Accurate Statistical Circuit Simulation in the Presence of Statistical Variability

Plamen Asenov

Submitted in fulfilment of the requirements for

the degree of Doctor of Philosophy.

School of Engineering

College of Science and Engineering

March 2013

Dedication

I dedicate this thesis to my father and Professor, Asen Asenov, whose enthusiasm for and knowledge of semiconductor device physics is unsurpassed. He's read it already.

Abstract

Semiconductor device performance variation due to the granular nature of charge and matter has become a key problem in the semiconductor industry. The main sources of this 'statistical' variability include random discrete dopants (RDD), line edge roughness (LER) and metal gate granularity (MGG). These variability sources have been studied extensively, however a methodology has not been developed to accurately represent this variability at a circuit and system level. In order to accurately represent statistical variability in real devices the GSS simulation toolchain was utilised to simulate 10,000 *20/22nm* n- and p-channel transistors including RDD, LER and MGG variability sources. A *statistical* compact modelling methodology was developed which accurately captured the behaviour of the simulated transistors, and produced compact model parameter distributions suitable for advanced compact model generation strategies like PCA and NPM. The resultant compact model libraries were then utilised to evaluate the impact of statistical variability on SRAM design, and to quantitatively evaluate the difference between accurate compact model generation using NPM with the Gaussian V_T methodology. Over 5 million dynamic write simulations were performed, and showed that at advanced technology nodes, statistical variability cannot be accurately represented using Gaussian V_T. The results also show that accurate modelling techniques can help reduced design margins by elimiating some of the pessimism of standard variability modelling approaches.

Acknowledgements

A big thank you to my supervisors Prof. Scott Roy and Dr. Campbell Miller who spent numerous hours helping, guiding and being frustrated by me. Next on the list is Dr. Dave Reid, who's statistics knowledge is a resource I tapped numerous times. Thanks also to everyone else in the Device Modelling Group and Gold Standard Simulations for all the help and support over the years. You've all been great friends as well as colleagues. Aside from this I'd like to acknowledge the contribution of Dave New and Yves Laplance from ARM.

Thank you to the rest of my family, my mother and sister who supported and encouraged me throughout. And finally thank you to my fiancée Claire, who always believed I would finish this thesis, even when I did not.

This work wouldn't have been possible without funding from the EPSRC.

Publications

Peer Reviewed Conference Papers

- **VARI 2010 (Talk):** P. Asenov, D. Reid, C. Millar, S. Roy, Z. Liu, S. Furber, A. Asenov. Generic Aspects of Digital Circuit Behaviour in the Presence of Statistical Variability in Proc. VARI 2010
- **ESSDERC 2010 (Talk):** P. Asenov, N. A. Kamsani, D. Reid, C. Millar, S. Roy, A. Asenov. Combining Process and Statistical Variability in the Evaluation of the Effectiveness of Corners in Digital Circuit Parametric Yield Analysis In Proc. ESSDERC 2010
- **DATE 2011 (Poster):** Michael Merrett, Plamen Asenov, Yangang Wang, Mark Zwolinski, Scott Roy, Campbell Millar, Dave Reid, and Asen Asenov. Modelling Circuit Performance Variations due to Statistical Variability: Monte Carlo Static Timing Analysis
- **VARI 2011 (Talk):** P. Asenov, D. Reid, C. Millar, S. Roy, A. Asenov. Introducing Statistical Variability into SRAM SNM Simulations – a Comprehensive Study
- **SISPAD 2011 (Talk):** P. Asenov, D. Reid, C. Millar, S. Roy, A. Asenov. The Effect of Compact Modelling Strategy on SNM and Read Current in Modern SRAM
- **VARI 2012 (Talk):** P. Asenov, D. Reid, U.Kovac, S. Roy, C. Millar, A. Asenov. Advanced Statistical Compact Model Extraction for SRAM SNM Simulation
- **ESSDERC 2012 (Talk):** P. Asenov, D. Reid, S. Roy, C. Millar, A. Asenov. An Advanced Statistical Compact Model Strategy for SRAM Simulation at Reduced VDD
- **IRPS 2013 (Talk):** L. Gerrer, S. M. Amoroso, P. Asenov, J. Ding, F. Adamu-Lema, S. Markov A. Asenov. Comprehensive Statistical Simulation of Reliability of nanoscale MOSFET Devices and Circuits

- **ICICDT 2013 (Invited Talk):** Extension of ESSDERC 2012 paper

Tutorials/Workshops

- **DFM&Y Workshop 2011 (Talk):** P. Asenov, D. Reid, C. Millar, S. Roy, A. Asenov. Investigating the Effect of Adaptive Body Biasing and Adaptive Voltage Scaling on Parametric Yield in the Presence of Process and Statistical Variability

Journal Papers

- **Solid State Circuits (Invited):** P. Asenov, D. Reid, S. Roy, C. Millar, A. Asenov. An Advanced Statistical Compact Model Strategy for SRAM Simulation at Reduced VDD

Contents

List of Figures

List of Tables

Nomenclature

ABB Adaptive Body Biasing

AVS Adaptive Voltage Scaling

BSIM Berkeley Short channel IGFET Model

CLT Central Limit Theorem

DIBL Drain Induced Barrier Lowering

GLD Generalized Lambda Distribution

HPC Hight Performance Computing

KDE Kernel Density Estimate

LER Line Edge Roughness

MGG Metal Gate Granularity

MPV Most Probable Vector

NPM Non Linear Power Method

PCA Principal Component Analysis

RDD Random Discrete Dopants

SCE Short Channel Effects

SNM Static Noise Margin

SoC System On Chip

SPICE Simulation Program with Integrated Circuit Emphasis

SRAM Static Random Access Memory

STA Static Timing Analysis

Chapter 1

Introduction

1.1 Motivation

The importance of the semiconductor industry is highlighted by the fact that, even during the recent financial crisis, it achieved 7.3% growth between 2012 and 2013. The basis of such growth is the ability of the semiconductor industry to increase transistor density, reduce the cost per function, and increase system performance for each successive technology generation. The trend of doubling transistor density every 18 months, first identified by Gordon Moore in 1965, became known as *Moore's Law* [1], and was later translated into a set of "scaling rules" by Robert Dennard in 1974 [2].

The aggressive scaling rules outlined by Dennard have continued to form the basis of semiconductor development. However, as individual transistor dimensions have reduced below $100\,nm$, technology scaling has become increasingly problematic and expensive. Much of the difficulty and cost involved relates to manufacturing process improvements and developments. For example modern $22\,nm$ physical gate length transistors are printed using a $193\,nm$ lithography process [3]. But in addition to these historically well recognised problems, physical "statistical" variability effects, related to the granularity and discreteness of matter and charge, lead to significant variability in the performance of identically manufactured devices [4], causing loss of yield and increased design optimisation/verification/validation costs including the need for more onerous

simulation techniques. Traditional Static Timing Analysis (STA) techniques have been shown to be inaccurate, with more fundamental dynamic SPICE simulation required to verify the timing performance along the critical paths of a system [5].

One of the main components of modern System-on-Chip (SoC) applications is Static Random Access Memory (SRAM). The interest in SRAM stems from the fact that 20-40% of all program instructions reference memory [6], and, on-chip SRAM cache is the only sufficiently fast storage system for the quantities of data required by the processor [7]. SRAM density, and thus memory size, has to increase relative to processor speed and number of cores. One of the many advantages of transistor scaling is that the SRAM cell footprint area achieves a reduction by a factor of two per technology generation, which allows for a potential doubling of SRAM density. In order to achieve this by definition $\times 2$ reduction in cell area and subsequent increase in SRAM density at every technology generation, foundry level designers attempt to optimise the SRAM cell until the minimally sized transistors can be achieved for a cell design which provides the required yield. As the effective magnitude of statistical variability is inversely proportional to transistor area, the minimal dimensions of SRAM transistors leaves SRAM cells acutely vulnerable to statistical variability. The huge number of SRAM cells in modern memory arrays necessitates the simulation of SRAM performance beyond 5σ.

Although the problem of statistical variability has been recognised, it has become increasingly important with each subsequent technology generation. As the magnitude of statistical variability increases, so does its influence on circuit performance, power consumption and design yield. While traditional circuit design techniques sometimes include statistical variability analysis, these have generally been limited to modelling threshold voltage variability as a Gaussian distribution. The impact of these assumptions and simplifications on circuit simulation accuracy has not been thoroughly investigated, as few strategies have been proposed to accurately model statistical variability effects.

1.2 Aims and Objectives

While the impact of variability related to process and layout can be ameliorated via the maturity of manufacturing process and design for manufacture (DFM) tools, the statistical variability associated with the discreteness of charge and matter can only increase as any device architecture is scaled. It is becoming increasingly important to accurately propagate this statistical device variability up the design flow to circuit and system designers.

The main aim of this research is therefore to investigate ways to propagate device level statistical variability information to designers in such as way as to provide power/performance and yield predictions which can inform and aid the design and design evaluation process.

In order to achieve this aim we will have to accomplish the following objectives:

• Develop a methodology to integrate both process and statistical device variability into industrially relevant SPICE circuit simulations. This objective is made more complex by the differing natures of process and statistical variability, and by the fact that the pre-existing industry standard technique (corner simulation) used to analyse process variability has significant limitations when applied to statistical variability and, indeed, for modern process variability.

• Perform an extensive analysis of the accuracy of the circuit simulations using the developed this methodology as a function of the level of accuracy of the underlying device models.

• Apply this methodology to key digital system circuit components. Most importantly, the methodology should be able to analyse general digital standard cells (Boolean logic gates, adders, etc.), SRAM cells (which typically account for 60% of any digital system's circuit area) and latches.

1.3 Outline

This thesis is organised as follows. Chapter 2 outlines the sources of variability and their impact on device and circuit performance. The link between physical device performance and circuit simulation — the MOSFET *compact*

model — is introduced. Common compact modelling techniques, including different methodologies for including variability, are discussed in detail. After this, standard circuit simulation techniques, including the tradeoff between predictive accuracy and computational time, are discussed. The methodologies for introducing MOSFET variability in circuit simulations are summarised and critiqued.

In Chapter 3 the simulation methodology and the full tool flow which will be used for this work is outlined. An overview of the GSS 3D atomistic simulator GARAND is presented. Additionally, the compact model fitting tool, Mystic, is described, and the development of the statistical compact modelling strategy is thoroughly outlined. Advanced statistical compact model generation strategies, used to generate an effectively infinite number of compact models which reproduce the statistical behaviour of the extracted model set, are described. Finally, the circuit simulation tool required to use the generated compact models, *RandomSpice*, is discussed.

In Chapter 4 a 20/22nm technology generation template transistor, used in this study, is introduced, and the uniform compact model, provided by GSS, is described. The statistical compact model extraction results are presented, highlighting the accuracy of the extraction strategy, and some of the difficulties inherent in statistical compact modelling. The extracted compact model parameter distributions are used for statistical compact model generation and a number of generation strategies are tested. In order to benchmark the accuracy of the generation strategies, generated device parameters are compared to the extracted parameter data. The compact model subsampling problem, which leads to the requirement for compact model generation strategies, is also discussed in detail.

The most accurate compact modelling strategy presented, Non-linear power method (NPM), is used for the purpose of statistical variability aware SRAM cell and system simulation in Chapter 5. The purpose of this chapter is to evaluate the errors introduced into SRAM simulation though the use of traditional Gaussian V_T models, using NPM simulations as a reference. The initial simulations were focused on the standard d.c. figures of merit of SRAM performance. In order to evaluate the effect of statistical variability on SRAM

in a more industrially relevant case, the second half of the chapter focuses on the results of a joint project created in partnership with ARM Ltd. A full SRAM system is simulated and large scale NPM based simulations are used to evaluate the accuracy of a standard industry margining method.

In Chapter 6 the impact of statistical variability *and* process variability on digital logic pipelines is quantitatively assed. A full netlist, including parasitic interconnect and layout information, is used as a test circuit. The impact of different types of individual variability sources are investigated. Finally a case study is performed to evaluate the impact of statistical variability on circuit performance and yield enhancement techniques ABB and AVS.

Finally, Chapter 7 outlines the main conclusions of this work. The main results are re-iterated and suggestions for future work are proposed.

Chapter 2

Background

2.1 MOSFET Scaling

Over the last 40 years of MOSFET scaling, following Moore's law [8], the semiconductor industry has delivered increasing performance and reduced cost per function for Complementary Metal Oxide Semiconductor (CMOS) integrated circuits and systems. Achieving scaling at each new technology node is necessary for semiconductor manufacturers to ensure a competitive advantage, a point identified by Moore himself as early as 1965 [1]. As can be seen in Figure 2.1, extracted from the 2011 executive summary of the International Technology Roadmap for Semiconductors (IRTS) [9], scaling is projected to continue over the next 15 years with the physical gate lengths of transistors reaching below the $10\,nm$ mark. The period of 'happy scaling' however, has come to an end [10]. Two main reasons contribute to the increased difficulty in scaling device dimensions; (i) increasing manufacturing difficulty - for example, the manufacturing costs associated with lithography have increased by 7 orders of magnitude in four decades [11], and (ii) the ever increasing variability in transistor performance in extremely scaled transistors.

Due to the commercial importance of semiconductor scaling there is a large amount of research regarding bulk MOSFET variability at the $90\,nm$ [12] technology generation and below [13, 14, 15, 16, 17, 18]. However, due to the extremely high level of statistical variability in contemporary conventional (bulk)

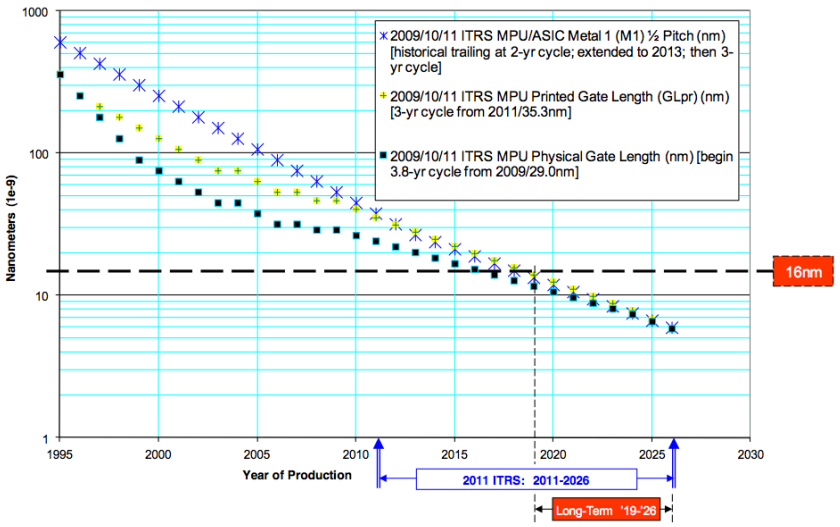


Figure 2.1: Semiconductor Device Predicted Trends, ITRS Executive Summary 2011 [9].

MOSFETs, there is a significant amount of research on alternative structures such as FinFET [19, 20, 21] and Silicon-On-Insulator (SOI) [22] devices, which will enable the continuation of CMOS scaling. The main advantages of these alternative transistor structures are improved performance, reduced leakage, and a reduction of the impact of major statistical variability sources.

At the 22 nm technology node, variability has become so problematic that the industry leading Intel has introduced novel TriGate (FinFET) transistors, despite increased development and manufacturing costs [23, 24] and the huge amount of work required in re-designing cell libraries and IP blocks for FinFETs. It is believed that FinFETS were mainly introduced to improve the stability and reliability of Static Random Access Memory (SRAM) arrays, which occupy a significant portion of chip area in modern System on Chip (SoC) applications [25]. The usage of minimal geometry transistors leaves SRAM acutely sensitive to purely statistical variability. The vast number of SRAM cells in memory arrays (for example, in a 4 Mb cache there are almost 38 million SRAM cells [16], each with 6 transistors) provides a strong motivation to study SRAM cell performance at deviations of 5 sigma and greater from the mean.

The rest of this Chapter will provide an overview of the origin and impact of different sources of variability present in modern CMOS transistors. After this the concept of *Compact Models*, used to represent complex transistor behaviour in circuit simulation, is introduced. The technology for extracting compact models and the methodologies for capturing variability in compact models, and therefore in circuit simulations, are then introduced and discussed. Finally circuit simulation techniques are introduced and variability aware simulation methods are discussed.

2.2 Variability Classification

When considering variability present in modern CMOS transistors and circuits, it is important to define a consistent terminology. In this thesis we will use the following classification:

- *Process Variability* is the variability associated with the manufacturing of devices, manifested as slow parameter drift across chip, across wafer and wafer-to-wafer, this is depicted in Figure 2.2(a,b,c).
- *Systematic Variability* is a sub-class of process variability, which is layout dependent, and is introduced by lithography, strain and well proximity effects, as is depicted in Figure 2.2(d).
- *Statistical Variability* is due to the discreteness of charge and granularity of matter. In the absence of process and systematic variability in a hypothetical "perfect" manufacture process, statistical variability would still be present. This is depicted in Figure 2.2(e).

The remainder of this section presents and discusses the sources and the impact of these types of variability. It is important to understand and accurately capture these variability effects as they have a significant impact on circuit and system performance, power and overall product yield. Ideally, accurate transistor level variability information should be available to circuit designers significantly before technology introduction, for the purpose of circuit optimisation in the presence of variability. The design process usually involves performance

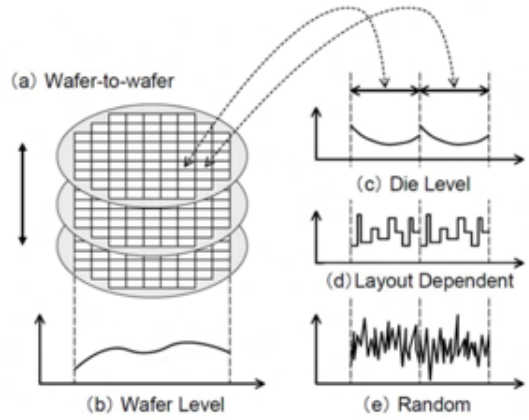


Figure 2.2: Different sources of variability in CMOS devices and their impact [13].

optimisation, yield improvement, and even device/circuit co-design to optimise a transistor technology simultaneously with circuit design, and hence, the earlier in the design process that suitably accurate variability information is introduced, the more effective optimisation tends to be.

2.3 Process Variability

Historically the dominant sources of variability, introduced through process variations, known as 'process' or 'global' variability, which result in a slow parameter drift from wafer-to-wafer, across wafer and across die, as shown in Figure 2.2(a,b,c). Process variability is a complex but well understood and researched problem, with papers on ion implantation and oxide thickness variation dating back to as early as 1974 [26]. The problem of process variability, however, has not been eliminated as it dynamically evolves with each technology generation. Changes in device structure, processing, materials and conditions, some of which are shown in Figure 2.3, have been introduced due to the industry requirement for continuous scaling. Many of the manufacturing process changes have been driven by the need to control Short Channel Effects (SCE) in deep sub-micron devices, to ensure the required performance and density gain of each subsequent technology generation. As a result of these

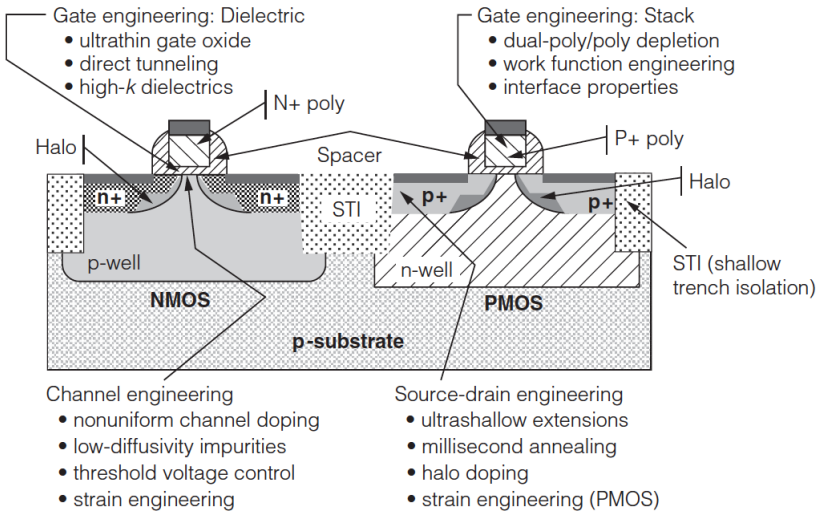


Figure 2.3: CMOS cross section with major sources of process variability [14].

changes, new sources of process variability are constantly being introduced [14]. Some examples of these changes include: shallow trench isolation (STI), strain, the use of high-κ gate dielectrics and halo doping. The introduction of these process steps have added sources of variability including:

- Variation due to the physical implant and anneal process, which includes variability introduced through halo implantation, accuracy and purity of dose [27] and variation in peak anneal temperature.
- Chemical metal polish (CPM) variability effects in the polishing of Shallow Trench Insulation (STI) [28] leading to gate height variation in both polysilicon and metal gates [29].
- Variation in film thickness impacting oxide thickness, gate stacks, wire and dielectric layer height, due to the deposition and growth process [30].
- Temperature non-uniformities in the critical post exposure bake (PEB) and etch steps [30].

Process variability is complex but generally well understood, characterised, and in many cases predictable and manageable [17, 18, 31]. It can be measured effectively through the use of appropriate test structures [32] and can be

effectively reduced, maturing the technology in co-operation between process and design techniques, as well as purely design based improvements [33].

There are also techniques available which aim to ameliorate process variability at a circuit level, across a die and from die to die, including adaptive voltage scaling (AVS) [34], adaptive body biasing (ABB) [35, 36] and gate length biasing [37]. These methods are effective due to the slow changing nature of process variability, meaning all local transistors will be similarly impacted. Due to this property of process variability, a chip or die level global adjustment in supply voltage (AVS) or body bias (ABB) can be applied to correct for the process variability present. The effect of process variability is modelled in circuit simulation through the use of compact model *process corners* [38, 15, 14], which will be discussed in Section 2.7.3.

In addition to the relatively small process variability, typically 5-10% change in performance across a wafer [39], present devices also exhibit significant layout dependent systematic and purely statistical on-chip variability.

2.4 Systematic Variability

Systematic variability, illustrated in Figure 2.2(d), is mainly related to device shape and strain [18]. One of the principle sources of systematic variability is lithography. This is due to the fact that modern devices with physical dimensions between $25\,nm$ and $35\,nm$, are still printed with $193\,nm$ light sources [40]. In dense patterns proximity and fringing effects become problematic, as each printed shape is influenced by surrounding patterns. This leads to drawn gates which do not perfectly match the mask shape, yielding systematic errors and variability in device dimensions [41]. A good example is the well proximity effect [42], related to the implantation of the well in the substrate, which causes ions to react with the photo-resist boundary and can cause a variation of threshold voltages in the devices within $1\mu m$ to $2\mu m$ of the well edge [43].

The problem is largely mitigated through optical proximity correction (OPC) [44]. This technique, shown in Figure 2.4, involves distorting the lithography mask so that the drawn structures reproduce the designed shapes more accur-

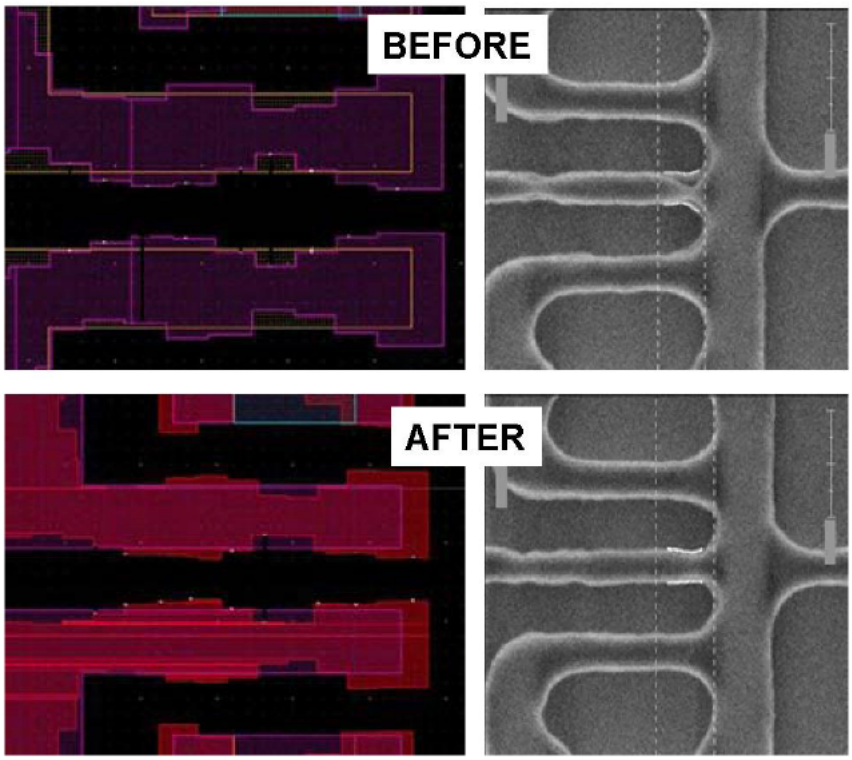


Figure 2.4: Lithography induced variability and OPC based improvement, the left images show the predicted drawn shapes and the right images show Scanning Electron Microscope (SEM) images of the manufactured patterns [49].

ately. This process is only possible because the lithography induced distortion can be calculated and as a result, the mask required to invert the transfer function of the distortion can be constructed [40] and the correct device shapes can be drawn.

Although OPC can mitigate systematic lithographic variability effects, the introduction of strain [45] exacerbates this source of variability. The application of strain is not uniform and introduces additional variation which is dependent on physical channel length [46].

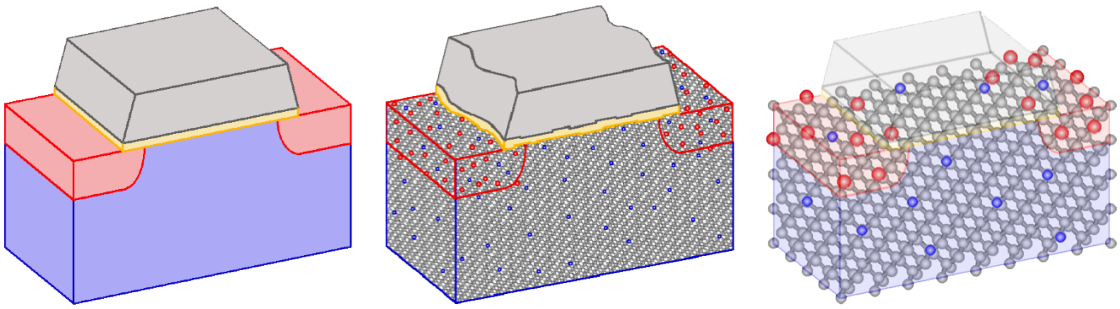
Overall, due to its nature, systematic variability can be mitigated and managed using design for manufacture (DFM) tools [47] as well as via the push towards regularised design [48].

2.5 Statistical Variability

Statistical variability, which is exacerbated by aggressive scaling, is caused by the granularity of charge and matter. In present bulk technologies, the dominant statistical variability sources are random discrete dopants (RDD) [50, 34], line edge roughness (LER) [51] and polysilicon/metal gate granularity (P/MGG) [29]. Unlike process and systematic variability, statistical variability is truly random and produces differences in performance of otherwise microscopically identical transistors, as illustrated in Figure 2.2(e). Though statistical variability has been under investigation for a long period of time, it has only become industrially relevant as devices have reached the deep submicron region, in fact, it was been recently demonstrated that the magnitude of purely statistical variability in the 45nm technology generation exceeded that of process variability [13, 12]. Due to its *atomistic* nature, for a set technology type, statistical variability is impossible to reduce, and must be taken into account during both the transistor and circuit design process.

2.5.1 Random Discrete Dopants

The cause of Random Discrete Dopant (RDD) variability is clearly illustrated in Figure 2.5, where the "ideal" transistor, shown in Figure 2.5 (a), assumes continuously doped regions, well defined boundaries between the p-n junctions, uniform oxide thickness and ideal parallel gate structure. Moving to a "realistic" device, shown in Figure 2.5 (b), the silicon crystal lattice can be seen, with the position of dopant atoms represented in red and blue in the channel and source/drain regions. Random dopants are introduced predominantly through ion implantation, and are activated and redistributed through annealing [52]. In a large enough device the random dopant configuration will not have a significant effect due to averaging, as indicated by the Central Limit Theorem (CLT, see [53]). However, as device dimensions are scaled down, the reduced number of active dopants in the channel region and their random positions (see Figure 2.6) have a significant impact on device performance [34, 50, 54].



(a) (b) (c)

Figure 2.5: (a) Ideal transistor with continuous structure, (b) realistic transistor showing silicon lattice and dopants, (c) scaled transistor at 4.2nm emphasising the small number of random dopants at these dimensions [59].

The first order effect of random dopant variation is a random shift in the threshold voltage (V_{th}) of a device. Large scale simulation studies have shown that the RDD induced V_{th} distribution is close to Gaussian, however deviations are seen in the tails of the distribution for a sufficiently large sample size [54]. It has also been shown that in a $22\,nm$ bulk CMOS technology, the standard deviation of the threshold voltage can be as large as $75\,mV$ [55]. Although most of the published work relating to random dopants concentrates on their electrostatic effect on threshold voltage [54, 56, 57], random dopants also cause transport variability due to variations in ionised impurity scattering from transistor to transistor. It has been shown that for bulk transistors down to the 22nm technology generation RDD are the dominant source of statistical variability [58].

2.5.2 LER

Line edge roughness stems from the molecular structure of the photo-resist used in the lithography process. Its main effect on MOSFET operation is associated with local variations in the gate length as illustrated in Figure 2.5 (b) where the transistor has non-uniform gate edges along the channel width. A realistic example of LER can be seen in Figure 2.7, with local non-uniformity of the channel clearly visible as well as decorrelated source and drain edges.

With present $193\,nm$ lithography resists, the minimum limit of LER is

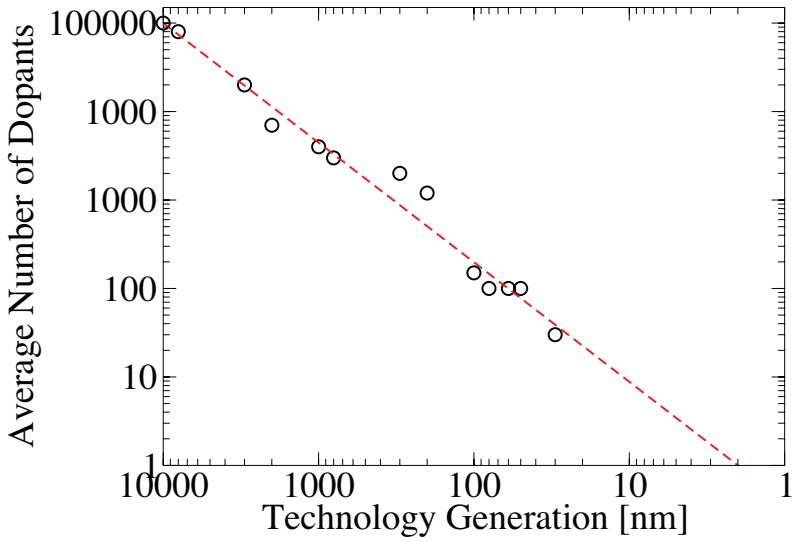


Figure 2.6: Average number of dopants in the channel as a function of technology node, extracted from[33].

around 5nm with RMS amplitude, $\Delta \approx 2\,nm$ and correlation length $\Lambda \approx 20\,nm$ [51]. This can have a significant effect on effective channel length, introducing $\pm 10\%$ variations in minimum channel length in a nominally 25nm channel length transistor. It is clear that the importance of LER will increase as a source of variability as transistor dimensions shrink. LER introduces significant variability in subthreshold current as well as threshold voltage. A worrying effect of LER is the degradation in $I_{\text{on}}/I_{\text{off}}$ ratio, due to enhanced short channel effects which can have a serious impact on device and circuit performance [60]. A measurement based study with a 90nm technology showed that LER effects become significant below 85nm channel length, and cause a four order of magnitude increase in the variance of the off-current [61]. In simulation studies where individual sources of statistical variability have been evaluated, it has been shown that at small enough device dimensions, around 14nm for bulk MOSFETs, LER supplants RDD as the dominant source of variability [55]. Furthermore, LER effects have been shown to be the dominant source of variability in Multi-Gate devices such as FinFETs [62].

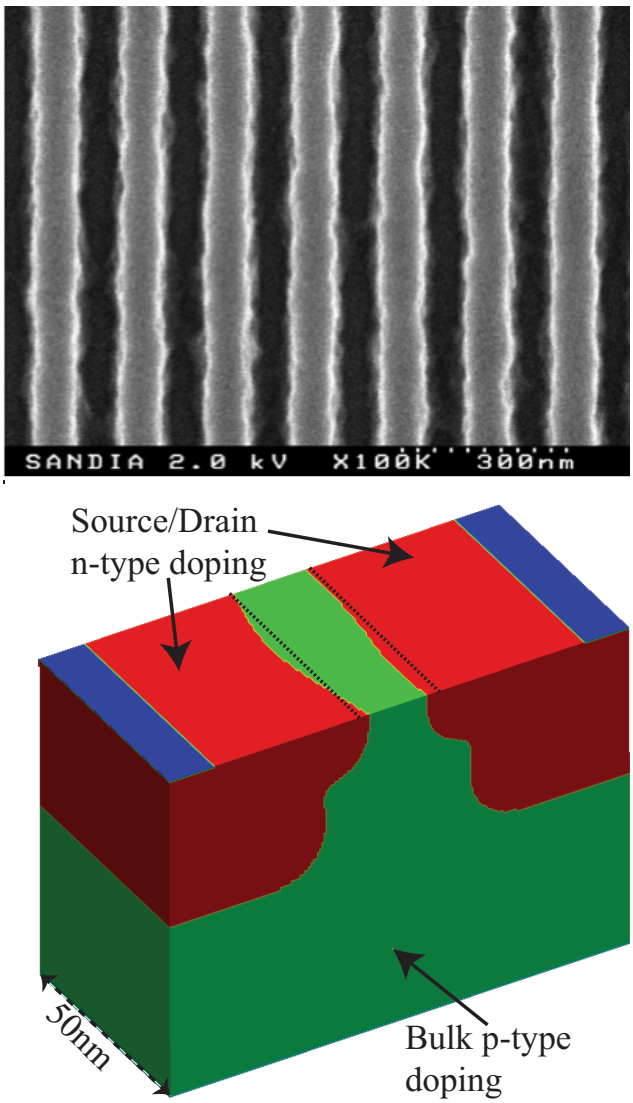


Figure 2.7: Example of LER photo-resist from SANDIA Labs, also shown is a realistic device with LER variability effects, the red regions represent the source/drain of the device, while the green region indicates substrate. The non-uniform nature of the channel edge is clearly seen.

2.5.3 Polysilicon/Metal Gate Granularity

Metal Gate Granularity (MGG) is a problem associated with the "gate first" process technology [63], where the gate metal is deposited before any high temperature annealing process. During high temperature processing, the nominally amorphous metal gate material becomes poly-crystalline. In metal gates this results in the formation of grains of differing metal work function, while in polysilicon gates the boundaries formed between grains cause Fermi level pinning and doping non-uniformity due the to rapid diffusion along grain boundaries. Both of these effects have a serious impact on the performance of the devices. The effect of P/MGG on device performance is highly dependant on the material grain size [64], with respect to overall gate size. In the case of metal gate, large grain size with respect to overall gate size, produces a bimodal impact on transistor performance, where different device instances are dominated by different grain work functions. Significantly small grain size, relative to overall gate area, will result in self-averaging of the grain work functions and have a small impact on overall device performance.

In the gate last processes, where the gate is deposited post annealing and the other high-temperature processes, the gate material remains effectively amorphous and does not introduce significant statistical variability effects.

2.5.4 Combined Statistical Variability

While it is important to understand and study the individual sources of statistical variability, it is paramount to model and evaluate the combined impact of RDD, LER and P/MGG. The impact of these sources of statistical variability on the threshold voltage of bulk MOSFETs has been shown to be relatively statistically independent [4]. Therefore the statistical addition of the individual variability sources, in respect to threshold voltage variability, is determined by Equation 2.1,

$$\sigma V_{T_{TOTAL}} = \sqrt{\sigma V_{T_{RDD}}^2 + \sigma V_{T_{LER}}^2 + \sigma V_{T_{P/MGG}}^2} \tag{2.1}$$

Where $\sigma V_{T_{TOTAL}}$ is the standard deviation of the threshold voltage distribu-

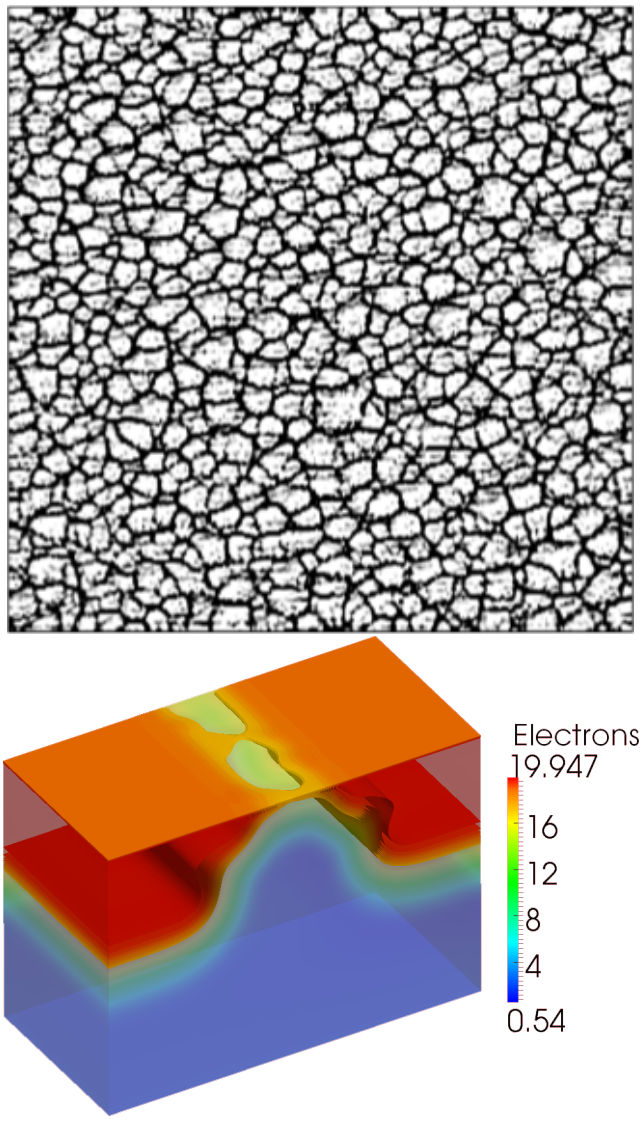


Figure 2.8: SEM micrograph of typical polysilicon grain from [52], and a simulated device exhibiting MGG.

tion of the device subject to RDD, LER and P/MGG, $\sigma V_{T_{RDD}}$is the standard deviation of the threshold voltage distribution of the device subject to RDD, $\sigma V_{T_{LER}}$is the standard deviation of the threshold voltage distribution of the device subject to LER and $\sigma V_{T_{P/MGG}}$ is the standard deviation of the threshold voltage distribution of the device subject to P/MGG. In order to fully analyse the impact of statistical variability in realistic devices in statistical circuit simulation, all sources of variability have to be included. For this purpose, a 3D Drift Diffusion simulator with density gradient quantum corrections - GARAND, fully described in Chapter 3 - will be employed.

Figure 2.9 shows 10,000 microscopically different $25\,nm$ n-channel transistors with a width of $25\,nm$, simulated at a high drain bias of $1\,V$ in the presence of RDD, LER and P/MGG using GARAND. This figure illustrates the statistical variability challenges facing circuit and system design at advanced technology generations. I_{off} distribution spans 5 orders of magnitude, there is significant I_{on} variability, and threshold voltage standard deviation is $75\,mV$.

Additional variability effects are related to the introduction of high-κ materials. Use of such materials increases effective oxide thickness and reduces direct tunnelling gate leakage, as well as reducing the random dopant induced statistical variability (which is inversely proportional to effective oxide thickness). However, imperfections in the high-κ to Si interface can lead to the formation of traps and decreased reliability [65]. This source of statistical variability will not be considered as part of this work.

In this thesis the main focus is on statistical variability, as process variability and systematic variability are well captured by traditional modelling and simulation techniques. Due to the extremely large number of transistors in modern SoC applications, it is difficult to capture and understand the impact of the performance of extreme devices, which could have a large impact on system performance. As statistical device variability has become more dominant, it becomes increasingly important to propagate variability information to circuit and system designers and enable a variability aware power/performance/yield optimisation.

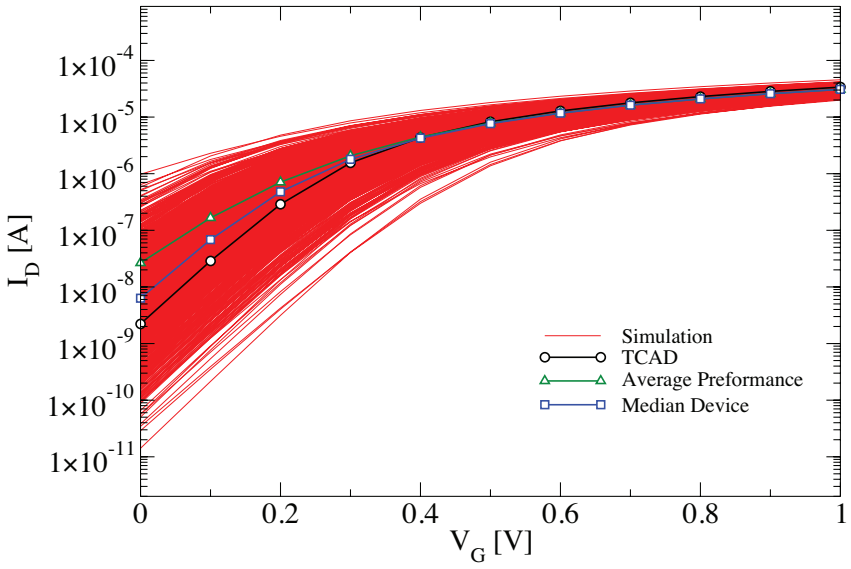


Figure 2.9: Transfer characteristics of 10,000 simulated 25nm gate length and width devices with RDD, LER and MGG simulated using the GSS 3D device simulator GARAND. Simulation drain bias is 1V. The plot also shows the initial uniform device (TCAD), average device performance and the median device, for reference.

2.6 Impact of Variability On Circuit Design and Verification

The impact of variability on design and verification can be split into two distinct categories, the effect of process variability and the effect of statistical variability. Process variability has historically been the dominant variability problem in traditional digital design. Typically manifested as a slow parametric drift across wafer, process variability effects all local transistors in a similar way. Local mismatch is minimal, but circuits on opposite sides of a chip, across the wafer, or from die-to-die, may have a large performance/power discrepancy. The impact of statistical variability, which causes differences on a local device level, is not well captured using the simulation techniques which have matured in industry to capture process variability. The most important effects of statistical variability on digital logic manifests in the form of delay and leakage power variability [66]. The Central Limit Theorem, outlined in [53], dictates that in long logic paths, with many additive stochastic delays, variability in signal propagation delay generally follows a Gaussian distribution. Further to this, the Law of Large Numbers, states that the magnitude of variability in delay introduced through statistical variability reduces as a function of path length.

Variability in digital circuit leakage power, introduced through statistical variability, has an exponential dependence on threshold voltage variability, as outlined in section 2.5. This can be extremely problematic, especially in low power applications and mobile devices, where an order of magnitude increase in leakage power can have a catastrophic effect on battery life. The leakage problem is very important for large volume commercial applications. For example, speculation in 2012 that Apple Inc. might replace the Intel microprocessors in their laptops with a lower power alternatives (media speculation [67]). In addition, one of the motivating factors behind the introduction of FinFET and SOI devices is their superior subthreshold slope and lower variability [19, 24, 20, 21], resulting in lower leakage and static power consumption.

The impact of statistical variability on analogue and analogue -like systems

- for example SRAM and latch registers - is more complex [52, 16, 68, 69], as many of these circuits rely for their operation on balanced transistor pairs. While local transistor mismatch under the impact of process variability is not significant, statistical variability can locally introduce critical variation in otherwise 'identical' devices, and can adversely impact the intended operation of these circuits. This is the main motivating factor behind the large amount of research into the performance of scaled SRAM at current and future technology nodes. This subject will be elaborated further in Section 2.9.

The main design concern in the presence of statistical variability is that instead of designing for a single device/circuit performance, or simply verifying design at pre-defined process variability corners, designers have to take into account a distribution of possible device performances, and ensure that designs still function in the most extreme device combinations. It becomes important to evaluate the projected yield during the design process in comparison with the design yield requirements. This is specifically important for high yield, low cost products, where profit margins can be relatively small, and a lower than required yield can cause a significant financial loss. The level of computational complexity involved in yield evaluation increases significantly when designs have to be simultaneously evaluated over different system performance indicators - including timing, power, leakage and density.

In order to be able to accurately determine circuit yield, the corresponding circuit simulations require accurate variability information. Different circuit simulation techniques are available with the standard trade-off between computational time, overall accuracy and predictive power. In order for these simulation techniques to be predictive, transistor performance in the presence of all variability effects, has to be accurately modelled. *Compact models* are designed for this purpose, and act as the link between silicon measurement (or TCAD simulation of device characteristics), circuit simulation and design.

2.7 Compact Modelling

Any nodal analysis based simulator is limited by the accuracy of the representation of its circuit components. While "simple" elements like resistors and capacitors can generally be described analytically within limits, complex non-linear circuit components like MOSFETs must be represented by *compact models*. A compact model is a set of related quasi physical equations which describe the operation of the required circuit element, with a set of tuneable parameters, which the circuit simulator can treat as a "black box" to which it supplies input nodal voltages and receives as outputs terminal currents. A MOSFET compact model has to capture both steady state and transient performance of the represented device in all possible modes of operation, including: drain bias dependence, gate bias dependence, body bias dependence, temperature dependence, as well channel length and width dependence. Aside from the basic behaviour of an ideal transistor, process variability, systematic variability and statistical variability have to be captured in order to fully represent realistic device performance and ensure accurate circuit simulations.

The compact model most often used to represent bulk MOSFET devices is the Berkeley Short-channel IGFET Model (BSIM), which was developed in the late 1980s to incorporate complex short channel effects not captured well in other models of the time. The motivation behind BSIM was "*to develop a semi-empirical model which can cope with rapid changes and advancements in technology*" [70], while avoiding the complexity of directly modelling all the underlying physical effects in solid-state transistors. Numerous iterations of the BSIM model have introduced more improvements and developments, closely matching technological advances including enhanced mobility models, halo doping effects and the inclusion of stress and noise models.

More recently, a family of alternative 'surface potential' based compact models have been introduced, in an attempt to provide a better physical representation of advanced CMOS devices. The most popular of these models, PSP [71], has been extensively benchmarked with respect to BSIM, and has shown little advantage[72], especially where statistical variability is considered [73]. Compact models are also available for Tri-Gate architectures (BSIM-

CMG [74]) and SOI devices (BSIM-IMG [75], UTSOI [76]).

In this thesis we will focus on the BSIM compact models due to their wider adoption in industry.

2.7.1 Compact Model Extraction

The challenge of compact model extraction is to accurately represent complex circuit elements, like MOSFETs, for the purposes of circuit simulation. Compact model extraction using a standard pre-defined model type is a multi-stage process, which requires representative device performance data under all usual modes of operation and possible device geometries. This data can be obtained through device simulation or direct silicon measurement.

Traditionally, initial compact models provided by foundries are based on TCAD simulated performance. The accuracy of these models is then improved when silicon data is available so the compact model is accurately representative of the technology, and further model updates are released as process changes are introduced or more accurate physical measurement data is available. An example of a BSIM4 compact model, fitted to simulated transistor $I_D - V_G$ characteristics can be seen in Figure 2.10.

The recommended BSIM4 extraction process is outlined in the BSIM4 manual [77]. Initially this involves supplying the model with physical and structural transistor parameters including channel doping concentration, oxide thickness, junction depth, source-drain resistance and the nominal temperature of the simulation/measurement data. These parameters form the initial model and are generally not significantly altered during the extraction process as the model attempts to maintain the physical meaning for these parameters. Additional parameters are slowly introduced until the required level of accuracy is met. Compact models are usually fitted to model channel length dependance, body bias dependance, gate and drain bias dependance as well as temperature dependance [78]. Due to the large amount of data required and the potentially large number of parameters involved (there are over 100 compact model parameters in BSIM4), this can be a long process which yields a model that represents the average behaviour of the transistor.

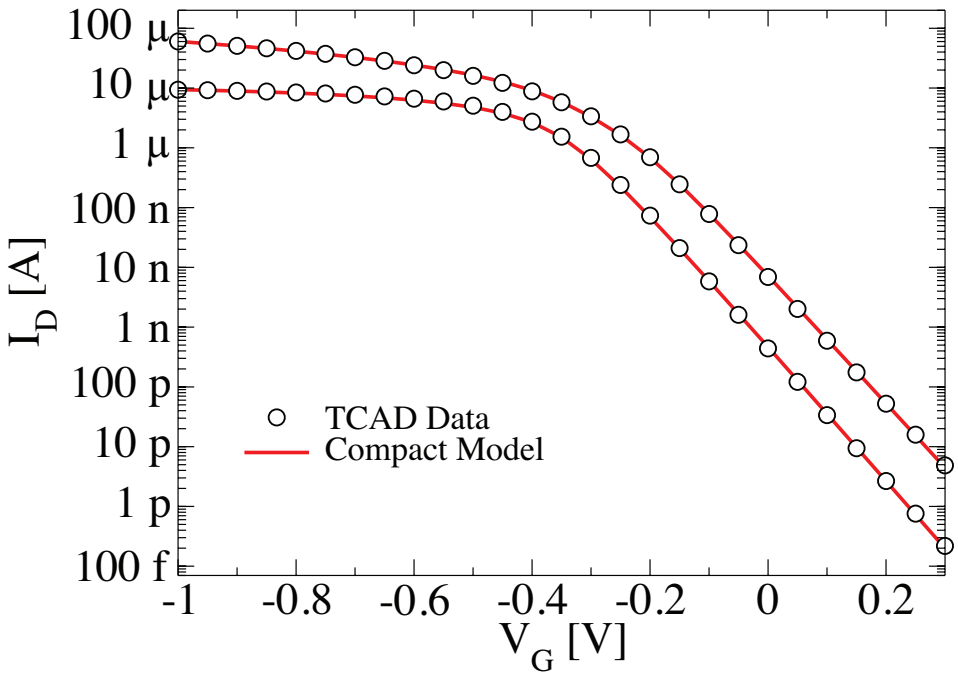


Figure 2.10: $I_D - V_G$ data from simulation compared to a fitted compact model, the two curves represent low drain ($V_D = 50mV$) and high drain ($V_D = 1.0V$) bias conditions.

