,(t) = \int_0^t \left| \frac{d\mathbf{r}(u)}{du} \right| du. \tag{A.2}$$

We can define a unique inverse function, $\widetilde{t}(s)$ which maps a given value of s on t as $s\,(t)$ is a positive-defined and monotonically increasing function. Thus we can rewrite the $\mathbf{r}(t)$ in terms of s as

$$\mathbf{r}(s) = \widetilde{f}(s)\hat{i}_y + \widetilde{g}(s)\hat{i}_z \tag{A.3}$$

where $\widetilde{f}(s) = f\left[\widetilde{t}\,(s)\right]$ and $\widetilde{g}(s) = g\left[\widetilde{t}\,(s)\right]$.

The unit vector along the tangent to the $\mathcal{I}_0$ is

$$\hat{t}(s) = \frac{d\mathbf{r}(t)/dt}{|d\mathbf{r}(t)/dt|} = \frac{d\mathbf{r}(t)}{dt} \times \left[\frac{ds(t)}{dt}\right]^{-1} = \frac{d\mathbf{r}\,(s)}{ds} = \frac{d\widetilde{f}(s)}{ds}\hat{i}_y + \frac{d\widetilde{g}(s)}{ds}\hat{i}_z = t_y(s)\hat{i}_y + t_z(s)\hat{i}_z \tag{A.4}$$

This allows us to define a unit vector normal to the curve $\mathcal{I}_0$ as

$$\hat{n}(s) = -t_z(s)\hat{i}_j + t_y(s)\hat{i}_z = n_y(s)\hat{i}_y + n_z(s)\hat{i}_z. \tag{A.5}$$

Let us define another curve $\mathcal{I}_\eta$ such that it is at a distance η along the normal (as defined by $\hat{n}(s)$) to the curve $\mathcal{I}_0$. The relationship between the Cartesian and curvilinear coordinate system is given by

$$\begin{aligned} y &= \widetilde{f}(s) + \eta\, n_y(s) \\ z &= \widetilde{g}(s) + \eta\, n_z(s) \end{aligned} \tag{A.6}$$

We emphasize here that the relationship between the (y, z) and (s, η) coordinate system is strictly valid only in a very small region around the interface to preserve the one-to-one

mapping. The Jacobian of the coordinate transformation can be written as

$$J(s,\eta) = \det \begin{bmatrix} \frac{\partial y}{\partial s} & \frac{\partial y}{\partial \eta} \\ \frac{\partial z}{\partial s} & \frac{\partial z}{\partial \eta} \end{bmatrix} = \begin{bmatrix} t_y - \eta \frac{\partial t_z}{\partial s} & -t_y \\ t_z + \eta \frac{\partial t_y}{\partial s} & +t_z \end{bmatrix} \tag{A.7}$$

$$= 1 + \eta \left(t_z \frac{\partial t_y}{\partial s} - t_y \frac{\partial t_z}{\partial s} \right) \tag{A.8}$$

We notice that $J(s,0) = 1$, and in fact the length of curve $\mathcal{I}_0$ is $D_0 = \int_0^{D_0} J(s,0)ds$. Since $\Delta(s,x)$ is perturbation to the nominal interface it is reasonable to expect them to be small and thus we approximate the $J(s,\eta)$ by its average value, $\overline{J}(s,\eta)$, along the curve $\mathcal{I}_\eta$,

$$\overline{J}(s,\eta) = \frac{\int_0^{D_0} J(s,\eta)ds}{\int_0^{D_0} ds} = \frac{D_\eta}{D_0}. \tag{A.9}$$

We recall the Fourier series expansion expresses the $\xi_n(\mathbf{r})$ as a function of (s_η, η). As seen from fig.4.1 one to one relationship must exist between the $\mathcal{I}_\eta$ and $\mathcal{I}_0$ and hence we must be able to express s_η as function of s ($\widetilde{s_\eta}(s)$). Thus the Fourier series expansion can be written as

$$\xi_n(\eta, s) = \sum_l \phi_{n,l}(\eta)\, e^{i2\pi l \widetilde{s_\eta}(s)/D_\eta}. \tag{A.10}$$

For small η we introduce a second approximation of by linearizing $\widetilde{s_\eta}(s)$ as

$$\widetilde{s_\eta}(s) \approx \frac{D_\eta}{D_0} s \tag{A.11}$$

that fulfills the conditions $\widetilde{s_\eta}(0) = 0$ and $\widetilde{s_\eta}(D_0) = D_\eta$.

Bibliography

[1] D. Kahng and M.M.Atalla, "Silicon-silicon dioxide field induced surface device," in *IRE-AIEE Solid-state Device Res.Conf.,(Carnegie Inst. of Technol., Pittsburgh, PA*, 1960.

[2] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, 1965.

[3] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, pp. 256–268, Oct 1974.

[4] C. Auth, C. Allen, A. Blattner, D. Bergstrom, M. Brazier, M. Bost, M. Buehler, V. Chikarmane, T. Ghani, T. Glassman, R. Grover, W. Han, D. Hanken, M. Hattendorf, P. Hentges, R. Heussner, J. Hicks, D. Ingerly, P. Jain, S. Jaloviar, R. James, D. Jones, J. Jopling, S. Joshi, C. Kenyon, H. Liu, R. McFadden, B. McIntyre, J. Neirynck, C. Parker, L. Pipes, I. Post, S. Pradhan, M. Prince, S. Ramey, T. Reynolds, J. Roesler, J. Sandford, J. Seiple, P. Smith, C. Thomas, D. Towner, T. Troeger, C. Weber, P. Yashar, K. Zawadzki, and K. Mistry, "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density mim capacitors," in *2012 Symposium on VLSI Technology (VLSIT)*, pp. 131–132, June 2012.

[5] J. Colinge, "Multi-gate SOI MOSFETs," *Microelectronic Engineering*, vol. 84, pp. 2071–2076, Jun 2007.

