Large-Scale Simulations of Intrinsic Parameter Fluctuations in Nano-Scale MOSFETs

David T. Reid

Submitted to the University of Glasgow, Department of Electronics and Electrical Engineering, in fulfilment of the requirements for the degree of Doctor of Philosophy.

June 2010

Dedication

For Thomas Leith, OBE – may you snooze peacefully in front of the big war movie in the sky!

Abstract

Intrinsic parameter fluctuations have become a serious obstacle to the continued scaling of MOSFET devices, particularly in the sub-100 nm regime. The increase in intrinsic parameter fluctuations means that simulations on a statistical scale are necessary to capture device parameter distributions. In this work, large-scale simulations of samples of 100,000s of devices are carried out in order to accurately characterise statistical variability of the threshold voltage in a real 35 nm MOSFET. Simulations were performed for the two dominant sources of statistical variability – random discrete dopants (RDD) and line edge roughness (LER). In total ∼400,000 devices have been simulated, taking approximately 500,000 CPU hours (60 CPU years). The results reveal the true shape of the distribution of threshold voltage, which is shown to be positively skewed for random dopants and negatively skewed for line edge roughness. Through further statistical analysis and data mining, techniques for reconstructing the distributions of the threshold voltage are developed. By using these techniques, methods are demonstrated that allow statistical enhancement of random dopant and line edge roughness simulations, thereby reducing the computational expense necessary to accurately characterise their effects. The accuracy of these techniques is analysed and they are further verified against scaled and alternative device architectures. The combined effects of RDD and LER are also investigated and it is demonstrated that the statistical combination of the individual RDD and LER-induced distributions of threshold voltage closely matches that obtained from simulations. By applying the statistical enhancement techniques developed for RDD and LER, it is shown that the computational cost of characterising their effects can be reduced by 1–2 orders of magnitude.

Acknowledgements

I'd like to thank a multitude of good people and the odd shady character, all of whom without which this work would not have been possible. First of all, I'd like to thank my supervisors: Scott Roy, for just generally knowing everything and Campbell Millar, who conned me into doing a PhD in the first place. I'd also like to thank Asen Asenov, for his patience and encyclopaedic knowledge of MOSFETs, all the other members of the Device Modelling Group for all their help with my silly questions and my family and friends, for listening to me rant and/or bore them about my work.

In particular, special thanks to my parents for all their love and support over the years. And finally, to Xu Gu for her love and for occasionally arguing about science with me.

And to anyone I have missed, I'm sorry, you know who you are and thank you!

Publications

Conference Papers

- **SISPAD 2008 (Talk):** Dave Reid, Campbell Millar, Scott Roy, Richard Sinnott, Gordon Stewart, Graeme Stewart and Asen Asenov. Prediction of random dopant induced threshold voltage fluctuations in nanoCMOS transistors. *Proc. SISPAD 2008*, pp 21-24.

- **Silicon Nanoelectronics Workshop 2008 (Poster):** Dave Reid, Campbell Millar, Scott Roy, Richard Sinnott, Gordon Stewart, Graeme Stewart and Asen Asenov. An accurate statistical analysis of random dopant induced variability in 140,000 13nm MOSFETs.

- **IEDM 2008 (Invited Talk):** A. Asenov, S. Roy, A. R. Brown, G. Roy, C. Alexander, C. Riddet, C. Millar, B. Cheng, A. Martinez, N. Seoane, D. Reid, M. F. Bukhori, X. Wang, U. Kovac. Advanced simulation of statistical variability and reliability in nano CMOS transistors.

- **ULIS 2009 (Talk):** Dave Reid, Campbell Millar, Gareth Roy, Scott Roy and Asen Asenov. Efficient Simulations of 6σ V_T Distributions Due to Random Discrete Dopants. *Proc. ULIS 2009*, pp 23-26.

- **ESSDERC 2009 (Talk):** Dave Reid, Campbell Millar, Gareth Roy, Scott Roy and Asen Asenov. Understanding LER-induced Statistical Variability: A 35,000 Sample 3D Simulation Study. *Proc. ESSDERC 2009*, pp 423-426.

- **IEDM 2009 (Talk):** Dave Reid, Campbell Millar, Gareth Roy, Scott Roy and Asen Asenov. Statistical enhancement of combined simulations of RDD and LER variability: What can simulation of a 10^5 sample teach us?

Journal Papers

- Campbell Millar, Dave Reid, Gareth Roy, Scott Roy and Asen Asenov. Accurate statistical description of random dopant induced threshold voltage variability. *IEEE Electron Device Letters*, 29(8):946-948, August 2008.

- Urban Kovac, Dave Reid, Campbell Millar, Gareth Roy, Scott Roy and Asen Asenov. Statistical simulation of random dopant induced threshold voltage fluctuations due to statistical variation for 35nm channel length MOSFET. *Microelectronics Reliability*, 48:1572-1575, August 2008.

- Dave Reid, Campbell Millar, Scott Roy, Gareth Roy, Richard Sinnott Gordon Stewart, Graeme Stewart and Asen Asenov. Enabling Cutting-Edge Semiconductor Simulation through Grid Technology. *Philosophical Transactions of the Royal Society A*, 367:2573-2584, June 2009.

- Dave Reid, Campbell Millar, Gareth Roy, Scott Roy and Asen Asenov. Analysis of Threshold Voltage Distribution due to Random Dopants: A 100,000 Sample 3D Simulation Study. *IEEE Transactions on Electron Devices*, 56(10):2255-2263, October 2009.

- Dave Reid, Campbell Millar, Scott Roy and Asen Asenov. Understanding LER Induced MOSFET V_T Variability - Part I: 3D Simulation of Large Statistical Samples. *IEEE Transactions on Electron Devices*. Submitted.

- Dave Reid, Campbell Millar, Scott Roy and Asen Asenov. Understanding LER Induced MOSFET V_T Variability - Part II: Reconstructing the Distribution. *IEEE Transactions on Electron Devices*. Submitted.

Contents

List of Figures

List of Tables

Nomenclature

CDF Cumulative Distribution Function

CIC Cloud In Cell

DD Drift/Diffusion

DG Density Gradient

DIBL Drain-Induced Barrier Lowering

EOT Equivalent Oxide Thickness

IPF Intrinsic Parameter Fluctuations

KDE Kernel Density Estimation

LER Line Edge Roughness

MC Monte Carlo

NEGF Non-Equilibrium Green's Function

PDF Probability Density Function

QQ Quantile-Quantile

RDD Random Discrete Dopants

SCE Short Channel Effects

SOI Silicon On Insulator

SSR Statistically Significant Region

TCAD Technology Computer Aided Design

UTB Ultra Thin Body

V_T Threshold Voltage

Chapter 1

Introduction

The phenomenal growth of the semiconductor industry has been driven by the continuous scaling of transistors and the corresponding increase in complexity of integrated circuits. The trend in the growth of the number of transistors on a chip was first observed in 1965 by Gordon Moore [8] and formulated at the 1975 IEDM [9] in what has come to be known as Moore's Law. Moore's Law effectively states that the number of transistors on a chip will quadruple every three years [10], and has, in effect, become a self-fulfilling prophecy for the semiconductor industry, as the continuation of Moore's law now guides research and development.

In 1974, Robert Dennard proposed the scaling rules that became the basis for the aggressive scaling of CMOS [11] that we have seen over the last 4 decades. MOSFETs have now reached deep sub-micron dimensions ($< 100\,nm$), and despite a multitude of issues related to continued scaling, the pace of Moore's Law continues unabated. The progress of semiconductor technology is now generally guided by the International Technology Roadmap for Semiconductors [12], which has been published since 1992, and details the design parameters and technological innovations necessary to continue Moore's Law.

There has been a great deal of research into CMOS scaling below 100 nm [13, 14] and while bulk MOSFETs with gate lengths as small as 5 nm have been reported [15], these devices have generally failed to meet the requirements of

the ITRS. It is widely expected that bulk MOSFETs will be replaced with alternative architectures well before channel lengths of 5 nm reach mass production. Indeed, the ITRS currently predicts the end of bulk CMOS in 2015 with gate lengths of 17 nm and the focus for extreme scaling has moved away from conventional architectures. As a result there has recently been extensive research into nanometer scale devices with Silicon-on-Insulator (SOI), FinFET and multi-gate architectures, which have better electrostatic integrity and scaling properties [16, 17, 18].

Problems arise at such small scales due to imperfections in the fabrication process and from random statistical variations in the fundamental atomic structure of devices. This is due to the fact that charge and matter are fundamentally discrete. It is from these variations and imperfections that intrinsic parameter fluctuations (IPF) of transistors arise and it is recognised that these variations will pose a serious obstacle to the continued scaling of CMOS [19]. Various techniques have been developed to improve variations arising from process tools, which are largely systematic and therefore predictable. These include techniques such as optical proximity correction. The nature of random variations however means that these cannot be eliminated no matter how much the fabrication process is improved and they must instead be accounted for during the design phase and minimized through device and circuit design.

Although alternative device architectures show less statistical variability of transistor parameters than the conventional architecture, bulk MOSFETs are still the workhorse of the semiconductor industry. In addition, the lifetime of bulk CMOS was extended in the 2008 update of the roadmap [12, 20]. The ITRS also indicates that bulk MOSFETs will still be in use after the introduction of ultra-thin body (UTB) SOI devices and it is eminently clear that bulk devices are and will continue to be an essential foundation for the semiconductor industry. This is why this work is primarily concerned with intrinsic parameter fluctuations in bulk MOSFETs.

It has also been suggested that Silicon in the transistor channel may be replaced by Silicon Germanium ($SiGe$), Germanium (Ge) or III-V compound semiconductors such as Gallium Arsenide (GaAs), Indium Gallium Arsenide (InGaAs) and Indium Antimonide (InSb) [21]. However, in the past, the

advantages of III-V-based transistors in increasing drive current and mobility have been outweighed by the advantages of Silicon, namely its material strength and the quality of its native oxide, SiO_2. To make alternative channel materials commercially viable, the difficulties in the development of a good quality gate dielectric and the integration of $SiGe$, Ge or III-V channels onto Si substrates must be overcome.

Recently, there has also been a lot of research into so-called 3D integration [22], in which there are multiple active layers in a chip. Apart from increasing the integration density, this has the advantage of reducing the average wire length and consequently reducing signal propagation times. Significant performance gains are therefore likely and it has been shown that power can be reduced by $\sqrt{N}$ [23] and frequency increased by $\sqrt{N^3}$ [24], where N is the number of layers. The fabrication process will not be straightforward however and there are likely to be issues with yield, alignment and reliability.

As stated above, the scaling of MOSFETs to deep sub-micron dimensions means that the fundamental discrete nature of charge and matter starts to become an issue. Intrinsic parameter fluctuations occur, for example, because the placement of dopants in devices cannot be exactly controlled and variations in the device structure start to influence device behaviour. These variations are random in nature and understanding the effect they have on device performance is an important problem. In order to understand and improve the operation of MOSFETs, it is common to use numerical modelling and simulations [25]. This allows the transistor characteristics to be better understood and designs to evolve, thus improving device performance. Understanding statistical variations, in particular the tails of parameter distributions, is a key problem in large integrated systems and enabling correct predictions of chip yields and performance and is an area that still requires significant research effort.

1.1 Aims and Objectives

With continued scaling comes increasing variability and coupled with the ever increasing number of devices per chip, it becomes increasingly difficult to predict the behaviour of intrinsic parameters with sufficient accuracy. Although MOSFETs were previously studied as single idealised devices, the influence of IPF has forced the use of statistical simulations in order to accurately predict the statistical nature of the operational characteristics of the device [26]. Such predictive simulations are essential in order to meet yield, power and performance targets and to ensure the long term reliability of chips.

The aim of this project is to study intrinsic parameter fluctuations in detail using large scale statistical simulations, which are enabled by Grid computing technology. The goal is to understand the statistical properties of nano-scale bulk MOSFETs subject to intrinsic parameter fluctuations in detail. This will enable the shapes of parameter distributions to be accurately deduced and related to the underlying device structure.

To achieve this aim, the project has the following objectives:

- Develop/adapt tools to facilitate and administer the simulation of 100,000s of devices on large compute resources.
- Simulate large ensembles of MOSFETs subject to two of the key variability sources in deep sub-micron devices – random discrete dopants (RDD) and line edge roughness (LER) – both individually and in combination.
- Through data mining and statistical analysis, relate the underlying device physics to intrinsic parameter fluctuations.
- Develop physically informed statistical enhancement methodologies to reduce the computational cost of characterising variability.

1.2 Outline

The remainder of this thesis is organised as follows. Chapter 2 reviews the background associated with the problem of intrinsic parameter fluctuations

and transistor variability. The main sources of variability, including random discrete dopants and line edge roughness, are described. The physical origins of the different sources are discussed, along with their effect on MOSFET parameters. The simulation techniques commonly used to study intrinsic fluctuations are also outlined and the strengths and weaknesses of each approach compared.

In Chapter 3, the simulation methodology is described in more detail. An overview of the Glasgow 3D atomistic simulator is given and the implementation of the sources of variability considered in this work is described. The extensive computational efforts undertaken here require the use of grid technology, and some aspects of this technology are outlined. The primary test bed device studied in this work is introduced and details of the device structure and characteristics are given.

Chapter 4 presents the results obtained from the simulation of samples of 100,000 devices with random discrete dopants. These results are statistically analysed and the factors contributing to random dopant fluctuations are investigated by employing data mining techniques. Further analysis is carried out to show how the distribution of threshold voltage can be reconstructed from the underlying device properties. A statistical enhancement methodology for random dopant simulations is developed and the accuracy of this approach is analysed by validating results against the original simulation data.

The impact of line edge roughness is investigated in Chapter 5, where the results of simulations of samples of 10,000–25,000 devices are presented. The corresponding device parameters contributing to LER-induced variability are deduced and two approaches for reconstructing the distribution of threshold voltage are proposed. Details of statistical enhancement methodologies for LER simulations are also given. The width dependence of LER-induced variability is analysed and the developed statistical enhancement methodology is applied and verified against simulation data. Finally, the impact of LER on alternative device architectures is investigated and the enhancement methodology again verified through comparison to simulation data.

In Chapter 6, the results of simulations of the combined effects of random dopants and line edge roughness are presented. Again, samples of 100,000

devices have been simulated in order to accurately capture the true shape of the distribution of threshold voltage. The methodologies developed in the previous two chapters are revisited and used to reconstruct to distribution of V_T due to combined RDD and LER-induced fluctuations. The accuracy of the statistical enhancement techniques is verified and the saving in computational cost by employing these techniques is demonstrated.

Finally, the conclusions for this work are drawn in Chapter 7. The main results are summarised and suggestions are made for possible extensions to this work that may be carried out in the future.

Chapter 2

Background

The continued scaling of conventional bulk MOSFETs has been one of the primary focuses of semiconductor research since Dennard proposed his generalized scaling rules [13]. In the last decade, the semiconductor industry has devoted much effort to scaling MOSFETs into the nanometer regime and MOSFETs in the 45 nm technology generation with channel lengths of 35 nm are currently in mass production [27]. Physical gate lengths below 30 nm are expected in the 32 nm technology generation [28], and according to the ITRS [12] physical gate lengths down to 17 nm are predicted for the end of applicability of bulk MOSFETs. SOI and eventually multi-gate devices are expected to take over at sub-16 nm technology generations. However, despite extensive research into alternative architectures, bulk MOSFETs remain the present device architecture of choice due to the difficulties in the large-scale integration of these new device architectures [29]. The cost and risk associated with transitioning to new architectures is a hugely important consideration for integrated device manufacturers (IDMs) and foundries.

One of the significant problems associated with further scaling is the statistical variability of transistor parameters. Transistor dimensions are now measureable in atomic-scale units and, as a result, self-averaging of atomic scale fluctuations and imperfections in the transistor structure is no longer taking place. MOSFET fabrication processes cannot be controlled precisely at atomic scales, meaning that the number and position of individual atoms

is random. Consequently, variations in MOSFET performance occur due to factors such as random dopant placement, atomic scale interface roughness and gate morphology and structure. For example, only a few tens of atoms in the channel dominate the behaviour of a sub-0.1 μm gate length MOSFET. Variations in the physical channel length also mean that short channel effects (SCE) become increasingly important, particularly at high drain voltages. This results in drain induced barrier lowering (DIBL), which reduces the height of the potential barrier between the source and drain and leads to threshold voltage (V_T) lowering and increased sub-threshold leakage current. In particularly bad cases, the barrier may be degraded to such an extent that punch-through occurs and significant drain current flows through the device regardless of the applied gate voltage. Such variations also make it difficult to scale the supply and threshold voltages, since in order to decrease leakage, V_T must be increased, however this reduces the drive current and thus adversely affects timing. The chip performance requirement also prevents supply voltage scaling, and has led to a significant increase in static power dissipation, to the point where static power is now a significant fraction of the total power dissipation [30].

Leakage [27] and V_T variability [31] have been significantly improved by the introduction of high-κ metal gates, however these introduce new sources of variability into the picture. It is clear that intrinsic parameter fluctuations are already a major challenge facing the semiconductor industry and will continue to be a problem in the future.

In this chapter, some of the primary sources of statistical variability in bulk MOSFETs are described and the particular sources investigated in this work are examined in further detail. Details of some of the common simulation techniques used to study statistical variability are given. Specifically, drift/diffusion, Monte Carlo and non-equilibrium Green's functions approaches are described, with a focus on their applicability for the large-scale simulations to be carried out in this work.

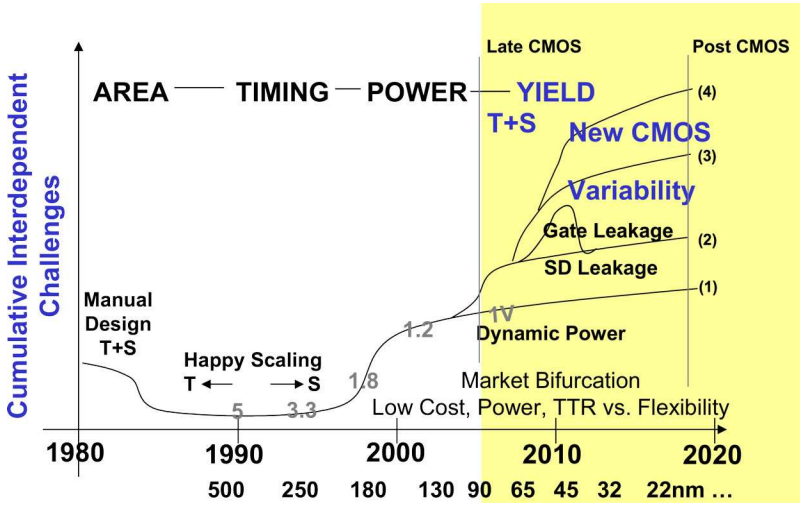


Figure 2.1: Design challenges for CMOS. After [1].

2.1 Intrinsic Parameter Fluctuations

It is commonly acknowledged that variability in device characteristics will be one of the major challenges as MOSFETs are scaled into the nanometer regime [1]. The "happy scaling" years are over, and design in the nanometer regime faces a great many challenges. The cumulative challenges facing the continued scaling of CMOS technology are shown in Figure 2.1. The impact of these challenges on the semiconductor industry is dramatic: in the 2007 ITRS [12] MOSFETs with physical gate lengths of 13 nm were predicted to reach mass production around 2013. However, in the 2008 update, this has been replaced with 18 nm and 13 nm postponed until 2017. In addition, for high performance (HP) applications, the lifespan of bulk MOSFETs has been extended by 4 years until 2016 and the introduction of multi-gate devices delayed by 4 years from 2011 until 2015. In the 2009 roadmap update, the down scaling of physical gate length for HP transistors was further shifted by a year and the end of bulk MOSFETs is now forecast for 2015. There are many factors that contribute to these difficulties, and variability of intrinsic parameters is one of the most significant.

At the device level, there are both systematic and statistical sources of variation. Systematic variations arise, for example, due to imperfections in

lithography process and are largely deterministic and although complex are predictable. As a result, they fall under a measure of control and methods exist to compensate for their effects [32]. For example, distortions that occur during the photo-lithography process can be corrected for using Optical Proximity Correction (OPC). In addition, strain, which is widely used to enhance transistor performance, causes both microscopic statistical variations and larger scale deterministic variations that primarily arise from the circuit layout. In this case, the layout dependent fluctuations are more significant than the local variations and as a result, strain can be effectively treated as a systematic source of variability and analysed using TCAD tools and SPICE [33, 34]. A more significant problem, however, stems from entirely statistical intrinsic parameter fluctuations (IPF) and the inherent randomness associated with them. Statistical fluctuations arise due to the fundamental discrete nature of charge and matter and the inability to precisely control, for example, the placement of dopant atoms in a device. These sources already account for more than 50% of the total variability in the current 45 nm technology generation [35, 36], and are expected to have a significantly greater influence at the 32 nm and future technology generations [37]. It is thus clear that statistical fluctuations are of critical importance to the future of CMOS scaling and integration [38].

In the past, the fabricated dimensions of transistors in CMOS chips have been much larger than atomic scales. It was thus entirely reasonable to use a continuous approximation of the device structure in order to simulate device behaviour, since average values taken over arbitrary volumes at transistor scale would not vary significantly. As transistor dimensions have approached atomic scale, where there may only be a few tens of dopant atoms in the active device region, this approximation is no longer valid [39]. If we consider a $0.25\,\mu m$ technology transistor (in production $\sim$ 1997) with a physical gate length $L_g = 200\,nm$, junction depth $x_j = 100\,nm$ and an average channel doping $N = 10^{18}\,cm^{-3}$, this would contain approximately 4000 dopants in the channel region. Assuming that the number of dopants follows a Poisson distribution, the standard deviation of the number of dopants ($\sqrt{4000}$) represents $\sim 1.6\%$ of the total doping. In comparison, a 65 nm technology transistor in production in 2007, with $L_g = 35\,nm$, $x_j = 18\,nm$ and $N = 2 \times 10^{18}\,cm^{-3}$ will

have approximately 40 dopants in the channel region. The standard deviation of the number of dopants is then $\sqrt{40}$, which represents $\sim$ 16% of the total doping - clearly a much more significant fluctuation. Furthermore, at 3σ this represents a variation of nearly 50% in the number of dopants, compared to 5% for the 200 nm transistor. It is clear that in the sub-100 nm regime, describing a transistor as being continuously doped and having smooth interfaces is no longer sufficient. It is essential to realise that it is not sufficient to study just one idealised device and to incorporate these atomic scale fluctuations into models and simulation tools. In this context, it is essential to study ensembles of devices that are microscopically different and the statistics associated with these ensembles in order to make predictions that can be used at higher levels of abstraction in the design hierarchy.

Not only are the absolute variations in device parameters larger as devices are scaled, but due to the huge numbers of transistors integrated in modern chips, it is also more likely that very rare devices will be encountered than in the past. These statistically rare devices must be factored in the higher levels of design. By way of example, consider a probability that a transistor is completely non-functional of 1 in 10 million. This translates to 200 non-functional devices on the latest 2 billion transistor chips. In addition, there will also be a significant fraction of devices that operate poorly. This already profoundly affects SRAM design [40], and, in logic circuits causes statistical timing problems [41] and hard digital faults [42]. In both cases, statistical variability restricts threshold and supply voltage scaling causing static and dynamic power dissipation problems [43].

As illustrated in Figure 2.2, there are several sources of statistical variability, which arise from the fundamental discrete nature of charge and matter. Variations in the number and position of dopant atoms (i.e. random discrete dopants) affect the electrostatics and carrier transport in devices. The molecular nature of the photoresists used to pattern devices causes line edge roughness (LER) and variations in the gate geometry. Variations also arise from the structure of the gate stack. For example, there will be variations in the local thickness of the gate oxide. In addition, in high-κ/metal gates, the metal is composed of crystal grains with different orientations and sizes. This

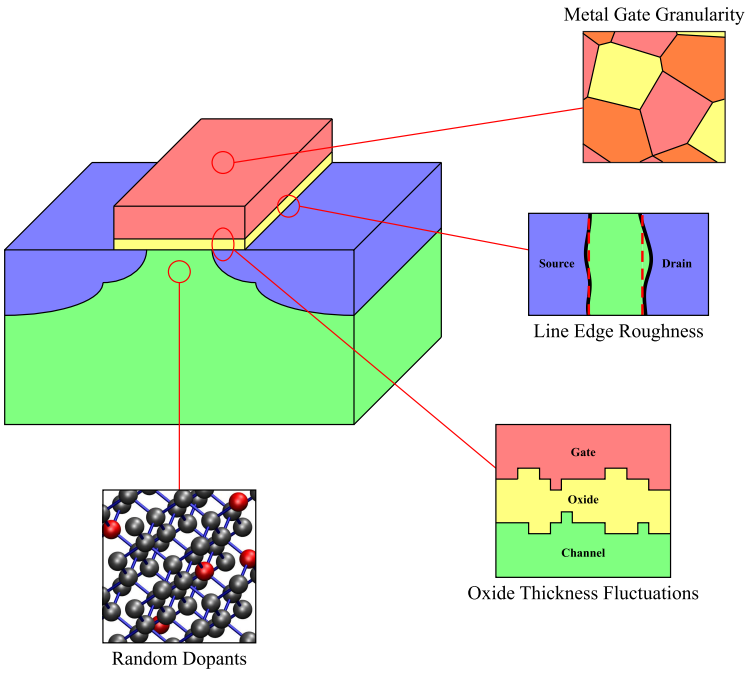


Figure 2.2: Illustration of some of the key sources of statistical variability in bulk MOSFETs.

leads to local variations in the gate work function, which affects the threshold voltage [44, 45].

Random discrete dopants (RDDs) in the channel and source/drain regions are the dominant source of statistical variability in contemporary bulk MOSFETs, which continue to be the CMOS workhorse at the 45 nm and 32 nm technology generations [46, 47, 48]. While RDDs are currently the primary source of statistical variability, the contribution of LER is becoming more important due to the fact that LER scaling lags the ITRS requirements [3]. While new device architectures such as SOI and FinFETs tolerate low channel doping, which reduces RDD variability, they are highly susceptible to the effects of LER. Recent simulation results predict that LER induced variability could overtake RDD variability in bulk; SOI and double gate devices, particularly at high drain voltages [49].

In the design of state of-the-art SRAM and flash memories, occurrences of devices more than 6-7σ from the mean now play important roles, creating

the necessity for detailed knowledge of the exact shape of statistical distributions far out into their tails [50]. Traditionally, simulations of small statistical samples (approximately 200 devices) have been used to identify and study the different sources of statistical variability, including random dopant effects [51, 52, 53, 54, 55]. However, using such small sample sizes it is not possible to accurately analyse the shape of the distribution or to accurately determine the impact of variability as far as 6 or 7σ from the mean. The question remains open as to what extent the different sources of variability, and random discrete dopants in particular, are capable of producing noticeable variations at 6 or 7σ.

Statistical variations will inevitably have a negative impact on the overall yield of a fabrication process and on the robustness of a design, making it essential that parameter fluctuations are properly incorporated into the design process so that feasible design margins can be established and, where possible, the design optimized to account for statistical variability [56]. Predictive simulation is therefore of great importance, as variations can be introduced into device models in a controlled manner that would be difficult and expensive, if not impossible, in a real experimental setting. It also allows individual sources of variability to be studied in isolation and thus in greater detail.

It should, of course, be acknowledged that in advanced emerging device architectures variability due to particular sources may be reduced or may disappear completely. For example, in devices that tolerate low channel doping, such as FinFETs, random dopant induced variability is drastically reduced [53]. It is also the case however, that other sources of variability will emerge in these devices, such as the impact of roughness and variations in fin shape on FinFET devices. Since transistors will continue to shrink, regardless of their architecture, fewer and fewer atoms will determine the behaviour of individual devices, meaning that statistical variability will continue to be a major factor in design and fabrication.

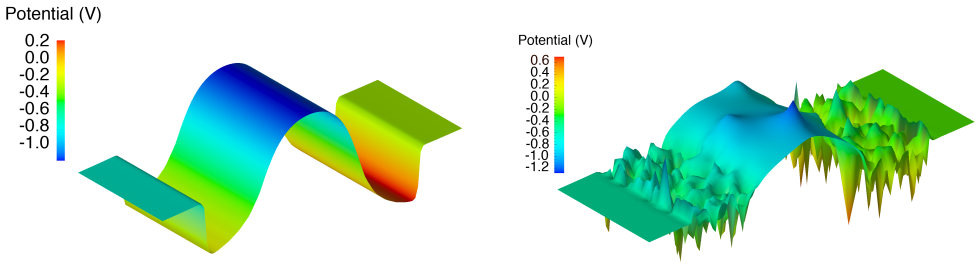
While 3D device simulation is an essential tool in predicting statistical device behaviour, it is still a computationally intensive task and it is impractical to consider simulating the large numbers of devices that are contained in cutting edge processors. Yet, in order to make valid predictions about the

impact of statistical variability on yield, it is necessary to have a detailed understanding of how the tails of the device parameter distributions behave. It is for this reason that we wish to study methods by which it may be possible to predict device behaviour, even for very rare devices, through statistical enhancement of a much smaller data set, facilitated by an understanding of the physics involved, rather than employing a purely brute force approach. It should be expected however that some degree of brute force will be involved, since without a good, physically accurate approximation of the underlying distributions involved, it will not be possible to verify the developed, physics led, statistically enhanced approaches.

2.1.1 Random Dopants

As discussed above, the scaling of transistor dimensions to sub-100 nm lengths results in a relatively small number of dopants in the active region of the device. Statistical averaging of the doping concentration is no longer a valid assumption and the behaviour of the device is determined by the number and individual position of dopant atoms. Local variations in the doping profile of the device lead to inhomogeneity in the potential profile. Consequently, certain parts of the device will turn on before others, resulting in a lowering of the average threshold voltage as compared to a continuously doped device, as well as variation around the mean value [57]. The inhomogeneity in the potential is demonstrated in example potential profiles obtained for continuously doped and atomistic devices, which are presented in Figures 2.3(a) and 2.3(b) respectively.

Random discrete dopants (RDD) are one of the primary sources of variability in bulk MOSFETs [51] and have been shown to contribute 60-65% of the total variability in measurements of 65- and 45 nm bulk Silicon devices [19]. In terms of predicting device characteristics through simulation, it has previously been reasonable to use a continuous approximation, where a given device is continuously doped, and as a result the carrier concentration, potential profile and interfaces are all smooth. This approximation is valid only as long as variations in number and position of dopants in the device have negligible effects



(a) (b)

Figure 2.3: Potential profiles of (a) a continuously doped device and (b) an atomistic device.

on the device characteristics.

The fabrication of modern MOSFETs involves several implantation steps [58], which are required for threshold voltage control; well implantations; source and drain implantations and extensions; and pocket implants to reduce short channel effects. Dopants are implanted at high energy and scatter many times before coming to rest. Thermal annealing then allows the implanted dopant atoms to replace Si atoms in the crystal lattice and become electrically active. Dopants diffuse during the annealing process, adding further randomness to the distribution. The net result of this process is that each device will have a particular doping distribution specific to that device due to the random nature of the scattering processes during implantation and diffusion during annealing. In general, the final positions of dopants are treated as being uncorrelated, although it has been suggested that Coulombic interactions during annealing may lead to correlations in dopant position and thus a more ordered distribution of dopants than would otherwise be expected [59]. A sketch of a 4.2 nm MOSFET, which represents an extreme scaling scenario and is of a comparable size to small biological features such as ion channels [60], is shown in Figure 2.4 with the silicon lattice overlaid and dopant positions indicated. This clearly indicates the scale of the problem, in that only a handful of dopants determine the behaviour of the device at this scale.

Since the doping distribution that results from the fabrication process is specific to a particular device, each device will have a slightly different thresh-

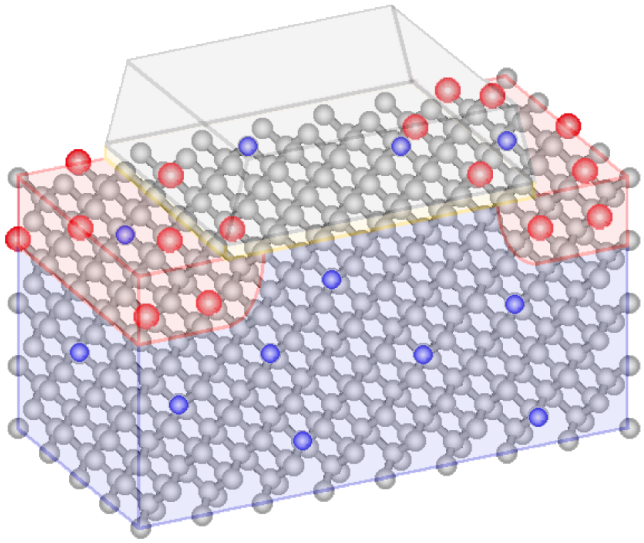


Figure 2.4: Sketch of a 4.2 nm gate length MOSFET, with silicon crystal lattice and dopant positions superimposed. After [2].

old voltage, as determined by the influence of individual dopant atoms on the potential. This effect was predicted several decades ago, in the early seventies [61, 46], and was subsequently confirmed experimentally in the late 80s and in a number of studies since [47, 62, 63, 64]. The effects of random dopants on MOSFET characteristics has been studied extensively both analytically [65, 66, 67] and numerically in 2D [66, 68, 48] and in 3D [69, 39]. In general, the doping concentrations are obtained from continuous doping profiles and dopant positions are generated from this using techniques such as the one described in Section 3.1.1 [68, 69, 70, 57]. It is also possible to model the distribution of dopant atoms using an atomic scale process simulator, which models the fabrication process and traces the actual implantation of dopant atoms in a realistic manner using Monte Carlo procedures [71, 72]. This, however, vastly increases the computational expense and in most instances the straightforward approach of generating dopants from the continuous doping profile is sufficiently accurate [2].

Through predictive 3D simulation, it has been shown that the introduction of random dopants causes a lowering of the average threshold voltage of the ensemble, as compared to the same device with continuous doping [69, 73].

It has also been shown that random dopant induced V_T variations increase significantly at very small (sub-20 nm) channel lengths [51]. This occurs because although the doping must be increased in smaller devices to overcome short channel effects, the physical size of the device also reduces to such an extent that the total number of dopants in the active region is also reduced. It is also important to note that it is not just the number of dopants that is important, but also the position of the dopants. In some cases the position of dopants has been neglected [54] and this leads to an underestimation of the variability and to the truncation of the lower tail of the distribution [74]. This is partially due to the fact that it is dopants close to the interface that have the greatest impact on the threshold voltage [75, 76]. The effect of dopant position on the threshold voltage is fully examined in this work in Chapter 4. Since the dopants close to the interface exert the greatest influence on V_T variability, devices that have lowly doped and undoped channels, such as SOI [77] and FinFET devices [78], can achieve a significant reduction in the variability caused by random dopants [79].

It has also been shown that the incorporation of quantum effects in 'atomistic' simulations results in an increase in both the variation and lowering of the threshold voltage due to the increase in equivalent oxide thickness (EOT) associated with the quantum mechnical charge distribution in the channel [80, 81]. Introducing quantum mechanical (QM) effects into the simulation also allows discrete dopants to be properly resolved. It should be noted that in sub-100 nm MOSFETs, the introduction of quantum effects into the simulation ordinarily results in a *positive* shift in the threshold voltage, due to quantization in the direction normal to the interface. However, the additional incorporation of random dopants results in a negative shift in V_T that depends on the doping concentration and may compensate for the positive shift associated with QM effects [73].

2.1.2 Line Edge Roughness

Although random discrete dopants are currently the primary source of statistical variability in conventional bulk MOSFETs, the contribution from line

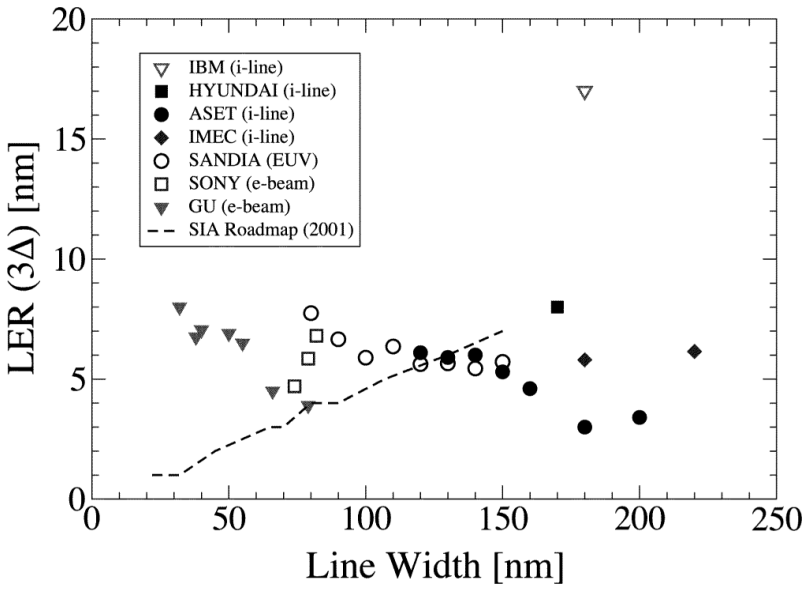


Figure 2.5: Data for LER reported by various labs that demonstrates the non-scaling of LER. The magnitude is on average about 5 nm. After [3].

edge roughness (LER) is increasing with continued downscaling of MOSFET devices. The problem is also exacerbated by the fact that LER scaling currently lags the requirements of the ITRS [82, 83, 3], as shown in Figure 2.5. Simulations of the LER formation process have demonstrated the difficulties associated with reducing the magnitude of the roughness [84]. A simulation study of the effects of LER on device performance has shown that if the magnitude of the roughness is not reduced below its current levels then LER will overtake random dopants as the dominant source of statistical variability at around 18 nm channel lengths [51]. LER is also expected to cause comparable fluctuations to RDD in SOI and double-gate devices at low drain voltage and to overtake RDD at high drain voltage [85, 86]. Recent 2D and 3D simulations of FinFETs have also shown that LER has a much greater influence on device performance than RDD [53].