The extracted 'nominal' compact model is representative of the uniform or ideal transistor behaviour. No variability information is included at this stage. It is vitally important to include variability information into the compact models to enable design verification as well as power/performance/yield optimisation of a design, as outlined in Section 2.6.

2.7.2 Variability Aware Compact Modelling

Due to their different natures, process and statistical variability have to be handled separately. For a small system or chip, process variability may be applied globally to the whole system, while statistical variability has to be injected on a transistor-to-transistor basis. Process variability is captured through 'Corner Models' [79] while statistical variability is traditionally captured through Gaussian V_T generation [80]. In addition to these methods, a novel statistical variability aware compact modelling approach, introduced by *Cheng et al.* [81], will be described below.

2.7.3 Corner Model Analysis

The impact of process variability on circuit performance is usually evaluated though a sequence of circuit simulations using compact models of devices which represent extreme device performance due to process variability known as *Corner Models*. These models are: typical n and p-MOSFETs, which represent the average or designed performance of the devices; fast n- and fast p-MOSFETs, which defines the maximum leakage/minimum delay corner in digital logic circuits; slow n- and slow p-MOSFETs, which define the maximum delay/minimum leakage corner in digital logic circuits; fast n- and slow p-MOSFETs and slow n- and fast p-MOSFETs, both of which describe the maximum mismatch corners [79]. Corner simulations are often performed at temperatures which produce worst performance for the circuit metric under investigation. The typical, slow and fast models for the transistors used in these simulations are based on foundry measurements of simple circuits such as ring oscillators [32]. Using these measurements it is possible to estimate process

corners at set standard deviations from the mean values of the measured distribution of the desired device parameters. The theory behind this simulation methodology, known as 'corner analysis', is that all possible combinations of extreme circuit performance, due to process variability, are fully represented, and therefore if circuit specifications are met within these simulations a close to 100% yield can be obtained. This form of analysis is sufficiently rigorous until the point where technology scaling causes statistical variability to have a significant impact on circuit performance.

Several methods have been proposed, to extend the applicability of the corner analysis method [82]. One of these is the 'corners and statistical' approach. This approach attempts to include statistical variability into the corner analysis methodology, by injecting statistical variability into corner models, thus including the effect of both sources of variability in the simulations [83]. This approach has been shown to be overly pessimistic, as it over-emphasises the possibility of having a circuit with a combination of poor process and statistical variability. In addition, it does not allow for Power/Performance/Yield (PPY) analysis as the overall performance distribution is not simulated, but simply the performance distribution at each corner.

2.7.4 V_{T} Base Variability Simulations

For a long time, the standard method of introducing statistical variability into compact models, and thus circuit simulation, has been through threshold voltage parameters. This is due to the fact that, in technology generations up to $90\,nm$, the first order impact of statistical variability can be captured relatively well as shift in the threshold voltage. The simplest way this can be modelled at the circuit level is through a Monte-Carlo generation strategy where a Gaussian distribution injected into a threshold voltage equivalent compact model parameter. Gaussian V_T methods are popular as it can be easily implemented, and due to the fact that it greatly simplifies analytical techniques as they assume all statistical variability effects can be captured in a single parameter, which is defined by its first two moments - the mean and standard deviation. This method, however, does not capture complex 1st

and 2nd order effects of statistical variability, such as on-current variability, Drain Induced Barrier Lowering (DIBL) variability, off current variability and subthreshold slope variability. These effects can have a significant impact on circuit performance, especially in non-digital logic circuits like SRAM and analogue systems [84]. Another, equally serious problem with Gaussian V_T based simulation, is that there is good evidence from measurements which shows that the distribution of the threshold voltage begins to deviate from Gaussian [54] at large values of σV_T. Papers have recently been published which show the disadvantages of using this strategy [85]

The errors introduced into circuit simulation as a result of simply representing statistical variability as Gaussian V_T variation will be thoroughly investigated through the course of this thesis through comparison with more rigorous methods. One possible solution for evaluating the effect of statistical variability or 'mismatch', originating from research in the analogue design domain, has been the introduction of 'variability injectors' [86]. This strategy involves introducing a voltage source at the gate of the MOSFET to simulate a threshold voltage shift, and a current source in parallel with the MOSFET, to take into account variability in the current factor $\beta = \frac{W}{L}\mu C_{ox}$. The advantage of this method is that on-current variability can be modelled as a second order effect to threshold voltage variability. It is also a relatively simple method to implement without altering the underlying compact model. The disadvantages of this method include the inability to capture correlation between 1st order effects and complex 2nd order effects, like DIBL and drain bias dependence of subthreshold slope, which have become increasingly important in scaled devices [84], as well as the increased computational complexity of two extra circuit elements (the voltage source and current source) per MOSFET in the system.

Another proposed extension to the Gaussian V_T methodology involves modelling DIBL as an independent variable [87]. This is possible as it has been shown that, for bulk devices, DIBL is uncorrelated to low drain threshold voltage and can be effectively modelled with a log-normal distribution. The results show that including DIBL in SRAM simulation has a large impact on predicted cell performance. Although these methods have shown improvements

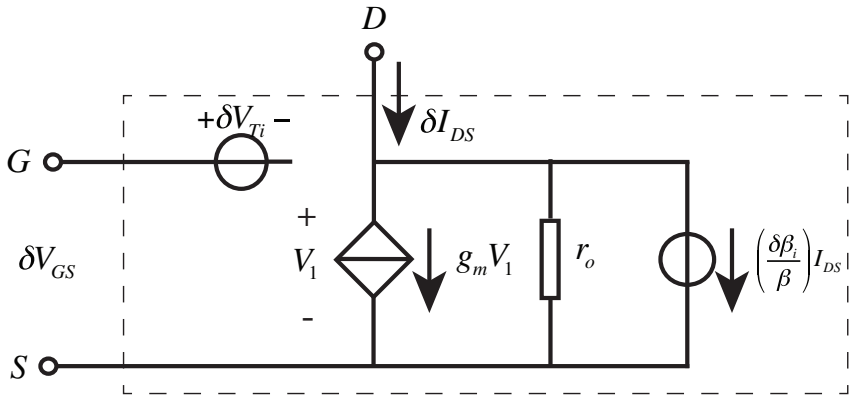


Figure 2.11: Equivalent MOSFET model with sources of current and voltage variations [86]

compared to the basic Gaussian V_T methodology, they still fail to capture variability in on-current, subthreshold slope and the decorrelation between these figures of merit and threshold voltage. It becomes clear that a more accurate compact modelling strategy is required.

2.7.5 Statistical Compact Models

An accurate statistical compact modelling approach was introduced by Cheng et al. [81] in 2010. The proposed statistical compact model extraction strategy is a two stage process, depicted in Figure 2.12, and requires a statistical set of device transfer characteristic (I_d−V_g ensemble), obtained from measurement or simulation. The initial stage involves extracting a standard 'nominal' compact model which captures device operation in the absence of variability. The 2nd stage begins with a sensitivity analysis of the compact model parameters, which leads to the selection of a subset of the parameters to be used in the statistical compact modelling stage. The selected parameters are then extracted for each device in the statistical device ensemble. This process produces a *compact model library* with a number of accurately extracted transistors, equal to the number of simulated measured devices. An enhanced version of this approach will form the basis of this work and will be further described in Section 3.4.

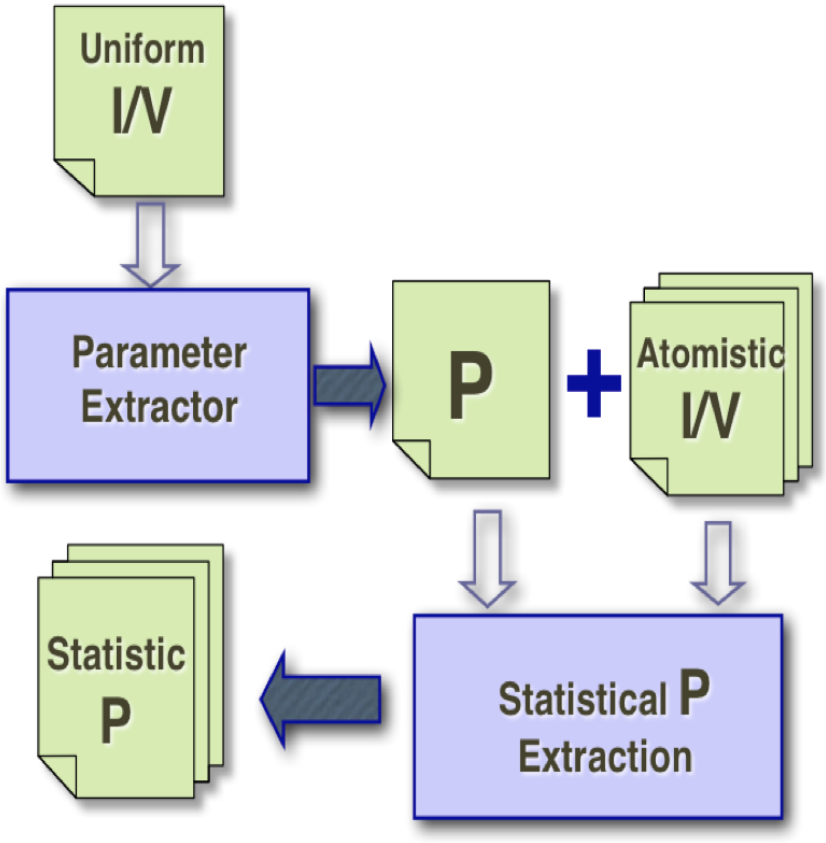


Figure 2.12: Two stage compact model extraction strategy using Mystic compact model extraction tool [81]

2.8 Circuit Simulation Techniques

Circuit simulation is used for multiple purposes: functional logic verification [88], timing closure and power analysis – a combination of which can allow circuit yield calculation [89], an important metric in determining the success of a design and the profitability of a product. If the maximum latch-to-latch delay in a digital system, D_{max} is:

$$D_{max} = T - L \tag{2.2}$$

with clock period T and latch setup time L, then timing closure states that in order for a system to fulfil timing requirements, the maximum delay (often extracted by predictive simulation) has to be smaller than the value of D_{max}. However, in the presence of variability, latch-to-latch delay (D) becomes a statistical distribution, as variable MOSFET performance causes stochastic delays within a system. This distribution can be evaluated through statistical circuit simulation, and can therefore be used to predict yield. Similar analysis can be performed to extract the power performance of a circuit (ideally simultaneously), and a distribution for this power performance can be obtained. The correlated distributions for delay and power can be combined to provide Power-Performance(delay)-Yield (PPY) trade-off information which can determine the practical viability of a design and aid in the optimisation/redesign strategy of a system. There are two main methods for circuit simulation, these are known as static and dynamic [90].

2.8.1 Static Timing Analysis

Custom circuit design and simulation is usually founded on basic circuit building blocks, or standard cells, which encapsulate basic functions and are combined to produce the desired system functionality. Static circuit simulation or static timing analysis (STA) [91] tools like PrimeTime [92] use calibrated look-up tables of cell level delays within a circuit to calculate maximum delay paths and estimate power consumption. Part of the setup process of this technique consists of characterising each cell in the standard cell library at multiple

input slews and output loads to represent all predicted operating conditions of the cell. This is a relatively long and computationally intensive process, but it only has to be done once per technology release life cycle. STA analysis is then based on cell switching profiles and the calculation and addition of delays/power of individual standard cells.

STA handles process variability by using five different sets of MOSFET models within the standard cell look-up table calibration. These models are calibrated to extreme process variability; these 'process corners' are described thoroughly in section 2.7.3.

Traditionally STA analysis has not taken into account statistical variability, however recently techniques like Statistical STA (SSTA) [93] have been proposed to handle the additional statistical variability introduced through the scaling process. Alternatively, system paths which limit performance, known as 'critical paths' [94] can be identified with global STA analysis, after which they can be separated from the rest of the system and thoroughly investigated using more accurate circuit simulation techniques [5]. Recent publications have shown that in the presence of statistical variability, STA calculated critical path importance can change order, and paths which are nominally not critical become important [95]. This effect is shown in Figure 2.13, which depicts critical path occurrence at different levels of statistical variability in a benchmark circuit, showing that at higher variability levels the nominally critical path is only critical $\sim 60\%$ of the time, and a total of 15 paths may be "critical" due to different statistical variably effects.

STA analysis is less accurate than dynamic circuit simulation, and has been shown to be increasingly pessimistic [96] in its predictions compared with real hardware at deep submicron technology nodes. The pessimism in STA analysis is dependant on the type of corners used in the analysis. However, as the timing calculations are several orders of magnitude less CPU intensive than dynamic circuit simulation, they allow the tractable analysis of complex billion transistor systems.

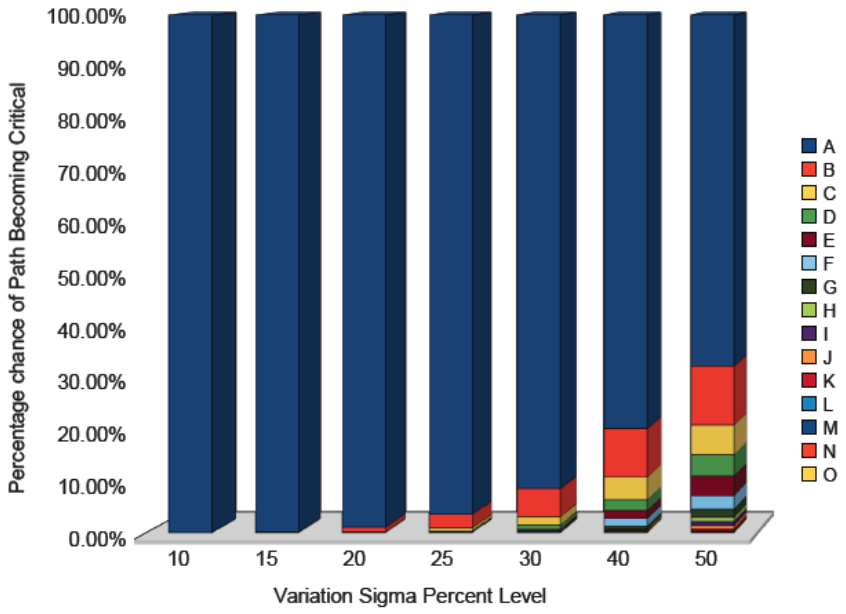


Figure 2.13: Possible critical path as a function of statistical variability, increasing statistical variability shows more paths in the system can be critical [95]. A to O represent different paths within the system which may be critical under the influence of statistical variation.

2.8.2 SPICE circuit simulation

Dynamic circuit simulation is the device level simulation of a circuit using a simulator like the well known Simulation Program with Integrated Circuit Emphasis (SPICE) or SPICE derivatives. These include Cadence's 'Spectre' [97], Mentor Graphic's 'Eldo' [98], Synopsys' 'Hspice' [99] and the open source simulator derived from Berkley's SPICE3, ngSPICE [100]. SPICE simulation involves the assembly of a coupled set of first-order differential equations which model the behaviour of a set of interconnected circuit elements. Time derivatives of these equations are then replaced by integration formulae which discretise time, and transform the nonlinear differential equations at each time point into a set of time independent non-linear algebraic equations. These equations are then iteratively solved using the Newton-Raphson method, until adequate precision is achieved [101].

SPICE simulations require a circuit description file, or netlist, which consists of circuit components and node connections. Netlists are usually extracted from a VLSI design post place/route and layout optimisation steps. These netlists contain all circuit elements, with compact models for MOSFETs as well as parasitic interconnect models to an accuracy level specified by the user. SPICE is capable of a range of analyses including transient, steady state DC and noise analysis. Relative to the accuracy of compact models and interconnect parasitic models, SPICE is the most accurate method of circuit simulation currently available. The two main disadvantages of SPICE simulation are the CPU time intensive nature of SPICE simulation and limitations on the size of circuit which can realistically be simulated.

During the course of this thesis ngSPICE will be used. This open source software is a updated version of Berkeley's SPICE3, and is somewhat slower and more limited than its industrial equivalents. However, it is not limited by license costs, so an arbitrary number of instances can be used in parallel on a high performance computer (HPC) cluster, which dramatically speeds up simulation of large statistical data sets.

A set of simulators which aim to bridge the gap between SPICE and STA are FastSPICE simulators. These simulators enable simulations of larger cir-

cuits than traditionally handled by SPICE simulators, as well as offering a significant speedup of simulation time, although this is achieved at the tradeoff of overall simulation accuracy.

2.9 Variability and SRAM

The impact of statistical variability on circuit performance in deep sub-micron technologies has become an important topic of research. Although the impact of statistical variability on digital logic is important and is predicted to be increasingly relevant in extremely scaled technologies where statistical variability will be further exacerbated [102], the impact of statistical variability is most significant on the Static Random Access Memory (SRAM) system. This is partially due to the fact that a significant portion, over 60% [25], of the chip area in modern System on Chip (SoC) applications can be occupied by SRAM. Unlike digital logic circuits, where timing delay variations along the depth of a pipeline can typically average out, the SRAM system requires methods of correction and redundancy to overcome the SRAM cell's inherent susceptibility to statistical variability [103].

The interest in SRAM stems from the fact that 20-40% of all program instructions reference memory [6], and, as on-chip SRAM is the only sufficiently fast storage system for the quantities of data required by the processor [7], SRAM density, and thus memory size, has had to increase relative to processor speed and number of cores. One of the many advantages of transistor scaling is that the SRAM cell footprint area achieves a reduction of a factor of two per technology generation, as is shown in Figure 2.14, which allows for a potential doubling of SRAM density.

Simulation of an entire SRAM system is limited to STA analysis due to the huge number of transistors ($>$ 6 million in a 1 Mbit SRAM array) present in the system. Although STA can be useful for the purpose of design verification and critical path identification, it cannot accurately capture the effects of statistical variability on an SRAM design as all SRAM cells are assumed to be identical. There are two approaches to this problem; initially SRAM

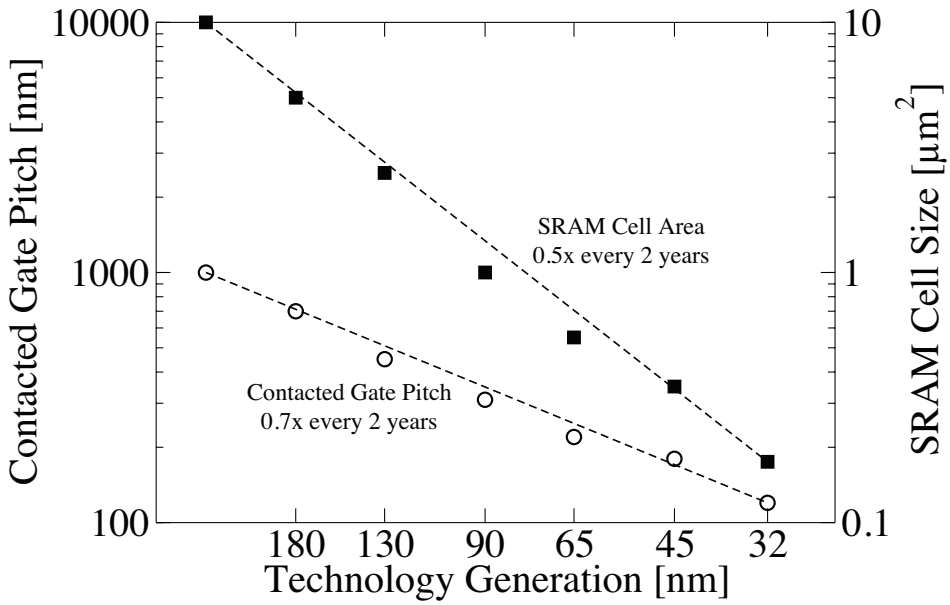


Figure 2.14: Plot of SRAM cell size and gate pitch as a function of technology node [104].

cell design is evaluated on a standard cell level, through static margining techniques like Static Noise Margin (SNM) [105]. These methods, which will be outlined comprehensively in Chapter 5, are used for initial benchmarking of cell performance due to the relatively quick simulation time, as only a single 6 transistor cell is simulated. To evaluate SRAM system performance more accurately, SRAM critical paths can be extracted, including, word line pulse generation, addressing logic, sense amplifier, pre-charge/line buffer and multiplexer circuitry, which can be dynamically simulated with a small number of cells or a single cell. The results of such simulations give a more realistic indication of SRAM cell and system performance, at the trade-off of much longer simulation times than the static SRAM cell simulations.

Process variability is introduced into SRAM simulation and evaluation through the use of 'process corner' simulations [106]. Statistical variability is traditionally introduced through Gaussian V_T simulation [107]. The Gaussian V_T representation of statistical variability is popular due to the relative ease of simulation, as commercial SPICE simulators include Gaussian gener-

ation, as well as enabling statistical margining techniques like Most Probable Vector (MPV) [52], which can drastically reduce simulation time, an important consideration when attempting to assess SRAM performance deep into the tails of performance metric distributions.

It has recently been shown, through measurement and simulation at the 65nm technology node [84], that Gaussian V_T based statistical simulations approaches are not sufficiently accurate to capture the impact of statistical variability on SRAM performance, with DIBL contributing significantly to SRAM stability. The errors introduced through Gaussian V_T simulation of both static and dynamic metrics of SRAM performance at the 20/22nm technology node will be analysed and evaluated as part of this thesis.

2.10 Summary

The main sources and effects of process and statistical variability on MOSFET performance have been discussed in Chapter 2. The importance of including these effects in the circuit design/verification/optimisation steps has been outlined. For this purpose the link between physical device performance and circuit simulation - the MOSFET compact model - has been introduced. The two main types of circuit simulation currently in use in both research and industry have been described, as well as a number of methodologies for introducing variability into these circuit simulation techniques. Finally the importance and need of accurate statistical compact models has been established, and some variability aware compact modelling techniques have been discussed.

Chapter 3

Simulation Methodology

This chapter outlines the simulation methodology employed in this work. The utilised tool chain, co-developed by GSS and the Device Modelling Group, will be employed for the purpose of physical simulation of statistical variability, statistical compact model extraction, and statistical circuit simulation. While the main contributions of this work, including the development of an accurate statistical compact modelling strategy suitable for compact model generation methodologies, the benchmarking of compact model generation methodologies, as well as the study of the impact of statistical variability on SRAM operation, will be detailed in Chapters 4 and 5, the tools with which this work is performed will be described in the remainder of this chapter.

3.1 The Simulation Tool chain

The tool chain employed in this thesis is illustrated in Figure 3.1. It consists of the 3D statistical atomistic simulator GARAND, the statistical compact model extractor Mystic and the statistical circuit simulation engine RandomSpice. Efficient statistical simulation using the software is enabled by the use of a fully automated cluster job submission and management system that allow simultaneous execution of thousands of statistical device simulations, compact model extractions and circuit simulation tasks on clusters with large number of processors. The job submission and management is interfaced to the data-

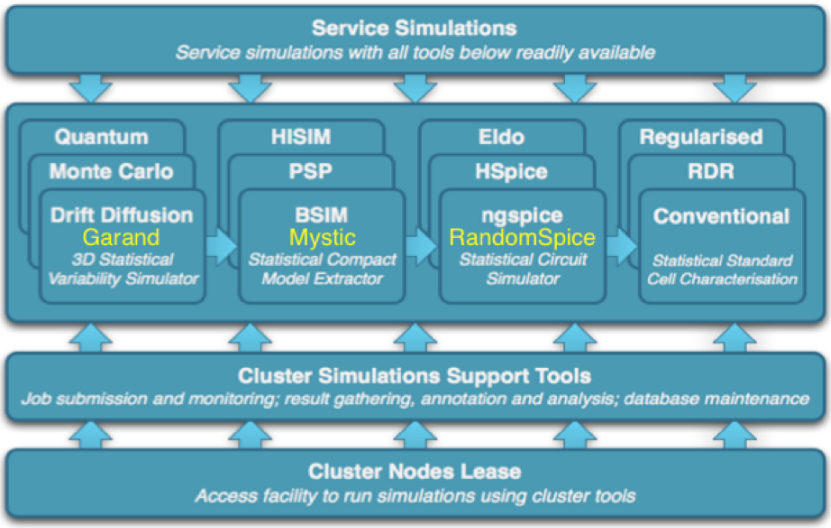


Figure 3.1: Full tool chain flow, from TCAD simulation (left) to statistical circuit simulation (right) [108].

base automatically harvesting and annotating the results of statistical device simulation, compact model extraction and circuit simulation. This tool chain has recently been made commercially available through GSS due to industrial interest in these capabilities which underlines the relevance of this work.

3.2 Physical Simulation of Variability

The purpose of this thesis is to develop an efficient and accurate statistical compact model extraction methodology capable of evaluating the effect of statistical variability on circuit performance. Considerable statistical device variability data is required to fulfil this purpose. This can be obtained in two different ways:

- Comprehensive device measurement: This requires the development of dedicated test structures and entails great expense in terms of development of test structures and measurement, as well as requiring a relatively well developed technology. It also presents difficulties in separating out the effects of process and statistical variability from the measured data,

although methods have been proposed for this purpose [109]. Aside from this, most designs begin well before a technology node is implemented, when statistical measurements are not available. This reduces the effectiveness of this methodology of characterisation.

- Simulation: For a long time TCAD simulation has been the standard for technology design and device development. While simulation in the presence of statistical variability is difficult and computationally intensive, 3D device simulations are required in order to accurately predict the impact of these unavoidable sources of transistor level variability on device and circuit performance. Due to computationally intensive nature of such simulations it is important to have access to massively parallel High Performance Computing (HPC) facilities in order to simulate numerous devices simultaneously.

For the purpose of this thesis physical device simulation will be employed for the the generation of nominal and statistical transistor characteristics used in statistical compact model extraction and circuit simulation. The statistical 3D 'atomistic' simulator GARAND, has been specifically designed for the simulation of statistical variability and reliability in contemporary and future CMOS transistors. The main features of GARAND include: Drift diffusion (DD), Monte Carlo (MC) and Non equilibrium Green's Function (NEGF) modules, the best available physical models [108] allowing atomic scale precision including simultaneous density gradient quantum corrections for electrons and holes, described in Section 3.2.2, mobility models that take into account the discreteness of dopants [50, 54]. It is capable of modelling all sources of statistical variability known to effect modern device performance including RDD [54], LER [51], MGG [64, 110] and Trapped discrete charges [111].

3.2.1 Basic Drift Diffusion Simulation

For this work we will use the Drift-Diffusion (DD) simulation engine of GARAND. This involves modelling the lowest-order transport system obtained, after substantial simplification of the Boltzmann transport equation (BTE). Uni polar

Drift-Diffusion requires that the current continuity equation (Equation 3.1),

$$\nabla.\overrightarrow{J_n} = 0 \tag{3.1}$$

where $\overrightarrow{J_n}$ is the current density vector, is solved self-consistently with the Poisson equation (Equation 3.2)

$$\nabla \cdot (\epsilon \nabla \psi) = q(n - p + N_A^- - N_D^+) \tag{3.2}$$

where ϵ is the intrinsic permitivity, ψ is the electrostatic potential, n and p are the electron and hole carrier densities and N_A^- and N_D^+ are the ionised acceptor and donor doping concentrations. In the simulation of MOSFETs this system of equations is solved for the majority carriers in the device as they dominate device performance.

In the DD approximation the current has two components, drift current and diffusion current. Given for an n-channel MOSFET, these are given by Equations 3.3 and 3.4,

$$\overrightarrow{J_{n,drift}} = qn\mu_n E = -qn\mu_n \nabla\psi \tag{3.3}$$

$$\overrightarrow{J_{n,diff}} = qD_n \nabla n \tag{3.4}$$

where μ represents mobility and D the diffusion coefficient. The inherent simplifications employed in classical DD limit the application and accuracy of this method in the simulation of scaled devices [112]. In order to improve accuracy GARAND applies Density Gradient (DG) quantum corrections to capture the impact of quantum effects in contemporary and future transistors and most importantly in order to accurately resolve the impact of all of the individual dopants in the simulation [58].

3.2.2 Density Gradient Corrections

In order to extend the applicability of the drift-diffusion simulation for aggressively scaled technology nodes GARAND applies DG corrections to capture

non-local quantum effects. Incorporating these effects improves the accuracy of the simulations of devices in the deep submicron regime where the non-local quantum effects begin to have a significant impact on device behaviour. DG introduces a dependence on the carrier density to the current density equations utilised in classical DD simulation. This adds an extra expression to the DD system of equations (Equation 3.5),

$$J_n = qD_n\nabla n - q\mu_n n\nabla\psi + 2qn\mu_n\nabla\left(b_n\frac{\nabla^2\sqrt{n}}{\sqrt{n}}\right) \tag{3.5}$$

where b is a term that expresses the magnitude of the density gradient dependence and has the general form $b = \frac{\hbar^2}{4m^*qr}$). The inclusion of density gradient corrections captures some of the lowest order quantum effects like confinement and even, to some extent tunnelling [112]. The combination of drift-diffusion and density gradient has been shown to be sufficiently accurate for the simulation of bulk silicon devices to the 20/22nm technology generation [58]. Devices with shorter channel lengths or more complex materials like Silicon Germanium ($SiGe$) require *ab initio* Monte-Carlo simulation to model localised transport effects which can have a significant effect on on-current [113].

The combination of of Poisson's equation (Equation 3.2), the current continuity equation (Equation 3.1) and the density gradient equations are solved self consistently using a modified Gummel iteration method [114]. In the simulator, the discrete Poisson 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.

3.2.3 Including Variability Sources with GARAND

As has been outlined in Chapter 2, the main sources of statistical variability relevant to contemporary and future CMOS transistors include Random Discrete Dopants (RDD), Line Edge Roughness (LER) and Polysilicon/Metal Gate Granularity (P/MGG). The methodologies through which these sources of variability are introduced into the 3D device simulation domains are described below.

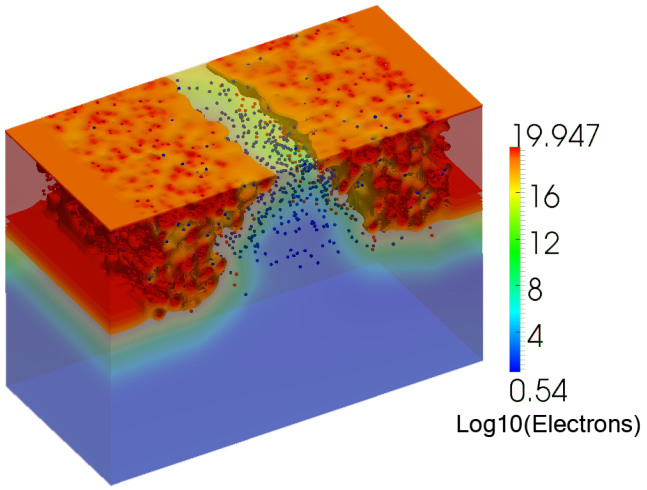


Figure 3.2: An example of a 25nm bulk n-channel device with RDD, the scale is logarithmic.

Random Dopants - The physical position of the random dopants are based on the complex doping profiles and process methodologies specific to each individual device technology. GARAND employs a method first described by *Frank et al.* [115] where random numbers are generated for each silicon lattice site to determine if a dopant is present or not. The rejection technique that selects whether a dopant should be placed at a particular lattice is site based on the probability given by the ratio of the doping and *Si* concentration at that site. Once dopant positions have been determined GARAND employs a charge assignment scheme which spreads the single dopant charge onto the surrounding mesh nodes. This is done using the Cloud-in-Cell (CIC) approach, in which the fraction of the dopant charge assigned to a particular mesh point corresponds to the distance between the dopant and mesh point. An example of the electron concentration contours of a 20/22nm bulk n-channel device generated with RDD is shown in Figure 3.2, where the blue points correspond to acceptors in the bulk and the red circles correspond to donors from the source/drain doping.

Line Edge Roughness is introduced using a 1D Fourier synthesis method first presented in [116]. This involves the generation of lines from a Gaussian or exponential power spectrum [117], which are fitted to measured or published

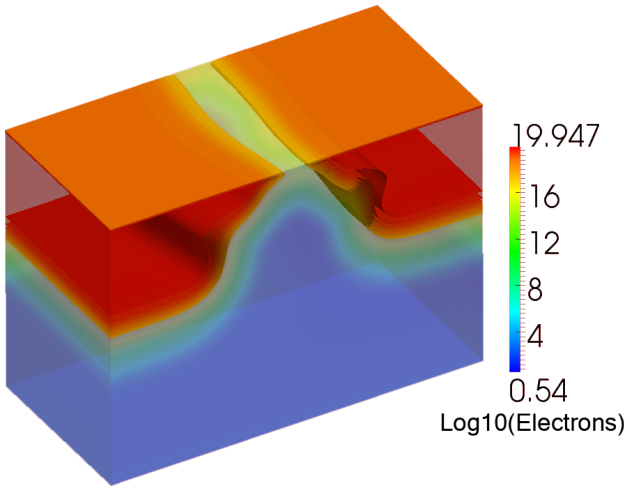


Figure 3.3: An example of a device with LER, the non-uniform gate shape is represented by the non-uniform source/drain shapes.

LER data. LER for a given technology is characterised by the line being generated by the auto-correlation function, the RMS amplitude (Δ) and the correlation length (Λ). An example of a device in the presence of LER is shown in Figure 3.3, where it can be seen that the source doping follows the variable gate edge and causes a difference in effective channel length along the channel.

Polysilicon/Metal Gate Granularity are generated based on measurements and published data [118], with a randomised grain pattern assigned to each different transistor. Grains are assigned with one of two possible work functions based on a pre calculated probability, and in regions below grain boundaries Fermi-level pinning is introduced due to interface states. An example of a device with variable MGG is shown in Figure 3.4. The effect of the different grain work function on the surface of the device channel is clearly shown in the figure, with the grain boundaries forming percolation paths across the channel.

In order to evaluate the performance of realistic devices, all the above variability sources in have to be included in GARAND simulations.

3.2.4 Cluster Computing Facilities

A large number of simulations, or samples, have to be performed to evaluate the effect of variability on a specific technology. This is due to the fact that

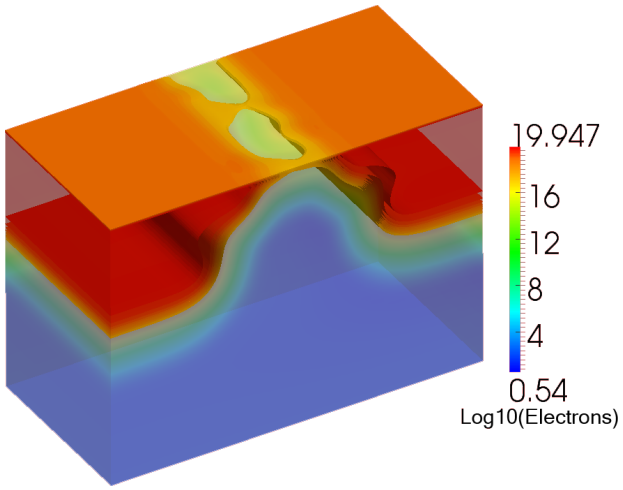


Figure 3.4: An example of the impact of MGG, two current paths are formed along grain edges.

the standard error of the sample distribution is a function of the square root of the number of samples, as shown in Equation 3.6:

$$E_{std} = \frac{\sigma}{\sqrt{n}} \tag{3.6}$$

Where E_{std} is the standard error, σ is the standard deviation and n is the sample size. This means that for a device with mean threshold voltage of 200 mV and $\sigma = 70$ mV, 1000 samples give a standard error of 2.2 mV and 10,000 simulations give a standard error of 0.7 mV.

As previously stated full 3D numerical simulation is computationally intensive, with the Id-Vg characteristics of a single device taking up to 8-9 hours, and a requirement for datasets in the range of 1000 to >10,000 devices, simulation on a single machine would become prohibitive. Utilising HPC facilities allows for massive parellisation as there is no interdependence between the different device simulations. The speedup of the simulations is relative to the number of free processors within the HPC cluster. This enables simulation ensemble sizes of >10,000 over a relatively short period of 2-3 days. HPC facilities can also be employed in the statistical model extraction and circuit simulation stages of the simulation process.

3.3 Extraction using Mystic

The resultant simulation data from GARAND can be directly utilised for the purpose of compact model extraction. The extraction tool, Mystic, provides a scriptable language with access to multiple optimisers including Levenberg-Marquardt [119], Bounded Trust Region [120] and derivative free optimisation methods like Constrained Optimisation BY Linear Approximation (COBYLA) [121].

Developing extraction strategies using Mystic relies on a deep understanding of the behaviour and limitations of a compact model and the underlying device physics. The inherent flexibility of the tool allows for multiple extraction strategies which achieve the required statistical compact model accuracy using different parameters and device operation targets. One of the principal components of this work will be the development of an accurate and reliable compact modelling strategy which is capable of capturing statistical variability, and producing data suitable for use with advanced compact model generation strategies. This extraction strategy will be described in Section 4.2.1, with extraction results analysed in Section 4.3.

3.3.1 Nominal Compact Model Extraction for BSIM4

The nominal compact model extraction stage involves extracting a base model into which variability can be injected. This model must take into account drain bias dependence, gate bias dependence, temperature dependence as well as channel length and width dependence. We refer to this as the *uniform model* as it represents idealised device performance under uniform doping conditions as well as idealised device geometry. The model can be based on average device measurement from silicon or preferably, TCAD simulation data. When the devices in a foundry Process Design Kit (PDK), used for physical design, are based on TCAD results, they can be made available for design use significantly before the technology is fully developed and physical silicon is available and this can give a significant competitive advantage to early technology adopters.

For the purposes of this work we extract uniform models based on TCAD

Drain Bias	$(1.00V, 0.05V)$
Channel Lengths	(200 nm, 150 nm, 100 nm, 50 nm, 40 nm, 30 nm, 25 nm, 20 nm)
Body Bias	$(0\,V, -0.2\,V, -0.4\,V, -0.6\,V, -0.8\,V, -1.0\,V)$
Device Widths	(30 nm, 25 nm, 20 nm)
Temperature	(-40°C, -27°C, 125°C)
Capacitance	C-V fitting data

Table 3.1: Required data for accurate uniform compact model extraction.

simulation which has been performed using GARAND. These simulations are based on a Template device, fully described in Chapter 4. Variability is not considered a this stage of the extraction process so continuous 2D simulations are performed in order to extract an accurate uniform model. Simulations are required at high drain and low drain bias, at multiple channel lengths, at various body bias levels and different temperature levels, shown in Table 3.1. The simulations capture the transfer characteristics (I_DV_G) and output characteristics (I_DV_D) of the device.

Before model extraction can begin, BSIM4 requires that basic structural and physical parameters are supplied. These parameters are introduced and described in Table 3.2. In order to aid the nominal optimisation process we also introduce sensible initial conditions for some of the more physical BSIM4 parameters. For example the BSIM4 parameter $VTH0$ is initialised at the long channel low drain threshold voltage for the device. $RDSW$, the source drain resistance, is extracted using the 'TMC' method [122], which utilises multiple channel length simulations in order to plot the transistor resistance as a function of the channel length. The source drain resistance is the estimated using a projection to a channel length of 0nm.

3.3.1.1 Target Extraction Strategy

A combination of local optimisation and a group extraction strategy is employed in order to obtain a complete nominal set of BSIM4 parameters which accurately capture the behaviour of the target device over a large range of operation conditions. This is based on the simulation of a set of transistors with continuous doping profiles and different channel lengths, focusing on those

Input Parameter Name	Physical Meaning
$TOXE$	Gate oxide thickness and dielectric constant
$NDEP$	Doping concentration in channel
$TNOM$	Temperature at which data is taken
L_{drawn}	Mask level channel length
W_{drawn}	Mask level channel width
XJ	Junction depth

Table 3.2: Prerequisite input parameters prior to extraction process.

critical to long channel behaviour, the threshold voltage in the short channel regime, and drain current response in the presence of high fields.

It is the goal of the extraction strategy to retain the physical relevance of as many of the compact model parameters as possible. If possible the physical parameters introduced in the early stages of the extraction are unchanged and most of the optimisation process is achieved through the 'fitting' parameters available. In order to achieve optimal results, parameters are *targeted* at the specific region of device operation where they are expected to have the greatest effect.

The challenges involved with the development of an accurate nominal compact modelling strategy are strongly related to the specific technology in question and the compact model implementation in use. Due to the fact that there are as many effectively 'physical' parameters as phenomenological 'fitting' parameters in the BSIM4 implementation it is often difficult to disentangle correlated parameters in order to provide a stable solution. The complexity of the physical effects in modern short channel transistors negate any of the long channel simplifications that have previously been applied through years of compact model development, with complex quantum mechanical and non-equilibrium transport effects having a serious impact on device performance. While advanced models like BSIM4 have the ability to model most of these effects, it is difficult to correctly identify their relative importance upon transistor performance.

Another challenge arises from the large number of targets in the optimisation process, as well as the complex correlation between these targets. All of

this combines to produce a huge and highly complex parameter search space with multiple minima and many possible intermediate solutions. Steering a numerical optimisation to an global minimum solution with physically relevant parameters which are suitable for statistical extraction is a complex process which usually involves some amount of compromise and many iterations. The models extracted in the course of this thesis have been specifically targeted at the nominal channel length of 25nm, and are designed to be robust in the range of 20-50nm in order to capture LER effects. However, deviation from the nominal behaviour increases as devices depart further from the expected channel length. A simple solution to this problem could be the extraction of multiple nominal models for different channel lengths as required. As the principal target of the simulations performed in this thesis will involve SRAM cell simulation, where the channel lengths are close to the nominal channel length of 25nm, this will not be considered.

3.4 Statistical Compact Model Extraction

Statistical compact modelling in relation to statistical variability introduced by the discreteness of charge and matter has been pioneered by Cheng et al. *[81]*. This is a two stage process, depicted in Figure 3.5. The first stage is the equivalent to the modified standard compact modelling process previously described in Section 3.3.1. It involves extracting a uniform model into which variability will be injected.

The second stage begins with an analysis of the available parameter set. A subset of the compact model parameters are chosen based on a sensitivity analysis, and are re-extracted for each device in an ensemble of measurements or simulations. This creates a statistical library of devices which can be used directly for circuit simulation in the form of lookup tables. The accuracy of these models depends on the number of parameters extracted and the specific extraction strategy employed. In the strategy proposed by *Cheng et al. [81]* statistical parameter extraction is based on a global optimisation using a least squares algorithm. Compact models extracted using this methodology

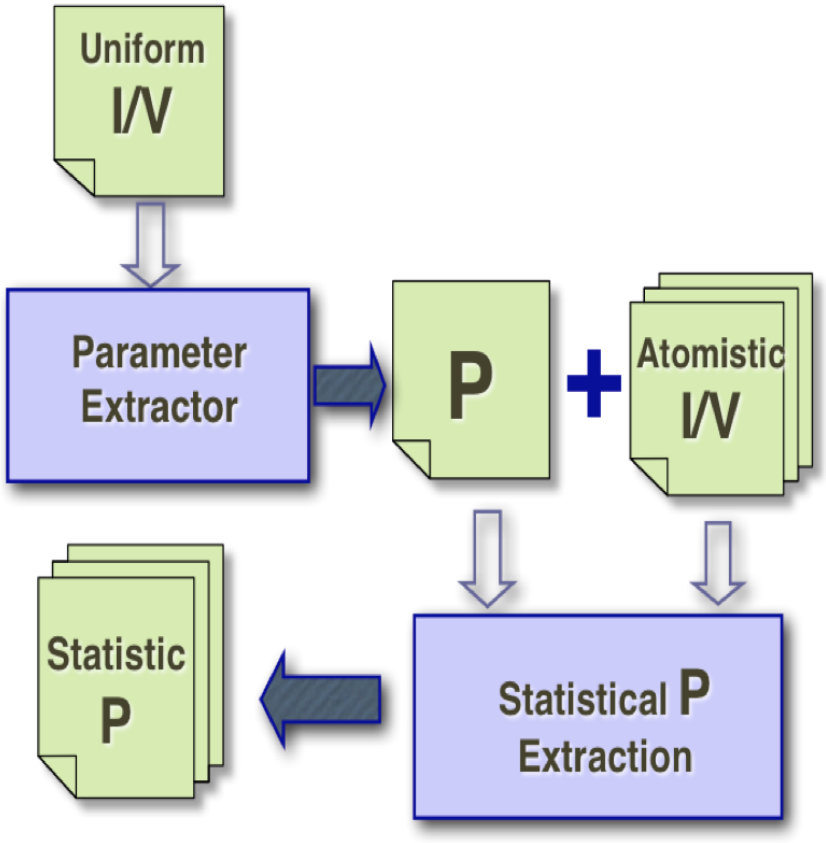


Figure 3.5: Two stage compact model extraction strategy using Mystic compact model extraction tool [81].

show good agreement with target data, however they do not always capture device performance figures of merit accurately. In addition, the extracted parameter distributions obtained are often unsuitable for advanced accurate compact model generation strategies.

The limitation of the direct use of extracted statistical compact models is due to the limited number of devices physically simulated/measured. This can cause problems with sub sampling in Monte-Carlo based circuit simulations, introducing un-physical artefacts in simulated circuit performance metrics. This limitation can be overcome through the use of compact model generation methodologies. Generation strategies use extracted statistical parameter distributions and attempt to generate new randomly selected devices which have parameter distributions that replicate the extracted parameter distributions obtained from direct extraction, whilst taking into account correlations between the extracted parameters. If a generation strategy is accurate, an essentially infinite ensemble of devices, which exactly reproduce the statistical performance of the target sample, can be generated for the purposes of circuit simulation. For the current generation strategies to be viable, extracted parameter distribution must be uni-modal and continuous. One of the major components of this work is to develop a physically based statistical compact modelling strategy, which retains the physical meaning of compact model parameters whilst producing viable distributions for advanced model generation strategies. This strategy will be introduced and discussed in Section 4.2.1. Generation strategies will be further discussed in Section 3.5.

3.5 Statistical Compact Model Generation

The motivation behind the development of a figure of merit based statistical extraction approach becomes clear when one considers the generation of arbitrary statistical compact models. After an ensemble of statistical compact models has been extracted, it is crucial to be able to accurately propagate this information to circuit simulation. In order to avoid problems associated with sub sampling (as there will always be a finite number of devices which

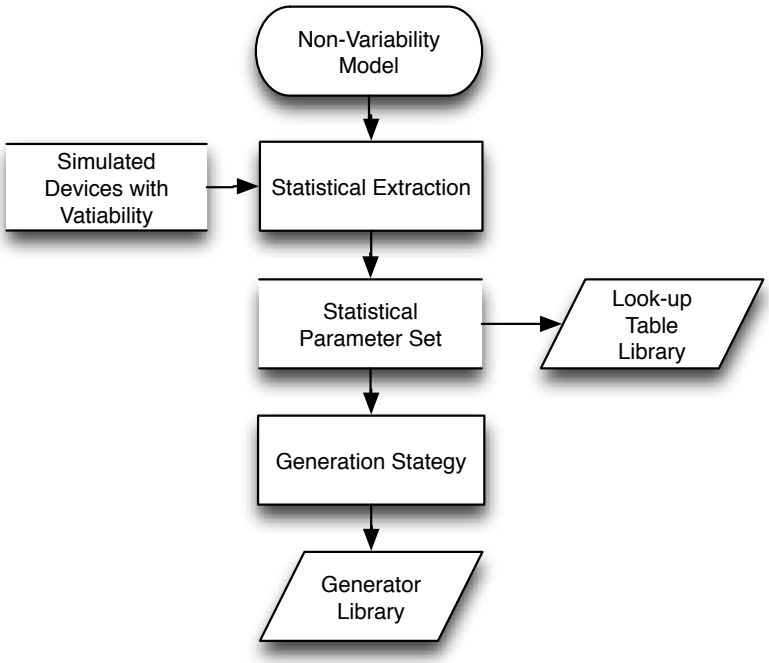


Figure 3.6: The flow from non-variability compact model to variability aware compact model generator libraries.

can be simulated or measured, this is illustrated in Section 4.5) and to facilitate advanced statistical enhancement techniques, it is desirable to have the ability to generate an effectively unlimited ensemble of devices that accurately reproduces the statistical performance of the underlying technology. For this to be achieved, a generation strategy must be implemented which captures the complex distribution of all extracted statistical parameters, as well as the correlations between them. The flow diagram in Figure 3.6 shows the required steps to producing a generator library.

Four compact model generation approaches which will be considered in this work are Gaussian V_{T}, skewed Gaussian V_{T}, Principal Component Analysis (PCA) [123, 124] and the Non-Linear Power Method (NPM) [125]. For the PCA and NPM generation strategies to be reliable, the distributions of extracted parameters must be continuous and uni-modal. This requirement can be problematic when using the method of statistical parameter extraction proposed by Cheng et. al. [81], where parameters may be unintentionally used to fit physical effects for which they were not designed, introducing non-physical correlations between parameters. The physically based extraction approach developed as part of this research, which will be extensively outlined in Section 4.2.1 avoids these problems by targeting parameters specifically to the regions of device operation where they have a significant impact, and guarantees representative and well defined parameter distributions as each parameter is specifically selected and optimised to a single physical effect, assuming the extraction target is uni-modal.

Complex higher order compact model generation strategies are based on the fundamental assumption, that extracted parameter distributions provide an accurate representation of the population distribution of devices. In other words, the behaviour of the generated devices follows the distribution of the simulated and extracted devices. If the sample on which the strategy is based is too small, or is not a good representation of the population, the generated devices will also be not representative of the underlying population. The error in the sampling mean can be calculated from Equation 3.6, we can also estimate the error in the higher order moments using the estimators shown in Equations 3.7,3.8 and 3.9 [126, 127]

$$E_\sigma \approx \frac{1}{\sqrt{2(n-1)}} \tag{3.7}$$

$$E_{\gamma_1} \approx \sqrt{\frac{6}{n}} \tag{3.8}$$

$$E_{\gamma_2} \approx \sqrt{\frac{24}{n}} \tag{3.9}$$

where $\gamma_1$is skewness, defined as

$$\gamma_1 = \frac{\mu - \nu}{\sigma} \tag{3.10}$$

where μ is the mean,ν is the median and σ is the standard deviation, and γ_2 is kurtosis, defined as

$$\gamma_2 = \frac{\mu_4}{\sigma^4} \tag{3.11}$$

where μ_4 is the 4th moment about the mean.

