Statistical Circuit Simulations - From 'Atomistic' Compact Models to Statistical Standard Cell Characterisation

Noor 'Ain Kamsani

Submitted in fulfilment of the requirements for
the Degree of Doctor of Philosophy

School of Engineering
University of Glasgow

January 2011

Abstract

This thesis describes the development and application of statistical circuit simulation methodologies to analyse digital circuits subject to intrinsic parameter fluctuations. The specific nature of intrinsic parameter fluctuations are discussed, and we explain the crucial importance to the semiconductor industry of developing design tools which accurately account for their effects. Current work in the area is reviewed, and three important factors are made clear: any statistical circuit simulation methodology must be based on physically correct, predictive models of device variability; the statistical compact models describing device operation must be characterised for accurate transient analysis of circuits; analysis must be carried out on realistic circuit components. Improving on previous efforts in the field, we posit a statistical circuit simulation methodology which accounts for all three of these factors. The established 3-D Glasgow atomistic simulator is employed to predict electrical characteristics for devices aimed at digital circuit applications, with gate lengths from 35 nm to 13 nm. Using these electrical characteristics, extraction of BSIM4 compact models is carried out and their accuracy in performing transient analysis using SPICE is validated against well characterised mixed-mode TCAD simulation results for 35 nm devices. Static d.c. simulations are performed to test the methodology, and a useful analytic model to predict hard logic fault limitations on CMOS supply voltage scaling is derived as part of this work. Using our toolset, the effect of statistical variability introduced by random discrete dopants on the dynamic behaviour of inverters is studied in detail. As devices scaled, dynamic noise margin variation of an inverter is increased and higher output load or input slew rate improves the noise margins and its variation. Intrinsic delay variation based on CV/I delay metric is also compared using I_{ON} and I_{EFF} definitions where the best estimate is obtained when considering I_{ON} and input transition time variations. Critical delay distribution of a path is also investigated where it is shown non-Gaussian. Finally, the impact of the cell input slew rate definition on the accuracy of the inverter cell timing characterisation in NLDM format is investigated.

Acknowledgement

Completion of this thesis marks another important milestone in my life. Here, I would like to thank the people who have helped me through this journey.

First, I would like to thank my supervisors, Dr. Scott Roy and Prof. Asen Asenov. This work would not be possible without their endless encouragement and support to further understand the wonders of the semiconductor physical world. I am very grateful for the chance to meet and work with them whose charismatic personalities and skills have never failed to inspire me in achieving greater things in life.

Not to forget the members of Device Modeling Group who are always willing to offer help when needed. All the good time we spent together will always be remembered. I also would like to thank the Malaysian community in Glasgow for helping me to cope with all the distress for being thousand miles away from families. To my husband, who also happened to be one of my colleagues in Device Modeling Group, thank you for being supportive all the way. To my families, thank you so much for taking care of Neil Luqman while both of us are completing our studies here.

Finally, thank you to Ministry of Higher Education, Malaysia and Universiti Putra Malaysia for their financial support in completing this research at University of Glasgow, Scotland, U.K.

Thank you all.

"*Jasamu dikenang sepanjang hayat.*"

Publications

1. ***ESSDERC/ESSCIRC 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," *European Solid-State Device Research Conference/European Solid-State Circuits Conference (ESSDERC/ESSCIRC), 2010*: September 13-17, 2010.

2. ***ULIS 2010 (Talk)*** : N. A. Kamsani, B. Cheng, C. Millar, N. Moezi, X. Wang, S. Roy and A. Asenov, "Impact of Slew Rate Definition on the Accuracy of nanoCMOS Inverter Timing Simulations," *Ultimate Integration on Silicon (ULIS), 2010*: March 17-19, 2010.

3. ***ULIS 2010 (Talk)*** : D. Dideban, B. Cheng, N. Moezi, N. A. Kamsani, C. Millar, S. Roy and A. Asenov, "Effect of Input Slew Rate on Statistical Timing and Power Dissipation Variability in nanoCMOS, " *Ultimate Integration on Silicon (ULIS), 2010*: March 17-19, 2010.

4. ***ISCAS 2009 (Talk)*** : N. A. Kamsani, B. Cheng, S. Roy, and A. Asenov, "Impact of Random Dopant Induced Statistical Variability on Inverter Switching Trajectories and Timing Variability," *IEEE International Symposium on Circuits and Systems (ISCAS), 2009*: May 24-27, 2009.

5. ***FTFC 2008 (Talk)*** : N. A. Kamsani, B. Cheng, S. Roy, and A. Asenov, "Statistical Circuit Simulation with Supply-Voltage Scaling In Nanometre MOSFET Devices Under The Influence of Random Dopant Fluctuations," *7th edition of Faible Tension Faible Consommation (FTFC), 2008*: May 26-28, 2008.

6. ***DATE 2008 (Poster)*** : N. A. Kamsani, B. Cheng, S. Roy, and A. Asenov, "Statistical Circuit Simulation with the Effect of Random Discrete Dopants in Nanometer MOSFET Devices," *Design Automation and Test in Europe (DATE): Workshop W2, Impact of Process Variability on Design and Test, 2008*: March 10-14, 2008.

*** Paper 1-5 were accepted after undergone a peer-review process.

Table of Contents

List of Tables

List of Figures

Nomenclature

IPF	-	Intrinsic parameter fluctuations
MOSFET	-	Metal-oxide-semiconductor field-effect transistor
RDD	-	Random discrete dopants
SV	-	Statistical variability
SRAM	-	Static random access memory
SPICE	-	Simulation program with integrated circuit emphasis
CMOS	-	Complementary metal-oxide-semiconductor
TCAD	-	Technology computer-aided design
BSIM	-	Berkeley short-channel IGFET model
IGFET	-	Insulated-gate field-effect transistor
NLDM	-	Non-linear delay model
GIDL	-	Gate-induced drain leakage
OPC	-	Optical proximity correction
STI	-	Shallow trench isolation
LER	-	Line edge roughness
OTV	-	Oxide thickness variation
PSG	-	Poly-silicon granularity
SOI	-	Silicon on insulator
STA	-	Static timing analyses
ITRS	-	International technology roadmap for semiconductor
EOT	-	Equivalent oxide thickness
DG	-	Density gradient

DIBL	-	Drain-induced barrier lowering
RMS	-	Root-mean-square
NM	-	Noise margin
SM	-	Safety margin
PSP	-	Penn State - Philips
Hi-SIM	-	Hiroshima-university STARC IGFET model
STARC	-	Semiconductor technology academic research center
EKV	-	Enz - Krummenacher - Vittoz
HSPICE	-	SPICE simulator from Synopsys
Sentaurus	-	TCAD simulation platform for semiconductor devices from Synopsys
Aurora	-	Parameter extraction tool from Synopsys
LDD	-	Lightly-doped drain
SCE	-	Short-channel effects
FO	-	Fan out
FI	-	Fan in
IC	-	Integrated circuit
LIBERTY	-	Gate-level modeling technology to address standard cell library modeling requirement from Synopsys
ELDO	-	SPICE simulator from Mentor Graphics
SPECTRE	-	SPICE simulator from Cadence
VLSI	-	Very-large-scale integration
CCS	-	Composite current source
CUT	-	Cell under test
In	-	Indium
As	-	Arsenic
SiGe	-	Silicon germanium

I-V	-	Current-voltage characteristic
C-V	-	Capacitance-voltage characteristic
V_{GS}	-	Gate-to-source voltage
V_{DS}	-	Drain-to-source voltage
V_{BS}	-	Bulk-to-source voltage
x_j	-	Junction depth
V_{DD}	-	Power supply voltage
GND	-	Ground
W_n	-	Gate width of *n*-MOSFET
L_n	-	Gate length of *n*-MOSFET
V_T	-	Threshold voltage
C_{ox}	-	Oxide capacitance
σ	-	Standard deviation
μ	-	Population mean
V_{gsteff}	-	Effective gate-to-source voltage
V_{dseff}	-	Effective drain-to-source voltage
C_{gs}	-	Total gate-to-source (intrinsic and parasitic) capacitance
C_{gd}	-	Total gate-to-drain (intrinsic and parasitic) capacitance
C_{gb}	-	Total gate-to-bulk (intrinsic and parasitic) capacitance
C_{bs}	-	Total bulk-to-source (intrinsic and parasitic) capacitance
C_{bd}	-	Total bulk-to-drain (intrinsic and parasitic) capacitance
C_{gc}	-	Gate-to-channel capacitance
C_{ov}	-	Overlap capacitance
C_{of}	-	Outer fringe capacitance
C_{if}	-	Inner fringe capacitance
C_j	-	Junction capacitance
C_{DS}	-	Intrinsic drain-to-source capacitance

Q_B	-	Bulk charge
Q_{inv}	-	Inversion charge
Q_G	-	Gate charge
Q_D	-	Drain charge
Q_S	-	Source charge
X_{part}	-	Charge partitioning parameter
acde	-	Exponential coefficient for charge thickness in accumulation and depletion regions in CAPMOD = 2
noff	-	CV parameter in $V_{gsteff,CV}$ for weak to strong inversion
moin	-	Coefficient for the gate-bias dependent surface potential
C_L	-	Fixed load capacitor
L_d	-	Logic depth
T_{DHL}	-	Propagation delay of an inverter with falling-output transition
T_{DLH}	-	Propagation delay of an inverter with rising-output transition
T_P	-	Propagation delay of two successive inverters
SR	-	Slew rate
T_T	-	Input transition time
C_{12}	-	Coupling capacitance between the aggressor and victim's interconnect
C_{victim}	-	Capacitance at the victim interconnect to ground
CV/I	-	Delay metric
I_{ON}	-	Drain current at $V_{GS} = V_{DS} = V_{DD}$
I_{EFF}	-	Effective drain current which is the average of I_{D_H} and I_{D_L}
I_{D_H}	-	Drain current at $V_{GS} = V_{DD}$ and $V_{DS} = V_{DD}/2$
I_{D_L}	-	Drain current at $V_{GS} = V_{DD}/2$ and $V_{DS} = V_{DD}$

T_{MAX}	-	Maximum period in a given clock cycle
f	-	local clock frequency of chip
T_{CRIT}	-	Critical path delay
τ	-	Intrinsic delay
P_{DYN}	-	Dynamic power
P_{LEAK}	-	Leakage power
P_{SCC}	-	Short circuit power
P_{AVG}	-	Average power
T	-	Clock cycle

Chapter 1

Introduction

The key economic driver of the global semiconductor industry is its ability to continually increase the useful, reliable functionality of each square centimetre of a semiconductor substrate. This driver is related to the number of transistors which can be fabricated per unit area, and therefore to the size of each transistor. Since silicon began to be used extensively to make integrated circuits in the 1960s, many studies on the limitations of technology scaling in terms of economics, manufacturability, material properties (for instance, thermal dissipation) and physical limitations in the transistor operation, have been carried out. This work contributes to the understanding of the limitations associated with intrinsic parameter fluctuations (IPF), which are caused by the discreteness and granularity of a matter in small devices. Interestingly, such effects were first forecast in the 1970s [1], about 20 years before they became critical for the future of device scaling and integration [2][3].

Studies have shown that for conventional Si bulk-MOSFETs, the magnitude of the IPFs rapidly increase as device dimensions are reduced. This is partially due to the relative reduction in the number of random discrete dopants (RDD) in the MOSFET channel that control the electrical properties of the transistors [4]. It is also due to a reduction in the physical oxide thickness and printed gate length, whilst the atomic scale roughness and the line edge roughness remain constant, leading to large percentage of random oxide thickness and gate length variations [5]. In addition,

new processing steps introduced to increase device performance, such as the introduction of high-*k* materials, may also contribute to a larger IPFs in smaller devices [6]. Each source of IPF at the device level introduces statistical variability (SV) at the circuit level.

While the SV have affected analogue circuits and circuit design for a number of technology generations, they have now begun to cause problems in the digital circuit domain. Static Random Access Memory (SRAM) has been the first victim of SV effects in the digital domain due to its minimal transistor size. Failures in the operation of SRAM cells already affect manufacturing yield, and require the addition of redundant cells in the design process [7][8]. In contrast to SRAM, digital logic gates typically have greater device channel widths resulting in less statistical variability which scales typically as $1/\sqrt{(WL)}$. However, they have also started to suffer from SV effects [9][10]. Failures in the functions of an SRAM or digital logic cell clearly compromise the system that contains them. However digital systems also specify a target operating frequency at specified power consumption, and SRAM or digital logic cells which operate too slowly will also increase the parametric yield loss in the design. To overcome such effects, extra design margin is added during the design verification process, which is seen as a source of design waste if it is not properly managed.

In conventional physical implementation flows, process variability is handled using corner analysis: late (setup) analysis at weak, min-voltage, high-temperature conditions and early (hold) analysis at strong, max-voltage, low-temperature conditions. However, with advances in technology, more sources of variability, larger magnitudes of variability, and the possibility of correlations between sources, there are too many corners to be considered in designs using smaller devices. This makes the worst and best case validation technique very pessimistic in designs [11].

The technique of *statistical design* has been posited for the purpose of obtaining a more optimal design before real tape-out process. Successful tape-out in

65 nm technology employing such statistical design techniques has been reported recently [12]. However, the cost effectiveness of this technique is still questionable by the majority of design communities and in-depth analysis of the statistical design flow is still needed to understand at which design level this technique is best suited. To migrate from corner analysis into statistical design also raises challenges that need to be addressed properly in order to tackle the variability issues with confidence at every targeted digital design level. One of these challenges flagged was the statistical library characterisation with accurate representation of statistical variation in advanced technology for use in statistical tools. In order to achieve such accurate characterisation, a proper treatment is needed when considering the parameter variables especially the ones that are difficult to characterise, such as IPFs. Another challenges is the lack of suitable and robust statistical simulation and verification tools. Such tools must be capable of interfacing with the existing tools in a designated design flow.

1.1 Aim and Objectives

The aim of this research is to study in detail the impact of statistical variability on digital circuits and systems. We shall consider integrated circuit designs using well-scaled Si bulk-MOSFET devices which have been carefully calibrated to match state-of-the-art devices designed for the 45 nm technology node and beyond. Device level variability may be obtained directly from experimental measurements, or in our case, obtained from statistical 3-D numerical simulations carried out by the 3-D ‘atomistic’ device simulator developed at the University of Glasgow. We will investigate on statistical scale the performance variation of circuits which are subject to SV. This will be carried out using a hierarchical simulation technology integrating 'atomistic' compact models based on physical simulation of statistical variability into statistical SPICE circuit simulation tools.

- The first objective is to investigate the limitations of supply voltage scaling in digital circuits when subject to SV for 35 nm, 25 nm, 18 nm and 13 nm gate length devices. This topic is of the interest of circuit designers as supply voltage has always be one of the means of managing the total power consumption of integrated circuits. As the magnitude of IPFs increases at each subsequent technology generation, we will predict the minimum supply voltage for each particular technology based on developed models detailed in Chapter 4, considering the combined effects of RDD, LER, OTV, and PSG.
- The second objective is to study the accuracy of time dependent circuit simulations; comparing compact model simulation against physical device simulations. In past studies, the 'atomistic' compact models developed by the Device Modelling Group of University of Glasgow have mainly been used in static circuit analyses. In order to expand the work to transient circuit analyses, further calibration is needed to ensure the simulated device in the numerical simulation matches the SPICE circuit simulation using BSIM compact model for the well-scaled Si bulk-MOSFET devices. The second objective is addressed in Chapter 5.
- The third objective is to perform an exhaustive statistical study of the dynamic behaviour and performance of the most fundamental CMOS circuit, the inverter, and of chains of inverters, all subject to underlying statistical variability in their constituent MOSFETs. The comprehensive investigation should lead to a more detailed understanding of the noise susceptibility of the inverter when subject to device scaling and SV which is crucial for circuit designers in managing signal integrity of the designed circuit. This study would also evaluate delay distribution under different conditions of fan-in and fan-out (*FO/FI*), load and input slew rate to give a better insight into the statistical delay model to be incorporated into any statistical timing analysis tool. Lastly, we will also investigate delay

variation in more complex circuits subject to RDD and device scaling. This study will help to identify key areas for circuit optimisation when subject to SV for future technology generations. The third objective is addressed in Chapter 6.

- The fourth objective is to study the accuracy of different standard cell characterisation techniques in capturing the delay information of fundamental CMOS system building blocks, called standard cells in the industry terminology, for higher level of abstraction usage for the 45 nm technology node and beyond. This study will help to identify limitations in the current standard cell format, the Non-Linear Delay Model, which is still widely used at the 65 nm technology node. The last objective is addressed in Chapter 7.

In fulfilling these objectives, we will develop a set of simulation and analysis methodology and technology which can be applied to any small-to-medium scale circuit netlist, and form the foundation of a SV toolkit for statistical timing analysis. We trust that such technology will be of great assistance to designers trying to develop more robust and reliable circuits at the 45 nm technology node and beyond, in the presence of large CMOS SV.

1.2 Thesis Outline

The rest of the thesis is laid out as follows:

CHAPTER 2 - Background

An overview of device scaling and its major limitations is first given. Followed by the impact of scaling and intrinsic parameter fluctuations in digital logics when subject to device scaling is entailed. An overview of statistical design and its advantages and disadvantages is also given.

CHAPTER 3 - Statistical Simulation Methodology

The statistical simulation methodology employed in this study is described in detail. A brief discussion on devices used in this study is presented. It is followed by description of Glasgow 3-D 'atomistic' device simulator, statistical compact models and the statistical circuit simulation procedures.