LER arises due to the discrete molecular nature of the photoresists used during the fabrication process. During the fabrication process, the wafer is spin coated with photoresist before being selectively exposed to UV light through the photomask in order to pattern the transistor gates. The wafer is then

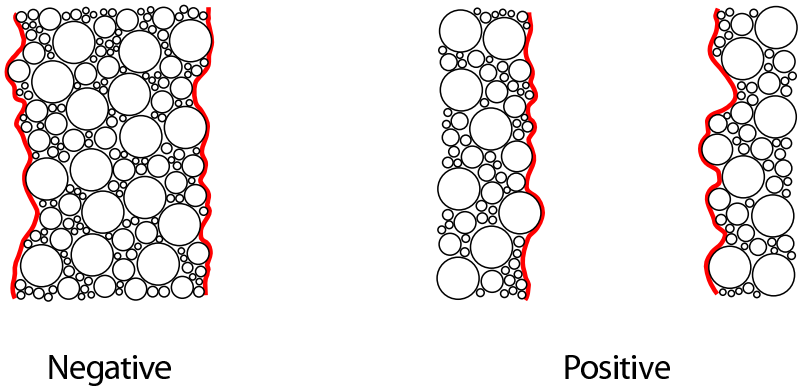


Figure 2.6: Roughness patterns in positive and negative resists. After [2].

baked and spin coated with the corresponding developer for the photoresist. Depending on the type of resist, either the exposed or unexposed areas are soluble in the developer. This leaves behind the image on the mask in the case of positive photoresist, or its inverse in the case of negative resist. Example roughness patterns corresponding to positive and negative resists are shown in Figure 2.6.

The exposed regions of the patterned layer are then etched and the remaining photoresist stripped. Roughness arises during this process mainly due to the polymer nature of the resist molecules. The resist molecules can form polymer aggregates due to intermolecular forces and entanglement of polymer chains [87, 88]. Roughness arises because the larger aggregates take longer to dissolve in the developer than the smaller molecules. *Yamaguchi et al.* [87] also demonstrated that the magnitude of the roughness is closely related to the polymer granule size and furthermore, the granule size depends on the molecular weight of the resist. For the particular resist studied by *Yamaguchi et al.*, ZEP520 (Nippon Zeon Co.), the reported granule diameter was ∼20-30 nm, with corresponding line width fluctuations of ∼2-3 nm. Additional roughness can also come from insufficiency of the lithographic system, however this contributes more to low frequency roughness [83] and is not a major contributor to high frequency roughness that occurs on the spatial scale of individual devices, which comes primarily from roughness due to polymer granules [89].

In addition, the patterned gate is used as a mask for the source/drain

implantation, which are thus self-aligned with respect to the gate edge, however the metallurgical PN junctions will not follow the gate LER exactly [90] due to smearing during thermal annealing and when the implantation angle is shallow. LER was not an issue in the past when transistor spatial dimensions were orders of magnitude larger than the roughness, however in deep sub-micron devices, the magnitude of LER is comparable to the channel length, resulting in significant variations in the channel length, which in turn degrades transistor performance. Indeed, in order to be able to reliably fabricate devices at the end of the roadmap, the ITRS indicates that sub-1 nm control of LER will be necessary.

The degradation in performance due to LER is due to the deviation of the channel length from its ideal uniform value. As a result, in particular regions, the channel will be shorter than average and in other regions longer than average. In regions where the local length is longer than average, the device is more difficult to turn on and, while this results in a reduction in leakage, it also reduces the drive current. Conversely, when the local length is shorter than average, the device turns on earlier, increasing leakage and degrading the noise margins. Shorter channel lengths are also problematic because shortening the channel has a stronger effect on leakage than lengthening it, due to the exponential dependence of the current. The short channel effects of the device are degraded due to lowering of the potential barrier in regions where the local length is short. At high drain this effect will be enhanced due to the further penetration of the drain electric field into the channel, leading to increased drain-induced barrier lowering (DIBL).

It is clear that LER has an important impact in deep sub-micron transistors that must be properly accounted for and that improvements in the photolithography process will be necessary. Current resists must be improved as it is apparent that with channel lengths less than 10 nm expected at the end of the roadmap, the presence of polymer granules 2-3 times the channel length in resist chemicals will have seriously detrimental effects on the reliability of the fabrication process. Additionally, although RDD induced effects can be reduced with undoped channels, LER continues to be an issue in such alternative architectures, with the problem being particularly acute in FinFETs [91]. It

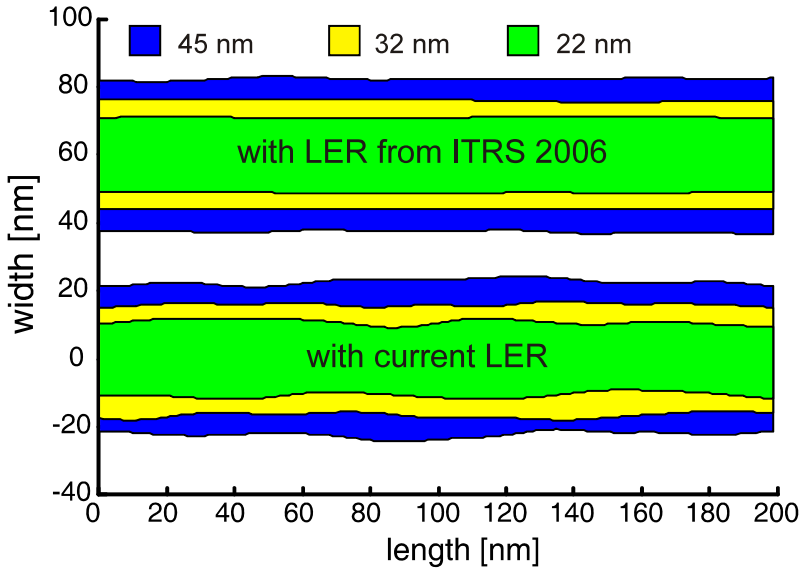


Figure 2.7: ITRS predictions and extrapolation of current LER for interconnects. After [4].

was however recently demonstrated by *Fukutome et al.* [92] that LER-induced V_T variations in bulk MOSFETs could be reduced through extension implantations parallel to the gate width direction of the device, a technique that may be useful as LER induced variability increases.

It should also be noted that LER is a phenomenon that affects interconnects as well as transistors. LER in interconnects results in resistance and capacitance variability, a factor that will significantly affect signal propagation between components [4]. The impact of LER on interconnects is shown in Figure 2.7. As stated already, LER scaling lags the ITRS and the roadmap forecast for LER may be overly optimistic. Figure 2.7 demonstrates interconnects for the 45-, 32- and 22 nm technology nodes, with roughness as forecast by the roadmap and unscaled roughness.

2.2 Simulation Techniques

Different techniques can be employed in the simulation of sub-0.1 μm MOSFET devices [5]. As shown in Figure 2.8, these range in complexity from

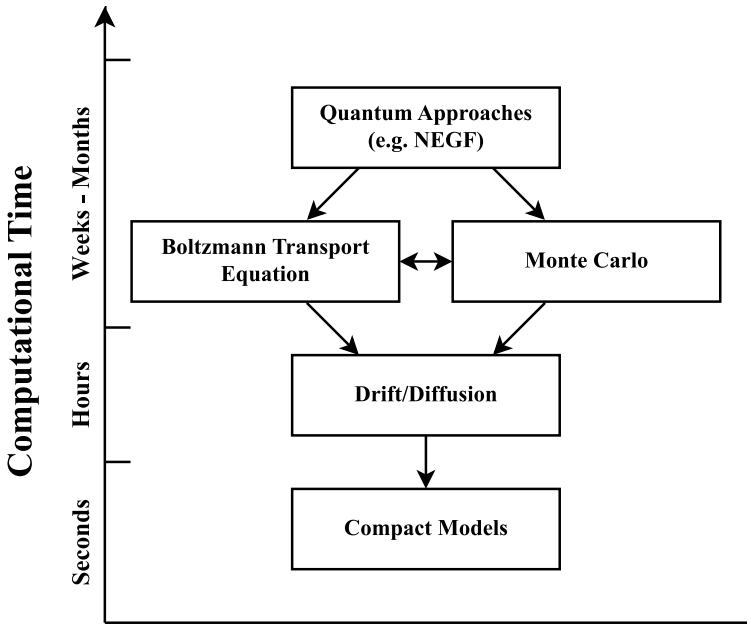


Figure 2.8: Hierarchy of computational techniques used to study MOSFETs. After [5].

quantum mechanical approaches, which are extremely computationally demanding, to compact models that attempt to model MOSFET characteristics semi-analytically and are computationally simple enough that many thousands of transistors can easily be included in a circuit-level simulation. Between these extremes lie techniques such as drift/diffusion simulations and Monte Carlo approaches, which attempt to capture important physical phenomena while introducing simplifying assumptions that reduce the complexity of the simulation. Since intrinsic parameter fluctuations arise from the granular nature of charge and matter, it is essential that a simulator accurately describes the physical system and is able to resolve individual dopant atoms and material non-uniformities. Intrinsic parameter fluctuations are by nature 3-dimensional, therefore, regardless of the modelling formalism employed, it is necessary to carry out 3D simulations in order to fully capture their effects on device characteristics.

It is also important to consider the statistical nature of intrinsic parameter fluctuations. It is of course impossible to exhaustively characterise a statistical

population for all but the most trivial of problems, and statistical variability is no exception. Furthermore, due to the extremely high level of integration on modern microprocessor chips, problematic statistically rare devices become, relatively speaking, a much more common occurrence due to the large numbers of devices involved. It is therefore necessary to simulate a sufficiently large statistical sample of devices so that the parameters characterising the statistical distributions can be extracted with an acceptable level of accuracy. With this in mind, it is necessary to employ modelling approaches that are efficient enough to allow large statistical ensembles to be simulated within a realistic time frame.

Despite their efficiency, compact models are only useful at higher levels of abstraction such as in circuit design, where transistors are, in essence, treated as black-box components. They do not capture the device physics accurately and rely on semi-empirical fitting in order to reproduce transistor characteristics. In addition, they lack predictive power at the device level. The excessive simplification of the physical processes involved and of the device itself result in an unacceptable loss of physical detail and they are not generally useful for characterising statistical variability that arises from the atomic structure of the device. For this reason, we do not consider compact models further in this work.

The information that is propagated from physical simulations to the next level in the design tool chain should also be considered. Although a complete description of the I-V characteristics of a device will always be more accurate, threshold voltage and leakage current are important and useful parameters for circuit design that can be obtained with significantly less computational effort than a full characterisation. Due to its importance in the operation of CMOS circuits, we primarily focus on threshold voltage variability in this work.

In this section, we review some of the common techniques used for simulation of MOSFET devices with respect to the requirements outlined above.

2.2.1 Drift/Diffusion

The drift/diffusion (DD) approach, which models the lowest-order transport system that can be obtained from the Boltzmann transport equation (BTE) [5], has been the backbone of numerical device simulation for over forty years. This began in the 1960s with the work of Scharfetter and Gummel [93], who proposed a robust discretisation of the DD equations and a self-consistent iterative method for solving them. Since then, a wealth of practical knowledge on the application of the DD equations has been accumulated [25, 94]. Despite the remarkable scaling of semiconductor devices since the 60s, many refinements and improvements to the DD model have been proposed (for example, more accurate mobility models [95, 96] and Density Gradient quantum corrections - see Section 2.2.1.1) and the DD approach remains a useful tool even for the simulation of deep sub-micron devices.

Here we consider the steady-state simulation of an n-channel MOSFET, which is a unipolar device. Apart from breakdown simulations, only the transport of electrons in the channel determines the device operation. In the DD approach, the steady-state current continuity equation (Equation 2.1) is solved self-consistently with Poisson's equation (Equation 2.2), which provides coupling of the current and charge distribution to the potential distribution and the corresponding field distribution [97]. Solving Poisson's equation yields the electrostatic potential, which can then be included in the current density equations.

$$\nabla \cdot J_n = 0 \tag{2.1}$$

$$\nabla \cdot (\epsilon \nabla \psi) = q(n - p + N_A^- - N_D^+) \tag{2.2}$$

In Equation 2.1, J_n is the current density. In Equation 2.2, ϵ is the permittivity, ψ is the electrostatic potential, n and p are the electron and hole distributions and N_A^- and N_D^+ are the ionised acceptor and donor distributions. It should be noted that different simulation techniques employ different techniques to update the mobile charge distribution. The current density is

expressed as the superposition of two components – the drift component (Equation 2.3), which is related to the electric field (E) and the diffusion component (Equation 2.4), which is related to the gradient of the electron density (n).

$$J_{n,drift} = qn\mu_n E = -qn\mu_n \nabla\psi \tag{2.3}$$

$$J_{n,diff} = qD_n \nabla n \tag{2.4}$$

where μ is the mobility, D is the diffusion coefficient and the other symbols have the same meaning as above. In the Boltzmann approximation, the mobility and diffusion coefficient are related via the Einstein relation:

$$D_n = \frac{k_B T}{q}\mu_n \tag{2.5}$$

where k_B is Boltzmann's constant and T is the temperature.

Significant simplifying assumptions are made in the derivation of the DD model from the BTE in order to obtain a closed system of equations and these limit the validity of the DD model [98]. As such, the DD model does not incorporate non-local effects and assumes that carriers are in thermal equilibrium with the lattice. This is not valid at high electric fields, however, where the carrier energy increases above the lattice energy and the effective carrier temperature is higher than the lattice temperature. The validity of the DD model can, however, be improved through empirical extensions such as better mobility models and quantum corrections. These improve the validity of the DD model at higher electric fields and take quantum confinement into account. However the validity of the DD approach is still limited to operating in regimes where the electric field varies slowly. DD assumes that carriers are able to instantaneously respond to changes in the electric field, however carriers are not massless and require a finite time and distance to equilibriate with the field and as a consequence, DD is unable to capture velocity overshoot effects [5].

Despite the shortcomings of the DD model, it is extremely useful due to its robustness, easy extension into 2 and 3 dimensions and efficiency. In simulations of sub-100 nm devices, it is still perfectly accurate in the sub-threshold

regime of operation, as the coupling between the current and Poisson's equation is weak and electrostatics mainly dominate the behaviour of the device. The application of DD methods is therefore ideal for the investigation of statistical threshold voltage and leakage current variability. Since the simulation of many thousands of devices is necessary to fully characterise the statistical distributions of MOSFET parameters, speed is of the utmost importance and the computational efficiency of the drift/diffusion model is an important factor.

2.2.1.1 Density Gradient

The drift/diffusion approximation can be extended to include quantum corrections, which improve the validity of the approach in aggresively scaled devices. The Glasgow DD simulator incorporates density gradient (DG) corrections, which were originally proposed in 1987 [99] by Ancona and Tiersten as a macroscopic description of some of the quantum mechanical behaviour of strong inversion layers at the Si/SiO_2 gate interface of transistors. DG theory allows some non-local quantum effects to be taken into account, thereby improving the accuracy of the DD approximation for highly scaled devices. Quantum effects are incorporated by adding a dependence on the gradient of the density to the equations of state for the electron gas. This in turn leads to an extra term in the expression for the current.

$$J_n = qD_n\nabla n - q\mu_n n\nabla\psi + 2qn\mu_n\nabla\left(b_n\frac{\nabla^2\sqrt{n}}{\sqrt{n}}\right) \tag{2.6}$$

This can be seen as a generalized drift/diffusion current equation with an additional driving force, due to an effective potential related to the gradient of the electron gas density. Since it arises from the gradient of the electron distribution, the additional term can be thought of as a "quantum diffusion" current, in the same way that $qD_n\nabla n$ is seen as a classical diffusion current. In Equation 2.6, J_n is the current density, q is the electronic charge, n is the electron density, μ is the mobility, ψ is the electrostatic potential and b is a term that expresses the magnitude of the density gradient dependence and has the general form $b = \frac{\hbar^2}{4m^*qr}$ [100]. In the relation for b, $\hbar$ is the reduced Planck constant, m^* is the effective mass and r is a parameter that depends on the

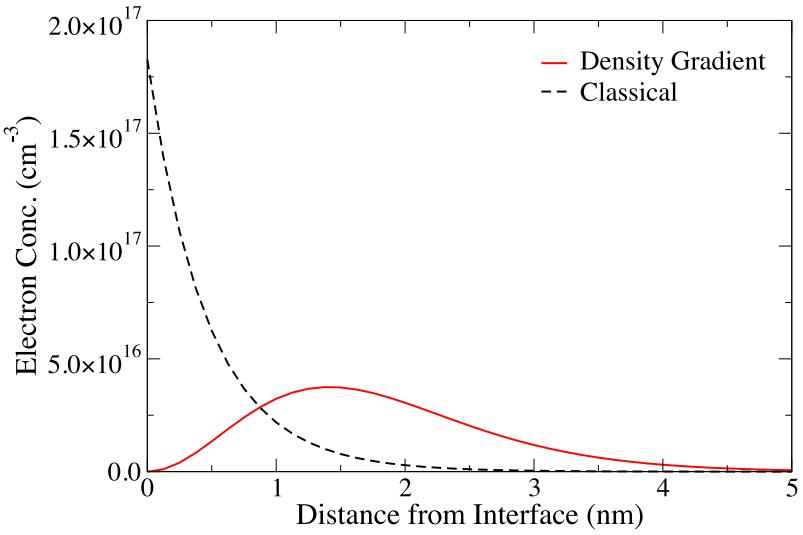


Figure 2.9: Comparison of the electron concentration obtained from classical and density gradient simulations.

temperature and the number of filled sub-bands [101]. For "high" temperatures (above ~ 77 K in Silicon), r approaches 3 and b becomes $\frac{\hbar^2}{12qm_n^*}$ [102].

By expressing the current density in terms of a quasi-Fermi potential, an expression for the quasi-Fermi potential can be obtained [103].

$$\phi_n = \psi - \frac{k_B T}{q} \ln(\frac{n}{n_i}) + 2b_n \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} \tag{2.7}$$

The inclusion of the additional driving term has the effect of pushing carriers away from the Si/SiO_2 interface, which is consistent with the 1-D solution of the Poisson-Schrödinger equation [104]. This can be seen in Figure 2.9, which shows the electron concentration obtained from classical and density gradient simulations. By deriving the density gradient theory from microscopic principles, rather than macroscopic, it has been shown [105] that the macroscopic theory represents the lowest order quantum effects arising from the microscopic physics and thus has a good range of applicability. Density gradient theory is therefore able to capture some aspects of quantum mechanics, such as quantum confinement and, to some extent, quantum tunnelling [106].

The shift in the peak electron concentration away from the interface occurs as a result of quantum confinement. This increases the equivalent oxide thick-

ness and leads to an increase in the threshold voltage. Quantum tunnelling is included only to a limited extent, as DG theory cannot account for electron coherence effects. However it was shown [106] that reducing the channel length resulted in degradation of the sub-threshold slope in DG simulations, a result that is consistent with the qualitative inclusion of tunnelling.

In addition, other simulation results by *Xiong et al.* [107] have also shown good agreement between density gradient theory and full Poisson-Schrödinger, with less than 10 mV difference in the threshold voltage obtained from both methods for a 20 nm double gate device.

Although not the original purpose of density gradient theory, the inclusion of density gradient corrections is also useful for correctly capturing the effect of random dopants on the potential distribution of the device. In classical drift/diffusion, discrete impurities cause sharp wells in the potential, resulting in charge trapping around the impurity. When density gradient is included, the same force that pushes the peak carrier concentration away from the Si/SiO_2 interface also has the same effect around discrete impurities and aids the correct resolution of random dopants. This will be discussed in more detail in Section 3.1.1.

The comparison of density gradient with more advanced simulation techniques demonstrates that DG is able to produce accurate results even for very small devices, meaning that the lower computational cost of DD after the inclusion of DG corrections can be exploited for large-scale statistical simulations for devices in the sub-0.1 μm regime.

2.2.2 Monte Carlo

Monte Carlo simulation techniques [108, 109] approach the problem of simulating carrier transport in a fundamentally different way to drift/diffusion methods. While DD methods represent a low order approximation of the Boltzmann transport equation, Monte Carlo methods are an indirect way of solving the BTE. The BTE is a difficult equation to solve and obtaining a direct solution is impossible in all but the most trivial of cases and Monte Carlo methods have gained popularity due to the fact that they avoid the attendant

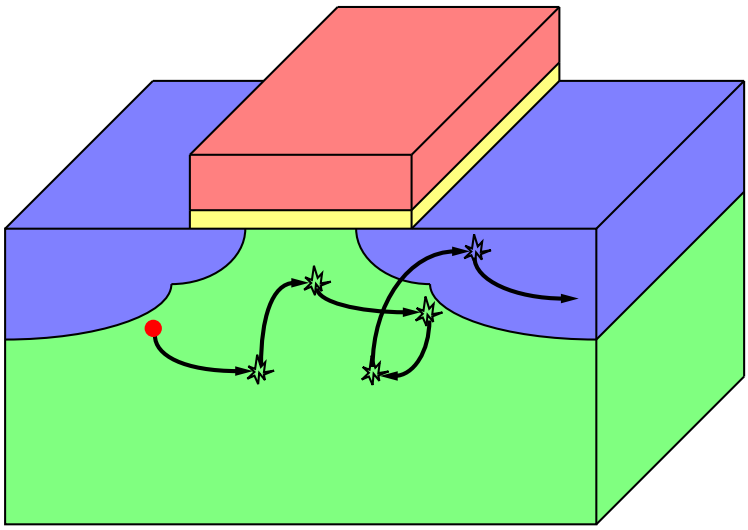


Figure 2.10: Illustration of the free flight and scattering of carriers in a Monte Carlo simulation.

problems of obtaining a direct solution. They instead seek to model the microscopic transport phenomena involved in semiconductor device operation, i.e. the free flight and scattering of carriers. It should be noted that Monte Carlo is a general statistical numerical method and is applicable to other problems in addition to semiconductor carrier transport [109].

Monte Carlo simulates the movement of carriers inside a semiconductor device, which can be thought of as a series of free flights and randomly selected scattering events at the end of each free flight. The movement of a single particle through the simulation domain is illustrated in Figure 2.10. Although the trajectory of particles during free flight is calculated using classical Newtonian physics, the scattering events incorporate quantum mechanics and, as such, Monte Carlo is a semi-classical model of the transport of particles. Scattering events can include interactions between carriers and phonons, fixed impurities and other carriers. The flight times and scattering events are determined by random numbers and by choosing appropriate probability distributions for the random numbers will correspond to the correct physics of the processes. Since large quantities of random numbers are required for Monte Carlo simulations, it is clear that a fast, high quality source of random numbers is important.

The typical flow of a Monte Carlo simulation is as follows. The initial state of the device is obtained from an analytical model or the output of another simpler simulation technique, such as drift/diffusion. Carriers are then propagated, scattering events selected and statistics on the particles gathered. The simulation can be carried out as a 'frozen field' simulation, where the field distribution inside the device does not change, or self-consistently, where Poisson's equation is used to update the driving field after each time step. This process is then repeated until the quantities of interest are estimated with sufficient precision.

One of the major advantages of Monte Carlo over drift/diffusion is its ability to correctly capture non-equilibrium carrier transport at high electric fields. The small spatial scales involved in deep sub-micron transistors mean that the electric fields are very high when the device is in the on state and the operating point is far from equilibrium. The carriers are not in thermal equilibrium with the lattice and non-local, non-equilibrium carrier transport becomes important. Since the actual microscopic physics of the carriers is taken into account, effects such as velocity overshoot are properly considered. This incorporation of accurate carrier transport mean that Monte Carlo simulations may, to some extent, go beyond the limits of the BTE and yield more realistic results, as they naturally include noise and fluctuations that would occur in real devices [5].

Despite the better physical accuracy of the Monte Carlo approach, one of the major drawbacks is the large computational expense, as it is necessary to populate the system with a sufficient number of carriers, such that the ensemble averages are suitably representative of the corresponding averages for the entire carrier gas [94]. Due to the small number of particles in the simulation ($\sim$10,000–50,000), there will still be noise in the statistical data however, and it may be important to apply variance reduction techniques in order to improve the accuracy of estimated averages. In order to accumulate good statistics, it is also necessary for the simulation time to be sufficiently long. In general, the requirement for long simulation times (on the order of days/weeks, rather than hours) means that Monte Carlo methods are not well suited for large-scale statistical simulation. In terms of statistical variability, they are more naturally suited to studying on-current variability than threshold

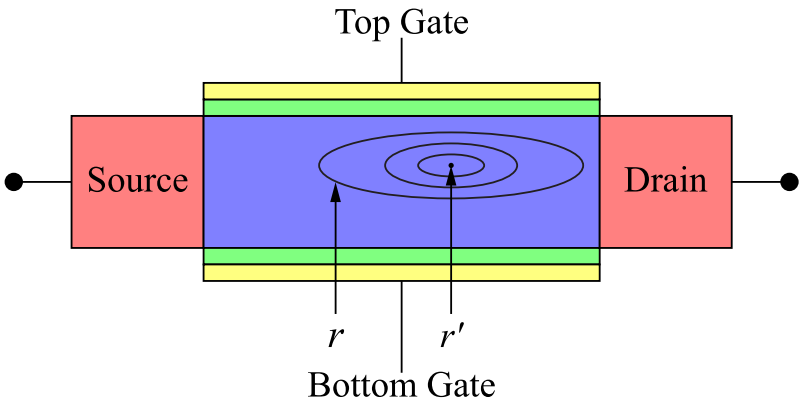


Figure 2.11: Illustration of the Green's function, which represents the resulting wavefunction at $\mathbf{r}$ in response to an excitation at $\mathbf{r}'$. After [6].

voltage variability, since in the sub-threshold regime the number of free carriers and scattering events is much lower, meaning that it takes much longer to accumulate reliable statistics [98] and increasing the computation time taken to obtain accurate estimates of the parameter of interest.

2.2.3 Non-Equilibrium Green's Functions

Despite the improved physical accuracy of Monte Carlo approaches, semi-classical approaches begin to lose validity at the extremes of scaling [110] and below channel lengths of ∼10 nm quantum effects begin to dominate. This requires a fundamentally quantum description of the system. This is distinct from the incorporation of quantum effects into semi-classical Monte Carlo simulations to describe particular quantum aspects of the system. Quantum transport approaches are effectively at the top of the hierarchy of simulation complexity and are extremely computationally expensive.

In the non-equilibrium Green's function (NEGF) approach, the device under study is represented by an appropriate Hamiltonian, for example, the effective mass Hamiltonian. The Green's function describes the response of the system at any arbitrary point $\mathbf{r}$ to an excitation at the point $\mathbf{r}'$ [6], as illustrated in Figure 2.11.

The electron and current densities can be obtained from the lesser Green's

function $G^<$ using the following equations.

$$n(E,x) = iG^<(E,x,x') \tag{2.8}$$

$$J(E,x) = -i\frac{e\hbar}{2m}(\nabla - \nabla')G^<(E,x,x')\Big|_{x-x'} \tag{2.9}$$

Where E is the energy, e is the electronic charge, $\hbar$ is the reduced Planck constant, m is the effective mass and ∇ and ∇' operate on x and x', respectively. More details on the calculation of $G^<$ and the related functions can be found in [6].

To solve Green's functions for the quantities of interest, it is necessary to invert very large matrices, making the simulation very memory intensive. For example, for a mesh containing n points, the corresponding Hamiltonian will contain $n \times n$ elements. This complicates the algorithm and makes it difficult to extend it into 2- and 3D as the matrices quickly become extremely large. As a result, the NEGF approach has primarily been used to study lower dimensional systems, although full scale 3D simulation studies have been carried out recently [111].

The typical flow of a NEGF simulation would be as follows. An initial estimate of the potential and electron concentration are obtained from, for example, a drift/diffusion simulation. The NEGF approach is used to determine the new electron density based on the initial potential. A quasi-Fermi potential is calculated from the old potential and new electron density, which is then used to solve Poisson's equation to obtain the new potential distribution. This process is repeated iteratively until self-consistent convergence of the potential, electron density and current are achieved.

Due to the immense computational complexity associated with the NEGF approach, it is infeasible to use this for large-scale statistical simulation of realistic devices. In addition, the devices of interest in this work are larger than those typically examined using the NEGF approach and a semi-classical approach is entirely sufficient. The NEGF approach does however have applicability in calibrating quantum corrections that are applied to traditional DD

and Monte Carlo simulations.

2.3 Summary

In this chapter, an overview was given of some of the key sources of statistical variability. The two major sources in contemporary bulk MOSFETs – random discrete dopants and line edge roughness – were examined in detail. Some of the simulation techniques commonly used to study intrinsic parameter fluctuations were also discussed. Drift/diffusion, Monte Carlo and Non-equilibrium Green's functions approaches were outlined, along with some of their advantages and disadvantages. In addition, density gradient quantum corrections, which are used to improve the accuracy and applicability of the drift/diffusion approach were summarised.

Computational efficiency was the most important consideration for the large-scale simulations in this project, thus while Monte Carlo and NEGF approaches result in better physical accuracy, they were not appropriate for this work. In the next chapter, the Glasgow 3D drift/diffusion simulator is described in more detail, along with how the sources of variability of interest for this study are incorporated. The 35 nm MOSFET that is the primary testbed in this research is described and details of the Grid technology employed in this study are given.

Chapter 3

Simulation Methodology

In this Chapter, we describe the Glasgow 3D "atomistic" simulator and how the sources of variability studied in this thesis are implemented in the simulator. There is considerable technical difficulty associated with the large-scale simulations that were carried out as part of this work. An overview of the Grid technology that facilitated these simulations is given. The transistor studied is described in detail, including information about the corresponding TCAD process and device simulation and the calibration of the simulations to the measured device characteristics. Finally, we outline specific aspects of the simulations, such as the choice of bias conditions and the issues of numerical accuracy and convergence.

3.1 The 'Atomistic' Simulator

The Glasgow "atomistic" simulator has been developed within the Device Modelling Group for a number of years and an overview of the operation of the simulator is given in this section. It should be noted that although no fundamental changes to the simulator were necessary for this work, some modifications were made in order to make the use of the simulator in a Grid environment easier (see Section 3.2).

The simulator self-consistently solves the nonlinear Poisson (Equation 3.1) and current continuity equations (Equation 3.2) in the drift/diffusion (DD)

approximation. The DD equations are discretised and solved in a 3D simulation domain. Since intrinsic parameter fluctuations occur due to physical phenomena that are inherently 3 dimensional in nature, this renders 2D and quasi-3D simulations unsuitable to fully capture their influence on the device. This is due, for example, to the fact that 2D and quasi-3D simulations will not properly capture the current flow around discrete impurities and the resulting percolation path between the source and drain, which allows early turn-on of the device [112].

The steady-state system of semiconductor equations suitable for the simulation of MOSFETs includes Poisson's equation (3.1), the current continuity equation (3.2) for the majority carriers in the transistor (electrons in this case) and the corresponding equations for the current in the drift/diffusion approximation (3.3).

$$\nabla \cdot (\epsilon \nabla \psi) = q(n - p + N_A^- - N_D^+) \tag{3.1}$$

$$\nabla \cdot J_n = 0 \tag{3.2}$$

$$J_n = qD_n \nabla n - q\mu_n n \nabla \psi \tag{3.3}$$

In Equation 3.1, ϵ is the permittivity,ψ is the potential, q is the electron charge, n is the electron concentration, p is the hole concentration, N_A^- is the ionised acceptor concentration and N_D^+ is the ionised donor concentration. In Equation 3.2, J_n is the current density, which can be obtained from Equation 3.3. In Equation 3.3, J_n is the electron current density, D_n is the electron diffusion coefficient and μ_n is the electron mobility.

These equations are discretised onto a non-uniform 3D Cartesian mesh using a finite difference scheme that was first proposed by Scharfetter and Gummel [113]. The Scharfetter-Gummel scheme assumes that the carrier distribution is exponential between mesh points and is stable regardless of the mesh size. In MOSFET simulations, a fine mesh is required to accurately capture the changes in the current density distribution in the inversion layer.

The mesh spacing h is usually in the range 1–0.5 nm, although finer steps are frequently used in the active regions of the device.

The use of fine meshing is problematic, however, as the inclusion of random dopants in the simulation leads to a dependence of the solution on the mesh step size, as described in [114] and in Section 3.1.1 below. This is due to artificial trapping of mobile charge in the potential wells associated with discrete dopants. This erroneous charge trapping is significantly reduced by the inclusion of Density Gradient (DG) quantum corrections [115, 103] in the simulator. Density Gradient corrections were originally developed in order to properly capture the behaviour of the inversion layer charge [99], which is essential for the correct modelling of IV curves in ultra small devices. The quantum corrections take quantum confinement into account, which raises the electron ground state, and thus also aid in the correct resolution of discrete dopants. The Glasgow simulator includes Density Gradient corrections for both electrons and holes, which enables the effect of discrete dopants on device characteristics to be correctly determined. The density gradient equation, which is solved self-consistently with Poisson's equation, is given by Equation 3.4.

$$2b_n \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} = \psi - \phi_n + \frac{k_B T}{q} \ln(\frac{n}{n_i}) \tag{3.4}$$

In Equation 3.4 b_n is defined as $\frac{\hbar^2}{12qm_n^*}$, where $\hbar$ is the reduced Planck constant, q is the electron charge and m_n^* is the effective mass of an electron. n is the electron distribution, ψ is the electrostatic potential, ϕ_n is the quasi-Fermi potential, n_i is the intrinsic carrier concentration and the other symbols have their conventional meanings.

Equations 3.1–3.3 are a system of coupled equations that is solved self-consistently using Gummel iterations [93]. The Gummel iteration is modified to include the Density Gradient equations and as implemented in the Glasgow simulator, the Gummel iteration consists of the following process, which is illustrated in Figure 3.1. First, the Density Gradient equations are solved self-consistently with Poisson's equation. This pair of equations is then solved self-consistently with the current continuity equations. Direct inclusion of

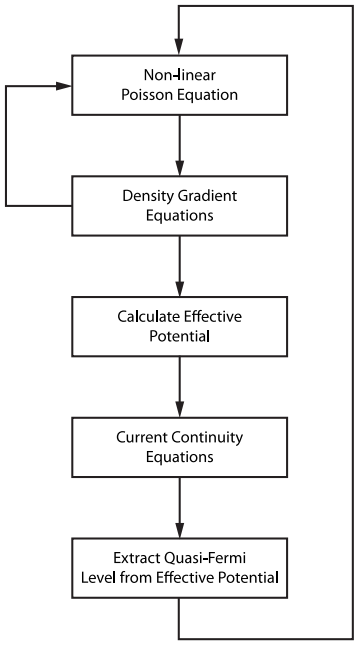


Figure 3.1: Flowchart illustrating how the Density Gradient equations are incorporated into the Gummel iteration. After [2].

the DG equations into the Gummel cycle was found to negatively impact the convergence properties of the algorithm, however solving the DG equations self-consistently with Poisson's equation improves stability and convergence [2].

In the simulator, Poisson's Equation and the density gradient equations are solved using a Successive Over-Relaxation (SOR) solver and the current continuity equations are solved using a BiCGSTAB solver. The more complex BiCGSTAB solver is required for the current continuity equations due to the non-diagonally dominant system of equations that arises from the Gummel discretisation scheme and the inclusion of sophisticated mobility models, which extend the validity of the drift/diffusion formalism.

As stated previously, the drift/diffusion formalism does not capture non-equilibrium transport effects at high electric fields and thus does not accurately predict the current in decananometer transistors. It also cannot capture current variations associated with Coulomb scattering from random discrete dopants. This does not, however, diminish the usefulness of the drift/diffusion approach for predicting threshold voltage variability, since it models the sub-

threshold regime well, where the behaviour of the device is primarily determined by electrostatics. The applicability of DD for the simulation of ultra small devices is also improved by the incorporation of quantum corrections.

DD is also favourable from the point of view of its computational efficiency. To demonstrate this, the Glasgow simulator was benchmarked on a 2.66 GHz Intel Xeon (Harpertown) machine. Although the simulator can use multiple cores in a machine, it was restricted to a single core for the purposes of the benchmark, as this is reflective of the cluster/grid environment where the large-scale simulation work was carried out. Simulations to find the threshold voltage were carried out for the 35 nm device (described in Section 3.3) with continuous, uniform doping and no sources of variability included. The mesh size for the device is $109 \times 36 \times 64$ nodes (251,136 nodes in total). For a classical simulation, without density gradient corrections, the overall simulation time was ∼2.5 minutes, with 55-60% of the time per (gate) voltage point spent solving Poisson's equation and 35-40% spent solving current continuity. With density gradient corrections included, the total simulation time was $\sim$ 32.5 minutes, with over 95% of the time per voltage point taken solving Poisson's and the density gradient equations and, correspondingly, less than 5% on solving current continuity. Although the inclusion of density gradient drastically increases the total simulation time, it is necessary in order to obtain accurate results for such small devices.

3.1.1 Random Dopants

The DD approach is a classical continuum simulation approach and the representation of random discrete dopants presents some difficulties. Special treatment is needed to consistently represent individual discrete dopants in the simulation domain. The statistical generation of the dopant distribution in individual devices also needs special care.

One of the major problems associated with including random dopants in DD simulations is artificial charge trapping on top of impurity atoms [72]. In classical drift/diffusion simulations, this occurs due to the use of Boltzmann or Fermi-Dirac statistics, which results in the electron concentration closely

following the electrostatic potential. Discrete dopants create deep Coulomb potential wells for majority carriers and in classical simulations, where the carrier concentration is locally related to the potential distribution, such wells are able to trap large numbers of carriers, resulting in artificial localisation of the majority carriers. This is unphysical however since, quantum mechanically, only certain discrete energy levels are allowed, and for individual dopants the ground state is near the conduction band edge [116, 117], meaning that carriers are less strongly localised near the dopant and can more easily escape the potential well.

In DD simulations, the dopant charge density associated with a particular mesh node is inversely proportional to the volume of the node, therefore, the requirement for a fine mesh in the simulation of ultra small devices leads to localisation of charge within the mesh cell. The result is a sharpening of the Coulomb potential well with the refinement of the mesh, which represents the singularities of the potential associated with a point charge more closely. Proposed solutions to this problem have included charge assignment schemes [118], splitting of the Coulomb potential into short and long range components [119] and the introduction of quantum corrections [120, 114, 121].