If we consider a W = L = 25nm device with mean threshold voltage of 200 mV and a standard deviation of threshold voltage of 70 mV, the standard error in the mean, variance, skewness and kurtosis threshold voltage as a function of sample size can be seen in Table 3.3. In order to minimise the error in the extracted parameter distribution moments, we perform this work with a sample size of 10,000 n-channel and p-channel devices.

3.5.1 Gaussian V_T

Gaussian V_T is the traditional way of introducing statistical variability in circuit simulation. This is a simple way of estimating the first order effects of variability. Gaussian V_T is a popular approach as it can be easily implemented, and enables simplified statistical analysis.

The Gaussian V_T generation methodology only requires the uniform model extraction set to be completed. Once a uniform model has been extracted,

Sample Size	Std. Error of Mean	Error in Variance	Error in Skewness	Error in Kurtosis
100	3.5%	7.1%	24.6%	49.0%
200	2.5%	5.0%	17.3%	34.6%
500	1.6%	3.2%	11.0%	21.9%
1000	1.1%	2.2%	7.7%	15.5%
10,000	0.35%	0.7%	2.4%	4.9%

Table 3.3: Standard error of mean and an estimate of error in variance, skewness and kurtosis of the threshold voltage of devices as a function of sample size.

a distribution of the threshold voltages of the statistical device data can be extracted. The assumption is then made that the threshold voltage data is Gaussian distributed so is completely described by its first two moments, the mean (μ_{VT}) and the standard deviation (σ_{VT}). Generated compact models are then centred around the BSIM4 parameter $VTH0$, with the standard deviation σ_{VT} calculated from the underlying technology.

The Gaussian V_{T} methodology is popular for multiple reasons, principally is the fact that a threshold voltage shift encapsulates the first order effect of statistical variability, a method entitled *idealisation of statistical chaos in a single variable* [52]. It is demonstrated in Section 4.2.2 that Gaussian V_{T} is easily introduced through a single BSIM4 compact model parameter, and most commercial SPICE -like simulators include Gaussian random number generation as a feature for statistical simulation, and due to the computationally light nature of this methodology, there is little overhead associated with circuit level Monte-Carlo simulation. Another motivating factor for the use of this method is that only one measurement is required per device, and only a small number of measurements are required to capture the two moments. Table 3.3 shows that 1,000 samples are adequate for a 1% error in the mean and 2% error in the standard deviation of the threshold voltage distribution. The small number of measurements required for the application Gaussian V_{T} generation has the effect reducing the overhead, in both cost and time, for statistical categorisation. Aside from this, the assumption that Gaussian V_{T} sufficiently captures

variability effects accurately, enables statistical simplifications and margining techniques which can drastically reduce the number of simulations required.

An extension of this methodology, which will also be considered in this thesis is generating models using a skew normal distribution, which allows the third moment of the input data to be modelled. The skewness of the generated distribution is again directly calculated from the simulated device ensemble and will reproduce the skewness of the sample distribution.

3.5.2 Uncorrelated Compact Model Parameter Generation

The uncorrelated compact model parameter generation approach involves generating models based on statistical compact model extraction. The approach aims to capture the individual parameter distributions, but not the correlations between the distributions of parameters. The most basic methodology involves calculating the mean and standard deviation of each of the selected statistical parameters and reproducing them using individual Gaussian distributions. The accuracy of this method can be improved by extending the generated distributions to the higher moments (skewness and kurtosis) of each individual extracted parameter distribution.

A significant problem associated with uncorrelated parameter generation approaches is that, while they have the ability to capture extracted parameter distributions correctly, not taking the correlations between parameters into consideration leads to the generation of non-physical devices which do not fall within the physical range of behaviour of the underlying technology. This is particularly problematic for the figure of merit based extraction as the extracted parameters closely follow the underlying physical effects, which intrinsically contain the correlations between the physical figures of merit.

The errors introduced through the uncorrelated parameter generating approaches have been demonstrated [124] and as a result they will not be specifically considered in this work. However, due to the prevalence in industry of Gaussian V_{T} it will be compared with more advanced generation strategies

3.5.3 Principal Component Analysis

The Principal Component Analysis (PCA) methodology has been proposed as a statistical method for compact model generation by *Kovac et al.* [123, 124], with promising results. PCA methods assume that variables follow a Gaussian distribution, and match the extracted parameter mean and standard deviation while retaining the correlations between parameters. This is done by finding the eigenvectors and eigenvalues of the covariance matrix of the random variables such as:

$$AV = \lambda V \tag{3.12}$$

where V represents the eigenvectors and λ are the eigenvalues of the covariance matrix of the extracted model parameters A. PCA can then transform uncorrelated Gaussian parameters using eigenvectors calculated to match the variance and correlations of the input data as shown in Equation 3.13.

$$X = VZ \tag{3.13}$$

where Z represents the uncorrelated Gaussian variables and X is the input data. The resultant variates follow a Gaussian distribution with a mean of 0 and variance of A. The final step is to scale X to the mean of the original parameter. PCA is effective and very stable as long as the extracted parameter sets are Gaussian or near-Gaussian. In order to capture complex short channel effects in advanced device architectures where extracted parameter distributions can be highly non-Gaussian including large amounts of both skewness and kurtosis, higher order generation methods are required.

3.5.4 Non-Linear Power Method (NPM)

NPM is a moment matching technique, which can be employed in order to accurately reproduce the first four moments of individual parameter distributions, producing a more accurate fit for non-Gaussian distributed parameters through mean, standard deviation, skew and kurtosis, whilst retaining the correlations between random variables. NPM is based on the transform-

ation of a normal (Gaussian) variable (Z_i) with zero mean and unit variance into a non-normal variable (Y_i) through a transform $Y_i = c_i^T Z_i$ where $c_i^T = (c_{0i} + c_{1i} + c_{2i} + c_{3i})$ are unknown constants and $Z_i^T = (1 + Z_i + Z_i^2 + Z_i^3)$. Using a 3rd order polynomial for Z_i allows control of the mean, standard deviation, skewness and kurtosis of the non-normal variable. Expressions are extracted for Y_i to determine the constants c_i^T. Substituting the central moments of Z_i into the moment formulas of Y_i, a system of non-linear equations is constructed. The system of equations is shown below

$$E[Y_i] = c_i^T E[Z_i] \tag{3.14}$$

$$VAR[Y_i] = E\left[\left(c_i^T Z_i\right)^2\right] - \left(E\left[c_i^T Z_i\right]\right)^2 \tag{3.15}$$

$$\gamma_{1i} = \frac{E\left[\left(c_i^T Z_i\right)^3\right] - 3E\left[\left(c_i^T Z_i\right)^2\right]\left(E\left[c_i^T Z_i\right]\right) + 2\left(E\left[c_i^T Z_i\right]\right)^3}{\left(VAR[Y_i]\right)^{\frac{3}{2}}} \tag{3.16}$$

$$\gamma_{2i} = \frac{E\left[\left(c_i^T Z_i\right)^4\right] - 4E\left[\left(c_i^T Z_i\right)^3\right]\left(E\left[c_i^T Z_i\right]\right) - 3\left(E\left[\left(c_i^T Z_i\right)^2\right]\right)^2}{\left(VAR[Y_i]\right)^2} \tag{3.17}$$

$$+\frac{12E\left[\left(c_i^T Z_i\right)^2\right]\left(E\left[c_i^T Z_i\right]\right)^2 + 6\left(E\left[c_i^T Z_i\right]\right)^4}{\left(VAR[Y_i]\right)^2}$$

where $E[x]$ is the mean value, $VAR[x]$ is the variance, γ_{1i} is the sample skewness and γ_{2i} is the sample kurtosis.

The system of equations is simultaneously solved through root finding to provide the constants c_i^T. In order to retain the correct correlations between the parameters it is necessary to calculate the intermediate correlation matrix between the non-normal random variables Y, this is done using Isserlis's theorem [128] and are calculated through the following expression:

$$\rho_{Y_iY_j} = E\left[c_i^T Z_i c_j^T Z_j\right] \tag{3.18}$$

where $\rho_{Y_iY_j}$ is the correlation between two model parameters and $\rho_{Z_iZ_j} = E[Z_iZ_j]$ is the intermediate correlation between two standard normal random variables. A total of $(N-1) \times \frac{N}{2}$ polynomial equations need to be solved, where N is the number of compact model parameters. The required variable Y_i is then generated using a combination of Singular Value Decomposition (SVD) of the intermediate matrix and the constants calculated by the NPM approach. The resultant generated distributions match the first four moments of the extracted parameter distributions as well as all the cross correlations between these parameters, therefore giving a clear theoretical advantage over PCA. The predicted effects of both these generation strategies on the accuracy of model generation and circuit simulation will be explored in Section 4.6.

Another possible approach is to use the Generalised Lambda Distribution (GLD, see [129]) which provides the ability to match a larger number of arbitrary distributions without a heavy reliance on their moments, employing a goodness-of-fit test. One problem with this method is the increased computational intensity of the parameterisation of the GLD, as well as the dependence on the goodness-of-fit test employed. Whilst an interesting alternative, the GLD based generation method will not be considered in the course of this work.

3.6 Circuit Simulation using RandomSpice

RandomSpice is an advanced statistical circuit simulation engine. It address the challenges associated with statistical circuit simulation in the presence of process and statistical variability, which necessitate accurate predictions of the statistics of transistor and circuit characteristics far into the tails of their distributions. RandomSpice provides the capability to probe very rare circuit instances, which are critical to yield calculations of multi-million transistor circuit blocks such as SRAM. At its core, RandomSpice is a Monte Carlo simulation engine, supporting multiple SPICE backends, including Synopsys

HSPICE, Mentor Graphics' Eldo and the open source simulator ngSPICE. RandomSpice also allows very large-scale parallel simulations to be performed on High-Performance-Computing (HPC) clusters which greatly simplifies the production of the very large simulation ensembles required to accurately study the impact of variability on design. In order to accurately reproduce statistical variability in MOSFET characteristics, compact model technology libraries are specifically generated for use with RandomSpice. Compact model parameter extraction can be directly from measurement or from TCAD simulation results. RandomSpice also supports compact model generation methods including Gaussian V_{T}, PCA and NPM.

3.6.1 Monte-Carlo Circuit Simulation Methods

The RandomSpice circuit simulation methodology involves the creation of a basic template SPICE netlist. MOSFETs which are to be randomised are labelled with a keyword which is compact model library specific. RandomSpice then generates and simulates the required number of randomised circuit instances with randomly generated compact models for each of the tagged transistors. For the purposes of this work ordinary Monte-Carlo simulation [130] will be performed, where no assumptions are made about the distributions of device or circuit performance metrics. One problem with Monte-Carlo simulation is that although it captures the distribution well close to the mean, it can take a very large number of simulations (on the order of 1-10 Million) to accurately capture the tails of output distributions to $5-6\sigma$. While, due to the increased transistor size and the opportunity for statistical averaging, it is not necessary to simulate normal digital logic circuits to this extreme, it is key to be able to simulate this far in the tails for SRAM analysis. This is due to the high density of SRAM cells (10^7 cells and $> 6 \times 10^7$ transistors in a $10Mb$ SRAM block) and the dependence of SRAM operation on balanced transistor pairs.

In order to evaluate SRAM performance to $5-6\sigma$ we fit SRAM performance figures of merit distributions from simulation with sample sizes in the range of 100,000-500,000 using the Freimer, Mudholkar, Kollia and Lin's (RMKL) parameterisation of the Generalised Lambda Distribution (GLD) described in

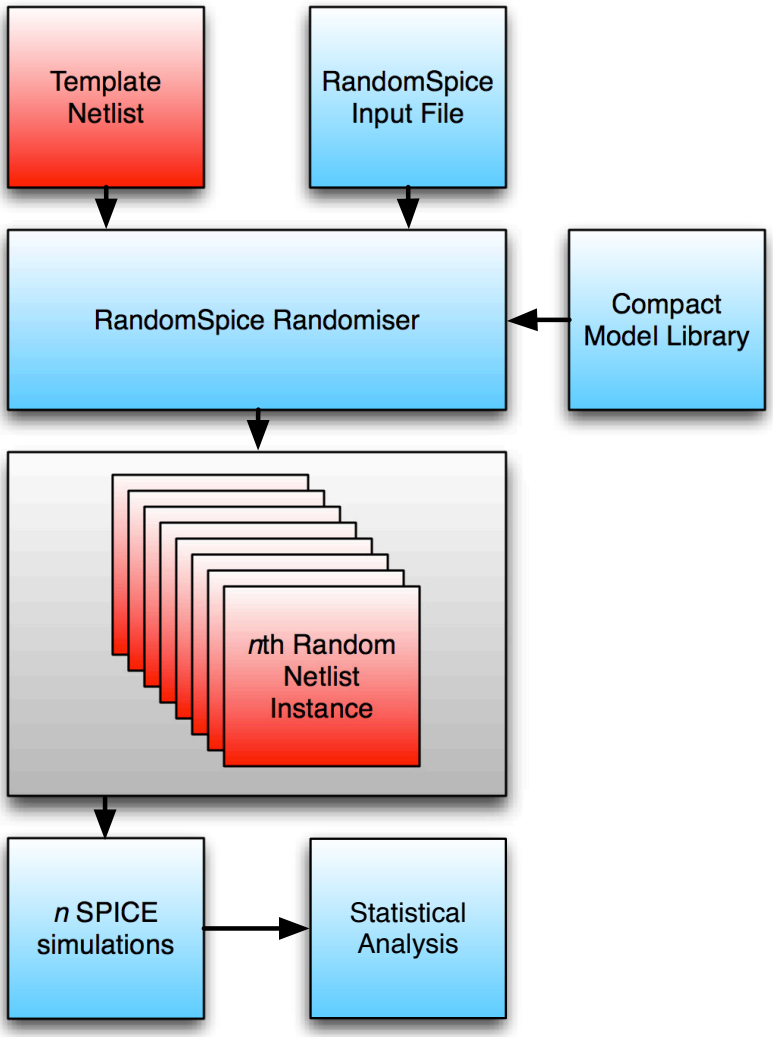


Figure 3.7: RandomSpice flowchart.

[131, 132]. Once a GLD fit has been extracted we have an analytical representation of the figure of merit and the performance at the required sigma value can be calculated. This methodology assumes that the simulated data is representative of the population distribution of the figure of merit, and that no higher order physical effects impact the tails of the distribution. The GLD methodology allows for a yield estimate based on a single figure of merit for device performance,

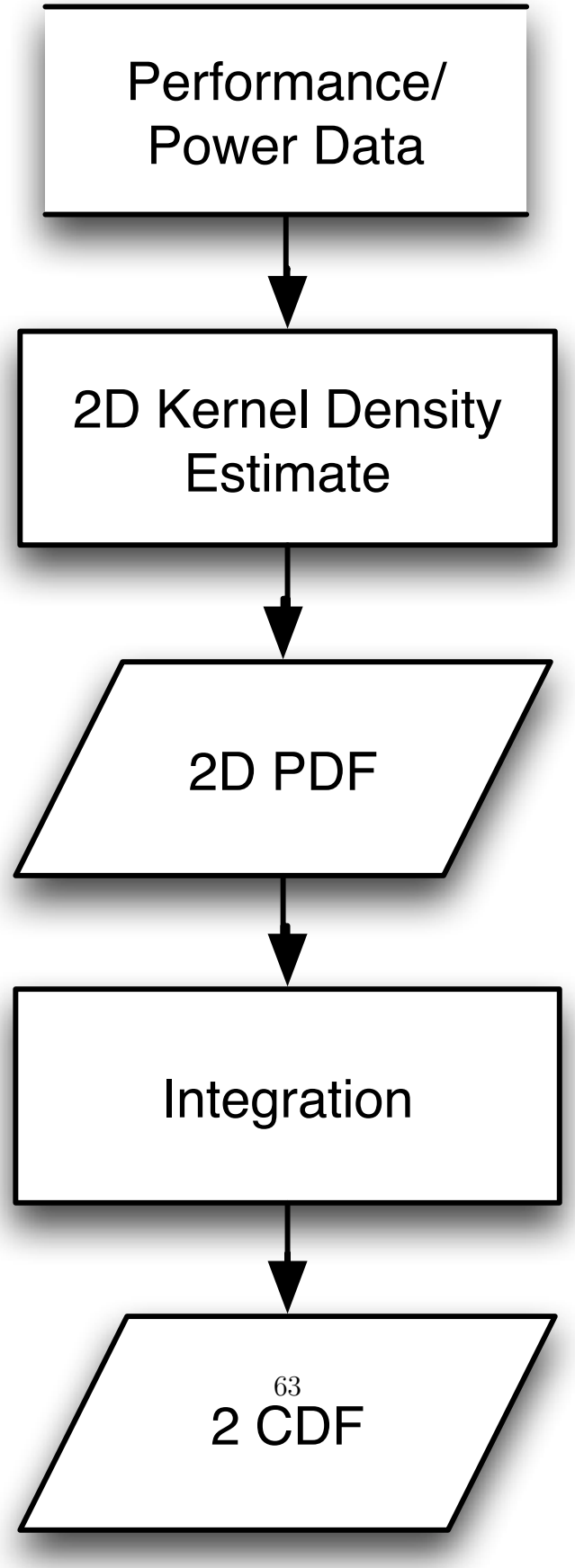
3.6.2 Performance/Power/Yield Analysis

In order to perform Performance/Power/Yield (PPY) optimisation it is required to extrapolate the cumulative distribution function (CDF) derived from the simulated data points in two correlated dimensions. This is achieved through a 2D kernel density estimate (KDE) [133]. This process replaces each data point with a 2D normal Gaussian distribution, these 2D Gaussian distributions then add to construct a 2D probability density function (PDF). Integrating along the 2D PDF we obtain the 2D CDF . The equi-potential lines along the 2D CDF represent the equi-yield contours for those data points. In order to perform PPY analysis of this form it is important to have the full distributions of performance and power as obtained by Monte-Carlo analysis.

The flow from performance/power data to yield estimation is show in Figure 3.8. This will be explored further in Chapter 6.

3.7 Summary

In this chapter the proposed flow from physical device level simulation to statistical circuit performance evaluation in the presence of variability through the intermediate like of statistical compact model extraction and generation has been outlined. The ability to simulate a large number devices under the effects of statistical variability, combined with the level of accuracy in the compact model extraction and generation, gives the methodology an advantage over most of the current statistical variability aware circuit simulation methods. Aside from this some post simulation analysis techniques are introduced, in-


Performance/
Power Data
2D Kernel Density
Estimate
2D PDF
Integration
63
2 CDF

cluding GLD projection and KDE analysis, which can be used to evaluate the success of a given design. Subsequent chapters show the application of this methodology in an advanced 20/22nm bulk technology generation, with evaluations on both SRAM and digital logic designs. In the next chapter we will focus on improving the methodology for statistical compact model extraction targeting the most important MOSFET figures of merit.

Chapter 4

20/22nm CMOS Technology Extraction Results

In order to advance the compact model extraction methodology under modern and industrially relevant high variability conditions, as well as to investigate the effect of statistical variability at a future technology node, we focus on a 25nm physical gate length bulk MOSFET, which is representative of the 20/22nm technology generation. The difficulties involved in developing 20/22nm bulk technology with a sufficient performance and yield were highlighted by Intel's decision to switch to FinFETs at this technology generation [23], due to the fact that these transistor dimensions push bulk MOSFET technology close to its physical limits. This makes the 20/22nm bulk technology a challenging application for the statistical compact modelling strategy.

This Chapter will introduce the design of a template transistor representative of the 20/22nm technology generation. The results of the physical 3D simulations of these devices will first be used to obtain uniform, and then statistical, compact models. The physical extraction strategy developed as part of this work will be outlined and the results will be benchmarked against simulated transistor data and standard extraction methods. The extracted compact models will then be used to inform the selection of an appropriate strategy for statistical compact model generation using the Gaussian V_T, PCA and NPM, strategies outlined in Chapter 3. The statistical behaviour of transistors gen-

erated with these strategies will be compared to the statistical behaviour of the simulated transistors. The work described in this chapter will result in the creation of a set of 20/22nm transistor libraries suitable for large scale statistical circuit simulation.

4.1 20/22nm Technology Generation Testbed Transistor

Both n- and p-channel bulk MOSFETs, with a physical gate length of 25nm, were designed using the GSS process simulator ION [108], This process simulator has been designed in order to allow the generation of realistic doping profiles that closely match those obtained from full process simulation. ION uses published data for the stopping distances of ions in matter, and associated projected range and straggle parameters, to create representative doping profiles.

The testbed transistors have been designed following the prescriptions of the ITRS-2010 update [9] for a high performance 20/22nm technology generation transistor, subject to realistic physical constraints. The devices feature a high-κ dielectric metal gate stack with 0.85nm Effective Oxide Thickness (EOT). The p-channel device templates are designed to complement as closely as possible the electrical characteristics of the corresponding n-channel devices.

The device structures and doping profile are shown in Figure 4.1 and the important geometric and electrical parameters are summarised in Table 4.1.

Full electrical transfer characteristics for the n- and p-channel transistors obtained through simulation using GARAND, are shown in Figure 4.3, alongs side the uniform compact model fit. Additionally the dependence of threshold voltage on channel length, for both the n- and p-channel transistors can be seen in Figure 4.2.

Parameter	n-MOS	p-MOS	Description
$L_G[nm]$	25	25	Physical gate length
$EOT[nm]$	0.85	0.85	Equivalent oxide thickness
$X_i[nm]$	15	22.5	Source/drain extension
$N_A[\times 10^{18} cm^{-3}]$	4.5	4.95	Channel doping concentration
$V_{DD}[V]$	1	1	Nominal supply voltage
$I_{OFF}[nA]$	100	100	Off current
$I_{ON}[\mu A]$	1351	1009	Drive Current
$Spacer[nm]$	24	24	

Table 4.1: Structural and electrical parameters for the 20/22nm technology generation transistors.

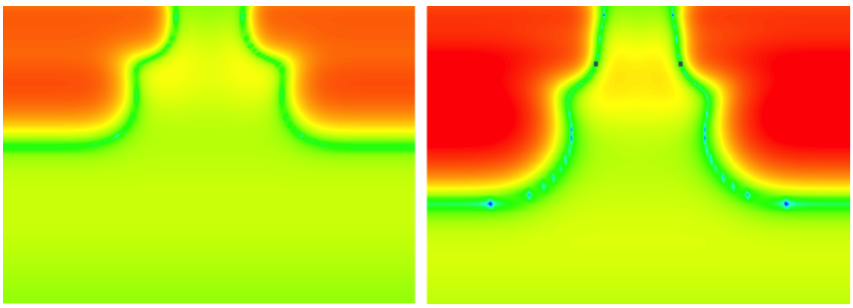
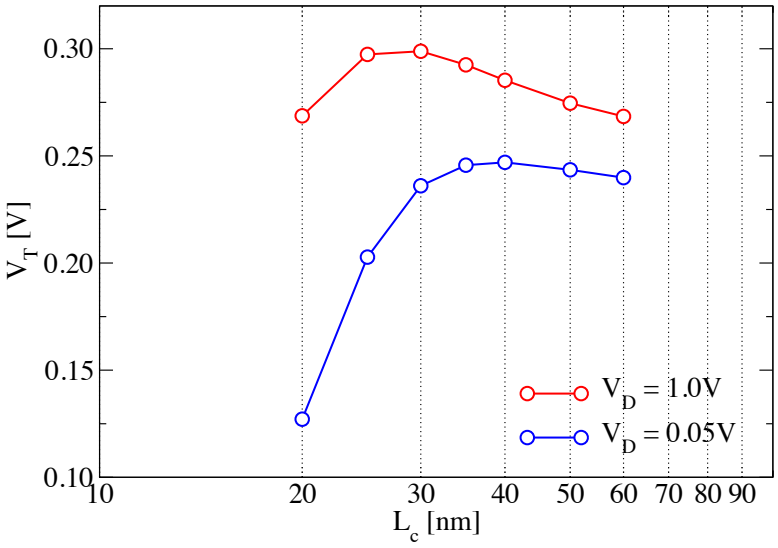
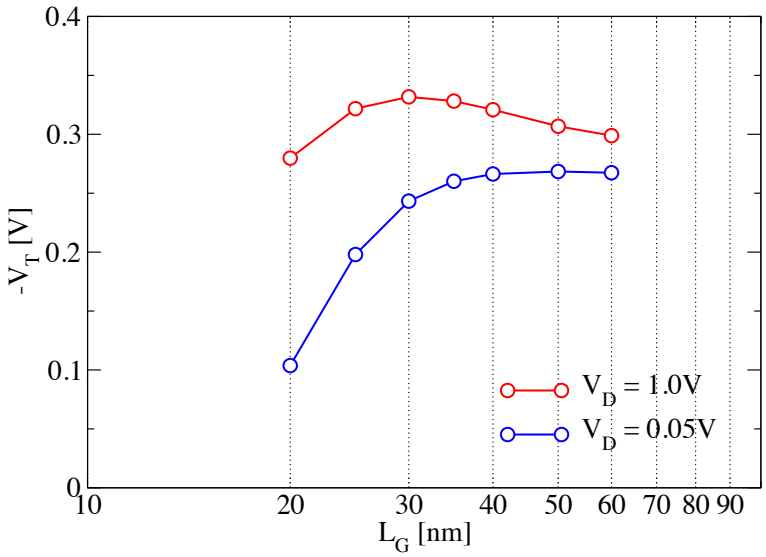


Figure 4.1: Net doping profiles for the template n-channel $25nm$ MOSFET (left) and p-channel $25nm$ MOSFET (right). The discontinuities are an artefact of the plotting tool.



(a)



(b)

Figure 4.2: Threshold voltage as a function of channel length, illustrating V_T rolloff of the (a) n-channel and (b) p-channel 22nm template bulk MOSFET at both high drain and low drain bias. Simulated devices have dimensions W = L = 25nm.

4.2 Nominal Compact Model Extraction Results

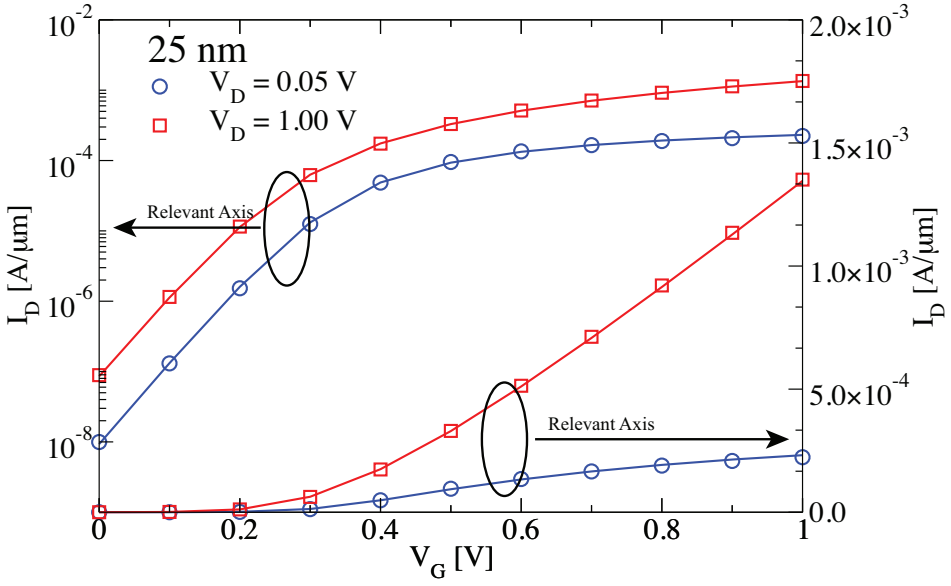
The 20/22nm BSIM4 model cards are based on a full set of electrical transfer characteristics obtained from the simulation of the template 25nm gate length MOSFETs. The presented extraction results were obtained using the methodology described in Section 3.3.1 and were provided by GSS. Figures 4.3 and 4.4 present the results obtained from the simulation the 25nm gate length n-channel and p-channel transistors using BSIM4 and SPICE compared to the GARAND simulation results.

These extracted models include accurate modelling of the substrate bias dependence at both low and high drain as illustrated in Figures 4.5 and 4.6, which show that the device body bias dependant behaviour of both n- and p-channel transistors is well modelled for substrate biases of 0, -0.2, -0.4, -0.6, -0.8 and -1.0V. To calculate a measure of the goodness-of-fit, we calculate average percentage relative error as described in Equation 4.1,

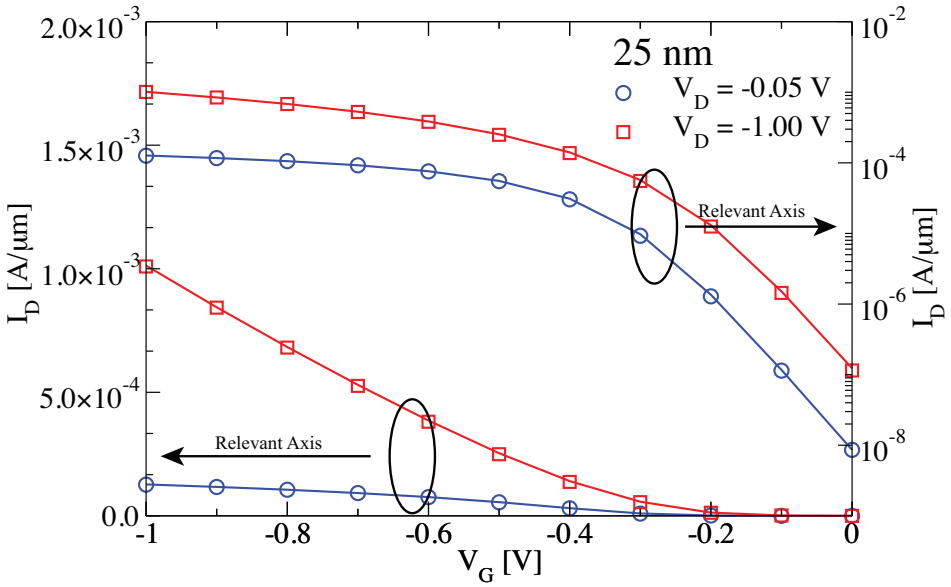
$$E_\mu = \frac{1}{n} \sum_{1 \longrightarrow n} E_n \tag{4.1}$$

where E_μ is the average percentage relative error and E_n is the individual percentage error at each of the 28 high and low drain simulated points from GARAND. At the nominal 25nm channel length, the average percentage relative fitting error of the transfer characteristics (I_dV_g), shown in Figures 4.5 (NMOS) and 4.4 (PMOS), is 2.5% for NMOS and 3.0% for PMOS, while the output characteristics (I_dV_d) fitting error is 1.4% for NMOS and 2% for PMOS. As it is difficult to ensure both length and body bias dependence these errors increase slightly to 4.5% at $\pm 5nm$ channel length for the I_dV_d and 3% for the I_dV_g characteristics.

The extraction has delivered a model which is reliable in the $20 - 40nm$ physical channel length range, and accurately captures drain bias and body bias dependence. With this first goal successfully achieved, we can begin to consider the statistical variability effects on the device performance, and the statistical model extraction stage can be initiated.

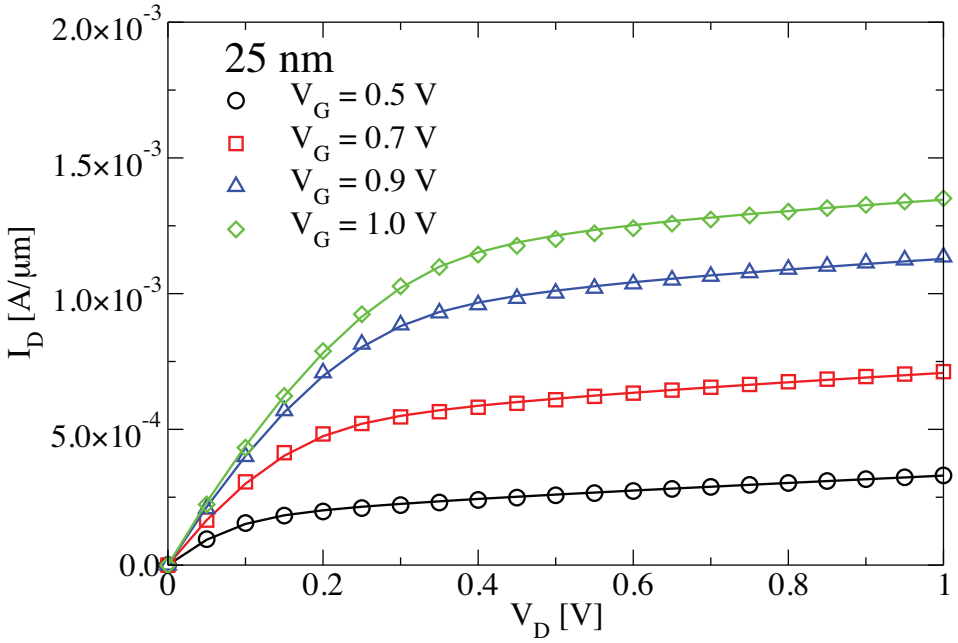


(a)

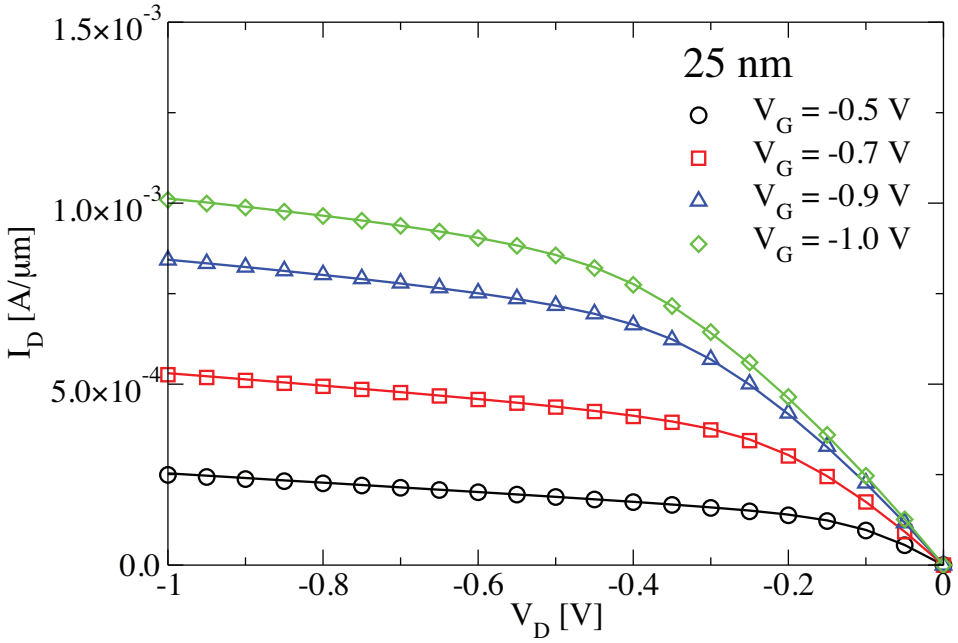


(b)

Figure 4.3: BSIM4 results of the 20/22nm (a) n-MOSFET transfer characteristics, (b) p-MOSFET transfer characteristics. GARAND simulation results are shown as solid lines, while extracted compact model values are denoted by symbols.

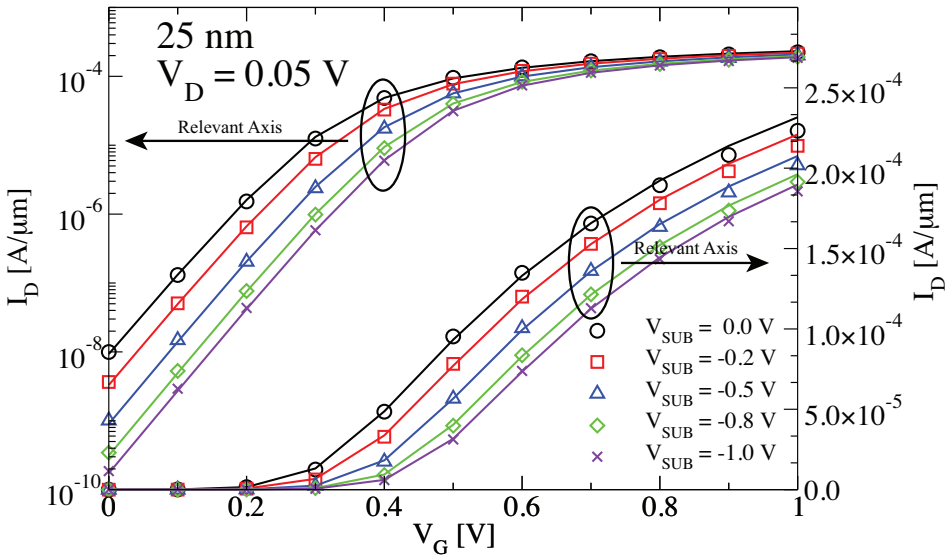


(a)

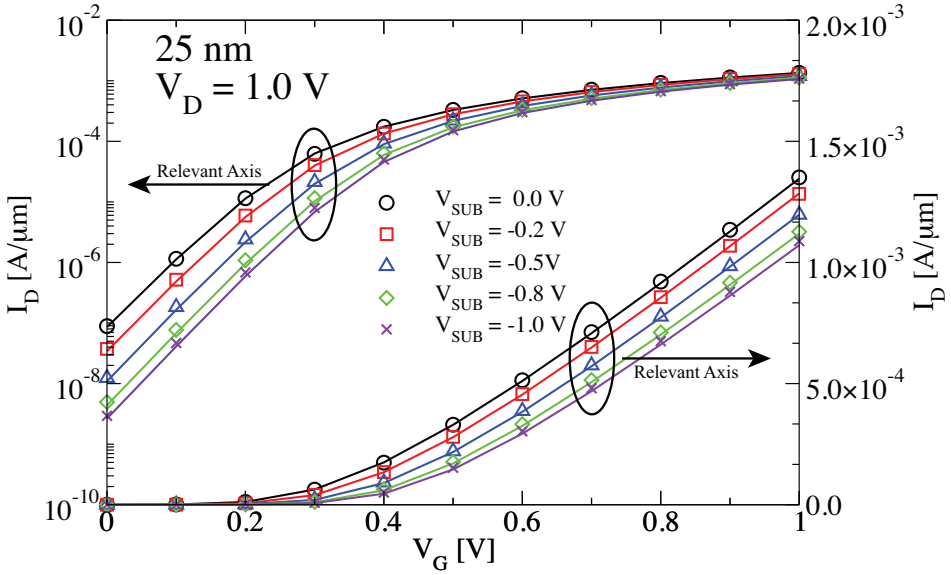


(b)

Figure 4.4: BSIM4 results of the 20/22nm (a) n-MOSFET output characteristics, (b) p-MOSFET output characteristics. GARAND simulation results are shown as solid lines, while extracted compact model values are denoted by symbols

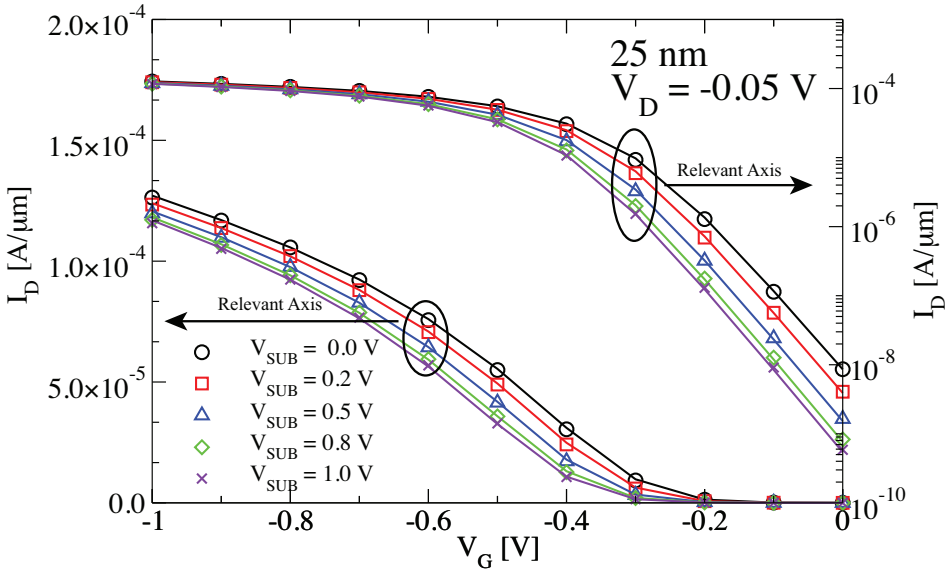


(a)

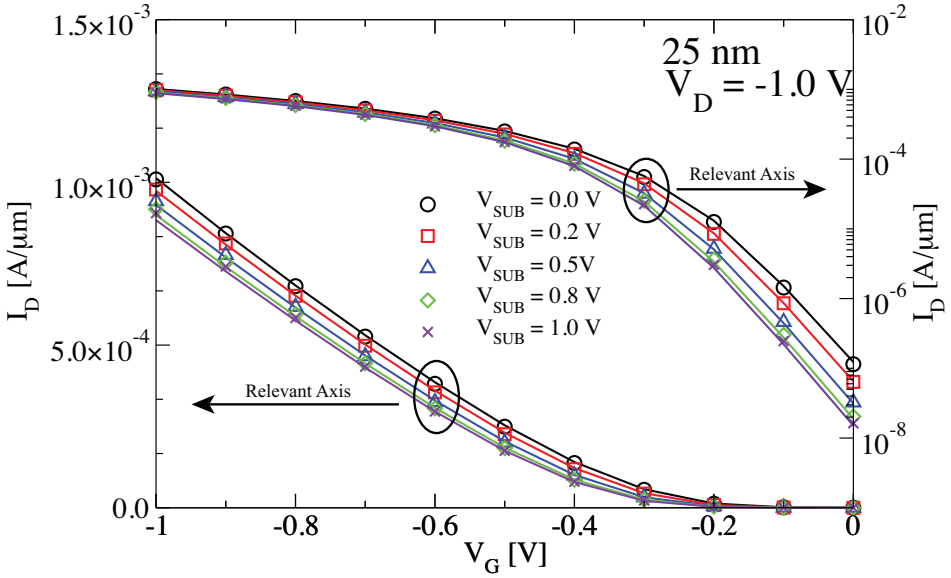


(b)

Figure 4.5: BSIM4 results of 20/22nm n-MOSFET at (a) VD = 0.05V and (b) VD =1.0V for substrate biases of 0, -0.2, -0.4, -0.6, -0.8 and -1.0V. GARAND simulation results are shown as solid lines, while extracted compact model values are denoted by symbols.



(a)



(b)

Figure 4.6: BSIM4 results of 20/22nm p-MOSFET at (a) VD = 0.05V and (b) VD =1.0V for substrate biases of 0, -0.2, -0.4, -0.6, -0.8 and -1.0V. GARAND simulation results are shown as solid lines, while extracted compact model values are denoted by symbols.

4.2.1 Figure of Merit Based Extraction with Mystic

In the strategy introduced by Cheng et al. [123], parameter selection is based on a numerical sensitivity analysis. This parameter set is then extracted on a per-device basis. The extraction is performed using a global optimisation of all parameters over all data points ($I_D - V_G$) in a single device. This is easily achieved using Mystic, as depicted as the second extraction stage in Figure 3.5. Once a uniform model is extracted and a parameter set has been selected Mystic performs the global extraction using one of its built-in optimisers.

This approach strives for minimal extraction errors with respect to the target device data, however, it does not guarantee continuous uni-modal distributions of the extracted parameter sets, which is desirable for model generation. The problem stems from the fact that there is a complex interdependence between the compact model parameters, and their non-linear impact on transistor performance. As all parameters are extracted simultaneously over the whole range of the data, highly correlated parameters can compensate for each other while reducing the error and can artificially influence the extraction process, leading to non-physical parameter values, with non mono-modal distributions or extreme outliers in certain parameter distributions. This creates problems when considering statistical model generation methods that require continuous uni-modal distributions of the underlying parameters. A few extreme outliers can make the calculated moments of an extracted parameter distribution moment unreliable [126], and can have a largely detrimental impact on generated device accuracy. The motivation behind the figure of merit based statistical parameter extraction approach introduced in this work is to accurately reproduce the target device behaviour, whilst also producing parameter distributions which are suitable for advanced statistical compact model generation strategies.

The physical extraction approach is rooted in device performance figures of merit which can be obtained through 3D simulation. Inspection of the distributions of the key device figures of merit, including high drain and low drain threshold voltage, on current, off current, subthreshold slope and DIBL, shows that they are continuous and can be accurately described with four moments -

mean, standard deviation, skew and kurtosis. These moments are conventionally used to accurately describe a uni-modal distribution [134], as higher order moments have a limited effect and increase complexity whilst being difficult to accurately evaluate without a very large quantity of data. The figure of merit based approach utilises this uni-modal nature, by adopting the figures of merit as compact model extraction targets or extracting each parameter individually, and looping through the extraction strategy self consistently, targeting only the region of transistor operation the parameter is intended to effect. This approach forgoes a parameter sensitivity analysis in favour of a more physical parameter selection methodology.

Parameter selection is based on the parameter's ability to capture a physical aspect of device operation and corresponding variability, or a specific figure of merit of the statistical device ensemble. The figures of merit used in the fitting procedure include: threshold voltage, DIBL, subthreshold slope, drain dependence of subthreshold slope, low drain on current, and high drain on current, low drain off current and high drain off current. If the behaviour of each figure of merit can be fitted with a single parameter, we can use a minimum of set 8 parameters.

The selected BSIM4 parameters are then extracted individually and self-consistently within the optimisation loop with respect to the figure of merit it has been selected to capture. The extraction process is depicted in Figure 4.7.

4.2.2 Physical Parameter Selection/Sensitivity Analysis

For the figure of merit extraction strategy to be viable, a parameter set has to be identified which can accurately represent the figures of merit of the simulated statistical device ensemble. This process has to be based on a good understanding of the compact model implementation as well as the physical characteristic being modelled. The selected parameter set for the 22nm statistical model extraction is shown in Table 4.2. This approach attempts to correlate compact model parameters with the figures of merit of the devices. Since the distributions of figures of merit are uni-modal and continuous, the extracted parameter set is expected to be more suitable for compact model

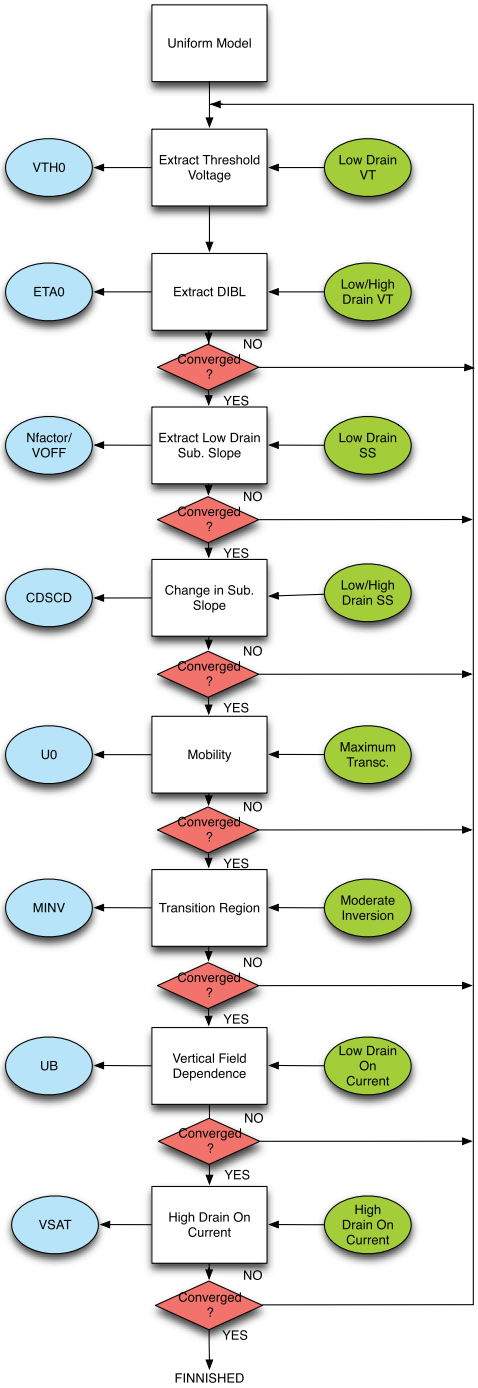


Figure 4.7: Figure of merit based extraction strategy flow chart.

Parameter	Description
$VTH0$	Low drain threshold voltage
$ETA0$	DIBL
$VOFF$	Channel charge at density ad $V_{DS} = 0V$
$NFACTOR$	Low drain subthreshold slope
$CDSCD$	Drain dependence of subthreshold slope
$MINV$	Moderate inversion fitting parameter
$U0$	Low field mobility
UB	Vertical field dependence
$VSAT$	Saturation Velocity

Table 4.2: Selected compact model set for figure of merit based statistical model extraction, corresponding physical effect is also described.

generation than a brute force global optimisation approach. Aside from this, in this method the gross physical effect of the compact model parameters is retained which can greatly aid the analysis of effects of device performance on circuit performance. In each case the parameters have been carefully selected in an attempt to simplify their impact on transistor performance. Ideally all parameters should respond linearly and should be completely decoupled in their effect. However, due to the complex physics of small transistors, and the inherent correlations between the effects that the parameters are intended to model, it is difficult to find set of parameters which are orthogonal in terms of their effects. A brief description of each parameter as well as its implementation in BSIM4 and its effect on the $I_d - V_g$ characteristics at high drain bias (1 V) and low drain bias(50 mV) follows.