[6] N. Loubet, T. Hook, P. Montanini, C. W. Yeung, S. Kanakasabapathy, M. Guillom, T. Yamashita, J. Zhang, X. Miao, J. Wang, A. Young, R. Chao, M. Kang, Z. Liu, S. Fan, B. Hamieh, S. Sieg, Y. Mignot, W. Xu, S. C. Seo, J. Yoo, S. Mochizuki, M. Sankarapandian, O. Kwon, A. Carr, A. Greene, Y. Park, J. Frougier, R. Galatage, R. Bao, J. Shearer, R. Conti, H. Song, D. Lee, D. Kong, Y. Xu, A. Arceo, Z. Bi, P. Xu, R. Muthinti, J. Li, R. Wong, D. Brown, P. Oldiges, R. Robison, J. Arnold, N. Felix, S. Skordas, J. Gaudiello, T. Standaert, H. Jagannathan, D. Corliss, M. H. Na, A. Knorr, T. Wu, D. Gupta, S. Lian, R. Divakaruni, T. Gow, C. Labelle, S. Lee, V. Paruchuri, H. Bu, and M. Khare, "Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET," in *2017 Symposium on VLSI Technology*, pp. T230–T231, June 2017.

[7] T. Grasser, T.-W. Tang, H. Kosina, and S. Selberherr, "A review of hydrodynamic and energy-transport models for semiconductor device simulation," *Proceedings of the IEEE*, vol. 91, pp. 251–274, Feb 2003.

[8] T. Grasser, H. Kosina, M. Gritsch, and S. Selberherr, "Using six moments of Boltzmann's transport equation for device simulation," *Journal of Applied Physics*, vol. 90, no. 5, pp. 2389–2396, 2001.

[9] R. Grassi, A. Gnudi, I. Imperiale, E. Gnani, S. Reggiani, and G. Baccarani, "Mode space approach for tight-binding transport simulations in graphene nanoribbon field-effect transistors including phonon scattering," *Journal of Applied Physics*, vol. 113, no. 14, p. 144506, 2013.

[10] A. Gehring and H. Kosina, "Wigner function-based simulation of quantum transport in Scaled DG-MOSFETs using a Monte Carlo method," *Journal of Computational Electronics*, vol. 4, pp. 67–70, Apr 2005.

[11] L. Lucci, P. Palestri, D. Esseni, L. Bergagnini, and L. Selmi, "Multisubband Monte Carlo Study of Transport, Quantization, and Electron-Gas Degeneration in Ultrathin SOI n-MOSFETs," *IEEE Trans. on Electron Devices*, vol. 54, pp. 1156 –1164, may 2007.

[12] M. Lenzi, P. Palestri, E. Gnani, S. Reggiani, A. Gnudi, D. Esseni, L. Selmi, and G. Baccarani, "Investigation of the transport properties of silicon nanowires using deterministic and Monte Carlo approaches to the solution of the Boltzmann Transport Equation," *IEEE Trans. on Electron Devices*, vol. 55, pp. 2086–2096, Aug 2008.

[13] S. Datta, "Nanoscale device modeling: the Green's function method," *Superlattices and Microstructures*, vol. 28, no. 4, pp. 253 – 278, 2000.

[14] M. P. Anantram, M. S. Lundstrom, and D. E. Nikonov, "Modeling of nanoscale devices," *Proceedings of the IEEE*, vol. 96, pp. 1511–1550, Sept 2008.

[15] E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev.*, vol. 40, pp. 749–759, Jun 1932.

[16] H. Kosina, M. Nedjalkov, and S. Selberherr, "A Monte Carlo method seamlessly linking quantum and classical transport calculations," *Journal of Computational Electronics*, vol. 2, pp. 147–151, Dec 2003.

[17] V. Sverdlov, T. Grasser, H. Kosina, and S. Selberherr, "Scattering and space-charge effects in Wigner Monte Carlo simulations of single and double barrier devices," *Journal of Computational Electronics*, vol. 5, pp. 447–450, Dec 2006.

[18] C. Grillet, D. Logoteta, A. Cresti, and M. G. Pala, "Assessment of the electrical performance of short channel InAs and Strained Si nanowire FETs," *IEEE Transactions on Electron Devices*, vol. 64, pp. 2425–2431, May 2017.

[19] E. Gnani, A. Gnudi, S. Reggiani, and G. Baccarani, "Deterministic solution of the 1D Boltzmann transport equation: Application to the study of current transport in nanowire FETs," *Microelectronics Journal*, vol. 44, no. 1, pp. 20 – 25, 2013. Special Issue of Microelectronics Journal on the IEEE International MOS-AK/GSA Workshop on Compact Modeling 2010.

[20] P. Osgnach, E. Caruso, D. Lizzit, P. Palestri, D. Esseni, and L. Selmi, "The impact of interface states on the mobility and drive current of In0.53Ga0.47As semiconductor n-MOSFETs," *Solid-State Electronics*, vol. 108, pp. 90 – 96, 2015. Selected papers from the 15th Ultimate Integration on Silicon (ULIS) conference.

[21] E. Caruso, D. Lizzit, P. Osgnach, D. Esseni, P. Palestri, and L. Selmi, "Simulation analysis of III-V n-MOSFETs: Channel materials, Fermi level pinning and biaxial strain ," in *IEEE IEDM Technical Digest*, pp. 7.6.1–7.6.4, 2014.

[22] D. Lizzit, D. Esseni, P. Palestri, and L. Selmi, "A new formulation for surface roughness limited mobility in bulk and ultra-thin-body metal-oxide-semiconductor transistors," *Journal of Applied Physics*, vol. 116, no. 22, 2014.

[23] P. Osgnach, A. Revelant, D. Lizzit, P. Palestri, D. Esseni, and L. Selmi, "Toward computationally efficient multi-subband Monte Carlo simulations of nanoscale MOSFETs," in *2013 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 176–179, Sept 2013.

[24] P. Toniutti, D. Esseni, and P. Palestri, "Failure of the scalar dielectric function approach for the screening modelling in double-gate SOI MOSFETs," *IEEE Trans. on Electron Devices*, vol. 57, no. 11, pp. 3074–3083, 2010.

[25] "The International Technology Roadmap for Semiconductors (ITRS)," 2015.

[26] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C. H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz, B. McIntyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D. Parsons, C. Prasad, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, and K. Zawadzki, "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging," in *2007 IEEE International Electron Devices Meeting*, pp. 247–250, Dec 2007.