CHAPTER 4 - Hard Logic Fault Related Supply Voltage Limitations due to MOSFET Variability

An analytic model is developed which predicts the minimum supply voltage for digital circuits in the presence of SV – the voltage at which steady state faults become unavoidable. Supply voltage limitations are discussed for devices subject to device scaling, based on collected data from the literature.

CHAPTER 5 - Accuracy of Transient Simulation Using BSIM Compact Models

Device characterisation of 35 nm gate length *n*- and *p*-channel MOSFETs, developed using careful TCAD calibration, is performed. The accuracy of the resulting BSIM compact models is evaluated against TCAD simulation. It is shown that BSIM compact models can be part of an accurate and computationally efficient methodology for performing accurate time dependent circuit simulations in the presence of variability.

CHAPTER 6 - Inverter Performance Variability Due To Random Discrete Dopants

Dynamic noise margin, timing and power variation are studied in detail for CMOS inverters and chains of inverters. At the end of this chapter, the impact of random discrete dopants (the major source of variability in bulk devices) on delay in inverters subject to device scaling is described.

CHAPTER 7 - Accuracy of Standard Cell Characterisation Techniques

The standard Non-Linear Delay Model (NLDM) approach to recording the timing information of a circuit building block or *standard cell* is evaluated for 35 nm and 25 nm gate length devices developed at the University of Glasgow.

CHAPTER 8 - Conclusions and Future Work

Lastly, conclusions of this research are drawn in this chapter and possible future work is laid out.

Chapter 2

Background

In this chapter, the purpose of device scaling and the major bottlenecks to scaling are discussed, including intrinsic parameter fluctuations. Then, a description of the primary sources of intrinsic parameter fluctuations and their impact on device characteristics is given. A discussion of statistical design, as a method of coping with the problems introduced by intrinsic parameter fluctuations follows.

2.1 Device Scaling

For four decades, Moore's law [13] has driven the semiconductor industry in the pursuit of smaller geometry/higher performance devices. The continued shrinking of horizontal and vertical features size improves device density on a chip and reduces the cost per function. However, the historical use of generalised scaling, which was achieved by reducing gate dielectric thickness and gate length, and increasing the channel doping is no longer achievable due to physical and technological limitations [14][15]. New technology boosters involving changes in device materials and processing have been adopted to comply with the speed and power requirements of Moore's law for advanced technology nodes, in association with geometrical scaling [16][17].

One of the critical problems of conventional scaling is that the oxide thickness scaling needed to provide sufficient drive current at reduced supply

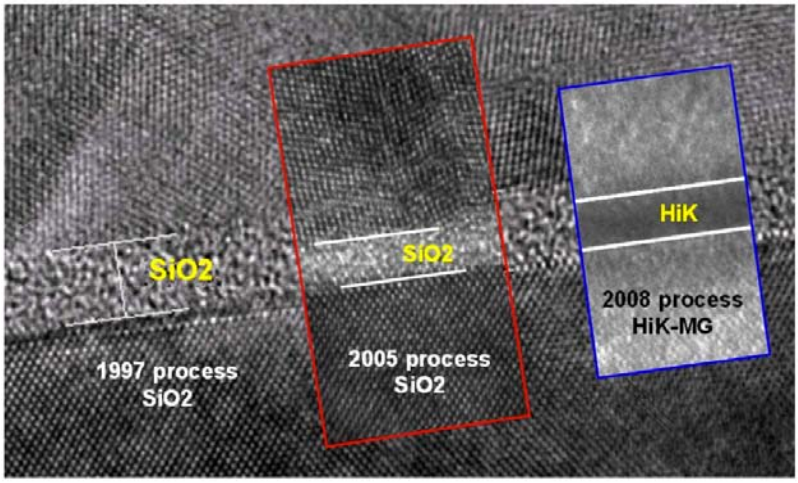


Figure 2-1 : Oxide thickness scaling reached atomic scale [18].

voltage in bulk-MOSFET devices, has reached a fundamental limit. By scaling the oxide thickness aggressively to ~1 nm, the direct gate tunnelling current through the oxide has become a significant issue. As a result, the power dissipation associated with the direct tunnelling gate current has become a major contributor to the overall chip leakage and standby power dissipation [19]. Further reduction of the oxide thickness will exponentially increase the tunnelling current and hence greatly affect the power dissipation, which is especially problematic for low-power applications, and is the major reason for the introduction of high-κ hafnium-based dielectrics at the 45 nm technology node [20]. Fig. 2-1 illustrates oxide thicknesses for different processes and materials for three technology nodes. Introduction of high-κ materials have enabled the use of physically thicker dielectrics while maintaining the scaling of device equivalent oxide thickness.

A second problem with conventional device scaling is the high-channel doping that bulk-MOSFETs require to control short-channel effects. Reduction of channel length without increasing the channel doping causes threshold voltage rolloff and punch-through. Even though threshold voltage scaling is desirable to increase the gate overdrive (V_{GS}-V_{th}) and hence increase switching speed, the subthreshold leakage current increases exponentially with a linear reduction in the threshold voltage. Large subthreshold leakage current may lead to unacceptably high power consumption. Use of shallow source and drain extensions, and lateral

nonuniform doping such as pocket implants compensate the threshold voltage rolloff and avoid punch-through [21]. However, the high doping concentration results in mobility reduction due to an increase in ionised impurity scattering and performance degradation [22]. Process induced strain has been introduced to compensate for the associate performance loss [23][24]. High channel doping also introduces direct band-to-band leakage in the drain region and severe gate-induced-drain-leakage (GIDL) effects [25][26].

2.2 Device Process Variability

Apart from the scaling obstacles discussed above, process variability also has become increasingly problematic in device scaling. It causes circuit layout or electrical parameters to vary from the designed values, and hence can lead to catastrophic or parametric yield losses. The device process variability can be categorised into global and local variations.

In global variation, the physical parameter variations induced by manufacturing processes such as the oxide layer thickness, gate length and doping concentration change gradually across the chip/wafer. This type of variation is related to the inaccuracy of process parameters and non-uniformity of the equipment used to fabricate the devices. However, this type of variation can be controlled by using more accurate process control or better manufacturing equipment and over time, as new technology matures, this type of variation may be greatly reduced.

Local variation, which is associated with the fluctuations of physical or electrical parameters of transistors within a die, arises due to the physics of manufacturing process. It can be divided into systematic and random variations as shown in Table 2-1 and illustrated in Fig. 2-2.

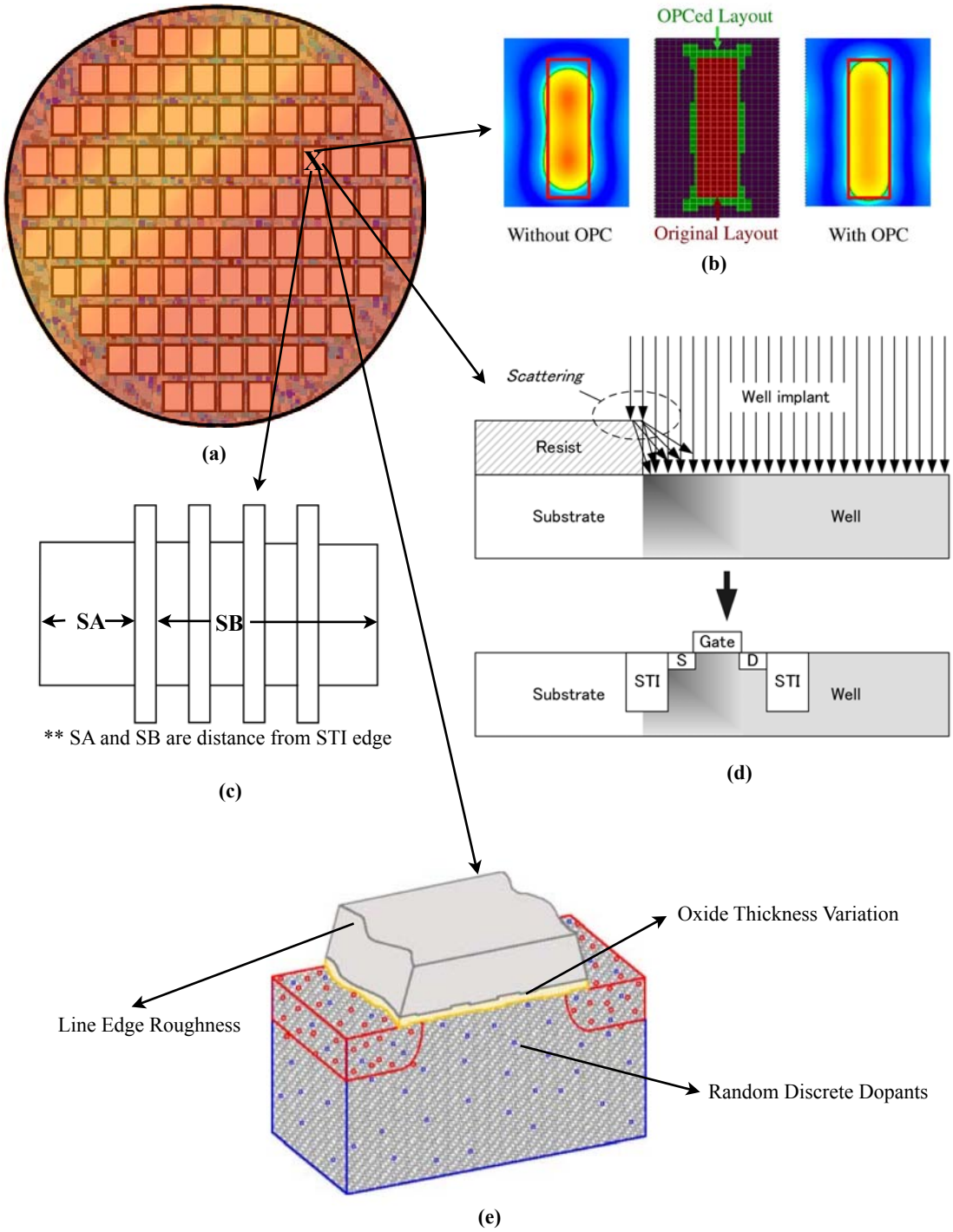


Figure 2-2 : Illustration of local variation on a die *X*, marks on the wafer in (a). Also known as with-in die variation. (b) Schematic representation of optical proximity error and optical proximity correction [27]. (c) Schematic representation STI induced stress in a layout [28]. (d) Schematic representation of well edge proximity effect [29]. (e) Schematic representation of random variation which includes line edge roughness, oxide thickness variation and random discrete dopants [30]. As can be seen from the figures (b-d) the variation can be estimated from a layout while (e) can randomly occurs in any transistor across the die *X*.

TABLE 2-1
Categorisation of device variation.

Local Variation	Causes
Systematic	Optical Proximity Effect
	Layout Mediated Strain
	Well Proximity
Random	Random Dopants
	Line Edge Roughness
	Poly-Si Granularity
	Interface Roughness
	High-κ Morphology

Systematic variation is the component of the physically varying parameters that follow a well understood behaviour and can be predicted or modelled up-front. Examples of systematic variations are the optical proximity effect [31], layout mediated strain [32] and the well proximity effect [29]. The optical proximity effect is the result of diffraction phenomena during patterning process of transistors, which results in structure irregularities where a printed width line is either narrower or wider than the designed layout, as illustrated in Fig. 2-2(*b*). This effect is more pronounced at smaller technology nodes because the wavelength of the light used for patterning is larger in comparison to the gate feature length [33][34]. For example at the 45 nm technology node, the printed feature length of the transistor is approximately 5.5 times smaller than the 193 nm light that prints it [35][36].

Strain engineering was first introduced in the 90 nm technology node to increase carrier mobility, and has now become an essential component of modern transistors [37][38]. However, the introduced strain is layout dependent, and as a result, varies the drive current in transistors with different geometrical layouts and spatial arrangements on the die. The strain-enhanced mobility strongly depends on the spacing between transistors, distances from the shallow trench isolation (STI) and different number and position of contacts [39][40].

The well proximity effect arises during the implant process where dopants scatter laterally from the edge of the photoresist mask and implanted in the silicon

surface in the vicinity of the well edge, as illustrated in Fig. 2-2(*d*). As a result, non-uniform doping concentration within the well causes the transistors which are near to the edge of the well to vary in their threshold voltage and drive current from devices that are located remotely from the edge.

All these systematic variations can either be eliminated by adopting more refined manufacturing techniques (such as optical proximity correction [27]) or accurately estimated as a function of circuit layout as shown in Fig. 2-2 (*b* to *d*). Accurate estimation of layout dependent variability allows it to be accounted for in the circuit design process, greatly reducing the design margin. However, random variations (shown in Fig. 2-2 (*e*)) cannot be eliminated due to more refined processing, or modelled deterministically, as they are a fundamental result of the discreteness of charge and matter. This type of variability must be margined in circuit and system simulations, and will be discussed next.

2.3 Intrinsic Parameter Fluctuations

The intrinsic parameter fluctuations (IPF) which arise from the discreteness of charge and the granularity of matter have become a serious threat to device scaling and integration. They have become prominent in extremely scaled devices as the physical device dimensions approach the atomic scale. In contrast to the other types of process variability, no tightening of process control or uniformity can mitigate the impact of IPF on bulk devices. The intrinsic parameter fluctuations will affect design, yield, and pose difficulties in circuit simulation and verification for future technology nodes.

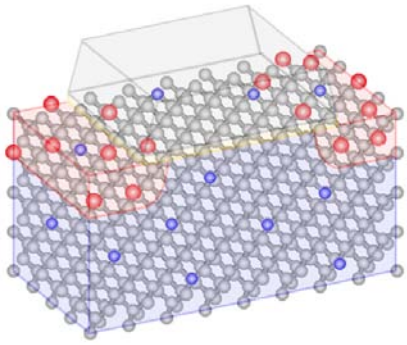


Figure 2-3 : Illustration of RDD in 4.2 nm channel length transistor [41]. Blue and red dots represent dopants while grey dots indicate the silicon lattice.

Random discrete dopants (RDD), introduced by the implantation process in the fabrication of transistors have been shown to be the main source of statistical variability in modern bulk MOSFETs. Experimental studies show 60-65% of the total threshold voltage variation in 65 nm and 45 nm bulk-MOSFETs results from RDD [43]. As devices scale, the number of dopants in the device channel decreases, and a small fluctuation in the number and arrangement of such dopants causes a significant change in device threshold voltage. The dopants induce potential variation locally in the channel and cause the devices to turn-on at different applied gate biases depending on the specific microscopic arrangement or number of dopants in the active region [4]. Fig. 2-3 shows discrete dopants in a hypothetical 4.2 nm gate length transistor.

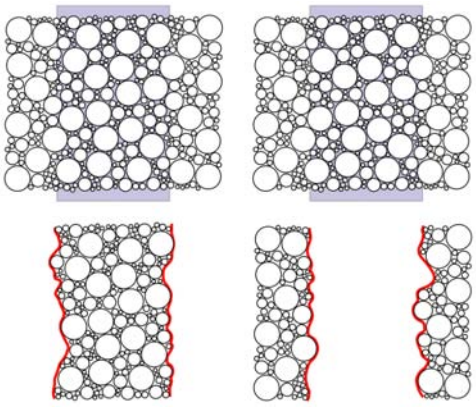


Figure 2-4 : Illustration of LER with positive (left) and negative (right) photoresist [41].

Another source of IPF is line edge roughness (LER), arising from the polymer nature of the photoresist used in the lithographic process as illustrated in Fig. 2-4. As devices scale, the magnitude of this molecular line edge roughness causes appreciable local fluctuations in the channel length across the width of a device [44]. It has been demonstrated that if the magnitude of this roughness cannot be scaled below the current levels, LER could become a dominant source of variability when the transistors are scaled below 20 nm channel length [5]. At high drain bias, the local regions of shorter channel length

induced by LER, lower the threshold voltage and have detrimental effect on the sub-threshold leakage current, which is exponentially dependent on local channel length.

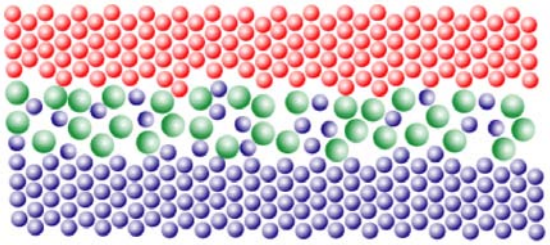


Figure 2-5 : Illustration of OTV at the Si/SiO_2 interface [41].

Another source of IPF is the oxide thickness variations (OTV) associated with Si/SiO_2 interface roughness at the channel-oxide and poly-oxide boundaries due to the molecular nature of the oxide and the poly-silicon as shown in Fig. 2-5. With device scaling, the oxide layer has now reached ~1 nm, equivalent to approximately five inter-atomic spacings [45] and a thickness roughness of the scale silicon lattice atomic spacings is approximately 0.28 nm [46] [47]. These fluctuations cause local potential variation across the channel and contribute to the total threshold voltage variation [48]. These fluctuations also cause significant variability in the gate tunnelling current as the tunnelling current is exponentially dependent on the oxide thickness [49].

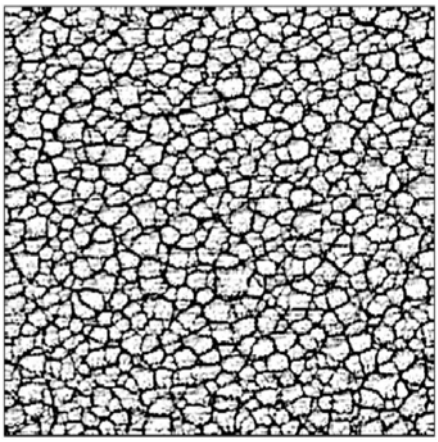


Figure 2-6 : SEM micrograph of typical PSG from bottom [52].