The threshold voltage of the devices is principally captured using the parameter $VTH0$. The effect of this parameter on transfer characteristics is shown in Figure 4.8. $VTH0$ produces a linear shift in the threshold voltage within the BSIM4 model, as shown in Equation 4.2,

$$v_{th} = VTH0 + (X) \tag{4.2}$$

where X represents the complex threshold voltage dependence on non-uniform channel doping, halo doping, short channel DIBL effects, and narrow width effects, all of which are extracted in the uniform compact model. In the ex-

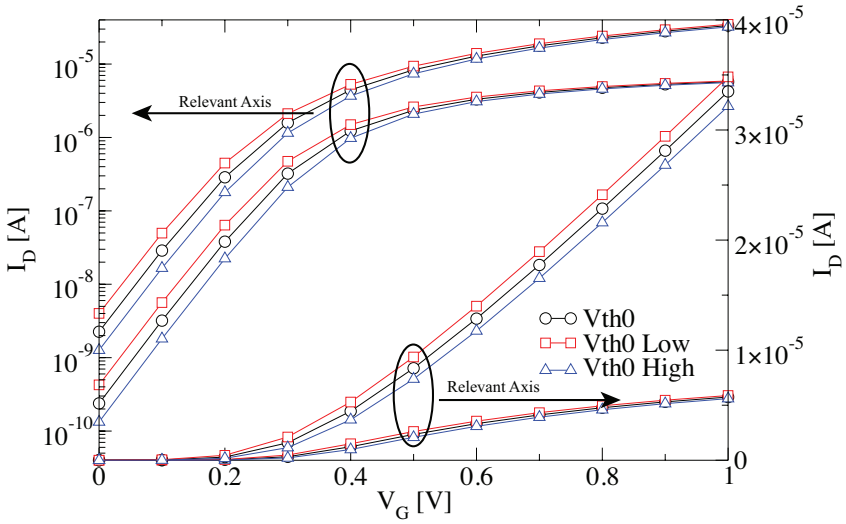


Figure 4.8: Effect of $VTH0$ on transfer characteristics, showing a linear shift of the $I_d - V_g$ curve with respect to gate voltage.

traction strategy $VTH0$ is targeted at the low drain bias threshold voltage of each individual device in the ensemble.

The effect of DIBL on device performance is captured using the parameter $ETA0$. This parameter affects the transfer characteristics as shown in Figure 4.9. The figure shows that ETA0 has little or no impact on low drain performance, but controls the threshold voltage at high drain bias. The extraction of $ETA0$ is therefore targeted to the high drain threshold voltage of the device, while $VTH0$ remains at the value extracted in the previous stage.

Off current is captured using the parameter $VOFF$. The effect of this parameter on the transfer characteristics is shown in Figure 4.10, where $VOFF$ changes the off-current of the device without affecting the subthreshold slope by extending the transition region between the subthreshold and linear regimes of the device.

Low drain subthreshold slope is captured using the parameter $NFACTOR$. This parameter changes the subthreshold slope of the transistor as shown in Figure 4.11. In order to insure correct slope and off current, $NFACTOR$ and $VOFF$ are often extracted together and target the combination of subthreshold slope and off current at low drain bias for each device.

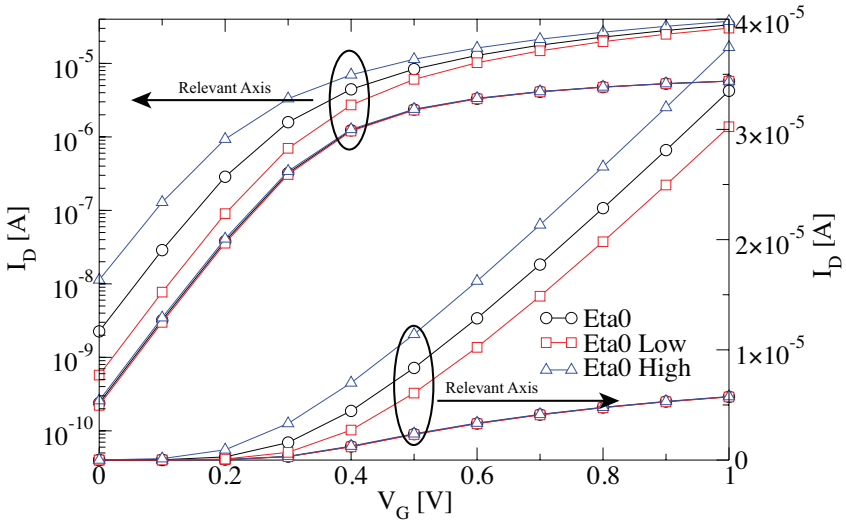


Figure 4.9: Effect of $ETA0$ on transfer characteristics. Little impact on low the low drain bias curve is seen, while the high drain bias performance shows a liner shift in threshold voltage.

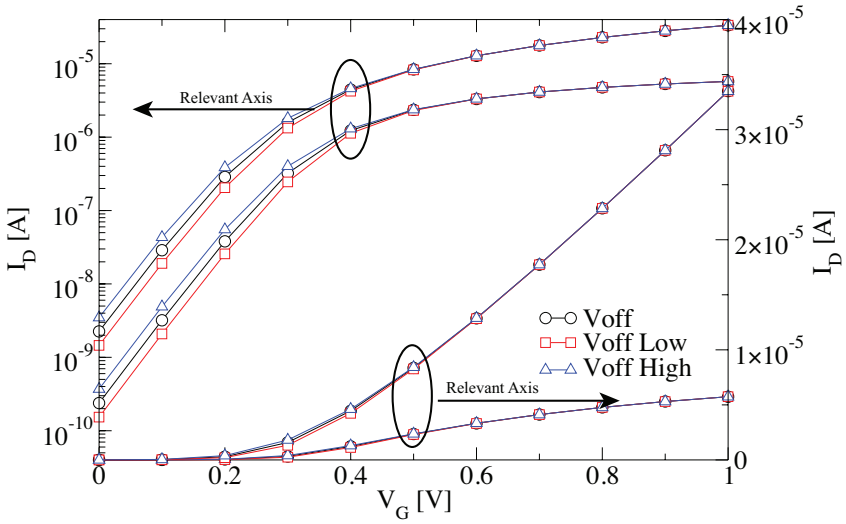


Figure 4.10: Effect of $VOFF$ on transfer characteristics, showing a parallel shift in the subthreshold behaviour, whilst not affecting the linear region of the transistor.

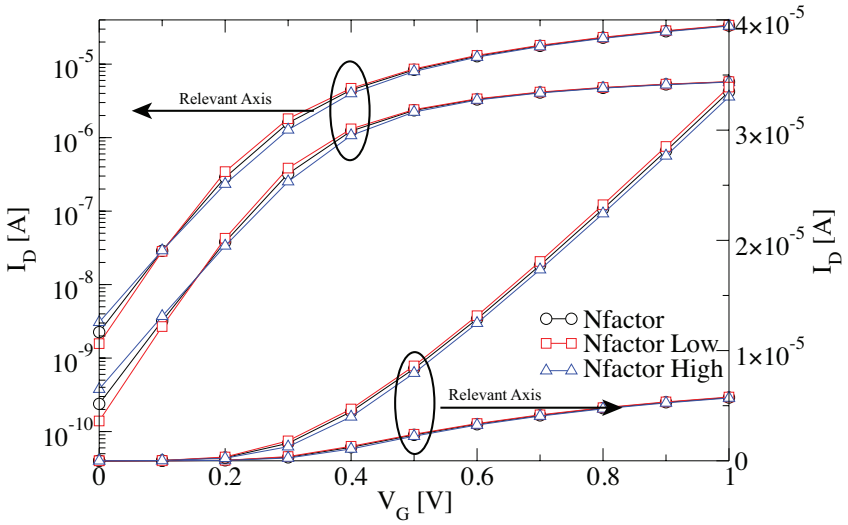


Figure 4.11: Effect of $NFACTOR$ on transfer characteristics, showing change in subthreshold slope at both low and high drain bias.

Drain bias dependence of the subthreshold slope (Δ slope) is captured using the parameter $CDSCD$. This parameter adjusts the high drain subthreshold slope of the device without affecting the low drain subthreshold slope as shown in Figure 4.12. It is important to capture the change in subthreshold slope as a function of drain bias, as it can have a significant impact on leakage and SRAM performance.

The behaviour of the device in the moderate inversion region is captured using the parameter $MINV$. The effect of this parameter is shown in Figure 4.13. $MINV$ is a purely phenomenological parameter introduced to improve the fit in the region linking the subthreshold and strong inversion regions. $MINV$ is targeted to the region around the threshold voltage point, specifically the transconductance.

Variation in the mobility, mainly due to RDD, is captured using the parameter $U0$. Similar to $VTH0$, $U0$ linearly scales the BSIM4 effective mobility, as shown in Equation 4.3 and Figure 4.14.

$$\mu_{eff} = \frac{U0 \times U_{\text{lenght}}}{U_{\text{vertical}}} \tag{4.3}$$

where U_{length} contains a complex expression for the length dependence of mo-

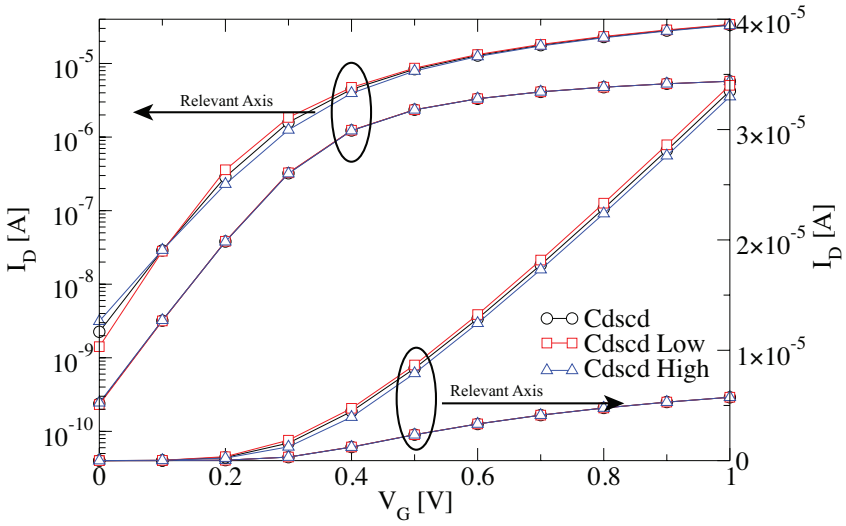


Figure 4.12: Effect of $CDSCD$ on transfer characteristics, controlling the high drain bias subthreshold slope and off current without affecting the low drain bias characteristics.

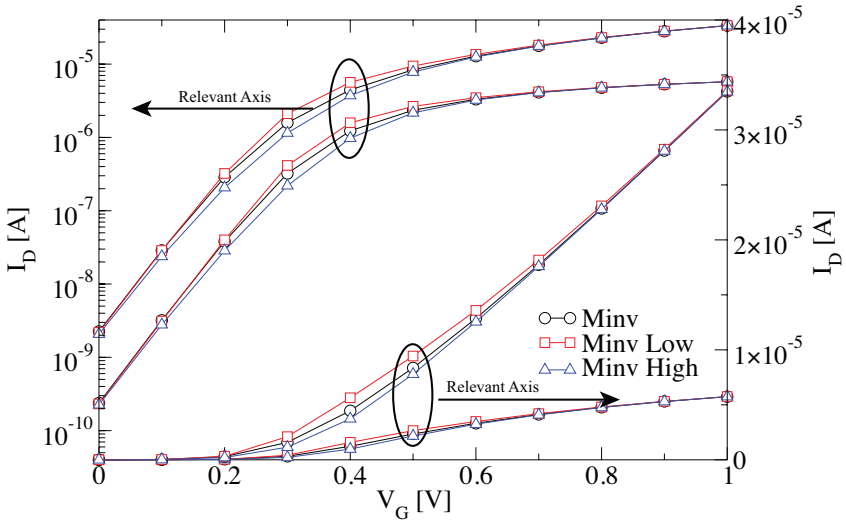


Figure 4.13: Effect of $MINV$ on transfer characteristics, included to allow control of the moderate inversion region in the transition between subthreshold and strong inversion.

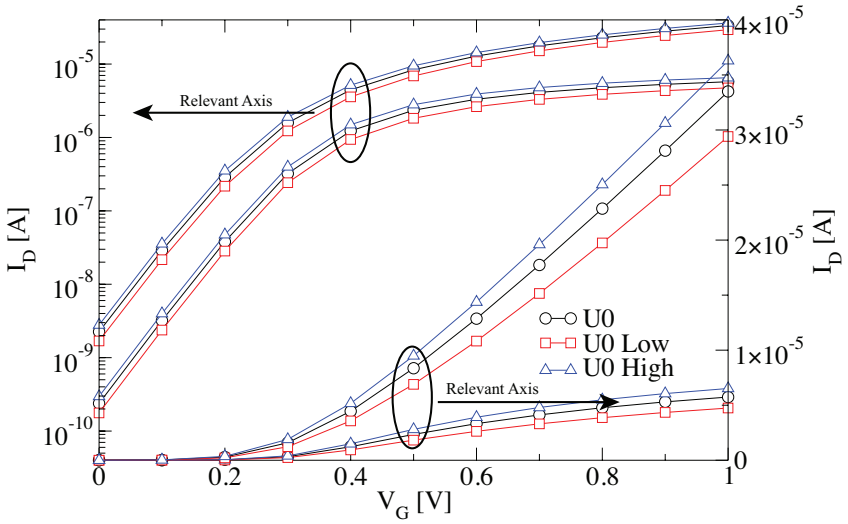


Figure 4.14: Effect of $U0$ on transfer characteristics, increasing or decreasing this parameter causes a vertical shift in the device characteristics.

bility and U_{vertical} contains the vertical field dependence of mobility.

Variation in low drain on-current is captured using the vertical field dependence parameter UB. The effect of this parameter on the transfer characteristics can be seen in Figure 4.15. UB impacts both low drain and high drain on-current, however we only fit to low drain on-current as $VSAT$ will be used to fit high drain on-current. The expression for $U_{vertical}$ can be seen in Equation 4.4:

$$U_{vertical} = 1 + (UA + UCV_{bseff})(A) + UB(A)^2 + UD(B)^2 \tag{4.4}$$

where

$$A = \frac{V_{gsteff} + 2V_{th}}{TOXE} \tag{4.5}$$

$$B = \frac{V_{th} \times TOXE}{V_{gsteff} + 2\sqrt{V_{th}^2 + 0.0001}} \tag{4.6}$$

and UA, UB and UC are fitting parameters, V_{bseff} is a function of body bias voltage, V_{gsteff} is a function of gate voltage, $TOXE$ represents the effective oxide thickness. UB is selected instead of UA as it has a quadratic dependence with gate voltage and can introduce "bending" to the low drain I_DV_G at high

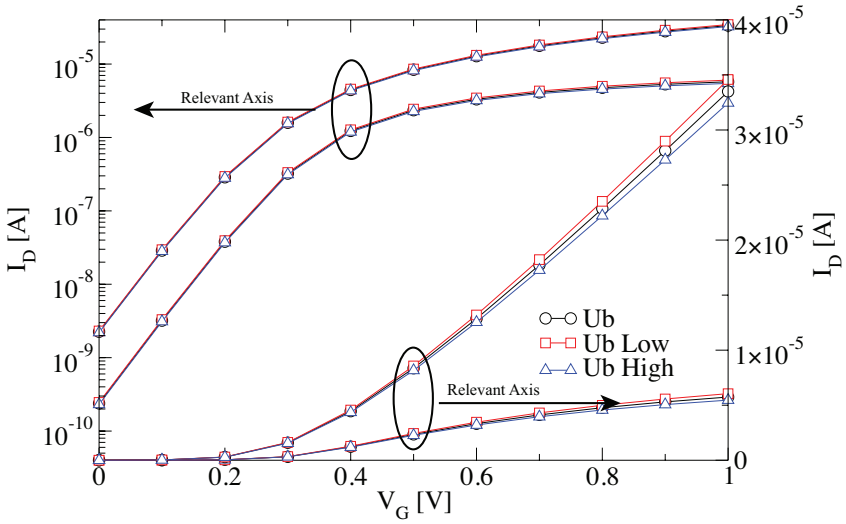


Figure 4.15: Effect of UB on transfer characteristics

gate bias. $U0$ and UB are usually extracted in subsequent stages to fully capture the shape of the low drain current above threshold voltage.

Variation in high drain on-current is captured using the saturation velocity parameter $VSAT$. The effect of this parameter on the transfer characteristics is shown in Figure 4.16. $VSAT$ is used to model high drain on-current directly as shown in Equation 4.8.

$$v = \frac{\mu_{eff}}{1 + \frac{E}{E_{sat}}} \qquad E < E_{sat} \tag{4.7}$$

$$v = VSAT \qquad E \geq E_{sat} \tag{4.8}$$

Where E is the electric field along the channel, E_{sat} is the electric field at the saturation velocity for the majority carriers and μ_{eff} is the effective mobility.$VSAT$ is used to calibrate the high drain on-current.

Despite the method's effectiveness, which will be outlined in Section 4.3, there are still a few difficulties and limitations related to this figure of merit extraction strategy. One of the principle assumptions is that the underlying device figures of merit have continuous mono modal distributions. For a technology or device where this does not hold true the extraction strategy, with a

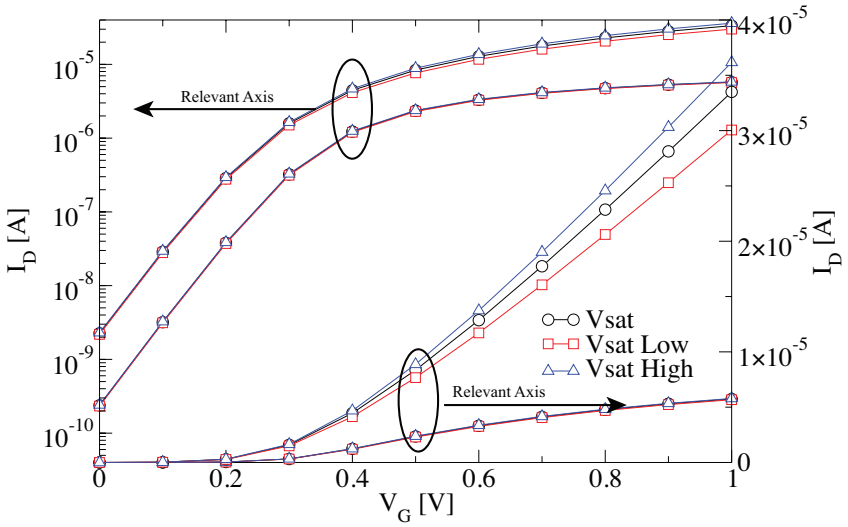


Figure 4.16: Effect of $VSAT$ on transfer characteristics

view to compact model generation, will not be applicable. An example of this is threshold voltage variability due to MGG, which is shown to have a bi-modal distribution when grain size approaches gate size [64]. An additional assumption is also made, that the uniform model provides a good initial condition for statistical extraction and has a performance which is reasonably representative of each device in the statistical ensemble. We also presume that the selected parameter set can fit all the statistical devices using the uniform model as a base model. The latter assumption can be stressed in extreme devices, where the physical properties (channel doping (RDD) and effective channel length (LER)) of the device can significantly differ from the nominal design point device. The ability to control a single figure of merit with a single parameter is limited by the implementation of the BSIM4 model and its underlying equations. Some of the phenomenological "fitting" parameters are applied through complex non-linear trigonometric functions, for example $MINV$ is introduced as an inverse tangent, as shown in Equation 4.9.

$$m* = 0.5 + \frac{\arctan(MINV)}{\pi} \tag{4.9}$$

Additionally, there is a high correlation between the parameters, many of which

are used in the same equations. While the local extraction of a single parameter per stage attempts to limit these interactions, it is difficult to extract the parameters independently.

A selection of problematic devices are explored in Section 4.4.

4.3 Statistical Compact Model Extraction Results

A figure of merit statistical extraction was performed for an ensemble of 10,000 20/22nm transistors simulated using GARAND and including statistical variability in the form of RDD, LER and MGG. In this section, we provide a detailed comparison of the device figures of merit calculated from the 10,000 fitted compact models (threshold voltage V_T, on-current I_{ON}, off-current I_{OFF}, and DIBL) at high and low drain bias, with the reference device figures of merit calculated from 3D physical simulation results. The simulation data is in the form of gate bias points swept from -0.3V to 1V at 0.1V steps, including both high and low drain bias conditions.

Once the compact model extraction is complete we can calculate and compare the figures of merit of the compact models with 3D device simulations. By comparing device figure of merit distributions, instead of simply looking at overall error in the fit, it becomes apparent whether device fits are consistent deep into the tails of the distributions. These comparisons are presented in the form of Quantile-Quantile plots (QQ plot). A QQ plot is a graphical method for assessing whether two samples are drawn from the same underlying distribution and is explained in detail in [135]. In this case, the reference distribution is a Gaussian distribution with the same mean and standard deviation as the data, which appears shown as a straight diagonal line on the QQ plot. The QQ plots show both the simulated device data and extracted compact model data compared to this Gaussian distribution. If the simulated device and compact model data match they should be identical across the entire QQ plot. The QQ plot is useful as it can clearly indicated mean shifts, increase in error deeper into the tails and the continuity of the distribution.

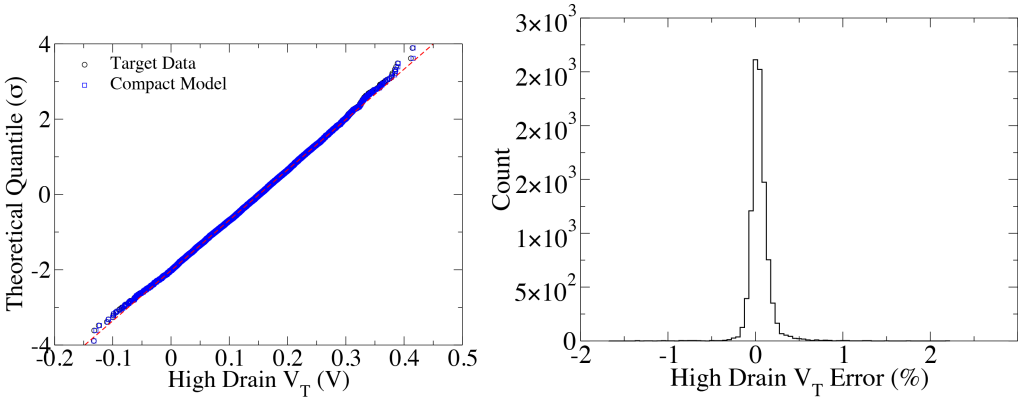


Figure 4.17: High drain threshold voltage distribution fit (left) and error distribution (right). The QQ plots shows the distribution is Gaussian for the simulated devices.

In addition to the QQ plots, Figures 4.17 to 4.23 also show histograms of relative error distributions for each figure of merit illustrating the low fitting error of these devices. As these figures show, the extracted compact models capture the distributions of the figures of merit of the simulated device ensemble, with extremely low errors (within 1% across all 10,000 devices) in the threshold voltage and on-current figures at both high and low drain. Off-current shows a higher error (close to 10 devices above 15% at high drain), though this was expected due to the logarithmic nature of this figure of merit, where a small amount of noise has a high impact on percentage error.

The average percentage relative error of the extracted models over all 28 high drain and low drain simulated data points is depicted as a histogram in Figure 4.24. The error mean is 2.2%, with a standard deviation of 1.16%, with a total of 302 (or 3%) of devices above 5% error, and 13 (or 0.1%) of devices above 10% error. It is important to note that the high error devices do not represent any extreme of any individual device figure of merit performance, but relate to unusual combinations of physical effects. Some of these devices will be investigated in Section 4.4.

Having demonstrated that the method is capable of capturing the device performance figures of merit, we now check that the correlations between them are also accurately retained. Figure 4.25 shows the correlations between the

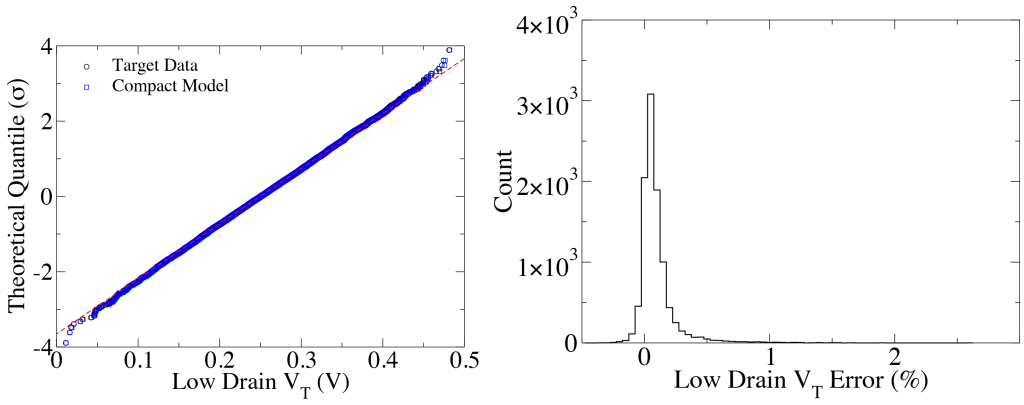


Figure 4.18: Low drain threshold voltage distribution fit (left) and error distribution (right). The QQ plots shows the distribution is Gaussian for the simulated devices.

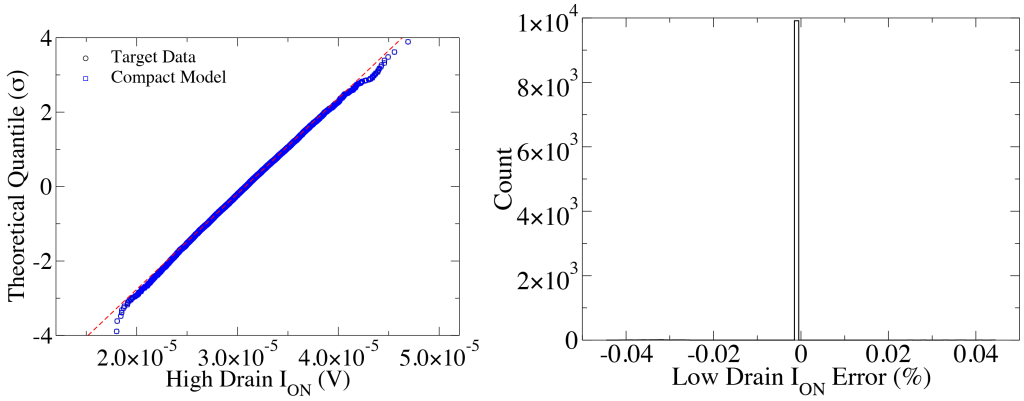


Figure 4.19: High drain I_{on} distribution fit (left) and error distribution (right). The QQ plots shows the distribution is Gaussian for the simulated devices. Error is minimal as this is the last figure of merit to be extracted.

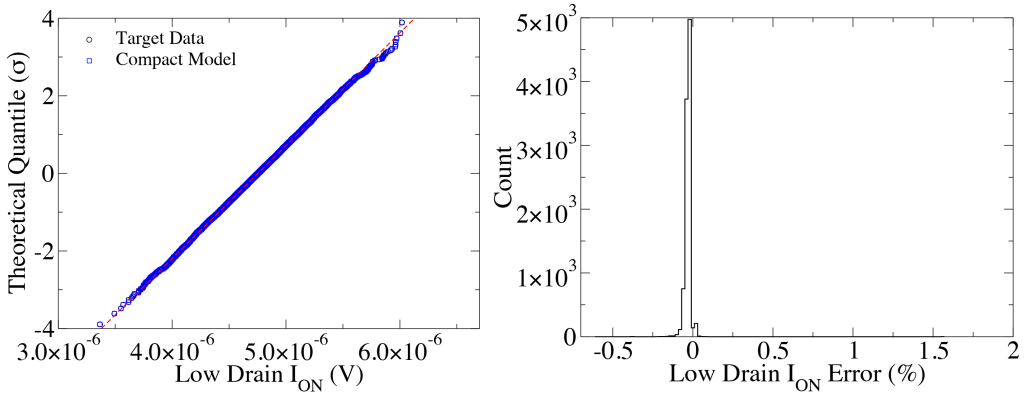


Figure 4.20: Low drain I_{on} voltage distribution fit (left) and error distribution (right). The QQ plots shows the distribution is Gaussian for the simulated devices.

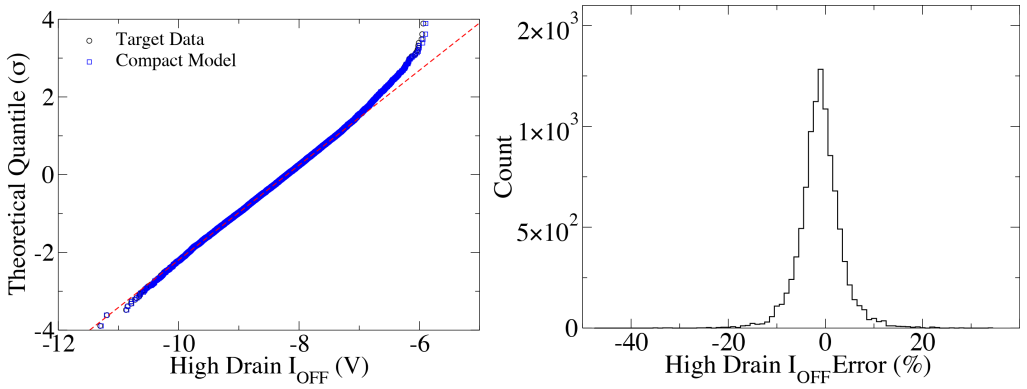


Figure 4.21: High drain I_{off} distribution fit (left) and error distribution (right). The QQ plots shows the distribution is skewed, this is due to the fact that the highest off-current represents device which have V_T close to 0V, and the behaviour is no longer logarithmic.

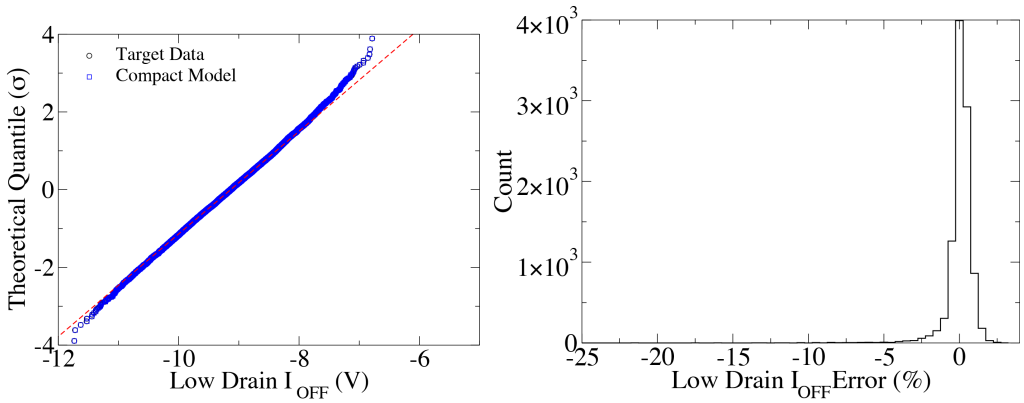


Figure 4.22: Low drain I_{off} distribution fit (left) and error distribution (right). The QQ plot shows elements of both skew and kurtosis.

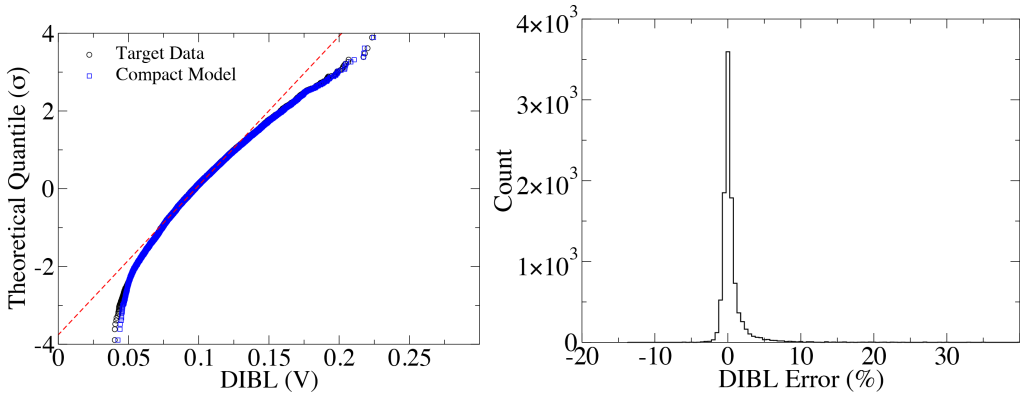


Figure 4.23: DIBL distribution fit (left) and error distribution (right). The QQ plot shows a large amount of skewness, at least partially due to the fact that the DIBL distribution is bounded - DIBL cannot produce a negative shift in threshold voltage.

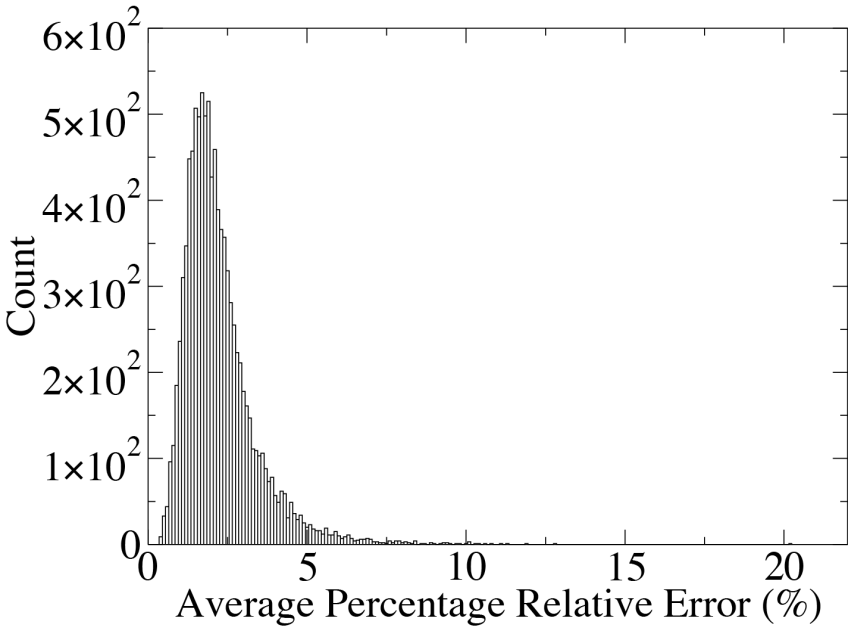


Figure 4.24: Average percentage relative error of fitted models.

simulated device figures of merit in the form of scatterplots and correlation coefficients, demonstrating that the extracted compact models almost perfectly capture the correlations between the simulated device figures of merit.

Having established that the extraction strategy captures the statistical variability present in the 20/22nm devices, the focus of the work now shifts on compact model generation, where it must be confirmed that the extracted model parameters are suitable for compact model generation strategies.

4.3.1 Extracted Parameter Distributions

Having performed the statistical compact model extraction, we can now examine the distributions of the parameters and the correlations between them. The distributions of the extracted model parameters are shown in Figure 4.26, represented in the form of QQ-plots, with the dashed line showing a Gaussian distribution. This information shows that the majority of the parameters are non-Gaussian distributed with large amounts of skew present (e.g. $ETA0$), while other parameters shown a large amount of kurtosis, ($CDSCD$). This

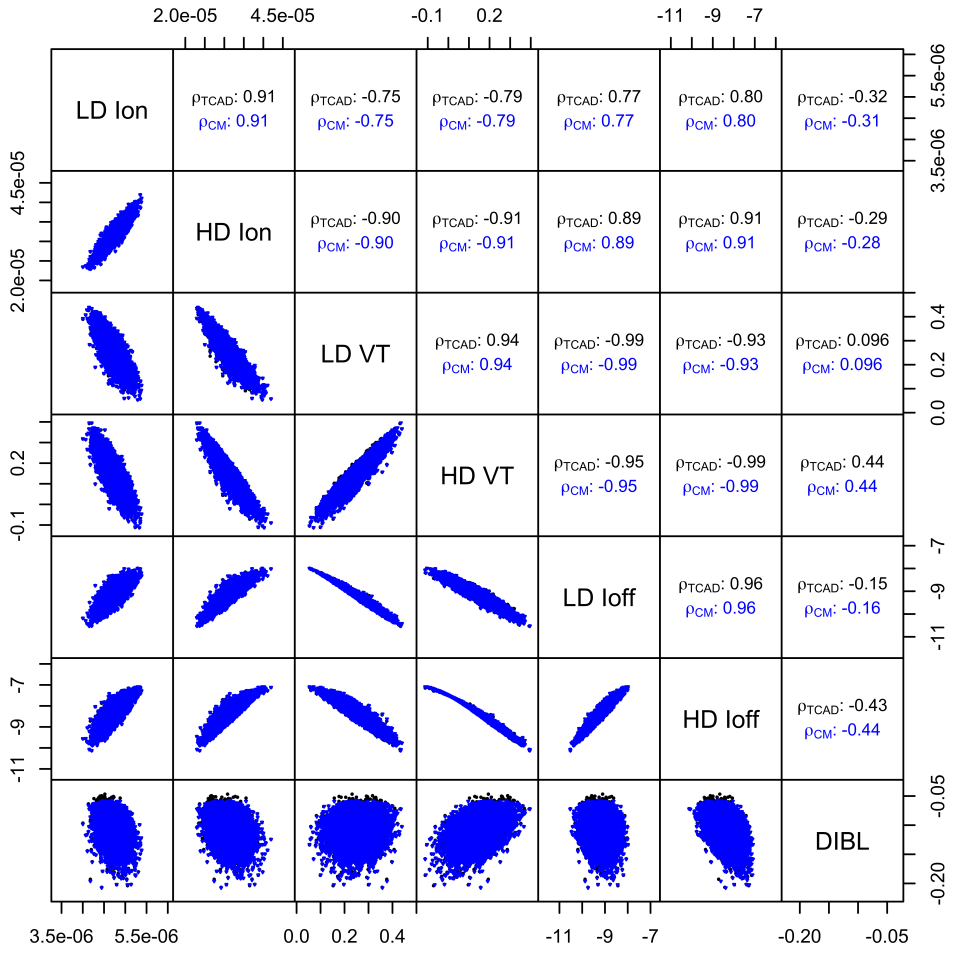


Figure 4.25: Correlations between device figures of merit. Black represents the 3D simulated device data and the blue extracted compact model data, the bottom half of the table shows scatter plots of the two data sets and the upper diagonal shows correlation coefficients. The table shows that the correlation between the figures of merit is complex the fact that the compact model captures this shows the underlying physics is being effectively captured.

is obvious when we consider the impact that skew and kurtosis has on a QQ-plot. The impact of skew is to increase the probability of extreme values in one tail while reducing the probability of extreme values in the other. This can be clearly seen in the QQ-plot of $ETA0$. The upper tail deviates from the Gaussian line in a direction which indicates more extreme values are present and the lower tail also deviates from the Gaussian line, but in this case we see that the values are less extreme than those we expect from a Gaussian distribution. Kurtosis manifests itself as an increase in the probability of extreme values in both tails. The QQ-plot of $CDSCD$ shows that, while the middle of the distribution is relatively Gaussian, the tails deviate, and both the upper and lower tail show more extreme values than the Gaussian distribution.

Figure 4.27 shows that there are also strong correlations between the parameters - for example the correlation coefficient between $VTH0$ and $U0$ is as high as 0.51, and that some parameters have complex non-linear correlations, for example $VSAT$ and $CDSCD$. This is not surprising as the figure of merit extraction strategy attempts to map parameters with figures of merit and it has already been shown, in Section 4.3, that the device figures of merit are highly correlated.

After the ensemble of statistical compact models has been extracted, it is crucial to accurately propagate this information to circuit simulation. In order to avoid problems associated with subsampling, as there will always be a finite number of devices which can be simulated or measured (as is discussed in Section 4.5), it is important to have the ability to generate an effectively unlimited ensemble of devices that replicate the statistical performance of the underlying technology. The compact model parameter distributions obtained from the figure of merit extraction strategy are all continuous and as such are suitable for the purpose of advanced statistical compact model generation strategies like PCA and NPM.

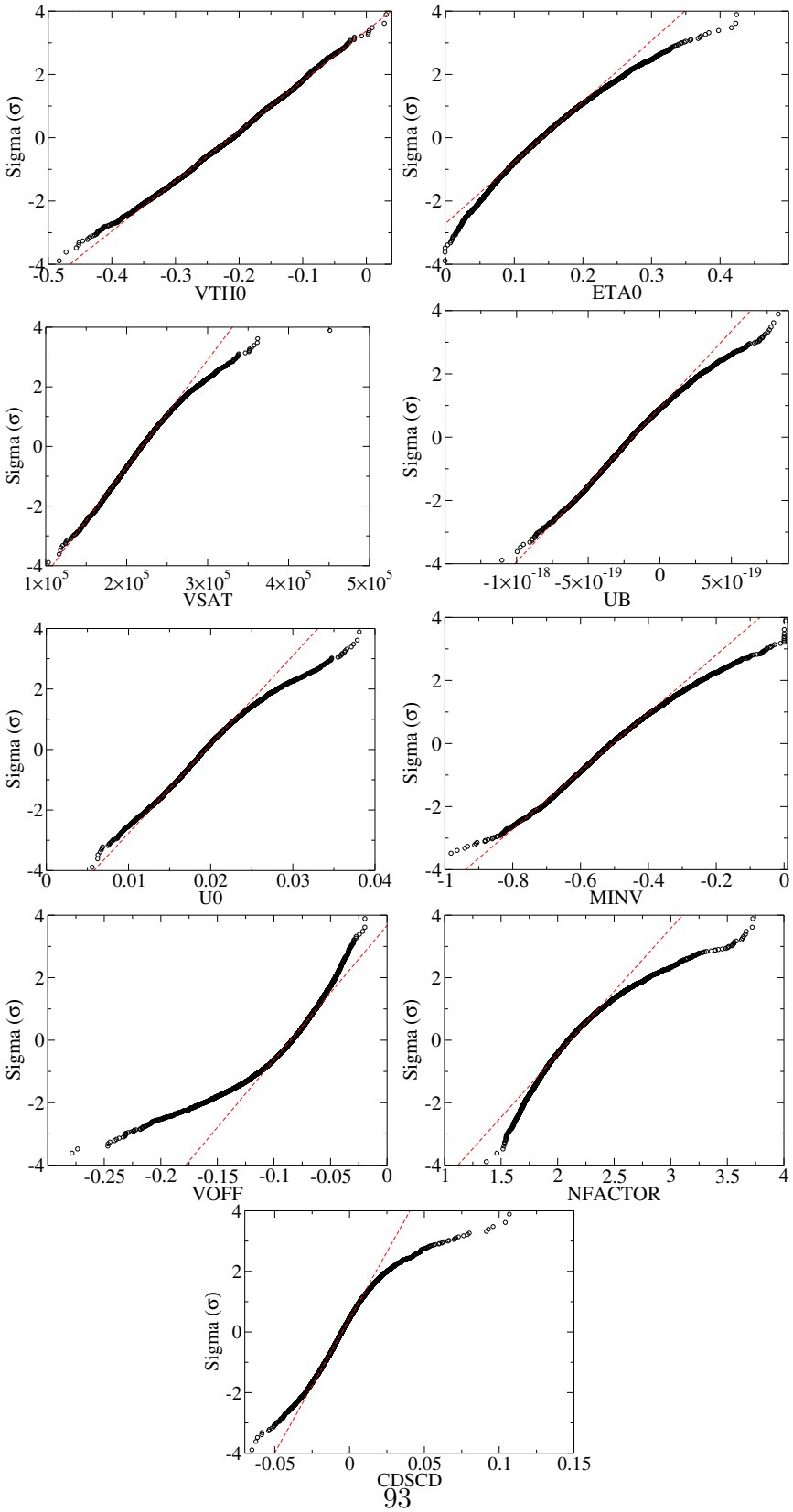


Figure 4.26: Extracted Parameter Distributions.

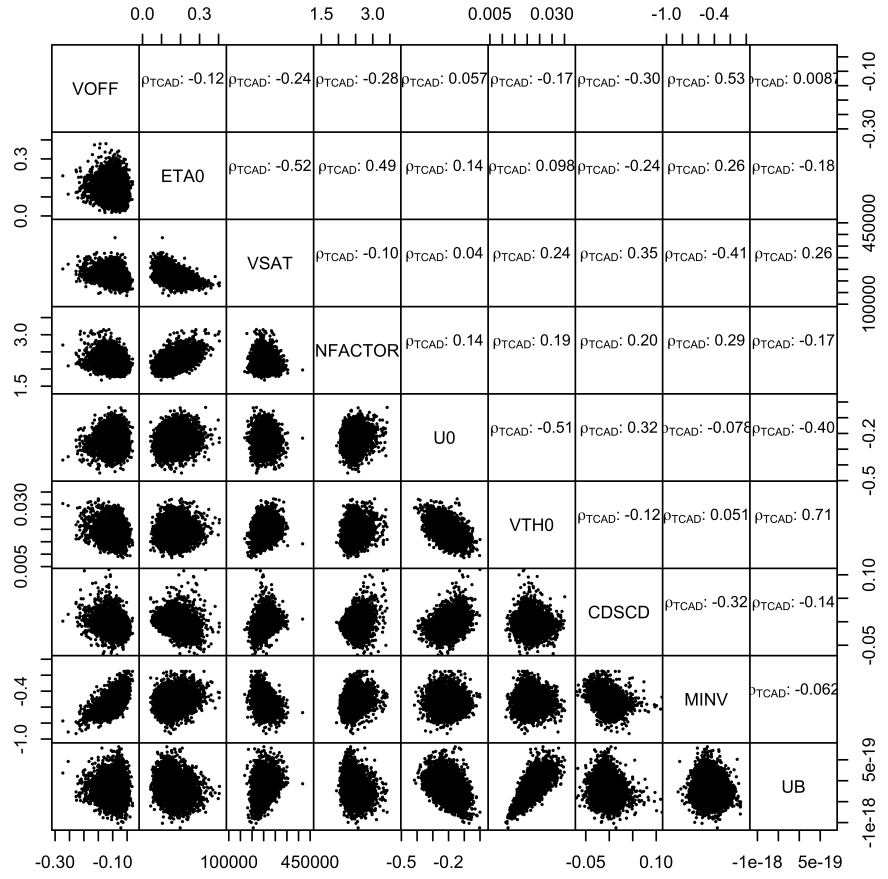


Figure 4.27: Correlations between the parameters, the bottom half of the table shows correlation scatter plots and the top half shows correlation coefficients.

4.4 Statistical Compact Modelling Challenges

During the development of this extraction strategy, it has become apparent that particularly rare combinations of statistical variability sources can result in extreme device characteristics, which present significant problems during the extraction of statistical compact model parameters.

Figure 4.28 compares the transfer characteristics of the uniform transistor with the transfer characteristics of three such 'anomalous' transistors. The fitting accuracy of the uniform transistor is 2% over the full range of the transistor characteristics and the use of a set of 9 carefully selected statistical compact model parameters yields a 2.7% average percentage relative error over the whole ensemble of 10,000 statistical characteristics. In contrast the errors for devices 9597, 6794 and 2040 of the statistical ensemble are 12%, 7% and 6% respectively. The reasons for the observed extreme device behaviour are analysed in the following sub-sections.

4.4.1 Device 9597

The transfer characteristics of device 9597 show a high on-current similar to that of the template transistor but a very low on/off-current ratio of only 60, compared to the anticipated ratio of 1000 based on the nominal transistor design. Aside from the fact that this device has a statistically rare low threshold voltage, its behaviour can be attributed to acute short channel effects that manifest as poor sub threshold slope and very large DIBL of $220mV$. The reasons for this can be understood from an analysis of the structure of the device. Figure 4.29 shows the electron concentration equi-contours in this transistor at V_T, at high drain (top) and low drain (bottom) bias conditions for a threshold voltage defined using a constant current criteria of $10nA$. The analysis shows that three random donor dopants protruding from the source form a current percolation path $3-4nm$ below the surface of the gate, reducing the effective channel length at this point and decreasing gate control over the current flowing through this region. This is the expected short channel behaviour of a 'buried channel' type transistor formed by the three rogue

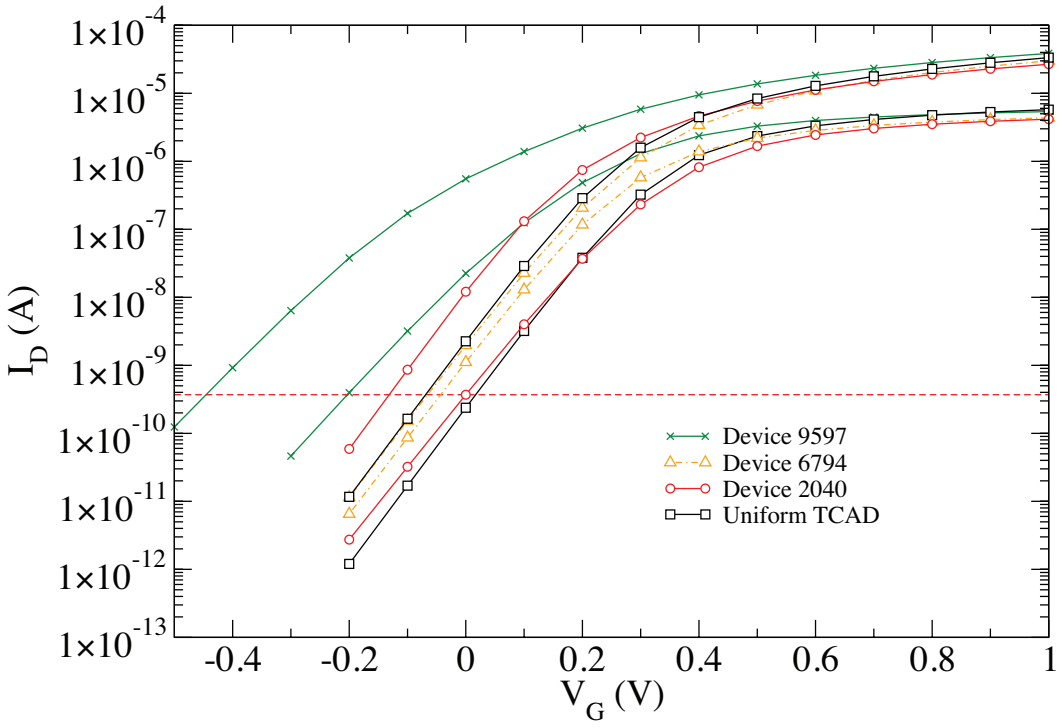


Figure 4.28: Transfer characteristics of the designed device and three extreme performance devices at high drain and low drain bias.