The inclusion of individual discrete dopants in the Glasgow DD simulator is achieved by using a charge assignment scheme to spread the single point charge of the dopant onto the surrounding mesh nodes. There are several charge assignment schemes that can be employed, such as nearest grid point (NGP), Gaussian smearing and Cloud-in-Cell (CIC) [122]. Since NGP leads to excessive charge trapping and Gaussian smearing relies on somewhat arbitrary choices of parameters [2], CIC is adopted in the simulator. The CIC approach spreads the single elemental charge q onto the eight mesh points neighbouring the dopant atom, as illustrated in Figure 3.2. The charge assigned to a particular mesh point is given by Equation 3.5:

$$\rho(x, y, z) = w_x w_y w_z \frac{1}{V} \tag{3.5}$$

where ρ is the charge, w_x, w_y and w_z are weighting factors that depend on the location of the dopant atom and V is the volume associated with the mesh

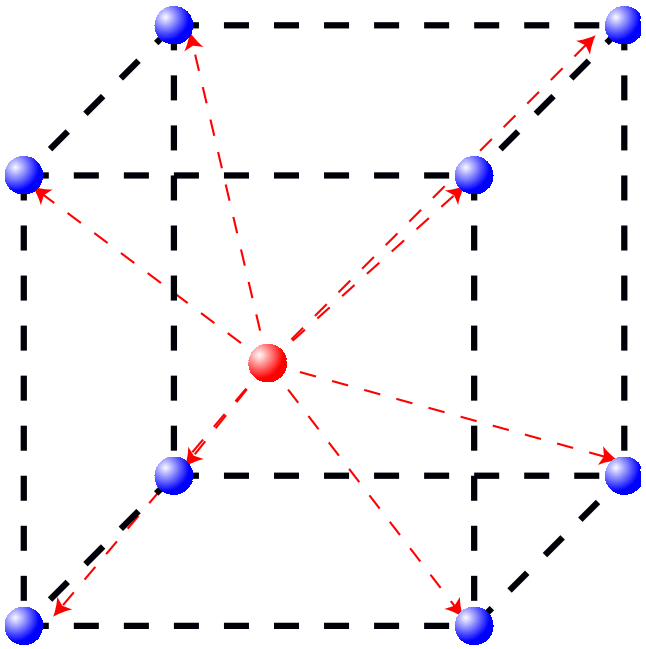


Figure 3.2: The Cloud-in-Cell charge assignment scheme. The charge from the dopant atom is split among the eight neighbouring mesh points. After [2].

node, given by $V = h_x h_y h_z$. For the Cloud-in-Cell approach, with a uniform mesh in x, y and z, the weighting factors are given by Equation 3.6:

$$w_x = \begin{cases} 1 - \frac{|x - x_p|}{h} & |x - x_p| \leq h \\ 0 & otherwise \end{cases} \tag{3.6}$$

where x is the dopant x position, x_p is the mesh point position and h is the mesh step size. The fraction of the dopant charge assigned to a particular mesh point therefore corresponds to the distance between the dopant and the mesh point.

While artificial charge trapping is reduced by employing the Cloud-in-Cell charge assignment scheme, there is still a dependence on the mesh step size. For dopants that are near mesh points, CIC also behaves more like NGP, as a larger proportion of the charge is assigned to the nearest mesh point. Although there are other more complex charge assignment schemes, for example Triangular Shaped Cloud (TSC), which spread the charge beyond the mesh cell containing

the dopant, these schemes become more computationally complex and are not necessarily desirable in simulations of MOSFET devices, for example, at the Si/SiO_2 interface.

Since charge assignment does not eliminate artificial charge trapping, it is necessary to adopt an additional approach to reduce charge localisation. Due to the somewhat arbitrary choice of cut-off point and the possibility of double counting in approaches that split the Coulomb potential into short and long range components, density gradient quantum corrections are used in the Glasgow simulator. The inclusion of density gradient corrections results in a significant decrease in the amount of charge that becomes localised around discrete impurities. This reflects the effect of quantum confinement in the potential wells, which reduces the sharp peaks associated with discrete impurities in classical DD simulations. By reducing the amount of charge that becomes trapped around impurities, there are more carriers that are free to contribute to the current flow, resulting in a decrease in the resistance associated with the region containing discrete dopants.

The inclusion of density gradient corrections also helps to alleviate the dependence of the charge on the mesh step size. The effective quantum potential is much less sensitive to the mesh step size than the corresponding classical potential, thus avoiding further increases in the amount of charge localisation around impurity atoms due to reductions in the mesh size.

As well as correct resolution of individual discrete dopants, it is also important to correctly determine where dopants should be placed in the devices in order to represent dopant distributions that arise from realistic doping profiles. Since modern CMOS devices have small dimensions and complex doping profiles and require non-uniform meshing, simple approaches based on randomly generated 3D co-ordinates or selecting dopant numbers from a Poisson distribution are insufficiently precise to accurately determine the discrete dopant distribution.

The Glasgow simulator employs a method first described by *Frank et al.* [7]. In this method, random numbers are generated for every Silicon lattice site to determine if the atom is a dopant or not. This is not an onerous task, since in deep sub-micron devices, the actual number of Si lattice sites is of a manage-

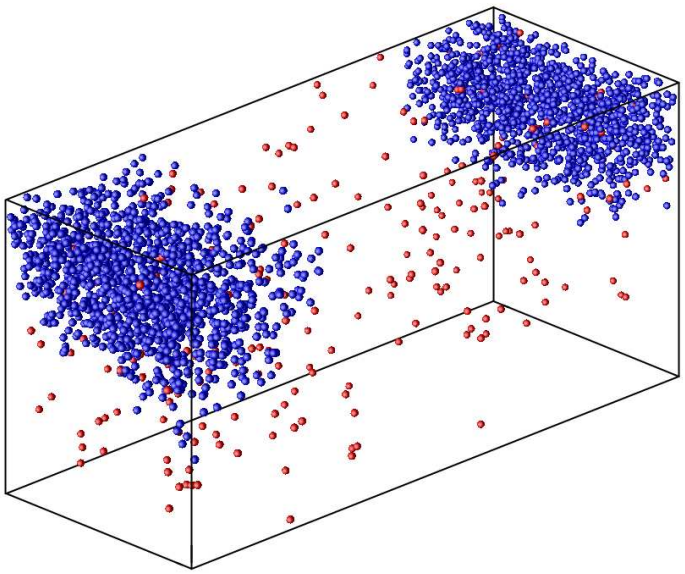


Figure 3.3: Sample random dopant distribution obtained using the method described in [7]. Discrete acceptors are indicated in red and discrete donors in blue.

able order of magnitude. For example, a small cubic volume with side 50 nm, which may represent the simulation domain of an ultra small device, contains only $\sim 6.25 \times 10^6$ Si atoms. The generation of $10^6 - 10^7$ random numbers is a straightforward computational task compared to the full 3D simulation of the device. Thus, for each lattice site, dopants are introduced using a rejection technique that selects whether a dopant should be placed at a particular lattice site based on a probability given by the ratio of the doping and Si concentration at that site. In regions of uniform doping, this is equivalent to a Bernoulli trial for each lattice site, therefore the distribution of the number of dopants that arises is a Binomial distribution. For a large number of trials, as is the case here, the Binomial distribution can be approximated well by a Poisson distribution, hence the common observation that the number of dopants closely follows a Poisson distribution. A sample random dopant distribution obtained using this method is shown in Figure 3.3.

It should also be noted that in real devices, there may be a possible cor-

relation in the discrete dopant position due to Coulomb interactions between the dopants during high temperature annealing. This is not included in the simulator and is currently ignored this work [59].

3.1.2 Line Edge Roughness

Line edge roughness is introduced into the simulator using a method based on 1D Fourier synthesis, as detailed in [3]. Previous methods used to model LER include 2D simulations coupled with the statistics of the channel length [123, 124, 125, 126] and 3D simulations that employed square wave approximations of the gate edges [127, 128]. More complex approaches to modelling LER also incorporate atomic scale process simulation [71] and the effects of strain related variations [129], however these are neglected in this study.

The Fourier synthesis method employed here generates lines from a Gaussian power spectrum. The corresponding auto-correlation function, which is characterised by two parameters – the RMS amplitude (Δ)[1] and the correlation length (Λ) – is fitted to that obtained from the analysis of SEM micrographs of extreme ultraviolet (EUV) [130] and electron beam lithography [131]. Values obtained for Λ indicate that the autocorrelation length lies in the range 20-30 nm, while 3Δ has remained in the range $\sim 3-5$ nm.

The Gaussian power spectrum used in the simulator is given by Equation 3.7:

$$S_G(k) = \sqrt{\pi}\Delta^2\Lambda \exp(-\frac{k^2\Lambda^2}{4}) \tag{3.7}$$

where $k = i(\frac{2\pi}{Ndx})$. N is the number of mesh points in Fourier space, with spacing dx and $0 \leq i \leq \frac{N}{2}$. Note that the power spectral density (S) and the auto-correlation (R) functions can be related using the Wiener-Khintchine theorem [132], which states that the power spectral density is the Fourier transform of the auto-correlation function, given by Equation 3.8.

[1]Note that the values commonly quoted as "LER magnitude" are usually defined as 3Δ.

$$S_x(\omega) = \int_{-\infty}^{\infty} R_x(\tau) \exp(-\jmath\omega\tau) d\tau \tag{3.8}$$

A set of N discrete elements is used for the Fourier space representation of the line $[L(k)]$, with amplitudes generated from the power spectrum and the phase of each element being selected randomly in order to generate randomly varying lines in real space (Equation 3.9).

$$L(k) = S_G(k) \angle rand(0, 2\pi) \tag{3.9}$$

For inputs which are real numbers, the discrete Fourier transform obeys the symmetry $X_k = X^*_{N-k}$, where the $*$ represents complex conjugation. Furthermore, the elements X_0 and $X_{N/2}$ (the Nyquist frequency) are purely real. Enforcing these conditions on $L(k)$ therefore ensures that the inverse DFT of the randomly constructed spectrum (i.e. the LER pattern, $l(x)$) is purely real. This also means that only $N/2$ elements of $L(k)$ are independent, with elements $N/2 + 1$ to N being determined by complex conjugation. The corresponding real space line is thus obtained by taking the inverse Fourier transform of $L(k)$, as in Equation 3.10. An example of a resulting random line is shown in Figure 3.4.

$$l(x) = \mathcal{F}^{-1}\{L(k)\} \tag{3.10}$$

Both gate edges are generated using this method, and an example doping profile with LER introduced in this way is shown in Figure 3.5. It should be noted that in our simulations, the metallurgical PN junctions are assumed to follow the gate edges. Although this assumption may not be valid in certain situations, for example when the implantation angle is very shallow [90], it is reasonable given that the correlation length is typically longer than the junction depth in deep sub-micron devices. It is also possible that thermal annealing during doping activation can cause smearing of the PN junctions.

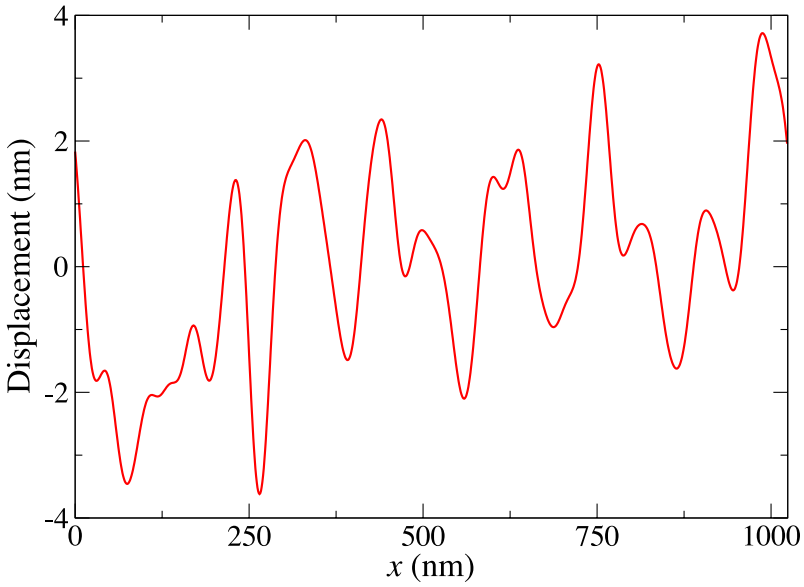


Figure 3.4: Example of a random line generated by the above algorithm.

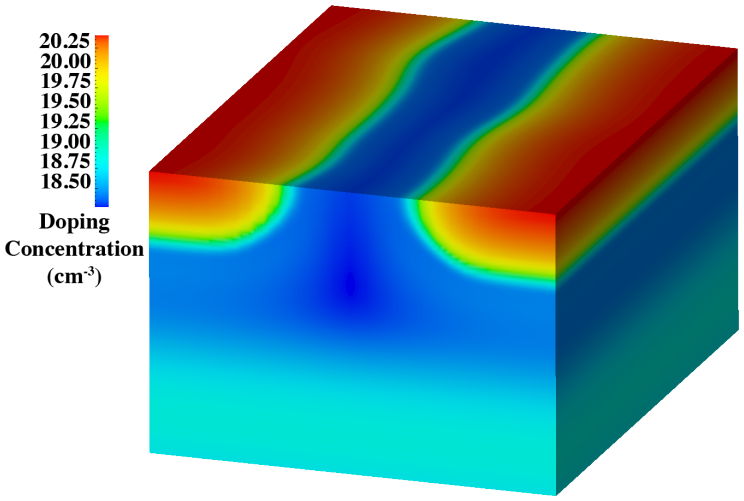


Figure 3.5: Doping profile in an example device with LER introduced.

3.2 Grid Technology

In order to facilitate computation on the scale required for this work, it was necessary to employ advanced grid technology. This is important due to both the large numbers of computational resources deployed and for storing, managing and analysing the corresponding output data produced by the simulator.

In total, over 500,000 CPU hours (approximately 60 CPU years) worth of computational time was required for the 3D simulations carried for this work. The computational resources were provided by ScotGrid [133] and by the Device Modelling Group's (DMG) own in-house cluster. These resources totalled approximately 1,000 cores at the time these simulations were carried out and consist of a mixture of high performance AMD Opteron and Intel Xeon systems.

Significant technical challenges are associated with simulation on this scale in terms of job submission, job tracking and data management, which had to be overcome. Since each simulated device is independent, the most straightforward method for large scale simulation is a simple task farming approach, which requires the ability to submit thousands of jobs in a single batch. On the ScotGrid system, job submission was performed using the Globus software toolkit [134]. The functionality of Globus on ScotGrid was however limited to single job submission and monitoring, rendering it unsuitable for large scale task farming. The Ganga frontend [135] has been developed at CERN to overcome this limitation and the maximum number of concurrent jobs is limited only by the system. In this case, the system limits the number of jobs to 1,000 per user, but since our simulator itself can create and simulate a statistical sample of devices within a single job execution the facilities provided by Ganga proved sufficient to perform the necessary amount of simulations. While this method of submission alleviates many of the issues associated with large parallel job submissions, Ganga is not without problems. Ganga automates the process of job submission and monitoring, but since it is a front-end to the existing system, some of the deficiencies of the grid middleware are still evident; for example job submission is extremely slow (Approximately 1-2 hours to submit 500 jobs). Additionally, it was also found that Ganga sometimes

fails to properly track jobs (due to both bugs in Ganga and resource broker problems, in our experience), resulting in it becoming impossible for the user to control the jobs, and requiring administrator intervention to cancel the execution of such rogue jobs. On the DMG's in-house cluster, job submission is achieved using Sun Grid Engine (SGE). SGE inherently supports large-scale parallel submission in the form of array jobs, making submission to this system more straightforward.

The issues encountered with Ganga and middleware stack on ScotGrid resulted, in some instances, in the complete loss of a set of 500 jobs. Although the majority of issues were related to job submission rather than execution, in some cases batches of jobs were lost partway through the job, resulting in wasted run time and incomplete output data. This complicates the resubmission procedure as devices that successfully completed before the job failed should not be rerun and must be excluded from the job when it is resubmitted. Other intermittent failures occurred during the simulations due to issues with both the software and hardware on the system and it was usual to have to resubmit up to 25% of the jobs for a batch. In addition, while the simulation work was on-going, Ganga was upgraded, which led to status monitoring issues whereby Ganga would successfully submit jobs, but could not obtain status information or control the jobs. The issues encountered with job submission also frequently resulted in partial submission of the batch, which again led to incomplete output data. It is estimated that of all the jobs submitted to ScotGrid during this work, approximately 30-40% had to be resubmitted at least once. It should also be noted that it is important to track failed and/or numerically unstable simulations in order to preserve the statistical integrity of the samples. While, for example, it is feasible and correct for duplicate atomistic profiles to occur due to the finite number of lattice sites in the system, it is important to ensure that duplicate random number generator seeds are not included in the ensemble.

Due to the issues encountered in the submission and tracking of the large numbers of jobs, a data management system based on a PostgreSQL database backend was developed to alleviate some of these problems. In addition, it was clear that managing the large numbers of output files produced was difficult

exp-id	device-no	vth	ioff	ndop	date-run
1	1	0.092104	4.90992e-08	2661	2007-11-19

location	dopants
node092.beowulf.cluster	[Encoded binary data]

Table 3.1: Example record of the output data from a simulation with random dopants.

and that simple file-based storage approaches were inadequate to properly archive and manage the data. The simulator was modified to store the output data directly in the database, which facilitated some semblance of live job status monitoring, the facility for which is not provided by Ganga/Globus. This also aided in co-ordinating the resubmission of jobs that failed due to issues with the grid software or hardware problems. The use of a database also allows the data to be stored in a self consistent manner along with useful metadata and derived data, and there is a total of approximately 41,000,000 rows of output and derived data stored from this work. Device results can, for example, be cross-referenced with records of the details of the experiment and simulator input files. In addition, this is useful for subsequent analysis, as complex data mining can be performed directly on the data via SQL queries and direct interfaces to tools such as 'R' [136] and 'Python' [137]. The structure of the database is demonstrated in Table 3.1, which shows an example record of the output for a simulation with random dopants.

3.3 Device Characteristics

In order to accurately characterise the effects of intrinsic parameter fluctuations, the device under investigation should be realistic and representative of a recent technology generation. Although now slightly dated, the device under study in this thesis is a 35 nm physical gate length n-Channel MOSFET, originally published by Toshiba [138], which is representative of devices from the 65 nm technology node. The device is widely used within the Glasgow Device Modelling Group in various variability simulation studies. This allows the

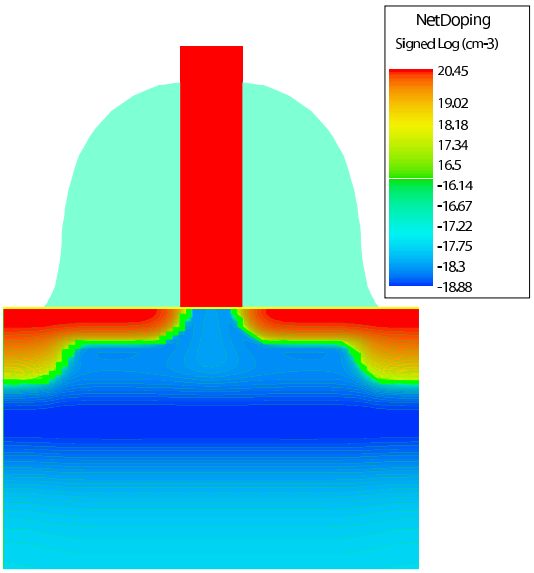


Figure 3.6: Doping profile of the 35 nm device under investigation in this thesis.

results obtained in this work to be directly compared against existing results for the same device and gives confidence that the obtained results are accurate. The device has a poly-silicon gate with Oxynitride gate dielectric and $EOT = 0.88\,nm$, and features retrograde Indium channel doping, which helps reduce threshold voltage variability due to random dopant effects and avoids degradation of carrier mobility. The device also includes Boron source/drain haloes to help control short channel effects (SCE) and prevent punch-through. Shallow source/drain extensions are also used in order to control SCE, with $x_j = 20\,nm$. The drive current reported by *Inaba, et al.* [138] for the device was $676\,\mu A/\mu m$ at $V_D = 0.85\,V$ and $I_{OFF} = 100\,nA/\mu m$. The 2D net doping profile of the device is shown in Figure 3.6, and the 1D profile of the channel doping and the source/drain extensions in Figure 3.7.

To ensure the accuracy of the simulations, the "atomistic" simulator is calibrated to match the characteristics obtained from TCAD simulation (using Taurus [139]) of a wide, continuously doped device. The TCAD simulations themselves have also been calibrated to real device measurements. The atomistic simulator cannot be directly calibrated to the device measurements as it

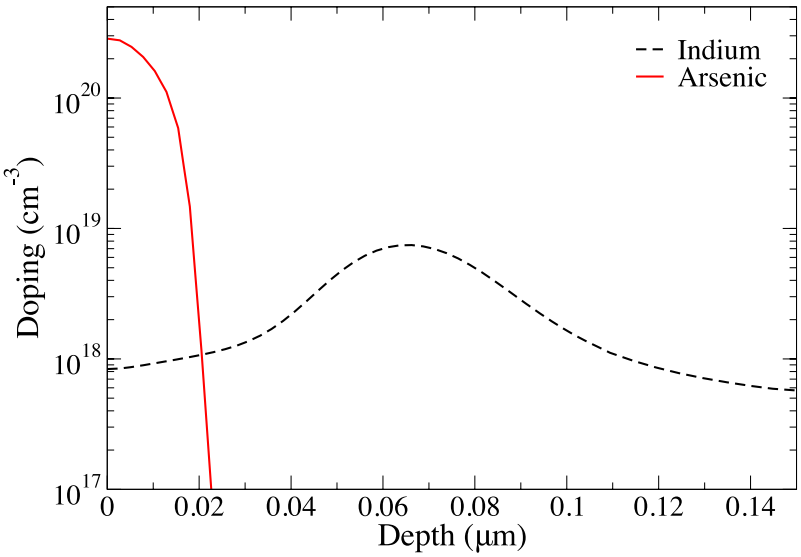


Figure 3.7: Vertical doping profile (Indium) and source/drain extension profile (Arsenic) of the 35 nm device.

does not include external resistances, which are associated with the metal contacts. Instead, the TCAD simulations, which do include external resistances, are calibrated to the experimental measurements. The external resistances are then removed from the TCAD simulations, allowing the calibration of the "atomistic" simulator against the TCAD results [2].

The calibration procedure primarily involves tuning the parameters associated with the mobility models used in the simulations. In this case the mobility model used is the Caughey-Thomas model [95], in which the mobility depends on the doping concentration and the electric field. Additional parameters, such as those associated with density gradient quantum corrections, can also be adjusted in order to accurately match the experimental data. Full details of the calibration for this device can be found in [2].

Figure 3.8 compares the simulated $I_D V_G$ characteristics with those obtained from measurement and from TCAD simulation of the 35 nm MOSFET using Taurus and shows that the simulator reproduces the device characteristics well.

In order to assess the potential impact of intrinsic parameter fluctuations on future scaled devices, the 35 nm MOSFET used here is scaled down to a physical gate length of 13 nm, which reflects the predicted device dimensions

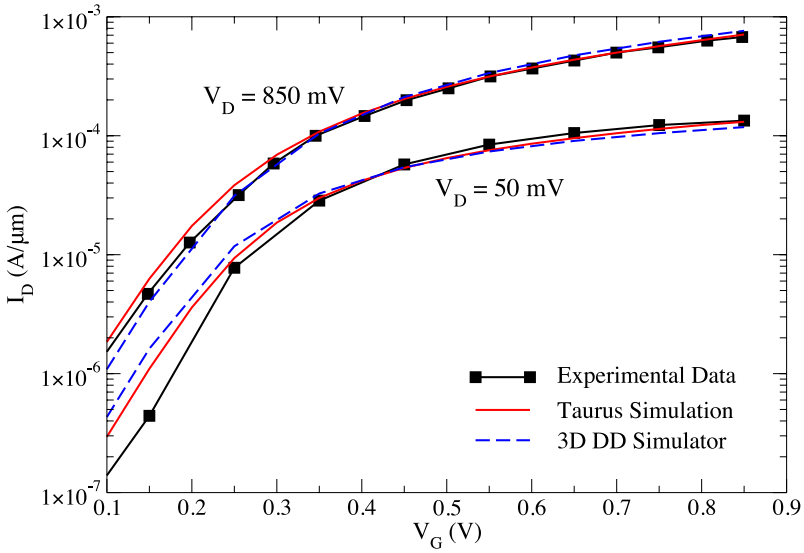


Figure 3.8: Comparison of the $I_D V_G$ characteristics of the 35 nm device obtained from drift/diffusion simulation, TCAD simulation and experimental measurement.

Device	**EOT (nm)**	**x_j (nm)**	**Doping (cm^{-3})**
35 nm	0.88	20	2×10^{18}
13 nm	0.44	8	1.6×10^{19}

Table 3.2: Comparison of basic device parameters including EOT, x_j, and surface doping concentration in the channel.

near the end of the roadmap [51]. The scaling is based on generalised scaling rules [11] and a summary of the equivalent oxide thickness (EOT), extension depth (x_j) and surface doping is given in Table 3.2. The scaling is performed with the intention of preserving the device structure and doping profile shape from the base 35 nm device. The surface doping concentration in the channel is also kept as low as possible, in order to reduce fluctuations due to random discrete dopants, while maintaining suitable control over short channel effects. The scaling also maintains an off current less than $1\,\mu A/\mu m$.

3.4 Simulation Details

In the simulation work that follows, we focus solely on threshold voltage variability. Examining V_T variation is a common way of characterising the effects of statistical variability on MOSFET devices [140], as V_T variations are important for the operation of many types of CMOS circuits and are also related to leakage current variations in the device, which are causing the leakage power crisis in modern VLSI chips. The drift/diffusion formalism is suitable for this study as it provides the required computational efficiency when compared to other methods. DD simulations provide accurate results in the sub-threshold region of operation, where electrostatics dominate, thus the use of threshold voltage to quantify the effect of statistical variability is a natural choice.

For these simulations, we define the threshold voltage using a current criterion, as given by Equation 3.11.

$$I_{V_T} = 10^{-8} \frac{W_{eff}}{L_{eff}} \tag{3.11}$$

The threshold current is scaled by the width and length of the device in order to allow a consistent comparison between the different simulated devices.

In order to ensure accurate results from the simulator, tolerances of 10^{-8} are enforced on the Poisson and current continuity solvers, while an overall accuracy of 10^{-4} is enforced on the Gummel iteration. If these tolerances are not sufficiently high, the simulator has convergence difficulties with the current and tends to take more Gummel iterations for a particular gate voltage and may not be able to determine the current accurately enough to extract the threshold voltage.

Finally, it should also be noted that in this work we focus exclusively on n-Channel transistors, as they have been shown to have greater variability than p-Channel devices [141].

3.5 Summary

In this chapter, some of the underlying features of the Glasgow 3D drift/diffusion simulator were described. Details of the drift/diffusion and density gradient equations were given, along with details of the numerical algorithms used to solve the equations. The approaches used to incorporate random dopants into the simulator, including the charge assignment scheme and dopant placement algorithm were described, as well as the algorithm used to generate line edge roughness patterns. It was necessary to employ advanced Grid technology in order to manage the computational resources and data generated by the simulations performed in this study, and an overview of this technology was given.

The primary test bed device used in this work was a 35 nm MOSFET originally published by Toshiba. The device structure, doping profile and calibration of the "atomistic" simulator to TCAD and experimental measurements were detailed. Specifics of the simulations carried out, such as the threshold voltage criterion used and the accuracy used in the solvers were also given.

In the next chapter, we begin the simulation study with an investigation of statistical variability associated with random discrete dopants, which are the primary source of statistical variability in contemporary bulk MOSFETs.

Chapter 4

Random Discrete Dopants

Random discrete dopants are the dominant source of statistical variability in contemporary bulk MOSFETs, which continue to be the CMOS workhorse at the 45 and 32 nm technology generations [46, 47, 48]. In the design of state of-the-art SRAM and flash memories, occurrences of devices more than 6-7σ from the mean now play important roles, creating the necessity for detailed knowledge of the exact shape of statistical distributions of device parameters far out into their tails [50]. The small sample sizes ($\sim$ 200 devices) that have been traditionally used to characterise statistical variability are insufficient to accurately analyse the shape of a distribution or to accurately determine the magnitude of variability as far as 6 or 7σ from the mean. It is an open question, however, as to what extent the different sources of variability, and random discrete dopants in particular, are capable of producing noticeable variations at 6 or 7σ.

In this chapter we present a detailed 3D simulation study of random dopant induced threshold voltage variation using unprecedented statistical samples of more than 10^5 microscopically different devices. Simulations are performed for conventional bulk n-channel MOSFETs with 35 and 13 nm channel lengths, with sample sizes of 100,000 and 140,000, respectively. Ensembles of this size allow us to predict, with a high level of statistical confidence, the correct shape of the real distributions of parameter fluctuations caused by RDD. A careful statistical analysis of these results then reveals the underlying physics that

Device	**Mean** (mV)	**St. Dev.** (mV)	**Skew**	**Kurtosis**
35 nm	225.9±0.1	30.28±0.07	0.1597±0.008	0.0486±0.02
13 nm	225.9±0.2	81.79±0.16	0.2177±0.007	0.1212±0.02

Table 4.1: Statistical moments for the simulated devices with standard errors computed by bootstrapping (see Appendix A). The mean of V_T has been normalised to a typical value for high performance devices.

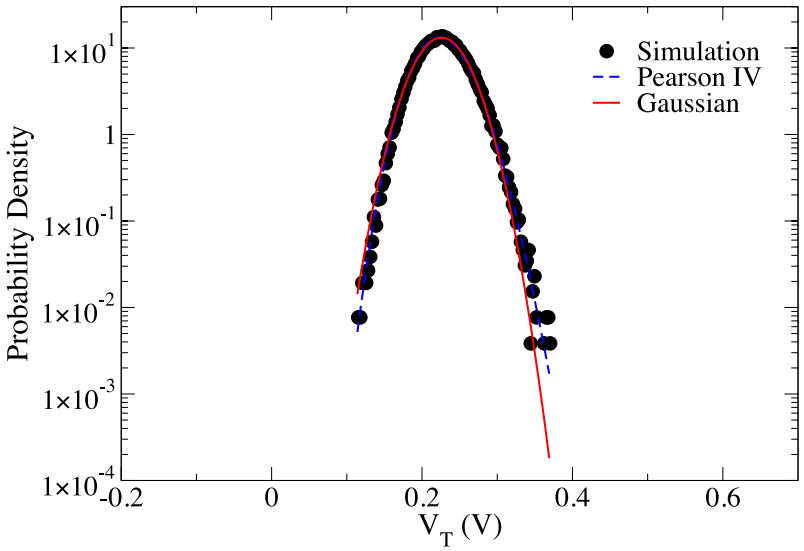
shapes the distribution. Based on this, we then demonstrate a methodology whereby the distribution of V_T can be calculated semi-analytically and show how the procedure can be statistically enhanced. Finally, we present an analysis of the accuracy of this approach.

4.1 Simulation Results

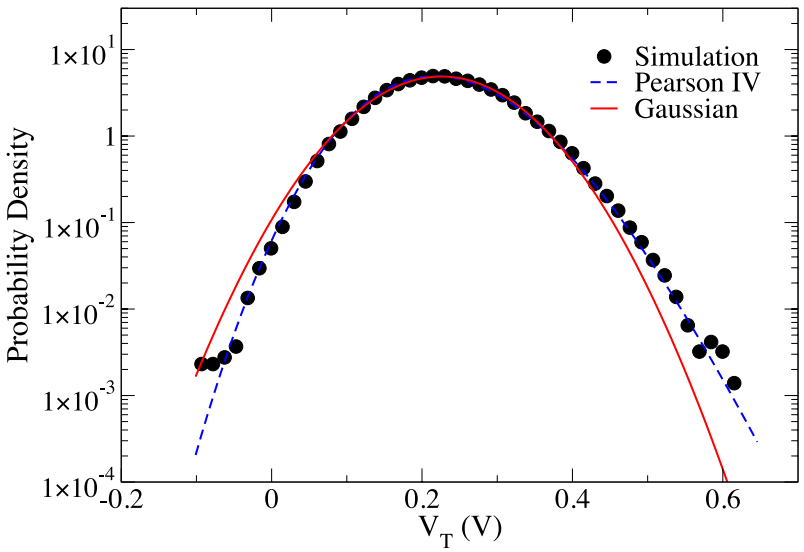
The statistical distributions of the threshold voltage (V_T) obtained from simulations at a low drain voltage of $100\,mV$ are shown in Figure 4.1 (See Section 3.4 for full details of the simulation parameters). Note that, where possible, data for the 35 and 13 nm devices are shown on the same axis scales in order to allow the relative variations in the two devices to be compared. A visual inspection of the data indicates that the distribution of V_T is asymmetric and that the asymmetry is more pronounced in the case of the 13 nm MOSFET. Similar asymmetry has recently been observed in experimentally measured V_T distributions in a large statistical sample of transistors corresponding to the 65 nm technology generation [32], and in flash memory devices [142]. The calculated values for the mean value of V_T, and its standard deviation, skew and kurtosis[1] for both devices are presented in Table 4.1. It is clear from the non-zero values of the skew and kurtosis that the distributions of V_T are asymmetric and that the asymmetry increases with decreasing channel length, confirming the conclusion drawn from visual inspection.

In order to highlight the deficiencies of the common assumption that random dopant induced V_T variability follows a Gaussian distribution, we approx-

[1]Note that this is Fisher's kurtosis, which is defined as $\frac{\mu_4}{\mu_2^2} - 3$ and is 0 for a Gaussian distribution.



(a)



(b)

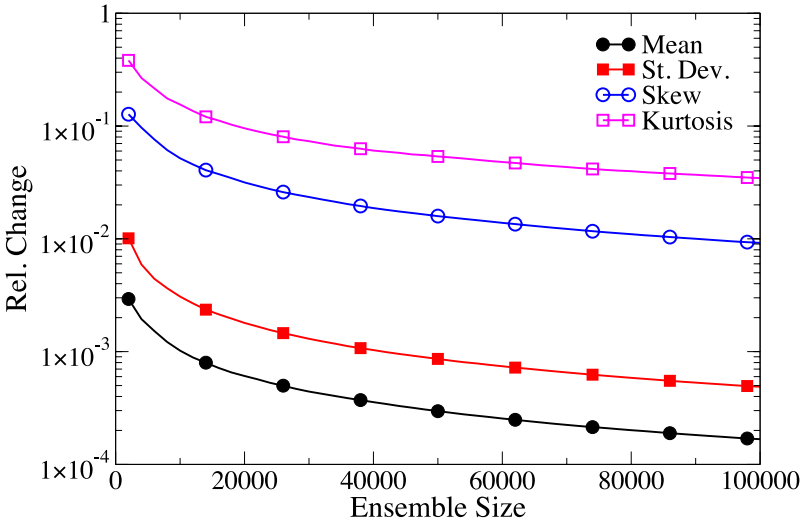
Figure 4.1: Distributions of V_T for the (a) 35 nm and (b) 13 nm devices. Type IV Pearson and Gaussian distributions are shown for comparison. It is clear that the Pearson IV produces a better fit across the entire distribution. For the 35 nm device, the χ^2 error for the Pearson IV is 0.38 vs. 2.4 for the Gaussian and 0.18 vs. 1.5 for the 13 nm device.

imate the distribution of the simulated data for both devices with a Gaussian and a Type IV Pearson distribution. The Type IV Pearson distribution has both skew and kurtosis [143], and is described by Eq. 4.1.

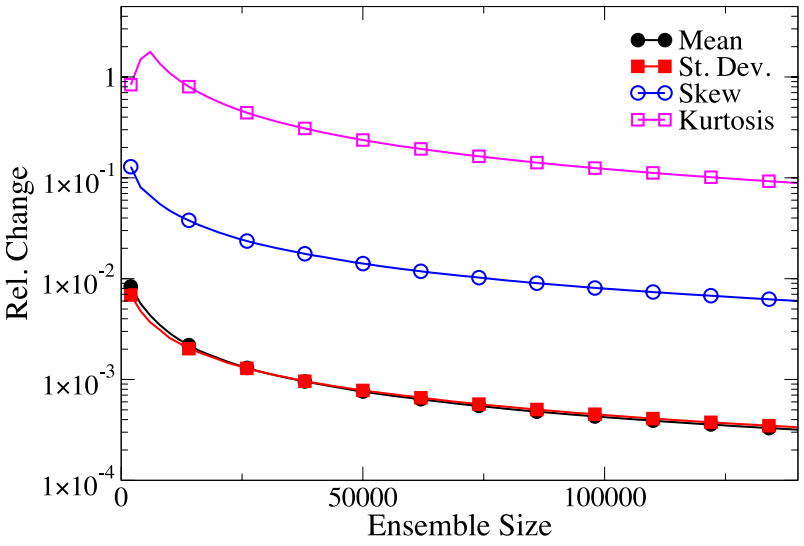
$$P(V_T) = k\left(\frac{(V_T-\lambda)^2}{a^2}+1\right)^{-m} e^{-\nu\, tan^{-1}\left(\frac{V_T-\lambda}{a}\right)} \tag{4.1}$$

The parameters of the Pearson distribution k, m, ν, a and λ are fitted to the data by performing a non-linear curve fit which minimises the total χ^2 error over the full range of data. The mean μ and the standard deviation σ of the Gaussian distribution are simply calculated from the raw device data. The Gaussian and Pearson distributions obtained for the 35 and 13 nm devices are shown, along with the simulated distributions in Figures 4.1(a) and (b). The semi-logarithmic plot provides a detailed view of the discrepancy between the Gaussian distribution and the raw statistical data as one approaches larger values of σ in the tails of the distribution. It is clear that the Pearson distribution provides a significantly better fit to the simulation data, due to its larger number of degrees of freedom. It should be noted that despite the excellent fit provided by the Pearson IV, there is no physical meaning associated with the use of this distribution. It is, however, a useful tool to demonstrate the error introduced by the assumption of a symmetrical distribution, however further analysis is necessary in order to relate the distribution of V_T to the statistics of the underlying physical mechanisms.

Although it is clear that the tails of the distribution of V_T in small devices cannot be accurately represented by a Gaussian distribution, it is important to determine how accurate any characterisation of the threshold voltage actually is. In order to assess the errors in the statistical description of the raw data, the relative change in the first four moments of the two distributions of V_T is calculated as the ensemble size increases. The results are presented in Figure 4.2. The clear convergence observed in all four moments gives a high degree of confidence that the statistical characterisation provides accurate results. As expected, for a given sample size, the error in the estimate of the skew and the kurtosis is significantly higher than the error in the mean and the standard deviation indicating that very large samples are required to accurately deter-



(a)



(b)

Figure 4.2: Relative change in the first four statistical moments of the distributions of V_T as a function of sample size for the (a) 35 nm device and (b) 13 nm device.