The granular structure of the polysilicon (poly-Si) gate has also been identified as another important source of IPF, termed poly-silicon granularity (PSG). These fluctuations are most likely caused by Fermi-level pinning at the boundaries between grains due to a high density of defect states [50][51]. The Fermi-level pinning of grain boundaries at the poly-Si/gate-oxide interface induces fluctuations in surface potential within the MOSFET channel and causes a variation in threshold voltage and current characteristics from one device to another. The magnitude of these

fluctuations depend on the unique location of the poly-Si grain boundaries in the gate with respect to the channel in each individual transistor [52].

As the sources of variability described above result from the atomicity of the charge and the granularity matter, the introduction of new materials and processes is unlikely to eliminate them – although it may be possible to adjust their relative magnitudes and improve the resistance of devices to some sources of variability. In addition, the introduction of new materials or processes may also *introduce* new sources of IPF. For instance, the introduction of high-κ dielectrics and metal gates can introduce additional variability due to local fluctuations in the composition of the high-κ dielectric [53]. Overall, at the 65 nm and 45 nm technology nodes, it has been experimentally shown that RDD in the channel and source/drain regions is the major source of IPF in contemporary bulk MOSFETs [43]. Alternatives to bulk CMOS devices, such as silicon on insulator (SOI), can significantly reduce the IPF caused by RDD, although such devices are still subject to LER, OTV, and RDD in the source and drain regions, and adoption of such new device structures is non-trivial due to: material quality issues (for instance, the uniformity of the silicon layer in fully depleted SOI and the quality of the back interface [54][55][56]); floating body effects not observed in bulk-CMOS (i.e. the 'kink' effect [57][58][59][60]); and self-heating effects due to thermal insulation of the active region of the transistors from the substrate, leading to increased device temperatures and altered *I-V* characteristics [61][62]. Hence, as long as bulk devices can still remain functional and scalable, information on the statistical variability caused by the IPF sources in circuit performances must be made available to support the design process, this will allow designers to deal with variability issues which will become critical in the design cycle in achieving optimal designs.

2.4 Impact of IPF in Digital Circuits

Moore's Law continues to drive the exponential increase in the number of transistors on a silicon die. However, due to restrictions in the scaling of supply voltage – required in order to retain sufficient circuit and system speed – power densities have begun to become prohibitive, resulting in complex design trade-offs between system power, speed, transistor budget and yield. IPFs have a significant contribution to the power crisis. The variations in threshold voltage and leakage current directly responsible for the increased margins in the power / speed / yield design trade-off. IPF have already started to affect the performance and yield of digital systems [43][63][7][64][19][79][65].

SRAMs in particular are strongly affected due to a small design margin, as they are designed to have the highest density possible, and typically use minimal width transistors. The presence of transistor variability and subsequent SRAM drive load and pass transistor mismatch, further reduces their functionality margin. Exotic memory cell designs have been proposed to cope with the variability, including the topology transition from 6T- to 8T- and 10T-SRAM cells [66][67][8], which of course come at the expense of larger area overhead. However, the efficiency of these new topologies still needs to be evaluated against simple 6T-SRAM device sizing strategies in coping with the variations present in the 45 nm generation and beyond.

Standard CMOS logic on the other hand, it is usually designed with larger transistor widths than SRAM, and therefore has better susceptibility to statistical variability. Even so, standard CMOS logic will also inevitably face problems in power / speed / yield trade-offs due to increasing device variability.

To address power dissipation issues while maintaining system speed, several approaches have been proposed. One approach is to compensate the use of low supply voltage with extreme pipelining architecture to maintain high throughput [68]. In this approach a long data path is shortened by breaking the logic into smaller data paths and flip flops are inserted between the pieces of logic. Shorter logic depth

and increased pipelining compensates for the increased gate delay resulting from lower power operation. However shorter logic depths automatically result in increased delay variations, as the number of gates in the logic path are reduced: it is known that the delay variation is inversely proportional to the square root of logic depth [69].

Another approach is to employ dynamic voltage and frequency scaling across the system: *in situ* circuitry is used to monitor the clock frequency requirement and the supply voltage is adjusted accordingly to conserve energy on-the-fly [70][71]. When employing this design approach, the design must be verified over a wide range of supply voltages and clock frequencies. This technique imposes several challenges in the performance verification process, because the current industrial standard cell format, the non-linear delay model (NLDM) has the following deficiencies : 1) It is not robust in evaluating the cell at various supply voltage values due to the usage of linear derating factor which is not valid at low supply voltage [72] (the derating factor is used to obtain delay values when the operating condition of the cell is out of its characterised conditions) and 2) It does not well capture the effects of changing supply voltages on device variability as will be demonstrated later in this thesis.

Both of these proposed approaches to control system speed and power are influenced by the variability, mandating that IPF must be taken into account in circuit or system optimisation, trading off between performance, power and yield. The new approaches to system design must take into account the increasing influence of IPF on performance, power dissipation and yield. *Therefore, the development of tools and methodologies to help designers to trade-off between timing, power and yield in the presence of acute statistical variability must become an integral part of the circuit and system design and verification.*

2.5 Statistical Circuit Design

In order to qualify for volume production, a circuit design must meet critical performance specifications. Exhaustive functional and performance verifications are performed at every design level to ensure correct design implementation before signing-off the design for tape-out.

The *de facto* methodology to determine the performance spread in the presence of process variability is to run multiple static timing analyses (STA) at different process conditions – known as corner analysis. In this approach, logic circuits are designed for functionality under worst-case and best-case conditions. However, at the 65 nm technology and beyond, where the variability has become an important issue, the ability to predict circuit performance under process variation has deteriorated. This is due to the increasing complexity of the semiconductor fabrication processes, extreme lithography, strain variation and the rising role of statistical variability.

In the timing verification, a design margin is usually allocated in the verification process to account for any unpredictable variation in the physically fabricated silicon. The design margin not only accounts for unpredictable variation arising in the manufacturing process but also for other components of uncertainty such as clock jitter, noise, *etc.* which are either unpredictable, or too complicated to predict at any particular point in the design process. These design margins increase in magnitude with each technology node due to the increasing number of sources of variability and their increasing magnitudes. The margins result in over-design and if not managed properly, leads to greater waste in the trade-off between silicon area, system speed, yield and power consumption.

In corner analysis, devices are assumed to have parameters that yield the worst circuit performance. Corner analysis guarantees good yield, but leads to pessimistic design, and statistical design has been proposed to enable further optimisation of a design before tape-out [74]. In statistical design, the circuit

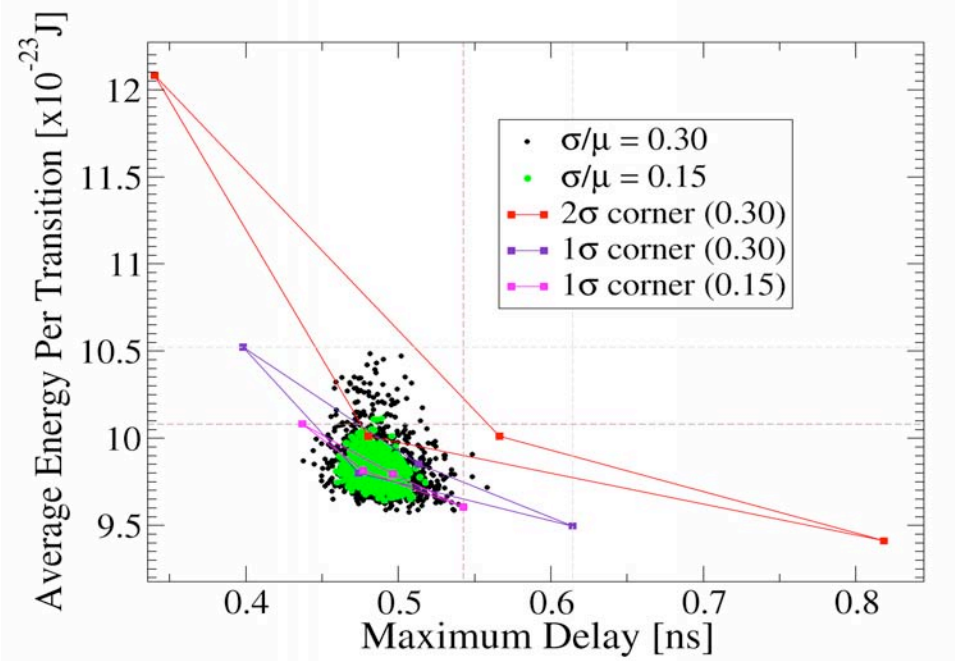


Figure 2-7 : Adder circuit simulation using 130 nm technology which shows the how pessimistic the corner analysis can be in comparison to the statistical analysis [73].

performance parameters which cannot be modelled deterministically are statistically modelled rather than being lumped into a design margin. Fig. 2-7 illustrates the disadvantage of corner analysis over statistical analysis in the presence of IPFs. A huge power/speed design margin between both analyses occurs due to the uncorrelated nature of the IPFs inherent in the transistors in the circuit. In a statistical design philosophy, circuit designers should be able to reach a more optimal design because information on the *distribution* of the performance of a circuit design is made available to them, whereas corner analysis only flags a pass / fail status for the circuit in fulfilling its specifications.

Although statistical design promises advantages, it is still immature and has clear limitations. Firstly, the characterisation of the global, local, systematic and random variation sources is time consuming and difficult in practice. Secondly, statistical design is computationally expensive because accurate performance distributions can only be found by running Monte Carlo simulations. Although techniques have been proposed to alleviate this high computational effort [75][76],

these techniques fail to accurately predict the tails of the performance distribution, which is critical in the correct estimation of design yield. It should be noted, for example, that SRAM designs typically require design to 6σ [77], and hence, correct estimation of the tail is necessary in obtaining an efficient functional design. For these reasons, there is still no a clear industry consensus regarding the direction of statistical design, and industry is loathe to incur the training and transitioning costs associated with a change in methodology until there is more clarity.

In this thesis, therefore, *we will evaluate different aspects of statistical simulation methodology which employs physical atomistic device simulation to account all the IPF due to intrinsic variations, up to statistical SPICE circuit simulation*. The methodology will be applied to circuits employing 35 nm, 25 nm, 18 nm and 13 nm gate length devices (equivalent to the 65 nm, 45 nm, 32 nm and 22 nm technology nodes).

In addition, past studies using the 'atomistic' compact models have mainly focused on static circuits analyses. *We will expand the scope of such studies to investigate the impact of variability on the transient performance of circuits – allowing us to obtain accurate speed and power dissipation data for simple circuit configurations*. To enable such studies, in Chapters 3 and 4, we will outline the proposed statistical methodology, and will evaluate the accuracy of the static simulation results. In Chapter 5 we will present the *I-V* and *C-V* BSIM4 compact model fitting results for the developed devices compared against 2-D TCAD simulation to ensure the accuracy of the dynamic behaviour of the devices. We apply the transient analysis methodology to foundational circuits, and discuss the results obtained in Chapter 6. Then in Chapter 7, we discuss the importance of our results to the present industry methods of capturing circuit timing data, the Non-Linear Delay and Current Source Models which are designed to capture the timing of standard cells.

Although the literature contains a number of studies which have investigated the effect of IPF in circuits, they: 1) neglect the correlations between device

parameters (e.g. off-current, threshold voltage, on-current) that occur due to the each specific source of IPF [78] 2) consider unrealistic, simplistic and outdated device structures [79][80] and, 3) ignore the 3-D nature of the device physics involved in correctly modelling the underlying variations [80][81]. Thus, we believe that our approach will produce more accurate and useful results than previous studies, allow separation of the various effects and their causes (due to the systematic nature of our approach), and have greater predictive power.

2.6 Summary

In this chapter, the purpose of device scaling and some of its major bottlenecks have been discussed. A classification of the major variability sources has been presented. Focusing on the statistical variability, description of random discrete dopants, line edge roughness and oxide thickness variation – which are the primary sources of the intrinsic parameter fluctuations – and their impact in degrading the speed and power requirements of CMOS circuits have been detailed. Next, the impact of scaling and IPF on digital logic domain was discussed and the importance of developing tools to help designers to perform the timing, power and variability trade-off analyses that are needed for good circuit and system designs was emphasised.

Chapter 3

Statistical Simulation Methodology

3.1 Introduction

A number of studies have been carried out at the University of Glasgow to quantify the effect of different sources of intrinsic parameter fluctuations (IPF) on device operation [5][30][82][4]. However, it would be computationally prohibitive to perform detailed, device level, physics-based simulations on any circuit larger than a single inverter. In order to carry out simulations to investigate the statistical properties of circuits and systems we will employ a hierarchy of simulation tools to make the problem more computationally tractable, and develop a methodology of statistical simulation that will be appropriate for circuit and system research, and is also applicable to industrial simulations.

Fig. 3-1 shows a schematic hierarchical flow diagram of the tools used in this research. The process starts with the development of the MOSFET structure using the commercial Sentaurus Process tool. This tool carries out physics-based process modelling which can accurately model semiconductor fabrication processes such as implantation, annealing and etc. Then, device characteristics (*I*-*V* and *C*-*V* curves) for an ideal, smoothly doped device are generated using the Sentaurus Device tool, which uses a finite element discretisation method to solve the semiconductor transport equations. After generating the uniform/ideal device characteristics, the

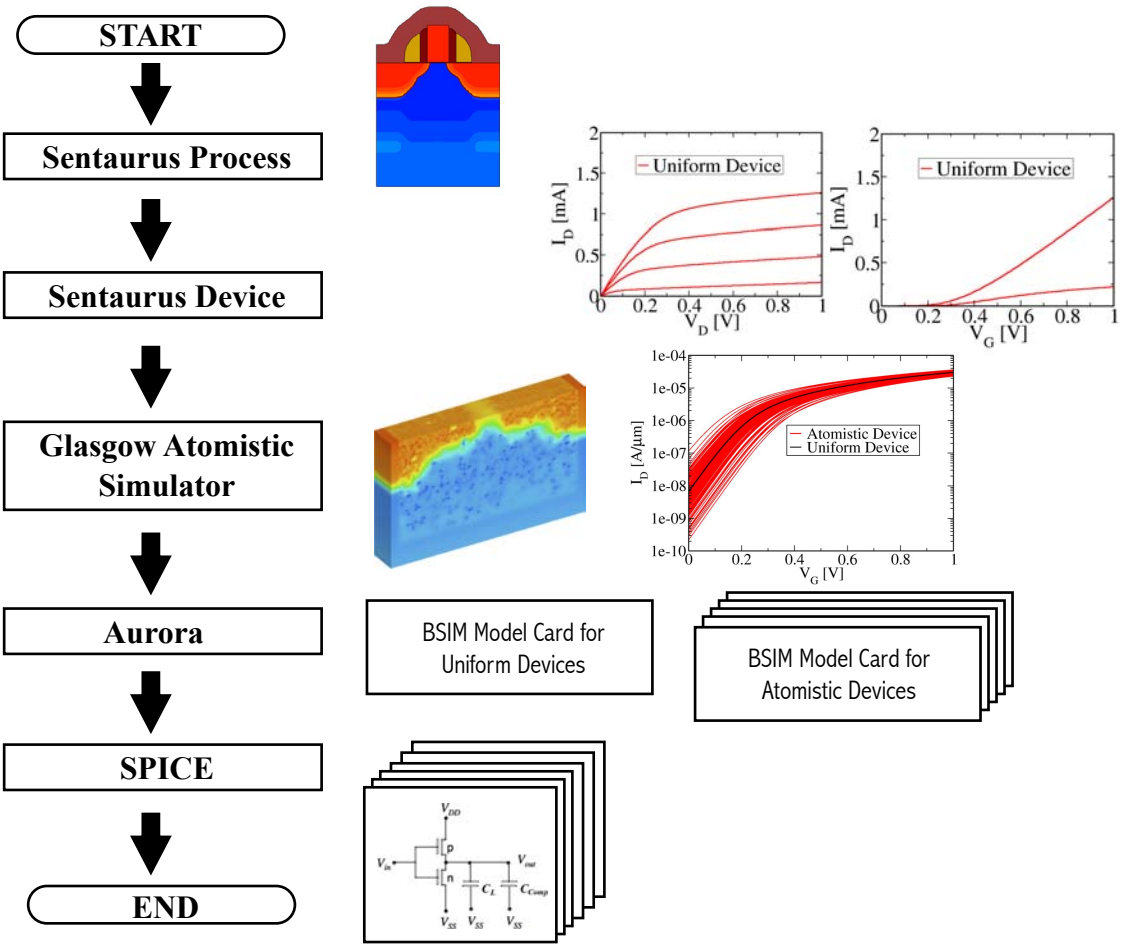


Figure 3-1 : Schematic flow diagram of tools used in this research. Figures beside of the flowchart are the products being supplied into the next tool chain to enable statistical circuit simulation studies.

developed doping profile is transferred into the Glasgow Atomistic Device simulator, a bespoke 3-D drift-diffusion based device simulator will be used to predictively simulate ensembles of MOSFETs subject to IPF. The simulator is calibrated to match the current-voltage characteristics obtained from Sentaurus Device. The result of these simulations will be *I-V* and *C-V* curves for each member of the ensemble. These *I-V* and *C-V* curves will contain the information needed to perform analysis of circuits employing the devices, and the ensemble of curves will contain the statistical information needed at the circuit level (assuming enough members of the ensemble are available to allow the appropriate statistical accuracy). Section 3.2 below describes in detail the devices used in this study, whilst section 3.3 describes the key properties of the 3-D device simulation tool.