[27] Y. C. Yeo, X. Gong, M. van Dal, G. Vellianitis, and M. Passlack, "Germanium-based Transistors for Future High Performance and Low Power Logic Applications," in *IEEE IEDM Technical Digest*, pp. 2.4.1–2.4.4, 2015.

[28] J. Del Alamo, "Nanometre-scale electronics with III-V compound semiconductors," *Nature*, vol. 479, pp. 317 –323, 2011.

[29] G. Doornbos and M. Passlack, "Benchmarking of III-V n-MOSFET Maturity and Feasibility for Future CMOS," *IEEE Electron Device Lett.*, vol. 31, pp. 1110 –1112, oct. 2010.

[30] S. M. Goodnick, D. K. Ferry, C. W. Wilmsen, Z. Liliental, D. Fathy, and O. L. Krivanek, "Surface roughness at the Si(100)-SiO_2 interface," *Phys. Rev. B*, vol. 32, pp. 8171–8186, Dec 1985.

[31] T. Yamanaka, S. J. Fang, H.-C. Lin, J. P. Snyder, and C. Helms, "Correlation between inversion layer mobility and surface roughness measured by AFM," *Electron Device Letters, IEEE*, vol. 17, pp. 178–180, April 1996.

[32] A. Pirovano, A. Lacaita, G. Ghidini, and G. Tallarida, "On the correlation between surface roughness and inversion layer mobility in Si-MOSFETs," *Electron Device Letters, IEEE*, vol. 21, pp. 34–36, Jan 2000.

[33] O. Bonno, S. Barraud, D. Mariolle, and F. Andrieu, "Effect of strain on the electron effective mobility in biaxially strained silicon inversion layers: An experimental and theoretical analysis via atomic force microscopy measurements and Kubo-Greenwood mobility calculations," *Journal of Applied Physics*, vol. 103, no. 6, p. 063715, 2008.

[34] R. E. Prange and T.-W. Nee, "Quantum Spectroscopy of the Low-Field Oscillations in the Surface Impedance," *Phys. Rev.*, vol. 168, pp. 779–786, Apr 1968.

[35] S. Jin, M. Fischetti, and T.-W. Tang, "Modeling of Surface-Roughness Scattering in Ultrathin-Body SOI MOSFETs," *IEEE Trans. on Electron Devices*, vol. 54, pp. 2191–2203, Sept 2007.

[36] D. Esseni, P. Palestri, and L. Selmi, *"Nanoscale MOS Transistors - Semi-Classical Transport and Applications"*. Cambridge University Press., 1st ed., 2011.

[37] C. Jungemann, A. Emunds, and W. Engl, "Simulation of linear and nonlinear electron transport in homogeneous silicon inversion layers," *Solid State Electronics*, vol. 36, no. 11, pp. 1529 – 1540, 1993.

[38] M. V. Fischetti and S. E. Laux, "Monte Carlo study of electron transport in silicon inversion layers," *Phys. Rev. B*, vol. 48, pp. 2244–2274, Jul 1993.

[39] F. Gámiz, J. B. Roldán, J. A. López-Villanueva, P. Cartujo-Cassinello, and J. E. Carceller, "Surface roughness at the Si-SiO2 interfaces in fully depleted silicon-on-insulator inversion layers," *Journal of Applied Physics*, vol. 86, no. 12, pp. 6854–6863, 1999.

[40] D. Esseni, "On the modeling of surface roughness limited mobility in SOI MOSFETs and its correlation to the transistor effective field," *IEEE Trans. on Electron Devices*, vol. 51, pp. 394–401, March 2004.

[41] T. Ishihara, K. Uchida, J. Koga, and S. Takagi, "Unified Roughness Scattering Model Incorporating Scattering Component Induced by Thickness Fluctuations in Silicon-on-Insulator Metal–Oxide–Semiconductor Field-Effect Transistors," *Japanese Journal of Applied Physics*, vol. 45, no. 4B, pp. 3125–3132, 2006.

[42] S. Jin, M. Fischetti, and T.-W. Tang, "Modeling of electron mobility in gated silicon nanowires at room temperature: Surface roughness scattering, dielectric screening, and band nonparabolicity," *Journal of Applied Physics*, vol. 102, pp. 083715–083715–14, Oct 2007.

[43] I. M. Tienda-Luna, F. G. Ruiz, A. Godoy, B. Biel, and F. Gámiz, "Surface roughness scattering model for arbitrarily oriented silicon nanowires," *Journal of Applied Physics*, vol. 110, no. 8, pp. –, 2011.

[44] M. Poljak, V. Jovanović, D. Grgec, and T. Suligoj, "Assessment of Electron Mobility in Ultrathin-Body InGaAs-on-Insulator MOSFETs Using Physics-Based Modeling," *IEEE Trans. on Electron Devices*, vol. 59, no. 6, pp. 1636–1643, 2012.

[45] S. Koba, Y. Ishida, R. ans Kubota, Y. Tsuchiya, H. ans Kamakura, N. Mori, and M. Ogawa, "The Impact of Increased Deformation Potential at MOS Interface on Quasi-Ballistic Transport in Ultrathin Channel MOSFETs Scaled down to Sub-10 nm Channel Length," in *IEEE IEDM Technical Digest*, pp. 12.1.1–12.1.4, 2013.

[46] D. Lizzit, D. Esseni, P. Palestri, P. Osgnach, and L. Selmi, "Performance benchmarking and effective channel length for nanoscale InAs, $In_{0.53}Ga_{0.47}As$, and sSi n-MOSFETs," *IEEE Trans. on Electron Devices*, vol. 61, pp. 2027 – 2034, June 2014.

[47] E. G. Marin, F. G. Ruiz, A. Godoy, I. M. Tienda-Luna, C. Martínez-Blanque, and F. Gámiz, "Theoretical interpretation of the electron mobility behavior in InAs nanowires," *Journal of Applied Physics*, vol. 116, no. 17, p. 174505, 2014.