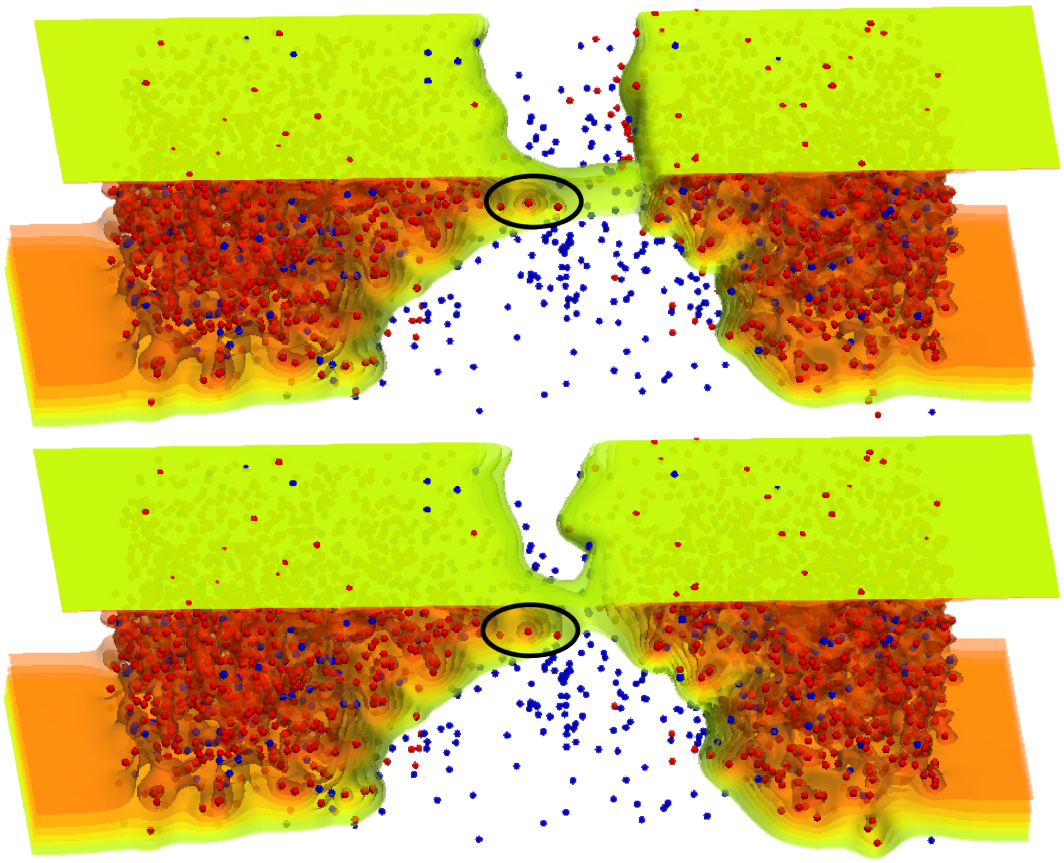


Figure 4.29: Electron concentration contours for nMOS Device 9597 at high drain (top) and low drain (bottom) at a constant current criterion. Acceptor dopants (blue) and donor dopants (red) are also shown. The source is on the left and drain is on the right.

donors.

4.4.2 Device 2040

Perhaps the most interesting of the anomalously behaving transistors, and the one with the most counter-intuitive behaviour, is device 2040. This device displays close to expected characteristics, aside from a larger DIBL than the uniform transistor, the high drain voltage subthreshold slope is steeper ($68mV/dec$) than low drain ($85mV/dec$). Figure 4.30 shows the electron concentration contours of the device at threshold voltage at high and low drain bias conditions.

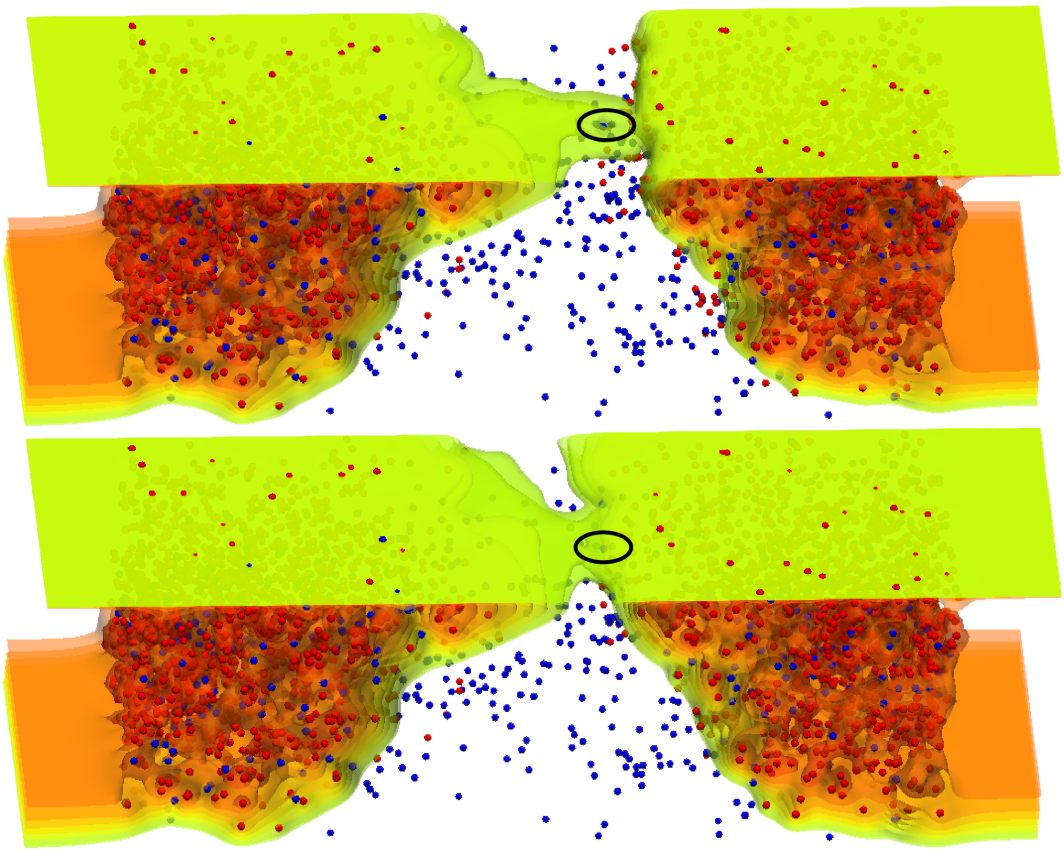


Figure 4.30: Electron concentration contours for Device 2040 at high drain (top) and low drain (bottom) at a constant current criterion. Acceptor dopants (blue) and donor dopants (red) are also shown. The source is on the left and drain is on the right.

At both drain bias conditions the device has a percolation path at the same position. However at high drain, where the depletion region between the drain and channel expands, the percolation path becomes narrower due to the critical role of few strategically placed acceptors in the vicinity of the potential barrier maximum, which is shifted towards the source end of the channel. This acts in concert with a strategically placed acceptor near the drain end of the percolation path that inhibits the penetration of the drain field into the percolation channel. This improves the gate control in the critical region of the potential barrier maximum, improving the subthreshold slope.

4.4.3 Device 6794

The transfer characteristics of device 6794, also shown in Figure 4.28, indicate that despite the very small DIBL – only $30mV$ –the transistor has a low drain threshold voltage below the threshold voltage of the uniform transistor, resulting in an increase in leakage. Figure 4.31 shows that the region of the percolation path is an area where there are few dopants, which leads to a reduction in the gate control of the current flow in this region. A current percolation path forms at both high and low drain voltage. However at high drain voltage three acceptor dopants near the drain are exposed and dramatically reduce the impact of the drain voltage on the potential barrier along the current percolation path and which leads to the extremely low DIBL in this transistor.

In all cases the extreme behaviour can be explained by the presence of only few (1-3) strategically placed acceptors or donors. This shows that a very small number of dopants can dramatically alter the characteristics of transistors at the $20/22nm$ CMOS technology generation and the accurate resolution of every dopant in 3D simulations is of great importance in order to correctly predict the statistical behaviour of a device technology. Finally we have highlighted the importance of random dopants at the drain end of the channel, which can alter the electrostatic influence of the drain bias.

4.5 Subsampling Issues

Subsampling is a problem specifically associated with statistical compact modelling in the presence of variability. The fact that only a finite number of devices can be measured or simulated results in a limited number of variable transistors available for the purpose of circuit simulation using a direct substitution method [81]. The result is that, simulation with a finite set of transistors can affect the circuit performance, but its impact is dependent on the number of devices available in the model library, circuit size and number of simulations. For example an inverter circuit simulated with a library of 10 n-channel and 10 p-channel transistors will only produce a possible combination of 100 unique

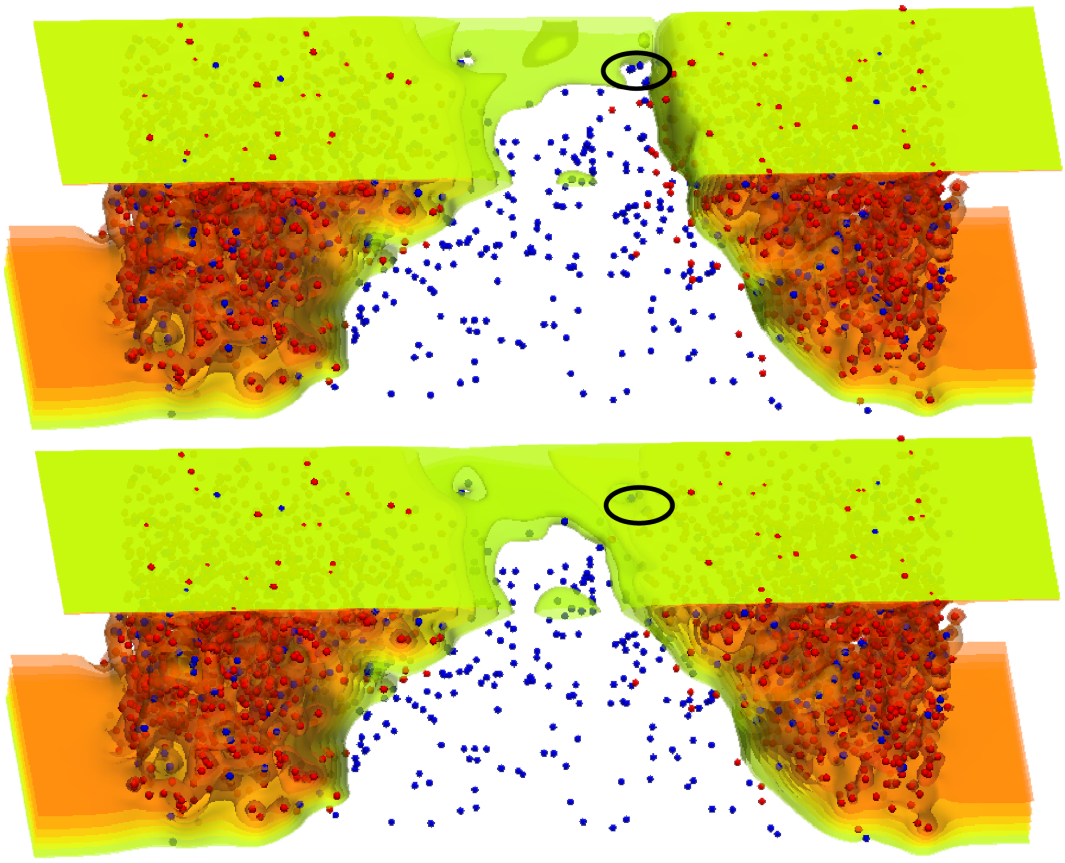


Figure 4.31: Electron concentration contours for Device 6794 at high drain (top) and low drain (bottom) at a constant current criterion. Acceptor dopants (blue) and donor dopants (red) are also shown. The source is on the left and drain is on the right.

inverters. Therefor if 1000 simulations are performed the output variable will be a poor representation of the expected output due to the small sample used.

In order to quantify the errors introduced through subsampling we create a number of lookup table libraries with 200 devices and 1000 devices, and compare with circuit performance distributions from simulations with 10,000 models and a set of "infinite" NPM generated devices. The relative accuracy of the simulations will indicate the requirement for compact model generators.

As we expect the problem to be most obvious in a small circuit, our "worst possible case" is selected to be inverter pulldown time, using a minimal sized inverter ($W_{pu} = 2 \times L$; $W_{pd} = 1 \times L$), where the pulldown transistor is minimally sized. If there is a limited number of transistors in the library we expect to see two separate effects:

- Bounding - the best/worst performance will be defined by the best/worst device in the system. Further to this near the tails there will be binning especially if the simulation sample size is much greater than the device ensemble.

- The circuit performance will be biased depending on how representative the sample of devices (the sample distribution) is of the overall device performance distribution (the population distribution).

As the device ensemble increases, the probability that the sample distribution will accurately represent the population distribution increases. This effect is illustrated in Equation 3.6, where we see standard error in the mean reduces as $\frac{1}{\sqrt{n}}$, where n is the number of independent samples.

Inverter pull-up delay is defined as the time between the input transition changing by 50% (high to low) of V_{DD} to the output changing (low to high) by 50% of V_{DD}. The simulation results are shown in Figure 4.32. 10,000 inverter simulations are performed with each library. The NPM generated results are treated as the reference, as they use an essentially unlimited set of generated devices, and are verified against the performance of the simulations directly using the 10,000 model sets, and are expected to accurately reproduce the population distribution. We see that the simulations with 200 models capture the

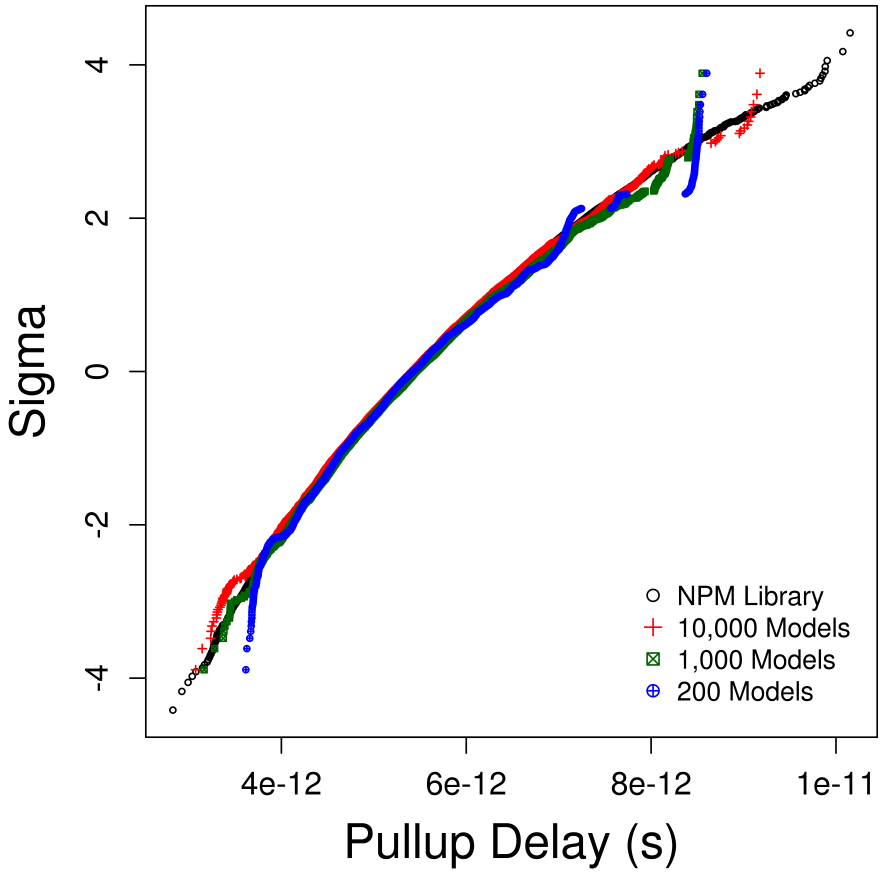


Figure 4.32: Inverter pull-up delay simulation results, using NPM simulations as a benchmark we see that 200/1000 model simulations accurately represent the distribution around $\pm 2\sigma$, however show binning and bounding in the upper tail.

mean of the distribution well, and are reliable out to $\sim 2\sigma$, however, beyond this point we see discontinuities, especially in the upper tail of the distribution, which represents binning. There is also bounding in the distribution – not surprising as we have performed 10,000 simulations with only 200 pull up transistor models. Simulations with 1000 models also capture the mean of the distribution well and push the failure point of the simulations closer to $\sim 3\sigma$, however subsampling again becomes evident in the tails of the distribution. Although 10,000 model simulations capture the distribution well, we see that there is some discontinuity in the tails. Also we see that NPM generates more extreme devices and produces a more continuous distribution, clearly demonstrating that subsampling can lead to large errors and artefacts, especially in the tails of circuit performance distributions. This could be particularly problematic in circuits like SRAM where it is desirable to simulate deep into the tails in order to study rare event failure and yield predictions.

The results clearly show the need for accurate compact model generation strategies capable of reproducing continuous distributions of device behaviour. As demonstrated, even though this statistical extraction strategy produces accurate models, problems associated with subsampling limit the applicability of the simulation using just the extracted compact models in high sigma and rare event analysis.

4.6 Statistical model Generation Accuracy

In order to avoid the problems associated with subsampling demonstrated in the previous section we consider compact model generation strategies which produce an effectively infinite ensemble of devices which have the same statistical properties as the extracted models. The strategies we will be comparing in this section are all outlined in Section 3.5, and include Gaussian V_T as well as NPM and PCA based on the extracted compact model ensembles. The accuracy of these generation strategies will be benchmarked against a large simulated device ensemble using the most important device figures of merit for reference.

4.6.1 Gaussian V_T Generation

The Gaussian V_T methodology, outlined in Section 3.5.1, is the most common method of of introducing statistical variability into circuit simulation and design evaluation in the presence of variability [52, 68, 69]. It involves generating a unique threshold voltage value for each transistor in the circuit, based on a Gaussian distribution extracted from the underlying technology. This is an effective way of estimating the first order effects of variability through a method entitled *idealisation of statistical chaos in a single variable [52],* as it assumes all effects of statistical variability manifest as a shift in threshold voltage, thus can be captured in a single parameter. Gaussian V_T is popular as it can be easily implemented, and due to the fact that it greatly simplifies analytical techniques. The Gaussian V_T methodology has several limitations, including the incorrect assumption that threshold voltage variability is Gaussian distributed [136], as well as the inability to capture the impact of variability effects on on-current, off-current and perhaps most importantly DIBL, which has been shown to have a significant impact on SRAM performance [85]. In order to demonstrate the validity of the Gaussian V_T methodology we compare the distribution of device figures of merit from the previous 10,000 extracted device ensemble with 10,000 devices generated using a Gaussian V_T method. This comparison is presented in Figure 4.33 in the form of QQ plots.

The figures show that MOSFETs generated using the Gaussian V_T methodology do not accurately reproduce the statistical range of behaviour of the underlying technology. Gaussian V_T does manage to capture the distribution of high drain on-current well. This is not surprising, however, as this figure of merit has a Gaussian distribution. We see the Gaussian V_T generated devices do not capture short channel effects well, as evidenced from the device DIBL, as there is simply no mechanism for introducing variability into the drain bias dependence of threshold voltage via V_T alone.

We also consider the correlations between generated device figures of merit and compare with those for the 3D device simulation data. The results are shown in Figure 4.34, and show an expected 1-to-1 correlation between all figures of merit of Gaussian V_T generated devices. This clearly does not match

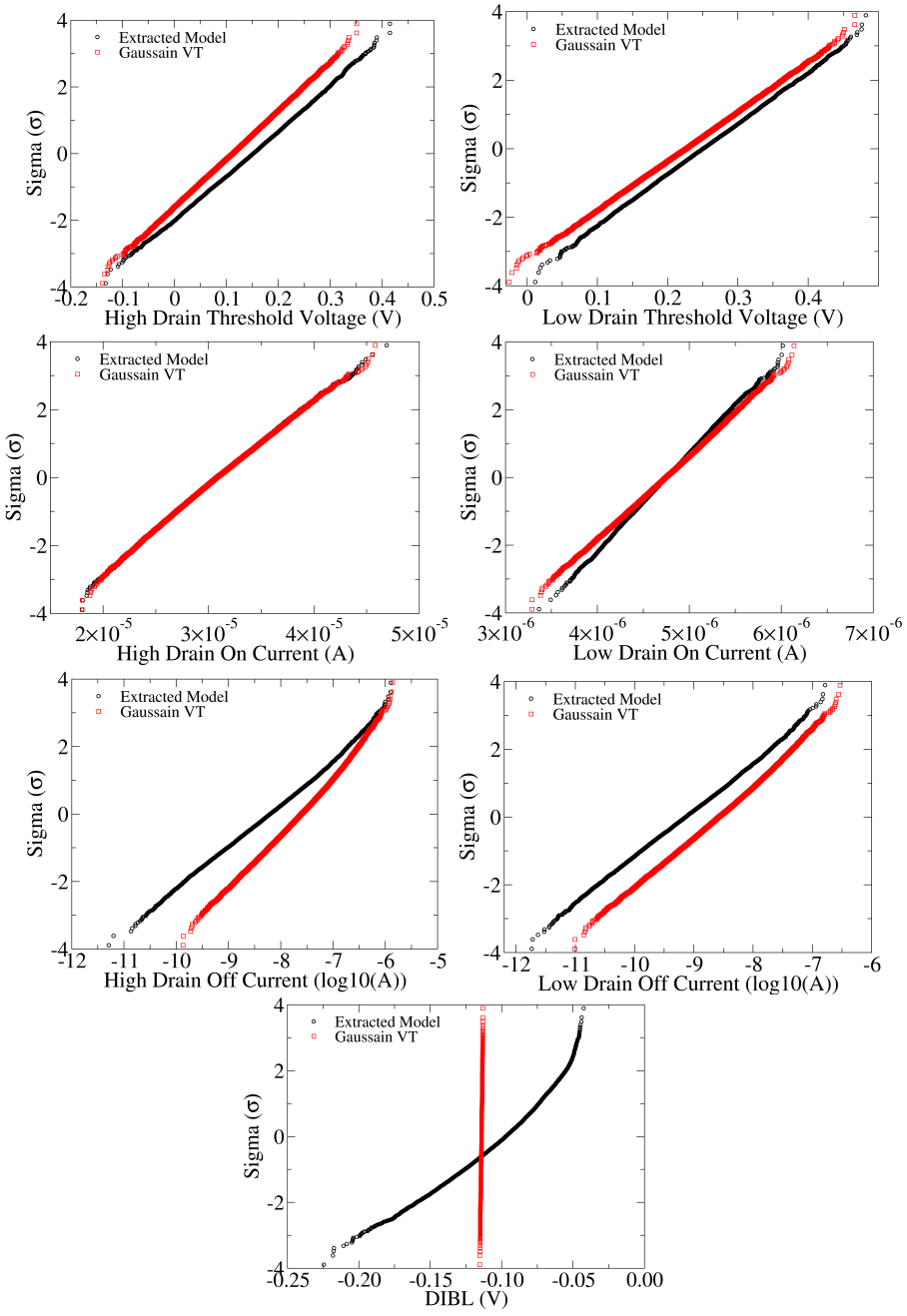


Figure 4.33: QQ plots comparing GARAND simulated device figures of merit with Gaussian V_T generated device figures of merit. The results show Gaussian V_T devices do not reproduce the figures of merit of the target data.

the simulated device data.

4.6.2 Principal Component Analysis Generation

The PCA generation approach is outlined in Section 3.5.3. PCA operates under the assumption that the individual extracted parameter distributions are Gaussian distributed, whilst retaining the correlation between the individual parameters. As we have seen in Section 4.3.1, this assumption is particularly inaccurate with many of the parameter distributions displaying a large amount of skewness and kurtosis. The impact of this incorrect assumption underlying PCA is difficult to predict, due to the large number of parameters and the complex correlations between them. In order to evaluate the errors introduced through the use of a PCA generation strategy we compare device figure of merit distributions from the 10,000 extracted device ensemble with 10,000 devices generated using PCA based generation. The resultant QQ plots can be seen in Figure 4.35.

The effect of the Gaussian distributed parameter assumption inherent to PCA is clearly seen in the parameter correlation plot shown in Figure 4.37 and the correlations between the figures of merit of the PCA generated devices shown in Figure 4.36. The figures show that PCA manages to capture the distribution of most device figures of merit well, however high drain off-current and DIBL distributions generated deviate significantly from simulated device data. This can be explained by considering the extracted parameter distributions, shown in Figure 4.26. The two parameters used to target these figures of merit ($CDSCD$ for high drain off-current and $ETA0$ for DIBL) both exhibit a significant amount of skew and kurtosis which *can not* be modelled with the PCA generation approach. Of greater concern is the distribution of low drain on-current, where PCA is producing devices which have extremely low on-current. In this case we are seeing the generation of devices which are non-physical or are approaching the boundaries of parameter applicability.

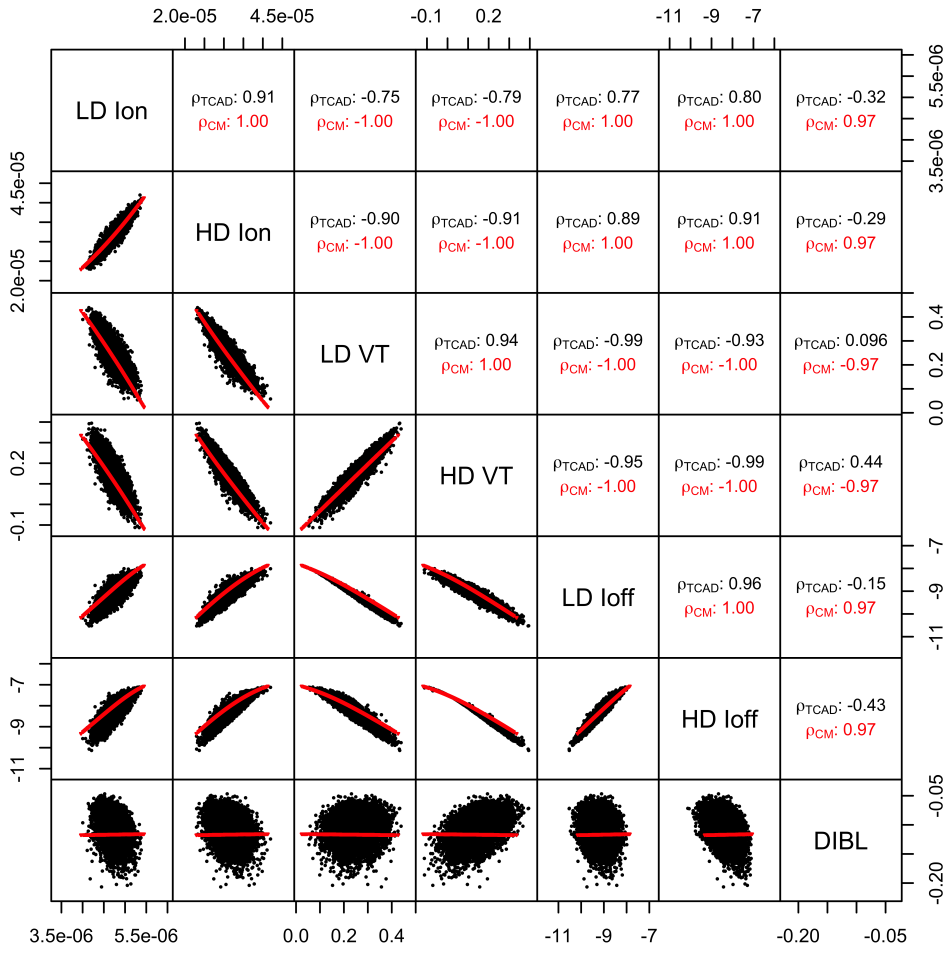


Figure 4.34: Correlations between device figures of merit, the black represents the 3D simulated device data and the red shows the Gaussian V_T generated compact model data, the bottom left of the table shows correlation scatter plots and the top right shows correlation coefficients, note the 1:1 correlation of all Gaussian V_T figures of merit.

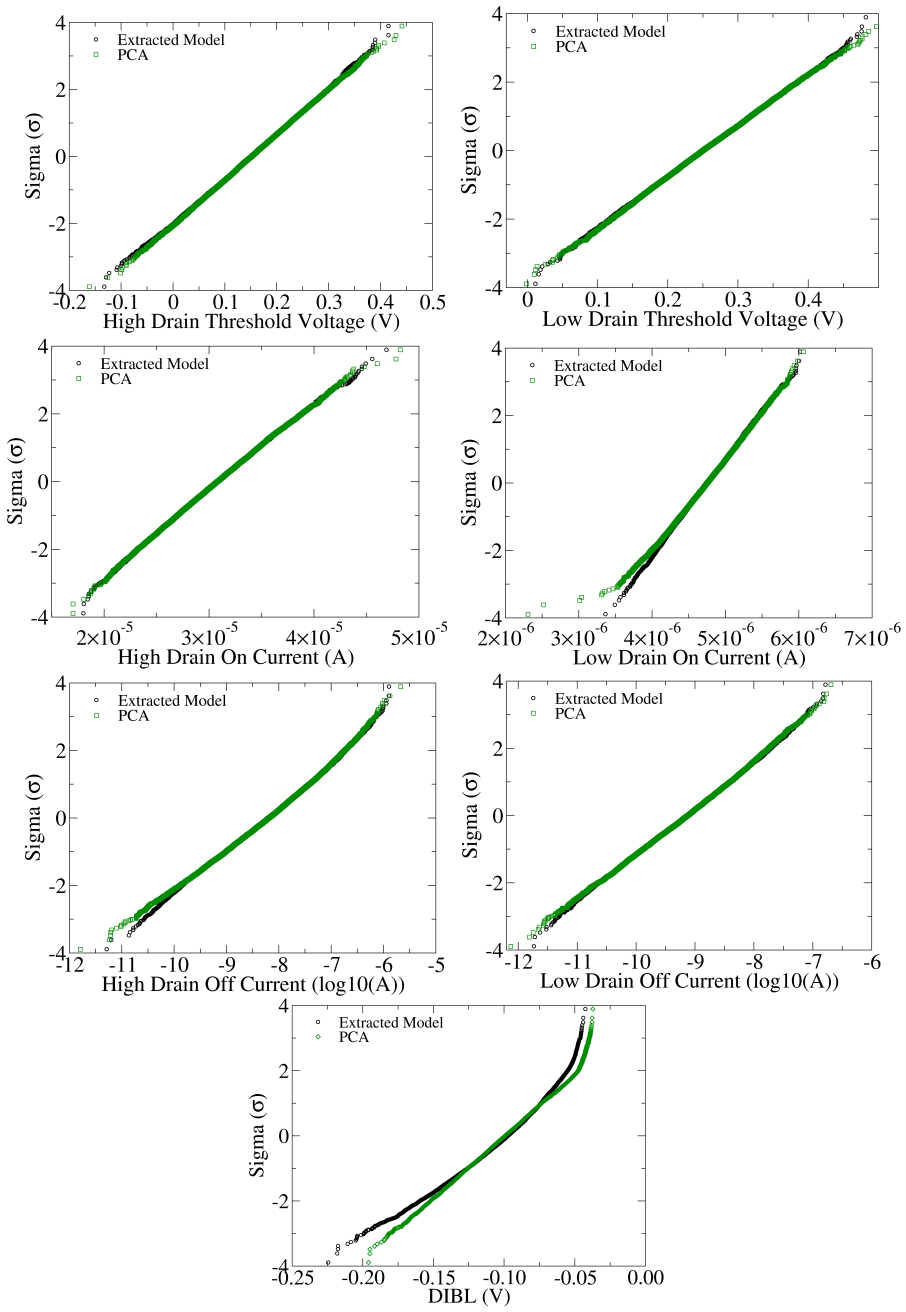


Figure 4.35: QQ plots comparing GARAND simulated device figures of merit with PCA generated device figures of merit, the results show PCA devices match relatively well over most figures of merit, however DIBL is not accurately captured and low drain on-current has some un-physical outliers.

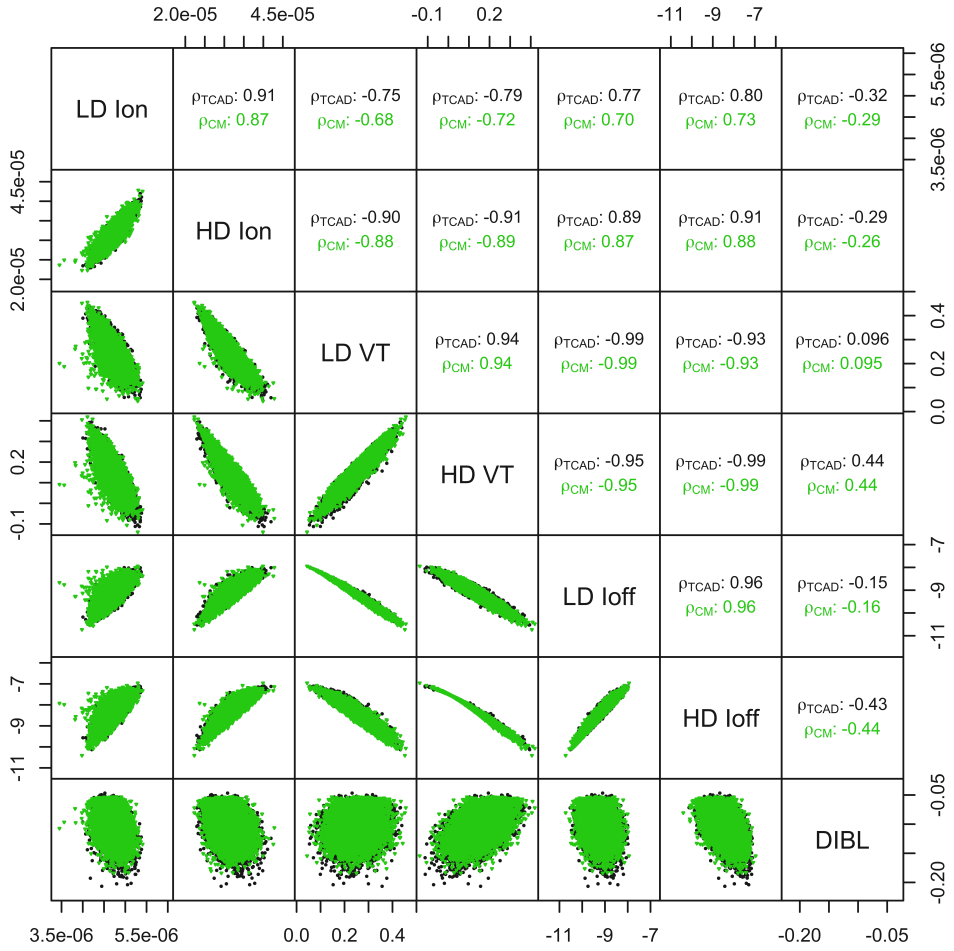


Figure 4.36: Correlations between device figures of merit, the black represents the 3D simulated device data and the green shows the PCA generated compact model data, the bottom left of the table shows correlation scatter plots and the top right shows correlation coefficients, note that PCA captures the correlation coefficient well, aside from some un-physically low on-current values at low drain bias.

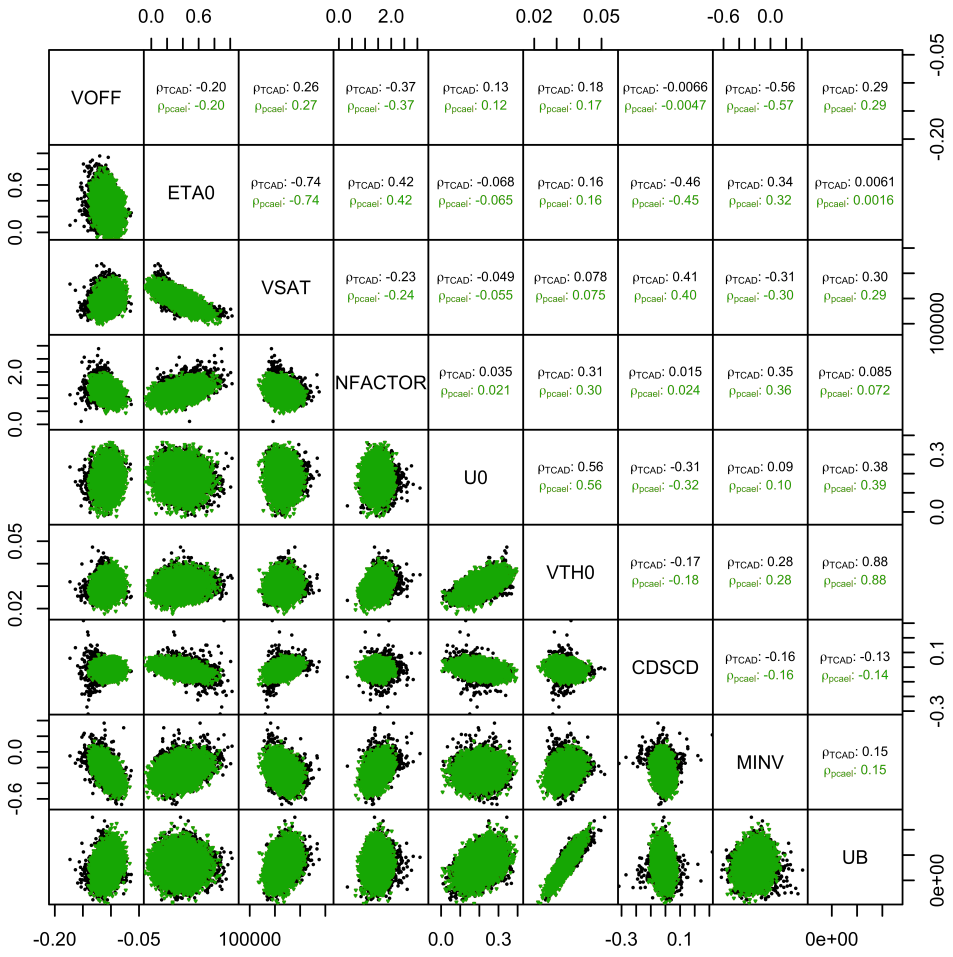


Figure 4.37: PCA correlation scatter plot and correlation coefficients, extracted parameter correlation scatterplot and coefficients are shown as a reference.

It is clear that the PCA compact model generation method fails to capture the complex non linear correlations between certain parameters, and due to this limitation can produce devices which are highly non-physical. It is possible that even a small number of such artificially extreme devices could significantly impact on circuit simulations introducing effects that are not representative of the underlying technology. These results show that for the extracted parameter set, where parameters are of non-Gaussian nature, PCA is not a reliable and robust generation methodology.

4.6.3 Non-Linear Power Method Generation

The NPM generation approach is described in Section 3.5.4. We consider NPM as it is able to capture the non-Gaussian nature of extracted compact model parameter distributions for nano-scale devices. In this implementation of NPM we consider the first four moments of the extracted parameter distributions and the correlations between them. 10,000 NPM generated devices have been compared to 10,000 simulated devices. The results of this analysis are shown in Figure 4.38, and the correlations between the figures of merit are also shown in Figure 4.39. The accuracy of the generated models is apparent from the data given in the NPM parameter correlation plot shown in Figure 4.40, which shows that NPM reproduces the complex non-linear correlation between the extracted parameters.

The results clearly demonstrate that the NPM method of compact model generation is capable of accurately reproducing the range of physical behaviour in a large statistical ensemble of devices in the presence of statistical variability.

4.7 Summary

In this chapter we have presented the results of a full set of 3D simulations of statistical variability in a 20/22nm CMOS bulk device technology including variability sources RDD, LER and MGG. The bulk of the original work in this

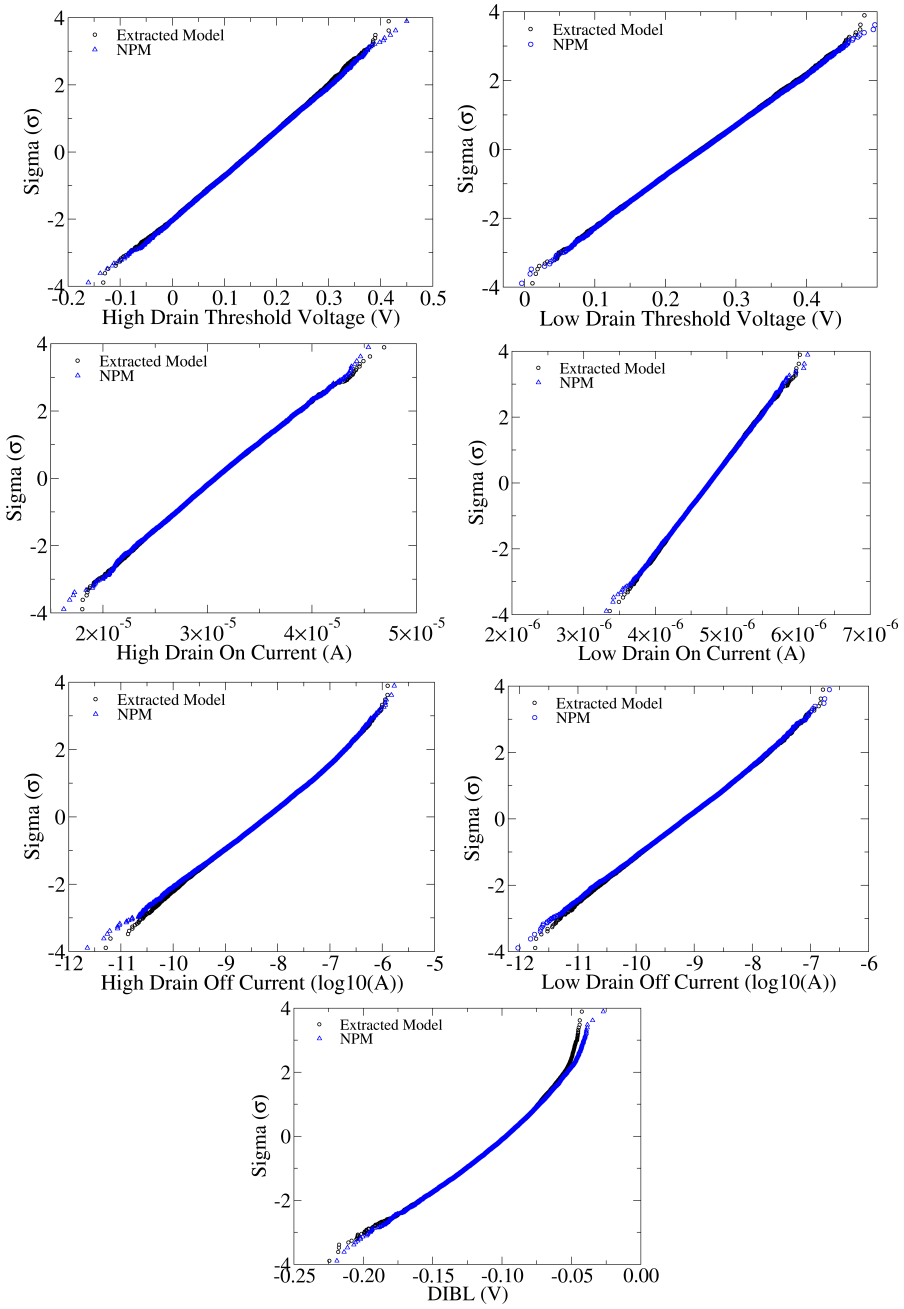


Figure 4.38: QQ plots comparing GARAND simulated device figures of merit with NPM generated device figures of merit, the results show NPM devices match well over all figures of merit. DIBL distribution struggles to match the lower tail as the compact model is unable to produce extremely low DIBL devices.

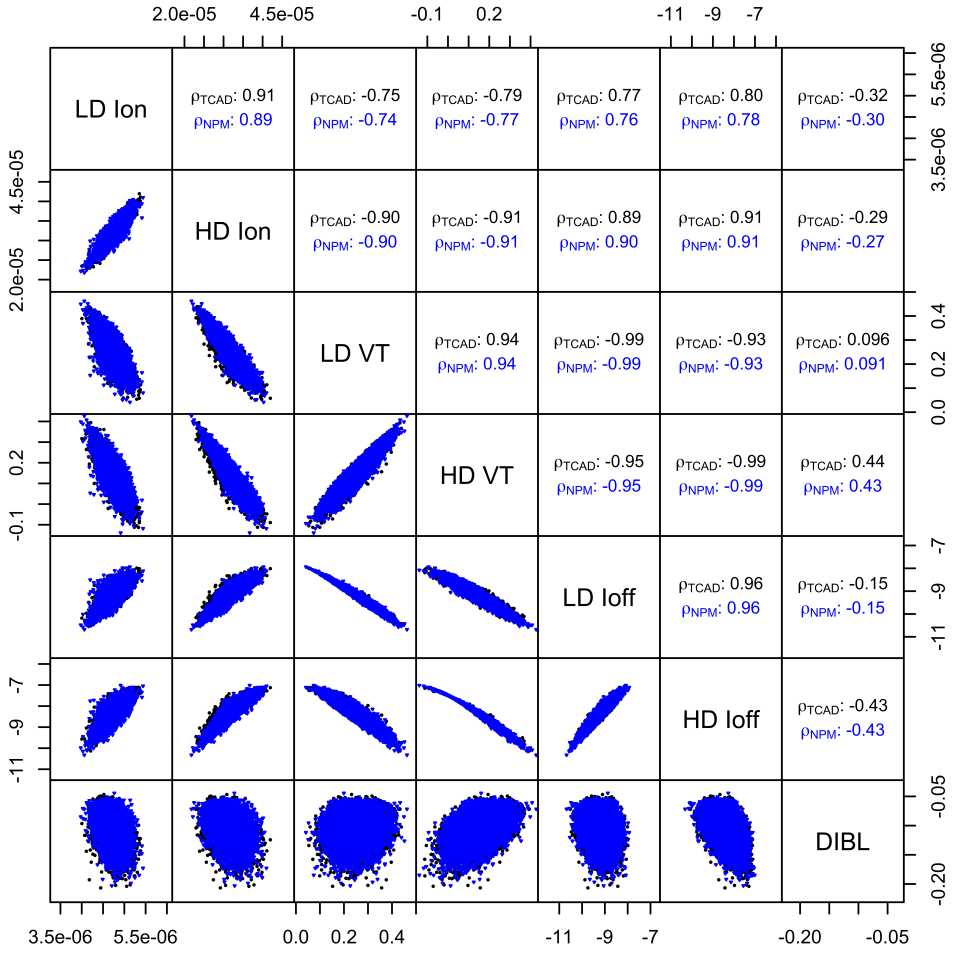


Figure 4.39: Correlations between device figures of merit, the black represents the 3D simulated device data and the blue shows the NPM generated compact model data, the bottom left of the table shows correlation scatter plots and the top right shows correlation coefficients, note that NPM captures the correlation coefficient well.

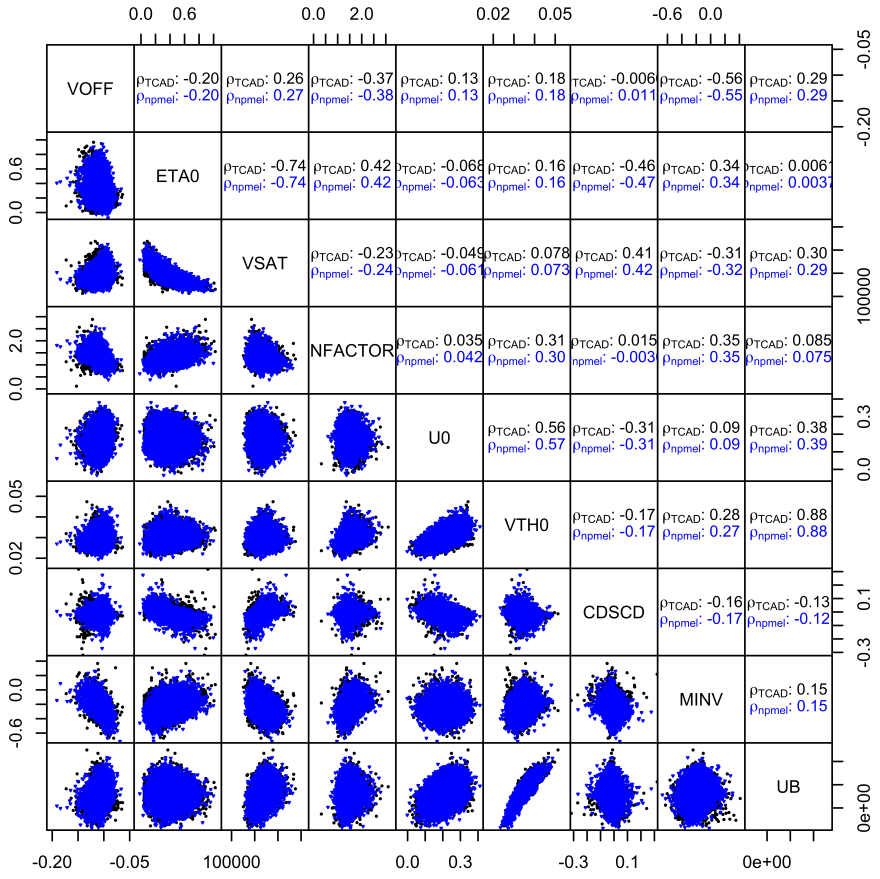


Figure 4.40: NPM correlation scatter plot and correlation coefficients, extracted parameter correlation scatterplot and coefficients are shown as a reference.

chapter involved the development of an accurate and robust compact model extraction strategy based on key transistor figures of merit. The extraction strategy is capable of accurately capturing the effects of statistical variability on device performance, and furthermore, is suitable for providing the necessary input to advanced compact model generation strategies such as PCA and NPM.

This compact model extraction strategy was applied to a large statistical dataset of 10,000 simulated devices and was shown to accurately capture all important figures of merit of device performance. The extracted compact model parameters were then used to inform statistical compact model generation using three strategies - Gaussian V_T, PCA and NPM. The devices generated using these strategies were benchmarked against the 3D atomistic simulations, and clearly showed that NPM generation is the only method capable of accurately reproducing the behaviour of the underlying devices.

At this point we have a statistical compact model extraction and generation methodology, with the ability to accurately generate an essentially unlimited number of devices, for the purpose of statistical circuit simulation. In the following chapters wen will use these methods to evaluate the impact of statistical variability on circuit and system performance. As SRAM is the most sensitive circuit with respect to statistical variability, we will be considering this first.

Chapter 5

Statistical SRAM Simulation

A significant portion, over 60% [6], of the chip area in modern System on Chip (SoC) applications is occupied by SRAM. Unlike digital logic circuits, where timing delay variations along the depth of a pipeline can typically average out, the SRAM system requires methods of correction and redundancy to overcome the SRAM cell's inherent susceptibility to statistical variability [103]. In order to increase SRAM density, foundry designers attempt to optimise cells until the smallest transistors can be found for that cell design, whilst providing the required yield. As, to first order, statistical variability is inversely proportional to transistor area, the minimal dimensions of SRAM transistors leaves SRAM cells significantly more vulnerable to statistical variability than random logic circuitry. The huge number of SRAM cells in modern memory arrays necessitates the simulation of SRAM performance up to and beyond 5 sigma as, assuming no methods of recovery or redundancy, which can be very expensive to implement, a single SRAM cell failure can a cause whole SRAM block to fail.