In order to carry out statistical circuit simulations effectively, the *I-V* and *C-V* curves for each device must be translated into a compact model, to be used in SPICE. Aurora, a parameter extraction tool for semiconductor devices, is used to extract the parameters of the BSIM compact models. Statistical information on the electrical characteristics of the devices is then encapsulated in an ensemble of BSIM compact models, to be used in circuit simulation. Section 3.4 below describes the choice of compact model used, and the details of how these compact models are efficiently extracted from the large *I-V*, *C-V* dataset.

At the next level, that of circuit simulation, circuits will be investigated using SPICE (or equivalent) circuit simulation and a Monte Carlo technique – a series of circuit simulations will be carried out with the devices in each nominal circuit replaced by random members of the device ensemble. This set of simulations will give the detailed distributions of any circuit parameters of interest, with the accuracy of the distributions dependent on the number of repeat simulations of a given nominal circuit, and the size of the device ensemble.

A limitation of this Monte Carlo technique is the sample size of 'atomistic' compact models that can be generated, a number limited by the foundational device simulations which are the most computationally burdensome part of the procedure. In order to reduce the computational effort in generating a large number of the compact models, statistical enhancement techniques can be applied as reported in [83][84]. Such statistical enhancement techniques, although possible, were not required for the results shown in later chapters.

3.2 MOSFET Devices Under Study

Two template devices are considered in this work. One is based on a research device, fabricated and reported by Toshiba in 2001 [85] which represents the 65 nm technology node and the other is a device design developed at the University of Glasgow which closely matches recently published state-of-the-art 45 nm technology

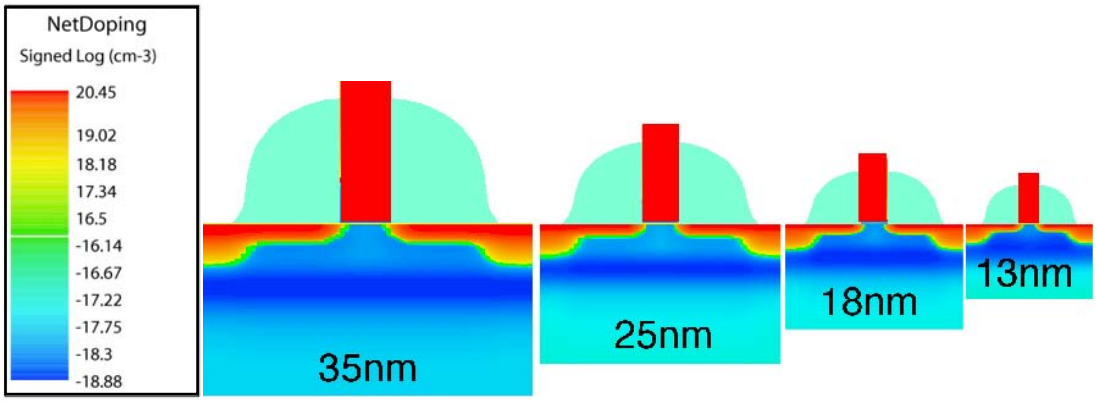


Figure 3-2 : Cross-section of the scaled conventional devices from a template of Toshiba device with 35 nm gate length, taken from [177].

generation counterparts [36]. Both template devices have a metallurgical channel length of 35 nm, and are used as the starting points to predict the physical scaling of smaller gate length devices. The Toshiba device template and its scaled devices are used in the investigations performed in Chapters 4 and 6. The University of Glasgow designed devices incorporate strain induced mobility enhancement and updated values of the oxide thickness to match the 2007 ITRS roadmap and state-of-art industrial devices. They are used in the characterisation studies performed in Chapter 5 and 7.

Fig. 3-2 shows the cross-section of the reference Toshiba template MOSFET and its scaled versions used in this study. The cross-section shows the doping profile of the device. It has a complex doping profile featuring retrograde In channel doping (shown in light blue/turquoise colour), As source/drain and Si-gate doping (shown in red), and source/drain pockets which are heavily doped with Boron (shown in dark blue colour) to reduce short-channel effects.

Generalised scaling rules are used to obtain the structural and doping parameters for the scaled devices of Fig. 3-2, closely following the prescription of the 2005 ITRS in terms of equivalent oxide thickness (EOT), junction depth x_j, doping and power-supply voltage V_{DD}. As can be seen in Fig. 3-2, the channel-doping concentration at the interface increases while the source/drain doping concentration is remains constant as device dimensions reduce – it is already close

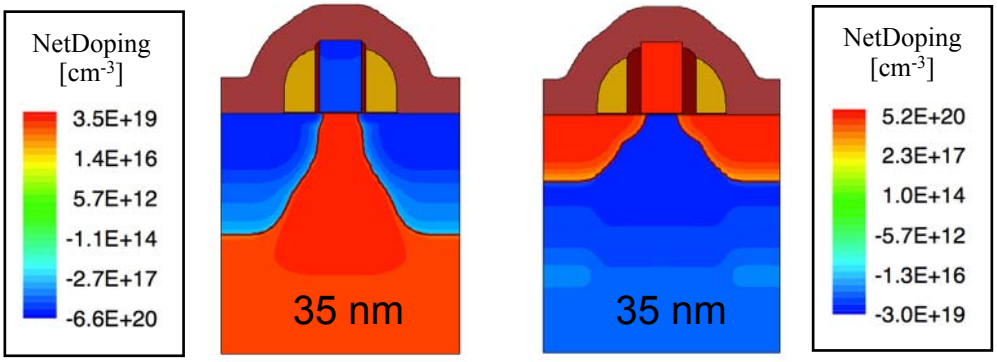


Figure 3-3 : Cross-section of the *p*-MOSFET (left) and *n*-MOSFET (right) device doping profiles simulated using Sentaurus to model a standard modern process flow. These devices are enhanced with strain engineering to match the performance of 45 nm technology generation counterparts [88].

to the solid solubility limit [87]. A full description on the scaling and calibration processes of the device can be obtained in [177].

Fig. 3-3 illustrates the cross-section of the 35 nm gate length *p*- and *n*-channel transistors developed using TCAD process simulation, carefully calibrated to published data [36]. A cap (contact etch stop) layer is deposited on the source/drain in order to introduce strain into the channel region. Tensile nitride capping introduces tension into the *n*-channel MOSFETs, and a compressive nitride contact-etch-stop layer and SiGe source/drain areas are used to introduce compressive stress in *p*-channel MOSFETs. The effect of these cap layers increases carrier mobility in the devices. A more detailed description of these device structures and the device processing used to create them is described elsewhere in [88]. These efforts in developing realistic device structures and the careful calibration of device designs to published electrical results give us confidence that the variability information extracted from simulations of these devices will be relevant and useful.

3.3 The Glasgow 'Atomistic' Device Simulator

In this section, the Glasgow ‘atomistic’ device simulator will be briefly described. There are numerous techniques that can be employed to study the characteristics of modern semiconductor devices, including: full quantum transport,

Monte-Carlo device simulation (where here the Monte Carlo approach is used in the analysis of charge transport within a device) and drift-diffusion [89]. Each differs in the implemented physical models in the simulation, trading off computational effort against the ability of the simulator to accurately predict all the properties of future generations of highly scaled devices.

A 3-D drift-diffusion simulator, which has been developed over a number of years at the University of Glasgow, is used in this study [4]. The drift-diffusion simulator self-consistently solves the Poisson and current-continuity equations to obtain the terminal currents at any applied bias. This technique assumes that transport is in local equilibrium with the applied field, and hence well captures device electrostatics. It can reliably predict sub-threshold current in deca- and nano-meter scale devices since the main mechanism of charge transport in this regime is through diffusion and the corresponding injection is exponentially sensitive to the potential distribution. However, the effect of non-equilibrium carrier transport is not well captured by the drift-diffusion approach and the on-current magnitude and its variability are underestimated. Therefore, results in this study which rely primarily on the magnitude of the saturation current in MOS devices should be considered ‘best case’ results, with realistic variability almost certainly higher. For example, it is well-known, that the drift-diffusion underestimates the drain current variability above threshold by about 45% [90] thus, the statistical simulation performed using the extracted compact models will also underestimate the drain current variation above threshold. A full 3-D Monte-Carlo device modelling treatment is necessary to correctly estimate on-current variability where the scattering rate of the particles can be taken into account, or a hybrid technique such as that described in [91] where the Monte Carlo device modelling simulator continually updates the mobility estimates used in the drift-diffusion simulator. The Glasgow drift-diffusion simulator employs density gradient (DG) quantum corrections [92] for both electrons and holes to account for the quantisation effects which causes the peak of the charge distribution

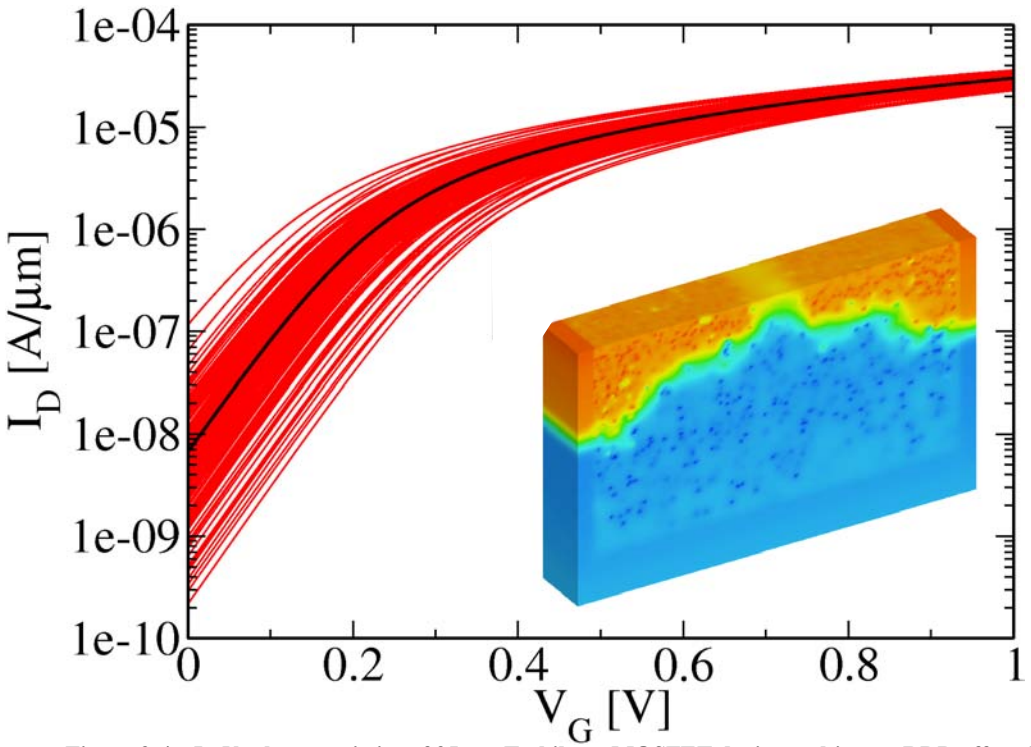


Figure 3-4 : I_D-V_G characteristics of 35 nm Toshiba *n*-MOSFET devices subject to RDD effect (shown in red lines). Black line shows the I_D-V_G characteristic of the uniform device. Inset showing 3-D 'atomistic' potential profile of the Toshiba 35 nm MOSFET. Potential varying in the channel and source/drain region which indicates the presence of dopants. Taken from [41].

in the channel to shift away from the interface due to the steep potential well in reduced channel length devices.

The 3-D Glasgow device simulator has been used to simulate the effects of random discrete dopants (RDD), line edge roughness (LER) and oxide thickness variations (OTV) which are the identified sources of intrinsic parameter fluctuations described in Chapter 2. Random discrete dopant effects are included based on a continuous doping profile of the reference and scaled devices described above. Based on this profile, dopants are introduced randomly using a rejection technique [93]. LER is introduced by using one-dimensional Fourier-synthesis, generating random gate edges from a power spectrum corresponding to a Gaussian autocorrelation function [106]. The oxide thickness variation effect is simulated by using Fourier synthesis to generate a random 2-D surface from a power spectrum corresponding to an exponential autocorrelation function [94][95]. Full implementation of the simulation of sources of variation is described elsewhere in

[41]. Typical results obtained from the simulator are shown in Fig. 3-4, which graphs the I_D-V_G characteristics of an ensemble of 200, Toshiba 35 nm gate length *n*-MOSFETs simulated in the presence of random discrete dopants at high drain bias, V_{DS} = 1 V. Variations in off-current, on-current and threshold voltage across the ensemble are clear, and distributions of these, and other parameters of interest can be obtained from the simulation data. These ensembles of realistic *I-V* curves are key in developing statistical compact models and thus in performing statistical circuit simulation. Inset of Fig. 3-4 shows the potential distribution of the 35 nm Toshiba device simulated using the 3-D Glasgow 'atomistic' simulator where the potential is shown non-uniform (by the colour contrast) in the presence of dopants in the channel and source/drain regions. These dopants cause the electrostatic and transport behaviour of an ensemble of macroscopically identical devices to differ in its characteristics when subject to different number and position of the dopants in the devices.

3.4 Statistical Circuit Simulation

3.4.1 'Atomistic' Compact Models

In this study, the BSIM4 was selected as the compact model of choice. It is widely used, and familiar to circuit designers, having served as an industrial standard since its introduction in 1997. It is actively updated, and has a flexible model parameter extraction flow, making it efficiently to work with. Although the BSIM compact model is able to replicate the current-voltage and capacitance-voltage characteristics of nominal bulk-MOSFET devices accurately, no compact model is able to replicate the effect of IPF accurately in its formulation, due to the complexity of IPF and because IPF was never considered as part of the physical underpinnings of any extant compact model family. However, we have discovered that the flexibility of the BSIM model makes it possible to capture the effects of IPF

accurately by making use of compact model parameters originally aimed at other effects.

There are two strategies that can be adopted during a parameter extraction process: global and local optimisation. In global optimisation, the optimisation algorithm finds one set of model parameters which best fit the available measured data. In local optimisation, parameters are extracted independently of one another. The generation of our ensemble of 'atomistic' compact models is performed in two stages using a combination of global and local strategies with the commercial Aurora tool. At the first stage, extraction of a complete set of BSIM model parameters over the complete operating range of a nominal, continuously doped device is performed. At the second stage, parameter extraction is done for each member of the an ensemble of microscopically different devices. However, at this stage only a few selected BSIM model parameters are chosen and re-extracted for each device in the ensemble. This small subset of the BSIM model parameters represent the effect of the sources of variation.

The choice of the model parameters used in the second stage of the extraction procedure, depends on the sources of variability being investigated, their physical effect on the *I-V* curves of the devices being studied, the precise parameters available in the compact model employed, and the required accuracy of the resultant ensemble of compact models. Fig. 3-4, displays the I_D-V_G characteristics of an ensemble of 200, Toshiba 35 nm gate length *n*-MOSFETs at high drain bias. It can be seen that compact model parameters relevant to device off-current, subthreshold slope, threshold voltage and on-current would be of most use in capturing the effect of atomistic variability on these devices.

From our knowledge of the BSIM model, and the nature of the variations shown in Fig. 3-4, we choose seven parameters to fully capture the effects of RDD. These parameters are:

a) *dsub* - DIBL coefficient exponent in subthreshold, which is used to account for DIBL variations.

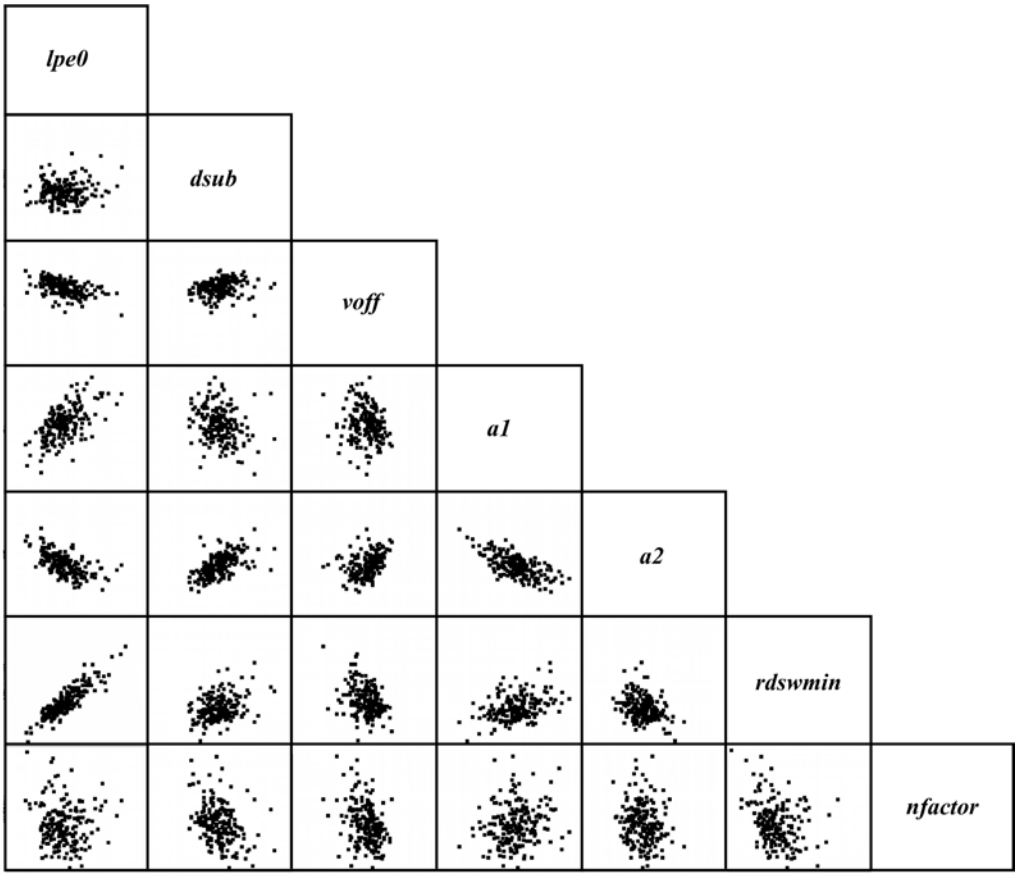


Figure 3-5 : Scatter plots between two mapped parameters.

b) *a1* - First non-saturation effect factor, a mobility parameter which helps to capture current variations.

c) *a2* - Second non-saturation effect factor, also a mobility parameter which helps to capture the current variations.

d) *rdswmin* - Resistance per unit width at high V_{GS} and zero V_{BS}, which accounts for resistance variations in the channel affecting the current variations.

e) *nfactor* - Subthreshold swing factor which is used to account for the subthreshold slope variations.

f) *voff* - Offset voltage in the subthreshold regime which is used to account for the subthreshold slope variations.

g) *lpe0* - Lateral non-uniform doping parameter at V_{BS}=0, which is used to account for the threshold voltage variations.