[48] F. Stern and W. E. Howard, "Properties of Semiconductor Surface Inversion Layers in the Electric Quantum Limit," *Phys. Rev.*, vol. 163, pp. 816–835, Nov 1967.

[49] S. Mori and T. Ando, "Intersubband scattering effect on the mobility of a Si (100) inversion layer at low temperatures," *Phys. Rev. B*, vol. 19, pp. 6433–6441, Jun 1979.

[50] T. Ando, A. B. Fowler, and F. Stern, "Electronic properties of two-dimensional systems," *Reviews of Modern Physics*, vol. 54, pp. 437–672, Apr 1982.

[51] Z. Stanojevic, O. Baumgartner, L. Filipovic, H. Kosina, M. Karner, C. Kernstock, and P. Prause, "Consistent low-field mobility modeling for advanced MOS devices," *Solid State Electronics*, vol. 112, pp. 37–45, 2015.

[52] D. Ferry and S. Goodnick, *Transport in Nanostructures*. Cambridge Studies in Semiconductor Physics and Microelectronic Engineering, Cambridge University Press, 1997.

[53] R. Baum, "The correlation function of Gaussian noise passed through nonlinear devices," *IEEE Trans. on Information Theory*, vol. 15, pp. 448–456, Jul 1969.

[54] A. D. Poularikas, *The Transforms and Applications Handbook*. CRC Press LLC, 2000.

[55] E. D. Siggia and P. C. Kwok, "Properties of electrons in semiconductor inversion layers with many occupied electric subbands. I. Screening and Impurity Scattering," *Phys. Rev. B*, vol. 2, pp. 1024–1036, Aug 1970.

[56] A. V. Oppenheim and G. C. Verghese, *Signals, Systems and Inference.* Prentice Hall, 2015.

[57] A. Paussa, F. Conzatti, D. Breda, R. Vermiglio, D. Esseni, and P. Palestri, "Pseudospectral methods for the efficient simulation of quantization effects in Nanoscale MOS transistors," *IEEE Transactions on Electron Devices*, vol. 57, pp. 3239–3249, Dec 2010.

[58] P. Palestri, C. Alexander, A. Asenov, V. Aubry-Fortuna, G. Baccarani, A. Bournel, M. Braccioli, B. Cheng, P. Dollfus, A. Esposito, D. Esseni, C. Fenouillet-Beranger, C. Fiegna, G. Fiori, A. Ghetti, G. Iannaccone, A. Martinez, B. Majkusiak, S. Monfray, V. Peikert, S. Reggiani, C. Riddet, J. Saint-Martin, E. Sangiorgi, A. Schenk, L. Selmi, L. Silvestri, P. Toniutti, and J. Walczak, "A comparison of advanced transport models for the computation of the drain current in nanoscale nMOSFETs," *Solid-State Electronics*, vol. 53, no. 12, pp. 1293 – 1302, 2009. Papers Selected from the Ultimate Integration on Silicon Conference 2009, ULIS 2009.

[59] M. Lundstrom, *"Nanoscale MOS Transistors - Semi-Classical Transport and Applications"*. Cambridge University Press., 2nd ed., 2009.

[60] C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev. Mod. Phys.*, vol. 55, pp. 645–705, Jul 1983.

[61] S. Jin, A. T. Pham, W. Choi, Y. Nishizawa, Y.-T. Kim, K.-H. Lee, Y. Park, and E. S. Jung, "Performance evaluation of InGaAs, Si, and Ge nFinFETs based on coupled 3D drift-diffusion/multisubband Boltzmann transport equations solver," in *2014 IEEE International Electron Devices Meeting*, pp. 7.5.1–7.5.4, Dec 2014.

[62] P. Lugli and D. K. Ferry, "Degeneracy in the ensemble Monte Carlo method for high-field transport in semiconductors," *IEEE Transactions on Electron Devices*, vol. 32, pp. 2431–2437, Nov 1985.

[63] H. Frohlich, "Theory of electrical breakdown in ionic crystals," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 160, no. 901, pp. 230–241, 1937.

[64] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: Part I-effects of substrate impurity concentration," *IEEE Trans. on Electron Devices*, vol. 41, pp. 2357–2362, Dec 1994.

[65] O. Badami, E. Caruso, D. Lizzit, P. Osgnach, D. Esseni, P. Palestri, and L. Selmi, "An Improved Surface Roughness Scattering Model for Bulk, Thin-Body, and Quantum-Well MOSFETs," *IEEE Trans. on Electron Devices*, vol. 63, pp. 2306–2312, June 2016.

[66] C. L. Hinkle, A. M. Sonnet, R. A. Chapman, and E. M. Vogel, "Extraction of the effective mobility of $In_{0.53}Ga_{0.47}As$ MOSFETs," *IEEE Electron Device Letters*, vol. 30, pp. 316–318, April 2009.

[67] Y. Xuan, Y. Q. Wu, T. Shen, T. Yang, and P. D. Ye, "High performance submicron inversion-type enhancement-mode InGaAs MOSFETs with ALD Al_2O_3, HfO_2 and

HfAlO as gate dielectrics," in *2007 IEEE International Electron Devices Meeting*, pp. 637–640, Dec 2007.

[68] T. B. Bahder, "Eight-band k·p model of strained zinc-blende crystals," *Phys. Rev. B*, vol. 41, pp. 11992–12001, Jun 1990.

[69] I. Vurgaftman, J. R. Meyer, and L. R. Ram-Mohan, "Band parameters for III–V compound semiconductors and their alloys," *Journal of Applied Physics*, vol. 89, no. 11, pp. 5815–5875, 2001.

[70] S. Kim, M. Yokoyama, N. Taoka, R. Iida, S. Lee, R. Nakane, Y. Urabe, N. Miyata, T. Yasuda, H. Yamada, N. Fukuhara, M. Hata, M. Takenaka, and S. Takagi, "High Performance Extremely Thin Body InGaAs-on-Insulator Metal-Oxide-Semiconductor Field-Effect Transistors on Si Substrates with Ni-InGaAs Metal Source/Drain," *Applied Physics Express*, vol. 4, no. 11, p. 114201, 2011.