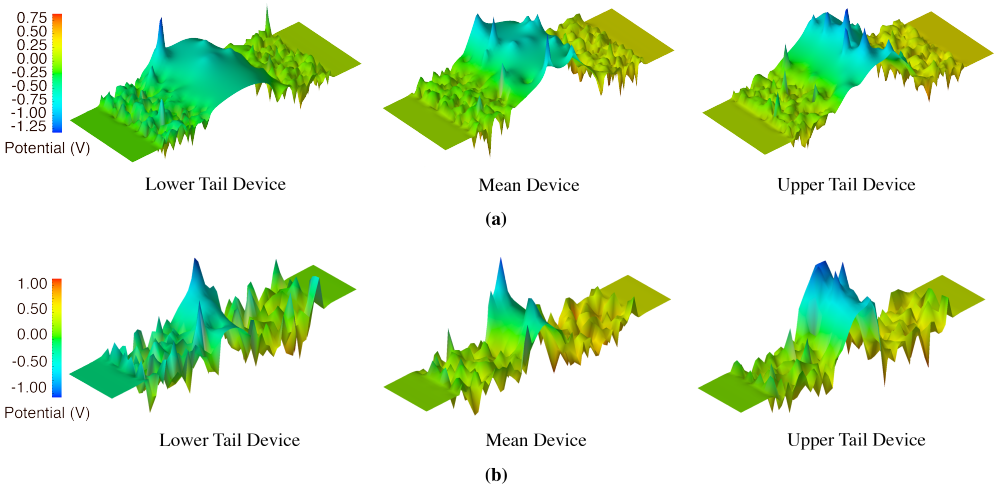


Figure 4.3: Raw electrostatic surface potential profiles for devices in the lower part, middle and upper part of the distributions for (a) 35 nm devices and (b) 13 nm devices.

mine the higher order moments of the distributions and to correctly quantify the asymmetry.

In order to better understand the physical mechanisms whereby random discrete dopants affect V_T it is also instructive to study the surface potential distributions of devices from across the distribution at identical gate voltages. Since the channel determining the current flow is close to the Si/SiO_2 interface, it is reasonable to assume that it is dopants in the vicinity of the interface that will have the greatest impact on the threshold voltage. Devices close to the mean V_T, and from the upper and lower tails of the distributions of V_T for both 35 and 13 nm transistors are selected and the surface potential plots for the selected devices are shown in Figures 4.3(a) and (b). At both channel lengths the behaviour of the devices with higher V_T is determined by the clustering of dopants across the channel width at the location of the maximum of the potential barrier between the source and the drain. At this position the dopants have the maximum impact on V_T by almost completely blocking the current path. Conversely, the behaviour of the transistors from the low end of the distribution of V_T is determined by the lack of dopants in the part of the channel near the potential barrier maximum, creating an open current path

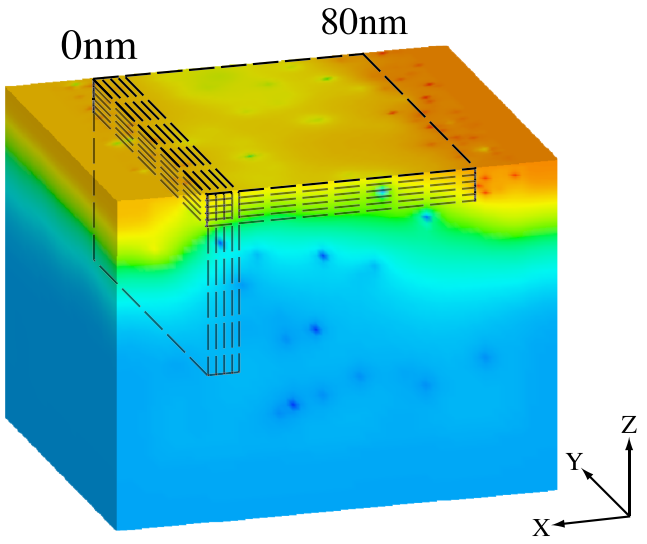


Figure 4.4: The device is divided into 1 nm slices in both the X and Z axes. The number of dopants in each box is used to calculate the correlation between position and threshold voltage. The extent of the SSR shown in Figure 4.6(a) is indicated.

responsible for the low V_T. We can see that in the high V_T devices, there are no current channels through valleys in the surface potential profile, since the valleys are blocked by spikes in the potential associated with random discrete dopants. Comparing this to the low V_T transistors, it is clear that there are many more available paths through which current flows, due to the lack of dopants. This results in early turn-on of the device and thus lower threshold voltage. In mean V_T devices, there are both valleys through which current can flow and spikes which obstruct current flow at the position of the potential barrier maximum.

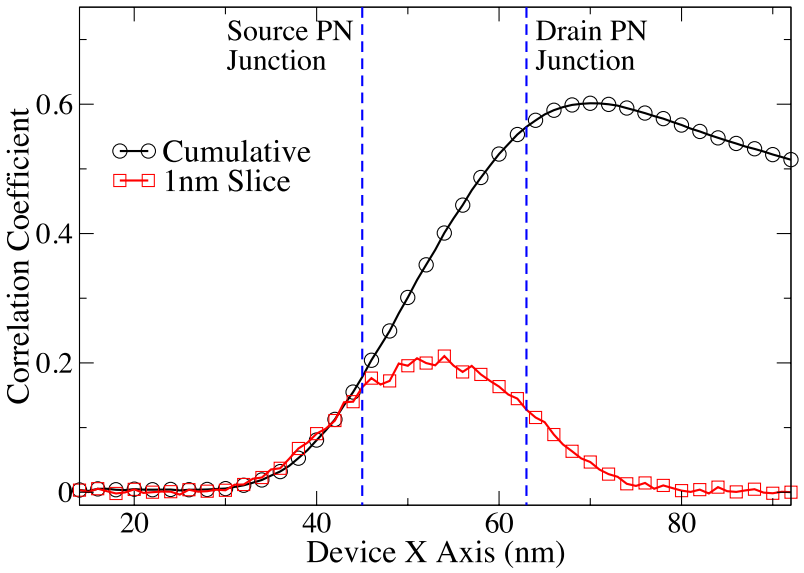
4.2 Statistical Analysis

Early theoretical analysis has suggested that asymmetry in random dopant induced V_T variation [144, 145] can be attributed to the Poisson distribution governing the number of dopants in the gate depletion region [57]. It is known that the number of dopants in the channel of a device affects the threshold

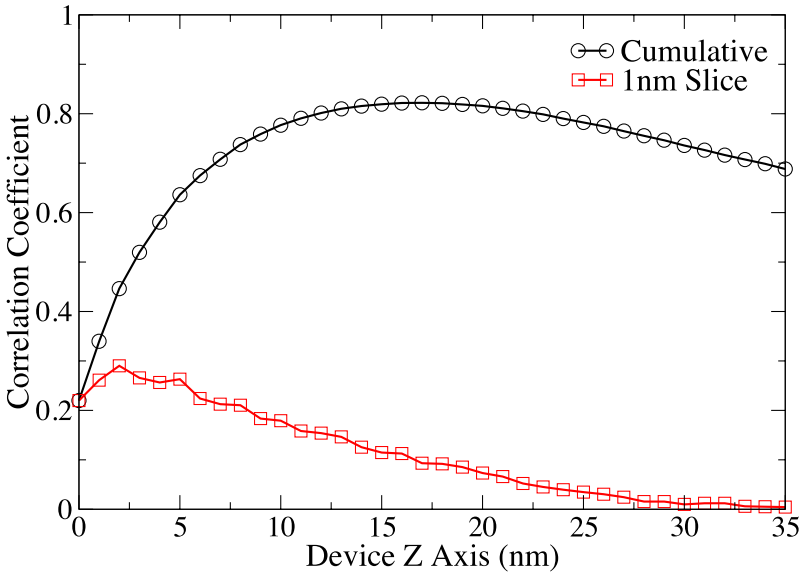
voltage, as discrete dopants cause a localised increase of the potential in the channel. However, the number of dopants is not the only source of V_T variation. The position of dopants [146, 69] must also be considered, as many different configurations of dopant position will occur for a fixed number of dopants. Modelling V_T variation using only total dopant number neglects variation due to dopant position and leads to an unphysical truncation of the lower tail of the distribution [74]. Thus, the first step in a more detailed analysis is to determine in which region variations in local dopant density have the greatest effect on the threshold voltage. This will help us to define the statistically significant region (SSR) of the transistor, in which random dopants dominate the statistical behaviour of the device ensemble.

In order to accomplish this we have calculated the correlation coefficient between V_T and the total dopant number in a series of 1 nm deep horizontal slabs bounded by the source and the drain, starting from the oxide interface and ranging down through the device body. A similar procedure is repeated for 1 nm wide vertical slabs ranging through the channel from source to drain, as illustrated schematically in Figure 4.4. The correlation between V_T and dopant x position is shown in Figure 4.5(a) for both 1 nm thick slices d nm from the interface and for slices d nm thick (i.e. the cumulative sum of the 1 nm slices up to and including that position). From this figure we see that the largest calculated correlation is for dopants between the PN junctions of the device. Figure 4.5(b) shows the correlation between threshold voltage and dopant z position, again for 1 and d nm thick slices and clearly indicates that dopants near the interface make the most significant contribution to V_T fluctuations. By combining the 1D correlation in x and z from the 1 nm slices, a two-dimensional map of the correlation between the position of an individual dopant within the SSR and V_T can be constructed, which is plotted in Figures 4.6(a) and (b) for both the 35 and 13 nm devices.

The SSR is bounded by the metallurgical junctions of the source and drain and extends approximately 20 nm down from the interface in the 35 nm device and approximately 10 nm from the interface in the 13 nm transistor. These values compare closely with the depletion depths for these devices, which are $\sim 25\,nm$ and $\sim 9\,nm$ in the 35 and 13 nm devices, respectively. This conclusion

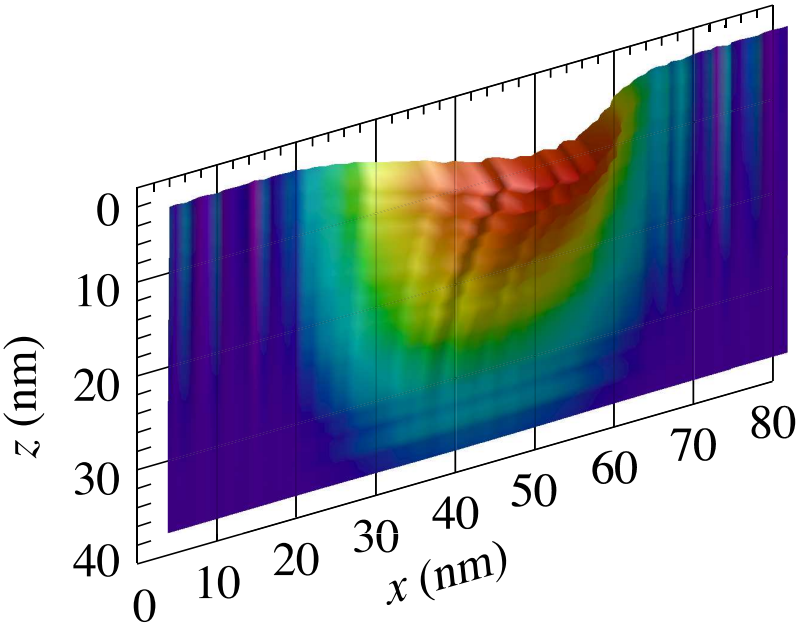


(a)

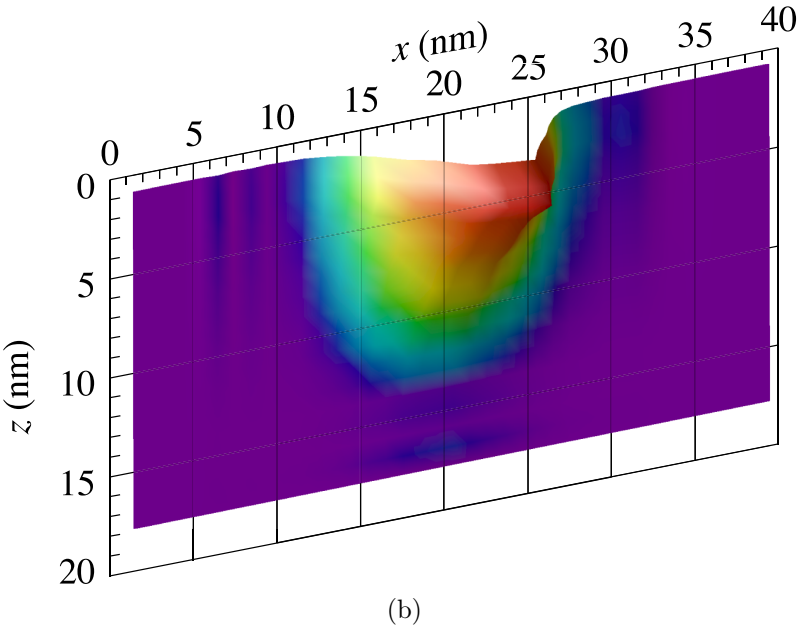


(b)

Figure 4.5: Correlation between dopant position and threshold voltage for the 35 nm device (a) in the X axis and (b) in the Z axis. Note that $z = 0\,nm$ is at the oxide interface.



(a)



(b)

Figure 4.6: The two dimensional correlations of dopant position and V_T for (a) the 35 nm device and (b) the 13 nm device. The statistically significant region can be determined visually from these plots.

is consistent with previous theoretical studies [26] and experimental observations [47] that it is dopants in the channel that have the greatest influence on V_T fluctuations. It is interesting to note that for the two simulated devices, the maximum correlation between dopant position and V_T is not at the oxide interface, but approximately 1.5 nm below it. This is due to the density gradient quantum corrections used in the simulations, which force the maximum of the carrier distribution away from the surface, and determining the position at which the device is most sensitive to the presence of random dopants within the channel. The density gradient carrier distribution is consistent with that obtained from the self-consistent solution of the 1D Poisson-Schrödinger equation [104].

By choosing devices with a fixed number of dopants within the SSR (N_{SSR}) it is possible to estimate the distribution of the threshold voltage caused solely by the random position of dopants. Figures 4.7(a) and (b) illustrate the evolution of the distribution of V_T as a function of N_{SSR} for the 35 and 13 nm transistors respectively. As illustrated in Figure 4.8 both the mean and standard deviation of the threshold voltage distributions increase linearly with N_{SSR}. For densely populated samples with a constant number of dopants in the SSR (around the mean value $\overline{N_{SSR}} = 44$ for the 35 nm and $\overline{N_{SSR}} = 20$ for the 13 nm transistor) the calculated skew and kurtosis are small, leading to the conclusion that the distributions of threshold voltages due to random dopant position for fixed N_{SSR} are Gaussian. In order to verify this hypothesis, we use the Mann-Whitney test [147], which tests the null hypothesis that two samples are drawn from the same underlying population. Several of the positional distributions for the 35 nm device were tested against 10,000 samples randomly generated from a Gaussian distribution with the same mean and standard deviation as the data and the statistics of the p-values obtained can be seen in Table 4.2. Taking the standard statistical significance level of $\alpha = 0.05$, we see that there are no p-values close to or below this level, therefore we accept the null hypothesis that the distributions are Gaussian.

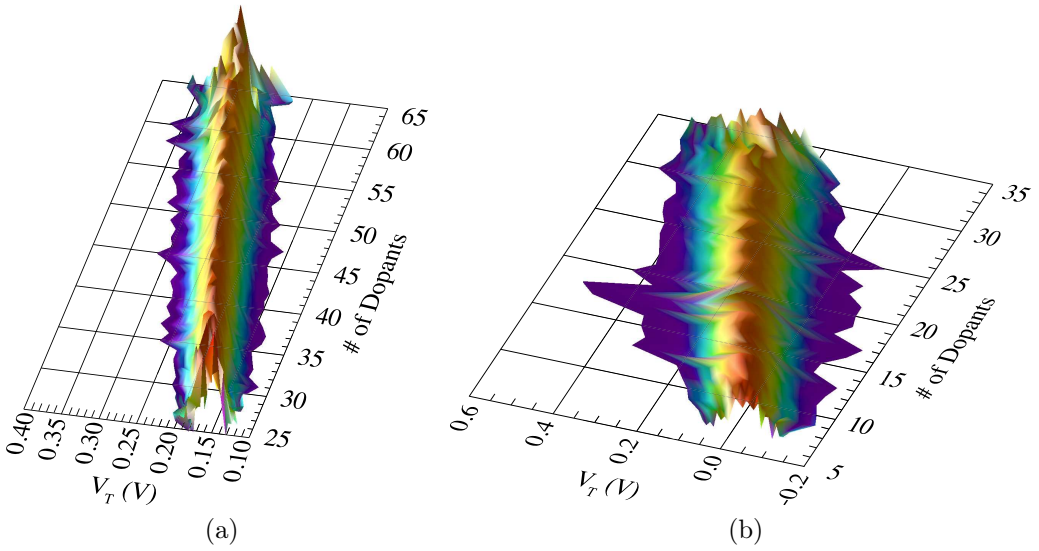


Figure 4.7: The distribution of V_T as a function of number of dopants, N_{SSR}, in the SSR for (a) the 35 nm transistor and (b) the 13 nm transistors. For a fixed N_{SSR}, the distribution of V_T is determined by dopant position. Note the increasing mean and standard deviation as a function of N_{SSR}.

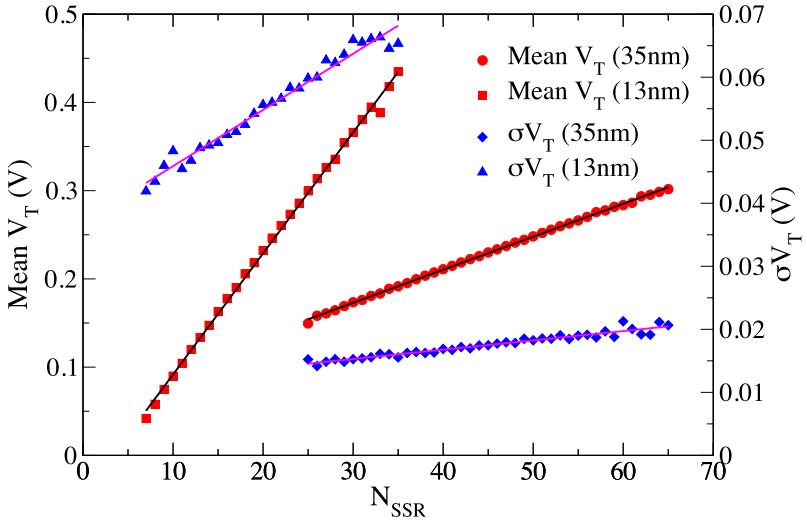


Figure 4.8: The dependence of the V_T mean and standard deviation as a function of N_{SSR} for both devices. The linear dependence allows positional effects on V_T to be extrapolated out to larger values of σ.

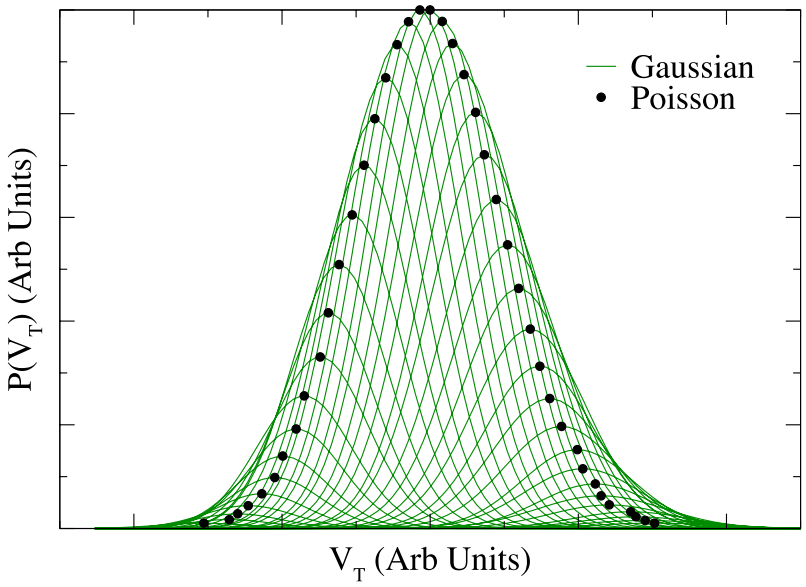
N_{SSR}	Mean	St. Dev.	Min	Max
40	0.856	0.108	0.330	0.999
41	0.765	0.145	0.210	0.999
42	0.771	0.145	0.262	0.999
43	0.790	0.141	0.244	0.999
44	0.798	0.138	0.246	0.999
45	0.518	0.148	0.119	0.999
46	0.836	0.120	0.351	0.999
47	0.847	0.113	0.365	0.999
48	0.733	0.143	0.210	0.999

Table 4.2: Statistics of the p-values obtained by conducting Mann-Whitney tests for the positional distributions for the 35 nm device against 10,000 random Gaussians. As there are no p-values below 0.05, we accept the null hypothesis that the positional distributions are Gaussian.

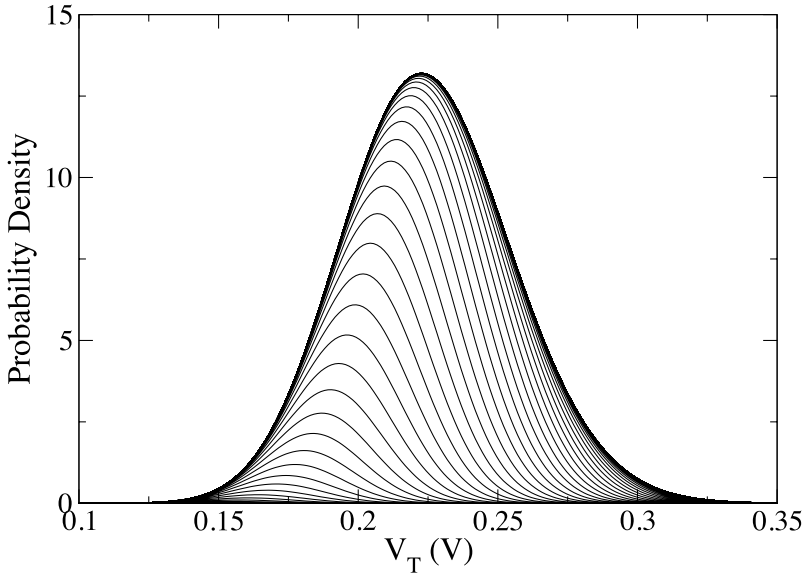
4.3 Constructing the Distribution of V_T

As previously stated, there are two components that contribute to RDD-induced V_T variation – the number of dopants in the SSR and the position of these dopants. It should be noted that since the device has a finite number of sites in the underlying Si crystal lattice, it is both feasible and correct for duplicate dopant distributions to occur and these should not be excluded from consideration.

The number of dopants in the SSR can be described by either a Binomial distribution [148] or a Poisson distribution [68]. The Binomial distribution converges to a Poisson distribution for a sufficiently large number of trials and we consequently assume that a Poisson distribution governs the number of dopants in the SSR as, in this case, there are on the order of 10^6-10^7 trials, which is the approximate number of lattice sites at which a dopant can be placed in the Silicon crystal. We have already shown that the variation in V_T due to position (i.e. when N_{SSR} is fixed) is Gaussian. Although there is a different Gaussian for each N_{SSR}, this can also be seen as a single Gaussian with varying μ and σ that is moved through the Poisson distribution. This represents a convolution operation and the complete distribution of V_T is thus the discrete convolution of a Poisson distribution with a mean value $\overline{N_{SSR}}$



(a)



(b)

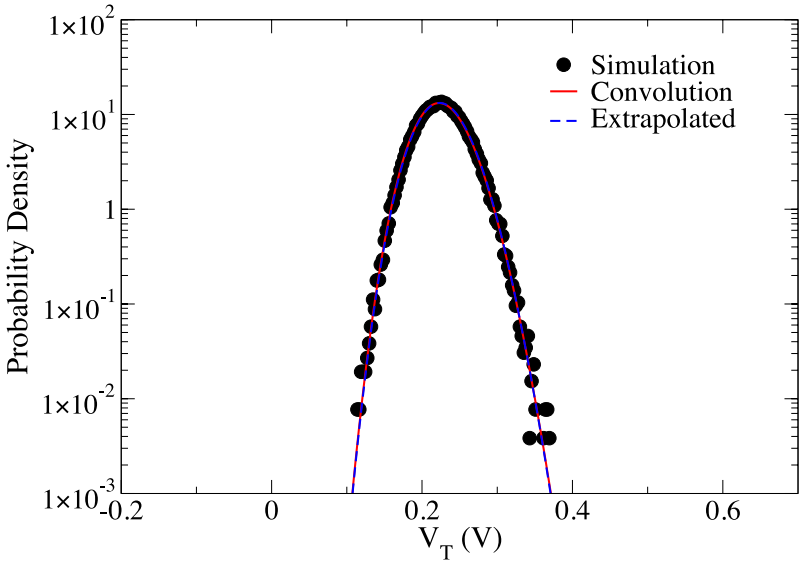
Figure 4.9: (a) Illustration of how the variation that comes from each fixed value of N_{SSR} contributes to the total variation. The Gaussians are weighted by the corresponding probability from the Poisson distribution. (b) Illustration of how the overall distribution (for the 35 nm device) converges as the summation in Equation 4.2 progresses.

$[P(i, \overline{N_{SSR}})]$, governing the number of dopants in the SSR, and the Gaussian distribution of V_T for fixed N_{SSR} $[G(V_T, \mu(N_{SSR}), \sigma(N_{SSR}))]$, as expressed by Eq. 4.2. Note that the limits of the summation, N_{SSRmin} and N_{SSRmax}, should be selected such that the probability of that number of dopants occurring is low enough to be effectively zero.

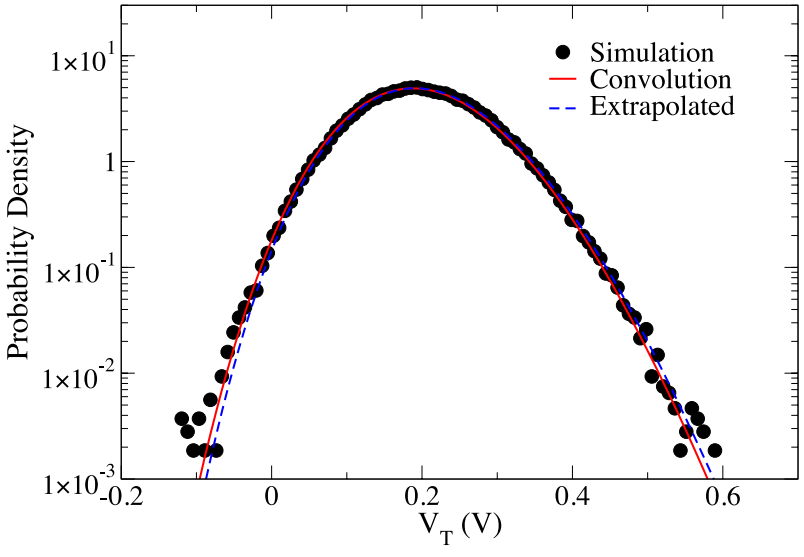
$$P(V_T) = \sum_{i=N_{SSRmin}}^{N_{SSRmax}} P(i, \overline{N_{SSR}}) \cdot G(V_T, \mu(i), \sigma(i)) \tag{4.2}$$

Figure 4.9(a) illustrates how the two components contributing to the RDD-induced variation combine to give the total variation. From this, it is clear that for different values of N_{SSR}, there are particular combinations of dopants that give rise to the same threshold voltage. For example, particular configurations of dopants for $N_{SSR} = 43$ and $N_{SSR} = 44$ may both give rise to a V_T of $220\,mV$. The total probability of obtaining any given threshold voltage is thus the sum of the probabilities arising from each of these positional distributions, weighted by the probability of that particular number of dopants occurring (as given by the Poisson distribution).

Figure 4.9(b) shows how the distribution of V_T converges as the summation in Equation 4.2 progresses, for the 35 nm device. Figures 4.10(a) and (b) show the semi-logarithmic distribution of V_T for both the 35 and 13 nm devices. The convolution is calculated by direct extraction of the values for $\mu(N_{SSR})$ and $\sigma(N_{SSR})$ from the simulation data and by fitting straight lines for these functions and extrapolating (see Section 4.3.1). It can clearly be seen that both of the calculated convolutions match the distribution of data obtained from simulation extremely well. Figures 4.11(a) and (b) show Quantile-Quantile (Q-Q) plots [149] of the data. Q-Q plots are a graphical statistical technique for testing the null hypothesis that data follows a certain distribution. The reference distribution appears as a straight line and any departures from this line indicate a difference between the data and the reference distribution. Figure 4.11(a) compares the simulation data to a Gaussian distribution with the same mean and standard deviation as the data, clearly demonstrating the assumption that the data follows a Gaussian distribution is incorrect, with the

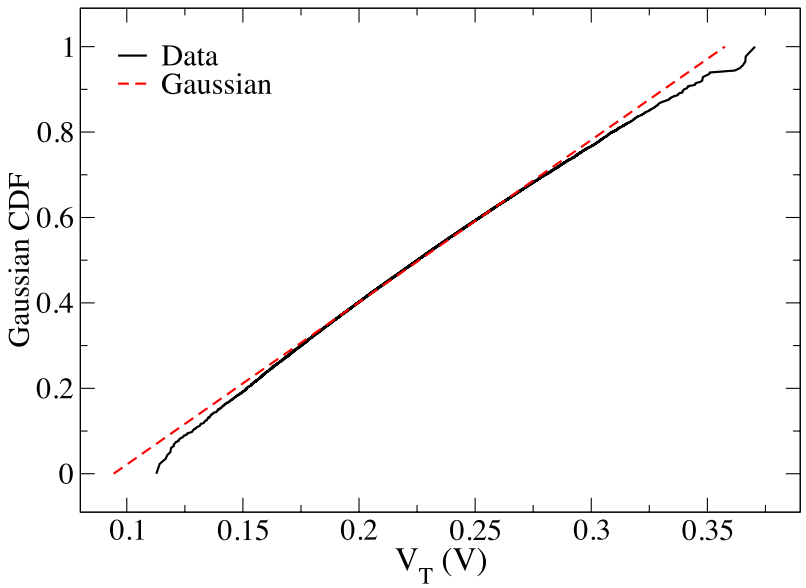


(a)

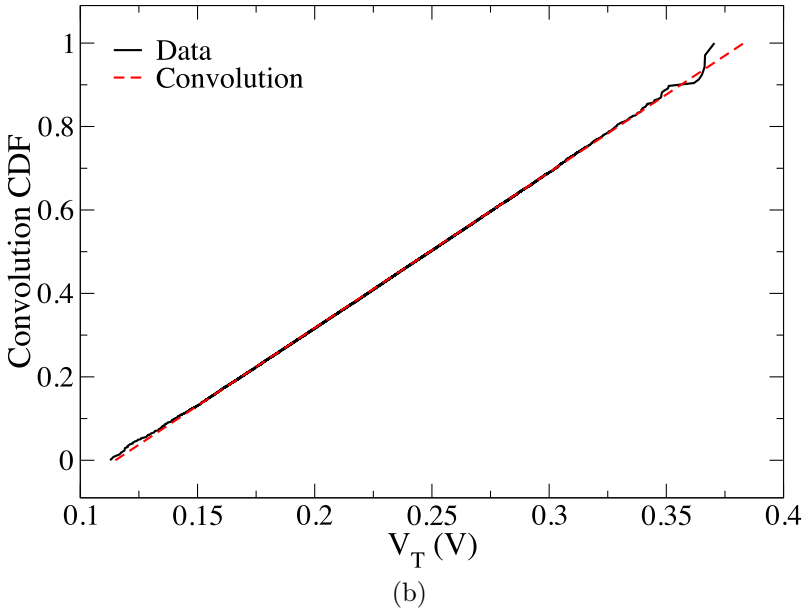


(b)

Figure 4.10: Comparison between the simulated V_T distribution and the convolution. The convolution is calculated directly from the simulation data and by extrapolating the functions $\mu(N_{SSR})$ and $\sigma(N_{SSR})$ shown in Figure 4.8. Both methods show excellent agreement with the simulation data.



(a)



(b)

Figure 4.11: Quantile-Quantile plots comparing the simulation data for 35 nm with (a) a Gaussian distribution and (b) the semi-analytical convolution.

deviation from Gaussian increasing significantly in the tails. The simulation data is compared to the distribution generated using our convolution model in Figure 4.11(b) and it is can be seen that both closely match.

From this analysis it becomes clear that the asymmetry in the random dopant induced threshold voltage distribution is due to two factors: firstly the Poisson distribution for a fixed value of $\overline{N_{SSR}}$ is asymmetric with positive skew and this asymmetry increases as $\overline{N_{SSR}}$ is reduced and secondly because the standard deviation $\sigma(N_{SSR})$ increases with N_{SSR}, causing the upper tail of the distribution to become extended.

4.3.1 Statistical Enhancement

As Figure 4.8 clearly shows, the mean $\mu(N_{SSR})$ and the standard deviation $\sigma(N_{SSR})$ of the positional distributions depend linearly on the number of dopants. It is straightforward to calculate a linear regression for these functions and from this it is possible to extrapolate the values of $\mu(N_{SSR})$ and $\sigma(N_{SSR})$ in order to calculate values from the distribution of V_T to arbitrary values of σ_{V_T}. It should of course be noted that these extrapolations are only valid within the limits of the physics included in the simulator. Within these limits, this approach can be used to obtain distributions that are equivalent, in terms of information content, to those obtained from brute force simulation of very large statistical samples. While in Figure 4.8 the regression fits are calculating using all of the available simulation data, this need not be the case. For example, in the 35 nm MOSFET case, the extrapolated results in Figure 4.10(a) are obtained as follows. For this device, $\overline{N_{SSR}} = 44$, and we have selected V_T distributions for $N_{SSR} = 42$ and $N_{SSR} = 46$. This yields 3 points with which to calculate the regressions for $\mu(N_{SSR})$ and $\sigma(N_{SSR})$. We then extrapolate for $N_{SSR} = 1$ to 200 to obtain the positional Gaussians and calculate the distribution of V_T using Equation 4.2. The distribution obtained from Equation 4.2 using the extrapolated values actually better fits the simulation data than the convolution calculated with the Gaussian distributions extracted from simulation. This is evident in the reduction of the calculated χ^2 error, which decreases from 0.94 for the distribution from extracted values

to 0.55 for the extrapolated. This improvement in the statistical error is due to the fact that the linear extrapolation removes noise from the values of $\mu(N_{SSR})$ and $\sigma(N_{SSR})$ at small and large N_{SSR}, where the sampled populations of V_T become small.

Thus, a possible method for the statistical enhancement of random dopant induced variability simulations includes the following steps:

1. **Determine the SSR** – An SSR must be selected for the given device in order to avoid the decorrelating effect of dopants that do not significantly influence the threshold voltage. The effect of the selection of SSR is examined in Section 4.4.

2. **Estimate $\overline{\mathbf{N_{SSR}}}$** – The value of $\overline{N_{SSR}}$ can be estimated by generating a large sample of devices and determining the average number of dopants in the previously selected SSR. It should be noted that while N_{SSR} can only take non-negative integer values, $\overline{N_{SSR}}$ can be any positive real number. It is not necessary to perform any simulations at this point – only the device structure need be generated. Note also that $\overline{N_{SSR}}$ can be determined by integrating the continuous doping profile in the SSR and that atomic level process tools that accurately model the implantation process, if available, may be useful for this process.

3. **Select devices for simulation** – From the large sample of generated devices, select devices with, for example, $\overline{N_{SSR}}-\Delta$, $\overline{N_{SSR}}$ and $\overline{N_{SSR}}+\Delta$ dopants in the SSR and simulate n of these devices. The effect of the choice of Δ and n is also examined in Section 4.4.

4. **Estimate $\mu(\mathbf{N_{SSR}})$ and $\sigma(\mathbf{N_{SSR}})$** – The linear functions that characterise the positional Gaussian distributions can be estimated from the simulation results.

5. **Calculate $\mathbf{V_T}$** – Equation 4.2 can be used to calculate the distribution of V_T.

The process of generating atomistic devices is significantly less computationally demanding than their simulation, and as we have shown with the above

example, this methodology is capable of accurately predicting the distribution obtained via pure "brute force" simulation. Since only a subset of devices are simulated with this approach, it could be used to dramatically speed up the accurate evaluation of V_T fluctuations. This method is also independent of the actual continuous doping profile in the simulated devices.

4.4 Error Analysis

In order to employ the statistical enhancement methodology described in the previous section, several assumptions must be investigated, including the selection of the SSR, the choice of Δ and the sample size n. In order to examine the robustness of the semi-analytical method, we have analysed the impact of these three factors on the accuracy of the final calculated distribution. The analysis is performed on the simulations results obtained for the 35 nm MOSFET and the simulation data obtained from the large ensemble is used as the "gold standard" to which all other results are compared.

4.4.1 Choosing the SSR

In this simulation study, the SSR was determined by directly calculating the correlation between dopant position and threshold voltage. Unfortunately such an approach requires the simulation of a very large statistical sample to accurately calculate the correlation. In order to make the statistical enhancement methodology more efficient, it would be useful to determine the SSR based on structural device parameters. We therefore examine how critical the choice of SSR is to the accuracy of the calculated distribution of V_T. This can be achieved by varying the length and depth of the SSR in turn, as shown in Figure 4.12, and then assessing the impact that the change in the SSR boundaries has on the accuracy of the calculated distribution of V_T. Throughout this analysis, it is assumed that the SSR encompasses the entire width of the device.