At this second stage of parameter extraction, no prior assumptions about the parameter distributions nor the correlations between parameters are made. A direct parameter extraction procedure is used and the statistical compact model parameters are obtained by fitting the *I-V* curves against the atomistic simulation results using the 7 parameters described above in the uniform/ideal device's compact model. As a result, the extracted compact models accurately encapsulate the IPF introduced by the RDD simulated in the 3-D 'atomistic' device simulator with mean RMS error of 1.16% [83]. Fig. 3-5 shows the scatter plots of the extracted 7 parameters for 200 devices subjected to IPF. Some of the mapped parameters have a strong correlation with the other parameters (shown by the increasing/decreasing pattern of the plotted points in the Cartesian axes). These correlations should be preserved at the statistical compact model generation in order to maintain the correct behaviour of the device operation in circuit simulation.

3.4.2 Wider-Sized Transistor Model

In a circuit simulation, the transistor width may vary from a minimum-size to any arbitrary number to suit the needs of circuit designers. However, the single device extraction strategy employed in generating the 'atomistic' compact model is based on simulation of square, minimal sized devices. Simulation of wider devices in a circuit by naively changing the width parameter of an extracted compact model will not reproduce the true effects of the IPF distribution of a larger size transistor.

To overcome this limitation, simulation of a wider sized device is performed by slicing the wider gate into a number of square devices as shown by the fine black lines in Fig. 3-6 (*left*). Fig. 3-6 (*left*) shows a simplified layout of an inverter, while Fig. 3-6 (*right*) shows the corresponding schematic diagram of the CMOS inverter. The square-sized transistors are connected in parallel to form the wider sized transistor. Each square, minimal sized transistor is correctly simulated using the

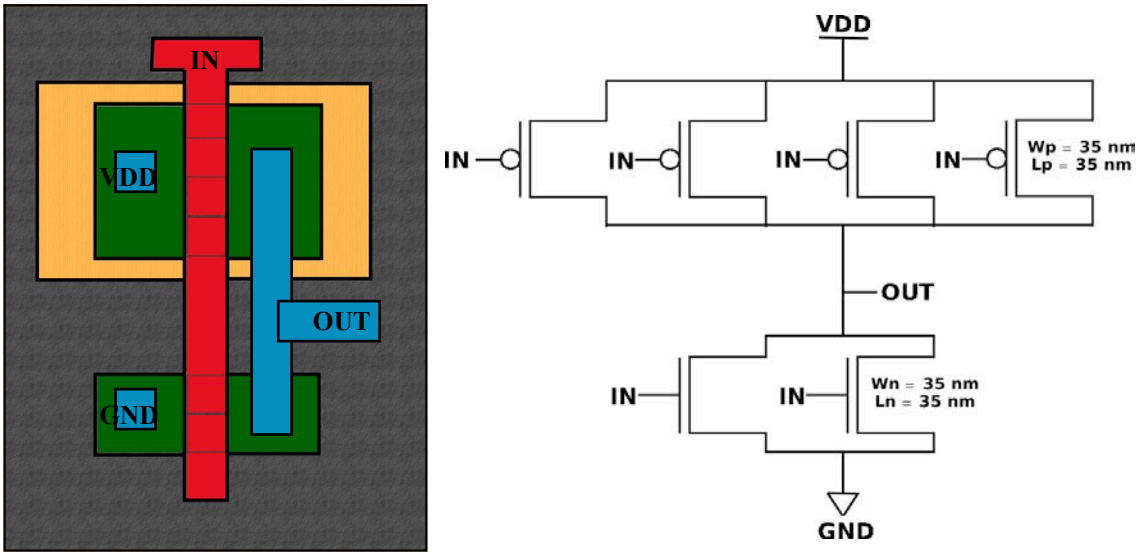


Figure 3-6 : A simplified layout of an inverter (left). Its corresponding representation in schematic diagram of the inverter (right) .

characterised 'atomistic' compact models. As discussed in Chapter 2, IPF has no spatial correlation in a circuit layout and can randomly occur in neighbouring transistors, so each square, minimal sized transistor is randomly chosen from the ensemble of 'atomistic' compact models. It should be noted that even though this approach is accurate in capturing the effect of fluctuations in circuit level, it increases the size of the circuit under test, as wide transistors are substituted for a series of parallel minimal sized transistors. The technique therefore has limitations due to the maximum number of components that SPICE can simulate, and the increased memory and data storage footprint of the larger circuit.

We can also observe from Fig. 3-6 (*b*) that by adopting this approach, a wider-sized transistor can only be in the form of an integer number of the minimum gate length size of the device. However, as stated earlier, a transistor's width can vary including fractional values of the minimal transistor width. In order to eliminate this limitation, another approach is to generate a set of 'atomistic' compact model equipped with width-dependence model which will require an approximation for all the distribution of the selected parameters which are sensitive to the channel width. However, this technique will require more careful analysis of the 3-D

physical simulation result and certainly require more fitting procedures to generate such compact models. This approach is not covered in this study.

3.5 Summary

In this chapter, the statistical circuit simulation methodology adopted in this study was described, including: the 35 nm physical gate length devices and simulation tools calibrated and used to provide foundational, predictive device parameters for the tool-chain and the BSIM compact models employed. The template devices are based on state-of-the art 35 nm gate length MOSFET with electrical characteristics that have been calibrated against published data [36][85]. The scaling includes strain-engineered devices and follows the ITRS prescriptions. Using this approach based on calibrated device, gives confidence that the statistical data obtained from the Glasgow Atomistic Device Simulator closely reproduce the actual statistical data of the prototyped devices. The scaled set of transistors were the closest devices that could be publicly used by the group based on close relationship with industrial/research partners which reflect currently manufactured devices in the semiconductor industries and the predicted future-scaled devices beyond the year 2007 - when this research began. Several devices have been used previously in the literature which were unrealistic in terms of their doping profile and structure; and obsolete in terms of technology nodes [79][80]. This has resulted in results that are significantly more realistic than any other work in the field.

The key properties of the 3-D Glasgow Atomistic Simulator also have been discussed, including use of density gradient quantum corrections [92], an essential feature in predicting the correct behaviour of decananometer MOSFETs where quantum effects start to play important role. This simulator captures well the subthreshold regime and threshold voltage of the simulated transistors but underestimates the on current and its variation [90]. This is because the drift-diffusion method cannot capture non-equilibrium transport effects. The Monte Carlo method is needed in order to capture the real transport behaviour in the

decananometer scale transistors. However, simulation of one semiconductor device in order to obtain one current-voltage point takes approximately 2 weeks of simulation time and it is computationally prohibitive for statistical variability studies. There are several device modelling groups which are developing Monte Carlo simulation methods [198][199] but none has successfully applied it for statistical variability studies. At the University of Glasgow some progress have been made in using Monte Carlo simulation for statistical variability studies [91][200] [201] however it is still immature for large scale production simulations. Whilst the augmented drift-diffusion technique we employ does not capture the on-current as well as full Monte Carlo simulation, it is the most accurate and practical technique presently published in the literature.

Next, generation of BSIM 'atomistic' compact models was carried out using a 2-stage extraction strategy where in the first stage, a full set of BSIM parameters are extracted based on the uniform device characteristics. In the second stage, 7 parameters are chosen to encapsulate the variation in the electrical characteristics observed in the microscopically different devices subject to statistical variability. In the literature, several attempts have been made to study the impact of statistical variability on circuits by varying parameters in the compact model. However, the approaches are either making an assumption that the distribution of a chosen parameter, e.g. threshold voltage, is Gaussian [142][143][144] or neglect correlations between the chosen device parameters to reflect the underlying physics of statistical variability [78]. Therefore, our approach produces more accurate and predictive result for the aimed technology node as each of the compact model is fitted to 3-D device simulation result subject to statistical variability.

Lastly, the statistical circuit simulation employed in this study has been described. An ensemble of compact models which are macroscopically identical but microscopically different are randomly chosen to be used for the individual transistor instances in circuit. A practical difficulty with this approach, the generation of wider-sized transistors was discussed and a solution is described.

Having the capability to run circuit simulations with the generated model cards, this work enables the transition to a higher level of abstraction which is the characterisation of statistical standard cells. Whilst there are more mature system analysis tools reported in the literature to analyse systems subject to device variability from IMEC [202] the results of this work presently provide the only practical systems analysis methodology to give device accuracy of better than 2% accuracy.

Chapter 4

Hard Logic Fault Related Supply Voltage Limitations Due To Statistical MOSFET Variability

4.1 Introduction

As described in the introductory chapters, statistical variability, introduced by the discreteness of charge and granularity of matter, has become a major concern associated with CMOS transistors scaling and integration [96][97]. It already critically affects SRAM scaling [79][98], and introduces leakage and timing issues in digital logic circuits [99][100][101].

Variability is the main factor restricting the scaling of the supply voltage, which for the last three technology generations has remained constant, adding to the looming power crisis [102][103]. It is very important to understand properly how variability will affect the scaling of the supply voltage in future technology generations, and this is the problem which will be the subject of investigation in this chapter.

Several attempts [104][105] have been made to predict the limitations of supply voltage scaling due to variability. Most of these are based on simple analytical models of the nature of the dominant source of variability in bulk MOSFETs – threshold voltage variability introduced by random discrete dopants (RDD) [4]. However, comprehensive numerical simulations have shown that in addition to

being over-simplistic, these simple models significantly underestimate the RDD induced variability of modern decananometer scale CMOS transistors [92] and therefore result in over-optimistic predictions for the limits of supply voltage scaling. It has also became clear that other sources of variability, among which are line edge roughness (LER) [106] and poly silicon granularity (PSG) [52], may become as important, or more important, than RDD as devices continue to scale [52] [5].

Results from recent and comprehensive, statistical 3D simulations for the statistical variability in bulk CMOS devices [52][5] can be used to study the hard limitations that variability imposes on the supply voltage of future technology generations. The most serious limitations are those which bound the logical failure (non-switching) of the most robust digital circuit component, the CMOS inverter. The analyses of this chapter deal with the conditions under which CMOS inverters fail, and thus define the limits of any digital logic. Our predictions are based on an analytical model for inverter variability which is carefully tested and validated with respect to statistical circuit simulations.

In section 4.2 the analytical model for the statistical variability of an inverter, based on a simple but accurate expression for the current in decananometer MOSFETs is presented. In section 4.3 the analytical model is validated to statistical SPICE simulations, based on statistical compact models extracted from comprehensive 3D physical simulations of variability. The predictions for the hard logic fault limitations on the supply voltage are presented in section 4.4.

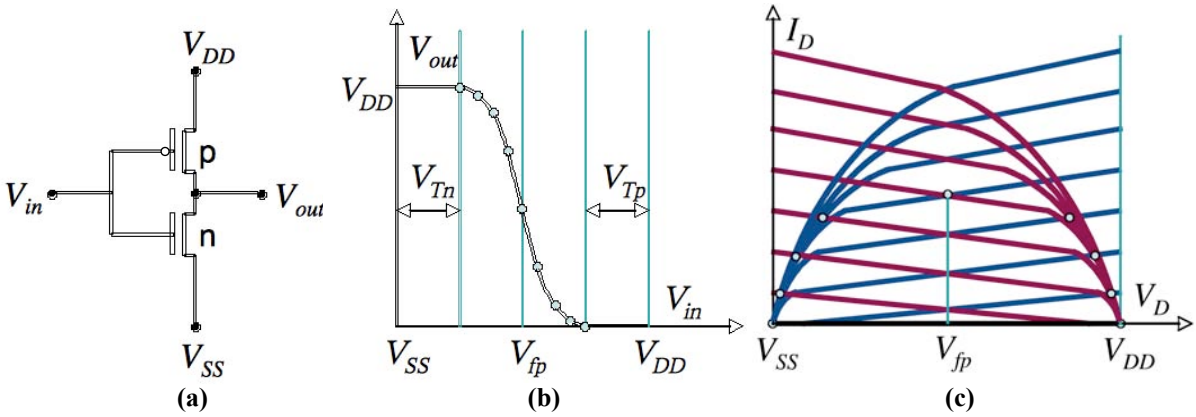


Figure 4-1 : CMOS inverter. (a) Schematics; (b) Transfer characteristics; (c) Definition of the transfer characteristics.

4.2 Inverter Variability Model

The transfer characteristic of a CMOS inverter, illustrated in Fig. 4-1 (*b*), is defined as a solution of the equation $I_{D,n}(V_{in}, V_{out}) = I_{D,p}(V_{DD} - V_{in}, V_{DD} - V_{out})$, where $I_{D,n}(V_G,V_D)$ and $I_{D,p}(V_G,V_D)$ are the currents flowing through the *n*-channel and *p*-channel MOSFETs respectively, where V_G and V_D are the gate and drain voltages of the MOSFETs and V_T, (in Fig. 4-1 (*b*)) the threshold voltage. These conditions are illustrated in Fig. 4-1 (*c*) in which the output characteristics of the two transistors are superimposed. The flip voltage of the inverter V_{fp} is defined as the value of the input voltage V_{in} at which the output voltage is equal to one half of the supply voltage $V_{out} = V_{DD}/2$. In a well-balanced inverter at $V_{in} = V_{fp}$ the two transistors are in saturation and therefore V_{fp} can be determined by equating their saturation currents as shown in Eqn. 4-1 under the approximation that the saturation current $I_{Dsat}(V_G)$ depends only on the gate voltage.

$$I_{Dsat,n}(V_{fp}) = I_{Dsat,p}(V_{DD} - V_{fp}) \tag{4-1}$$

The MOSFET current in saturation can be approximated by the product of the channel width W, and the sheet carrier charge density Q and average carrier velocity v_{av} at the source end of the channel $I_{Dsat} = Wv_{av}Q$. The sheet charge density at the source is given by $Q = C_{ox}(V_G - V_T)$, where C_{ox} is the effective gate capacitance. In decananometer MOSFETs the average velocity at the source is given by the product of the injection velocity v_{in} and the ballisticity factor B, $v_{av} = v_{in}B$

[107]. This results in the following expressions for the saturation currents of the *n*- and the *p*-channel transistors in the inverter at flip voltage conditions

$$I_{Dsat,n} = W_n v_{in,n} B_n C_{ox}(V_{fp} - V_{T,n})/L \tag{4-2}$$

$$I_{Dsat,p} = W_p v_{in,p} B_p C_{ox}(V_{DD} - V_{fp} - V_{T,p})/L \tag{4-3}$$

Substituting Eqn. 4-2 and 4-3 into Eqn. 4-1 and solving in respect of V_{fp} where

$$k_{np} = \frac{W_n v_{in}^n B_n}{W_p v_{in}^p B_p} \tag{4-4}$$

$$V_{fp} = \frac{\left(V_{DD} + k_{np} V_{Tn} - V_{Tp}\right)}{1 + k_{np}} \tag{4-5}$$

In a well balanced inverter $k_{np} = 1$, $V_{Tn} = V_{Tp}$, and $V_{fp} = V_{DD}/2$. Thus,

$$\sigma V_{fp}^{V_{Tn}} = \frac{\partial V_{fp}}{\partial V_{Tn}} \sigma V_{Tn} = \frac{k_{np} \sigma V_{Tn}}{1 + k_{np}} \tag{4-6}$$

$$\sigma V_{fp}^{V_{Tp}} = \frac{\partial V_{fp}}{\partial V_{Tp}} \sigma V_{Tp} = \frac{\sigma V_{Tp}}{1 + k_{np}} \tag{4-7}$$

where σV_{Tn}, σV_{Tp} are the standard deviations of the threshold voltages of the *n*- and *p*-channel MOSFETs respectively. Since intrinsic parameter fluctuations are purely random and uncorrelated, it is reasonable to assume that there is no correlation between the *n*- and *p*-channel MOSFETs intrinsic threshold voltage variations. This assumption gives the following expression for the standard deviation of the inverter transition point

$$\sigma V_{fp} = \frac{\sqrt{k_{np}{}^2 \sigma V_{Tn}^2 + \sigma V_{Tp}^2}}{1 + k_{np}} \tag{4-8}$$

Eqn. 4-8 indicates that in a well balanced inverter *(k_{np} = 1)* the standard deviation of the flip voltage is determined only by the standard deviations of the threshold voltages of the *n*- and *p*-channel MOSFETs and does not depend on the detailed shape of the current voltage characteristics.