[71] Y.-C. Yeo, T.-J. King, and C. Hu, "Metal-Dielectric Band Alignment and its Implications for Metal Gate Complementary Metal-Oxide-Semiconductor Technology dielectrics," *Journal of Applied Physics*, vol. 92, 2002.

[72] D. Shahrjerdi, E. Tutuc, and S. K. Banerjee, "Impact of surface chemical treatment on capacitance-voltage characteristics of GaAs metal-oxide-semiconductor capacitors with Al2O3 gate dielectric," *Applied Physics Letters*, vol. 91, no. 19, p. 063501, 2007.

[73] W. Zheng, K. H. Bowen, J. Li, I. Dabkowska, and M. Gutowski, "Electronic Structure Differences in ZrO_2 vs HfO_2," *The Journal of Physical Chemistry A*, vol. 109, no. 50, pp. 11521–11525, 2005. PMID: 16354043.

[74] H. Ishii, A. Nakajima, and S. Yokoyama, "Growth and electrical properties of atomic-layer deposited ZrO_2/Si-nitride stack gate dielectrics," *Journal of Applied Physics*, vol. 95, no. 2, pp. 536–542, 2004.

[75] S. Kim, M. Yokoyama, N. Taoka, R. Iida, S. Lee, R. Nakane, Y. Urabe, N. Miyata, T. Yasuda, H. Yamada, N. Fukuhara, M. Hata, M. Takenaka, and S. Takagi, "Electron Mobility Enhancement of Extremely Thin Body $In_{0.7}Ga_{0.3}As$-on-Insulator Metal-Oxide-Semiconductor Field-Effect Transistors on Si Substrates by Metal-Oxide-Semiconductor Interface Buffer Layers," *Applied Physics Express*, vol. 5, no. 1, p. 014201, 2012.

[76] K. Uchida, M. Saitoh, and S. Kobayashi, "Carrier transport and stress engineering in advanced nanoscale transistors from (100) and (110) transistors to carbon nanotube FETs and beyond," in *IEEE IEDM Technical Digest*, pp. 1–4, 2008.

[77] S. W. Wang, T. Vasen, G. Doornbos, R. Oxland, S. W. Chang, X. Li, R. Contreras-Guerrero, M. Holland, C. H. Wang, M. Edirisooriya, J. S. Rojas-Ramirez, P. Ramvall, S. Thoms, D. S. Macintyre, G. Vellianitis, G. C. H. Hsieh, Y. S. Chang, K. M. Yin, Y. C. Yeo, C. H. Diaz, R. Droopad, I. G. Thayne, and M. Passlack, "Field-effect mobility of

inas surface channel nMOSFET with low D_{rmit} scaled gate-stack," *IEEE Transactions on Electron Devices*, vol. 62, pp. 2429–2436, Aug 2015.

[78] K. J. Kuhn, "Considerations for Ultimate CMOS Scaling," *IEEE Trans. on Electron Devices*, vol. 59, pp. 1813–1828, July 2012.

[79] N. Waldron, C. Merckling, L. Teugels, P. Ong, S. A. U. Ibrahim, F. Sebaai, A. Pourghaderi, K. Barla, N. Collaert, and A. V. Y. Thean, "InGaAs Gate-All-Around Nanowire Devices on 300mm Si Substrates," *IEEE Electron Device Lett.*, vol. 35, pp. 1097–1099, Nov 2014.

[80] S. Barraud, J. M. Hartmann, V. Maffini-Alvaro, L. Tosti, V. Delaye, and D. Lafond, "Top-Down Fabrication of Epitaxial SiGe/Si Multi-(Core/Shell) p-FET Nanowire Transistors," *IEEE Trans. on Electron Devices*, vol. 61, pp. 953–956, April 2014.

[81] H. Mertens, R. Ritzenthaler, A. Chasin, T. Schram, E. Kunnen, A. Hikavyy, L.-A. Ragnarsson, H. Dekkers, T. Hopf, K. Wostyn, K. Devriendt, S. A. Chew, M. S. Kim, Y. Kikuchi, E. Rosseel, G. Mannaert, S. Kubicek, S. Demuynck, A. Dangol, N. Bosman, J. Geypen, P. Carolan, H. Bender, K. Barla, N. Horiguchi, and D. Mocuta, "Vertically Stacked Gate-All-Around Si Nanowire CMOS Transistors with Dual Work Function Metal Gates," in *IEEE IEDM Technical Digest*, pp. 19.7.1–19.7.4, 2016.

[82] M. Rau, E. Caruso, D. Lizzit, P. Palestri, D. Esseni, A. Schenk, L. Selmi, and M. Luisier, "Performance Projection of III-V Ultra-Thin-Body, FinFET, and Nanowire MOSFETs for two Next-Generation Technology Nodes," in *IEEE IEDM Technical Digest*, pp. 30.6.1–30.6.4, 2016.

[83] M. Karner, O. Baumgartner, Z. Stanojević, F. Schanovsky, G. Strof, C. Kernstock, H. Karner, G. Rzepa, and T. Grasser, "Vertically Stacked Nanowire MOSFETS for Sub-10 nm Nodes: Advanced Topography, Device, Variability, and Reliability Simulations," in *IEEE IEDM Technical Digest*, pp. 30.7.1–30.7.4, Dec 2016.

[84] B. A. Korgel, "Semiconductor nanowires: Twins cause kinks," *Nature Materials*, vol. 5, no. 7, pp. 521–522, 2006.

[85] S. D. Suk, M. Li, Y. Y. Yeoh, K. H. Yeo, K. H. Cho, I. K. Ku, H. Cho, W. Jang, D. W. Kim, D. Park, and W. S. Lee, "Investigation of nanowire size dependency on TSNWFET," in *IEEE IEDM Technical Digest*, pp. 891–894, Dec 2007.

[86] W. Lu, P. Xie, and C. M. Lieber, "Nanowire Transistor Performance Limits and Applications," *IEEE Trans. on Electron Devices*, vol. 55, pp. 2859–2876, Nov 2008.

[87] S. Kim, M. Luisier, A. Paul, T. B. Boykin, and G. Klimeck, "Full Three-Dimensional Quantum Transport Simulation of Atomistic Interface Roughness in Silicon Nanowire FETs," *IEEE Trans. on Electron Devices*, vol. 58, pp. 1371–1380, May 2011.