The selection of the SSR has a direct effect on the Poisson distribution of N_{SSR} through the mean value $\overline{N_{SSR}}$, as different numbers of dopants will fall

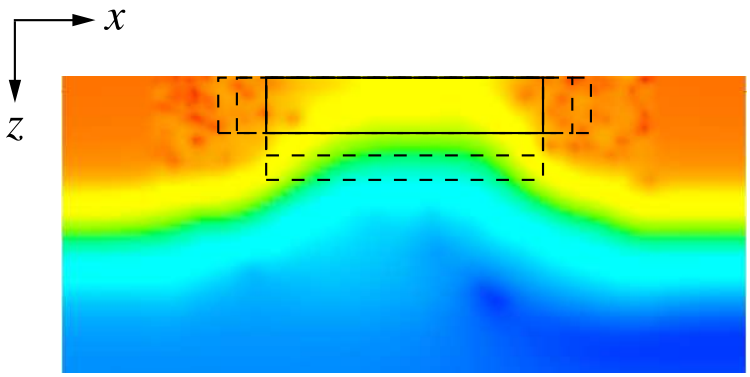
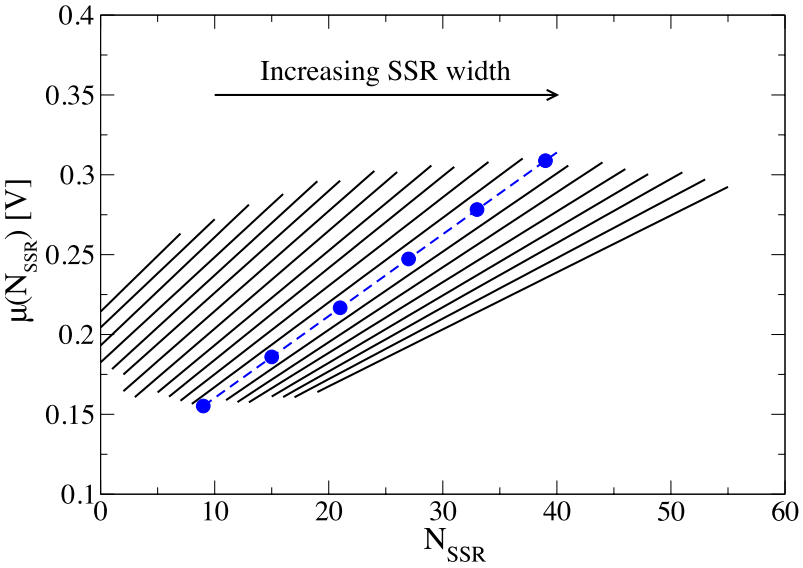


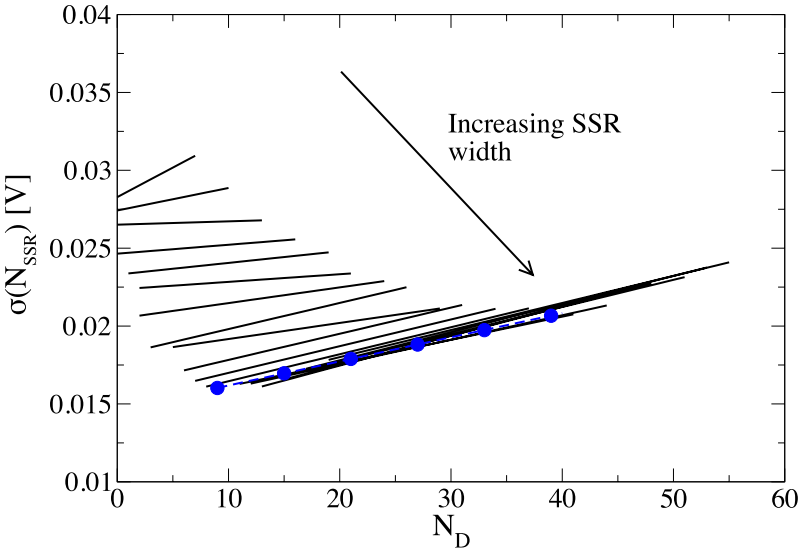
Figure 4.12: Demonstration of how the SSR is varied to determine the effect of SSR size on the error of the constructed distribution.

in different configurations of the SSR for the same device. This changes both the location and shape of the Poisson distribution, since the moments depend on the mean value. It would thus be reasonable to expect that changes in this value will have a significant effect on the accuracy of the final distribution. However, since for each particular choice of SSR a given device may have a different value of N_{SSR}, it will fall into a different bin, changing the parameters of the positional distributions. This causes the functions $\mu(N_{SSR})$ and $\sigma(N_{SSR})$ to vary and they must be re-extracted. The effect of SSR length is examined by fixing the SSR to the centre of the channel and varying the length from 2 to 40 nm. The corresponding linear dependences of $\mu(N_{SSR})$ and $\sigma(N_{SSR})$ are shown in Figure 4.13, clearly indicating that changes in the SSR result in different slopes and y-axis intercepts for $\mu(N_{SSR})$ and $\sigma(N_{SSR})$. Varying the depth of the SSR (while keeping the length fixed) also results in similar trends for $\mu(N_{SSR})$ and $\sigma(N_{SSR})$.

Next, we examine the effect of SSR width on the χ^2 error of the calculated distribution, which is used as a measure of the accuracy of the calculation. For comparison, the lowest χ^2 value obtained from the SSR estimated in Figure 4.6(a) was 0.55. From the results shown in Figure 4.14(a), it is clear that a very narrow SSR will lead to a poor prediction of the distribution of V_T, since it is obviously unphysical to expect the small number of dopants in this narrow region to completely dominate the fluctuation of V_T. We also expect that a

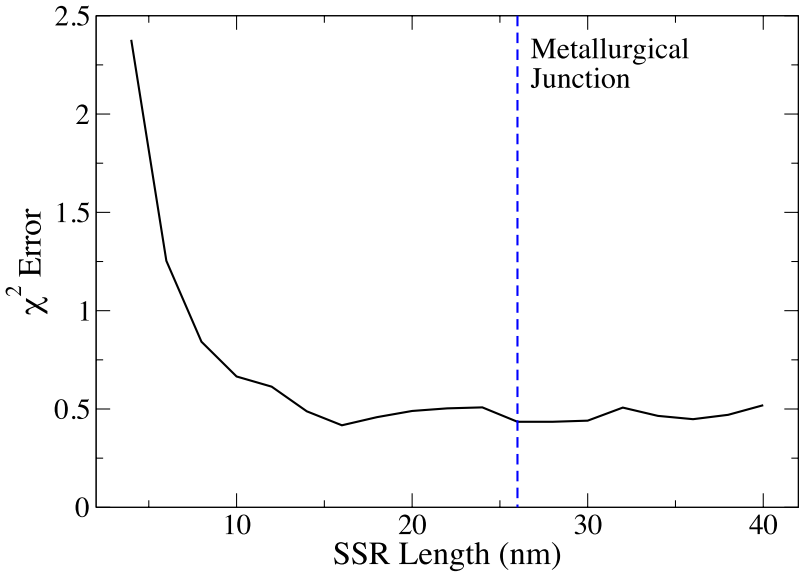


(a)

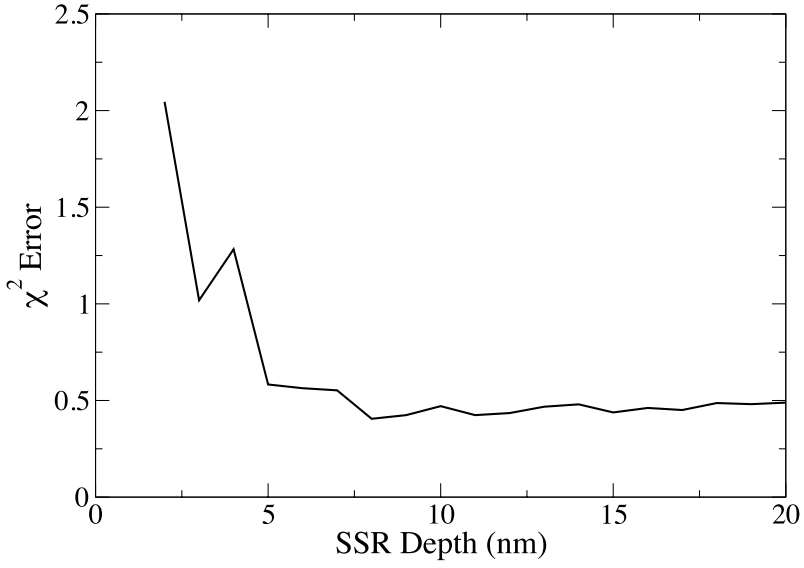


(b)

Figure 4.13: Functions (a) $\mu(N_{SSR})$ and (b) $\sigma(N_{SSR})$ for SSR lengths from 2 nm to 40 nm. The dashed lines indicate where the SSR bounds correspond to the metallurgical PN junctions.



(a)



(b)

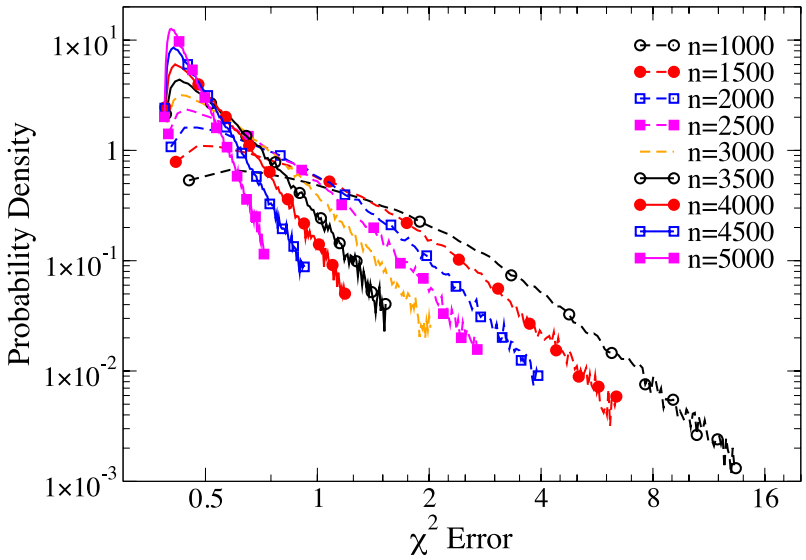
Figure 4.14: Dependence of the χ^2 error of the extrapolated distribution on (a) the length of the SSR and (b) the depth of the SSR. The dashed line in (a) represents where the edges of the SSR overlap the metallurgical junctions.

very large SSR would result in a poor estimation of the distribution of V_T since the SSR would then include large numbers of dopants that do not contribute significantly to V_T variations and which cause some degree of de-correlation. Once the SSR reaches a width sufficient to cover of the majority the channel, the calculated error becomes very close to the previous optimum value and there is little improvement in the error beyond this. In practice, after this point, changes in the linear functions actually compensate for the changes in the Poisson distribution, resulting in a constructed distribution that still accurately matches the simulated distribution. This suggests that the methodology is rather robust to the choice of the SSR, which can therefore be based solely on knowledge of the transistor structure. Based on the information presented in Figure 4.14(a), it is reasonable to select the metallurgic PN junctions near the oxide interface as a reliable estimate of the lateral bounds of the SSR. Therefore the length of the SSR may be determined from the 2D doping profile of the device, which in the case of the 35 nm device investigated here is $\sim 26\,nm$, with the SSR centered in the channel. Similarly, we can examine the effect of SSR depth on the final χ^2 error value. Figure 4.14(b) indicates that the SSR must be sufficiently large to capture the depletion layer under the gate. The theoretical maximum depletion layer width can be estimated using Equation 4.3 and in the case of the 35 nm device is $\sim 25\,nm$ and as Figure 4.14(b) indicates, the error drops off well within this depth, thus this value can safely be used as the vertical extent required for the SSR.

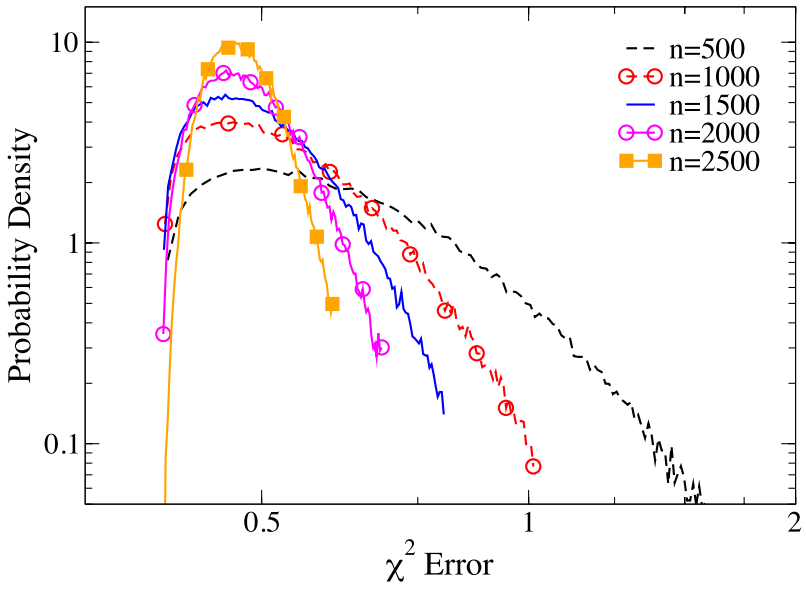
$$W_d = \sqrt{\frac{4\epsilon_{Si}k_BT\,ln\left(\frac{N_a}{n_i}\right)}{q^2N_a}} \tag{4.3}$$

4.4.2 Impact of Sample Size

In the statistical enhancement methodology described in Section 4.3.1, n devices with $\overline{N_{SSR}} - \Delta$, $\overline{N_{SSR}}$ and $\overline{N_{SSR}} + \Delta$ dopants in the SSR are selected for simulation in order to establish the functions $\mu(N_{SSR})$ and $\sigma(N_{SSR})$. It is useful to examine the effect of both n and Δ on the accuracy of the calculated distribution in order to determine the optimal subset of devices to simulate.



(a)



(b)

Figure 4.15: Distribution of χ^2 errors for different sample sizes for (a) $\Delta = 2$ and (b) $\Delta = 8$. The errors are calculated for 100,000 random convolutions drawn from the original data. Note that the x-axis is different in both plots.

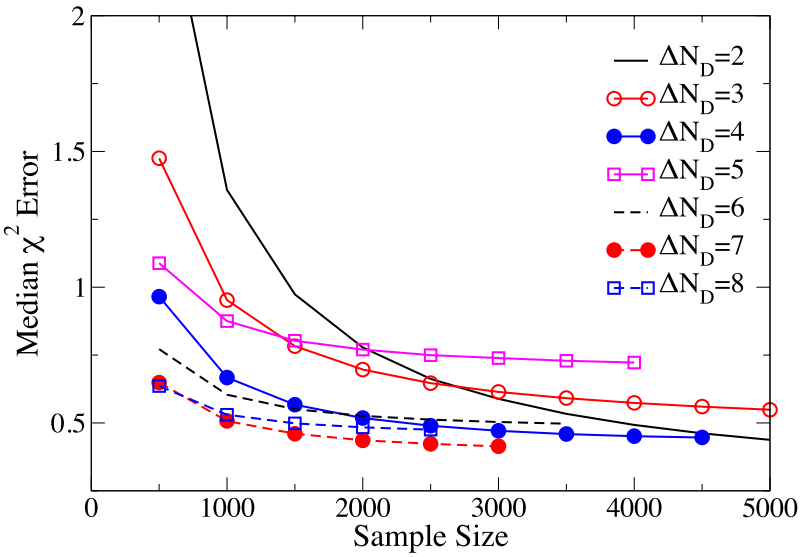


Figure 4.16: Median χ^2 error as a function of sample size for $\Delta = 2 - 8$. The errors are small for sample sizes above $\sim 2,000$ for most values of Δ.

Again, the original simulation data is used as the "gold standard" against which the errors are calculated.

First, we examine the distribution of errors for the calculated distribution for different sample sizes, with $\Delta = 2$ and $\Delta = 8$. Errors are calculated by repeated random sub-sampling of the simulation data – a type of cross-validation. In this case, sub-samples of n devices having $\overline{N_{SSR}} - \Delta$, $\overline{N_{SSR}}$ and $\overline{N_{SSR}} + \Delta$ dopants in the SSR are drawn from the data and the convolution calculated for these random sub-samples. The χ^2 error is determined by comparing this random convolution to that generated from the original large ensemble. This procedure is repeated 100,000 times to yield a distribution of the errors for the given sample size and value of Δ.

The distributions of errors are shown in Figures 4.15(a) and (b) respectively. From this, it is clear that the average and spread of the error is generally larger for a given sample size and $\Delta = 2$, compared to $\Delta = 8$. Indeed, the accuracy is rather poor for $\Delta = 2$ and a small sample size. If we consider the linear regression used to determine $\mu(N_{SSR})$ and $\sigma(N_{SSR})$, it is obvious that when Δ is small, noise in the values of μ and σ will have a greater influence on the regression than when Δ is large. As Figure 4.15(b) shows, a relatively accurate

distribution can be obtained with a sample size of only 1,000 when $\Delta = 8$.

The median χ^2 error as a function of sample size is shown in Figure 4.16 for $2 \leq \Delta \leq 8$, which confirms these observations and also indicates that for larger values of Δ, the improvement in error is small above sample sizes of $\sim 2,000$. Using this information, it can be seen that an accurate prediction of the distribution of V_T can be obtained with, for example, $\Delta = 8$ and $n = 2,000$.

From this analysis, it is clear that by employing the methodology described in Section 4.3.1 the evaluation of V_T fluctuations can be made considerably more efficient. The distribution calculated using the statistical enhancement methodology accurately reproduces the location, spread and shape of the distribution obtained from the brute force approach. We can see that the same level information content obtained through the brute-force simulation of 100,000 different devices can be obtained from a comparatively small sample of only 6,000 – a reduction of more than an order of magnitude in terms of the amount of computational time necessary to characterise the effect of random discrete dopants.

4.5 Summary

In this chapter, the statistical distribution of threshold voltage due to random discrete dopants has been accurately characterised through the simulation of statistical samples with size $> 10^5$. The results show that the distribution of threshold voltage in ultra-small devices is asymmetric and that the asymmetry increases with scaling. The simulation of large statistical samples allows the variation and shape of the distribution to be accurately analysed.

Through data mining and statistical analysis of the data generated from simulation, we show that there are two components that contribute to threshold voltage variation due to random dopants – the number of dopants in the SSR and their random positions. The total variation in the threshold voltage can be calculated from the distributions extracted from devices with identical numbers of dopants in the SSR, and closely matches the brute force simulated distribution. This approach can be statistically enhanced in order to reduce

the computational cost of accurately characterising random dopant induced V_T fluctuations and the statistically enhanced methodology is shown to be robust. The effect of the selection of the parameters involved on the accuracy of the approach is also investigated.

In the next chapter, we continue the investigation into V_T variability by studying the influence of line edge roughness. While random dopants currently dominate threshold voltage variability, line edge roughness is also a major contributor to the total variability, and may overtake random dopants as the dominant source in future devices [51].

Chapter 5

Line Edge Roughness

The detailed distribution of V_T due to random discrete dopants (RDD), which are the primary source of statistical variability in current conventional (bulk) CMOS transistors [48, 47] has been analysed in the previous chapter. The contribution of line edge roughness (LER) to statistical variability, however, is becoming more important due to the fact that LER scaling currently lags the requirements of the ITRS [3]. As described in Section 2.1.2, LER occurs due to the molecular nature of the photoresists used during device fabrication and results in non-uniformity of the gate edges. As a result, due to local short channel effects, particular sections of the gate may start to conduct before others, leading to variation in the threshold voltage of the device. While new device architectures such as Silicon-on-Insulator (SOI) and FinFETs tolerate low channel doping, which reduces RDD variability, they are highly susceptible to the effects of LER. Recent simulation results predict that LER induced variability could overtake that induced by RDD in bulk; SOI and double gate devices, particularly at high drain voltages [49, 51].

To study the effects of LER, we have simulated very large device ensembles with unique patterns of LER in each device at both high and low drain bias conditions, for the same 35 nm bulk MOSFET for which random dopant effects were studied in Chapter 4. Smaller ensembles of a selection of other devices have been simulated in order to confirm the trends observed in the simulation of the 35 nm device and in order to assess the impact of LER induced vari-

Device	$\mathbf{V_D}$ **(mV)**	# **Simulated**
Toshiba 35 nm Bulk	100	25,000
Toshiba 35 nm Bulk	800	10,000
45 nm Bulk	100	1,000
32 nm SOI	100	1,000
22 nm DG	100	1,280

Table 5.1: Summary of the devices and associated drain voltages simulated to study LER induced V_T variability.

ability on new device architectures. The devices simulated in this Chapter are summarised in Table 5.1.

In order to introduce LER into the simulations, we utilise a method based on 1D Fourier synthesis which generates lines using a Gaussian power spectrum (see Section 3.1.2 for details). The corresponding autocorrelation function is characterised by two parameters – the RMS amplitude (Δ) and the correlation length (Λ). Note that quoted values for LER magnitude usually correspond to 3Δ. Values for LER magnitude have remained in the range of $4-5\,nm$ due to the fact that LER is not scaling as predicted by the ITRS roadmap [3]. In our study we have adopted a slightly pessimistic value of $3\Delta = 5\,nm$. The correlation length is a measure of the distance over which the fluctuations are correlated, which is determined by fitting the autocorrelation function (in this case a Gaussian autocorrelation function) to measured gate edge patterns [150]. Thus, in these simulations, values of $\Delta = 1.6667\,nm$ and $\Lambda = 30\,nm$ have been used to generate random source/drain and gate edges introduced by the roughness of the resist during fabrication. In this work, we assume that the PN junctions follow the same pattern as the gate edge. This may not be the case when the implantation angle of the doping process is very shallow [90]. Smearing of the PN junction LER can also occur as a result of thermal annealing during doping activation.

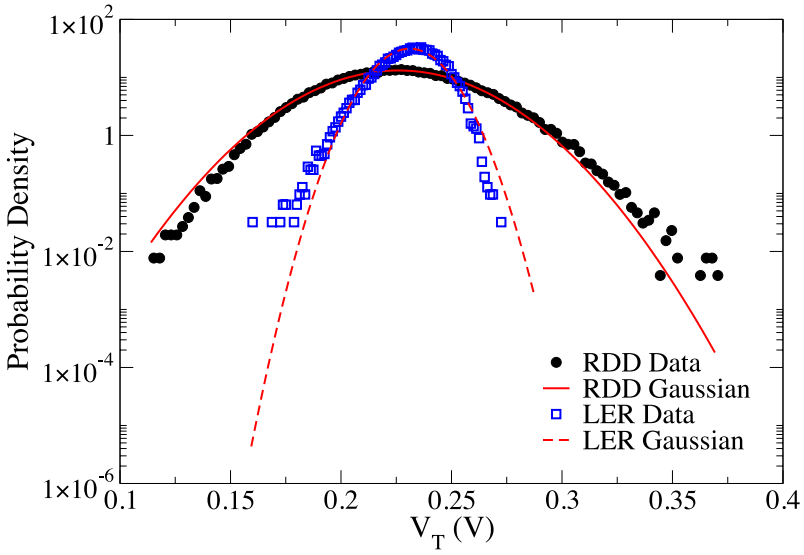


Figure 5.1: Comparison of the histograms of V_T obtained from RDD only and LER only simulations at $V_D = 100\,mV$.

5.1 35 nm MOSFET Results

In order to quantify the relative contributions of RDD and LER to the overall variation in V_T, the distributions of V_T introduced by RDD and LER at low drain ($V_D = 100\,mV$) are compared in Figure 5.1. It is clear that while RDD is the dominant source of statistical variability in this device, LER introduces statistical variability of a comparable magnitude, with the standard deviation being $\sim 40\%$ of the RDD-induced standard deviation. There is also a small shift in the central tendency between the distributions, with $\overline{V_T}$ being $5\,mV$ higher for LER than for RDD. The reduction in $\overline{V_T}$ due to RDD occurs due to the lowering of the potential barrier in the regions between dopants, leading to current percolation and early turn on in some parts of the device. Since the carrier concentration is exponentially dependent on the potential, the local potential lowering in the regions without dopants results in a strong increase in the carrier concentration and the current density in such percolation paths and in an overall reduction of $\overline{V_T}$ [57].

Comparison of the extracted distributions with reference Gaussian distributions having mean (μ) and standard deviation (σ) values calculated from

Statistic	$V_D = 100mV$	$V_D = 800mV$	RDD
Minimum (mV)	159.4±5.2	65.54±8.6	112.7±1.5
Maximum (mV)	271.9±2.0	234.6±3.7	370.5±2.3
Mean (mV)	231.1±0.1	174.3±0.2	225.9±0.1
St. Dev. (mV)	12.75±0.06	19.18±0.15	30.28±0.07
Skew	-0.407±0.02	-0.431±0.03	0.159±0.008
Kurtosis	0.255±0.06	0.363±0.12	0.0486±0.02

Table 5.2: Summary of the statistical moments and the standard errors of the data for LER simulations at $V_D = 100\,mV$ and $V_D = 800\,mV$.

the data clearly shows that the LER induced distribution of V_T is skewed in the opposite direction when compared to the RDD induced distribution. This is consistent with the descriptive statistics extracted from the distribution, presented in Table 5.2. In particular, we note that the skew is significantly greater for variation induced by LER than for that due to RDD. The negative skew observed in the LER case is attributable to the lowering of the threshold voltage due to early turn-on in devices with narrow paths across the channel. Similar negative skew has been observed experimentally and attributed to LER by *Miyamura et al.* [151].

The distributions of V_T at low and high drain voltages are shown in Figures 5.2(a) and (b). The two distributions are plotted using a semi-logarithmic scale and indicate that both threshold voltage variation and threshold voltage lowering increase with drain voltage. This can be attributed to the influence of drain induced barrier lowering (DIBL). In this case, the increase in drain voltage results in a negative shift of $57\,mV$ ($\sim 25\%$) of the mean threshold voltage and an increase in the standard deviation by $6.4\,mV$ ($\sim 50\%$). Gaussians with mean and standard deviation extracted from the data are shown in both plots to illustrate the departure from the common assumption that V_T follows a Gaussian distribution. Table 5.2 provides a detailed summary of the statistical parameters of the distributions. The skew and kurtosis values in the table show that the drain voltage slightly increases the asymmetry. While the asymmetry of the distribution of V_T may be related to early turn-on in the narrowest parts of the channel, further analysis (see Section 5.2) has suggested

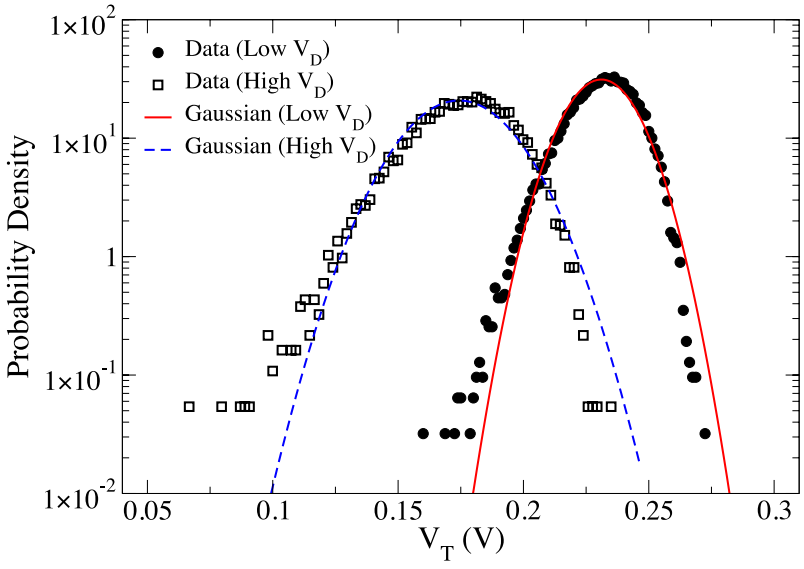


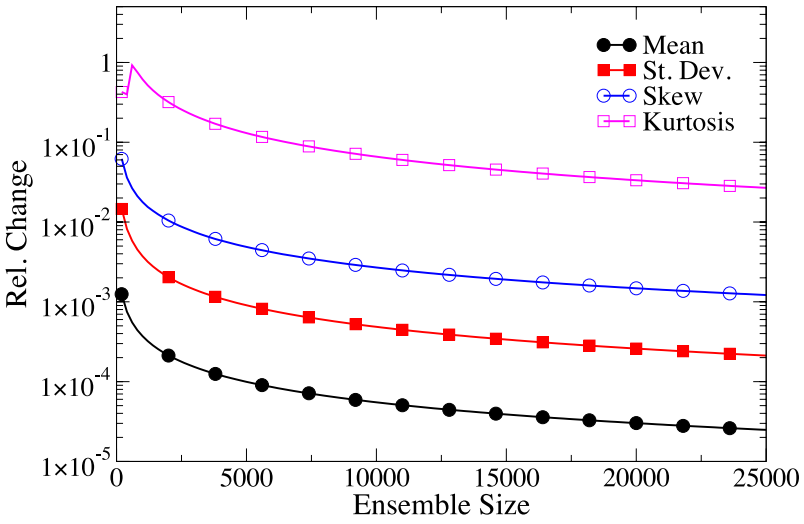
Figure 5.2: Comparison of the histograms of V_T obtained for LER simulations at $V_D = 100\,mV$ and $V_D = 800\,mV$. Gaussian distributions with the data mean and standard deviation are shown for comparison.

that there is a stronger correlation between the threshold voltage and the *average* channel length compared to the correlation with the minimum channel length.

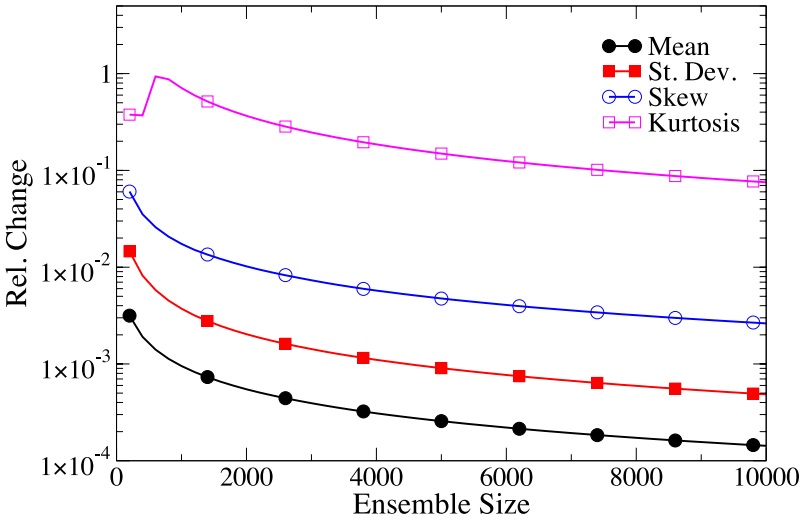
The relative change in the first four statistical moments of the distributions of V_T as a function of sample size is shown in Figure 5.3 to give an indication of the level of statistical accuracy attained by the simulation of such large statistical samples. The convergence of the statistical moments to well defined values can be clearly seen. As with RDD, it is clear that the shape of the distribution cannot be accurately determined from small samples of a few hundreds of devices.

5.2 Statistical Analysis

It is reasonable to expect that the threshold voltage in any particular transistor from the sample will be dominated by the shortest path(s) across the channel since, due to the short channel effects, the dependence between shortest path



(a)



(b)

Figure 5.3: Relative change in the first four statistical moments of the distribution of V_T as a function of sample size for (a) $V_D = 100\,mV$ and (b) $V_D = 800\,mV$.

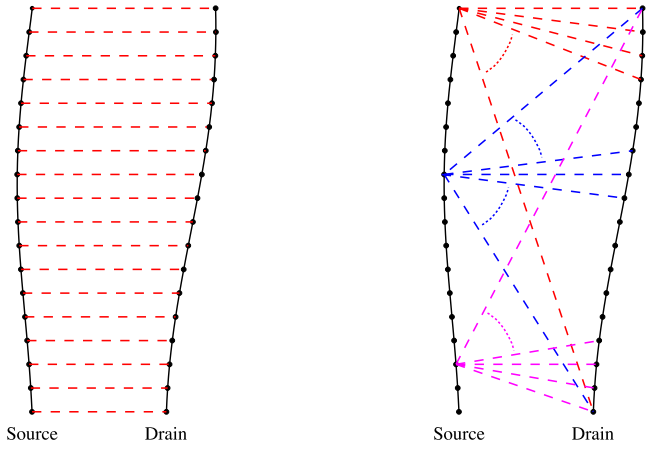


Figure 5.4: Demonstration of how the minimum distance across the channel is computed. In (a) the distance from source to drain is calculated normal to the width direction of the channel at each mesh point. In (b) the distance from each point of the source line to every point of the drain line is calculated.

and current, particularly in the sub-threshold region, is non-linear. In order to examine the correlation between V_T and minimum channel length (L_C) it is necessary to clarify how the minimum value of L_C is extracted. Figure 5.4 illustrates the position of the source/drain metallurgical junctions at the interface for a particular simulated device. First, the distance across the channel at each point along the channel width is calculated, as shown in Figure 5.4(a). Then the distance between each point on the source line and every point on the drain line is calculated, as shown in Figure 5.4(b), in order to check for short paths between extrusions that are offset along the line. We found that since the LER patterns vary relatively slowly there is little difference between the two methods and we consequently use the first method, as it is computationally and conceptually simpler. The average and maximum values for L_C are also computed in this manner.

Figure 5.5 compares the distributions of V_T as a function of the minimum, maximum and average value of L_C. It is visually clear that both the minimum and maximum values of L_C are strongly correlated with the threshold voltage, with calculated correlation coefficients of $\rho_{min\,L_C,V_T} = 0.92$ and $\rho_{max\,L_C,V_T} =$

0.88. However, there is an even stronger correlation between average channel length and V_T yielding a correlation coefficient of $\rho_{avg\,L_C,V_T} = 0.994$. In order to examine the origins of this correlation, devices with uniform gate edges and different channel lengths are simulated. This yields devices with channel length equal to the average at every point along the width of the device. Figure 5.6 compares the channel length dependence of the threshold voltage obtained from 'uniform edge' simulations with the scatter plot of average L_C against V_T at low and high drain voltages. The uniform edge results lie extremely close to the upper boundary of the distributions of V_T, indicating that such simulations can be used to accurately predict the effects of LER on a given device. Several curve fits were also performed to this data in order to allow the prediction of LER effects to large values of σ_{V_T}. The closest agreement was obtained for function of the form $f(x) = \alpha - \beta \exp(-\gamma x)$. This is plotted in Figure 5.6, showing an excellent fit to the simulation data. It should be noted that in a well-designed device, the influence of reverse short channel effects would be evident. The results shown here indicate comparatively poor V_T roll-off behaviour in the simulated 35 nm device.

We also examine the relationship between off current and average value of L_C, shown in Figure 5.7. As expected, a similarly strong correlation exists between the leakage current and average L_C. This is not surprising, given that $\log I_{OFF} = \log I_{V_T} - \frac{V_T}{S}$. The results for uniform edge simulations are also plotted, showing an excellent prediction of I_{OFF} variability for a given channel length.

5.3 Constructing the Distribution of V_T

From the simulation results, we have established that there is a close relationship between threshold voltage and the mean channel length. In this section we analyse the factors that contribute to LER induced variability and investigate how the distribution of V_T can be calculated semi-analytically.

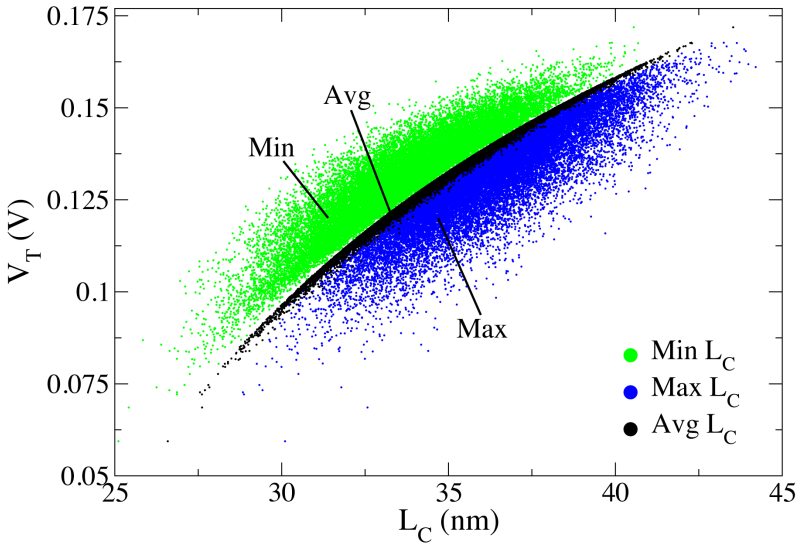


Figure 5.5: Scatterplot of V_T against minimum, maximum and average L_C for each device. An almost direct, although non-linear, relationship between V_T and average L_C can be seen.

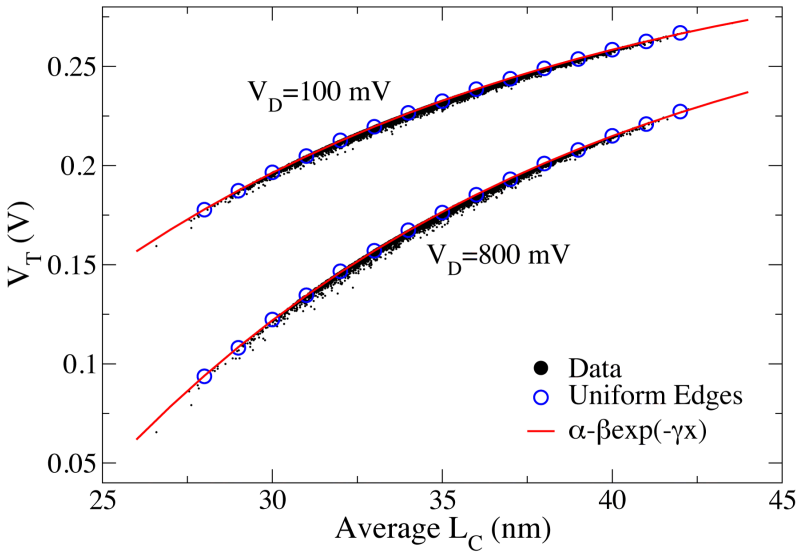


Figure 5.6: Scatterplot of V_T against average L_C for low and high drain voltages. The results of simulations with constant channel lengths are also plotted along with a curve fit of the form $\alpha - \beta\, exp(-\gamma x)$.

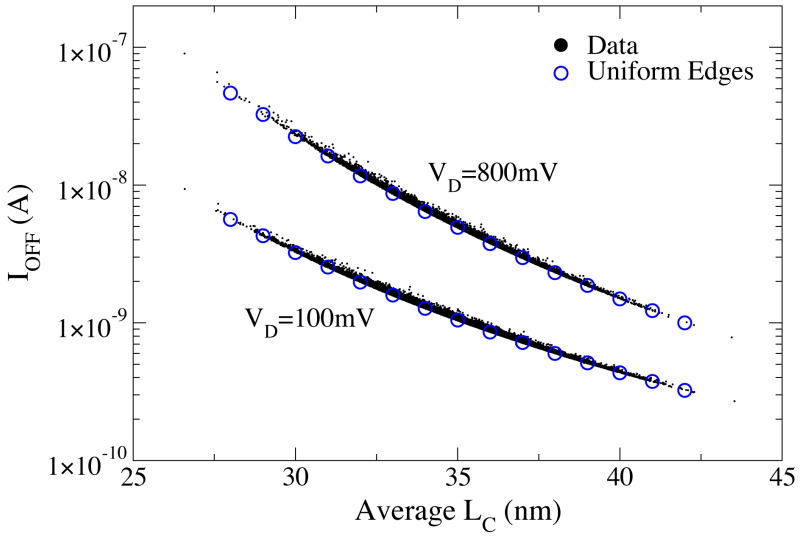


Figure 5.7: Scatterplot of I_{OFF} against average L_C. The results for constant channel lengths are also plotted.