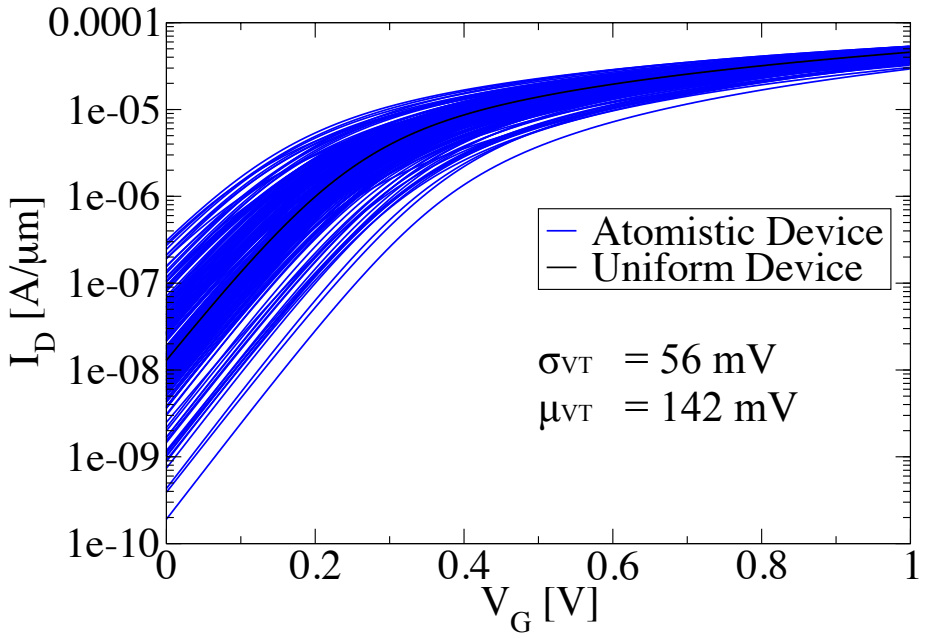


Figure 4-2 : Current-voltage characteristics of the simulated 200 microscopically different 18 nm *n*-channel MOSFETs with W_n=L_n at V_D=1 V.

4.3 Validation

Validation of the prediction of Eqn. 4-8 is made using the standard deviations of the flip voltage obtained from statistical Monte Carlo Spice circuit simulations of inverters constructed from members of scaled device ensembles with gate lengths 35 nm, 25 nm, 18 nm and 13 nm. The transistors are scaled versions of a prototype 35 nm MOSFET developed and published by Toshiba [85], against which TCAD process and device simulations are meticulously calibrated [108]. The scaling, which is described in detail elsewhere [5], is based on the guidance of the 2005 edition of the International Technology Roadmap for Semiconductors [ITRS] for high performance devices. Key design parameters of the scaled devices are summarised in Table 4-1.

TABLE 4-1

Key design parameters of the scaled devices.

Channel length [nm]	35	25	18	13	9
Equivalent Oxide Thickness [nm]	0.88	0.65	0.5	0.43	0.35
Junction depth, x_j [nm]	20	13	9	8	6

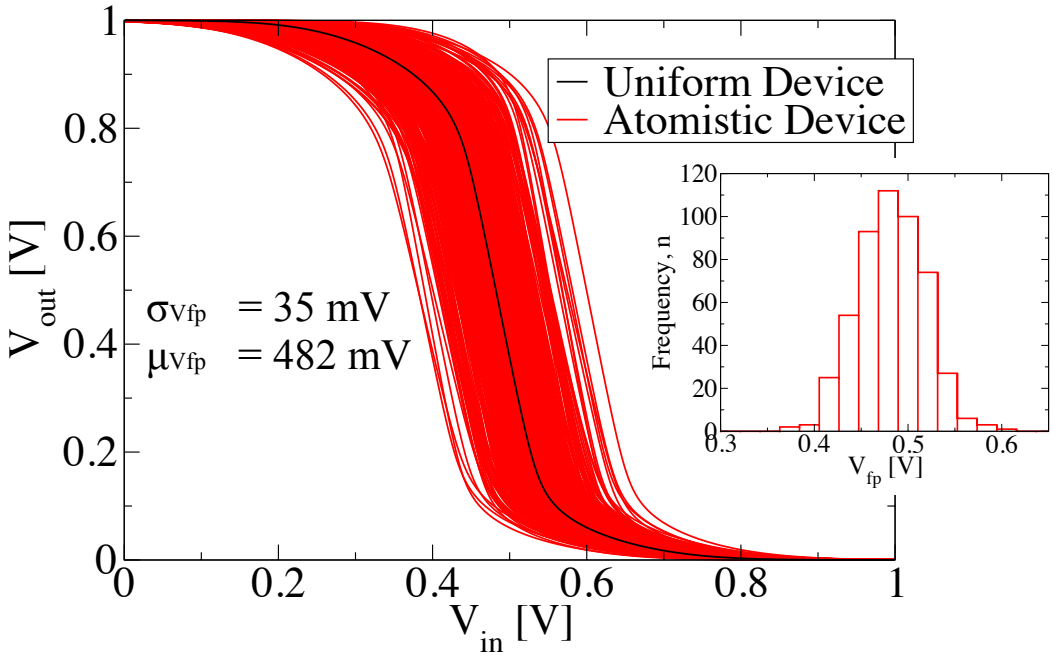


Figure 4-3 : Transfer characteristics of 500 statistically different minimal size inverters built with random occurrences of 18 nm *n*- and *p*-channel MOSFETs randomly selected from statistical samples of 200 microscopically different transistors with characteristics illustrated in Fig. 4-2. Inset showing the distribution of the flip voltage, V_{fp} extracted from the transfer characteristics for 18 nm devices.

Recent trends in physical gate length scaling have deviated from 2005 ITRS predictions and therefore the reader must match the physical gate length of the simulated transistors to the changing physical gate length targets in forthcoming technology generations. Also, oxide thickness predictions were updated in more recent ITRS editions. Discussion on this updated information on oxide thickness that will affect the results presented in this paper is also presented in later sections.

In the validation, statistical variability introduced only by RDD is considered. At each channel length, samples of 200 MOSFETs with microscopically different random dopant distributions were simulated with Glasgow 3D 'atomistic device simulator employing density gradient quantum corrections for electrons and holes simultaneously. The standard deviation of the threshold voltage was extracted for each of the channel lengths following the procedures described in [4]. Fig. 4-2 illustrates the 200 simulated current voltage characteristics of the 18 nm *n*-channel MOSFET ensemble.

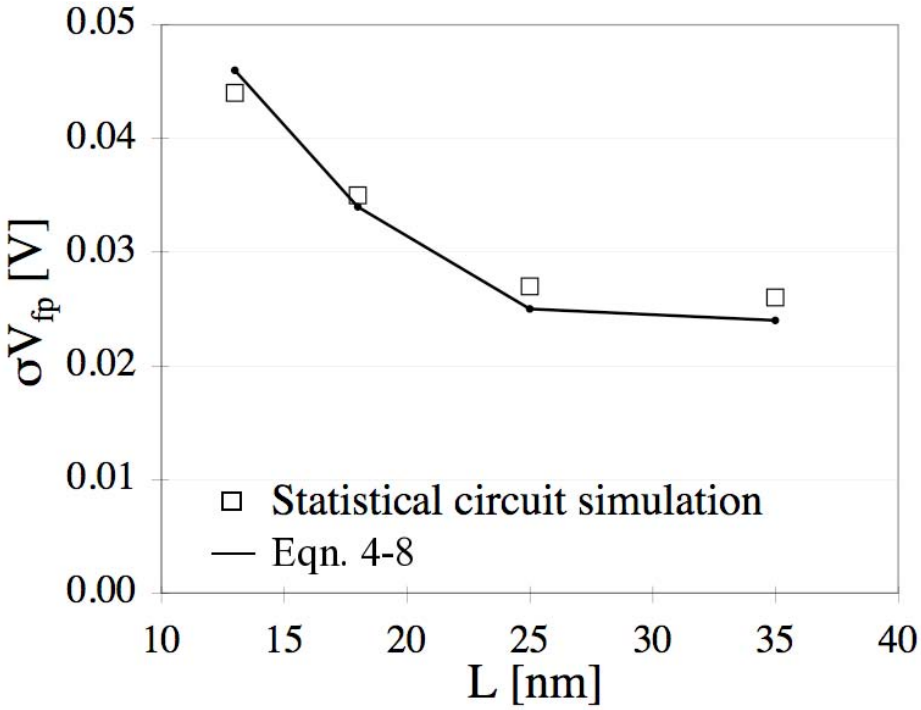


Figure 4-4 : Standard deviation of the flip voltage σV_{fp} extracted from the statistical simulation of inverters build from transistors with the different channel lengths, and the predictions of Eqn. 4-8.

Statistical sets of compact models are extracted from the simulated current voltage characteristics of each individual microscopically different transistor following the methodology described in Chapter 3. Statistical SPICE simulations of minimum size $(W_n = L_n)$ well balanced $(W_p = 2W_n)$ inverters were carried out for each channel length. Fig. 4-3 illustrates the static transfer characteristics of 500 statistically different, minimal size inverters built with random occurrences of 18 nm *n*- and *p*-channel MOSFETs selected from statistical samples of 200 microscopically different transistors with the characteristics illustrated in Fig. 4-2.

The standard deviation of the flip voltage σV_{fp} is extracted from the statistical inverter ensemble and compared, in Fig. 4-4, with the predictions of Eqn. 4-8, where σV_{Tn}, σV_{Tp} are obtained directly from the current–voltage characteristics obtained from each MOSFET ensemble. Excellent agreement is observed between the results from the statistical circuit simulation and Eqn. 4-8. This increase the confidence to use Eqn. 4-8 in order to make predictions for σV_{fp} based only on the statistical simulation results for σV_T without simulating the full current voltage characteristics

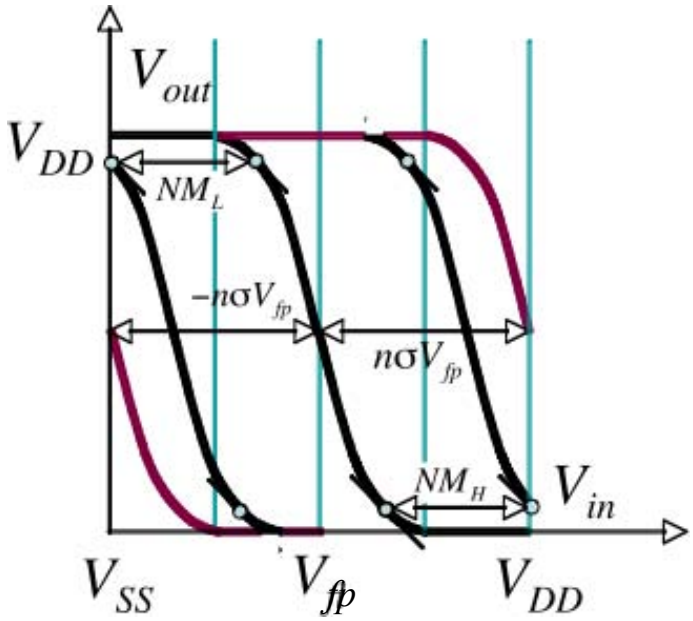


Figure 4-5 : Relationships between the design margin, the additional noise margin and the supply voltage.

of the devices from the statistical sample; extracting statistical equivalent circuit models; and performing statistical circuit simulation using these models.

4.4 Supply Voltage Scaling Limitations

As illustrated in Fig. 4-5, an integrated circuit must fail, as a result of hard digital fault, if a rare $n\sigma$ occurrence of the V_{fp} becomes equal to the supply voltage or to zero. The supply voltage limitation associated with the $n\sigma$ design margin (where n is a parameter chosen by circuit designer to fulfil a design specification) is $V_{DD,min} = 2n\sigma V_{fp}$ when the mean V_{fp} is $V_{DD}/2$. The allowable σ is normally constricted by an additional safety margin SM, defined as the V_{IN} between the high (NM_H) and the low noise margin (NM_L) points (at derivative of -1 of the transfer curve) where V_{IN} which falls within this region will result in undetermined output. In this case $V_{DD,min} = 2n\sigma V_{fp} + SM/2$. From the SPICE simulation of inverters constructed of transistors with continuous doping profiles, estimation is made on additional safety margin, which for all channel length devices is approximately equal to 0.17 V. Most of the results for the supply voltage limitations presented in this section do not include this

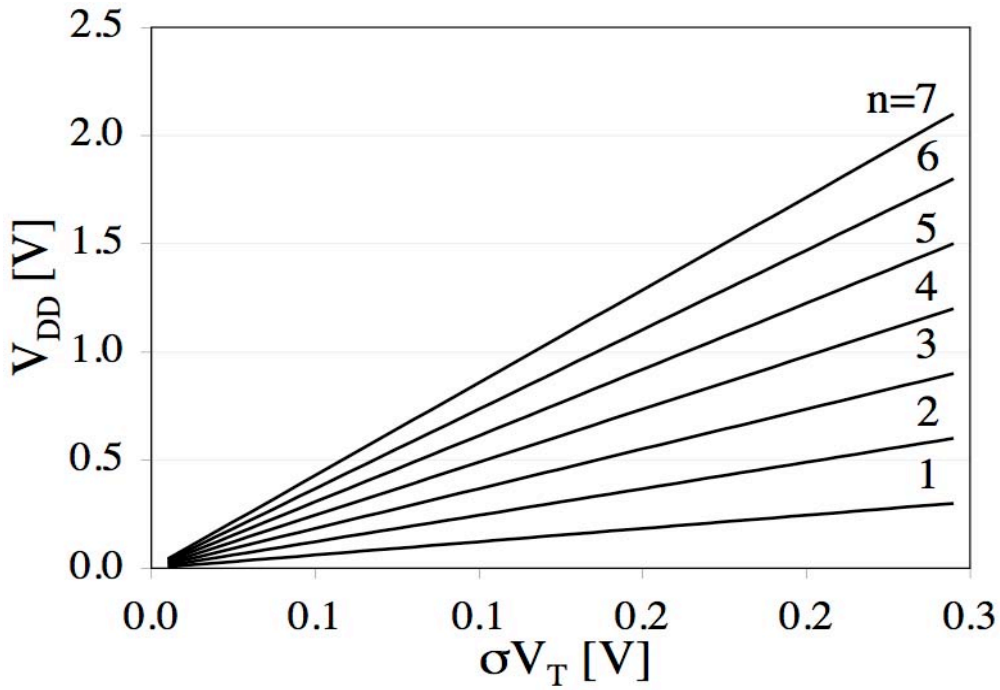


Figure 4-6 : Dependence of the minimum supply voltage on the standard deviation of the threshold voltage for a minimal size inverter and for different values of n defining the design margins.

quasi-constant safety margin correction. All the estimates are also based on the assumption that *n*- and *p*- channel MOSFETs have similar threshold voltage standard deviations for equal channel width. Since in a well balanced inverter $W_p = 2W_n$ assumption $\sigma V_{Tp} = \sigma V_{Tn} / \sqrt{2}$ is made.

The dependence of the minimum supply voltage on the standard deviation of the threshold voltage is plotted in Fig. 4-6 for a minimal size inverter and for different values of *n* defining the design margins. There is an assumption that both V_T and therefore V_{fp} follow normal distributions (an assumption which needs further careful testing, but is beyond the work of this thesis). From Fig. 4-6, σV_T in the range of 100 mV limits the supply voltage to approximately 1 V for the minimal size inverters (for 7σ design margin) particularly if the additional safety margins are included. In the rest of this section, the supply voltage limitations of bulk MOSFET CMOS implementations are reviewed.

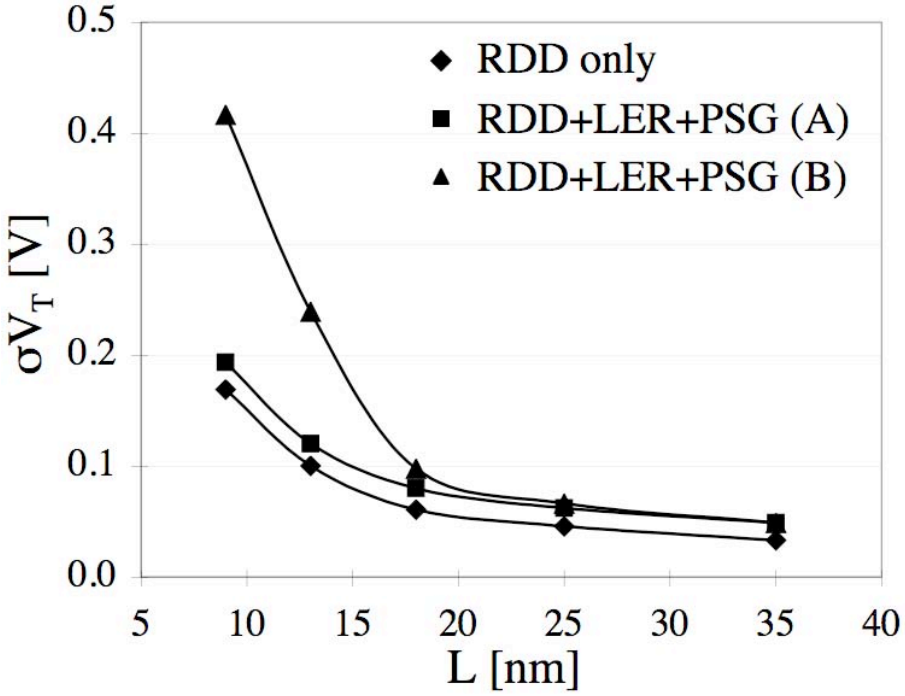


Figure 4-7 : Channel length dependence of σV_T taking into account only RDD and RDD, LER and PSG in combination: in scenario A, LER follows ITRS 2005 prescriptions; in scenario B, LER=4 nm taken from [52][5][109].