[88] R. G. Hobbs, N. Petkov, and J. D. Holmes, "Semiconductor Nanowire Fabrication by Bottom-Up and Top-Down Paradigms," *Chemistry of Materials*, vol. 24, no. 11, pp. 1975–1991, 2012.

[89] F. Wang, C. Wang, Y. Wang, M. Zhang, Z. Han, S. Yip, L. Shen, N. Han, E. Y. B. Pun, and J. C. Ho, "Diameter Dependence of Planar Defects in InP Nanowires," *Scientific Reports*, vol. 6, no. 32910, 2016.

[90] K. Moors, B. Sorée, and W. Magnus, "Modeling surface roughness scattering in metallic nanowires," *Journal of Applied Physics*, vol. 118, no. 12, p. 124307, 2015.

[91] Z. Stanojević and H. Kosina, "Surface-roughness-scattering in non-planar channels - The role of band anisotropy," in *Proc.SISPAD*, pp. 352–355, Sept 2013.

[92] S. Jin, S. M. Hong, W. Choi, K. H. Lee, and Y. Park, "Coupled drift-diffusion (DD) and multi-subband Boltzmann transport equation (MSBTE) solver for 3D multi-gate transistors," in *2013 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 348–351, Sept 2013.

[93] O. Badami, D. Lizzit, R. Specogna, and D. Esseni, "Surface roughness limited mobility in multi-gate FETs with arbitrary cross-section," in *IEEE IEDM Technical Digest*, pp. 36.1.1–36.1.4, Dec 2016.

[94] D. Lizzit, O. Badami, R. Specogna, and D. Esseni, "Improved surface-roughness scattering and mobility models for multi-gate fets with arbitrary cross-section and biasing scheme," *Journal of Applied Physics*, vol. 121, no. 24, p. 245301, 2017.

[95] M. Bescond, N. Cavassilas, and M. Lannoo, "Effective-mass approach for n-type semiconductor nanowire MOSFETs arbitrarily oriented," *Nanotechnology*, vol. 18, no. 25, p. 255201, 2007.

[96] D. J. BenDaniel and C. B. Duke, "Space-Charge Effects on Electron Tunneling," *Phys. Rev.*, vol. 152, pp. 683–692, Dec 1966.

[97] A. Paussa, R. Specogna, D. Esseni, and F. Trevisan, "Discrete geometric approach for modelling quantization effects in nanoscale electron devices," *Journal of Computational Electronics*, vol. 13, no. 1, pp. 287–299, 2014.

[98] E. Polizzi and S. Datta, "Multidimensional nanoscale device modeling: the finite element method applied to the non-equilibrium Green's function formalism," in *2003 Third IEEE Conference on Nanotechnology, 2003. IEEE-NANO 2003.*, vol. 1, pp. 40–43 vol.2, Aug 2003.

[99] H. U. Baranger and J. W. Wilkins, "Ballistic structure in the electron distribution function of small semiconducting structures: General features and specific trends," *Phys. Rev. B*, vol. 36, pp. 1487–1502, Jul 1987.

[100] S. Jin, T. W. Tang, and M. V. Fischetti, "Simulation of silicon nanowire transistors using Boltzmann transport equation under relaxation time approximation," *IEEE Transactions on Electron Devices*, vol. 55, pp. 727–736, March 2008.

[101] D. Lizzit, *"Advanced Models for Simulation of planar and Gate-All-Around Nanoscate MOSFETS"*. PhD thesis, Universitá Degli Studi di Udine, 2016.

[102] T. Wang, T. H. Hsieh, and T. W. Chen, "Quantum confinement effects on low-dimensional electron mobility," *Journal of Applied Physics*, vol. 74, no. 1, pp. 426–430, 1993.

[103] Z. Stanojević, M. Karner, M. Aichhorn, F. Mitterbauer, V. Eyert, C. Kernstock, and H. Kosina, "Predictive physical simulation of III/V quantum-well MISFETs for logic applications," pp. 310–313, Sept 2015.

[104] M. V. Fischetti, "Monte Carlo simulation of transport in technologically significant semiconductors of the diamond and zinc-blende structures. I. Homogeneous transport," *IEEE Trans. on Electron Devices*, vol. 38, no. 3, pp. 634–649, 1991.

[105] S. Monaghan, P. Hurley, K. Cherkaoui, M. Negara, and A. Schenk, "Determination of electron effective mass and electron affinity in HfO_2 using MOS and MOSFET structures," *Solid State Electronics*, vol. 53, no. 4, pp. 438 – 444, 2009.

[106] M. V. Fischetti, D. A. Neumayer, and E. A. Cartier, "Effective electron mobility in Si inversion layers in metal-oxide-semiconductor systems with a high-k insulator: The role of remote phonon scattering," *Journal of Applied Physics*, vol. 90, no. 9, pp. 4587–4608, 2001.

[107] M. Auf der Maur, M. Povolotskyi, F. Sacconi, A. Pecchia, and A. Di Carlo, "Multiscale simulation of MOS systems based on high-κ oxides," *Journal of Computational Electronics*, vol. 7, no. 3, pp. 398–402, 2008.

[108] H. Jia, Y. Zhang, X. Chen, J. Shu, X. Luo, Z. Zhang, and D. Yu, "Efficient field emission from single crystalline indium oxide pyramids," *Applied Physics Letters*, vol. 82, no. 23, pp. 4146–4148, 2003.

[109] T. Vasen, P. Ramvall, A. Afzalian, C. Thelander, K. A. Dick, M. Holland, G. Doornbos, S. W. Wang, R. Oxland, G. Vellianitis, M. J. H. van Dal, B. Duriez, J. R. Ramirez, R. Droopad, L. E. Wernersson, L. Samuelson, T. K. Chen, Y. C. Yeo, and M. Passlack, "InAs nanowire GAA n-MOSFETs with 12-15 nm diameter," in *IEEE Symposium on VLSI Technology - Technical Digest*, pp. 1–2, June 2016.