5.3.1 Components of Variation

Based on the extracted relationship between V_T and the mean channel length ($\overline{L_C}$), we investigate how the distribution of threshold voltage can be constructed from a reduced sample. From Figure 5.6 it is clear that although V_T and $\overline{L_C}$ are very highly correlated, there is a small variation in the threshold voltage for a fixed average channel length. This is due to the fact that although different gate shapes can give the same average channel length, the SD effects are not identical and may result in different threshold voltages. Since there are two factors contributing to the V_T variation, it is reasonable to assume that the overall distribution of V_T will be the convolution of the distributions resulting from the two sources, in a similar way to the random dopant induced distribution (see Section 4.3).

To analyse the variation in devices with identical average channel lengths, we extract the distribution of threshold voltage for a small segment of average channel length, e.g. from 35 nm to 36 nm. However, even for a small 1 nm segment, there is still a macroscopic variation in V_T caused by the dependence of V_T on $\overline{L_C}$, which must be separated from V_T variation for a fixed

channel length. To separate these two contributions, we normalise the threshold voltage in a segment using the function originally fitted to the results of the uniform edge simulations [$V_T(\overline{L_C})$] and then determine the distribution of the remaining contribution, as illustrated graphically in Figure 5.8. The 'sub-distributions' extracted using this method can be seen in Figure 5.9. The probability density functions (PDFs) have been estimated using kernel density estimation (KDE) [152] rather than histograms. A KDE is a continuous estimate of the PDF that is useful when the sample size is small, as it has the effect of smoothing the histogram. The PDF is estimated by placing a kernel (usually a Gaussian) at each occurrence in the sample and summing the corresponding probabilities. Examining the 'sub-distributions' shows that they are negatively skewed, and that the standard deviation changes when progressing to larger values of $\overline{L_C}$. While the mean of these distributions can be determined approximately using the uniform edge data, the standard deviation requires closer attention. The dependence of the standard deviation of the sub-distributions as a function of average L_C is shown in Figure 5.10. Linear and decaying exponential curve fits are also shown and with the exponential fit being more favourable. This is related to weaker dependence of the threshold voltage on channel length variations at longer average channel lengths. As a result, at larger channel lengths, different microscopic realisations of a particular average channel length have a smaller impact on the corresponding V_T variations. While no definite trends for the skew and kurtosis could be determined for the sub-distributions, in Section 5.3.2 we show that the shape of these distributions has little influence on the accuracy of the constructed distribution.

Having determined how V_T depends on $\overline{L_C}$, the next step necessary for the re-construction of the distribution of V_T is to determine the distribution of $\overline{L_C}$ itself. The distribution of $\overline{L_C}$ obtained by generating 10^6 random line pairs is shown in Figure 5.11. Visually, the distribution appears Gaussian, which is confirmed by the skew and kurtosis, which are 0.0033 and -0.071 respectively. This was further verified using the Mann-Whitney test [147]. As detailed in Section 4.2, this tests whether two samples are drawn from the same underlying population. This allows the simulation data to be tested by comparing with

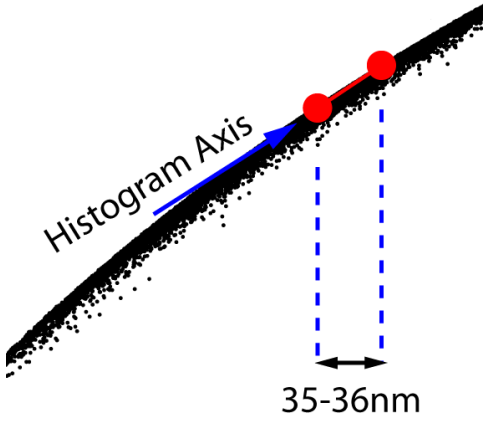


Figure 5.8: Illustration of how the distribution of V_T for a small segment of $\overline{L_C}$ is extracted.

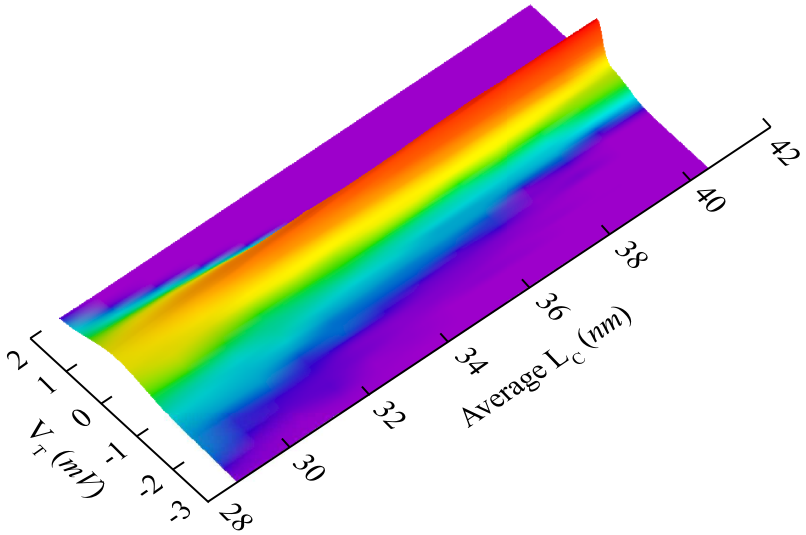


Figure 5.9: The 'sub-distribution' of V_T extracted for successive 1 nm segments of $\overline{L_C}$ in 3D.

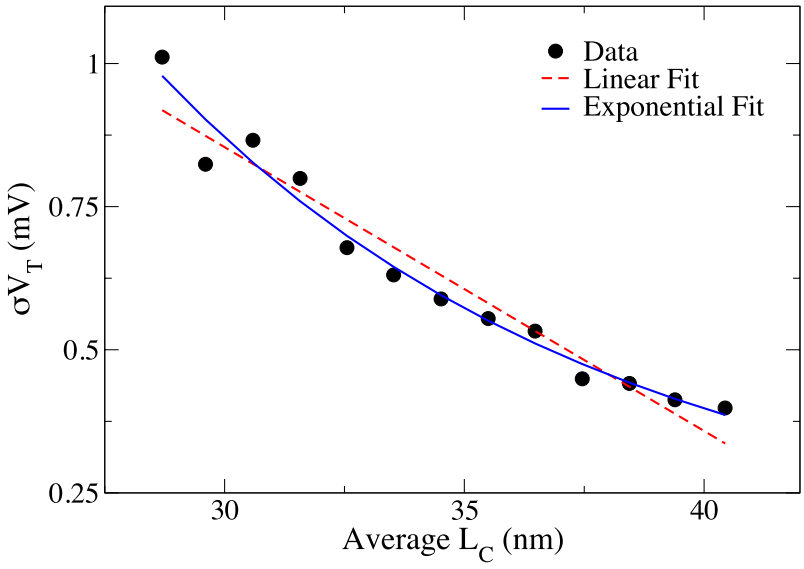


Figure 5.10: Standard deviation of the 'sub-distributions' of V_T. Linear and decaying exponential curve fits are also shown.

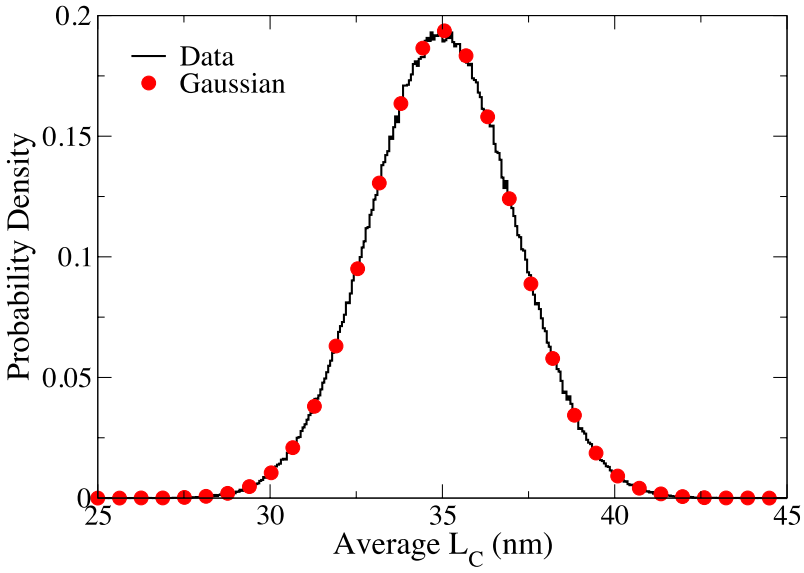


Figure 5.11: Distribution of $\overline{L_C}$ in the simulated 35 nm device. 10^6 random line pairs were generated and the average channel length calculated as shown in Figure 5.4(a). The distribution is compared to a Gaussian with the same mean and standard deviation as the data.

random deviates drawn from a true Gaussian distribution. The data for $\overline{L_C}$ was tested against 1,000 randomly generated Gaussians, with an average p-value of 0.635 and no p-values below the standard statistical significance level of $\alpha = 0.05$ were produced. We therefore conclude that the distribution is indeed Gaussian, with $\mu = 35\,nm$ and $\sigma = 2.059\,nm$.

5.3.2 Convolution Method

Knowing how $\overline{L_C}$ varies, it is possible to construct the distribution of threshold voltage by convolving the distribution of $\overline{L_C}$ with the sub-distributions extracted for small segments of mean channel length. Since we have not quantified the higher moments of the sub-distributions, and in order to simplify this procedure, we have approximated the sub-distributions for a given value of L_C using Gaussian distributions. The approach parallels the method previously used to construct the distribution of threshold voltage due to random dopants, and is expressed mathematically in Equation 5.1, which can be used to determine the probability for a particular threshold voltage.

$$P(V_T) = \sum_{\overline{L_C}=0}^{\infty} G(\overline{L_C}, \mu_{\overline{L_C}}, \sigma_{\overline{L_C}}) \cdot G(V_T, \mu_{V_T}(\overline{L_C}), \sigma_{V_T}(\overline{L_C})) \tag{5.1}$$

Where G is a Gaussian distribution, $\mu_{\overline{L_C}}$ is the mean of $\overline{L_C}$, $\sigma_{\overline{L_C}}$ is the standard deviation of $\overline{L_C}$, the curve fitted to the uniform edge simulations (i.e. $V_T(\overline{L_C})$) is used as $\mu_{V_T}(\overline{L_C})$ and the decaying exponential curve fitted to the standard deviation of the extracted sub-distributions (see Figure 5.10) is used as $\sigma_{V_T}(\overline{L_C})$. The numerical evaluation of Equation 5.1 yields an excellent fit to the simulation data with a χ^2 error of 0.461. It should be noted that a step size of 0.1 nm or smaller is typically required for the numerical summation in order to achieve an accurate and smooth distribution. In order to reduce the computational complexity of evaluating the sum, the upper limit of the summation can also be reduced to a practical finite number, e.g. 100 nm.

While an exponential fit for $\sigma_{V_T}(\overline{L_C})$ provides a good match to the data, there is still a degree of uncertainty over the actual form of this function. We should therefore determine to what extent the choice of $\sigma_{V_T}(\overline{L_C})$ affects the

constructed distribution. In order to analyse the impact, we set $\sigma_{V_T}(\overline{L_C})$ to a constant value chosen to be approximately the standard deviation ($\sim 0.5\,mV$) at the nominal channel length for the device (35 nm). Calculating the distribution for this value results in a χ^2 error of 0.459, indicating that changes in $\sigma_{V_T}(\overline{L_C})$ have a relatively small impact on the final distribution of V_T. Having verified that using a constant value for $\sigma_{V_T}(\overline{L_C})$ does not significantly degrade the accuracy of the calculated distribution, we can also analyse the impact of the magnitude of the constant chosen to represent $\sigma_{V_T}(\overline{L_C})$.

The distribution was calculated for values of $\sigma_{V_T}(\overline{L_C})$ from $0.01\,mV - 10\,mV$ and the resulting errors can be seen in Figure 5.12. Obviously, choosing unrealistically large values for σ results in an extremely poor fit to the data, as would be expected, since large values of σ will dominate over the dependence of V_T on $\overline{L_C}$. To avoid errors, the value of σ should be small and in fact, it could be eliminated entirely by assuming a delta distribution. This is based on the observation that the error of the calculated distribution becomes constant for values of σ below $\sim 0.5\,mV$, and the assumption that this remains true in the limiting case where $\sigma \to 0$. In this limit, a Gaussian distribution becomes a Dirac delta function, located at $x = \mu$:

$$\lim_{\sigma \to 0} G(x, \mu, \sigma) = \delta(x - \mu) \tag{5.2}$$

The Gaussian sub-distributions are thus replaced with Dirac delta functions. Since the sub-distributions have no width, as such, it is also necessary to replace the discrete sum over $\overline{L_C}$ with a continuous integral. The probability of obtaining a particular threshold voltage is then expressed by Equation 5.3.

$$P(V_T) = \int_0^\infty G(\overline{L_C}, \mu_{\overline{L_C}}, \sigma_{\overline{L_C}}) \cdot \delta(V_T - \mu_{V_T}(\overline{L_C}))\, dL_C \tag{5.3}$$

A distribution calculated using the above equation yields a χ^2 error of 0.464, which is very close to the value obtained using the previous approaches and is consistent with the minimum value of the error observed in Figure 5.12. This reduces the complexity of estimating the distribution, as it is not necessary to quantify the variation in V_T for a fixed value of $\overline{L_C}$. As with the previous

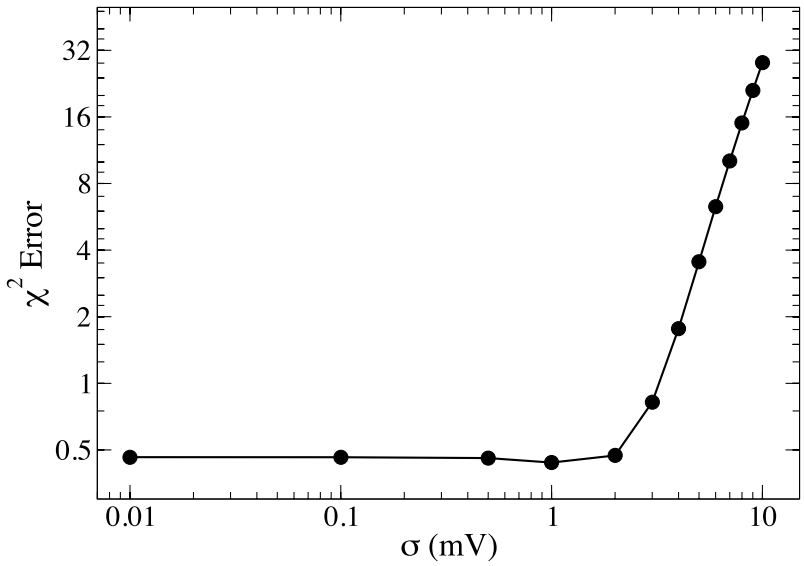


Figure 5.12: χ^2 error of the calculated distribution as a function of the value of σ used for the sub-distributions.

approach, the upper limit of the integral may be reduced in practice to a realistic maximum channel length.

By removing the need to estimate $\sigma_{V_T}(\overline{L_C})$ and by using a $V_T(\overline{L_C})$ dependence fitted to the simulations of devices with uniform edge, we reduce the computational effort necessary to accurately calculate the distribution of V_T to a small number of uniform simulations.

5.3.3 Transformation of Variable Method

In the previous section we have shown that the variation in V_T for a fixed value of $\overline{L_C}$ has little effect on the overall distribution of threshold voltage. The distribution of V_T is thus obtained by convolving the Gaussian distribution of $\overline{L_C}$ with a series of delta functions that are located according to the function $V_T(\overline{L_C})$. By convolving the distribution of $\overline{L_C}$ with delta functions, we are essentially using $V_T(\overline{L_C})$ to non-uniformly sample the distribution of $\overline{L_C}$. $V_T(\overline{L_C})$ can thus be seen as a function that provides a mapping between $\overline{L_C}$ and V_T. Using the following identity [153], this function can be used to transform one random variable (the average channel length) into another random variable

(threshold voltage):

$$P_Y(y) = \left[P_X(x) \left| \frac{dx}{dy} \right| \right]_{x=f^{-1}(y)} \tag{5.4}$$

where X and Y are random variables with probability density functions P_X and P_Y; and f is a function that relates the two variables, i.e. $Y = f(X)$. In this particular instance, we are transforming the random variable $\overline{L_C}$ into the random variable V_T using the fitted relationship $\alpha - \beta\, exp(-\gamma\overline{L_C})$ as the mapping function. This allows the distribution of V_T to be defined as follows:

$$P_{V_T}(V_T) = \frac{1}{\gamma(\alpha - V_T)} P_{\overline{L_C}}(\overline{L_C}(V_T)) \tag{5.5}$$

where $P_{\overline{L_C}}$ is the Gaussian distribution of the average channel length. Alternatively, this can be defined in terms of the CDF, which removes the need for the scaling factor, as the CDF is intrinsically normalized to 1:

$$P_{V_T}(V_T) = \frac{\partial}{\partial V_T} F_{\overline{L_C}}(\overline{L_C}(V_T)) \tag{5.6}$$

where $F_{\overline{L_C}}$ is the CDF of the distribution of $\overline{L_C}$. Using this method for calculating the distribution of V_T for the 35 nm device yields a χ^2 error of 0.46, which is consistent with the previous method. This method is preferable to the convolution method described in the previous section, as it avoids the computation of potentially complicated integrals and allows the utilisation of relatively complex relationships as the mapping function. It should also be noted that for the transformation of variable detailed above, the mapping function must be invertible. However, the transformation can be generalized to mapping functions that are not invertible, provided that there are a finite number of roots for each value of y in Equation 5.4.

Returning to the analysis of the final distribution, the transformation can be understood in terms of the density and mapping functions and it can be seen that the skew and kurtosis of the distribution of V_T are due to the non-linearity of the mapping function. This has the net result that, for a given 1D mesh of equally spaced points in $\overline{L_C}$, the corresponding points in V_T will have a non-uniform spacing. Where the rate of change of the mapping function is

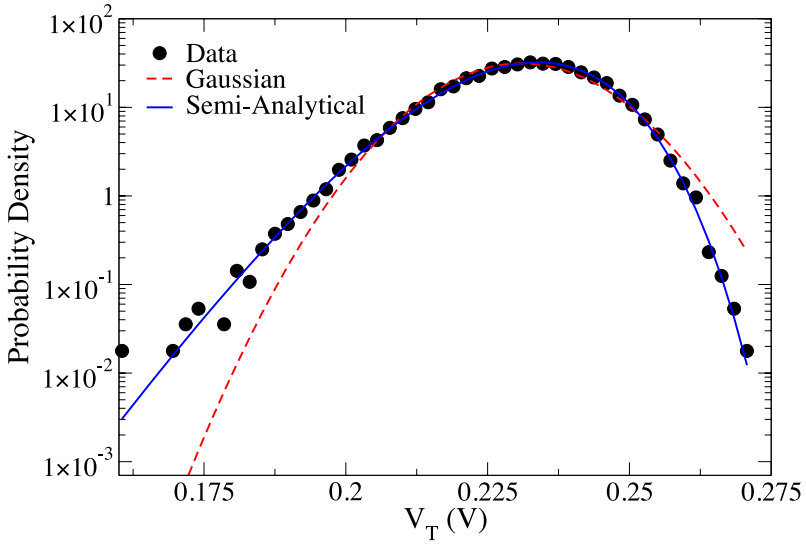
high, the spacing in V_T will be larger, introducing skew in the distribution. It is clear that in our case, due to the particular shape of the mapping function, the distribution will be skewed towards lower values of V_T.

5.3.4 Results

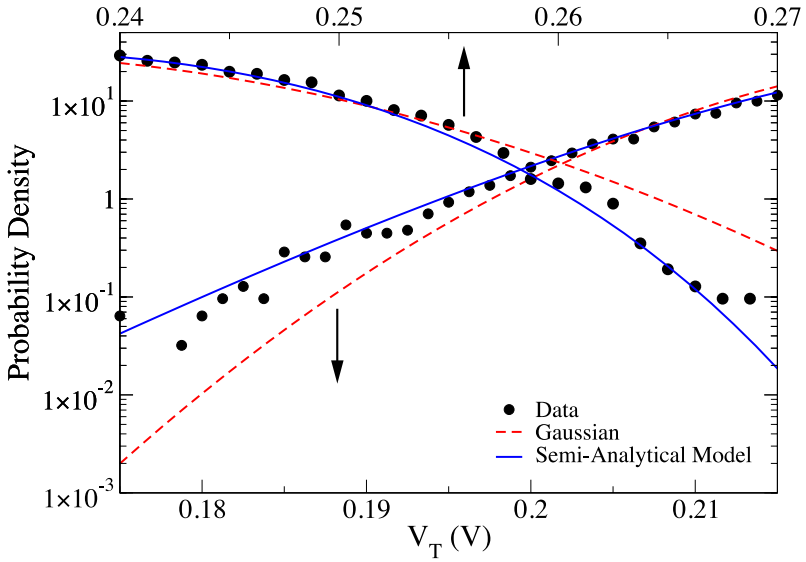
In this section we present the distributions of V_T calculated using the semi-analytical method from the previous section and compare them in detail with the simulation results. Here, the distributions are calculated using the transformation of random variable method.

The distribution of V_T at low drain voltage ($V_D = 100\,mV$) is calculated by transforming the distribution of $\overline{L_C}$, as detailed in Section 5.3.3, is shown in Figure 5.13. The entire range of the distribution is shown in Figure 5.13(a) and a magnified view of the tails is presented in Figure 5.13(b). It is clear that the distribution of V_T is considerably more skewed than that due to RDD and that a Gaussian distribution provides an extremely poor representation of the statistical variation of the threshold voltage due to LER. As a consequence, a circuit designer working under the assumption that the distribution is symmetric will introduce a significant error in their estimation of the number of devices with low and high V_T values, which will impact circuit performance and reliability. In this case, assuming a Gaussian distribution of V_T with the same mean and standard deviation as the simulated distribution results in the number of devices with $V_T < \mu - 3\sigma$ being underestimated by a factor of ~ 4 and the number of devices with $V_T > \mu + 3\sigma$ overestimated by a factor of ~ 29.

Although there is considerable noise in the tails of the simulated distribution, it is clear that the semi-analytical model presented here provides a much more accurate prediction of the variation. QQ plots of the simulation data against a Gaussian distribution are shown in Figure 5.14(a) and against the semi-analytical distribution in Figure 5.14(b). This indicates that the distribution obtained semi-analytically very closely matches both the shape and location of the underlying simulated distribution. It is also apparent in the Gaussian QQ plot that the data diverges quickly from the Gaussian and that it is a poor approximation even at relatively small values of σ.

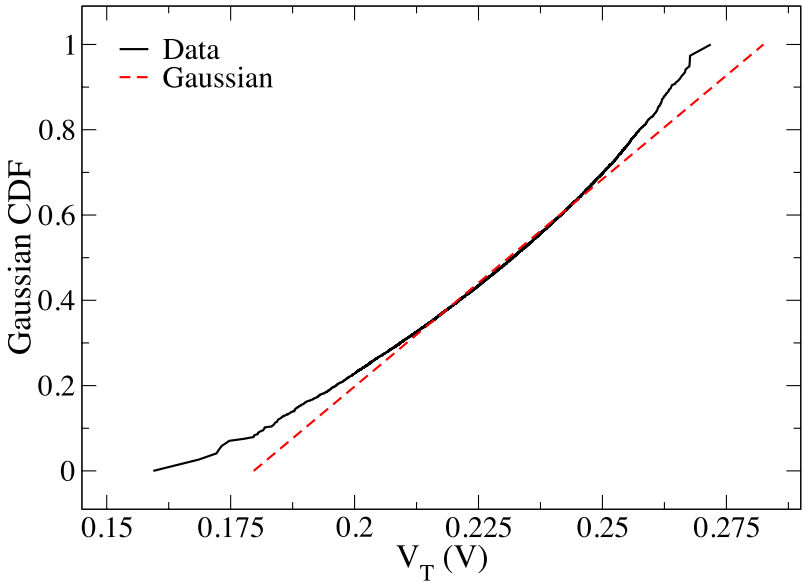


(a)

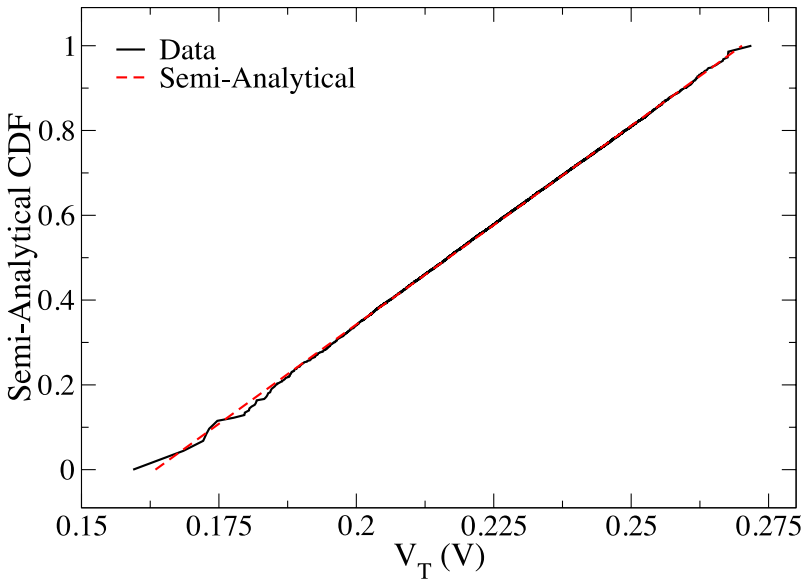


(b)

Figure 5.13: Comparison between the distribution of V_T due to LER obtained from simulation and those calculated using the semi-analytical method. (a) shows the entire distribution and (b) shows a magnified section of the tails. Note that in (b) the two tails are overlaid. Gaussian distributions shown for reference. The semi-analytical distribution is calculated using the method described in Section 5.3.3 and gives excellent agreement with the simulation data over the entire range of values.



(a)



(b)

Figure 5.14: Quantile-Quantile plots comparing the LER simulation data for 35 nm with (a) a Gaussian distribution and (b) the semi-analytical distribution.

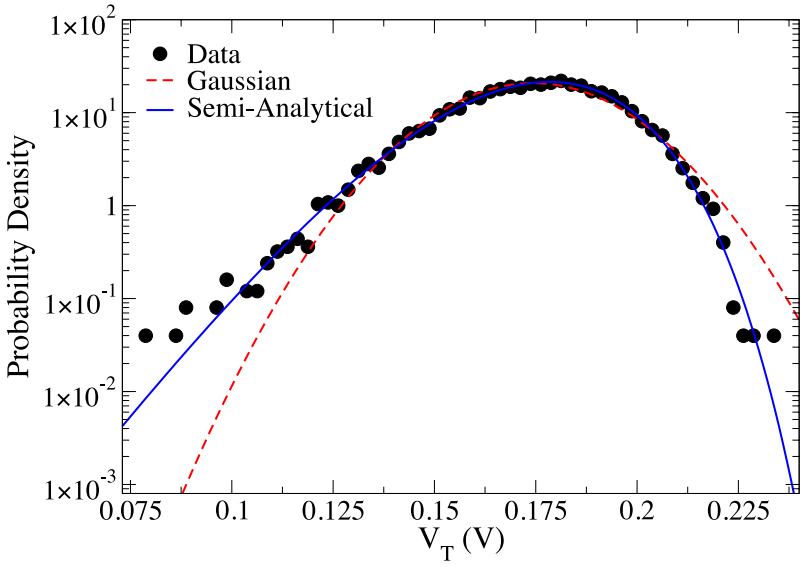


Figure 5.15: Comparison between the simulated V_T distribution due to LER and the semi-analytical distribution at high drain ($V_D = 800\,mV$), with a Gaussian distributions shown for reference.

The distribution of V_T at high drain voltage can also be calculated and produces a similarly good match with the simulation data, as can be seen in Figure 5.15. The distribution of $\overline{L_C}$ does not change with drain voltage, and it is only the mapping function $V_T(\overline{L_C})$ that changes due to the effects of DIBL. The high drain mapping function can be seen in Figure 5.6 and we note that as well as the downward shift that represents the threshold voltage lowering, the bending of the curve increases, which results in the extension of the lower tail of the distribution of threshold voltage and consequently an increase in the skew.

5.4 Width Dependence

In this section the effect of device width on LER induced V_T variation is investigated. While V_T variability due to random dopants generally scales with width as $\frac{1}{\sqrt{w}}$ [62], less is known about the width scaling of variation due to LER. In order to investigate this, samples of 1,000 devices were simulated for

Width	$\mathbf{w_0}$	$\mathbf{2w_0}$	$\mathbf{3w_0}$	$\mathbf{4w_0}$
# of Sims	25,000	1,000	1,000	1,000
Min (mV)	159.4±5.2	184.7±5.6	194.1±2.1	200.8±1.6
Max (mV)	271.9±2.0	256.0±0.5	253.8±0.8	252.5±1.4
Mean (mV)	231.1±0.1	231.1±0.3	230.8±0.3	230.7±0.2
St. Dev. (mV)	12.75±0.06	10.54±0.24	8.97±0.23	7.76±0.18
$\frac{\mathbf{St.Dev.}}{\sqrt{\mathbf{w}}}$**(mV)**	12.75	9.02	7.36	6.38
Skew	-0.407±0.02	-0.387±0.09	-0.363±0.10	-0.302±0.08
Kurtosis	0.255±0.06	0.0392±0.29	0.317±0.28	0.167±0.20

Table 5.3: Summary of the descriptive statistics and standard errors of the distribution of V_T for devices with widths 1-4. All results are for $V_D = 100\,mV$.

devices with channel widths 2, 3 and 4 times the minimal channel width (w_0) of 35 nm at a low drain voltage of $100\,mV$. The moments of the simulated distributions are given in Table 5.3. Values are also shown for the standard deviation scaled by $\sqrt{w}$ and from these results, we see that the standard deviation decreases with width more slowly than $\frac{1}{\sqrt{w}}$. While the error in the skew and kurtosis will be relatively high for the small samples under investigation here, the effect of channel width on the skew can be qualitatively assessed, and shows a decrease in the skew as the width increases. With regard to the kurtosis however, there is too much statistical noise to draw any conclusions on its behaviour with width from the simulation data.

In order to calculate the distribution of V_T semi-analytically, we require the distribution of the average channel length, and the mapping function $V_T(\overline{L_C})$. While Figure 5.6 indicates that the mapping function changes with drain bias, the threshold current scales with width and the shape of this function should not change. This is confirmed by extracting it for all four device widths (Figure 5.16). Since the mapping function does not change, the only factor remaining that can affect the threshold voltage variation is the distribution of $\overline{L_C}$. This distribution becomes narrower as the width increases, due to greater statistical averaging of the LER patterns. In the first instance, this leads to a narrower distribution of V_T, a prediction that is consistent with the values obtained from simulation, as demonstrated in the simulated and calculated distributions of V_T for the four different width devices, shown in Figure 5.17.

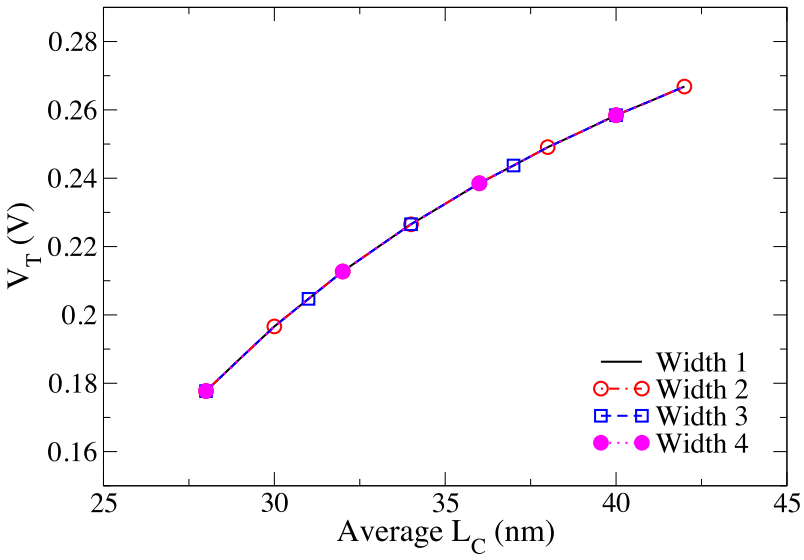


Figure 5.16: Comparison of the relationship between $\overline{L_C}$ and V_T for devices with widths 1-4, at $V_D = 100\,mV$.

The calculated distributions again provide excellent matches to the simulation data, further demonstrating the robustness of this method.

More subtly, the narrowing of the distribution of $\overline{L_C}$ also affects the skew of the distribution of V_T. As discussed already, the skew in the distribution of V_T is due to the non-uniform mapping of the random variable $\overline{L_C}$ into V_T. Given the same mapping function and the fact that the rate of change is higher at low values of $\overline{L_C}$, then a narrower distribution of $\overline{L_C}$ will cover less of the axis described by the mapping function. In particular, there will be fewer or no occurrences of the very short average channel lengths, which will lead to the distribution of V_T being less skewed. This is also consistent with the simulated values.

Since the distribution of V_T can be defined as a function of another random variable ($\overline{L_C}$), the width dependence of V_T variation can also be studied analytically. Given a Gaussian distribution G and a mapping function f, the moments about the mean of the transformed distribution can be expressed as

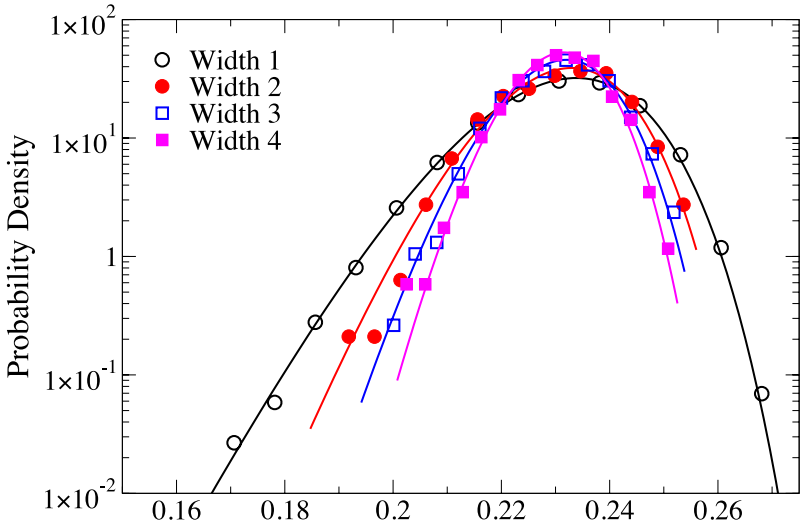


Figure 5.17: Comparison of the simulated and calculated distributions for devices with widths 1-4. Note that the width 1 distribution covers a much larger range of V_T as there are 25,000 devices for width 1 compared to 1,000 for widths 2-4. Symbols indicate the simulation data and lines the calculated distribution.

$$\mu_k = \int_{-\infty}^{\infty} (f(x) - \mu_1')^k G(x, \mu_{\overline{L_C}}, \sigma_{\overline{L_C}}(w)) dx \tag{5.7}$$

where k is the central moment number, μ_1' is the first moment about the origin (i.e. the mean) of the transformed distribution, $\mu_{\overline{L_C}}$ is the mean value of $\overline{L_C}$ and $\sigma_{\overline{L_C}}(w)$ is the standard deviation of $\overline{L_C}$ as a function of the transistor width. The standard deviation of V_T is then defined as $\sigma = \sqrt{\mu_2}$, the skew as $\gamma_1 = \frac{\mu_3}{\sigma^3}$ and the kurtosis as $\gamma_2 = \frac{\mu_4}{\sigma^4}$. Using these expressions, we can estimate how the distribution of V_T changes with width. Figure 5.18 shows the width dependence of the moments, obtained by evaluating the distributions of V_T for widths up to 20. By curve fitting, we find that the width dependence of the standard deviation, skew and kurtosis can be approximated using a function of the form $\frac{\alpha}{(x+\gamma)^\beta}$. As the skew and kurtosis decay, the distribution becomes more Gaussian-like, as we would expect from the effect of averaging. The standard deviation also decreases in the same way, consistent with the results

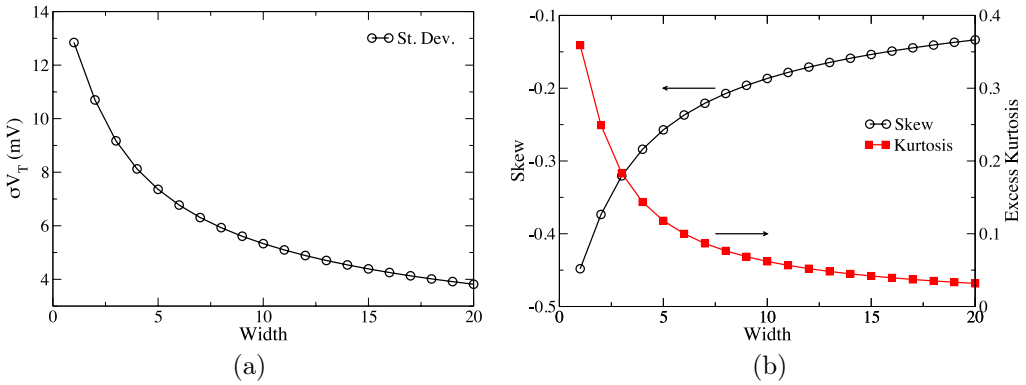


Figure 5.18: Dependence of the standard deviation, skew and excess kurtosis of the distribution of V_T on device channel width, all of which decay towards zero with increasing width.

	35nm Bulk	**45nm Bulk**	**32nm SOI**	**22nm DG**
# of Sims	25,000	1,000	1,000	1,280
Min (mV)	159.4±5.2	187.3±9.7	508.8±1.4	427.7±7.6
Max (mV)	271.9±2.0	351.9±1.0	541.6±0.7	529.3±1.4
Mean (mV)	231.1±0.1	292.2±0.8	528.3±0.2	499.7±0.4
St. Dev. (mV)	12.75±0.06	24.91±0.57	5.25±0.13	13.84±0.36
Skew	-0.407±0.02	-0.385±0.08	-0.512±0.07	-0.962±0.10
Kurtosis	0.255±0.06	0.112±0.25	0.199±0.16	1.44±0.44

Table 5.4: Summary of the statistical moments of the distribution of V_T at low drain in all four devices.

already shown and the expectation that variability decreases in wider devices.