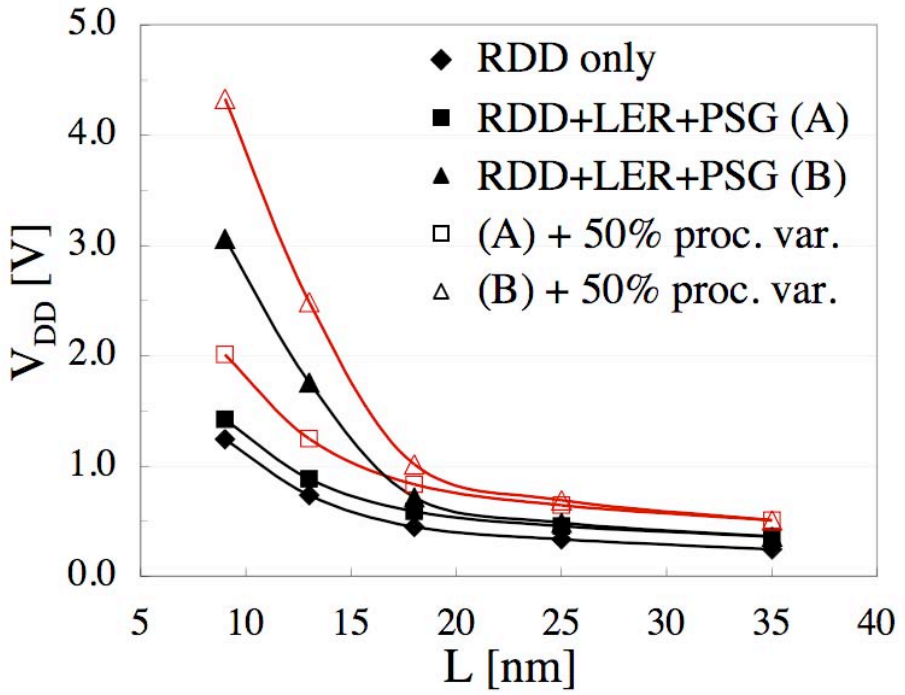


Figure 4-8 : Channel length dependence of the minimum allowable supply voltage corresponding to 6σ design margin for a minimum size inverter using solid symbols for the data for σV_T presented in Fig. 4-7. Open symbols examine the scenario when the simulated statistical variability is the same magnitude as the process induced variability.

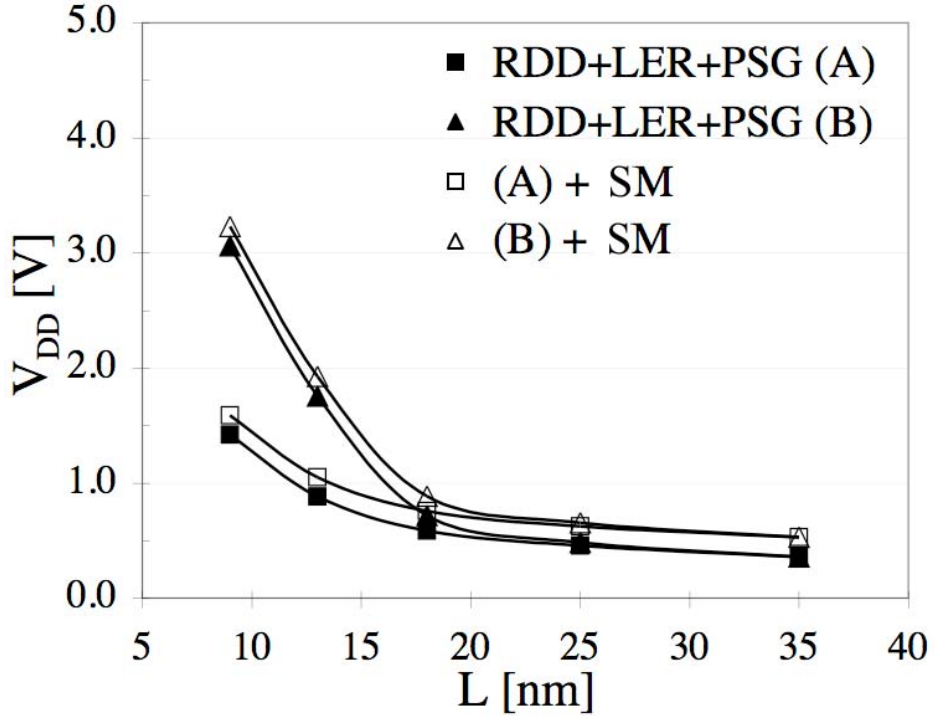
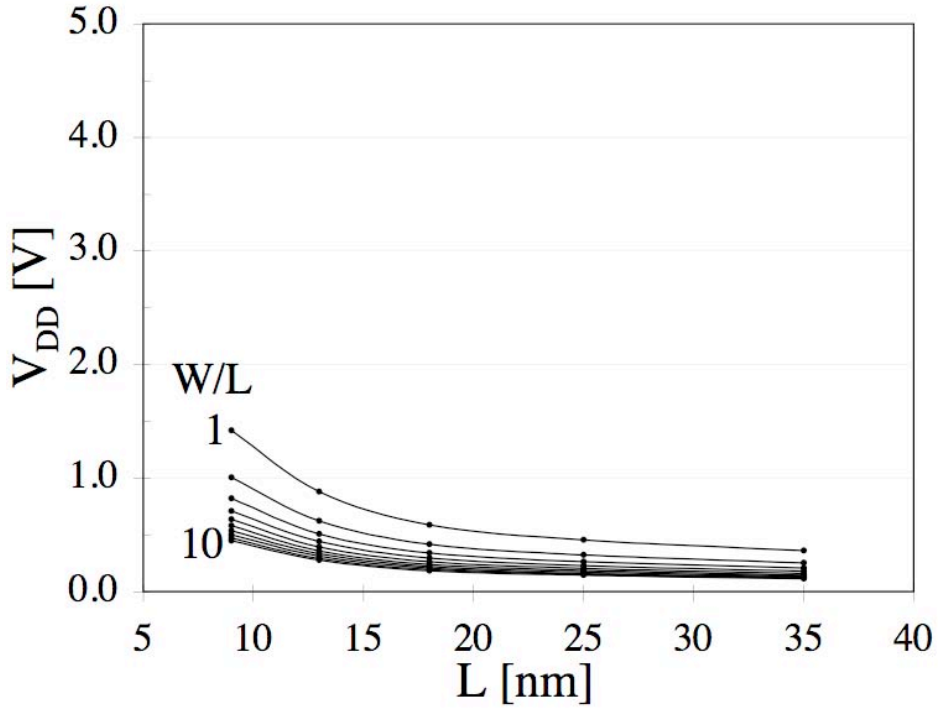
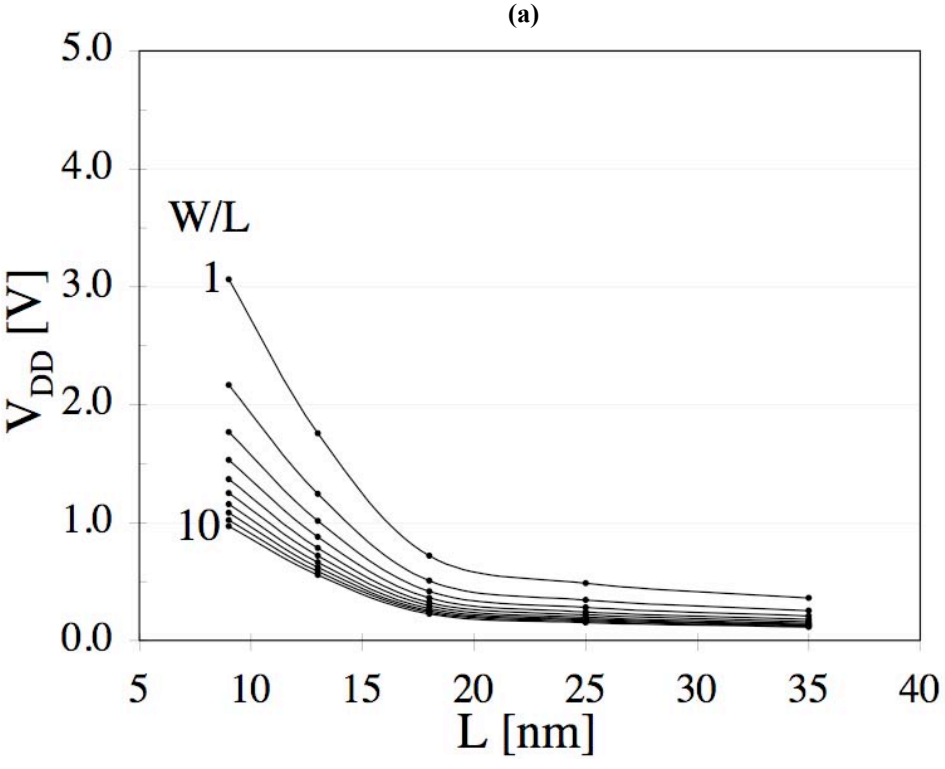


Figure 4-9 : Comparison of the 6σ supply voltage limitations for Scenarios (A) and (B) with (void symbols) and without (solid symbols) 170 mV noise margin added.

The simulated gate length dependence of σV_T for the scaled *n*-channel MOSFETs is illustrated in Fig. 4-7, which compares simulation results taking into account only RDD with results taking into account the simultaneous effect of RDD, LER and PSG [52][5][109]. In the second case which all effects are simulated, there are two scenarios for the LER being considered. In Scenario (A) LER follows the 2005 ITRS prescriptions as shown in Table 4-2. In the Scenario (B) LER is kept at 4 nm (when stating LER values, the 3σ value is usually quoted, i.e. σ = 1.3 nm in this case) for all channel lengths. This assumption is made based on the best lithography reported in 2005 [110] that includes e-beam lithography in research labs, and was allegedly limited by the fundamental nature of resist chemistry which is limited by 1 nm [111][112]. The LER scaling predicted in ITRS was simply an extrapolation based on the expected development of new generations of photoresist material [113]. From Fig. 4-7, for minimum size transistors, σV_T breaks the 100 mV ceiling at a channel length of approximately 15 nm for scenario A, and at approximately 18 nm



(a)



(b)

Figure 4-10 : Gate length dependence of the hard digital fault supply voltage limitations for transistors with different *W/L* ratios: a) LER follows the 2005 ITRS prescriptions; b) LER is kept at 4 nm for all channel lengths.

for scenario B. After these breaking points, σV_T increases much more rapidly with the reduction of the channel length in scenario B compared to scenario A.

TABLE 4-2

ITRS 2005 prescriptions for 3 sigma line edge roughness.

Channel length [nm]	35	22	18	13	9
Simulated LER [nm]	2.6	1.9	1.3	0.9	0.7

The gate length dependence of the minimum allowable supply voltage corresponding to a 6σ design margin for a minimum size inverter is plotted in Fig. 4-8 using solid symbols for the data of σV_T presented in Fig. 4-7. The void symbols represent results in which an assumption is made that the simulated variability in scenarios A and B are only half of the total device variability (statistical and systematic variability) – a typical situation at the 45 nm technology generation [43]. This ratio, however, is expected to change to a position where the statistical variability associated with the discreteness of charge and matter becomes more dominant in future technology nodes with excellent integration of Design for Manufacturing (DFM) techniques. For completeness, in Fig. 4-9, the 6σ supply voltage limitations for Scenarios (A) and (B) with (void symbols) and without (solid symbols) the additional 170 mV safety margin added are also compared. From the data presented in Figs. 4-8 and 4-9 it is clear that for bulk MOSFETs the hard logical faults limitation for the supply voltage breaks above 1 V for gate lengths smaller than 15 nm. Transition to ultrathin body SOI or multiple gate MOSFET architectures, which tolerate low channel doping and reduce the RDD related statistical variability, have been put forward as a way to allow supply voltage reduction for low power applications [114][115][116][117][118]. However, the results here show this will only be the case if the LER can be properly scaled according to roadmap projections.

It is fair to point out that all the above predictions are made for minimum size inverters, which may be rare in practical integrated circuits. Typically, primitive

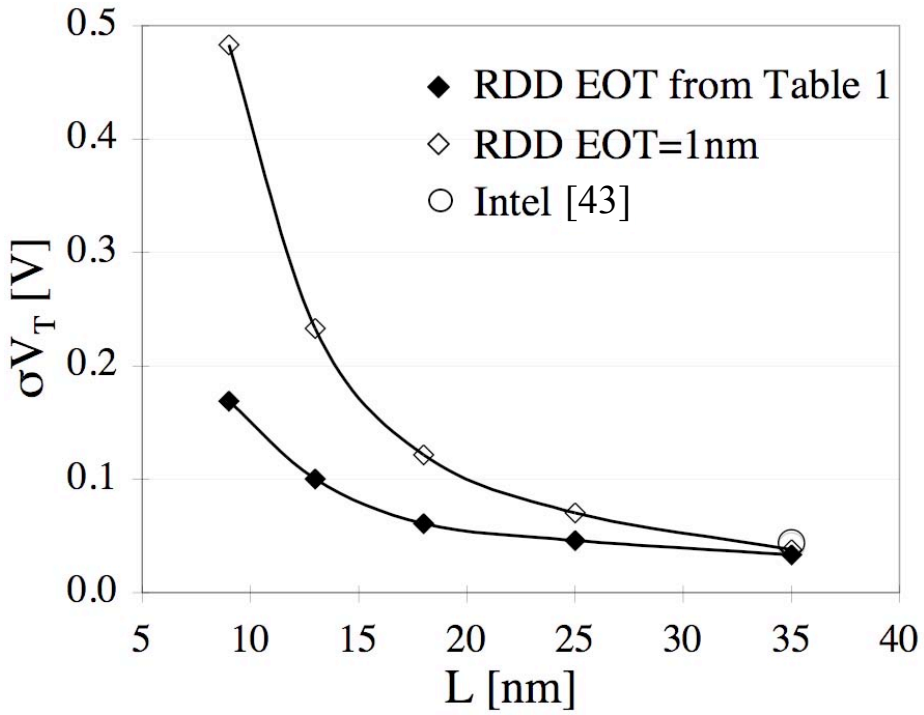


Figure 4-11 : Comparison of the RDD induced standard deviation of the threshold voltage σV_T for transistors with different gate lengths considering EOT from Table 4-1 and EOT = 1 nm taken from [119].

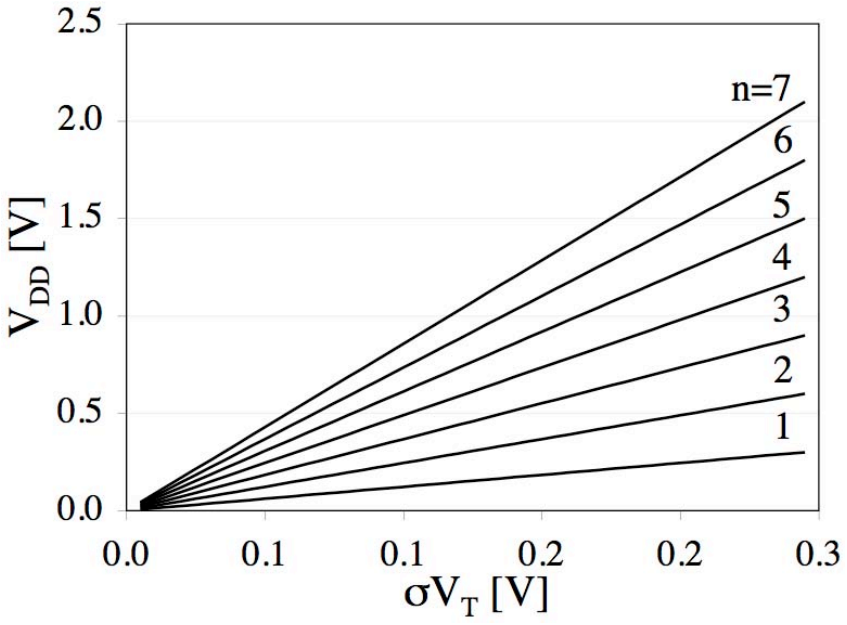


Figure 4-12 : Gate length dependence of the hard digital fault supply voltage limitations for transistors with different *W/L* ratios and EOT=1nm.

standard cells (i.e. NAND, NOR, INV) are designed for a wide range of drive strength (i.e. 1, 2, 8, 12, 20) which are primarily transistor-sized based [120][121]. Inverter with drive strength of 1 is chosen based on its optimum delay, power and area that can be obtained for a particular technology and inverter with drive strength of 4 is designed by increasing the devices size by a factor of 4.

Assuming that the statistical variability scales as 1/(square root) of the gate area, following the work in [122], in Fig. 4-10, the gate length dependence of hard digital fault supply voltage limitations based on *n*-MOS transistors with *W/L* ratio of 1 to 10 is examined. Fig. 4-10 (*a*) presents results that correspond to Scenario (A) while Fig 4-10 (*b*) presents results corresponding to Scenario (B). An increase in channel width relaxes the hard digital fault supply voltage limitations. In Scenario (A) this pushes the 1 V supply voltage floor to physical channel lengths below 10 nm. For Scenario (B) the floor remains higher, somewhere around the 14 nm range for *W/L* ratio increased larger than factor of 1.