[110] D. H. Kim and J. A. del Alamo, "Scalability of sub-100 nm InAs HEMTs on InP substrate for future logic applications," *IEEE Transactions on Electron Devices*, vol. 57, pp. 1504–1511, July 2010.

[111] P. Toniutti, P. Palestri, D. Esseni, F. Driussi, M. D. Michielis, and L. Selmi, "On the origin of the mobility reduction in n- and p-metal-oxide-semiconductor field effect transistors with hafnium-based/metal gate stacks," *Journal of Applied Physics*, vol. 112, p. 034502, 2012.

[112] A. C. Ford, J. C. Ho, Y.-L. Chueh, Y.-C. Tseng, Z. Fan, J. Guo, J. Bokor, and A. Javey, "Diameter-Dependent Electron Mobility of InAs Nanowires," *Nano Letters*, vol. 9, no. 1, pp. 360–365, 2009.

[113] O. Badami, E. Caruso, D. Lizzit, D. Esseni, P. Palestri, and L. Selmi, "Improved surface roughness modeling and mobility projections in thin film MOSFETs," in *2015 45th European Solid State Device Research Conference (ESSDERC)*, pp. 306–309, Sept 2015.

[114] Y. Hironori, M. Naoya, S. Jun, and T. Kimoto, "Mobility oscillation by one-dimensional quantum confinement in Si-nanowire metal-oxide-semiconductor field effect transistors," *Journal of Applied Physics*, vol. 106, no. 8, pp. 034312–1–14, 2009.

[115] K. Raleva, D. Vasileska, S. M. Goodnick, and M. Nedjalkov, "Modeling thermal effects in nanodevices," *IEEE Transactions on Electron Devices*, vol. 55, pp. 1306–1316, June 2008.

[116] R. Rhyner and M. Luisier, "Self-heating effects in ultra-scaled Si nanowire transistors," in *2013 IEEE International Electron Devices Meeting*, pp. 32.1.1–32.1.4, Dec 2013.

[117] D. Li, Y. Wu, P. Kim, L. Shi, P. Yang, and A. Majumdar, "Thermal conductivity of individual silicon nanowires," *Applied Physics Letters*, vol. 83, no. 14, pp. 2934–2936, 2003.

[118] M. Luisier, "Investigation of thermal transport degradation in rough Si nanowires," *Journal of Applied Physics*, vol. 110, no. 7, p. 074510, 2011.

[119] M. G. Pala and A. Cresti, "Increase of self-heating effects in nanodevices induced by surface roughness: A full-quantum study," *Journal of Applied Physics*, vol. 117, no. 8, p. 084313, 2015.

[120] E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev.*, vol. 40, pp. 749–759, Jun 1932.

[121] M. V. Fischetti and S. E. Laux, "Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects," *Phys. Rev. B*, vol. 38, pp. 9721–9745, Nov 1988.

[122] R. Lake and S. Datta, "Nonequilibrium Green's-function method applied to double-barrier resonant-tunneling diodes," *Phys. Rev. B*, vol. 45, pp. 6670–6685, Mar 1992.

[123] O. Badami, D. Saha, and S. Ganguly, "Efficient Wigner function simulation for nanowire MOSFETs and comparison to quantum drift-diffusion," in *2013 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 272–275, Sept 2013.

[124] D. Vasileska, S. M. Goodnick, and G. Klimeck, *Computational Electronics: Semiclassical and Quantum Device Modeling and Simulation*. CRC Press, 1st ed., 2010.

[125] T. T. T. Nghiêm, J. Saint-Martin, and P. Dollfus, "Electro-thermal simulation based on coupled Boltzmann transport equations for electrons and phonons," *Journal of Computational Electronics*, vol. 15, pp. 3–15, Mar 2016.

[126] M. Hinz, O. Marti, B. Gotsmann, M. A. Lantz, and U. Dürig, "High resolution vacuum scanning thermal microscopy of HfO_2 and SiO_2," *Applied Physics Letters*, vol. 92, no. 4, p. 043122, 2008.

[127] J. Wang, E. Polizzi, and M. Lundstrom, "A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation," *Journal of Applied Physics*, vol. 96, no. 4, pp. 2192–2203, 2004.

[128] M. Rau, T. Markussen, E. Caruso, D. Esseni, E. Gnani, A. Gnudi, P. A. Khomyakov, M. Luisier, P. Osgnach, P. Palestri, S. Reggiani, A. Schenk, L. Selmi, and K. Stokbro, "Performance study of strained III-V materials for ultra-thin body transistor applications," in *2016 46th European Solid-State Device Research Conference (ESSDERC)*, pp. 184–187, Sept 2016.

[129] S. Kim, M. Yokoyama, Y. Ikku, R. Nakane, O. Ichikawa, T. Osada, M. Hata, M. Takenaka, and S. Takagi, "Physical understanding of electron mobility in asymmetrically strained InGaAs-on-insulator metal-oxide-semiconductor field-effect transistors fabricated by lateral strain relaxation," *Applied Physics Letters*, vol. 104, no. 11, p. 113509, 2014.

[130] K. C. Saraswat and F. Mohammadi, "Effect of scaling of interconnections on the time delay of VLSI circuits," *IEEE Journal of Solid-State Circuits*, vol. 17, pp. 275–280, Apr 1982.

[131] L. H. Liang and B. Li, "Size-dependent thermal conductivity of nanoscale semiconducting systems," *Phys. Rev. B*, vol. 73, p. 153303, Apr 2006.

[132] A. I. Hochbaum, R. Chen, R. D. Delgado, W. Liang, E. C. Garnett, M. Najarian, A. Majumdar, and P. Yang, "Enhanced thermoelectric performance of rough silicon nanowires," *Nature*, vol. 451, pp. 163–167, 2008.

[133] D. Rideau, M. Feraille, L. Ciampolini, M. Minondo, C. Tavernier, H. Jaouen, and A. Ghetti, "Strained Si, Ge, and $Si_{1-x}Ge_x$ alloys modeled with a first-principles-optimized full-zone $\mathbf{k} \cdot \mathbf{p}$ method," *Phys. Rev. B*, vol. 74, p. 195208, Nov 2006.