5.5 Impact on Alternative Device Architectures

In order to confirm the trends observed in the simulations of the bulk 35 nm device and to examine the potential impact of LER on different device architectures, smaller ensembles of several other devices have been simulated. The devices include a low power 45 nm bulk MOSFET with an oxide thickness of 1.7 nm developed by ST Microelectronics [36]; a 32 nm ultra thin body SOI MOSFET with a body thickness of 7 nm and equivalent oxide thickness (EOT)

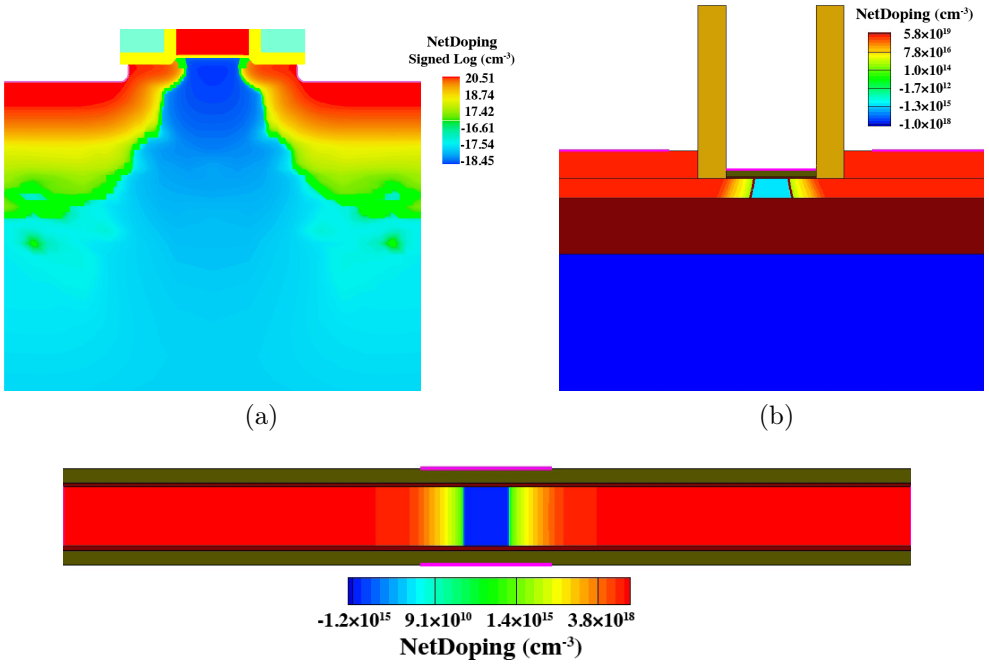


Figure 5.19: Doping profiles of the (a) 45 nm bulk device, (b) 32 nm SOI device and (c) 22 nm double gate device.

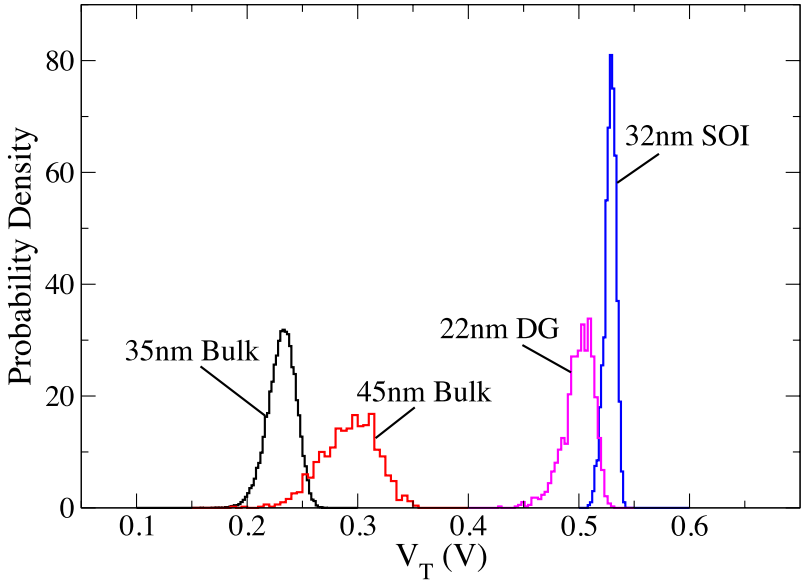


Figure 5.20: Comparison of the distribution of V_T due to LER in the four simulated devices at $V_D = 100\,mV$.

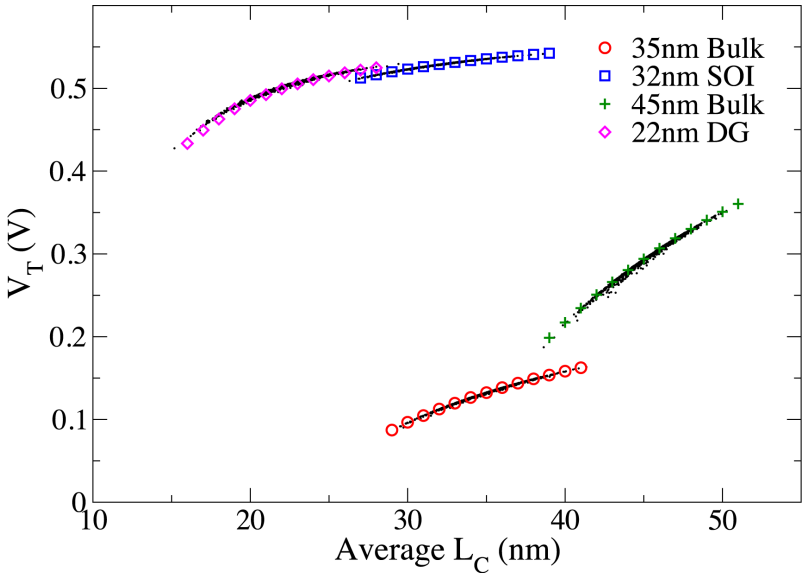


Figure 5.21: Comparison of the relationship between $\overline{L_C}$ and V_T in the four simulated devices at $V_D = 100\,mV$.

of 1.2 nm; and a 22 nm double gate MOSFET with a body thickness of 10 nm and EOT of 1.1 nm. The SOI and double gate devices were developed by the PULLNANO consortium [85]. The doping profiles of the devices are shown in Figure 5.19 and the distributions of threshold voltage obtained at low drain voltage ($V_D = 100\,mV$) are presented in Figure 5.20. For all devices, the shape of the distribution of V_T is similar, with negative skew being present for all four devices. It should be noted that the SOI device in particular exhibits good immunity to LER induced variability, having a standard deviation much lower than the other three devices. This is due to the fact that SOI devices have superior electrostatic integrity, which reduces the short channel effects.

Scatter plots illustrating the relationship between V_T and $\overline{L_C}$ in each device are shown in Figure 5.21. The same general relationship between $\overline{L_C}$ and V_T, as in the case of the 35 nm device, can be seen in all four devices. As this relationship provides the mapping between the two random variables ($\overline{L_C}$ and V_T), the shape of this function will directly affect the final distribution of V_T. The effect of these changes in shape can be seen in the skew and kurtosis values, given in Table 5.4, along with the other moments of the statistical distributions.

The values of the moments also confirm the visual observation that the SOI device has significantly better immunity to LER induced fluctuations. While there would appear to be no improvement in the V_T spread for the double gate device compared to the 35 nm bulk device, it should be noted that in these simulations both gates follow the same LER pattern, and therefore represent a worst case scenario. Modelling the bottom gate with different edges to the top gate would likely result in a reduction of the spread of V_T due to statistical averaging.

Since the simulation results for the three additional devices simulated here are similar to the results for the 35 nm device that we have studied in detail, the transformation of variable method developed for calculating the distribution of V_T should be equally applicable to these devices. The mapping function for all four devices can still be suitably modelled with an exponential function $f(x) = \alpha - \beta \exp(-\gamma x)$ and the distribution of $\overline{L_C}$ extracted from the simulation data. The distributions are computed for the three additional devices and the results are shown in Figure 5.22, where it can be seen that this method produces consistently good results across all of the devices, demonstrating the generality of the approach.

5.6 Summary

In this chapter, the effects of line edge roughness on threshold voltage variability have been studied in detail. The distribution of V_T has been accurately characterized at low and high drain voltages with ensembles of 10,000 or more devices. The results indicate that the V_T variations due to LER are negatively skewed.

We show that the dispersion of V_T for a given channel length does not contribute significantly to the overall V_T variation and that an accurate description of threshold voltage fluctuations can be formulated using only the results from 'uniform edge' simulations. We also demonstrate two semi-analytical methods whereby the distribution of V_T can be estimated. This methodology is shown to accurately reproduce the simulated distributions of V_T.

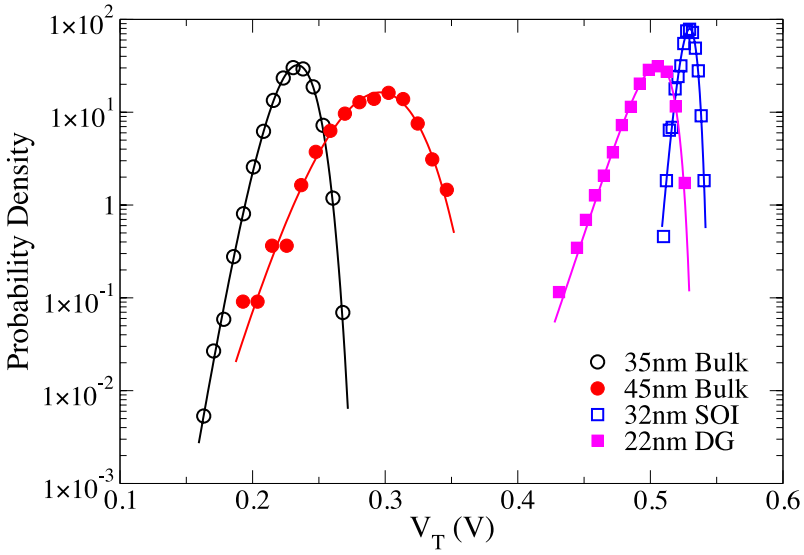


Figure 5.22: Comparison of the calculated and simulated distributions of V_T due to LER in the four simulated devices. Symbols indicate the simulation data and lines the calculated distribution.

The influence of LER in alternative device architectures is also investigated, showing that some of these architectures may present suitable immunity to the effects of LER. The methodology developed for calculating V_T fluctuations is shown to be equally applicable to other device architectures and produces good predictions of the variation in these devices. Finally, we investigate the effect of device width on LER-induced variations. Devices with widths 2-4 times the minimal are simulated and show that although the variation decreases with width, the standard deviation does not scale as $\frac{1}{\sqrt{w}}$. The distributions of V_T for wider devices are also calculated semi-analytically and the excellent match with simulation data further demonstrates the robustness and utility of this methodology.

In the next chapter, we examine the results of combined RDD+LER simulations and investigate how the individual distributions can be combined to match the simulation results.

Chapter 6

Combined Fluctuations

Random discrete dopant (RDD) and line edge roughness (LER) induced threshold voltage variability have been studied individually in Chapters 4 and 5. The detailed simulation study of these individual sources has allowed the development of methodologies for the accurate and efficient prediction of their effects on the threshold voltage. However RDD and LER are present simultaneously in real devices. For this reason, statistical enhancement of their individual simulations is only useful if the corresponding distributions can be reliably combined to produce the resulting distribution in real devices.

In order to examine the combined impact of RDD and LER on transistor variability, we have continued the simulation study of the 35 nm template MOSFET from the previous chapters. Simulations have been performed of 100,000 microscopically distinct devices, in which both RDD and LER are present. This allows the distribution of V_T due to the combined effects to be accurately characterized. We also examine how the overall distribution can be constructed from the distributions obtained from the simulations of the individual sources. We compare the statistical combination of the distributions of the raw RDD and LER simulation data and the combination of the statistically enhanced RDD and LER distributions with the simulation data for the combined variability sources.

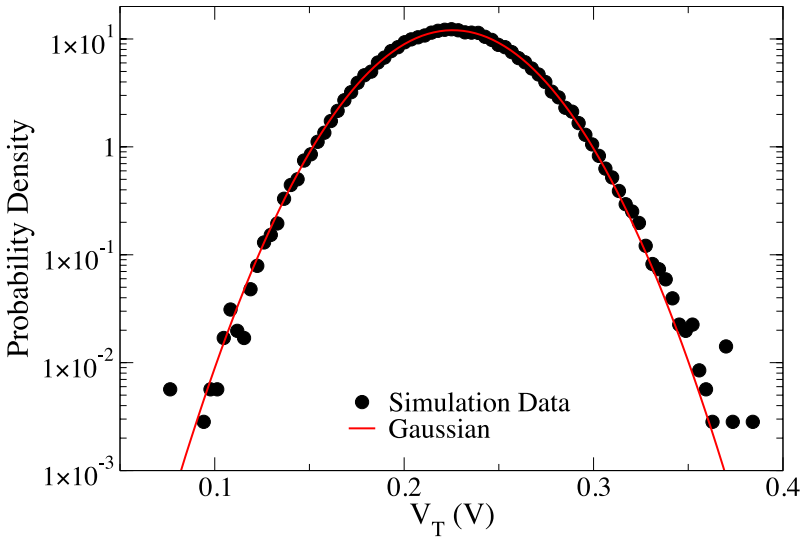


Figure 6.1: Distribution of V_T obtained for simulations of combined RDD and LER fluctuations at low drain ($V_D = 100\,mV$). The distribution is very close to Gaussian, but appears to deviate slightly in the upper tail.

6.1 Statistical Analysis

Simulations of combined fluctuations were performed at low drain bias ($V_D = 100\,mV$) and the semi-logarithmic distribution of threshold voltage is shown in Figure 6.1. Visual inspection of this distribution reveals that the fluctuations in V_T due to the combined effects of RDD and LER is very close to Gaussian. It appears that the opposing skews of the two individual distributions cancel out to a large extent. It is also apparent that there is a small deviation from Gaussian in the upper tail, which can be better observed in the QQ plot of the distribution against a Gaussian in Figure 6.2. The relative changes in the four moments of the distribution are shown in Figure 6.3, which shows, as with our previous simulation studies, that the large statistical ensemble provides a significant reduction in the error associated with the parameterisation of the distribution. Numerical values for the first four moments of the distribution are given in Table 6.1, along with the corresponding values from the individual simulations of RDD and LER induced variability. The skew and kurtosis figures confirm the assumption that the distribution is very close to Gaussian.

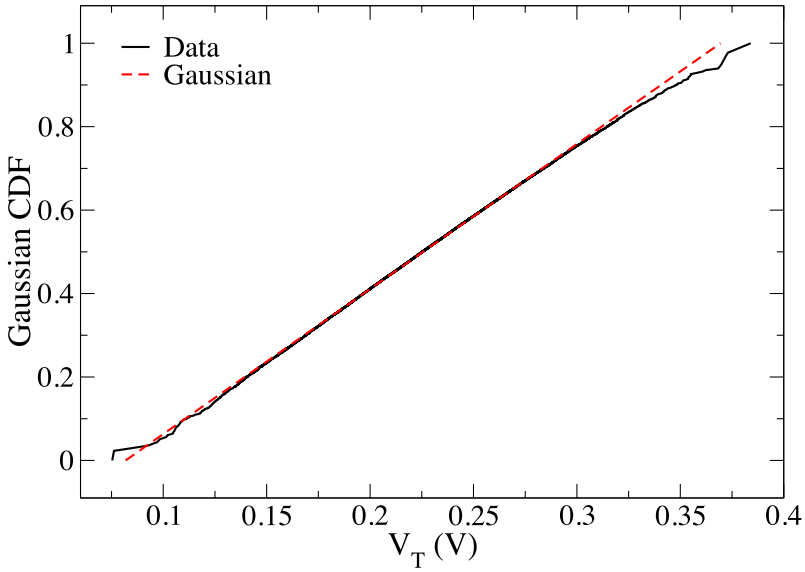


Figure 6.2: QQ plot of the V_T results from the combined RDD+LER simulations against a Gaussian distribution with the data mean and standard deviation. The upper tail deviation is more apparent.

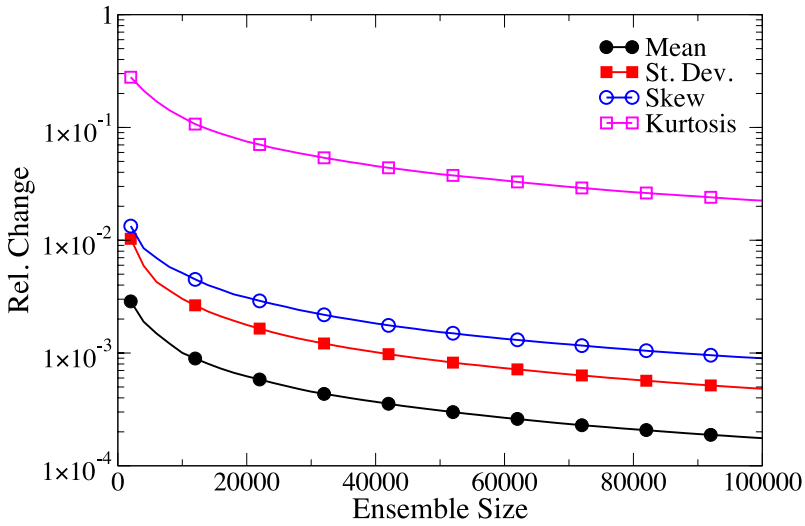


Figure 6.3: Relative change in the first four statistical moments of the distribution of V_T as a function of sample size for combined RDD and LER fluctuations.

Statistic	RDD	LER	RDD+LER
Minimum (mV)	112.7±1.5	159.4±5.2	75.35±6.6
Maximum (mV)	370.5±2.3	271.9±2.0	384.0±5.7
Mean (mV)	225.9±0.1	231.1±0.1	225.6±0.1
St. Dev. (mV)	30.28±0.07	12.75±0.06	33.08±0.07
Skew	0.159±0.008	-0.407±0.02	0.0623±0.008
Kurtosis	0.0486±0.02	0.255±0.06	0.0527±0.02

Table 6.1: Summary of the statistical moments and standard errors of the data for the combined RDD and LER simulations at $V_D = 100\,mV$.

To test further how close the simulated distribution is to a Gaussian, we use the Mann-Whitney test employed in the previous chapters. The simulated distribution was tested against 10,000 randomly generated Gaussians, and yielded a mean p-value of 0.334, with a standard deviation of 0.269. Approximately 12% of the p-values were less than $\alpha = 0.05$, which indicates that the distribution bears some similarity to a Gaussian but this is merely a coincidence in this case. Furthermore, skew is considered significant if the absolute value of the skew is greater than twice the standard error [154], indicating that in this case the skew is indeed significant. The kurtosis may also be considered significant in the same way, however the absolute value is much closer to the error margin and more likely to be a chance fluctuation.

It is therefore clear that in spite of the similarly between the simulated distribution and a Gaussian, the QQ plot of the data and the results of the Mann-Whitney tests indicate that we cannot conclude that the distribution of V_T due to the combined effects of RDD and LER is truly Gaussian in nature. Small positive skew similar to that present in our simulation results has also been observed in measurements performed on a test chip fabricated using a 65 nm SOI process having transistors with a physical gate length of 35 nm [155, 156, 32].

As observed in the previous study of random dopant effects, the asymmetry in the distribution of V_T due to RDD increases with scaling and it is likely that in smaller devices, the distribution due to the combined effects will also exhibit increasing asymmetry. It is also worth noting that the distribu-

tion studied here includes only the effects of random dopants and line edge roughness. While these are among the major sources of variability, polysilicon and metal gate induced variability can also have a strong effect in contemporary MOSFETs [52] and their incorporation will affect the shape of the final distribution.

It is worth noting that although the central limit theorem (CLT), which indicates that the sum of a sufficient number of random variables tends towards a Gaussian distribution, might be expected to apply here, in this case the number of distributions is small – in this work only two. The convergence towards Gaussian would therefore likely be slow were it not for the opposing skews of the two component distributions. As stated above, the skew is likely to increase as devices are scaled further and effects such as gate work function variability have been shown to be strongly non-Gaussian [157]. As a consequence, the CLT is unlikely to be particularly useful in this context. It is however likely that the CLT was more applicable in the past, since the distributions of interest were less asymmetric and the convergence towards Gaussian would be faster.

6.2 Combining RDD and LER Induced Distributions

Having studied random discrete dopants and line edge roughness in detail and developed methods for the statistical enhancement of their simulations in Chapters 4 and 5, we would like to examine how the distributions of V_T that arise from the individual sources can be combined.

By studying RDD and LER in isolation, we have obtained two random variables in V_T, which accurately represent the variation due to the two components. If the two random variables are statistically independent, then they can be straight-forwardly combined by convolving the corresponding density functions. In order to understand to what extent the assumption of statistical independence of RDD and LER is reasonable, we must examine the MOSFET fabrication process. Note that for simplicity, we refer to the fabrication of an

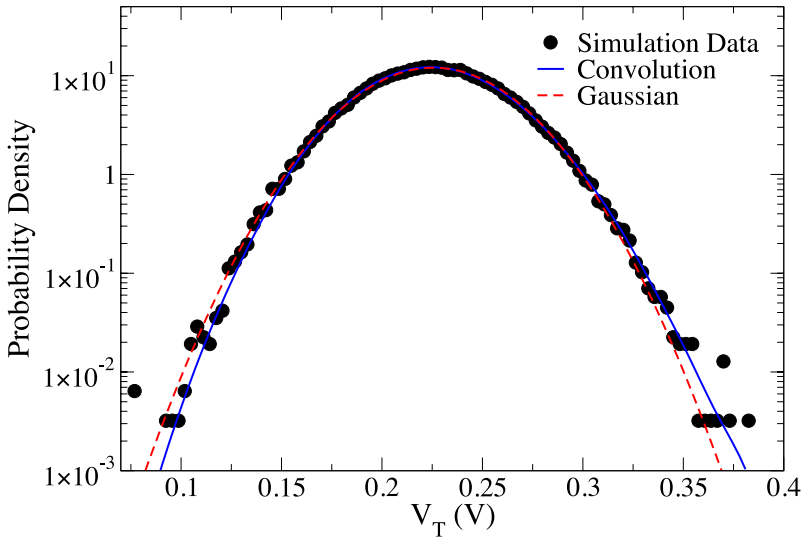
n-Channel MOSFET here.

It was already shown in Chapter 4 that random dopant effects are dominated by the dopants in a relatively small volume under the gate. The doping in this region is primarily determined by the P-well doping level and the P-doping implanted for threshold voltage adjustment. These doping steps occur before the gate formation and thus cannot be influenced by LER effects. The source and drain regions are implanted after patterning of the gate and are self-aligned, thus the LER pattern will be directly transferred to the junction edge. However, discrete donors in the source and drain do not have a significant effect of V_T variability [7]. The implantation of halo doping will also be affected by the LER of the gate, however the halo doping is used to control short channel effects and is generally implanted sufficiently deep in the device that the associated dopants are below the statistically significant region (Section 4.2). Therefore it is reasonable to assume that random discrete dopants and line edge roughness can be considered statistically independent. To test this assumption, we construct the distribution of V_T due to the combined effects of RDD and LER by convolving the individual distributions that arise from RDD and LER.

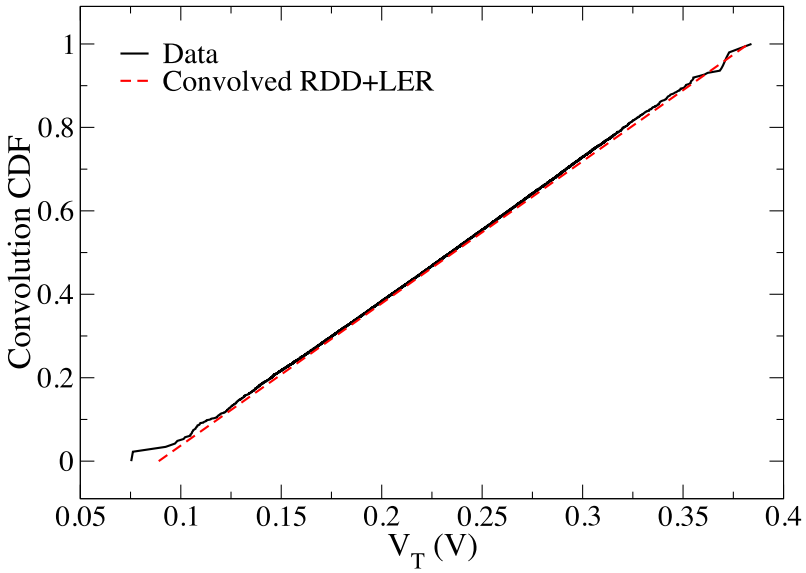
Figures 6.4(a) and (b) show the distribution of V_T constructed by the numerical convolution of the individual distributions obtained from simulations of RDD and LER is isolation. The resulting distribution is compared to the data obtained from simulations with combined RDD and LER. It can be seen that the convolution fits the data significantly better than the Gaussian, confirming that it is a better representation of the simulation data and that the two sources are indeed statistically independent. The QQ plot also indicates that the convolution of the two individual distributions accurately reproduces the shape of the distribution. In particular, there is an improvement in the upper tail, which is not captured well by the Gaussian.

One caveat that should be noted is the determination of the mean of the resulting distribution. The effect of convolving two (or more) density functions together is that the *cumulants*[1] of the distributions add together. Since

[1]Cumulants are quantities that add under convolution and are closely related to the moments of a distribution.



(a)



(b)

Figure 6.4: Comparison of the distributions obtained from simulation and by convolving the individual distributions obtained from simulations of RDD and LER in isolation. (a) Semi-logarithmic histogram and (b) QQ plot.

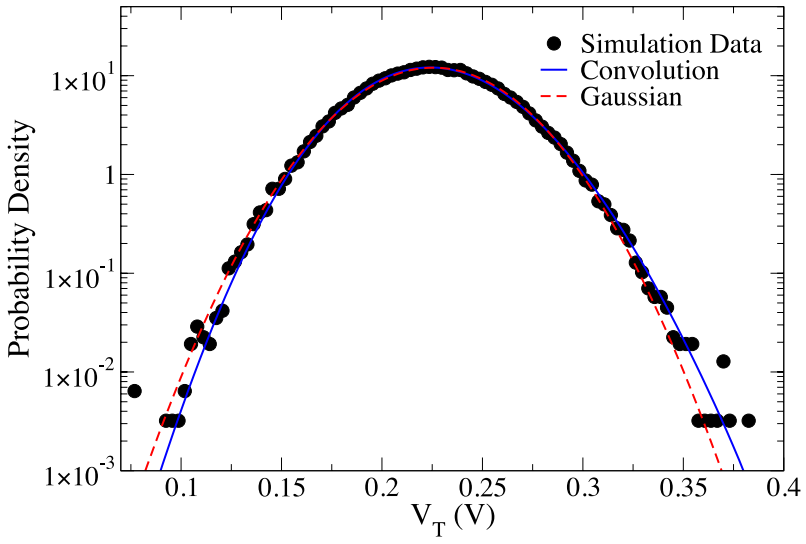
the mean is the first cumulant, ordinarily we should expect that $\mu_{RDD+LER} = \mu_{RDD}+\mu_{LER}$. However, addition of the mean values in this way is obviously unphysical. By examining the individual distributions, we note that μ_{LER} is close to the threshold voltage obtained from simulation of the continuously doped device ($0.232\,mV$), while μ_{RDD}, in contrast, shows a noticeable lowering. This is expected, as RDD simulations are known to have a lower average threshold voltage, compared to continuous doping simulations (see Section 2.1.1). There may be a small influence from LER due to the fact that devices with shorter average channel lengths will suffer from degraded short channel effects, however V_T lowering is dominated by the effects of RDD. For this reason, the convolved distribution is normalised to have the same mean value as the distribution due to RDD.

Although there is a small difference of $\sim 0.3\,mV$ between μ_{RDD} and $\mu_{RDD+LER}$, this is within the limits of the statistical error for the mean. The simulated distributions for RDD and RDD+LER both have a standard error of the mean of approximately $0.1\,mV$ and assuming that the sample mean is normally distributed around the population mean with a standard deviation equal to the standard error, the observed values for μ_{RDD} and $\mu_{RDD+LER}$ fall within 3σ of each other.

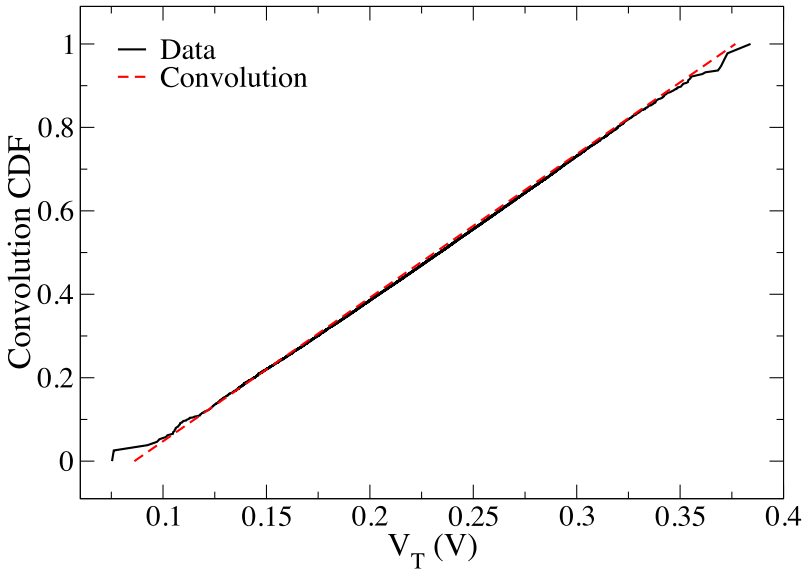
Having confirmed that the distribution due to combined RDD and LER can be accurately reproduced by convolving the distributions obtained from simulation, we examine the convolution of the semi-analytical distributions developed in the previous chapters. This is done in order to verify that their combination is consistent with both the raw simulated data for combined RDD+LER fluctuations and with the convolution of the individual simulated distributions.

The distribution of V_T obtained by convolving the semi-analytical distributions for RDD and LER is shown in Figure 6.5(a) and the corresponding QQ plot is shown in Figure 6.5(b). From both of these, it is clear that the results obtained by combining the statistically enhanced distributions also match the simulation data extremely well.

Since the distribution due to combined RDD and LER can be accurately reproduced using the statistically enhanced approaches detailed in Sections 4.3.1 and 5.3, it is possible to characterise the combined effects of random dopants



(a)



(b)

Figure 6.5: Comparison of the distributions obtained from simulation and by convolving the semi-analytical distributions for RDD and LER. (a) Semi-logarithmic histogram and (b) QQ plot.

and line edge roughness on threshold voltage and benefit from the reduced computational time needed to characterise the two sources individually by applying the developed computationally efficient statistical enhancement strategies. As detailed in Section 4.4, the effects of random dopants can be accurately characterised from the simulation (or potentially, measurement) of just 6,000 devices. Similarly, LER can be characterised by simulating a small number of devices, e.g. 20, with uniform gate edges and different channel lengths. In this way, a complete characterisation of the effects of RDD and LER, both in isolation and in combination, can be achieved at the expense of just $\sim 7,500$ CPU hours of simulation, compared to approximately 300,000 CPU hours to perform the complete low drain brute force characterisation of the 35 nm transistor – a reduction by a factor of ~ 40 times.

6.3 Summary

In this chapter, we investigated the combined effects of random discrete dopants and line edge roughness on threshold voltage variability. The distribution of V_T was accurately characterized by the simulation of a statistical ensemble of 100,000 devices. The results indicate that V_T variations due to the combined impact of RDD and LER are, in this case, close to Gaussian.

We show that the results obtained from the individual simulation of random dopants and line edge roughness can be combined by convolving the two distributions of V_T. The results closely match those obtained from combined simulations, indicating that the two sources of variability are statistically independent. We show also that the individual distributions obtained using the statistical enhancement methodologies proposed for RDD and LER in Chapters 4 and 5 can also be reliably combined to produce accurate predictions of the overall threshold voltage variability.

The excellent predictions of the combined variability yielded by the statistical enhancement methodologies allow a significant reduction in the computational effort necessary to accurately characterise the combined effects of random dopants and line edge roughness. As illustrated by the results shown

in this chapter, an accurate characterisation of combined RDD+LER variability can be obtained from the simulation of a few thousand devices, as opposed to several 100,000s of brute force simulations. Thus, a reduction of 1-2 orders of magnitude in the computational time necessary to accurately estimate the shape and tails of the distribution can be achieved.

Chapter 7

Conclusions and Future Work

The aim of this work was to investigate intrinsic parameter fluctuations in nano-scale MOSFETs in detail through large scale statistical simulations. These fluctuations arise due to the fundamental discrete nature of charge and matter. The operational characteristics of deep sub-micron MOS transistors vary due to the particular microscopic structure of a given device – for example the location of dopant atoms in the channel will vary from device to device. In order to properly account for these variations, it is necessary to incorporate knowledge about them in the design phase of the system. Accurate knowledge about transistor parameter distributions can be obtained from experimental measurements, however this is generally extremely expensive, time consuming and can only be done for a mature technology. Through careful and comprehensive modelling and calibration however, good predictions about statistical transistor behaviour can be made through simulation. The large number of transistors on modern chips demands statistical simulation on a very large scale in order to properly describe the tails of parameter distributions, whereby information about statistically rare devices, which have the greatest impact on circuit functionality, can be obtained.

To allow a detailed characterisation and analysis of statistical threshold voltage variability in MOSFET devices, almost 400,000 full scale 3D simulations were carried out for this work. Simulations were performed for two important sources of statistical variability – random discrete dopants and line

edge roughness, both individually and in combination. This allowed the exact shape of the distribution of V_T to be determined out to $\sim 5-6\sigma$. By employing statistical analysis and data mining techniques, it was possible, for the first time, to develop methodologies for statistical enhancement of simulations of RDD and LER. The distributions of V_T obtained using statistical enhancement were verified against those obtained from brute force simulation and excellent matches were obtained for a variety of device structures. The statistical combination of the two distributions was compared to simulations of combined RDD+LER effects and also showed excellent agreement. The statistical enhancement strategies were developed with the aim of reducing the computational time necessary to accurately characterise statistical variability. By applying the developed techniques, it was demonstrated that characterisation to a similar level of accuracy as obtained here through brute force simulation could be obtained from less than 10,000 CPU hours of simulation, compared to over 300,000 CPU hours needed for the brute force approach.

In Chapter 2, a description of some of the key sources of intrinsic parameter fluctuations in contemporary bulk MOSFETs was given. The particular sources investigated in this work – random dopants and line edge roughness – were described in detail along with their effects on MOSFET operational characteristics. Some of the common simulation techniques used to study intrinsic parameter fluctuations and their advantages and disadvantages were also discussed. Drift/diffusion, Monte Carlo and Non-equilibrium Green's functions approaches were outlined. Due to the requirement for simulation on extremely large statistical scales, computational efficiency was the most important consideration and for this reason drift/diffusion methods were exclusively used in this work. Quantum corrections that allow drift/diffusion to better capture the operation of nano-scale MOSFETs are included in the Glasgow simulator and were also outlined.

In Chapter 3, details of the simulation methodology were given. The operation of the Glasgow 3D atomistic simulator was briefly described and details were given on how random dopants and line edge roughness were introduced into the simulation. Due to the large scale of the simulations carried out, it was necessary to employ Grid technology to enable proper management of the

computational resources and output data. An outline of the corresponding methods and tools was given and some of the problems encountered were described. The characteristics and doping profile of the 35 nm MOSFET that was the focus of this study were described in detail. Finally, the specifics of the simulations carried out, such as the definition of threshold voltage used in this work, were given.

Chapter 4 described the results obtained from the large sale simulation of random dopant induced threshold voltage variability in 100,000 devices. The distribution of V_T was examined in detail and was shown to be positively skewed. For the 35 nm transistor studied, the distribution of V_T had a skew of $\sim$0.16, which increases with scaling, with a skew of $\sim$0.22 obtained for a 13 nm transistor. The non-Gaussian nature of the distribution has significant implications for yield estimation in large systems. The obtained distribution and random dopant positional data was then analysed in detail to determine the correlation between dopant position and threshold voltage. This allowed the statistically significant region of the device to be determined, which extends approximately from the source PN junction to the drain PN junction and less than the depletion width down from the interface. Further analysis determined the variation in V_T due to dopant number and dopant position and a methodology for reconstructing the distribution of V_T was developed. An error analysis of this methodology was then carried out and scenarios for statistical enhancement of random dopant simulations were demonstrated.

The simulation study was continued in Chapter 5 for LER induced threshold voltage variability. The results obtained from simulation of 35,000 devices were presented and the distribution of V_T again accurately characterised. In this case, the distribution of V_T was shown to be negatively skewed, with a skew of $\sim$0.41 for the 35 nm transistor. Statistical analysis was carried out and identified that LER induced variations in the threshold voltage were strongly correlated with the average channel length. This correlation can be characterised by simulating devices with uniform gate edges and varying channel lengths. Further analysis identified the contributing factors to LER induced variability and two semi-analytical methods for reconstructing the distribution of V_T were detailed. These semi-analytical methods allow the distribution of

V_T to be determined from the distribution of the average channel length and the aforementioned correlation and were compared to results from low and high drain voltage simulations, demonstrating an excellent match to the simulation data. Comparisons were also carried out for 35 nm devices with varying width and for alternative bulk, SOI and double gate architectures, all showing close agreement with the simulation data.

Finally, in Chapter 6 the combined effects of random dopants and line edge roughness were investigated. Simulation of an ensemble of 100,000 devices again allowed the true shape of the distribution of V_T to be obtained with statistical confidence. In this particular instance, the combined distribution is close to Gaussian, however this will not generally be the case. The statistical combination of the individual distributions of V_T arising from RDD and LER was examined and found to closely match the simulated distribution. The statistical enhancement techniques developed in the previous chapters were also used to construct the individual distributions and their combination compared with the simulation data. The excellent match obtained using the statistical enhancement techniques indicated that they could be applied to make accurate predictions of the individual and combined effects of RDD and LER with a reduction in computational cost of 1–2 orders of magnitude, compared to the standard brute force approach.

7.1 Future Work

There are several areas where the work presented here could be extended. First, we primarily focus on variations at low drain voltage here, and large scale simulation at high drain voltages would be useful in order to further validate and, if necessary, refine the statistical enhancement methodologies developed in this work. Although the behaviour of LER-induced variability at high drain voltage was briefly investigated in this work, the behaviour of RDD-induced and combined variability would also be worthy of further study at high drain voltage. In addition, investigations into the width dependence of V_T variability for RDD and combined fluctuations in realistic devices would

be a useful extension to this work. These extensions would potentially allow the techniques presented here to be further generalized.