This chapter will outline the basic structure and operation of a standard 6 Transistor (6-T) SRAM cell, followed by the definition of figures of merit for the cell, as well as a description of the methods that will be used to introduce statistical variability into our SRAM simulations. Once the overall impact of statistical variability on the static operation of SRAM cells has been evaluated, results from the dynamic simulation of a figure of merit of particular indus-

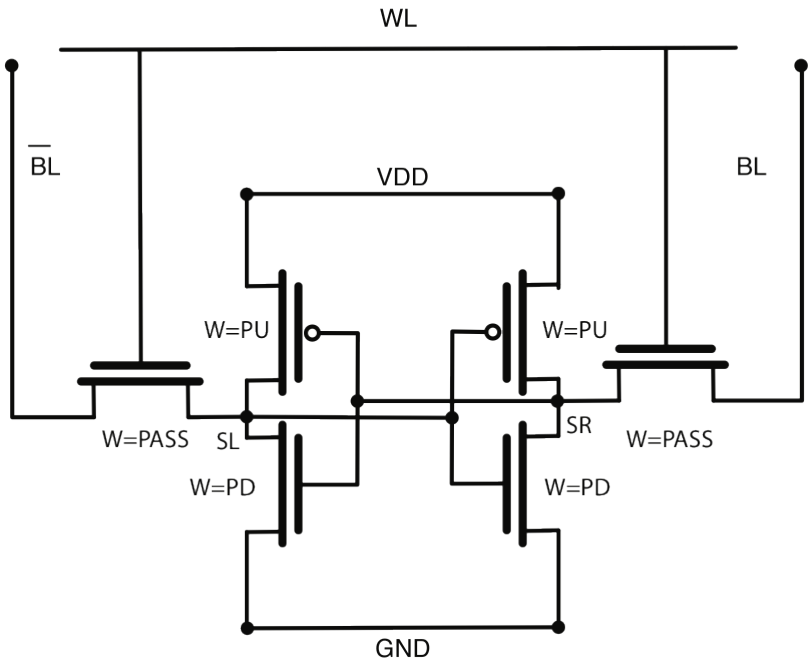


Figure 5.1: A 6-T SRAM cell transistor level schematic, *PU* denotes pull up, *PD* denotes pull down and *PASS* denoted pass transistors.

trial relevance will be presented. For the purpose of this simulation study, a full SRAM system, including peripheral circuitry, provided by memory design engineers at ARM Ltd, has been considered [137]. The results of these simulations in particular will demonstrate the industrial relevance and importance of accurate modelling of statistical variability at the circuit level.

5.1 The 6-T SRAM Cell

The schematic of a standard 6 Transistor (6-T) SRAM cell is shown in Figure 5.1. The cell consists of a cross-coupled inverter pair where the state of one node forces the state of the other, a configuration commonly known as a flip-flop (or latch). The operation is based on feedback between the input of one inverter and the output of the other inverter, and relies on symmetrical inverter pairs being balanced in their behaviour for optimal performance. Aside

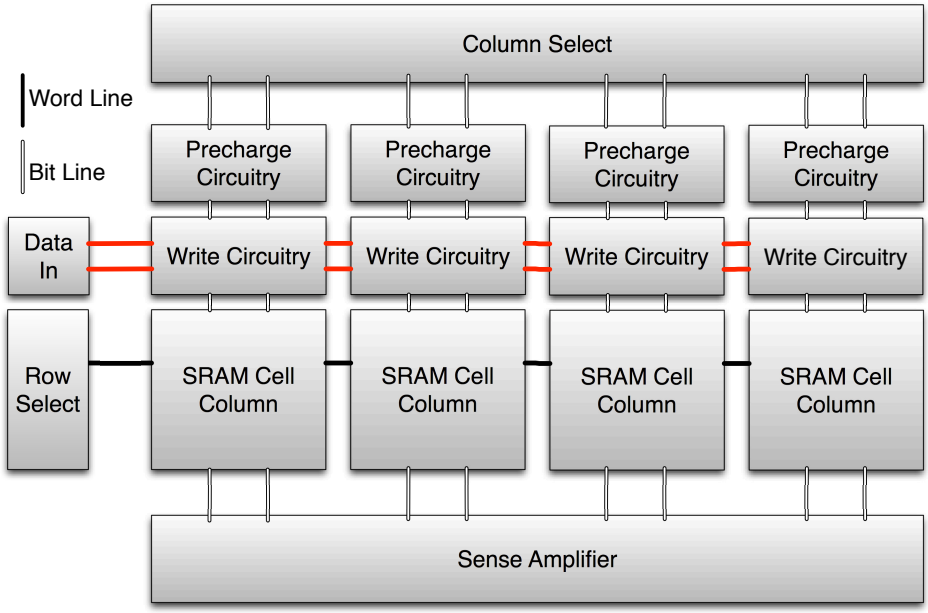


Figure 5.2: A block level SRAM system. Column and row select circuitry is driven by addressing circuitry which selects the required cells. Data in and write circuitry is used to write to cells and is disabled during read cycles when the sense amplifier outputs the stored data. Not shown is the clock circuitry or word line driver which determines the word line pulse width.

from the cross coupled inverter pair, two *pass*, or *access*, transistors connect external data lines to the internal nodes of the cell. The gates of the two pass transistors are controlled by the 'word line' (WL), with the pass transistor channels connecting the internal cell nodes to the bit lines (BL) of the column of cells. The bit lines are the external contact points to the bit cell where information is read from or written to the cell. To insure that the access time is equal for both nodes it is desirable for the pass transistors to be well matched, a requirement which can be significantly impacted by statistical variability.

Aside from the SRAM cell itself, there is a significant amount of peripheral circuitry including: word line pulse generation, addressing logic, sense amplifier, pre-charge/line buffer and multiplexer circuitry, all of which should be defined in order to accurately assess SRAM operation. A block level schematic of an SRAM system is shown in Figure 5.2, which makes clear its regular

structure. It is known that statistical variability impacts these peripheral circuit elements, however unlike the SRAM cell itself, this circuitry is mostly digital logic which is more strongly affected by process variability than statistical variability (this will be clearly demonstrated in Section 6.2). As such the impact of variability associated with the peripheral circuitry can be effectively estimated using process variability 'corner compact models'. The only exception is the sense amplifier, which sometimes includes a current-mirror or similarly sensitive circuit which can be highly susceptible to statistical variability. However, within SRAM, systems the sense amplifier is required to operate quickly and drive the output of the memory block, so very wide transistors are typically used (typical width for a 32nm technology SRAM cell transistor can be between 50 and 200nm, while sense amplifier transistors may be a wide as $1\mu m$). As device variability is relative to $\frac{1}{\sqrt{W}}$, where W is device width, extremely wide devices have a small amount of statistical variability. For these reasons, and for reduced simulation complexity, we will not be considering statistical variability in any of the SRAM array surrounding circuitry.

It has been shown that in sub 100nm technology generations statistical variability has a significant impact on SRAM yield and performance [52, 16, 68, 69], and it is clear that this must be factored into the SRAM design process from a very early stage [138]. The main reason behind the strong impact of statistical variability is rooted in the operation of the SRAM cell itself. An ideal SRAM cell is symmetrical, however, with increasing statistical variability, there is an increase in the mismatch between the cross coupled inverters and pass transistors, which can lead to one of many possible fail states for the cell. These include:

- stability failure - where the cell changes state when it should be 'holding'. This can occur due to thermal noise when the cell is idle or as it is being read.
- write failure - where the cell is unable to change state within the write timing requirements.
- read failure. The read operation starts with both bitlines pre-charged to

'1'. The side storing a '0' then has to cause a drop in the relevant bitline voltage. If the cell cannot create a large enough difference in the bit line voltages, which can be captured by the sense amplifier within the read timing requirements, the read operation fails.

- leakage failure - if a cell design leaks too much the static power requirements of a chip may not be met. This is especially important in mobile and low power designs.

Accurately modelling the complex cell behaviour which leads to these failure modes requires accurate device level variability information. Comparison with measurements has already shown that, at the 65nm technology node, V_T based simulation methods are not sufficient to capture the effects of variability in an SRAM cell [84, 85]. To corroborate and quantify this at the 20/22nm technology node we will compare cell simulations where variability is introduced via Gaussian V_T, with NPM based simulations generated using the accurate statistical models already extracted in Chapter 4.

5.2 SRAM Simulation Methods

During the SRAM cell design phase, before the surrounding circuitry is designed, SRAM cell performance is evaluated through steady state simulation of basic cell properties. These simulations are significantly less computationally intensive than analysis of a cell's dynamic properties, and offer good insight into the designed cell performance and potential disadvantages. It is relatively easy to expand the basic SPICE level simulation methodology to evaluate the impact of statistical variability on key SRAM cell figures of merit through Monte Carlo simulation. Initially we consider some of these static figures of merit to evaluate the impact of statistical variability on a standard cell level, using different methodologies for compact model generation.

In a successful SRAM design there is an essential trade-off between cell stability, write-ability, read-ability, leakage and cell area which can only be accurately assessed in the presence of statistical variability via statistical simulation. Static figures of merit exist, like cell leakage current [139], multiple

variations of write margin (WRM) [140], Static Noise Margin (SNM) [141], read current [142] and Access Disturb Margin (ADM) [143]. However, for the purpose of this work we will consider SNM and read current as representative of the first order effects of variability on SRAM cell performance. These two figures of merit are selected due to ease of simulation and prevalence in literature, and are described below.

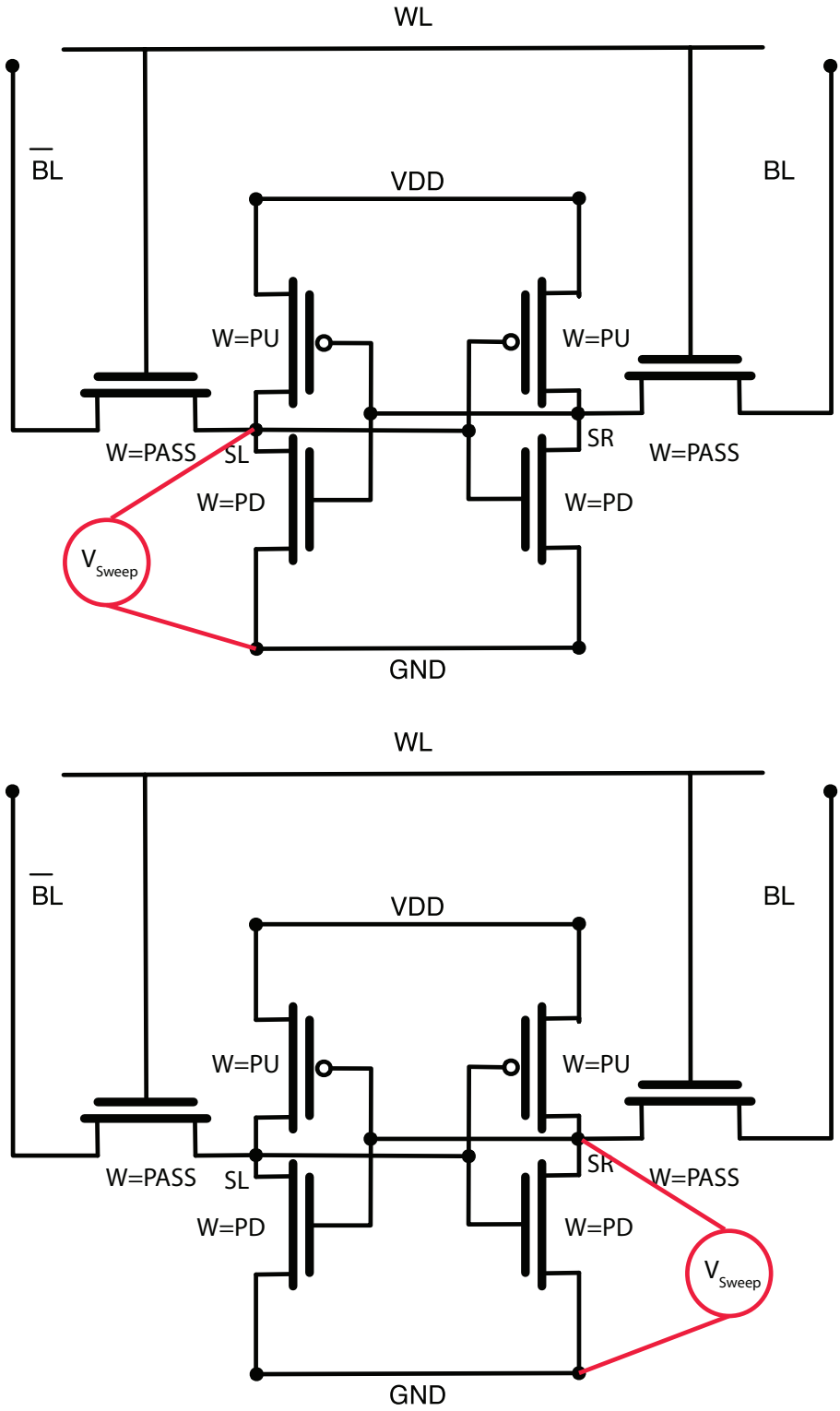
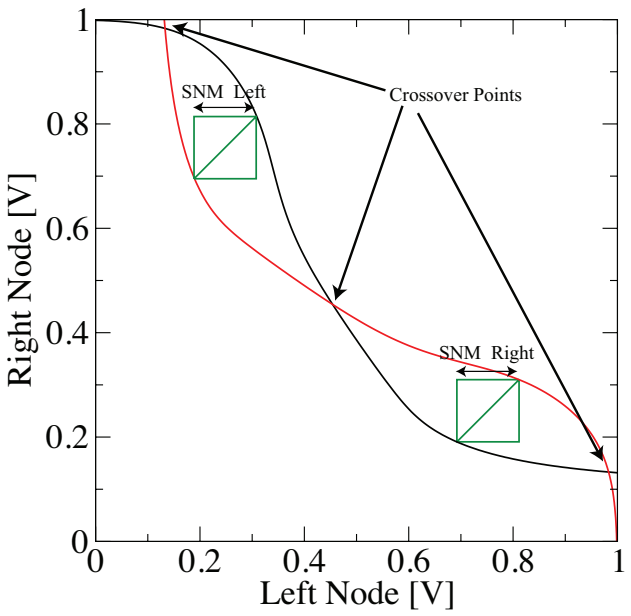


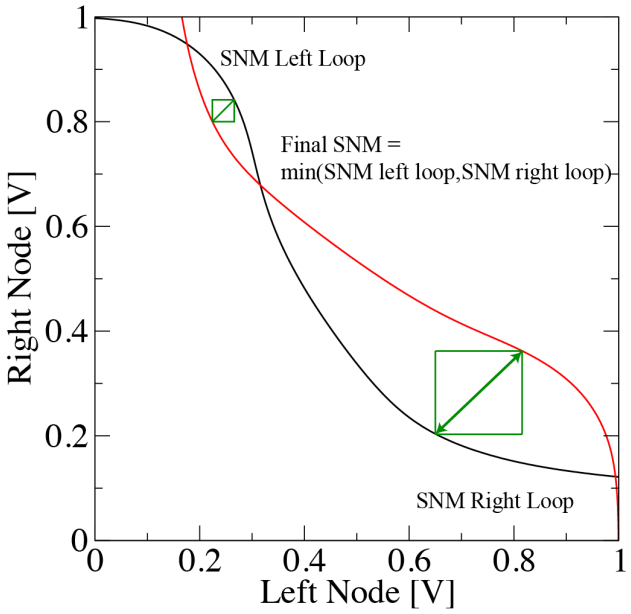
Figure 5.3: The two simulations required for SNM calculation, as the voltage on one node is swept, the voltage on the opposite node is measured.

5.2.1 SNM

SNM is a standard measure of the stability of a cell. Stability is worse when the cell is being read, and to simulate this, SNM simulation is performed with the bitlines and word line held high. The SNM voltage represents the maximum voltage which can be present, during the read operation, on either one of the internal nodes without causing the cell to change state. The SNM of a cell can be measured by connecting a voltage source to one of the internal cell nodes SL or SR, as shown in Figure 5.3. The applied bias from the voltage source is then swept from 0V to the supply voltage V_{DD} and the voltage at the opposite internal node is measured. This procedure is then repeated for the opposite internal node, the two simulations required are depicted in Figure 5.3. The largest square that can be fitted within each of the two loops of the "butterfly curve" (illustrated in Figure 5.4) is calculated and the static noise margin (SNM) of the cell is then defined as being the length of the side of the smaller of these squares as this represents the voltage required to cause the cell to change state. An example of the butterfly curves obtained during SNM measurement and calculation is shown in Figure 5.4. In a functional SRAM cell there are three possible cell states, storing a '1', storing a '0' or the cell metastable point (this is illustrated by the three crossover points in Figure 5.4 (a)). The first example shows an idealised cell without the presence of variability, this is indicated by the fact that the SNM square within each loop is equal and the crossover point between the two curves is at the same point within the x and y axis. This crossover point does not occur at $x = y = \frac{V_{DD}}{2}$ as the pull down transistor is much stronger than the pull up transistor in each inverter. This is due to the fact that reading a stored '0' is slower than reading a '1', as the bit lines are always pre-charged, so pull down transistors are significantly wider (and thus stronger) than pull up transistors. The second example shows a cell subject to statistical variability, this is obvious as the two 'loops' are unbalanced, and the crossover point occurs at $V(Left\,Node) = 0.32V$,$V(Right\ Node) = 0.68V$, this indicates the left node is far less stable than the right node.



(a)



(b)

Figure 5.4: SRAM SNM calculation (a) a cell without statistical variability with balanced transistors and a symmetrical 'butterfly curve', showing the three crossover points which represent the possible d.c. states for the cell, (b) a cell subject to statistical variability, with unbalanced transistors and asymmetrical 'butterfly curve'.

5.2.2 Static Read Current

Static Read current is a measure of the ability of the cell to cause a change in the voltage of the two bit lines (this must be large enough to trigger an output sense amplifier within the read cycle time for the cell to function). Read current is defined as the current flowing through the bit line to ground, through the node which is storing a '0' state. As the read operation is initiated with a '1' pre charged to both bitlines, the read current indicates the ability for the pull down and pass transistors to reduce this voltage sufficiently for the sense amplifier to detect the '0' stored inside the node that is being measured. The read current simulation setup for a bitcell is shown in Figure 5.5. A fail is defined as a read current which is not large enough to discharge enough of the bit line voltage for the sense amplifier to flip before the word line pulse passes. Traditionally, read current can be simulated with just the pulldown and pass transistors, however we simulate the full circuit to capture the impact of variability on all transistors in the circuit.

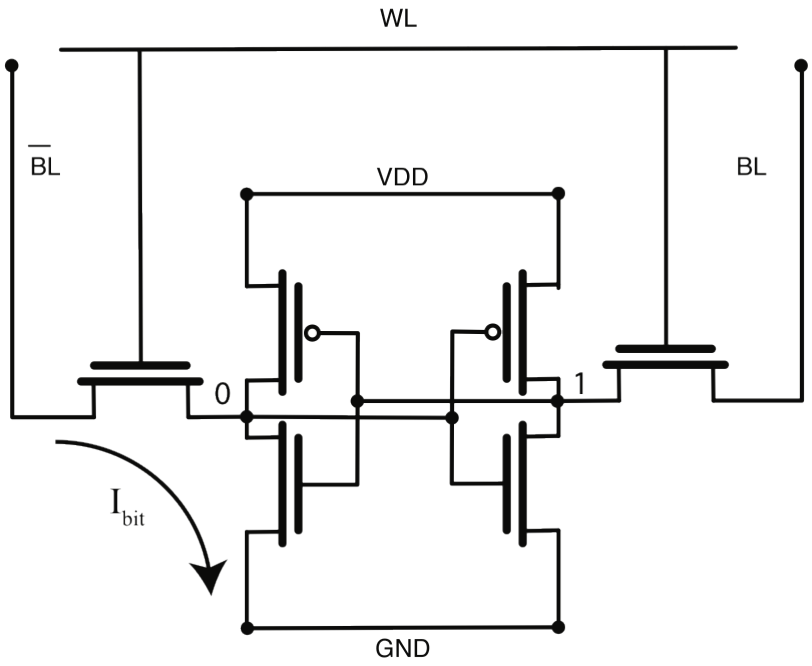


Figure 5.5: Read current definition. For the purpose of the simulation both the bit lines and the word line are held high.

Aside from the evaluation of circuit performance at nominal supply voltage, the V_{DD} dependence of these figures of merit is also important. The minimum safe operating voltage V_{DDmin} for a SRAM block determines the 'inactive' leakage of the SRAM block [144], as well as the amount of 'write assist' available [145]. Write assist involves the reduction of the supply voltage to the SRAM block. This reduces the stability of the cell, thus enhancing the write process. However, for the rest of the cells the reduction in V_{DD} during the write assist phase can be highly destructive.

Several techniques have been proposed to model the effects of statistical variability on SNM, read current and V_{DDmin}, however these are generally limited to modelling only the statistical variability of transistor threshold voltage [52, 68, 69, 146], or threshold voltage and uncorrelated DIBL parameters [87], and even at this level are based on the incorrect assumption that threshold voltage variability follows a Gaussian distribution [54]. In order to determine the impact of variability over the full range of device bias characteristics including the impact of variability on on-current, off-current and transconductance, we compare statistical SRAM SNM and read current simulations performed with statistically extracted BSIM4 models agains simulations which employ only threshold voltage variability.

Aside from the traditional static figure of merit simulations described above, the impact of different compact model generation strategies on predictions performance of the SRAM system will be investigated. For this purpose, we have been supplied with a test SRAM system schematic, including peripheral circuitry as part of a joint project with ARM Ltd. With the schematic of an entire SRAM system available, dynamic simulation can be performed as realistic timing simulations require the ability to model realistic input signals. Dynamic SRAM simulation of the whole system allows the analysis of realistic, full SRAM performance, including addressing, clock generation and sensing.

Initial simulations were performed with a minimal sized cell, based on the 20/22nm technology characterised in the previous chapter, with cell area $0.09\mu m^2$, scaled relative to Intel's 32nm $0.171\mu m^2$ cell [147]. However, as demonstrated by Figure 5.6, this cell was too sensitive to statistical variability, with SNM stability failures seen within the first 10,000 simulations even

at the typical/typical process corner. In order to study SNM as a function of V_{DD} at the 22nm technology node, a larger, more variability stable, 'high performance' cell was designed, with an increased cell area of $0.120\mu m^2$. The SNM distributions of the 'high performance cell and the 'minimal' cell are shown in Figure 5.6.

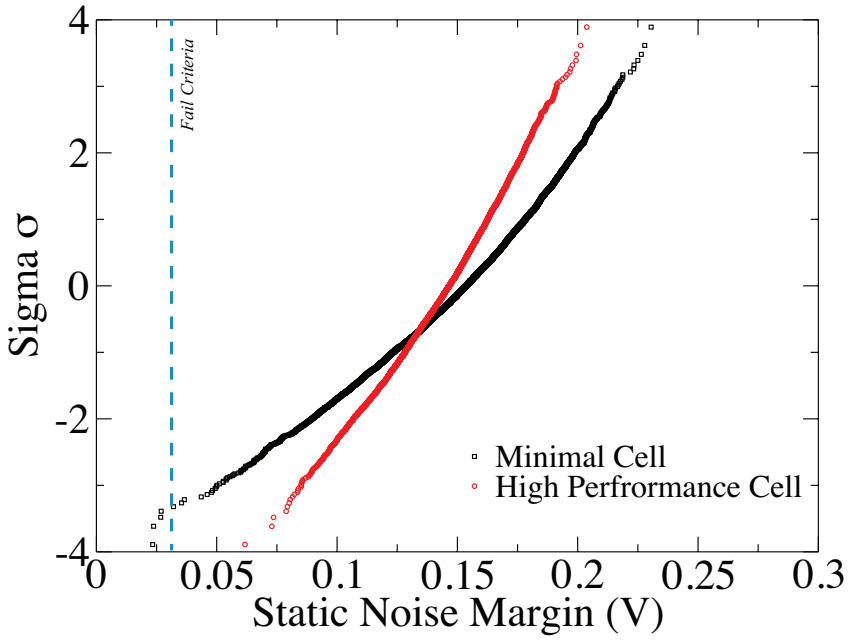


Figure 5.6: SNM of minimal cell and high performance cell, the fail criteria is set at 30mV.

5.3 SRAM Variability Simulations

Variability simulations in this section are performed using the GSS circuit simulation engine RandomSpice [108] with ngspice as a back-end simulator. The compact model and NPM generators used were described in Chapter 4, as well as the 20/22nm technology template transistor. The simulation methodology consists of generating randomised instances of a template netlist using each compact model generation strategy. This process is illustrated in Figure 5.7 and involves creating a copy of the template netlist where each transistor instance is replaced with a random compact model instance. This process is

repeated the required number of times to give an ensemble of statistically different circuits which are then simulated and their performance is evaluated. All of the simulations are performed using standard sampling and make no prior assumptions about expected circuit performance.

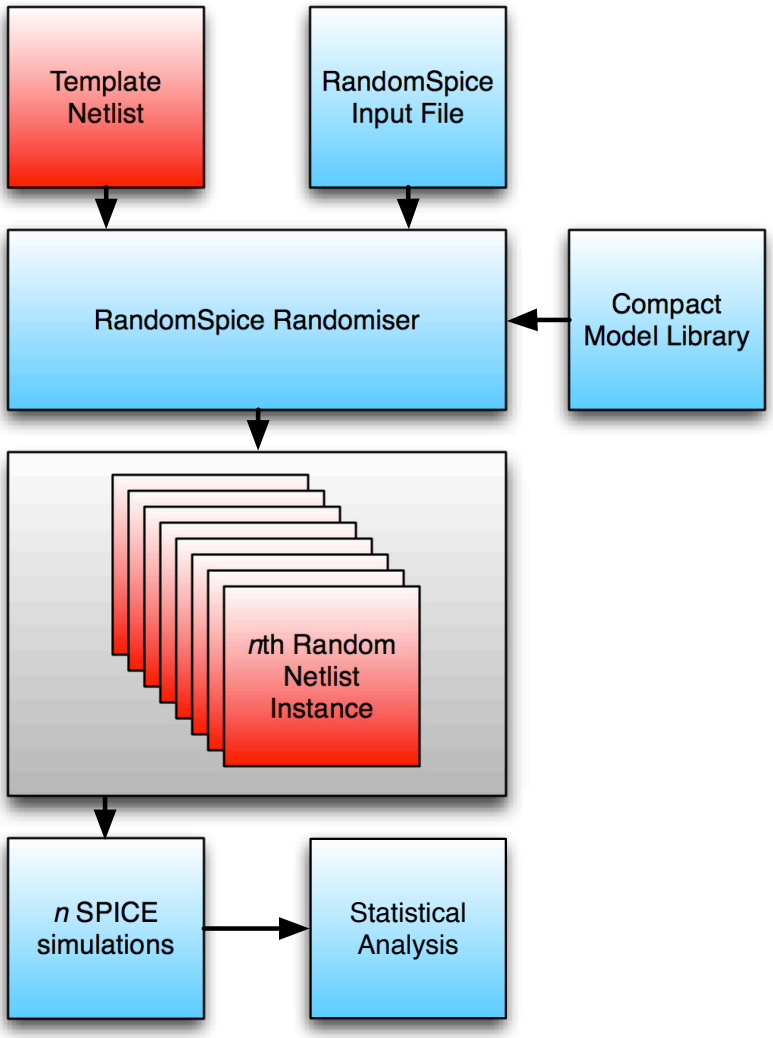


Figure 5.7: RandomSpice simulation flow, showing n statistically different circuit simulations. The different compact model generation strategies are introduced at a the compact model library level.

5.3.1 SNM Simulation

In order to evaluate the SNM performance of a cell design it is vital to characterise its response in the presence of statistical variability. While process variability can have a global impact on SNM performance, causing a slow drift across the SRAM block, across die and from wafer to wafer, the fact that statistical variability can introduce significant mismatch between the cross-coupled inverter pair within a single cell and on a cell-to-cell level, means that statistical variability can have a significant impact on SRAM SNM on a individual cell basis. As V_{DD} is reduced in a cell during the 'write assist' phase, or as a measure of reducing SRAM block leakage, especially during inactive phases, the minimum supply voltage at which all the cells in the SRAM block retain their state is an important metric for SRAM operation. We define minimum supply voltage at which the SRAM is still operable as V_{DDmin} and calculate this through SNM stability simulations. V_{DDmin} then becomes the minimum supply voltage for which the fail rate remains above a predefined minimum. It is obvious that V_{DDmin} will be highly dependent on the DIBL of the technology and, as is shown in Chapter 4, the extracted models and NPM generation strategy used in this study capture the variability in device DIBL accurately, where simpler methodologies such as PCA and Gaussian V_T fail to do so.

The distributions of SNM from 10,000 simulated SRAM cells generated from NPM and Gaussian V_T at a range of supply voltages $V_{DD} = 1V$ to $0.5V$ are shown in Figure 5.8. It is clear that even for as few as 10,000 samples, Gaussian V_T simulations under-estimate the effect of statistical variability on the SNM, compared to the NPM generated models which correctly capture additional aspects of statistical variability. The discrepancy between Gaussian V_T and NPM is worst at high supply voltage, due to the impact of DIBL, and at the lower tail of the distribution, which defines the worst performance of the circuit. This does not come as a surprise as in Section 4.6, Figure 4.33 we showed that Gaussian V_T generation cannot reproduce a distribution of DIBL. Although the deviation between NPM and Gaussian V_T simulated SNM appears relatively small in the Figure 5.8, the discrepancy increases as we move further into the tails of the distribution, which has a significant impact

on the estimation of SRAM yield.

A Generalised Lambda Distribution (GLD), is used to fit the distributions of SNM at $1V$ to estimate the failure rate of the cell at a give failure criterion. This is possible as a GLD can fit a wide variety of distributions and provides an analytical approximation of the probability distribution of the data. Although a GLD fitted distribution is limited by the accuracy of the simulated data, it is useful to highlight the difference between two distributions deep into their tails, which would otherwise take millions of simulations to reach. As Figure 5.9 shows, NPM simulations predict a failure rate of $\sim 5,000$ cells per billion, while the Gaussian VT simulations predict a failure rate of $\sim$ 400 cells per billion, indicating that, for this failure criterion, there is an order of magnitude difference in the predicted failure rate.

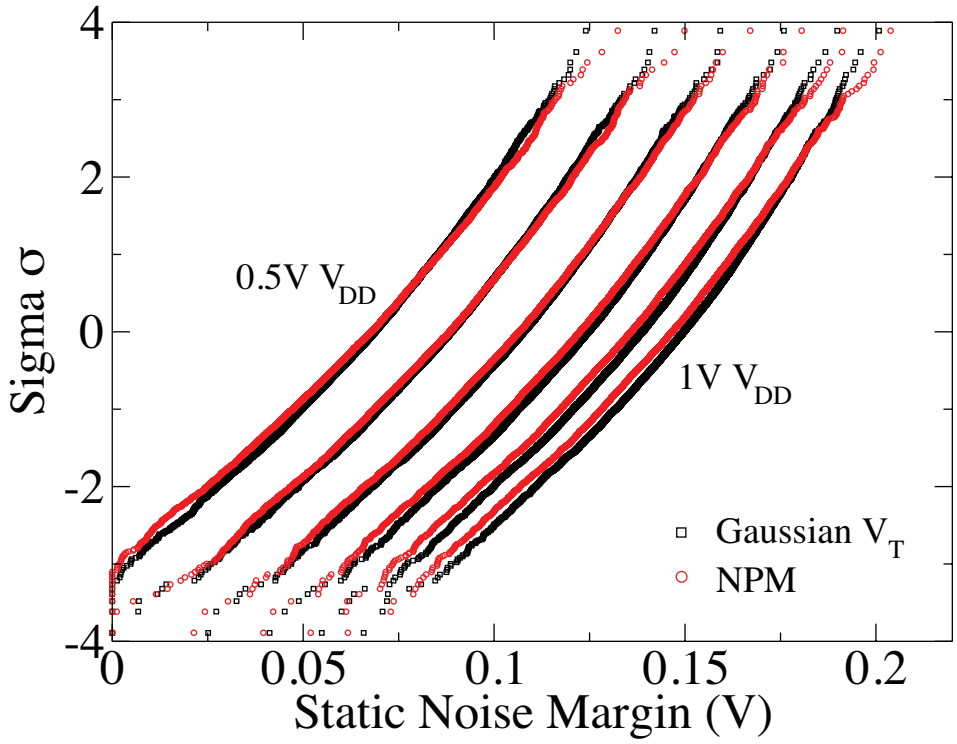


Figure 5.8: SNM at multiple V_{DD} levels - 1V, 0.9V, 0.8V, 0.7V, 0.6V and 0.5V, simulated with NPM models and Gaussian V_T models, showing a difference between the resultant SNM distributions.

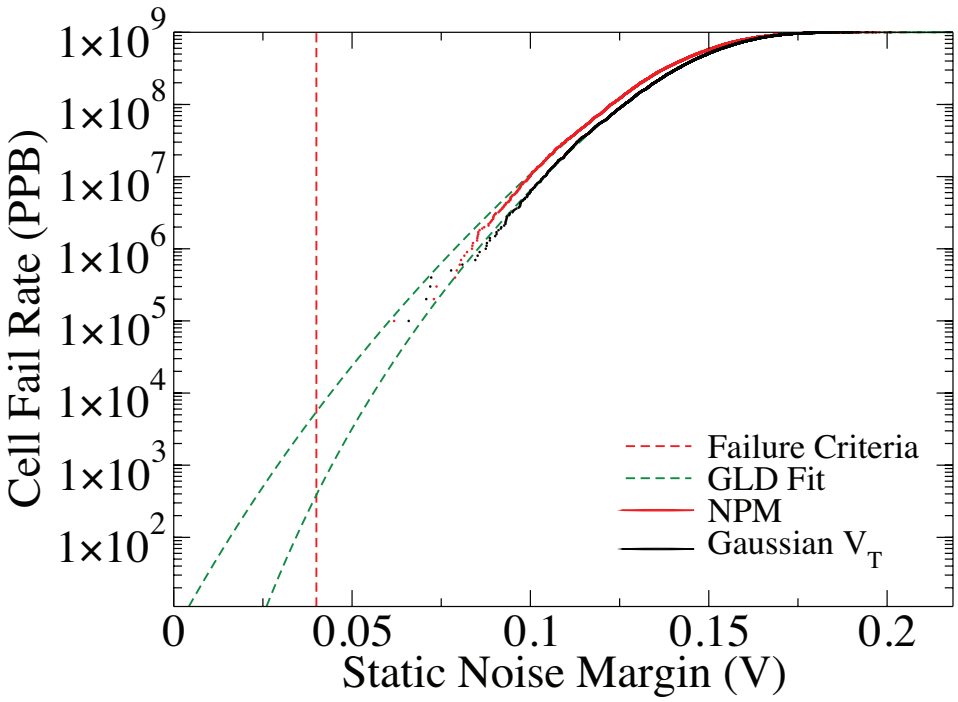


Figure 5.9: GLD based yield predictions at $V_{DD} = 1V$ showing difference between NPM and Gaussian V_T simulations.

As supply voltage is reduced, the correlation between the SNM at $V_{DD} = 1V$ and the SNM at the new V_{DD} steadily reduces from unity. This decorrelation is due to variability in the DIBL of the transistors. The NPM generated models accurately capture variability in DIBL, while Gaussian V_T distribution models have no way to capture this effect and model a constant DIBL for each device generated (this is illustrated in Figure 4.34). This divergence in the correlation coefficient, between the SNM at $V_{DD} = 1V$ and the SNM at this lower V_{DD}, is illustrated in Figure 5.10 and Figure 5.11. Figure 5.10 shows the evolution of the correlation coefficient between SNM at the nominal supply voltage and the reduced supply voltage, and Figure 5.11, which plots the SNM at $V_{DD} = 1V$ against the SNM at $V_{DD} = 0.5V$, shows the increased spread of SNM at low V_{DD} obtained using the NPM approach. NPM simulations show a correlation coefficient between SNM simulations at $V_{DD} = 1V$ and SNM simulated at $V_{DD} = 0.5V$ of 0.81, compared to a value of 0.93 predicted by Gaussian V_T simulations. The small amount of decorrelation present in the

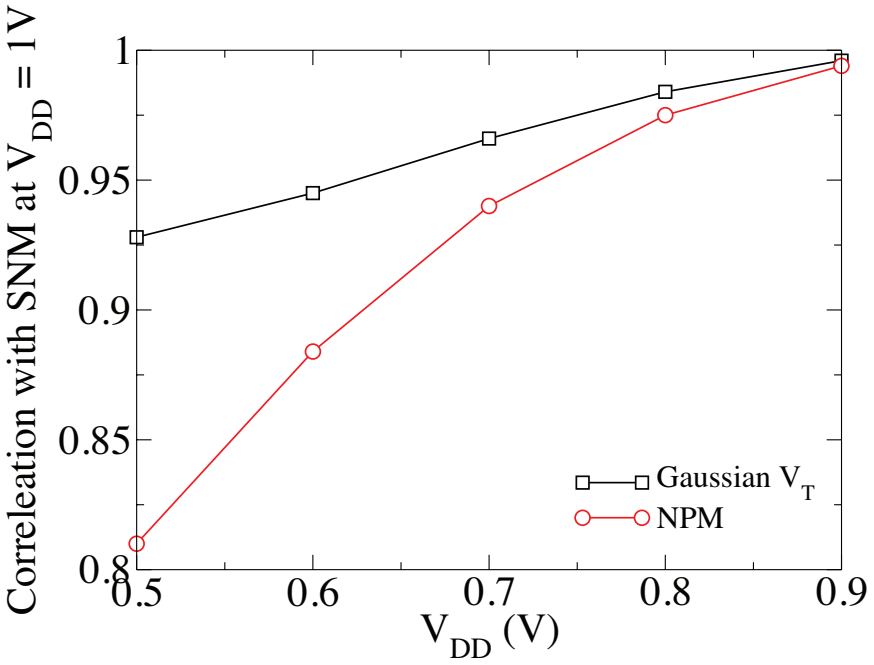


Figure 5.10: Correlation coefficients between SNM at $V_{DD} = 1V$ and lower supply voltages.

Gaussian V_T simulations is because, as supply voltage is reduced, transistor operation moves closer to the subthreshold regime and leakage current flowing through the pass gate starts to have an impact on SNM.

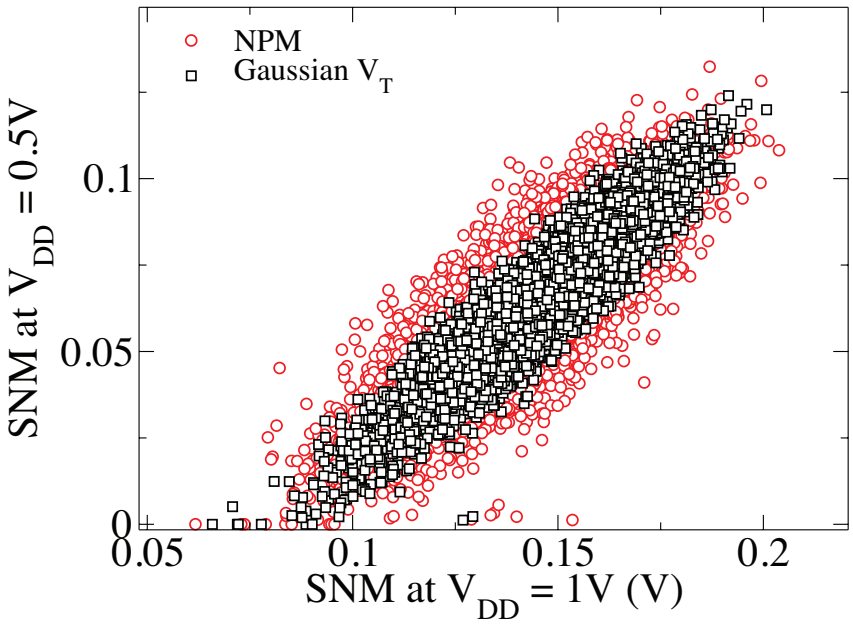


Figure 5.11: A scatter plot of SNM at $V_{DD} = 1V$ against SNM at $V_{DD} = 0.5V$.

To emphasise the error introduced through Gaussian V_T based simulation of SNM as a function of supply voltage, Figure 5.12 shows a set of non-extremal cells, chosen to have SNM in the range of 120-125mV at $V_{DD} = 1V$. As the supply voltage to these cells is reduced, the NPM simulations predict a much larger spread in SNM than the Gaussian V_T simulations. This result is most relevant when considering V_{DDMIN}. If V_{DDMIN} is estimated based on variability performance from Gaussian V_T simulations, SNM dependence on supply voltage will be incorrectly calculated.

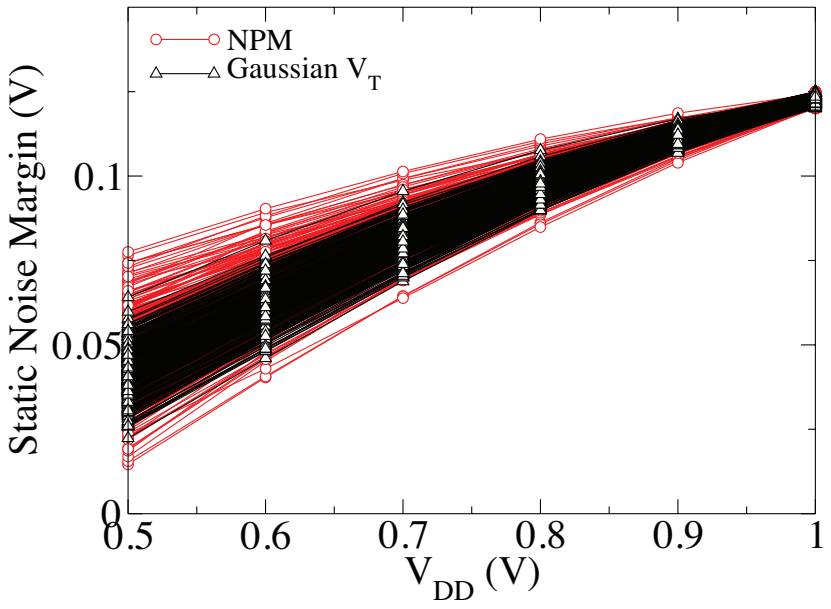
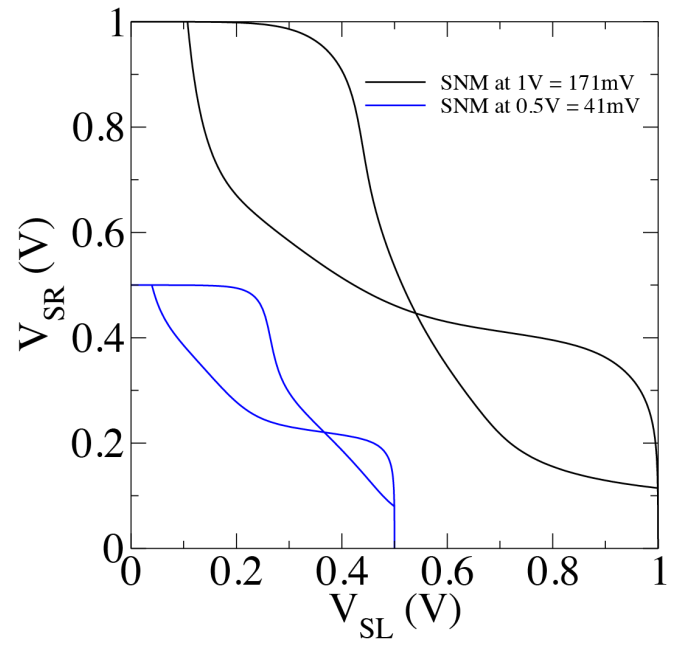


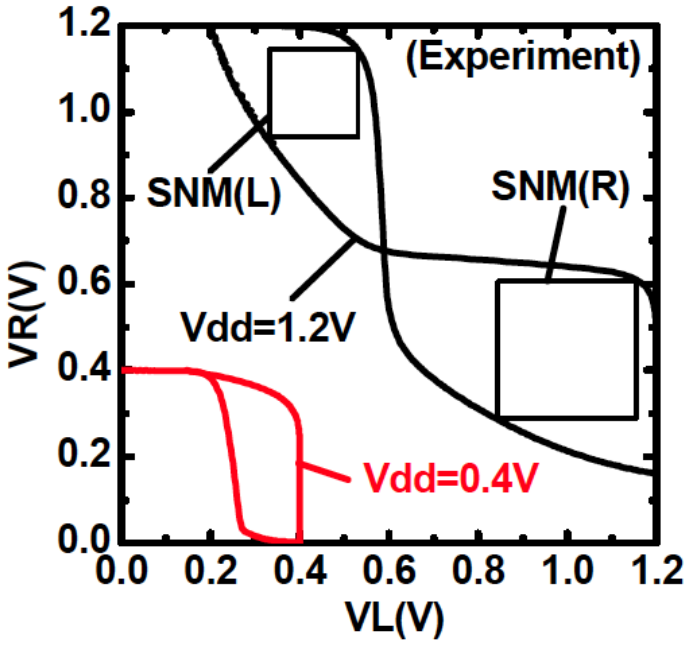
Figure 5.12: SNM as a function of VDD for cells with SNM between 120mV and 125mV at $V_{DD} = 1V$.

To illustrate cell imbalance (where one of the cross coupled inverters has a much weaker pull down than the other), and how it is impacted by supply voltage, Figure 5.13 (a) shows an instance of an SRAM cell simulated with the full compact model approach which has a very high SNM value of 171mV at $V_{DD} = 1V$. For cells of this range of SNM at supply voltage $V_{DD} = 1V$, Gaussian V_T simulations predict that the SNM of the cell at $V_{DD} = 0.5V$ will line the range of 80mV to 110mV (extrapolated from Figure 5.11). NPM simulations of the cell show that the actual SNM of this cell at $V_{DD} = 0.5V$ is 41mV, considerably lower than predicted from Gaussian V_T simulation. NPM simulation predicts that for a cell of this type, the SNM at $V_{DD} = 0.5V$ could be anywhere in the range of 40mV to 120mV, showing significantly greater spread. Another important effect to note is the change in shape of the SNM butterfly curve, from a well balanced cell at $V_{DD} = 1V$ to the heavily skewed curve observed at $V_{DD} = 0.5V$. This change in shape is due to the different amounts of DIBL in each of the transistors in the cell. Gaussian V_T simulations cannot reproduce this change in shape as the drain bias responses of all the transistors in the cell are identical. For reference, Figure 5.13 (b) illustrates

the same effect reported in experimental measurement of SNM in a 65nm technology [84], showing a comparable shape change associated with different drain bias dependent response of the individual transistors in the cell.



(a)



(b)

Figure 5.13: (a) An instance of a cell with extreme shift in SNM using full model based simulation, (b) is an example of SNM measurement of a 65nm technology cell from *Hiramoto et al.* [84].

Hiramoto et al. [84] perform an analysis of both NMOS and PMOS correlation between threshold voltage and low drain on-current. One of the important conclusions from the 65nm technology measurement analysis of the SRAM cell is that there is a decorrelation between device threshold voltage and on-current. Figure 5.14 shows equivalent plots for devices generated using NPM and Gaussian V_T, as well as the results of the Hiramoto measurements and clearly shows that NPM devices reproduce the trends obtained from physical device measurements giving confidence in the SRAM simulation results obtained.

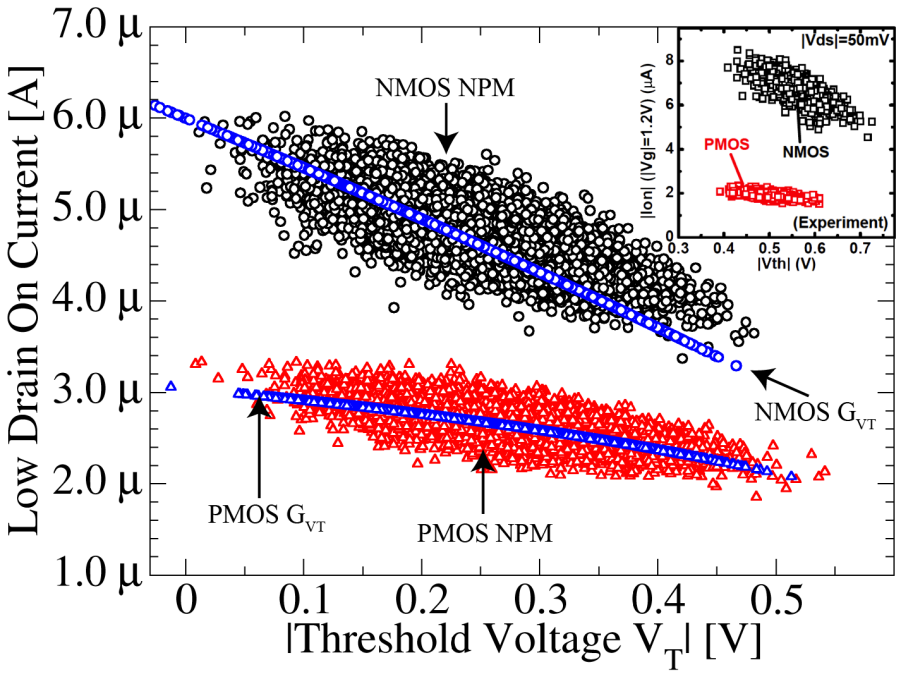


Figure 5.14: Correlation between threshold voltage and on-current for both PMOS and NMOS transistors using NPM and Gaussian V_T generation methodologies; direct device measurements from *Hiramoto et al.* [84] are inset. $V_{DS} = 50mV$ for both datasets.

5.3.2 Read Current simulation

All read current simulations were performed as outlined in Section 5.2. Simulations were performed at two supply voltage levels $V_{DD} = 1V$ and at $V_{DD} =$

$0.5V$. The lower supply voltage leads to lower read current, as transistor overdrive is reduced, which defines a more critical operating point for the SRAM system, and is especially relevant in low-power designs where a reduction in supply voltage may be required to meet power requirements.

Figure 5.15, shows QQ plots of the results of 100,000 simulations of cell read current using the Gaussian V_T and NPM generation approaches. The results show that at the nominal supply voltage of $V_{DD} = 1V$ the difference between the two sets of simulations is predominantly in the upper tail which, as shown in Table 5.1, leads to an increase in the mean and standard deviation of the distribution of read current. However, the lower tail of the distributions match well. The distributions change significantly at the lower supply voltage of $V_{DD} = 0.5V$. At this supply voltage some devices will be close to threshold and as a result of this the read current distribution begins to deviate from the expected Gaussian distribution. A similar effect has been shown in physical SRAM cell measurements [148], where non-Gaussian behaviour of the read current is observed from measurement at reduced supply voltage. QQ plots of the simulated read current distributions at $V_{DD} = 1V$ and $V_{DD} = 0.5V$ are shown in Figure 5.15, which shows that Gaussian V_T simulations overestimate the impact of variability on the distribution of read current. This is particularly evident in the $V_{DD} = 0.5V$ simulation set, as Gaussian V_T simulations overestimate the skew of the distribution. The results shows that, if the Gaussian V_T generation methodology is used to calculate minimum read current, the results could be overly optimistic.