[134] L.-W. Wang, A. Franceschetti, and A. Zunger, "Million-atom pseudopotential calculation of γ- X mixing in $\mathrm{GaAs/AlAs}$ superlattices and quantum dots," *Phys. Rev. Lett.*, vol. 78, pp. 2819–2822, Apr 1997.

[135] L.-W. Wang and A. Zunger, "Linear combination of bulk bands method for large-scale electronic structure calculations on strained nanostructures," *Phys. Rev. B*, vol. 59, pp. 15806–15818, Jun 1999.

[136] F. Chirico, A. Di Carlo, and P. Lugli, "Efficient self-consistent pseudopotential calculation of nanostructured devices," *Phys. Rev. B*, vol. 64, p. 045314, Jun 2001.

[137] D. Esseni and P. Palestri, "Linear combination of bulk bands method for investigating the low-dimensional electron gas in nanostructured devices," *Phys. Rev. B*, vol. 72, p. 165342, Oct 2005.

[138] J. L. P. J. van der Steen, D. Esseni, P. Palestri, L. Selmi, and R. J. E. Hueting, "Validity of the parabolic effective mass approximation in silicon and germanium n-mosfets with different crystal orientations," *IEEE Transactions on Electron Devices*, vol. 54, pp. 1843–1851, Aug 2007.

[139] V. Sverdlov, O. Baumgartner, H. Kosina, S. Selberherr, F. Schanovsky, and D. Esseni, "The linear combination of bulk bands-method for electron and hole subband calculations in strained silicon films and surface layers," in *2009 13th International Workshop on Computational Electronics*, pp. 1–4, May 2009.

[140] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, "Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations," *Phys. Rev. B*, vol. 74, p. 205323, Nov 2006.

[141] N. Neophytou and H. Kosina, "Atomistic simulations of low-field mobility in Si nanowires: Influence of confinement and orientation," *Phys. Rev. B*, vol. 84, p. 085313, Aug 2011.

[142] N. Neophytou and H. Kosina, "Effects of confinement and orientation on the thermoelectric power factor of silicon nanowires," *Phys. Rev. B*, vol. 83, p. 245305, Jun 2011.

[143] J. C. Slater and G. F. Koster, "Simplified LCAO Method for the Periodic Potential Problem," *Phys. Rev.*, vol. 94, pp. 1498–1524, Jun 1954.

[144] N. Ashcroft and D. Mermin, *"Solid State Physics"*. Saunders College Publishing, 1st ed., 1976.

[145] J.-M. Jancu, R. Scholz, F. Beltram, and F. Bassani, "Empirical spds^* tight-binding calculation for cubic semiconductors: General method and material parameters," *Phys. Rev. B*, vol. 57, pp. 6493–6507, Mar 1998.

[146] G. Klimeck, R. Bowen, T. B. Boykin, C. Salazar-Lazaro, T. A. Cwik, and A. Stoica, "Si tight-binding parameters from genetic algorithm fitting," *Superlattices and Microstructures*, vol. 27, no. 2, pp. 77 – 88, 2000.

[147] M. Cardona, P. Y. Yu, and M. Cardona, *"Fundamentals of Semiconductors: Physics and Materials Properties"*. World Scientific Publishing Co. Pte. Ltd., 1st ed., 1995.

[148] P. Vogl, H. P. Hjalmarson, and J. D. Dow, "A semi-empirical tight-binding theory of the electronic structure of semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 44, no. 5, pp. 365 – 378, 1983.

[149] T. B. Boykin, G. Klimeck, and F. Oyafuso, "Valence band effective-mass expressions in the $\mathrm{sp}^3d^5s^*$ empirical tight-binding model applied to a Si and Ge parametrization," *Phys. Rev. B*, vol. 69, p. 115201, Mar 2004.

[150] W. A. Harrison, *"Elementary Electronic Structure"*. World Scientific Publishing Co. Pte. Ltd., 1st ed., 1999.

[151] S. L. Richardson, M. L. Cohen, S. G. Louie, and J. R. Chelikowsky, "Electron charge densities at conduction-band edges of semiconductors," *Phys. Rev. B*, vol. 33, pp. 1177–1182, Jan 1986.

[152] L. Silvestri, S. Reggiani, E. Gnani, A. Gnudi, and G. Baccarani, "A low-field mobility model for bulk, ultrathin body SOI and double-gate n-MOSFETs with different surface and channel orientations-part I: Fundamental principles," *IEEE Transactions on Electron Devices*, vol. 57, pp. 1567–1574, July 2010.

[153] N. Neophytou, S. G. Kim, G. Klimeck, and H. Kosina, "On the bandstructure velocity and ballistic current of ultra-narrow silicon nanowire transistors as a function of cross section size, orientation, and bias," *Journal of Applied Physics*, vol. 107, no. 11, p. 113701, 2010.

[154] S. Lee, F. Oyafuso, P. von Allmen, and G. Klimeck, "Boundary conditions for the electronic structure of finite-extent embedded semiconductor nanostructures," *Phys. Rev. B*, vol. 69, p. 045316, Jan 2004.

[155] S. Mukherjee, K. Miao, A. Paul, N. Neophytou, R. Kim, J. Geng, M. Povolotskyi, T. C. Kubis, A. Ajoy, B. Novakovic, J. Fonseca, H. Ilatikhameneh, S. Steiger, M. McLennan, M. Lundstrom, and G. Klimeck, "Band structure lab, https://nanohub.org/resources/bandstrlab," 2015.

[156] M. V. Fischetti and S. Narayanan, "An empirical pseudopotential approach to surface and line-edge roughness scattering in nanostructures: Application to Si thin films and nanowires and to graphene nanoribbons," *Journal of Applied Physics*, vol. 110, no. 8, pp. –, 2011.

[157] T. Ando, "Screening Effect and Quantum Transport in a Silicon Inversion Layer in Strong Magnetic Fields," *Journal of the Physical Society of Japan*, vol. 43, no. 5, pp. 1616–1626, 1977.

[158] E. Kreyszig, *Advanced Engineering Mathematics: Maple Computer Guide*. New York, NY, USA: John Wiley & Sons, Inc., 8th ed., 2000.