The introduction of other sources of variability would also be a natural extension of this work. In particular, work function variability is expected to become a significant problem with the recent introduction of high-κ metal gate technology [45] and a study of this would be a good next step in continuing this work. As demonstrated in this work, large statistical samples are necessary in order to fully capture the shape of parameter distributions. By continuing the large scale simulations presented here with other sources of variability, the true shapes of the particular distributions can be deduced and ideally similar statistical enhancement methodologies could be developed and integrated with those presented here.

We have also considered only n-Channel devices in this work and since variability reported for p-Channel devices is generally lower [158, 141], it would be of interest to study the underlying causes of this difference in terms of the statistical device properties.

Appendix A

Statistics

A.1 Descriptive Statistics

Descriptive statistics are used to quantitatively describe various features of a dataset or probability distribution, such as location, dispersion and shape [159]. The location can be described by the mean (μ), median and mode, which specify the expected value, the 50-50 point and the most frequent value of the data. The dispersion of a dataset or distribution can be described by the variance (σ^2) and its square root, the standard deviation. This measures the spread of the data around its expected value. The shape of a distribution can be described by the parameters skew (γ_1) and kurtosis (γ_2), which measure the asymmetry and "peakedness" of the distribution, respectively. The quantities mean, variance, skew and kurtosis can be obtained from the moments of a probability density function, which are defined as follows:

$$\mu'_n = \int_{-\infty}^{\infty} (x-a)^n f(x) dx \tag{A.1}$$

Where μ'_n is the n'th moment about the value a of the distribution $f(x)$. The moments about the origin are given by Equation A.1 when $a = 0$, with the first moment about the origin being the mean. The central moments, μ_n, are given by Equation A.1 when $a = \mu'_1$, with the variance being the second central moment. Finally, the skew and kurtosis are given by the third and

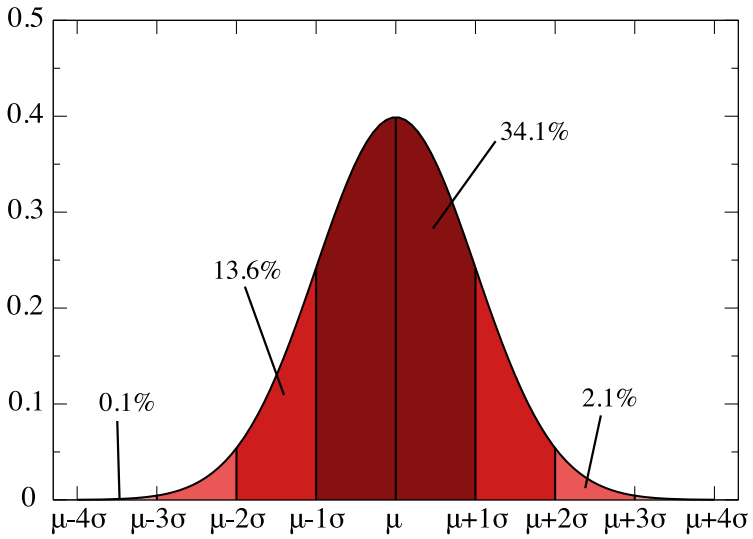


Figure A.1: Illustration of the proportion of occurences at different values of σ for a standard Gaussian distribution.

fourth standardized moments, which are defined as $\gamma_1 = \frac{\mu_3}{\sigma^3}$ and $\gamma_2 = \frac{\mu_4}{\sigma^4}$, respectively.

Statistical variability in MOSFETs has been primarily quantified using the standard deviation, which can also be interpreted graphically, as shown in Figure A.1. From the figure, it can be seen that for a Gaussian distribution $\sim$ 34.1% of occurences will be between μ and $\mu+\sigma$, obtained from $\int_{\mu}^{\mu+\sigma} g(x)dx = \frac{1}{2}Erf(\frac{1}{\sqrt{2}}) \approx 0.341$. Consquently, $\sim$ 68% of occurences will be within 1σ of the mean. Conversely, 32% of occurences will be more than 1σ from the mean. The value of σ therefore gives an indication of the rarity of a particular event. Table A.1 gives details of the number of occurences inside and outside a given value of σ for a Gaussian distribution.

Although the probability of an occurence at 6σ is very low, there are however more than 10^9 transistors on modern chips and it is inevitable that devices with deviations this far from the mean will occur. It is for this reason that circuit designs must properly account for the impact of devices with such extreme deviations.

nσ	**Fraction Inside nσ**	**Fraction Outside nσ**
1	0.683	3.17×10^{-1}
2	0.9545	4.55×10^{-2}
3	0.99730	2.70×10^{-3}
4	0.9999367	6.33×10^{-5}
5	0.999999427	5.73×10^{-7}
6	0.99999999803	1.97×10^{-9}
7	0.99999999999744	2.56×10^{-12}

Table A.1: Fraction of occurences inside and outside a given value of σ for a Gaussian distribution.

A.2 Mann-Whitney Test

Given two random samples X and Y with size m and n and density functions F and G, respectively, the Mann-Whitney test is a statistical method for testing the null hypothesis that $F = G$ [160], i.e. that the distributions of X and Y are the same. The alternative hypothesis is that $F \neq G$. The test was first proposed by Frank Wilcoxon in 1945 [161] and extended in 1947 by H. B. Mann and D. R. Whitney [147]. To test the null hypothesis, the samples from both X and Y are collectively ordered from smallest to largest, without distinguishing which sample they belong to. The ordered values are assigned ranks from 1 to $m + n$. If the null hyopthesis is assumed to be true, then the values $X_1, ..., X_m$ will tend to be randomly distributed throughout the ranked values, rather than being clustered, e.g. in the lower values.

To calculate the test statistic U, the ranks corresponding to sample X are summed. Assuming that $F = G$, the expected value of U is given by:

$$E(U) = \frac{m(m + n + 1)}{2} \tag{A.2}$$

Furthermore, the variance of U can also be calculated as follows:

$$Var(U) = \frac{mn(m + n + 1)}{12} \tag{A.3}$$

For large sample sizes, the test statistic U will be approximately Normal, with mean and variance given by Equations A.2 and A.3. This distribution

can then be used to determine whether the null hypothesis should be rejected or not, depending on whether U significantly deviates from its expected value. This can be determined by looking up p-values in standard tables of the Normal distribution.

More specifically, the null hyopthesis should be rejected if $|U - E(U) \geq c|$, where c is a constant that determines the significance level. The value of c can be calulcated by:

$$c = \sqrt{Var(U)}\Phi^{-1}(1 - \frac{\alpha}{2}) \tag{A.4}$$

Where α is the desired significance level and Φ^{-1} is the inverse CDF, also known as the quantile function, of the standard Normal distribution.

A.3 Bootstrap Resampling

Bootstrapping is a resampling technique that allows the errors in a statistical parameter to be estimated, which was introduced in 1979 by Bradley Efron [162, 160]. It is a computer-based method that makes estimates of statistical parameters using Monte Carlo simulation, which is useful when analytical forms for the parameters of interest are either unknown, too complex or do not exist. It can also be formulated in a parametric form, where assumptions are made about the underlying distributions, or as a nonparametric technique that makes no assumptions about the underlying distributions.

We begin with a random sample X with unknown distribution F and a parameter of interest, which depends on X. By way of example, we will use the skew γ_1. The sample X from unknown distribution F is replaced with a sample X^* from a known distribution, $\hat{F}$. The skew is then calculated from X^*. The choice of $\hat{F}$ is key for this technique and for the nonparametric bootstrap, $\hat{F}$ is chosen to be the sample density function. Thus, the sample X^* is a random sample from the initial sample X. It is important to note that sample size of X^* is the same as X and the values are obtained by sampling with replacement from the original data.

To obtain the distribution of the skew, a large number (N) of samples,

$X^{*(1)}$,...,$X^{*(N)}$, are drawn from the original data X and the skew calculated for each $X^{*(i)}$, yielding N estimates of γ_1 ($T = \gamma_1^{(1)}, ..., \gamma_1^{(N)}$), from which the distribution of the skew can be obtained. The standard error of the skew can then be estimated from the standard deviation of T.

Boostrapping is a general technique, and the parameter tested using this approach can be any parameter that can be meaningfully defined for X and all samples $X^{*(i)}$. Note also that the number of samples N can vary significantly and generally depends on the desired accuracy and the time and computing power available.

A.4 Interpretation of PDFs

A probability density function (PDF) describes the probability of a particular event or value occuring, and may be either discrete or continuous. For a discrete probability distribution, $p(x)$ must be ≥ 0 and the sum over all values must be 1. As a result, the probability is given by the y value of the distribution at any point x, i.e. $P[X = x] = p(x)$, and therefore $0 \leq p(x) \leq 1$. Note that a discrete probability distribution may also be referred to as a probability mass function (PMF).

For a continuous distribution, it is not possible to define the probability at a single point, since there are infinitely many points in a continuous function. The actual probability is therefore defined as the probability that a random value falls within a particular range, i.e. $P[a \leq x \leq b] = \int_a^b p(x)dx$. Continuous distributions must also be non-negative and obey the property $\int_{-\infty}^{\infty} p(x)dx = 1$. One consequence of defining the probability P over an interval is that the absolute value of the function $p(x)$ may be greater than 1, depending on the x-axis range. It is important to realise that this does not represent a probability greater than 1, but rather a probability density per unit x greater than 1. When the probability is evaluated, it must still obey the property $0 \leq P[a \leq x \leq b] \leq 1$.

Bibliography

[1] Gilbert Declerck. A look into the future of nanoelectronics. In *VLSI Symposium Technical Digest*, pages 6–7, 2005.

[2] Gareth Roy. *Simulation of Intrinsic Parameter Fluctuations in Nano-CMOS Devices*. PhD thesis, University of Glasgow, 2005.

[3] Asen Asenov, Savas Kaya, and Andrew Brown. Intrinsic parameter fluctuations in decananometer mosfets introduced by gate line edge roughness. *IEEE Transactions on Electron Devices*, 50(5):1254–1260, 2003.

[4] Tim Drysdale et al. Capacitance variability of short range interconnects. *Journal of Computational Electronics*, December 2007.

[5] Umberto Ravaioli. Hierarchy of simulation approaches for hot carrier transport in deep submicron devices. *Semiconductor Science and Technology*, 13(1):1–10, 1998.

[6] S. Datta. *Electronic Transport in Mesoscopic Systems*. Cambridge University Press, 1997.

[7] D. J. Frank, Y. Taur, et al. Monte carlo modeling of threshold variation due to dopant fluctuations. *VLSI Tech. Symp. Dig.*, pages 169–170, 1999.

[8] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38:114–117, 1965.

[9] G. E. Moore. Progress in digital integrated electronics. In *IEDM Digest of Technical Papers*, pages 11–13, 1975.

[10] H. S. P. Wong. Beyond the conventional transistor. *IBM Journal of Research and Development*, 46(2/3):133–168, 2002.

[11] R. Dennard, F. Gaensslen, et al. Design of ion-implanted mosfets with very small physical dimensions. *IEEE Journal of Solid State Circuits*, SC-9(5):256–267, October 1974.

[12] International technology roadmap for semiconductors, 2009.

[13] H. S. P. Wong, D. J. Frank, P. Solomon, C. Wann, and J. Welser. Nanoscale cmos. *Proceedings of the IEEE*, 87(4):537–570, 1999.

[14] D. J. Frank, R. Dennard, E. Nowak, P. Solomon, Y. Taur, and H. S. P. Wong. Device scaling limits of si mosfets and their application dependencies. *Proceedings of the IEEE*, 89(3):259–288, 2001.

[15] H. Wakabayashi, S. Yamagami, N. Ikezawa, A. Ogura, M. Narihiro, K. Arai, Y. Ochiai, K. Takeuchi, T. Yamamoto, and T. Mogami. Sub-10-nm planar-bulk-cmos devices using lateral junction control. In *IEDM Digest of Technical Papers*, pages 20.7.1–20.7.3, 2003.

[16] Thomas Skotnicki. Materials and device structures for sub-32 nm cmos nodes. *Microelectronic Engineering*, 84(9-10):1845 – 1852, 2007. INFOS 2007.

[17] Y. Jiang, T Liow, N Singh, L. Tan, G. Lo, D. Chan, and D. Kwong. Performance breakthrough in 8 nm gate length gate-all-around nanowire transistors using metallic nanowire contacts. In *VLSI Technology Symposium Technical Digest*, pages 34–35, 2008.

[18] T.-Y. Liow, K.-M. Tan, et al. 5 nm gate length nanowire-fets and planar utb-fets with pure germanium source/drain stressors and laser-free melt-enhanced dopant (melted) diffusion and activation technique. In *VLSI Technology Symposium Technical Digest*, pages 36–37, 2008.

[19] K. J. Kuhn. Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale cmos. In *IEDM Digest of Technical Papers*, pages 471–474, 2007.

[20] H. Iwai. Roadmap for 22 nm and beyond. *Microelectronic Engineering*, 86(7-9):1520–1528, 2009.

[21] S. Datta, T. Ashley, et al. 85nm gate length enhancement and depletion mode insb quantum well transistors for ultra high speed and very low power digital logic applications. In *IEDM Digest of Technical Papers*, pages 763–766, 2005.

[22] James Lu. Beol, sip and 3d integration technologies. *IEDM Short Course*, 2009.

[23] J. W. Joyner and J. D. Meindl. Opportunities for reduced power dissipation using three-dimensional integration. In *Proc. IITC*, pages 148–150, 2002.

[24] J. D. Meindl et al. Interconnecting device opportunities for gigascale integration (gsi). In *IEDM Digest of Technical Papers*, pages 23.1.1–23.1.4, 2001.

[25] Seigfried Selberherr. *Analysis and Simulation of Semiconductor Devices.* Spring-Verlag, 1984.

[26] Asen Asenov, A. R. Brown, John H. Davies, and Subhash Saini. Hierarchical approach to "atomistic" 3-d mosfet simulation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 1999.

[27] K. Mistry et al. A 45nm logic technology with high-k+metal gate transistors, strained silicon, 9 cu interconnect layers, 193nm dry patterning, and 100packaging. In *IEDM Digest of Technical Papers*, pages 247–250, 2007.

[28] S. Natarajan et al. A 32nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and $0.171\mu m^2$ sram cell size in a 291mb array. In *IEDM Digest of Technical Papers*, 2008.

[29] H. Kawasaki et al. Challenges and solutions of finfet integration in an sram cell and a logic circuit for 22 m node and beyond. In *IEDM Digest of Technical Papers*, pages 12.1.1–12.1.4, 2009.

[30] K. Roy. Device/circuit interactions at the 22 nm technology node. *IEDM Short Course*, 2008.

[31] S. Kamiyama, E. Kurosawa, S. Abe, M. Kitajima, T. Aminaka, T. Aoyama, K. Ikeda, and Y. Ohji. V_{th} fluctuation suppression and high performance of hfsion/metal gate stacks by controlloing capping-y_2o_3 layers for 22nm bulk devices. In *IEDM Digest of Technical Papers*, pages 17.3.1–17.3.4, 2009.

[32] Sani Nassif et al. High performance cmos variability in the 65nm regime and beyond. In *IEDM Digest of Technical Papers*, 2007.

[33] D. Pramanik, V. Moroz, and X. Wei Lin. Process induced layout variability for sub 90nm technologies. In *Proc. ICSICT*, pages 1849–1852, 2006.

[34] V. Moroz, L. Smith, X. Wei Lin, D. Pramanik, and G. Rollins. Stress-aware design methodology. In *Proc. ISQED*, 2006.

[35] H. Aikawa. Variability aware modeling and characterization in standard cell in 45 nm cmos with stress enhancement technique. In *VLSI Technology Symposium*, pages 90–91, 2008.

[36] A. Cathignol, B. Cheng, et al. Quantitative evaluation of statistical variability sources in a 45-nm technological node lp n-mosfet. *Electron Device Letters*, 29(6):609–611, June 2008.

[37] K. Ishimaru. 45nm/32nm cmos - challenge and perspective. *Solid State Electronics*, 52(9):1266–1273, September 2008.

[38] H. P. Tuinhout. Impact of parametric mismatch and fluctuations on performance and yield of deep-submicron cmos technologies. In *Proceedings of ESSDERC*, pages 95–101, 2002.

[39] Asen Asenov, A. R. Brown, et al. Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale mosfets. *IEEE Transactions on Electron Devices*, 50(9):1837–1852, September 2003.

[40] Binjie Cheng, Scott Roy, Gareth Roy, Fikru Adamu-Lema, and Asen Asenov. Impact of intrinsic parameter fluctuations in decanano mosfets on yield and functionality of sram cells. *Solid State Electronics*, 49(5):740–746, May 2005.

[41] A. Agarwal, K. Chopra, V. Zolotov, and D. Blaauw. Circuit optimization using statistical static timing analysis. In *Design Automation Conference*, pages 321–324, 2005.

[42] Noor Ain Kamsani, Binjie Cheng, et al. Statistical circuit simulation with supply voltage scaling in nanometer mosfet devices under the influence of random dopant fluctuations. In *FTFC*, 2008.

[43] T. Scotnicki. Nano-cmos and emerging technologies—myths and hopes. In *Proc. Int. Conf. SSDM*, pages 2–5, 2006.

[44] K. Ohmori, T. Matsuki, et al. Impact of additional factors in threshold voltage variability of metal/high-k gate stacks and its reduction by controlling crystalline structure and grain size in the metal gates. In *IEDM Digest of Technical Papers*, pages 409–412, 2008.

[45] Hamed Dadgour, Vivek De, and Kaustav Banerjee. Statistical modeling of metal-gate work function variability in emerging device technologies and implications for circuit design. In *Proc. ICCAD*, pages 270–277, 2008.

[46] R. W. Keyes. Effect of randomness in the distribution of impurity ions on fet thresholds in integrated electronics. *IEEE Journal on Solid State Circuits*, 10(4):245–247, 1975.

[47] T. Mizuno, J. Okumtura, and A. Toriumi. Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant num-

ber in mosfet's. *IEEE Transactions on Electron Devices*, 41(11):2216–2221, November 1994.

[48] P. A. Stolk and D. B. M. Klaasen. The effect of statistical dopant fluctuations on mos device performance. In *IEDM Digest of Technical Papers*, pages 627–630, 1996.

[49] Binjie Cheng, Scott Roy, et al. Evaluation of intrinsic parameter fluctuations on 45, 32 and 22nm technology node lp n-mosfets. In *Proc. of ESSDERC*, pages 47–50, 2008.

[50] R. Herald and P. Wang. Variability in sub-100nm sram design. In *ICCAD*, pages 347–352, 2004.

[51] Gareth Roy, A. R. Brown, et al. Simulation study of individual and combined sources of intrinsic parameter fluctuations in conventional nano-mosfets. *IEEE Transactions on Electron Devices*, 53(12):3063–3070, December 2006.

[52] Andrew Brown, Gareth Roy, and Asen Asenov. Poly-si-gate-related variability in decananometer mosfets with conventional architecture. *IEEE Transactions on Electron Devices*, 54(11):3056–3063, November 2007.

[53] E. Baravelli, M. Jurczak, N. Speciale, K. De Meyer, and A. Dixit. Impact of ler and random dopant fluctuations on finfet matching performance. *IEEE Transactions on Nanotechnology*, 7(3):291–298, 2008.

[54] T.-H. Yu, T. Ohtou, K.-M. Liu, W.-Y. Chen, Y.-P. Hu, C.-F. Cheng, and Y.-M. Sheu. Modeling and optimization of variability in high-k/metal gate mosfets. In *Proc. of SISPAD*, 2009.

[55] A. T. Putra, A. Nishida, S. Kamohara, T. Tsunomura, and T. Hiramoto. Impact of local poly-si gate depletion on vth variation in nanoscale mosfets investigated by 3d device simulation. In *Proceedings of ISDRS*, 2007.

[56] S. Bhunia, S. Mukhopadhyay, and K. Roy. Process variations and process-tolerant design. In *Intl Conf on VLSI Design*, pages 699–704, 2007.

[57] H.-S. Wong and Y. Taur. Three-dimensional "atomistic" simulation of discrete random dopant distribution effects in sub-0.1 μm mosfet's. In *IEDM Digest of Technical Papers*, pages 705–708, 1993.

[58] Y. Taur and T. Ning. *Fundamentals of Modern VLSI Devices*. Cambridge University Press, 1998.

[59] Z. Qin and S. T. Dunham. Atomistic simulations of effect of coulombic interactions on carrier fluctuations in doped silicon. *Materials Research Society Symposium Proceedings*, 765, 2003.

[60] Campbell Millar. *3D Simulation Techniques for Biological Ion Channel*. PhD thesis, University of Glasgow, October 2003.

[61] B. Hoeneisen and C. A. Mead. Fundamental limitations in microelectronics - i mos technology. *Solid State Electronics*, 15(7):819–829, 1972.

[62] K. R. Lakshmikumar, Robert Hadaway, and Miles Copeland. Characterization and modeling of mismatch in mos transistors for precision analog design. *IEEE Journal of Solid State Circuits*, SC-21(6):1057–1066, 1986.

[63] J. T. Horstmann, U. Hilleringmann, and K. Goser. Matching analysis of deposition defined 50-nm mosfets. *IEEE Transactions on Electron Devices*, 45(1):299–306, 1998.

[64] R. Difrenza, P. Llinares, and G. Ghibaudo. The impact of short channel and quantum effects on the mos transistor mismatch. *Solid State Electronics*, 47(7):1161–1165, 2003.

[65] K. Takeuchi, T. Tatsumi, and A. Furukawa. Channel engineering for the reduction of random-dopant-placement-induced threshold voltage fluctuation. In *IEDM Digest of Technical Papers*, pages 841–844, 1997.

[66] T. Hagiwara, K. Yamaguchi, and S. Asai. Threshold voltage deviation in very small mos transistors due to local impurity fluctuations. In *VLSI Technology Symposium Technical Digest*, pages 46–47, 1982.

[67] Y. Taur et al. Cmos scaling into the nanometer regime. *Proceedings of the IEEE*, 85(4):486–504, 1997.

[68] K. Nishinohara, N. Shiguo, and T. Wada. Effects of microscopic fluctuations in dopant distributions on mosfet threshold voltage. *IEEE Transactions on Electron Devices*, 39(3):634–639, March 1992.

[69] Asen Asenov. Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm mosfets: A 3-d 'atomistic' simulation study. *IEEE Transactions on Electron Devices*, 45(12):2505–2513, December 1998.

[70] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaasen. Device modeling of statistical dopant fluctuations in mos transistors. In *Proc. of SISPAD*, pages 153–156, 1997.

[71] M. Hane, T. Ikezawa, and T. Ezaki. Atomistic 3-d process/device simulation considering gate line edge roughness and poly-si random crystal orientation effects. *IEDM Tech. Dig.*, pages 241–244, 2003.

[72] T. Ezaki, T. Ikezawa, A. Notsu, K. Tanaka, and M. Hane. 3d mosfet simulation considering long-range coulomb potential effects for analyzing statistical dopant-induced fluctuations associated with atomistic process simulator. In *Proc. of SISPAD*, pages 91–94, 2002.

[73] Asen Asenov et al. Increase in the random dopant induced threshold fluctuations and lowering in sub-100 nm mosfets due to quantum effects: A 3-d density-gradient simulation study. *IEEE Transactions on Electron Devices*, 48(4):722–729, April 2001.

[74] S. Toriyama and N. Sano. Probability distribution functions of threshold voltage fluctuations due to random impurities in deca–nano mosfets. *Physica E*, 19:44–47, 2003.

[75] G. Slavcheva, J. H. Davies, A. R. Brown, and A. Asenov. Potential fluctuations in mosfets generated by random impurities in the depletion layer. *Journal of Applied Physics*, 91(7):4326–4334, 2002.

[76] Dave Reid, Campbell Millar, Gareth Roy, Scott Roy, and Asen Asenov. Analysis of threshold voltage distribution due to random dopants: A 100,000 sample 3d simulation study. *IEEE Transactions on Electron Devices*, 56(10):2255–2263, October 2009.

[77] O. Weber, O. Faynot, et al. High immunity to threshold voltage variability in undoped ultra-thin fdsoi mosfets and its physical understanding. In *IEDM Digest of Technical Papers*, 2008.

[78] E. Baravelli, A. Dixit, R. Rooyackers, M. Jurczak, N. Speciale, and K. De Meyer. Impact of line-edge roughness on finfet matching performance. *IEEE Transactions on Electron Devices*, 54(9):2466–2474, 2007.

[79] A. Asenov and S. Saini. Suppresion of random dopant-induced threshold voltage fluctuations in sub-0.1-μm mosfets with epitaxial and δ-doped channels. *IEEE Transactions on Electron Devices*, 46(8):1718–1724, 1999.

[80] Asen Asenov and Subhash Saini. Polysilicon gate enhancement of the random dopant induced threshold voltage fluctuations in sub-100nm mosfet's with ultrathin gate oxide. *IEEE Transactions on Electron Devices*, 47(4):805–812, April 2000.

[81] Yiming Li and Shao-Ming Yu. A study of threshold voltage fluctuations of nanoscale double gate metal-oxide-semiconductor field effect transistors using quantum correction simulation. *Journal of Computational Electronics*, 5:125–129, 2006.

[82] J. Bruley, T. Kane, and S. Boettcher. Measurement of ler in poly-silicon gates in mosfets by (s)tem. *Microscopy and Microanalysis*, 11:2092–2093, 2005.

[83] C. R. M. Struck, R. Raju, M. J. Neumann, and D. N. Ruzic. Reducing ler using a grazing incidence ion beam. *Proceedings of SPIE*, 7273:727346, 2009.

[84] E. Gogolides, V. Constantoudis, G. P. Patsis, and A. Tserepi. A review of line edge roughness and surface nanotexture resulting from patterning processes. *Microelectronic Engineering*, 83:1067–1072, 2006.

[85] B. Cheng, S. Roy, A. R. Brown, C. Millar, and A. Asenov. Evaluation of statistical variability in 32 and 22 nm technology generation lstp mosfets. *Solid State Electronics*, 53:767–772, 2009.

[86] T. Herrmann, W. Klix, et al. Line edge and gate interface roughness simulations of advanced vlsi soi-mosfets. In 101-104, editor, *Proc. of SISPAD*, 2007.

[87] T. Yamaguchi, H. Namatsu, M. Nagase, K. Yamazaki, and K. Kurihara. Nanometer-scale linewidth fluctuations caused by polymer aggregates in resist films. *Applied Physics Letters*, 71(16):2388–2390, 1997.

[88] M. Nagase, H. Namatsu, K. Kurihara, K. Iwadate, and T. Makino. Nanoscale fluctuations in electron beam resist pattern evaluated by atomic force microscopy. *Microelectronic Engineering*, 30:419–422, 1996.

[89] H. Namatsu, M. Nagase, et al. Influence of edge roughness in resist patterns on etched patterns. *Journal of Vacuum Science and Technology B*, 16(6):3315–3321, November 1998.

[90] H. Fukutome, Y. Momiyama, et al. Direct evaluation of gate line edge roughness impact on extension profiles in sub-50-nm n-mosfets. *IEEE Transactions on Electron Devices*, 53(11):2755–2763, November 2006.

[91] K. Patel, T.-J. King Liu, and C. J. Spanos. Gate line edge roughness model for estimation of finfet performance variability. *IEEE Transactions on Electron Devices*, 56(12):3055–3063, December 2009.

[92] H. Fukutome, Y. Hori, L. Sponton, K. Hosaka, Y. Momiyama, S. Satoh, E. Gull, W. Fichtner, and T. Sugii. Comprehensive design methodology of dopant profile to suppress gate-ler-induced threshold voltage variability in 20nm nmosfets. In *VLSI Technology Symposium Technical Digest*, pages 146–147, 2009.

[93] H. K. Gummel. A self-consistent iterative scheme for one-dimensional steady state transistor calculations. *IEEE Transactions on Electron Devices*, 11(10):455–465, October 1964.

[94] C. M. Snowden. Semiconductor device modelling. *Rep. Prog. Phys.*, 48:223–275, 1985.

[95] D. M. Caughey and R. E. Thomas. Carrier mobilities in silicon emprically related to doping and field. *Proceedings of the IEEE*, 55(12):2192–2193, December 1967.

[96] C. Lombardi, S. Manzini, A. Soporito, and M. Vanzi. A physically based mobility model for numerical simulation of nonplanar devices. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 7(11):1164–1171, 1988.

[97] S. E. Laux and M. V. Fischetti. Issues in modeling small devices. In *IEDM Digest of Technical Papers*, pages 523–526, 1999.

[98] 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*, 91(2):251–274, 2003.

[99] M. G. Ancona and H. F. Tiersten. Macroscopic physics of the silicon inversion layer. *Phys. Rev. B*, 35(15):7959–7965, 1987.

[100] M. G. Ancona. Density gradient theory analysis of electron distributions in heterostructures. *Superlattices and Microstructures*, 7(2):119–130, 1990.

[101] M. G. Ancona. Equations of state for silicon inversion layers. *IEEE Transactions on Electron Devices*, 47(7):1449–1456, 2000.

[102] M. G. Ancona. Finite temperature density gradient theory. In *Proc. IWCE*, pages 151–154, 1992.

[103] C. S. Rafferty, B. Biegel, M. G. Ancona, et al. Multidimensional quantum effects simulation using a density-gradient model and script-level programming technique. In *Proc. of SISPAD*, pages 137–140, 1998.

[104] S. Jallepalli, J. Bude, et al. Electron and hole quantization and their impact on deep submicron silicon p- and n-mosfet characteristics. *IEEE Transactions on Electron Devices*, 44(2):297–303, 1997.

[105] M. G. Ancona and G. J. Iafrate. Quantum correction to the equation of state of an electron gas in a semiconductor. *Phys. Rev. B*, 39(13):9536–9540, 1989.

[106] A. Agarwal, J. R. Watling, A. R. Brown, and D. K. Ferry. The use of quantum potentials for confinement and tunelling in semiconductor devices. *Journal of Computational Electronics*, 1(4):503–513, 2002.

[107] S. Xiong and J. Bokor. Sensitivity of double-gate and finfet devices to process variations. *IEEE Transactions on Electron Devices*, 50(11):2255–2231, 2003.

[108] Mark Lundstrom. *Fundamentals of Carrier Transport*. Cambridge University Press, 2nd edition, 2000.

[109] Carlo Jacoboni and Lino Reggiani. The monte carlo method for the solution of charge transport in semiconductors with applications to covalent materials. *Rev. Mod. Phys.*, 55(3):645–705, Jul 1983.

[110] J.-H. Rhew, Z. Ren, and M. S. Lundstrom. A numerical study of ballistic transport in a nanoscale mosfet. *Solid State Electronics*, 46(11):1899–1906, 2002.

[111] A. Martinez, N. Seoane, A. R. Brown, John R. Barker, and A. Asenov. 3-d nonequilibrium green's function simulation of nonperturbative scattering from discrete dopants in the source and drain of a silicon nanowire transistor. *IEEE Transactions on Nanotechnology*, 8(5):603–610, 2009.

[112] S. Roy, A. Lee, A. R. Brown, and A. Asenov. Applicability of quasi-3d and 3d mosfet simulations in the 'atomistic' regime. *Journal of Computational Electronics*, 2(2-4):423–426, 2003.

[113] D. L. Scharfetter and H. K. Gummel. Large-signal analysis of a silicon read diode oscillator. *IEEE Transactions on Electron Devices*, 16(1):64–77, January 1969.

[114] Gareth Roy, Andrew Brown, et al. Bipolar quantum corrections in resolving individual dopants in 'atomistic' device simulation. *Superlattices and Microstructures*, 34(3-6):327–334, 2003.

[115] Asen Asenov, Andrew Brown, et al. *Simulation of nano-CMOS devices: from atoms to architecture*, chapter Nanotechnolgoy for Electronic Materials and Devices. Springer, 2006.

[116] P. Y. Yu and M. Cardona. *Fundamentals of Semiconductors*. Springer, Germany, 1996.

[117] R. A. Smith. *Semiconductors*. Cambridge University Press, 1959.

[118] A. Asenov, M. Jaraiz, et al. Integrated process and device simulation of decananometre mosfets. In *Proc. of SISPAD*, pages 87–90, 2002.

[119] N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama. Role of long-range and short-range coulomb potentials in threshold characteristics under discrete dopants in sub-0.1um si-mosfets. In *IEDM Digest of Technical Papers*, page 275, 2000.

[120] Gareth Roy, A. R. Brown, A. Asenov, and S. Roy. Quantum aspects of resolving discrete charges in "atomistic" device simulation. *Journal of Computational Electronics*, (2):323–327, 2003.

[121] Z. Qin and S. T. Dunham. Modeling fermi level in atomistic simulation. *Materials Research Society Symposium Proceedings*, 717(C3.8), 2002.

[122] R. W. Hockney and J. W. Eastwood. *Computer Simulations Using Particles*. IoP Publishing, 1988.

[123] P. Oldgies, Q. Lin, et al. Modelling line edge roughness effects in sub 100 nm gate length devices. In *Proc. of SISPAD*, page 31, 2000.

[124] J. Wu, J. Chen, and K. Liu. Transistor width dependence of ler degradation to cmos device characteristics. In *Proc. of SISPAD*, pages 95–98, 2002.

[125] S.-D. Kim, H. Wada, and J. C. S. Woo. Tcad-based statistical analysis and modeling of gate line-edge roughness effect on nanoscale mos transistor performance and scaling. *IEEE Transactions on Semiconductor Manufacturing*, 17(2):192–200, 2004.

[126] A. T. Putra, A. Nishida, S. Kamohara, and T. Hiramoto. Random threshold voltage variability induced by gate-edge fluctuations in nanoscale mosfets. *Applied Physics Express*, 2009.

[127] T. Linton, M. Giles, and P. Packan. The impact of line edge roughness on 100nm device performance. In *Ext. Abs. Silicon Nanoelectronics Workshop*, pages 82–83, 1998.

[128] T. Linton, S. Yu, and R. Shaheed. 3d modeling of fluctuation effects in heavily scaled vlsi devices. *VLSI Design*, pages 103–109, 2001.

[129] X. Wang, S. Roy, and A. Asenov. Impact of strain on the performance of high-k/metal replacement gate mosfet. In *Proc. of ULIS*, pages 289–292, 2009.

[130] G. F. Cardinale, C. C. Henderson, et al. Demonstration of pattern transfer into sub-100nm polysilicon line/space features patterned with extreme ultraviolet lithography. *Journal of Vacuum Science and Technology B*, 17(6):2970–2974, 1999.

[131] D. MacIntyre and S. Thoms. High resolution studies on hoechst az pn114 chemically amplified resist. *Microelectronic Engineering*, 30:327–330, 1996.

[132] Simon Haykin. *Digital Communications*. Wiley, 1988.

[133] Scotgrid webpage, http://www.scotgrid.ac.uk/.

[134] I. Foster. Globus toolkit version 4: Software for service-oriented systems. In *IFIP International Conference on Network and Parallel Computing*, pages 2–13. Springer-Verlag, 2005.

[135] Ganga webpage, http://ganga.web.cern.ch/ganga/.

[136] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[137] G. van Rossum. Python language website.

[138] S. Inaba, K. Okano, et al. High performance 35 nm gate length cmos with no oxynitride gate dielectric and ni salicide. *IEEE Transactions on Electron Devices*, 49(12):2263–2270, 2002.

[139] Synopsys, Mountain View, California, USA. *Taurus Process and Device*, 2004.09 edition, September 2004.

[140] D. J. Frank, Ruchir Puri, and Dorel Toma. Design and cad challenges in 45nm cmos and beyond. In *Proc. ICCAD*, pages 329–333, 2006.

[141] K. Takeuchi, T. Fukai, A. T. Putra, A. Nishida, S. Kamohara, and T. Hiramoto. Understanding random threshold voltage fluctuation by comparing multiple fabs and tehcnologies. In *IEDM Digest of Technical Papers*, pages 467–470, 2007.

[142] A. Calderoni, P. Fantini, et al. Modelling the vth fluctuations in nanoscale floating gate memories. In *Proc. of SISPAD*, pages 49–52, 2008.

[143] J. Heinrich. A guide to the pearson type iv distribution. Technical report, University of Pennsylvania, 2004.

[144] Campbell Millar, Dave Reid, et al. Accurate statistical description of random dopant induced threshold voltage variability. *IEEE Electron Device Letters*, 29(8), August 2008.

[145] Urban Kovac, Dave Reid, et al. Statistical simulation of random dopant induced threshold voltage fluctuations for 35nm channel length mosfet. *Microelectronics Reliability*, 48:1572–1575, 2008.

[146] I. Mayergoyz and P. Andrei. Statistical analysis of semiconductor devices. *Journal of Applied Physics*, 90(6):3019–3029, September 2001.

[147] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60, 1947.

[148] Norman Gunther, Ayhan Mutlu, and Mahmud Rahman. Quantum-mechanically corrected variational principle for metal–oxide–semiconductor devices, leading to a deep sub-0.1 micron capacitor model. *Journal of Applied Physics*, 95(4):2063–2072, February 2004.

[149] NIST/SEMATECH. e-handbook of statistical methods - 1.3.3.24 quantile-quantile plot.

[150] S. Kaya, A. R. Brown, et al. Analysis of statistical fluctuations due to line edge roughness in sub-0.1μm mosfets. In *SISPAD*, 2001.

[151] M. Miyamura, T. Fukai, T. Ikezawa, R. Ueno, K. Takeuchi, and M. Hane. Sram critical yield evaluation based on comprehensive physical/statistical modeling, considering anomalous non-gaussian intrinsic transistor fluctuations. In *VLSI Technology Symposium Technical Digest*, pages 22–23, 2007.

[152] Emanuel Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[153] Ferrel G. Stremler. *Introduction to Communication Systems*. Addison Wesley, 3rd edition, 1992.

[154] B. G. Tabachnick and L. S. Fidell. *Using multivariate statistics*. Harper Collins, 1996.

[155] K. Agarwal, F. Liu, C. McDowell, S. Nassif, and K. Nowka. A test structure for characterizing local device mismatches. In *Symposium on VLSI Circuits Tech. Digest*, 2006.

[156] E. Leobandung, H. Nayakama, et al. High performance 65 nm soi technology with dual stress line and low capacitance sram cell. *VLSI Tech. Symp. Dig.*, pages 126–127, 2005.

[157] A. R. Brown, N. Idris, J. R. Watling, and A. Asenov. Impact of metal gate granularity on threshold voltage variability: A full-scale 3d statistical simulation study. *Electron Device Letters*, 2010. Submitted.

[158] A. Asenov, A. Cathignol, et al. Origin of the asymmetry in the magnitude of the statistical variability of n- and p-channel poly-si gate bulk mosfets. *Electron Device Letters*, 29(8):913–915, 2008.

[159] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2002.

[160] M. H. DeGroot and Mark J. Schervish. *Probability and Statistics*. Addison Wesley, 3rd edition, 2002.

[161] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

[162] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.