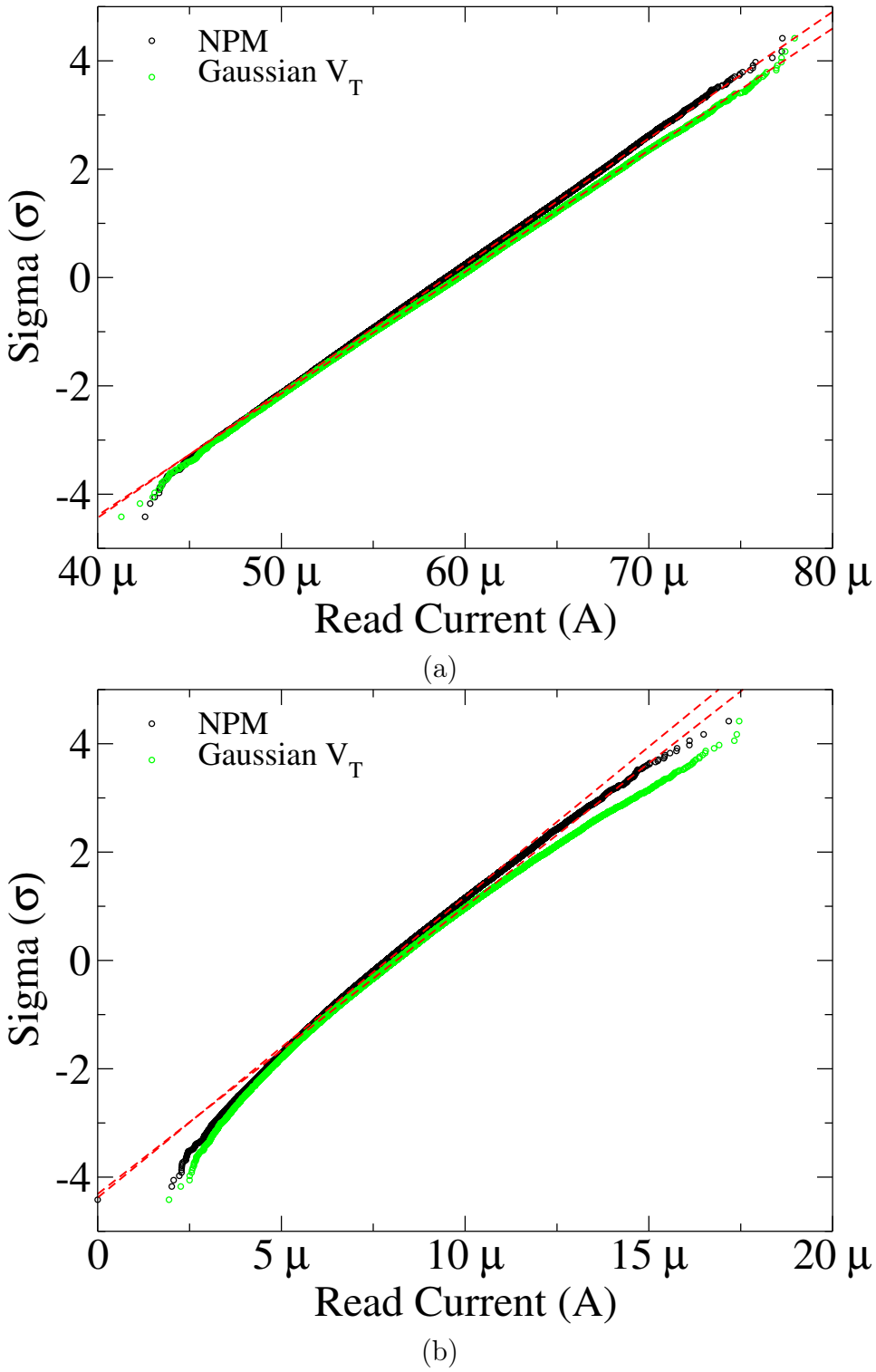


Figure 5.15: Read current distributions obtained from Gaussian V_T and NPM based simulation at (a) $V_{DD} = 1V$ and (b) $V_{DD} = 0.5V$. $V_{DD} = 0.5V$ simulations show significant skew.

Moment	$V_{DD} = 1V$			$V_{DD} = 0.5V$		
	NPM	GV_T	%Δ	NPM	GV_T	%Δ
Mean	$59.0\mu A$	$60.0\mu A$	1.7%	$7.93\mu A$	$8.20\mu A$	3.3%
Standard Deviation	$4.2\mu A$	$3.7\mu A$	4.6%	$1.78\mu A$	$1.89\mu A$	5.8%
Skewness	0.00	0.00	0.0%	0.22	0.29	24%
Kurtosis	2.95	2.96	0.0%	2.99	3.08	2.8%

Table 5.1: Moments of the simulated read current distributions obtained from Gaussian V_T and NPM based simulation.

The reason behind this overestimation of the impact of statistical variability can be explained when considering the Gaussian V_T generation strategy and the impact this strategy has on generated transistor performance. As the value of read current is defined by the on-current of the pass and pull-down NMOS transistors, we examine the distribution of this figure of merit of the statistically generated NMOS devices. This was previously investigated in Section 4.6, and showed that Gaussian V_T generation over-estimates the correlation between device performance figures of merit as all are directly determined by the threshold voltage. When we compare the distribution of low drain on-current generated using NPM and Gaussian V_T generation strategies, shown in Figure 5.16 in the form of a QQ plot, it is clear that the use of Gaussian V_T overestimates the variability in low drain on-current, with the standard deviation of this value 15% higher from Gaussian V_T compared to NPM generated devices. This accounts for the difference between NPM and Gaussian V_T SRAM read current simulations, and shows that for the purposes of read current simulation at reduced V_{DD}, Gaussian V_T is not appropriate to accurately capture the effects of statistical variability.

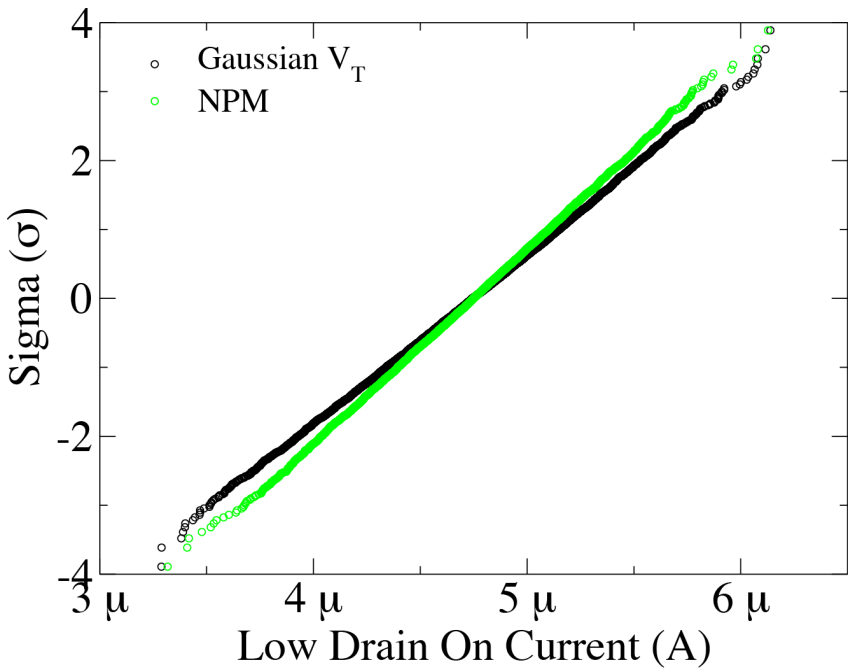


Figure 5.16: Low drain on-current distribution of devices generated with NPM and Gaussian V_T showing the artificially increased variance in the Gaussian V_T devices.

5.3.3 Dynamic Write Simulations

In order to evaluate the effect of statistical variability on SRAM in a more industrially relevant case, a joint project was undertaken in partnership with ARM Ltd, whose purpose was to evaluate the accuracy of an industry standard corner based simulation, in this case Most Probable Vector (MPV) analysis, described in [52], in comparison with full Monte-Carlo simulation using accurate compact models. The full Monte-Carlo approach is considered the benchmark, and for this purpose a very large number of simulations were performed (5 Million) in order to accurately capture the behaviour of the tails of the Dynamic Write Margin distribution.

The MPV method relies on the assumption that Gaussian V_T accurately captures all aspects of statistical variability, and involves calculating the standard deviation of the threshold voltage of each transistor in the circuit. Offsets

to the threshold voltage of each transistor are then calculated based on the equation below,

$$1\sigma M = \sqrt{\sum_{i=1}^{n} G_i^2 \sigma_i^2} \tag{5.1}$$

where $1\sigma M$ is the standard deviation of the mertic M, and G_i is the gradient of degradation of M, and σ_i is the standard deviation of the uncorrelated parameters, and i represents each parameter of the ensamble n. These offsets are applied in the direction which causes degradation in cell performance and allows the calculation of the most probable fail point of the circuit. We have already demonstrated that two of the assumptions underpinning the MPV approach are incorrect: that the threshold voltage is Gaussian distributed to high sigma [54], and the assumption that Gaussian V_T accurately represents statistical variability at a device and circuit level. The errors introduced through these assumptions will be evaluated through large scale Monte Carlo simulation using NPM model simulations as a reference.

A test memory system design was supplied by ARM for the purpose of the project, including addressing, word-line pulse generation, pre-charge, sense-amp and clock generation circuitry. The SRAM bitcell design is the same as that used in Sections 5.3.1 and 5.3.2. Unlike the previous steady-state simulations, used to classify the performance of the bitcell itself, dynamic write simulations model a realistic operating condition for the whole SRAM system, and as such can be directly related to actual SRAM system operation and yield. The simulation of dynamic write margin was chosen for this comparison due to its paramount importance in defining the word line pulse width of the SRAM system which, in effect, limits the operational speed of the whole SRAM block as the same word line pulse width is used in all modes of operation.

Due to the added complexity of the surrounding circuitry, increasing the number of transistors simulated from 6 in a single SRAM cell to close to 300 in the full system, as well as the increase in numerical complexity between performing steady state simulation and fine-grain transient simulation, simulation time is drastically increased, from fractions of seconds for the static cell simulation to minutes for a single random circuit instance. As a result of this,

in an industrial environment, it is impractical to run millions of simulations to verify designs at the high sigma values required for SRAM verification. Due to the HPC cluster facilities available for the purpose of this project, we were able to perform large scale simulations (>5 million) and to compare to the margining results obtained through MPV.

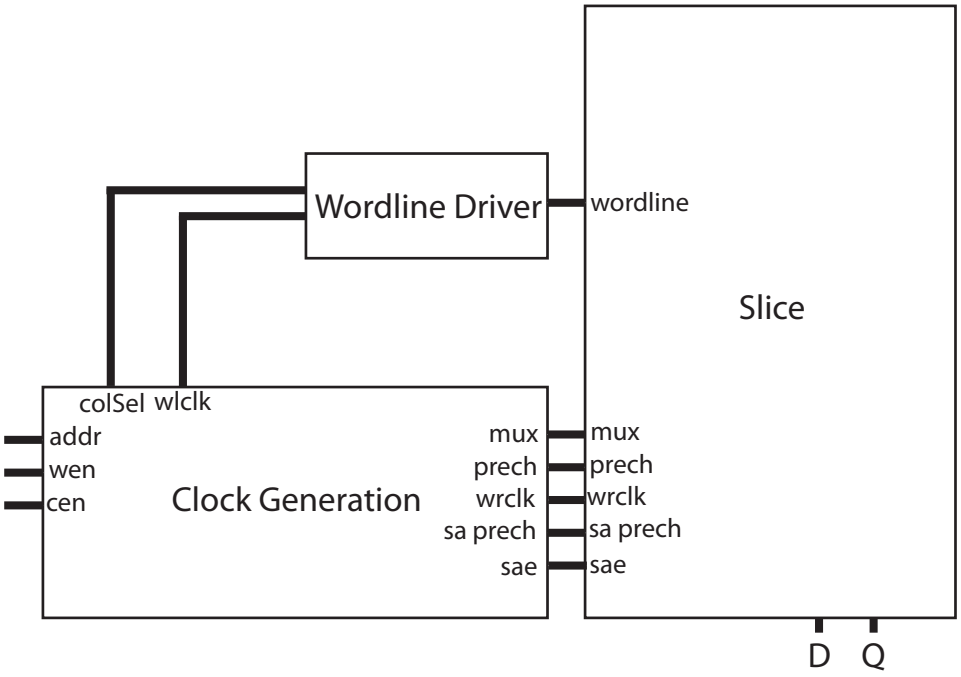


Figure 5.17: Block level dynamic write simulation circuitry.

The block level SRAM system description is shown in Figure 5.17. Instead of a full array of inactive cells, the single SRAM cell being written to, is simulated with extra capacitance added to the bitline and word line to account for the extra capacitance of the inactive cells also connected to the bitlines and word line. The dynamic write margin is defined as the time between the rising internal node, transitioning from '0' to '1', reaching 70% of V_{DD} before the word line falls to 50% of V_{DD}. The measurement is depicted in Figure 5.18. For the purpose of these simulations we only consider the effect of statistical variability on the SRAM cell, assuming that the word line pulse is constant. The assumption of a constant word line pulse is based on the premise that the surrounding digital circuitry is more significantly effected by process variability

as it is standard digital logic, with larger devices, and a global drift has a larger impact on digital circuit timing performance. This assumption is generally true in this case as we do not enable the sense amplifier, which may be highly sensitive to statistical variability, during the write operation.

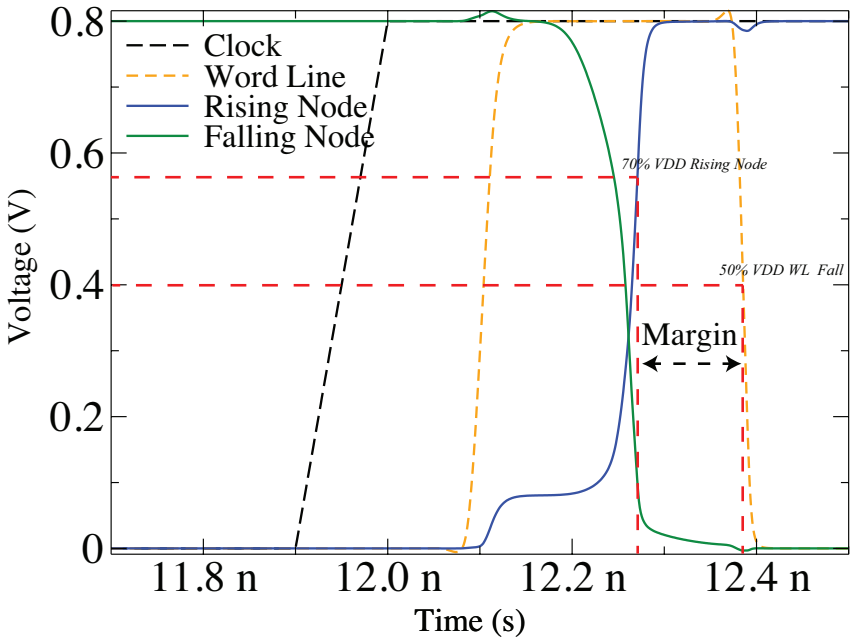


Figure 5.18: Dynamic write margin measurement.

Dynamic write margin is largely dependant on the *falling* node which transitions from '1' or high to '0' or low, this is due to the way the SRAM write operation is performed. Initially the bit lines are pre-charged to '1', then the cell node which is *rising* to '1' is actively driven. The bit-line on the internal node falling from '1' to '0' is slowly discharged through the pass gate to the bit-line. As a result, the rate at which the internal node is pulled down from '1' to '0' is dominated by the performance of the pass transistor, enabled by the word line, and the pull-up transistor, which is 'on' as the node is initially storing a '1'. The node voltage slowly discharges and starts to change the bias of the pull up and pull down transistors of the opposite inverter, creating a feedback loop, which slowly decreases the voltage on the falling node until the cell reaches its metastable point (this can be clearly seen on the rising

node voltage trace in Figure 5.18 after the plateau region before the quick rise transition) and the cell rapidly changes state. At this point the falling node and equivalent bitline are quickly discharged through the pulldown transistor and the write operation is completed. The dynamic write margin can be loosely broken into two components: the falling node discharge time, which dominates due to its relatively slow nature and is defined by the falling node pull-up to pass transistors, and the cell inverter pair metastable point, which is defined by the cross-coupled inverters forming the cell. In order simulate the circuit at its most likely failing point we perform all simulation and analysis a the $Slow_{NMOS} - Fast_{PMOS}$ process corner and at $Temp = -40°C$ as this represents the worst case write conditions.

The results of dynamic write simulation for Gaussian V_T, NPM and MPV approaches are shown in Figure 5.19 in the form of a logarithmic Empirical Cumulative Distribution Function (ECDF) plot [135]. In each result set there are 5 million NPM and Gaussian V_T simulations which characterise the results to roughly 4.5σ and allow analysis deep into the tails of the dynamic write distribution. The results show that, while MPV reproduces the results of Gaussian V_T simulations relatively well, both of these results are highly pessimistic in comparison with the accurate NPM model simulations. NPM simulations show a dynamic write margin fail rate closer to 4.4 σ, compared to 4.1 to 4.2σ predicted by Gaussian V_T and MPV. Converting this into parts per million, Gaussian V_T/ MPV simulations predict approximately 20 fails per million, while full model simulations show that the actual fail rate is closer to 4 fails per million, indicating that the MPV approach could lead to significant over design, in an attempt to meet overly pessimistic margins.

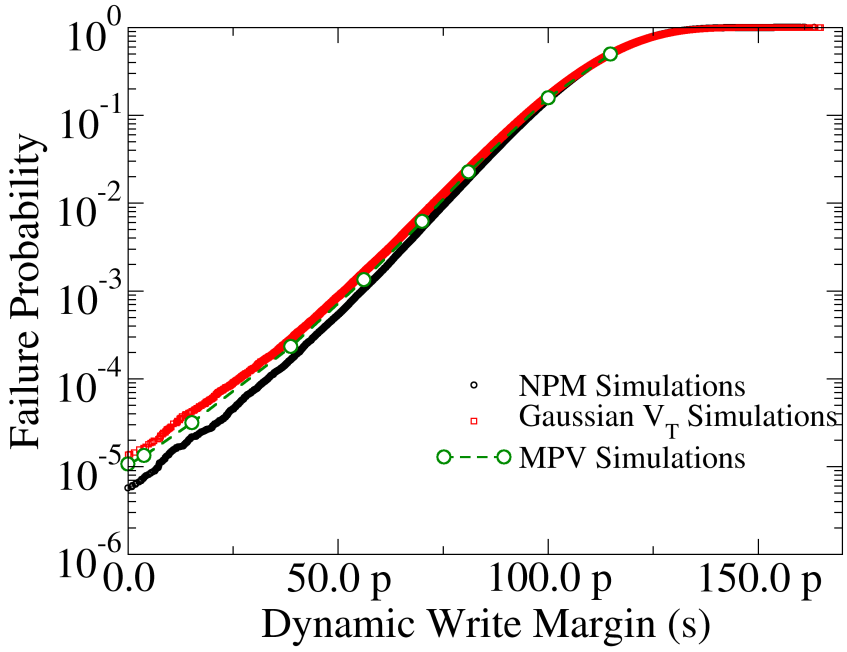


Figure 5.19: CDF plot of dynamic write margin obtained through Gaussian V_T, NPM and MPV simulation. 5 million Gaussian V_T and NPM simulations are performed, the CDF defined the probability of a cell performing up to and below a set write margin performance.

More significant differences are obtained when investigating the sensitivity of MPV to noise in the input variability information. During the initial simulation stage we had the advantage of knowing the exact standard deviations of the threshold voltages of each of the devices in the cell, as they were calculated directly from 3D statistical device simulation. To evaluate the sensitivity of MPV to noise in the input standard deviations of threshold voltage, we performed margining simulations with $\sigma V_T \pm 5\%$. The dashed lines in Figure 5.20 represent the results of these simulations. As can be seen in the tail of the distribution, the uncertainty of the fail rate increases exponentially. The small amount of uncertainty in σV_T causes the fail criterion to vary between 3.95 σ and 4.6 σ. This relates to a fail rate range of 2 to 40 parts per million or an uncertainty range greater than one order of magnitude. It is expected that this uncertainty will increase in a more realistic design where the fail criterion is closer to 5-5.5σ, due to the fact that the uncertainty increases exponentially

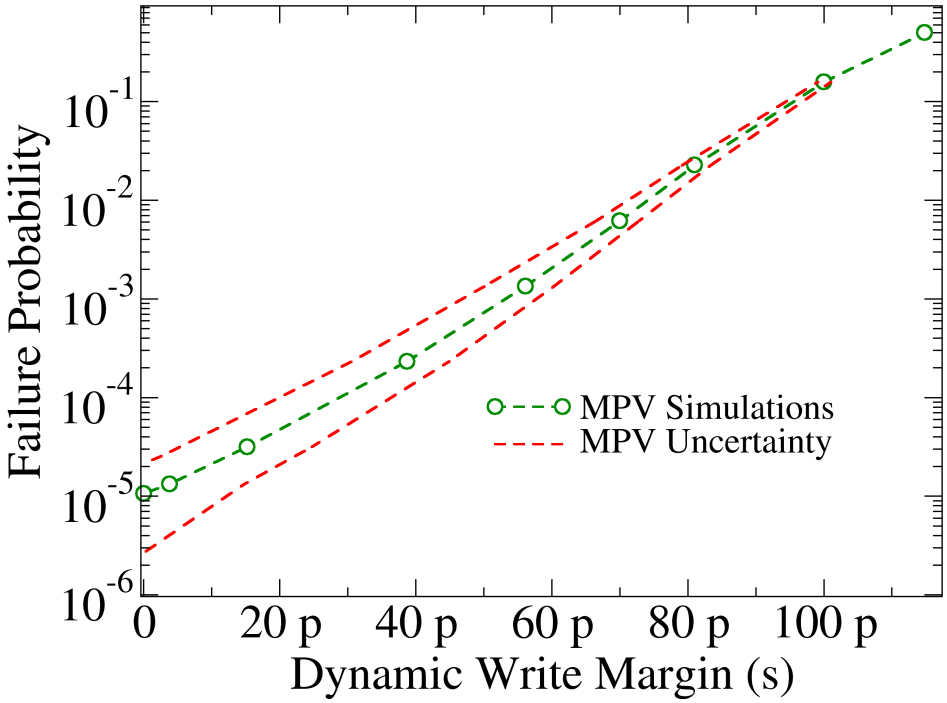


Figure 5.20: CDF plot of dynamic write margin obtained through MPV simulation, the red dashed lines represent MPV simulations with $\sigma V_T \pm 5\%$.

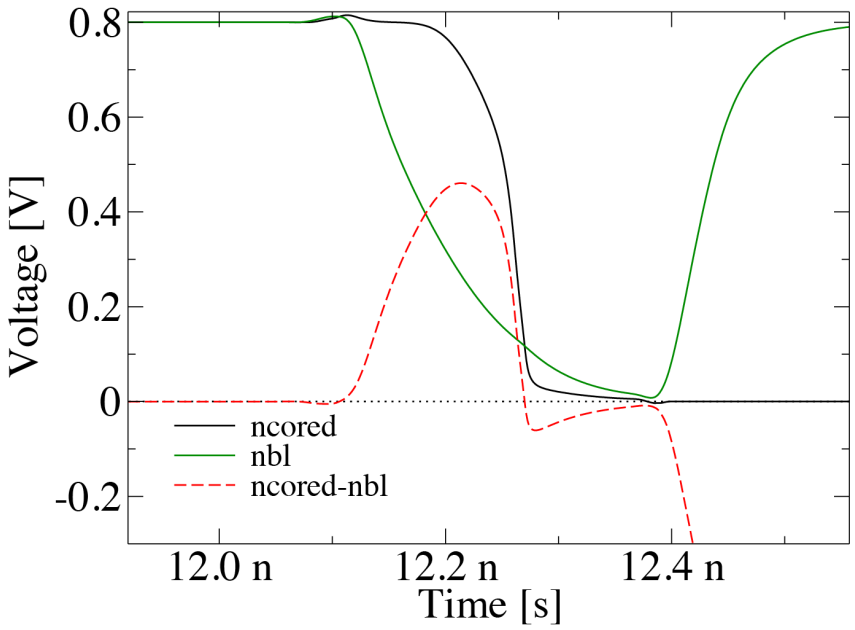
as a function of the distance from the mean.

Transistor	Relative Contribution (MPV)
Rising Pass Gate	0.20
Falling Pass Gate	0.90
Rising Pull Up	0.14
Falling Pull Up	0.31
Rising Pull Down	0.17
Falling Pull Down	0.02

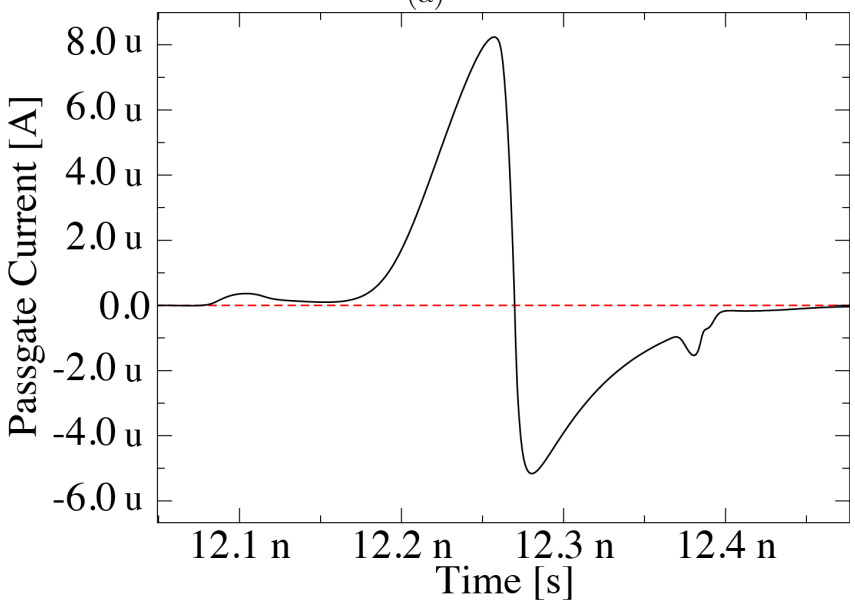
Table 5.2: Relative transistor contribution to variability in Dynamic Write simulation from MPV analysis.

In order to understand the difference between Gaussian V_T and NPM based simulations we analyse the most dominant transistor in the circuit as indicated by the MPV analysis shown in Table 5.2, the pass gate transistor on the falling SRAM cell node. The worst Dynamic Write margin is achieved when this

device is 'weak', meaning that it has a high threshold voltage and low on-current. This reduces write margin as the voltage on the falling node leaks through the pass transistor onto the bitline. Considering the bias conditions of the pass transistor during the write operation, the gate bias is '1' due to the word line pulse and the source and drain are '1' due to the pre-charged bit line and internal node state of the cell. The bit line voltage then begins to fall as the capacitance discharges. This builds up a potential difference across the source and drain of the pass transistor and current flows. The node voltage of the source and drain of the pass transistor in a simulation, without variability, is shown in Figure 5.21 (a). By analysing the bias conditions and the current flowing thorough the pass transistor, shown in Figure 5.21 (b), we reach the conclusion that although the peak effective drain bias of the critical pass gate is $0.4V$, the limiting factor on its operation is the low drain bias on-current.



(a)



(b)

Figure 5.21: (a) Source and drain voltages of the pass gate. The bitline voltage is represented by *nbl*, the node voltage is represented by *ncored* and the difference is represented by *ncored-nbl* (b) the current flowing through the pass gate, initially current flows onto the bit line, however as the cell reaches the metastable point the internal node quickly falls to '0' and current flow is reversed and flows from the bit line to ground through the internal cell pull down transistor.

Figure 5.22 shows the distribution of 10,000 n-channel passgate devices generated using the GSS 3D device simulator GARAND, plotting MOSFET threshold voltage against low drain on-current. In order to confirm that NPM captures and reproduces the GARAND simulated correlation between these two parameters 10,000 devices have been generated using NPM and added them to the plot. Finally, 10,000 devices have been generated using the Gaussian $V_T$method and included. The graph leads to two conclusions: (i) the Gaussian V_T approach overestimates the correlation between threshold voltage and on-current. The use of Gaussian V_T yields a correlation coefficient of 1, while simulations and NPM show a correlation coefficient of $\sim$ 0.7. This is in agreement with the results already seen in Section 4.6. (ii) There is a systematic offset between the Gaussian V_T devices and the GARAND/NPM generated devices. This becomes worst in the higher threshold voltage range, where Gaussian V_T simulation severely underestimates the on-current of the transistors. This leads to the result obtained in the Dynamic Write Margin simulations, where Gaussian V_T based circuit simulations under-estimate circuit yield as they predict passgates with high threshold voltages having much lower on-current than the simulated devices.

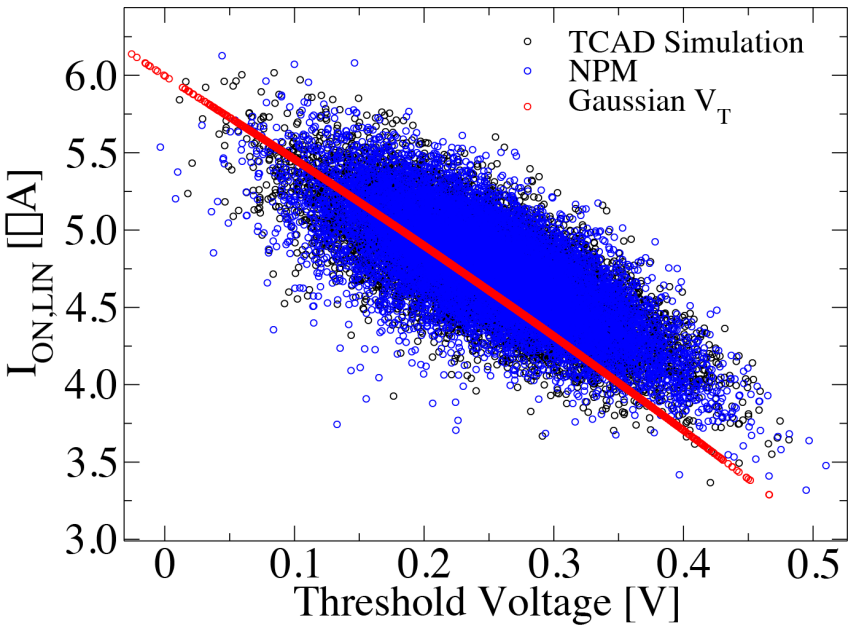


Figure 5.22: Generated device threshold voltage plotted against corresponding low drain on-current. Comparison between 3D device simulation using GARAND, NPM generated devices and Gaussian V_T generated devices.

The results clearly show that Gaussian V_T simulations cannot capture the complex impact of variability on SRAM Dynamic Write performance. The impact of Gaussian V_T simulation on simulation accuracy is also not easy to predict; SNM simulations show an underestimation of the effects of variability, while read current and dynamic write simulations show an overestimation of variability effects. This clearly demonstrates that there is no simple model that can be applied to correct the performance metric distributions obtained through Gaussian V_T simulation as the errors are application specific. The need for accurate compact models which capture and reproduce all effects of variability has been established.

5.4 Summary

As the SRAM cell and system is one of the most sensitive components in modern SoC applications, with respect to statistical variability, it has been

the initial focus of our simulation study in the impact of compact model accuracy on circuit simulation. The most common variant in modern industrial design, the 6T SRAM cell was introduced and two static figures of merit of cell performance, SNM and read current, were defined. Statistical simulations were then performed of these figures of merit using traditional Gaussian V_T and NPM methods for introducing statistical variability into the simulations. Simulations of SNM and read current showed that the use of simple Gaussian V_T models leads inaccuracy in SRAM circuit simulation results compared to full statistical compact models simulations using NPM. When considering a dynamic SRAM margin, in this case dynamic write margin, margining techniques based on Gaussian V_T distributions are no longer sufficiently accurate to model system performance and yield. Both of these conclusions highlight the importance of accurate statistical compact modelling for the purpose of SRAM performance, power and yield prediction.

Now that we have highlighted the impact of statistical variability on SRAM circuitry, in Chapter 6 we will consider the impact of statistical variability on digital logic circuit design and verification.

Chapter 6

Digital Circuit Simulation

Having established the importance of accurate statistical variability information for the purpose of SRAM simulation, we now consider the impact of statistical variability in an example of digital logic circuitry. Digital circuitry is inherently more resistant to statistical variability, as larger transistors are employed in the standard cells which form the basic building blocks of digital logic circuitry. The main challenges of digital logic performance are timing[149], static and dynamic power [150]. Individual cell variability in delay or power due to statistical variability is mutually independent. Thus in extremely long chains of logic cells, or *paths*, random fluctuations are additive and due to the Law of Large Numbers we expect these to converge to an average performance with a small amount of variance. As the longest paths within a system define the maximum delay, and thus limit the operating speed of the system, they are usually used to benchmark system performance.

It is difficult to perform SPICE level simulation of a full digital logic system, as this will characteristically involve millions of transistors. A circuit size out with the capability of most SPICE simulators. It is common for the slowest paths within the system to be found using an STA analysis tool, and only these paths are extracted for the purpose of statistical SPICE simulation. These *critical paths* [5] can then be evaluated in the presence of statistical variability as they are the most likely failure points of the system. However, it is important to note that STA fails to capture the impact of statistical variability at the full

system analysis stage, which can have an impact on the location and number of the critical paths discovered within the system [95]. In this section we compare the relative impact of process and statistical variability on a digital logic circuit, which is chosen to be representative of a critical path within a larger system, and discuss the necessity for accurate statistical models within digital circuit applications.

6.1 Adder Test Circuit

In order to investigate the effects of MOSFET variability on propagation delay, power and yield of digital logic circuits, a simple test circuit was designed by the nanoCMOS [151] project partners at the University of Manchester. Due to its multiple input and output layout, the one bit full adder (shown in Figure 6.1) is ideal for this purpose. The choice of such a small circuit with a minimal number of gates (13, made from only 4 standard cells – a NOT, NAND, OR and buffer cell – and a relatively small number of 52 transistors) allows for detailed analysis of which transistors are critical to circuit operation and which make a major contribution to its variability sensitivity. This also builds insight into the more general principals which govern the sensitivity of digital circuits to statistical variability. The adder was originally designed using a $130nm$ technology design flow. The SPICE netlist of the adder was extracted from the design flow, including information on interconnect parasitics in the form or extracted Resistive and Capacitive (RC) elements.

In order to quantitatively evaluate its impact on digital circuit performance, statistical variability is artificially injected into the simulation flow through the generation of transistor threshold voltages, assigned in random fashion, form a Gaussian distribution with appropriate values for the mean and standard deviation. The Gaussian distribution used in this analysis has been chosen to allow simplified and transparent semi-quantitative analysis of the general trends resulting from the ever-increasing variability in future technology generations. The amount of variability injected is determined by the standard deviation of the Gaussian distribution. For the purpose of this work σV_T is

Percentage Variation	nMOS $\sigma(mV)$	pMOS $\sigma(mV)$
10%	34.8	45.5
15%	47.1	59.6
20%	62.8	79.5
25%	78.5	99.3
30%	94.2	119.2
40%	125.6	158.9
50%	165.9	198.6

Table 6.1: Percentage variation and absolute variation in V_T

defined as a percentage of the threshold voltage of the n and p-type MOSFETs in the target technology. The value for V_{Tn} is approximately $330mV$ and for V_{Tp} is approximately $400mV$ respectively. The "amount" of statistical variability is defined as a percentage of the uniform threshold voltage. The simulated standard deviations apportioned to a square transistor ($L = W$) are summarised in Table 6.1. For comparison with modern technologies, the typical standard deviation of the threshold voltage for a square bulk transistor ($L = W$) in low power variant of the $45nm$ technology is $60mV$ [152] and is projected to increase to approximately $80mV$ in the $18nm$ technology [55].

By increasing the relative threshold voltage sigma it is possible to obtain a qualitative estimate of the impact of statistical variability on the Performance, Power and Yield (PPY) of a test circuit at steadily increasing levels of statistical variability. The PPY representation is a useful tool in the designer's arsenal when attempting to estimate the impact of scaling a design to a new technology generation, and will be especially useful in investigating the effect of statistical variability on a particular design. In particular, a design aim of close to 100% yield is likely to result in unacceptably low average circuit performance, and trading off between performance, power and yield may become the norm in the presence of significant statistical variability. RandomSpice generates a randomised model instance for each MOSFET found in the system, with an individual value of V_T. It then facilitates the simulation of these randomised circuits using a chosen back end simulator, in this case ngSPICE. RandomSpice also offers the opportunity to run any chosen number of

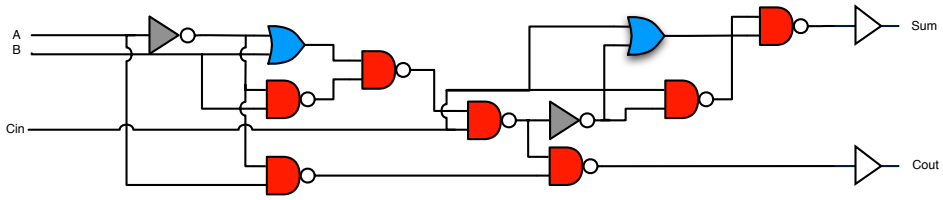


Figure 6.1: One bit ripple carry adder cell level design

randomised simulations in a cluster environment, allowing a sufficiently large ensemble of circuits to be simulated in order to achieve the required accuracy of the statistical analysis of digital circuits interrelating power, propagation delay and yield. By utilising large scale computing clusters, 1000s of simulations can be performed in parallel, allowing, in a short space of time, multiple levels of variability to be simulated at several supply voltages. In order to achieve accurate estimations of the impact of variability on the power consumption of the circuit, all 56 possible state transitions of the 3 input adder system, including state transitions which do not cause state changes in the output, were simulated to provide a realistic simulation test vector under all input/output conditions.

6.2 Single Variability Sources

Initially, four sets of simulations were carried out to demonstrate the effect of different sources of variability on delay, power and yield in the example digital circuit. When modelling process variability, the nominal value of V_T is extracted from the n-channel and p-channel values found in the compact models of the commercial design kit used to design the circuit under test. The standard deviations used in the RandomSpice technology libraries are again calculated as a percentage of the mean threshold voltage. In this analysis we do not consider across-chip process variation. In order to generate chip-to-chip process variability RandomSpice determines a new random mean value of V_T for all devices in the circuit. The V_T values selected for n- and p-channel devices as a result of chip-to-chip process variability can be uncorrelated or

correlated. For example, oxide thickness variations result in correlated n- and p-channel V_T variation, whereas fluctuations in ion implantation dose result in uncorrelated V_T variation. For combined process and statistical simulation, n and p mean V_T values are determined based on the process variability, then independent statistical variability is injected into each transistor in the circuit using a Gaussian distribution centred around the mean.

The results of the simulations with uncorrelated process variability, correlated process variability, statistical variability and correlated process combined with statistical variability are shown in Figure 6.2. Also visible in these figures are the 'process corners' from the technology used to design the Adder circuit, as well as KDE based 95% and 99% yield estimate contours. Figure 6.2 indicates that, when considering statistical variability only, corners are meaningless, introducing huge pessimism in both power and timing margins. This is due to the fact that corner simulations assume that all transistors are at 3σ. The probability of a device to be at $P^1_{3\sigma} = 0.0027$ and hence the probability of all 52 devices occurring at 3σ, where each value is independent, is $P^{52}_{3\sigma} = (P^1_{3\sigma})^{52} = 2.7 \times 10^{-134}$. When considering uncorrelated process variability only, corners are still very pessimistic – increasing delay requirements by approximately 30% and power requirements by approximately 10% (compared to the 99% yield contour). In the case of correlated process variability, comparison with yield contours indicate there is still significant pessimism introduced in both timing and power by corners with an increase in delay requirements by approximately 10% and power requirements by approximately 8%. In the case of statistical and correlated process variability the data illustrates that corners adequately model the margins required in both timing and power. However if a close to 100% yield is required, the process corners do not capture the full impact of both process and statistical variability. Another observation is the large penalty incurred by designing for 99% yield, compared to 95% yield. The graph shows that for 95% yield, design specifications could be 0.57 ns / 52 fJ. However, for 99% yield, these would need to increase to 0.63 ns / 55 fJ. This indicates there is a need for large-scale statistical analysis and power/performance/yield (PPY) calculation to optimise design parameters.

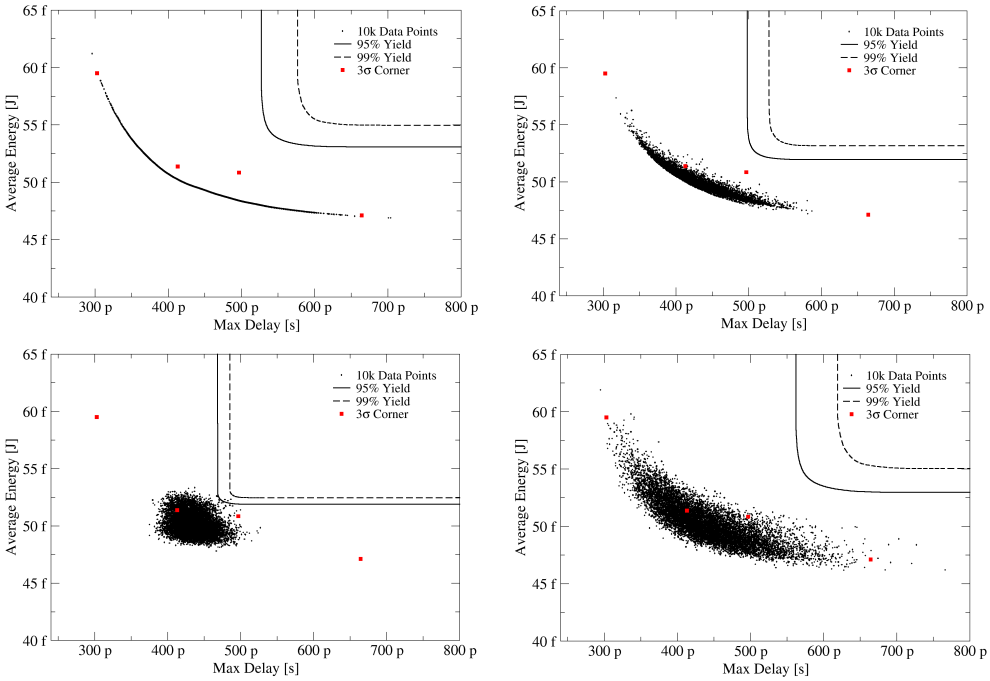


Figure 6.2: Scatter plots of delay and power for the various sources of variability. Correlated process only (top left), uncorrelated process only (top right), statistical only (bottom left) and correlated process and statistical (bottom right)

6.3 Case Study: ABB and AVS Analysis

There are techniques available to ameliorate process variability, including adaptive voltage scaling (AVS) [36], adaptive body biasing (ABB) [35] and gate length biasing [37]. In addition to the relatively small process variability, typically 1% per centimetre, present devices also exhibit significant layout dependent systematic and purely statistical on-chip variability.

In this case study we evaluate the relative effectiveness of AVS and ABB in the reduction of the negative impact of process variability, and to evaluate the impact these techniques have in the presence of statistical variability, on circuit/system performance and yield. Body biasing and voltage scaling are introduced into a statistical SPICE simulation methodology using the RandomSpice statistical circuit simulation tool.

Both process and statistical variability are introduced in the same way as in Section 6.2. Each simulation run consists of 10,000 random circuit instances including the desired types of variability; correlated or uncorrelated process variability with or without statistical variability. The circuit ensembles generated are then simulated with different amounts of body biasing (in this case up to $\pm 0.5V$ at $0.1V$ intervals) and different supply voltages (in this case $\pm 10\%$ of $VDD = 1.2V$ at intervals of $0.04V$). Using information from an initial ensemble of 10,000 simulations, an un-optimised circuit yield can be calculated and compared to the optimised yields obtained from simulations of ABB/AVS, in order to determine the relative gain obtained by applying these techniques. Yield calculations including ABB or AVS are based on the assumption that there is system level feedback, which internally adjusts ABB or AVS if the initial system is out with required performance bounds.

6.3.1 Results

In Figure 6.3 the effects of ABB and AVS on delay, power and yield are illustrated. The left column of plots show the initial circuit simulations without body biasing; the right plots then illustrate the performance trajectory of the same 10,000 circuits with an ABB of up to $\pm 0.5V$ (in steps of $0.1V$) and AVS

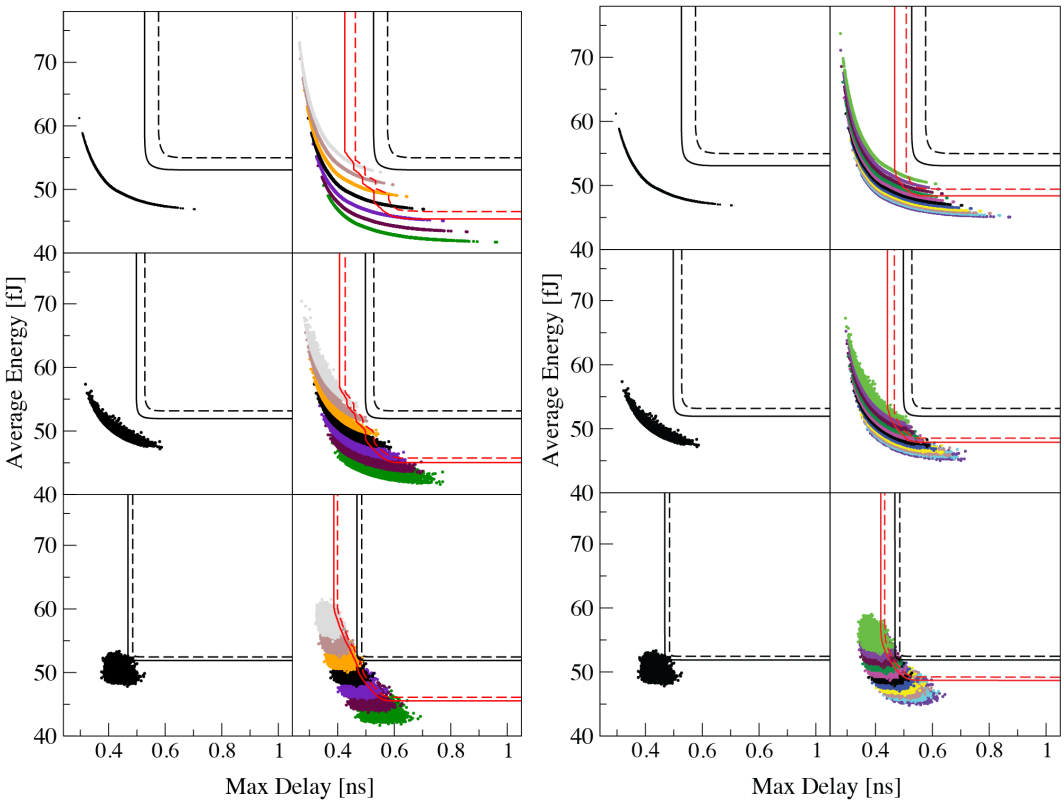


Figure 6.3: Power performance plots for nominal supply voltage and no applied body bias (left column) and with ABB (left set) and AVS (right set) applied (right column), for correlated process (top), uncorrelated process (middle) and purely statistical variability (bottom).

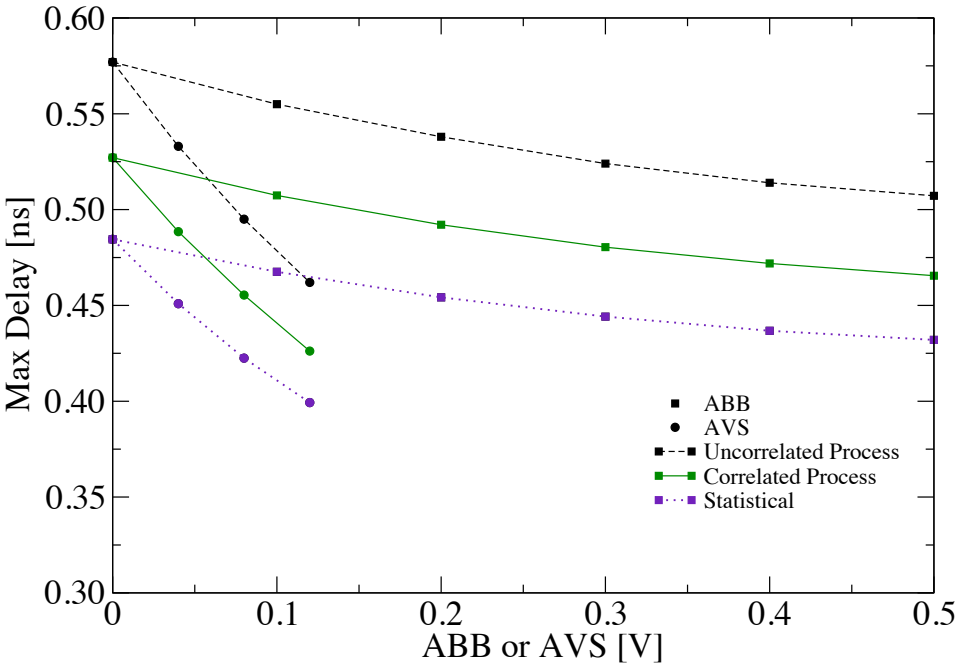
of VDD up to $\pm 0.12V$ (at steps of $0.04V$). The first conclusion to be drawn from these plots is that circuit performance is bounded by an 'optimal' Pareto front. This dictates the effect of ABB or AVS upon an individual circuit instance. If the performance of a circuit is initially poor, increasing body bias or supply voltage will be effective in reducing circuit delay, at the cost of increased energy. If, however, a circuit instance is already performing well, body biasing or an increase in supply voltage significantly increases the power consumption of the circuit with little corresponding increase in speed.

By choosing a fixed requirement for energy or delay, a comparison can be made of the relative amounts of body biasing or supply scaling required to achieve a given yield. In this example an upper limit on energy consumption

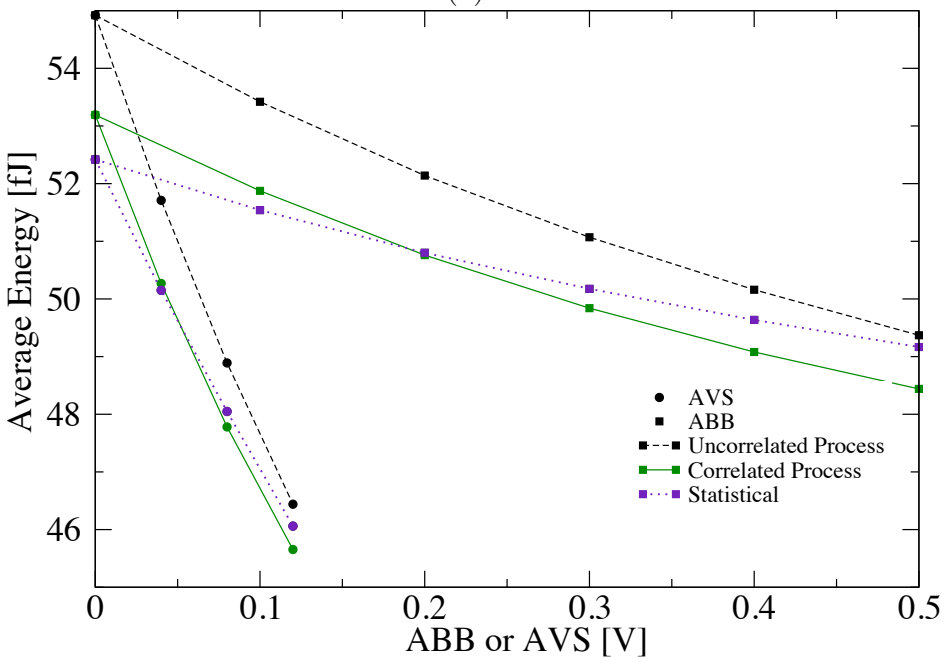
of $49fJ$ per transition, with no corrective measures applied, gives a yield of 26%. By applying up to $\pm 0.5V$ Adaptive Body Biasing this can be improved to 88%. The same effect can also be achieved by reducing the supply voltage of all instances of the circuit by $0.04V$.

In order to fully investigate the relative effect of ABB and AVS on the different types of variability, maximum delay and average power are extracted as a function of supply voltage or body bias at a constant yield of 99%, illustrated in Figure 6.4. This can be used to evaluate 'slack gain' at a desired yield. The initial observation is that a relatively small change in supply voltage causes the same effect as a large change in body biasing, leading to the conclusion that AVS is more effective at adjusting circuit performance. This of course does not take into account the relative difficulty/silicon area required to implement these approaches. An advantage of ABB is that it allows for much finer control in system performance. The delays shown in Figure 6.4 demonstrate that all 3 basic types of variability react in a similar manner to AVS and ABB. It should be noted, however, that correlated process variability shows the largest absolute improvement in delay while statistical variability shows the smallest. Also evident in the plot is the diminishing return with increased ABB and AVS, leading to the conclusion that an optimal trade off between the level of ABB or AVS applied, and the performance improvement obtained, can be achieved. The distribution of energies in Figure 6.4 reveal a more interesting result, showing that in the case of both ABB and AVS statistical variability is more resistant to improvement than the other sources of variability.

In the next stage of this analysis we examine the effect of ABB or AVS in the more realistic situation where a mixture of both statistical and process variability are present simultaneously. The results, presented in Figure 6.5, are of the same format as the previous results with unmodified simulations with 95% and 99% contours marked in the left column of the figures, and the same circuit instances simulated with applied ABB or AVS in the right column. Simulations were performed with correlated process variability as a reference; correlated process variability with 15% statistical variability ($\sigma = 15\%$ of V_T); and correlated process variability with 30% statistical variability ($\sigma = 30\%$ of V_T).



(a)



(b)

Figure 6.4: Minimum Delay (a) and Minimum Energy (b) at different levels of ABB/AVS at constant yield of 99%

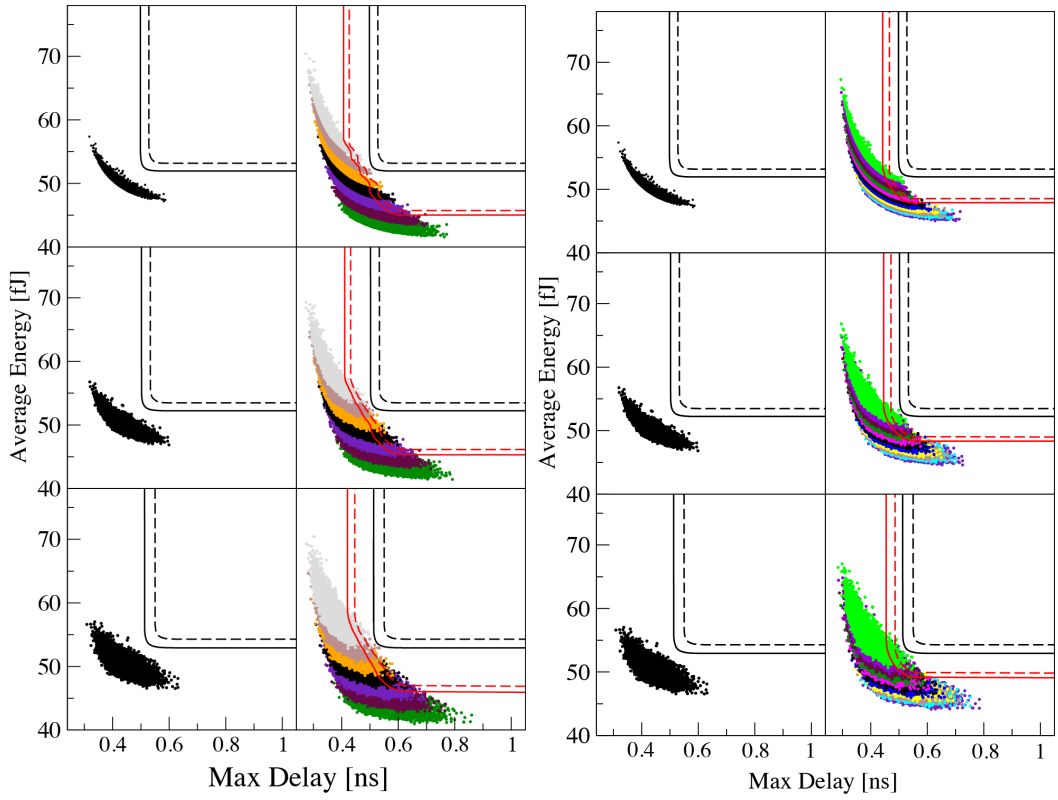


Figure 6.5: Power performance plots for nominal supply voltage and no applied body bias (left column) and with ABB (left set) and AVS (right set) applied (right column), for correlated process (top), uncorrelated process (middle) and purely statistical variability (bottom)

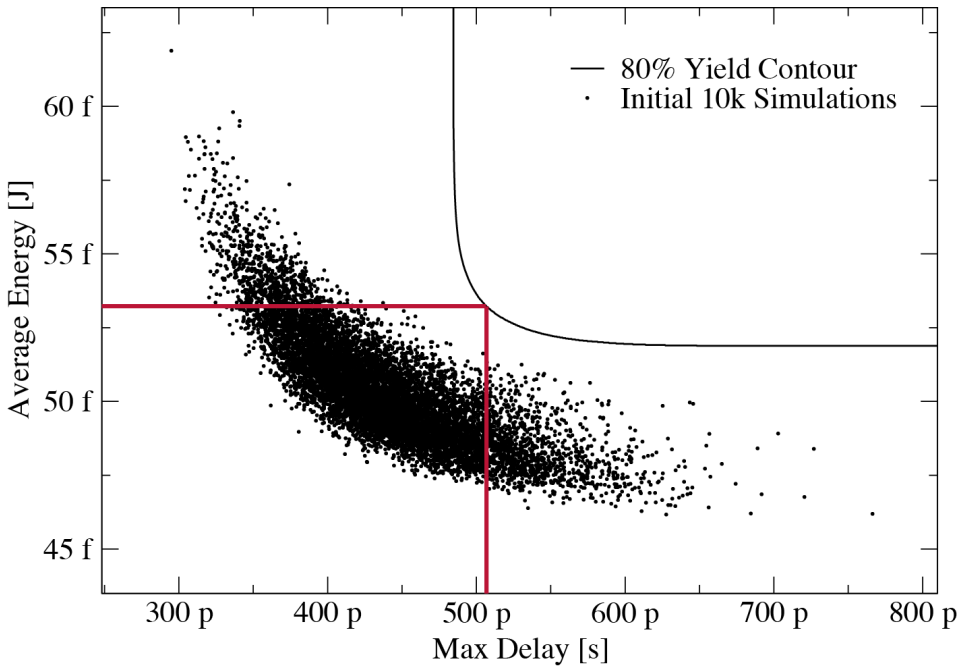


Figure 6.6: 10,000 Statistical simulations with correlated process and 30% statistical variability with 80% yield contour

As an illustrative 'case study' a set of design specifications are chosen to evaluate the relative effect of ABB and AVS on the 3 mixed sources of variability. As a baseline for comparison we choose the delay and power design cut-offs which provide an 80% yield from simulations containing correlated process with 30% statistical variability. These design cutoff points can be seen in Figure 6.6. Figure 6.7 depicts the maximum possible yield, given these delay and power cut-off points, with various amounts of applied ABB or AVS. As this figure clearly shows, the introduction of statistical variability has a significant impact on the maximum achievable yield. A small amount of statistical variability reduces yield by a few percent, but in the case where a large amount of statistical variability is introduced, the yield is greatly reduced, and ABB or AVS never fully recoups this loss. An additional conclusion that can be drawn from this information is the diminishing returns from increased amounts of ABB or AVS. The data reaches a point where increased body bias or supply voltage no longer increases yield, leading to the conclusion that, using these

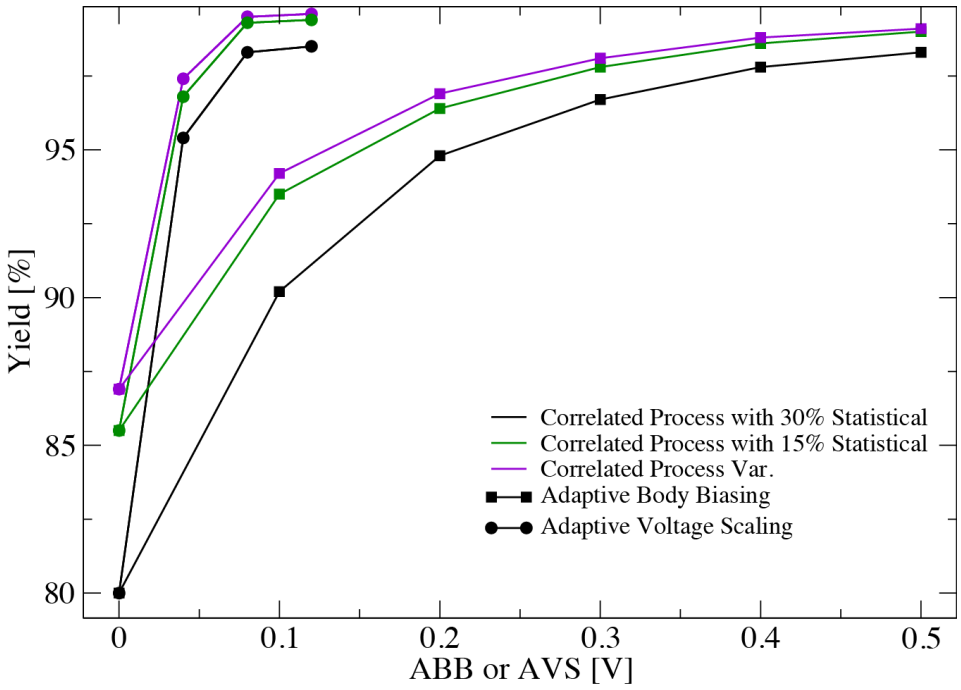


Figure 6.7: The effect of ABB and AVS on optimal parametric yield with different levels of statistical variability and correlated process variability

design specifications, no amount of body bias or voltage scaling will give ~100% yield in the presence of variability. This type of analysis allows a quantitative comparison between the merits of employing either adaptive body biasing or adaptive voltage scaling as a method of variability mitigation.

6.4 Summary

In this chapter a statistical SPICE simulation methodology, has been is used to analyse the impact of different types of variability on the performance, power and yield of a test circuit. The pessimism introduced in the estimation of yield using corner analysis has been demonstrated. ABB and AVS are introduced as a method for improving performance / power / yield. A methodology for extracting yield is introduced based on the assumption that each circuit can be optimised independently. We have demonstrated that while performance can be improved in the presence of variability, statistical variability is more resistant to the improvement of power consumption with both ABB and AVS. Finally a more realistic case study is presented where process and statistical variability are introduced and design cut-off points are chosen. This case clearly shows the diminishing returns possible with increased ABB or AVS and provides a method for the comparative evaluation of both optimisation methods so that an informed design choice could be made.

Chapter 7

Conclusions and Future Work

7.1 Summary

The aim of this work was to to investigate and develop techniques to propagate device level variability information, caused by the discreteness of charge and granularity of matter, to statistical circuit simulation to enable power, performance and yield predictions which can inform and aid the design and design evaluation process. The main sources of statistical variability in current CMOS transistors, including RDD, LER and MGG, have a significant impact on circuit and system performance, power and overall yield. In order to capture these effects and propagate them up the design flow from transistor level to circuit and system level, a methodology has to be developed to incorporate them in compact models and circuit simulation tools. Although methods have previously been proposed to model statistical variability in circuit simulations, they have generally been limited to Gaussian V_T distributions or uncorrelated variability injectors.

In order to accurately represent statistical variability in real devices the GSS simulation toolchain was utilised to simulate 10,000 *20/22nm* n- and p-channel transistors, including RDD, LER and MGG variability sources. A novel statistical compact modelling methodology was developed which accurately captured the behaviour of the simulated transistors, and produced compact model parameter distributions suitable for advanced compact model gen-

eration strategies like PCA and NPM. The resultant compact model libraries were then utilised to evaluate the impact of statistical variability on SRAM design, and to quantitatively evaluate the difference between accurate compact model generation using NPM with the current Gaussian V_T methodology. Over 5 million dynamic write simulations were performed. Results showed that at advanced technology nodes, statistical variability cannot be accurately represented using the Gaussian V_T methodology.

In Chapter 2 the sources of MOSFET variability and their impact on device and circuit performance were reviewed. The link between physical device performance and circuit simulation - the MOSFET *compact model* - was introduced. Common compact modelling techniques, including different methodologies for including variability, were discussed in detail. After this standard circuit simulation techniques, including the tradeoff between predictive accuracy and computational time, were outlined. The methodologies for introducing MOSFET variability in circuit simulations were summarised and critically assessed. The chapter also presents the motivation behind the accurate compact modelling strategy developed as part of this work, specifically for the purpose of SRAM design and verification, where performance metrics need to be accurately evaluated deep into the tails of their statistical distributions.

In Chapter 3 the GSS 3D atomistic simulator GARAND, the compact model fitting tool, Mystic, and the development of the statical compact modelling strategy were thoroughly described, focusing on establishing the physical links between parameter and device performance. Advanced statistical compact model generation strategies, used to generate an effectively infinite number of compact models which reproduce the statistical behaviour of the extracted model set, were described. Finally, the circuit simulation tool that can use the generated compact models, RandomSpice, was described. The chapter outlined a methodology which enable quick device design and simulation to circuit simulation for prototyping and evaluating future technologies, while also being applicable with physical transistor measurement.

In Chapter 4 the 20/22nm CMOS technology generation template transistor was introduced, and the uniform compact model, provided by GSS, was described. The statistical compact model extraction strategy was developed,

focusing on the accurate representation of key transistor figures of merit. Statistical compact model extraction results demonstrated the accuracy of the extraction strategy, showing an extremely low mean fitting error of 2.2% across 10,000 devices. The worst-fit problematic devices were analysed and highlighted difficulties inherent in statistical compact modelling at an advanced technology node where a combination of MGG, RDD and LER can cause for a significant spread in all device figures of merit, which push the boundaries of applicability of the uniform compact model. The extracted compact model parameter distributions were used for statistical compact model generation using multiple generation strategies. In order to benchmark the accuracy of the generation strategies, generated device parameters were compared to the extracted parameter data. The results clearly illustrated the deficiencies of the traditional Gaussian V_T approach, with generated device figures of merit showing 1-to-1 correlations, where the simulated devices showed significant decorrelation. It was demonstrated that the assumptions and simplifications inherent to PCA lead to non-physical devices, out of the tested generation strategies, highlighting the accuracy of the NPM compact model generation approach.

The most accurate compact modelling strategy presented, NPM, was used for the purpose of statistical variability aware SRAM cell and system simulation in Chapter 5. The purpose of the chapter was to accurately evaluate the errors introduced into SRAM simulation though the use of traditional Gaussian V_T models, using NPM simulations as a reference. Initial simulations target the standard d.c. figures of merit of SRAM performance. SNM simulations showed that Gaussian V_T simulation under-estimates the impact of transistor variability on cell stability, while read current simulations show Gaussian V_T over-estimating the impact of transistor variability on cell readability. In order to evaluate the effect of statistical variability on SRAM in a more industrially relevant case, the second half of the chapter focused on the results of a joint project created in partnership with ARM Ltd. In this project ensemble sizes of 5 million instances of a full SRAM system were simulated and NPM based simulations are used to evaluate the accuracy of a standard industry margining method, MPV. The results showed that, while MPV managed to reproduce

the results of the Gaussian V_T simulations, predicting approximately 20 fails per million, NPM simulations show that the actual fail rate is closer to 4 fails per million. The results clearly indicate that the Gaussian V_T is not accurate enough to predict SRAM yield at an advanced technology node.

In Chapter 6 the impact of statistical variability *and* process variability on digital logic critical paths was quantitatively assessed through a representative netlist, including parasitic interconnect and layout information, used as a test circuit. The impact of different individual and combined types of variability was investigated. The results show that for traditional logic timing and power metrics process variability dominates performance, however statistical variability introduces extra variation which significantly impacts yield and performance requirements. Finally a case study was carried out to evaluate the impact of statistical variability on circuit performance and yield enhancement techniques ABB and AVS, showing that although these methods can be used to improve yield, due to its stochastic nature, statistical variability still has a significant impact on overall yield.

7.2 Conclusions

Statistical variability is a key concern in the field of electronics and has a dramatic impact on circuit and system PPY. Although alternative device structures like FinFETs have been introduced which reduce the random variation in devices these effects will continue to increase in importance with continued scaling. The work has demonstrated that, although difficult, it is possible to capture stochastic variability effects accurately in compact models, and propegate the information to a circuit and system level. The results also show that accurate modelling techniques can help reduced design margins by elimiating some of the pessimism of standard variability modelling approaches.

7.3 Future Work

There are multiple areas upon which the work in this thesis could be expanded. The first of these would be to verify the methodology against physical device measurements. Ideally a sample size of devices larger than the simulated set would be ideal to test the accuracy of the NPM generators beyond the simulated device set. An extension of this would be to compare NPM generated SRAM simulations with physically manufactured SRAM measurements. For either of these options to be possible an industrial collaboration with a facility capable of large scale (>10,000) device measurements would be required. On an SRAM level large scale cell measurement (>5 million) to verify simulation accuracy could prohibitively due to the long measurement time and increased complexity of SRAM cell measurement. Post production monitoring may be a way to gather enough statistical information to verify the SRAM simulations on this scale, although it may be difficult obtain this data from manufacturers.

It would also be desirable to reduce the number of simulations required to accurately evaluate SRAM yield at 5σ and beyond. This is due to the fact that the computational power and time required may be prohibitive for most industrial applications, especially during the design optimisation phase where multiple sets of simulations could be required. For this purpose it would be important to apply statistical enhancement techniques like *importance sampling* or *statistical blockade* to the circuit generation phase. These techniques aim to produce accurate perfromance data deep into the tails of the distribution whilst reducing required simulation time, with respect to standard sampling. While these methodologies show promise they can be very sensitive to input parameters and can produce incorrect results which are difficult to verify. This is highlighted by the fact that there is no industry standard approach to statistical enhancement and, where present, enhancement techniques are developed and applied on a per-case basis.

A natural long term extension of the compact modelling methodology would be the development of a strategy to capture temporal degradation effects related to NBTI and PBTI. This should involve additional compact model extraction steps or circuit modification to model the impact of degradation on

device performance. This would allow for the designer to evaluate system performance over all possible operating conditions and estimate yield over the whole product lifecycle.

Finally variation in dynamic effects could also be an area into which this work could be expanded. This would be relatively simple, as the compact models have flexible and well defined capacitance parameters. This may be particulary important in 3D technologies like FinFET where parasitic capacitaces become more complex and may have an increased impact on circuit and system performance.

Bibliography

[1] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8), 1965.

[2] R. Dennard, F. H. Gaensslen, H. A. Yu, V. L. Rideout, E.. Bassous, and A. R. LeBlanc. Design of ion-impanted mosfets with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, 9(5):256–268, 1974.

[3] V S Basker et al. A 0.063um finfet sram cell demonstration with conventonal lithography using a novel integration scheme with aggresively scaled fin and gate pitch. *in Digest of Technical Papers, Symposium on VLSI Techonlogy*, pages 19–20, 2010.

[4] A. Cathignol, B. Cheng, D. Chanemougame, A. R. Brown, K. Rochereau, G. Ghibaudo, and A. Asenov. Quantitive evaluation of statistical variability sources in a 45-nm technological node lp n-mosfet. *IEEE Electron Device Letters*, 29(6):609–611, June 2008.

[5] T. Thiel. Have i really met timing? - validating primetime timing reports with spice. *Proceedings of Design Automation and Test in Europe*, 2004.

[6] J. L. Hennessy and D. A. Paterson. *Computer Arhitecture: a Quantative Approach*. Morgan-Kaufman, 1990.

[7] W. A. Wulf and S. A. Mckee. Hitting the memory wall, implications of the obvious. *ACM SIGARCH Computer News*, 23(1):20–24, March 1995.

[8] R. R. Schaller. Moore's law: past, present and future. *IEEE Spectrum*, 34(6):52–59, June 1997.

[9] ITRS. Itrs 2011 executive summary. http://www.itrs.net/Links/2011ITRS/2011Chapters/2011ExecSum.pdf, 2011.

[10] G. Declerck. A look into the future of nanoelectronics. *VLSI Technology, 2005. Digest of Technical Papers*, pages 6–10, June 2005.

[11] S. E. Thompson and S. Parthasarathy. Moore's law: the future of si mircoelectronics. *Materials Today*, 9(6), 2006.

[12] K. Nagase, S. Ohkawa, M. Aoki, and H. Masuda. Variation status in 100nm cmos process and below. *Proceedings of Conference of Microelectronic Test Structures*, (17):257–261, March 2004.

[13] K. Takeuchi and A. Nashida. Random fluctiuations in scaled mos devices. *in Proc. Simulation of Semiconductor Processes and Devices*, Sept. 2009.

[14] S. K. Saha. Modelling process varibility in scaled cmos technology. *IEEE Design and Test of Computers*, 27(2):8–16, March/April 2010.

[15] S. W. Director, P. Feldmann, and K. Krishna. Statistical integrated design. *IEEE Journal of Solid-State Circuits*, 28(3):193–202, March 1993.

[16] R. Heald and P. Wang. Variability in sub-100nm sram designs. *Proceedings of ICCAD*, pages 347–352, 2004.

[17] K. Okada. Statistical modeling of device characteristics with systematic variability. *IEICE Transactions on Fundementals*, E84-A(2):529–536, February 2001.

[18] C. Progler, A. Borna, D. Blaauw, and P. Sixt. Impact of lithography variability on statistical timing behaviour. *Design and Process Integration for Microelectronic Manufacturing II*, 5379:101–110, 2004.

[19] D. Hisamoto, W. Lee, J. Kedzierski, H. Takeuchi, K. Asano, C Kuo, E. Anderson, K. Tsu-Jae, J. Bokor, and C. Hu. Finfet-a self-aligned double-gate mosfet scalable to 20 nm. *IEEE Transactions on Electron Devices*, 47(12):2320–2325, December 2000.

[20] T. Yamashita, V. Basker, T. Standaert, C. Yeh, J. Faltermeier, T. Yamamoto, C. Lin, A. Bryant, and K. Maitra. Opportunities and challenges of finfet as a device structure candidate for 14nm node cmos technology. *ECS Transactions*, 34(1):81–86, 2011.

[21] B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C. Yang, C. Tabery, C. Ho, Q. Xang, T. King, J. Bokor, C. Hu, M. Lin, and D. Kyser. Finfet scaling to 10 nm gate length. *Proceedings of Electron Devices Meeting (IEDM)*, pages 251–254, February 2002.

[22] G. K. Celler and S. Cristoloveanu. Frontiers of silicon-on-insulator. *Journal of Applied Physics*, 93(9):4955–4978, 2003.

[23] INTEL. Intel finfet. http://www.intel.com/content/www/us/en/silicon-innovations/intel-22nm-technology.html.

[24] E. Karl, Y. Wang, Y. Ng, Z. Guo, F. Hamzaoglu, U. Bhattacharya, K. Zhang, K. Mistry, and M. Bohr. A 4.6ghz 162mb sram in 22nm tri-gate cmos technology with integrated active vmin-enhancing assist circuitry. *in Proc. International Solid-State Circuit Conference*, pages 230–232, 2012.

[25] B. Giraud, O. Thomas, A. Amara, A. Vladimirescu, and M. Belleville. *Planar Double-Gate Transistor*. Springer, 2009.

[26] W. Schemmert and G. Zimmer. Threshold-voltage sensitivity of ino-implated mos transistors due to process variations. *Electronics Letters*, 10(0):151–152, 1974.

[27] T. Tanaka, T. Usuki, Y. Momiyama, and T. Sugii. Direct measurement of vth fluctiation caused by impurity poisoning. *Digest of Technical Papers, Symposium on VLSI Techonlogy*, pages 136–137, 2000.

[28] R. Tian and X. Tang. Dummy-feature placement for chemical-mechanical polishing uniformity in a shallow trench insulation process. *IEEE Transactions on Computer Aided Design*, 21(1):63–71, 2002.

[29] M. Khare. High-k/metal gate technology: A new horizon. *Proceedings of Custom Integrated Circuits Conference*, pages 417–420, 2007.

[30] L. Pang, K. Qian, C. J. Spanos, and B. Nikolic. Measurements and analysis of variability in 45nm strained-si cmos technology. *IEEE Journal of Solid-State Circuits*, 44(8):2233–2239, August 2009.

[31] B. Nikoli and L. Pang. Measurements and analysis of process varibility in 90nm cmos. *Proceedings of Intrnatonal Conference on Solid-State and IC Technology*, 2006.

[32] K. Agrawal and S. Nassif. Characterizing process variation in nanometer cmos. *in Proc. Design Automation Conference*, pages 396–399, June 2007.

[33] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W. Shih, S. Siavakumar, G. Taylor, P. VanDerVoorn, and K. Zawadzki. Managing process variation in intel's 45nm cmos technology. *Intel Technology Journal*, 12(2):93–110, June 2008.

[34] T. Chen and S. Naffziger. Comparison of abb and asv for improving delay and leakage under the presence of process variation. *IEEE Transactions of VLSI Systems*, 11(5):888–899, October 2003.

[35] J. Tschanz, J. Kao, S. Nardendra, R. Nair, D. Antonaiadis, A. Chandrakasan, and V. De. Adaptive body biasing for reducing impacts of die-to-die and within die parameter variations on microprocessor frequency and leakage. *IEEE Journal of Solid-State Circuits*, pages 1396–1402, November 2002.

[36] J Tschanz, S Nardendra, R Nair, and Vivek De. Evectiveness of adaptive supply voltage scaling and body bias for reducing the impact of para-

meter fluctuations in low power and high performance microprocessors. *IEEE Journal of Solid-State Circuits*, pages 826–829, May 2003.

[37] P. Gupta, A. B. Kahng, P. Sharma, and D. Sylvester. Selective gate-length biasing for cost-effective runtive leakage control. *in Proc. Design Automation Conference*, June 2004.

[38] S. M. Burns, M. Ketkar, N. Menezes, K. A. Bowman, J. W. Tschanz, and V. De. Comparative analysis of conventional statistical design techniques. *in Proc. Design Automation Conference*, pages 283–243, June 2007.

[39] D. Boning and S. Nasif. *Design of High-Performance Microprocessor Circuits*. Wiley-Blackwell, 2000.

[40] L. Capodiechi. From optical proximity correction to lithography driven physical design (1996-2006): 10 years of resolution enhancement techonlogy and the roadmap enablers for the next decade. *Proceedings of SPIE Volume 6154*.

[41] G. Gildenblat, editor. *Compact Modling*. Springer, 2010.

[42] T.B. Hook. Lateral ion implant straggle and mask proximity effect. *IEEE Transactions on Electron Devices*, 50:1946–1951, 2003.

[43] P. G. Drennan, M. L. Kniffin, and D. R. Locasico. Implications of proximity effects for analog design. *in Proc. IEEE Custom Integrated Circuits Conference CiCC*, pages 169–176, 2006.

[44] A. K. Wong. Microlithography: trends, challenges, solutions, and their impact on design. *IEEE Micro*, 23(2):12–21, March/April 2003.

[45] D. A. Antoniadis, I. Aberg, C. Ni Chleirigh, O. M. Nayfeh, A. Khakifirooz, and J. L. Hoyt. Continuous mosfet performance increase with device scaling: The role of strain and channel material innovations. *IBM Journal of Research and Development*, 50(4.5):363–376, July 2006.

[46] X. Wang, B. Cheng, S. Roy, and A. Asenov. Simulation of strain enhanced variability in nmosfets. *Proceedings of Ultimate Integration of Silicon*, pages 89–92, 2008.

[47] C. Chiang and J. Kawa. *Design for manufacturability and yield for nanoscale CMOS*. Springer, 2007.

[48] A. J. Strojwas. Challenges in modeling layout systematic effects in compact device models. *MOS-AK.GSA Workshop*, 2010.

[49] K J Khun. Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale cmos. *Intel Technology Journal*, 2007.

[50] A. Asenov, G. Slavcheva, A. R. Brown, J. H. Davies, and S. Saini. Increase in the random dopant induced threshold fluctuations and lowering in sub-100nm mosfets due to quantum effects: A 3-d density-gradient simulation study. *IEEE Transactions on Electron Devices*, 48(4), 2001.

[51] A. Asenov, S. Kaya, and A. R. Brown. Intrinsic parameter fluctuations in decananometer mosfets introduced by gate line edge roughness. *IEEE Transactions on Electron Devices*, 50(5):1254–1260, May 2003.

[52] A. Singhee and R. Rutenbar, editors. *Extreme Statistics in Nanoscale Memory Design*. Springer, 2010.

[53] R. H. J. M. Otten and L. P. P. P. van Ginneken. *The Annealing Algorithm*. Kluwer Academic Publishers, 1989.

[54] D. Reid, C. Millar, G. Roy, S. Roy, and A. Asenov. Analysis of threshold voltage distribution due to random dopants: A 100,000 sample 3-d simulation study. *IEEE Transactions on Electron Devices*, 56:2255–2263, 2009.

[55] X. Wang, A. R. Brown, N. Idris, S. Markov, G. Roy, and A. Asenov. Statistical threshold-voltage variability in scaled decananometer bulk hkmg

mosfets: a full-scale 3-d simulation scaling study. *IEEE Transactions on Electron Devices*, 58(8):2293–2301, August 2011.

[56] Y. Li and C. Hwang. High-frequency characteristic fluctuations of nano-mosfet circuit induced by random dopants. *IEEE Transactions on Microwave Theory and Techniques*, 56(12):2726–2733, December 2008.

[57] N. Sano, K. Matsuzawa, M. Mukani, and N. Nakayama. On discrete random dopant modeling in drift-diffusion simulations: physical meaning of 'atomistic' dopants. *Microelectronics Reliability*, 42(2):189–199, February 2002.

[58] G. Roy, A. R. Brown, F. Adamu-Lema, S. Roy, and A. Asenov. 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.

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

[60] S. Xiong and J. Bokor. A simulation study of gate line edge roughness effects on doping profiles of short-channel mosfet devices. *IEEE Transactions on Electron Devices*, 51(2):228–232, February 2004.

[61] H. Kim, J. Lee, J. Shin, S. Woo, H. Cho, and J. Moon. Experimental investigation of the effect of lwr on sub-100nm device performance. *IEEE Transactions on Electron Devices*, 51(12):1984–1988, 2004.

[62] X Wang, A R Brown, B Cheng, and A Asenov. Statistical variability and reliability in nanoscale finfets. *Proceedings of Electron Devices Meeting (IEDM)*, 2011.

[63] H. Zhao, J. Huang, Y. Chen, J. H. Yum, Y. Wang, F. Zhou, F. Xue, and J. C. Lee. Effects of gate-first and gate-last process on interface quality of in0.53ga0.47as metal-oxide-semiconductor capacitors using atomic-layer-deposited al2o3 and hfo2 oxides. *Applied Physics Letters*, 95(25), GF-GL 2009.

[64] A. R. Brown, N. Idris, J. R. Watling, and A .Asenov. Impact of metal gate granularity on threshold voltage variability: A full-scale three-dimensional statistical simulation study. *IEEE Electron Device Letters*, 31(11):1199–1201, November 2010.

[65] G. Ribes, J. Mitard, M. Denais, S. Bruyere, F. Monsieur, C. Parthasaray, E. Vincent, and G. Ghibaudo. Review on high-k dielectrics reliability issues. *IEEE Transactions on Device and Materials Reliability*, 5(1):5–19, March 2005.

[66] X. Li, J. Le, and L. T. Pileggi. Projection-based statistical analysis of full-chip leakage power with non-log-normal distributions. *Proceedings of Design Automation Conference*, pages 103–108, 2006.

[67] http://techunwrapped.com/2011/08/11/apple-threatens-to-leave-intel-over-power-consumption/, March 2012.

[68] K. Agrawal and S. Nassif. The impact of random device variation on sram cell stability in sub-90-nm cmos technologies. *IEEE Transactions of VLSI Systems*, 16(1):86–97, January 2008.

[69] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl. The impact of intrinsic device fluctuations on cmos sram cell stability. *IEEE Journal of Solid-State Circuits*, 36(4):658–665, April 2001.

[70] B. J. Sheu, D. L. Scharfetter, P. Ko, and M. Jeng. Bsim: Berkley short-channel igfet model for mos transistors. *IEEE Journal of Solid-State Circuits*, 22(4):558–566, 1987.

[71] Psp model manual. http://pspmodel.asu.edu/, April 2012.

[72] T. Zhang. Comparison of psp and bsim4 mosfet model across various parameters. *Proceedings of German Microwave Conference*, pages 32–35, March 2010.

[73] N. Moezi, D. Dideban, B. Cheng, S. Roy, and A. Asenov. Impact of statistical parameter set selection on the statistical compact model accuracy: Bsim4 and psp case study. *Microelectronics Journal*, 2011.

[74] M. Dunga et al., editor. *BSIM-CMG: A compact model for multi-gate transistors.* FinFETs and Other Multi-Gate Transistors. Springer, 2008.

[75] Q. Chen. An exercise of et/utbb soi cmos modeling and simulation with bsim-img. *SOI Conference*, 2011.

[76] O. Rozeau, M. Jaud, T. Poiroux, and M Benosman. Surface potential based model of ultra-thin fully depleted soi mosfet for ic simulations. *in Proc. SOI Conference (SOI)*, pages 1–22, October 2011.

[77] Bsim4.6.0 mosfet model - user's manual. http://www-device.eecs.berkeley.edu/bsim/Files/BSIM4/BSIM460/doc/, 2012.

[78] W Liu. *MOSFET Models for SPICE Simulation: Including BSIM3v3 and BSIM4*, volume 1. Wiley-IEEE Press, 2001.

[79] G. Gasiot, M. Glorieux, S. Uznanski, S. Clerc, and P. Roche. Experimental characterization of process corners effect on sram alpha and neutorn soft error rates. *in Proc. Reliability Physics Symposium*, pages 3C.4.1 – 3C.4.5, 2012.

[80] H. Mahmoodi, S. Mukhopadhyay, and K. Roy. Estimation of delay variations due to random-dopant fluctuations in nanoscale cmos circuits. *IEEE Journal of Solid-State Circuits*, 40(9):1787–1796, September 2005.

[81] B Cheng, D Dideban, N Moezi, C Millar, G Roy, X Wang, S Roy, and A Asenov. Statistical-variability compact-modeling strategies for bsim4 and psp. *IEEE Design and Test of Computers*, March/April 2010.

[82] Y. Cao and C. McAndrew. Mosfet modeling for 45nm and beyond. *Proceedings of the 2007 IEEE/ACM international conference on Computer-aided design*, 2007.

[83] P. Asenov, N. A. Kamsani, D. Reid, C. Millar, S. Roy, and A. Asenov. Combining process and statistical variability in the evaluation of the effectiveness of corners in digital circuit parametric yield analysis. *in*

Proc. Solid-State Device Research Conference (ESSDERC), pages 130–133, September 2010.

[84] T. Hiramoto, M. Sazuki, X. Song, K. Shimizu, T. Saraya, A. Nishida, T. Tsunomura, S. Kamohara, K. Takeuchi, and T. Mogami. Direct measurement of correlation between sram noise margin and individual cell transistor variability by using device matrix array. *IEEE Transactions on Electron Devices*, 58(8), 2249-2256 2011.

[85] M. Suzuki, T. Saraya, K. Shimizu, A. Nishida, S. Kamohara, K. Takeuchi, S. Miyano, T. Sakurai, and T. Hiramoto. Direct measurements, analysis and post-fabrication improvement of noise margins in sram cells unilizing dma sram teg. *Digest of Technical Papers, Symposium on VLSI Technology*, pages 191–192, 2010.

[86] P. R. Kinget. Device mismatch and tradeoffs in the design of analog circuits. *IEEE Journal of Solid-State Circuits*, 40(6):1212–1224, 2005.

[87] M. Miyamura, T. Nagumo, K. Takeuchi, K. Takeda, and M. Hane. Effects of drain bias dependance on threshold voltage fluctuations and its impact on circuit characteristics. *Proceedings of Electron Devices Meeting (IEDM)*, pages 1–4, 2008.

[88] R.L. Wadsack. Fault modeling and logic simulation of cmos and mos integrated circuits. *Bell System Technical Journal*, 57:1449–1474, May-June 1978.

[89] D. M. H. Walker. *Yield simulation for integrated circuits*, volume 1. Springer, 1987.

[90] T. Sasao S. Hassoun. *Logic Synthesis and Verification*. Springer, November 2001.

[91] H. Jyu. Statistcal timing analysi of combination logic circuits. *Transactions on VLSI Systems*, 1(2):126, 1993.

[92] Synopsys. Synopsys primetime. http://www.synopsys.com/.

[93] I. Nitta, T. Shibua, and K. Homma. Statistical static timing analysis technology. *Fujitsu Scientific and Technical Journal*, 43(4):516–523, 2007.

[94] H. Chen and D. H. Du. Path sensitization in critical path problem. *IEEE Transactions on Computer Aided Design*, 12(2):196–207, 1993.

[95] M. Merrett, P. Asenov, Y. Wang, M. Zwolinski, S. Roy, C. Millar, D. Reid, and A. Asenov. Modelling circuit performance variations due to statistical variability: Monte carlo static timing analysis. *Proceedings of Design Automation and Test in Europe*, March 2011.

[96] J. A. G. Jess, K. Kalafala, S. R. Naidu, R. H. J. M. Otten, and C. Visweswariah. Statistical timing for parametric yield prediction of digital integrated circuits. *IEEE Transactions of Integrated Circuits and Systems*, 25(11):2376–2392, November 2006.

[97] Cadence. Spectre circuit simulator. http://www.cadence.com/.

[98] Mentor Graphics. Eldo circuit simulator. http://www.mentor.com/.

[99] Synopsys. Hspice circuit simulaton. http://www.synopsys.com/.

[100] ngSPICE. ngspice circuit simulator. http://ngspice.sourceforge.net.

[101] K. G. Nichols, T. J. Kazmierski, M. Zwolinski, and A. D. Brown. Overview of spice-like circuit simulation algorithms. *IEEE Proc-Circuits Devices and Systems*, 141(4), 1994.

[102] S. Borkar. Designing reliable systems from unreliable components: the challenges of transistor variability and degradation. *IEEE Micro*, 25(6):10–16, Nov/Dec 2005.

[103] D. Burnett, K. Erington, C. Subramanian, and K. Baker. Implications of fundamental threshold voltage variations for high-density sram and logic circuits. *Proceedings of VLSI Technology Symposium*, pages 15–16, 1994.

[104] H. S. Yang et al. Scaling of 32nm low power sram with high-k metal gate. *in Proc. Electron Devices Meeting (IEDM)*, pages 1–4, December 2008.

[105] E. Seevnick, F. J. List, and J. Lohstroh. Static-noise margin analysis of mos sram cells. *IEEE Journal of Solid-State Circuits*, 22(5):748–754, October 1987.

[106] H. Wang, M. Miranda, W. Dehaene, F. Catthoor, and K. Maex. Systematic analysis of energy and delay impact of very deep submicron process variability effects in embedded sram modules. *Proceedings of Design Automation and Test in Europe*, 2005.

[107] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene. Read stability and write-ability analysis of sram cells for nanometer technologies. *IEEE Journal of Solid-State Circuits*, 41(11):2577–2588, November 2006.

[108] Gold standard simulations. http://www.GoldStandardSimulations.com/, March 2012.

[109] G. Castaneda, A. Juge, G. Ghibaudo, D. Golanski, D. Hoguet, J. Portal, and B. Borot. Test structures for interdie variations monitoring in presence of statistical random variability. *in Proc. Microelectronic Test Structures (ICMTS)*, pages 36–42, 2012.

[110] A R Brown, J R Watling, and A Asenov. Intrinsic parameter fluctuations due to random grain orientations in high-Î^{o} gate stacks. *Journal of Computational Electronics*, pages 333–336, December 2006.

[111] A R Brown, V Huard, and A Asenov. Statistical simulation of progressive nbti degradation in a 45-nm technology pmosfet. *IEEE Transactions on Electron Devices*, 57(9):2320–2325, September 2010.

[112] T. Grasser, T. 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.

[113] C. Alexander, G. Roy, and A. Asenov. Random-dopant-induced drain current variation in nano-mosfets: A three-dimensional self-consistent monte carlo simulation study using "ab initio" ionized impurity scattering. *IEEE Transactions on Electron Devices*, 55(11):3251–3258, November 2008.

[114] 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.

[115] W. F. J. Frank and B. S. Berry. *Lattice Location and Atomic Mobility of Implanted Boron in Silicon*, volume 21. Taylor and Francis, 2974.

[116] S. Kaya, A. R. Brown, A. Asenov, D .Magot, and T. Linton. *Analysis of statistical fluctuations due to line edge roughness in sub 0.1um MOSFETs.* Simulation of Semiconductor Processes and Devices. Springer, 2001.

[117] S Xiong and J Bokor. Study of gate line edge roughness effects in 50 nm bulk mosfet devices. *Proceedings of Symposium on Photomask Technology*, pages 733–741, 2002.

[118] H. Fuketome, Y. Momiyama, T. Kubo, E. Yoshida, H. Morioka, M. Tajima, and T. Aoyama. Suppression of poly-gate-induced fluctuations in carrier profiles of sum-50nm mosfets. *In Proc. Electron Devices Meeting (IEDM)*, 2006.

[119] J. J. More. *Numerical Analysis*, volume 630 of *Lecture Notes in Mathematics.* Springer, 1978.

[120] M. N. Alexandrov, J. E. Dennis, R. M. Lewis, and V. Torczon. *Structural and Multidisciplinary Optimization.* Number 1. Springer, 1998.

[121] M. Marazzi and J. Nocedal. *Wedge trust region methods for derivative free optimization*, volume 91 of *Mathematical Programming.* Springer, 2002.

[122] C. C. McAndrew and P. A. Layman. Mosfet effective channel length, threshold voltage, and series resistance determination by robust optimization. *IEEE Transactions on Electron Devices*, 39(10):2298–2302, October 1992.

[123] B Cheng, D Dideban, N Moezi, C Millar, G Roy, X Wang, S Roy, and A Asenov. Statistical-variability compact-modeling strategies for bsim4 and psp. *IEEE Design and Test of Computers*, pages 26–30, 2010.

[124] B. Cheng, N. Moezi, D. Dideban, G. Roy, S. Roy, and A. Asenov. Benchmarking the accuracy of pca generated statistical compact model parameters against physical device simulation and directly extracted statistical parameters. *in Proc. Simulation of Semiconductor Processes and Devices*, pages 143–146, Sept. 2009.

[125] A. Lange, C. Sohrmann, R. Jancke, J. Hasse, B. Cheng, U. Kovac, and A. Asenov. A general approach for multivariate statistical mosfet compact modelling preserving correlations. *Proceedings of Solid-State Device Research Conference*, pages 163–166, September 2011.

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

[127] S. Ahn and J. A. Fessler. Standard errors of mean, variance, and standard deviation estimators. *EECS Department, The University of Michigan*, July 2003.

[128] J. V. Michalowicz, J. M. Nichols, F. Bucholtz, and C. C. Olsson. An isserlis' theorem for mixed gaussian variables: Application to the auto-bispectral density. *Journal of Statistical Physics*, 136(1):89–102, 2009.

[129] J. Ramberg and B. Schmeiser. An approximate method for generating asymmetric random variables. *Communications of the ACM*, 17(2):78–82, 1974.

[130] C. Z. Mooney. *Monte Carlo Simulation*, volume 116 of *Quantitative Applications in the Social Sciences*. Sage University, 1997.

[131] M. Freimer, G. Kollia, G. S. Mudholkar, and T. Lin. A study of the generalized tukey lambda family. *Communications in Statistics*, 17(10):3547–3567, 1988.

[132] R. A. R. King and H. L. MacGillivary. A starship estimation method for the generalised lambda distributions. *Australia and New Zeland Journal of Statistics*, 43(3):353–374, 1999.

[133] B. W. Silverman. *Density estimation for statistics and data analysis.* Chapman and Hall, 1986.

[134] J. S. Ramberg, P. R. Tadikamalla, E. J. Dudewicz, and E. F. Mykytka. A probability distribution and its uses in fitting data. *Technometrics*, 21(2):201–214, May 1979.

[135] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1, 1968.

[136] T. Mizutani, A. Kumar, and T. Hiramoto. Measuring threshold voltage variability of 10g transistors. *Proceedings of Electron Devices Meeting (IEDM)*, pages 25.2.1–25.2.4, 2011.

[137] ARM Ltd. 25nm sram system schematic. Personal Correspondance.

[138] H. Yamauchi. A discussion on sram circuit design trend in deeper nanometer-scale technologies. *IEEE Transactions on VLSI Systems*, 18(5):763–774, May 2010.

[139] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey. Sram leakage suppression by minimizing standby supply voltage. *Proceedings of Symposium on Quality Electronic Design*, pages 55–60, SRAM, Leakage, Decorr 2004.

[140] J. Wang, S. Nalam, and B. H. Calhoun. Analyzing static and dynamic write margin for nanometer srams. *Proceedings of Low Power Electronics and Design (ISLPED)*, pages 129–134, 2008.

[141] J. Lohstroh, E. Seevinck, and J. de Groot. Worst-case static noise margin criteria for logic circuits and their mathematical equivalence. *IEEE Journal of Solid-State Circuits*, 18(6):803–807, December 1983.

[142] T. Fischer, E. Amirante, P. Huber, T. Nirschl, A. Olbrich, M. Ostermayr, and D. Schmitt-Landsiedel. Analysis of read current and write trip voltage varibility from a 1-mb sram test structure. *IEEE Transactions on Semiconductor Manufacturing*, 21(4):534–541, November 2008.

[143] R. Wong, D. J. Frank, R. Mann, K. Sang-Bin, P. Croce, D. Lea, D. Hoyniak, L. Yoo-Mi, J. Toomey, M. Weybright, and J. Sudijono. Sram cell design for stability methodology. *Proceedings of VLSI Technology Symposium*, pages 21–22, 2005.

[144] E. I. Vatajelu and J. Figueras. Supply voltage reduction in srams: Impact on static noise margin. *Proceedings of Automation, Quality and Testing*, pages 73–78, May 2008.

[145] H. Pilo, C. Barwin, G. Braceras, C. Browning, S. lamphier, and F. Towler. An sram design in 65-nm technology node featuring read and write-assist circuits to expand operating voltage. *IEEE Journal of Solid-State Circuits*, 42(4):813–819, April 2007.

[146] F. Hamzaoglu, Y. Wand, P. Kolar, L. Wei, Y. Ng, U. Bhattacharya, and K. Zhang. Bit cell optimizations and circuit techniques for nanoscal sram desing. *IEEE Design and Test of Computers*, pages 23–31, January 2011.

[147] S. Natarajan et al. A 32nm logic technology featuring 2nd generation high-k metal-gate transistors, enhanced channel strain and a 0.171um2 sram cell size in a 291mb array. *In Proc. Electron Devices Meeting (IEDM)*, December 2008.

[148] J. Wang, P. Liu, Y. Gao, P. Deshmukh, S. Yang, Y. Chen, W. Sy, L. Ge, E. Terzioglu, M. Abu-Rahma, M. Garg, S. Seung Yoon, M. Han, M. Sani, and G. Yeap. Non-gaussian distribution of sram read current and design

impact to low power memory using voltage acceleration method. *Symposium on VLSI Techonlogy, Digest of Technical Papers*, pages 220–221, 2011.

[149] C. Visweswariah. First-order incremental block-based statistical timing analysis. *IEEE Computer-Aided Design of Integrated Circuits and Systems*, 25(10):2170–2180, 2006.

[150] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic. *Digital Integrated Circuits*. Prentica-Hall, 1996.

[151] University of Manchester. Meeting the design challenges of nanocmos electronics. http://www.cs.manchester.ac.uk/aboutus/news/archives.

[152] K. Bernstein, D.J. Frank, A.E. Gattiker, W. Haensch, B.L. Ji, S.R. Nassif, E.J. Nowak, D.J. Pearson, and N.J. Rohrer. High-performance cmos variability in the 65-nm regime and beyond. *IBM Journal of Research and Development*, 50(5/6):433–449, 2006.