The predictions for the scaling of the gate oxide thickness that guided the scaling of the devices used in this paper were based on the optimistic extrapolations of pre-2009 ITRS editions. With the introduction of high-κ gate dielectrics by Intel, 1 nm equivalent oxide thickness (EOT) has been achieved for 35 nm physical gate length transistors corresponding to the 45 nm technology generation [36]. This is larger than the 0.88 nm used in simulations performed in Glasgow for transistors with the same gate length. Therefore it is instructive to consider the pessimistic scenario when the EOT cannot be scaled below 1 nm. As illustrated in Fig. 4-11, in this case the RDD variability, which is inversely proportional to the oxide thickness remains the most important source of statistical variability in bulk MOSFETs. Note that the simulated and the estimated (from [43]) variability in 35 nm square transistors with 1 nm EOT is very close. The dependence of the minimum supply voltage on the MOSFET channel length corresponding to this scenario is illustrated in Fig. 4-12 for well balanced transistors with different *W/L* ratios of the driver transistor.

4.5 Summary

In this chapter, using statistical SPICE simulations, the impact of statistical variability on power supply voltage scaling in digital circuits was investigated. Statistical simulations were performed using the integrated 'atomistic' compact models of well scaled 35, 25, 18 and 13 nm gate length bulk MOSFETs, applying supply voltage levels prescribed by the ITRS. The minimum power supply voltage was evaluated for the ideal case and taking into consideration the safety margins and noise margin. An analytical model for the statistical variability of a CMOS inverter based on a simple model for the saturation current in decananometer scale MOSFETs was presented. The model was validated with respect to statistical circuit simulations of inverters with 35 nm, 25 nm, 18 nm and 13 nm physical gate lengths MOSFETs. The analytical model relates directly the inverter variability to the threshold voltage variability of the underlying MOSFETs. Results of comprehensive physical simulations of the threshold voltage variability of the scaled transistors were used to estimate the gate length dependence of the minimum supply voltage determined by hard logical failures of inverters at chosen design margins. Random Discrete Dopants (RDD), Line Edge Roughness (LER) and Poly Silicon Granularity (PSG) were considered as statistical variability sources in this study. In the simulations, two scenarios were explored with respect to LER scaling. In the first scenario the LER was scaled according to the requirements of the 2005 edition of the International Technology Roadmap for Semiconductors (ITRS). In the second scenario LER was kept at the present level [110]. For 6σ design margin of a minimum sized inverter, the minimum gate length which allows supply voltages below 1 V is in the neighbourhood of 15 nm, depending on the LER scaling scenario. For larger *W/L* ratios, the supply voltage floor is lower, moving the 1 V floor level to gate lengths of around 10 nm in a scenario which assumes continued LER scaling, and to 14 nm in a scenario which assumes that LER stays the same.

Restriction in the supply voltage scaling of future-scaled bulk CMOS devices due to the presence of statistical variability will counteract the advantage of geometry scaling as the dynamic power cannot be scaled any further. The restriction results from the circuit failing to function, in this case, the inverter is unable to invert its input logic level in the presence of statistical variability - not because of manufacturing defects which creates topological changes in the manufactured circuit. Although statistical variability can affect the actual operation of minimum size CMOS devices, this effect can be ameliorated simply by increasing the *W/L* ratio of the logic. However, this technique will reduce the advantages from the scaling in terms of increasing the circuit density. It also increases the output load capacitance and subthreshold leakage current in circuits of which contributes to larger dynamic and static components of power dissipation respectively. In modern digital electronic, especially mobile electronics, circuits not only have to operate correctly, but operate within a timing and power constraints to be commercially viable. The results of this chapter give the circuit designer a simple first order analytical technique to make informed choices balancing device width (and thus circuit size and silicon area) against reliability which can give first order results with minimal computational effort. This is a novel result of this work.

Chapter 5

Accuracy Of Transient Simulations Using BSIM Compact Models

5.1 Introduction

A compact model is a simplified, semi-analytical model describing a device operation which is used in circuit simulators such as SPICE to predict behaviour of a circuit design. A transistor compact model describes the transistor operation. The mathematical formulation of transistor compact model is based on semiconductor device equations which are the Poisson and current-continuity equations; and parameter values used in the formulation may represent a physical and non-physical information in order to get the best fit of the measured curves. Requirements of transistor modelling for circuit simulation are increasing due to device geometry scaling where inclusion of advanced physical effects in the model are necessary in obtaining accurate circuit simulation results and integration of more functions on a single chip prohibits increase in model execution time in a circuit simulator.

Compact models are the link between foundries and design houses. The electrical characteristics of devices manufactured using various foundry processes are captured using compact models so that designers can use those devices with confidence. The parameters used in the compact model are extracted from measurement data on devices of various sizes, and specific test structures [123]

[124][125]. Sub-micron devices are sensitive to manufacturing technologies and thus different foundries, employing different processing steps and recipes will produce devices with different characteristics for identical nominal gate lengths. Compact models must be able to relate the MOSFET operation to the transistor structure and geometry. It also must be flexible enough to accurately fit the differences in the measurement data resulting from the different processes used to fabricate particular devices in a particular foundry.

In general, there are 3 main types of compact models which aim to deliver the required properties for accurate circuit simulation, each with claimed benefits. These are charge based models (e.g. BSIM4), surface potential based models (e.g. PSP, HiSIM) and transconductance based models (e.g. EKV) [126][127]. The charge based models describe the drain current directly in relation to applied biases. While the surface potential based models describe the drain current in relation to surface potential at the source and drain. The surface potential at the source and drain are calculated by solving the Poisson equation iteratively as a function of applied biases (HiSIM) or by using an analytical approximation of the surface potential (PSP). Both model types use the charge sheet and the gradual channel approximations (which assume that potentials vary slowly across the channel allowing the 2-D problem to be solved as 2 separate 1-D problems). The advantage of surface-potential over charge based is the need for less fitting parameters to describe the drain current over *all* operating regions. The transconductance model describes the drain current in relation to inversion charge densities and is more applicable for analogue circuit simulation. Whilst these different models accommodate different needs in circuit simulations, the BSIM charge based model has historically been the model of choice in the digital design industry, and will be the model considered in this work.

In the next section of this chapter the BSIM formulation will be briefly discussed, with an emphasis on how it deals with internal transistor capacitances. Section 5.3 then gives a short description of transient simulations using BSIM

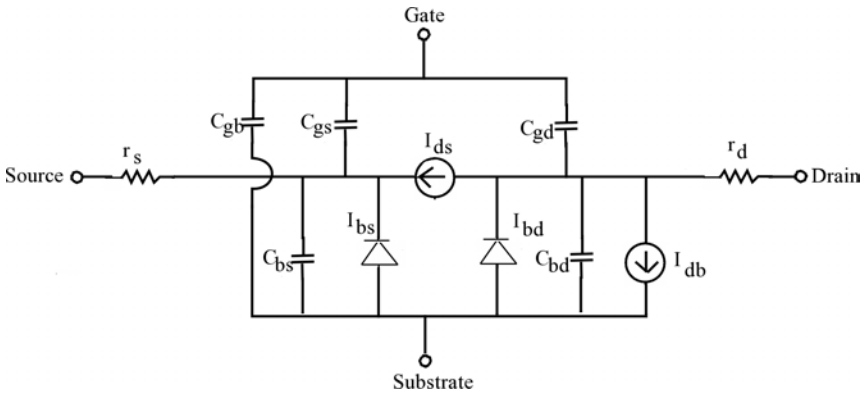


Figure 5-1 : MOSFET equivalent circuit model for transient analysis [136].

compact models with the SPICE circuit simulator. These two sections will give sufficient background to understand the context of the results presented in the remainder of the chapter. In section 5.4 the characterisation of BSIM4 models for 35 nm gate length devices aimed at digital circuit applications is presented. Then, the accuracy of dynamic behaviour of a simple inverter modelled using these compact models are compared with more *ab initio* TCAD simulation results in section 5.5.

5.2 BSIM Formulation

BSIM compact models have been developed at the University of Berkeley. Since the introduction of version BSIM3v3 in 1997, it has become a standard MOSFET compact model widely used in the design industry to model the complex behaviour of transistors in predicting circuit behaviour. The compact models are constantly being updated when advancing to a new technology node. BSIM4 has approximately 200 parameters to model the transistor behaviour with more added in every new technology generation. The latest compact models produced by the foundries that are publicly accessible to the academic community today are the 65 nm technology node transistor from the Taiwan Semiconductor Manufacturing Company.

In this section, the formulation of the BSIM compact model is discussed. In formulating a model for fast and accurate circuit simulation using the SPICE circuit simulator, the most desirable features of the model are: 1) that its description of the MOSFET drain current and all its derivatives with respect to terminal voltages must be continuous; 2) that it should require the smallest number of adjustable fitting parameters consistent with the physical effects to be captured, hence minimising the fitting process; 3) that there should be efficient computational convergence of the model equations to enable large circuit simulations over reasonable timescales.

5.2.1 Current-Voltage Relation

The drain current formulation in BSIM is a descendent of the Meyer model. In the Meyer model, the MOSFET's drain current in the subthreshold regime, ($V_{GS} < V_{TH}$) is described by an equation that approximates the diffusion current. Above threshold, ($V_{GS} > V_{TH}$) the drain current is described by two equations approximating the drift current in the linear ($V_{DS} < V_{DSAT}$) and saturation ($V_{DS} > V_{DSAT}$) regimes [128].

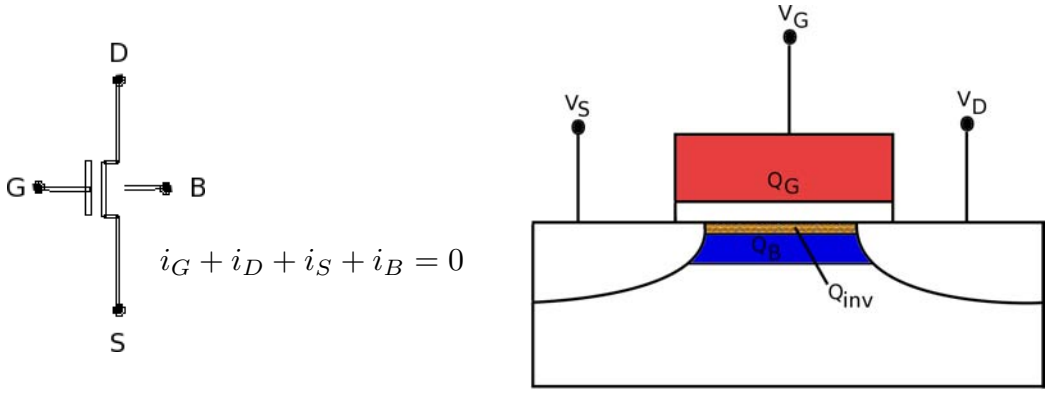
As different equations are used to describe the MOSFET drain current at different gate and drain biases, discontinuities in the drain current and its derivatives may occur at the transition points, $V_{GS} = V_{TH}$ and $V_{DS} = V_{DSAT}$. Discontinuities in the drain current characteristics are not desirable in circuit simulation because they cause non-physical results due to non-convergence of the current calculations. In order to eliminate these discontinuities, a smoothing function is applied at the transition points and the drain current equation in the BSIM3v3 model is reformulated to describe a continuous drain current from the subthreshold to strong inversion regimes [129]. The transition between the subthreshold and linear region is smoothed by transforming the gate voltage, V_{GS} into V_{gsteff}, while the transition between the linear and saturation region is smoothen by transforming the V_{DS} into V_{dseff}. The implementation of smoothing functions in the model, introduces nonphysical parameters.

Treatment of short channel, narrow width and non-uniform doping effects in small geometry MOSFETs are implemented by considering their effect on the threshold voltage, which becomes a function of a number of structural, electrical and fitting parameters including; body bias, effective gate length, oxide thickness, channel doping, etc. [42]. To improve model accuracy, fitting parameters have also historically been introduced into the mobility equations, parasitic resistance values and channel length modulation equations. Due to the introduction of a large number of fitting parameters, the model, whilst flexible, now offers less physical insight into device operation, and a complex hierarchical methodology is required to perform the extraction of these parameters. Further detail of the drain current formula is described in [28].

5.2.2 Capacitance-Voltage Relation

While the current-voltage relation solved by using Poisson and current density approximations, describes the steady-state current behaviour of a MOSFET device at different applied biases, accurate transient analysis of the MOSFET device also requires accurate modelling of its terminal capacitances. The terminal capacitances is used to describe the movement of the charges within the device with respect to time which is solved by using Poison and current-continuity approximations. These three equations namely Poisson, current density and continuity equations are important to be solved analytically in modelling accurate MOSFET device behaviour for circuit simulation.

In the Meyer approach, the four terminal MOSFET is assumed to be represented by a network of 5, two-terminal capacitances; C_{gs}, C_{gd}, C_{gb}, C_{bs} and C_{bd} as illustrated in Fig. 5-1 [128]. The non-linear capacitances C_{gs}, C_{gd} and C_{gb}, are expressed as a derivative of the gate charge (Q_G) with respect to its respective terminal voltage change, plus extrinsic capacitance components. Other intrinsic transcapacitance components are fixed to zero. However, the transient current derived from the Meyer capacitance model ($i = C\ dV/dt$) leads to a loss of charge



(a) **(b)**

Figure 5-2 : a) Charge conservation model. b) Simplified MOSFET cross-section with induced charge densities.

conservation, which can result in erroneous circuit simulation results [130][131] [132].

In the charge conservation law, the net change in the amount of electric charge in any volume of space must be equal to the net amount of charge flowing into the volume minus the amount of charge flowing out of the volume. Thus, the total currents that are flowing into and out of the devices must be equal to zero as illustrated in Fig. 5-2 (*a*). Fig. 5-2 (*b*) illustrates the charges that exist in a MOSFET device of which the quantity of the intrinsic charges need to be preserved at all operating region at all time which is the bulk charge (Q_B), channel charge (Q_{inv}) and gate charge (Q_G). The bulk (Q_B) and channel (Q_{inv}) charges can be analytically approximated at any gate potential from solution of the 1-D Poisson equation on the equivalent MOS capacitor structure. While the gate charge (Q_G) is $Q_G = -Q_B - Q_{inv}$. Based on this analytical expression of the gate charge, the Ward-Dutton model preserves charge conservation by introducing a charge partitioning scheme in the evaluation of the drain and source charges: $Q_{inv} = Q_D + Q_S$, $Q_D = X_{part} \times Q_{inv}$, $Q_S = (1 - X_{part}) \times Q_{inv}$ where $0 \leq X_{part} \leq 1$ [132].

The formulation of the BSIM capacitance model adopts this charge partitioning approach and uses charges as state variables in order to guarantee charge conservation in the MOSFET. All the intrinsic transcapacitances ($C_{i,j}$) are modelled as partial derivatives of the intrinsic charges with respect to terminal

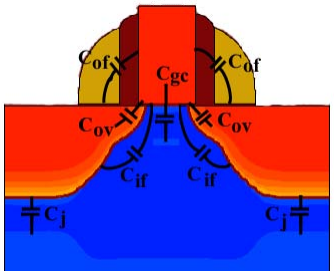


Figure 5-3 : MOSFET capacitances.

voltages as shown in Eqn. 5-1 where *i* and *j* stand for gate (*G*), drain (*D*), source (*S*) or bulk (*B*).

$$C_{i,j} = \frac{\partial Q_i}{\partial V_j} \qquad where\ i \neq j \tag{5-1}$$

The detailed expressions that describe the gate, bulk and channel charges are parameterised by the same threshold voltage (V_T), subthreshold slope (n), bulk-charge effect (A_{bulk}), oxide thickness (T_{ox}) and body bias coefficient (γ) variables that are used in the steady state current-voltage formulations. Additional fitting parameters also help to fit the *C-V* curves to measurement data. Further details on this charge formulation can be obtained in [133][28].

Fig. 5-3 shows the typical parasitic capacitances in a MOSFET device that result from its physical structure and which contribute to the 5, two-terminal capacitance values noted above. This parasitic components also are referred to as external capacitance components. The C_{gc} is related to the intrinsic capacitance components discussed above. In BSIM, the bias-independent overlap capacitance, C_{ov} is modelled using a parallel-plate approximation while the bias-independent outer fringe capacitance, C_{of} is modelled via a conformal transform. The inner fringe capacitance, C_{if} which is bias-dependent is not modelled in the BSIM capacitance model. The source/drain to bulk junction capacitance, C_j is divided into 3 components, bottom area capacitance (C_{AREA}), sidewall or peripheral capacitance along the 3 sides of junction's field oxide (C_{SW}) and sidewall or peripheral capacitance along the gate oxide side of the junction (C_{SWG}). All the parasitic capacitances are modelled as a function of device geometry and are treated as add-

on to the intrinsic gate capacitance description. For example, the total gate-to-drain capacitance is modelled as $C_{gd} = dQ_G/dV_D + 0.5C_{ov} + 0.5C_{of} + C_j$ where C_j is the junction capacitance related to the drain terminal [28].

5.3 SPICE Transient Simulation

SPICE is a circuit simulator use to enable prediction of a circuit behaviour by using compact models that represent each simulated circuit component. It translates the components and its network connection into equations to be solved. The SPICE simulator is heavily used in the analogue circuit design and standard cell characterisation of digital logics. A commercial tool HSPICE is used in this work.

Fig. 5-1 shows the equivalent circuit of MOSFET in SPICE transient simulation. The four-terminal transistor is described by 5 capacitances representing the gate, source and drain capacitances, parasitic source and drain resistances, current sources representing the d.c. effects and diodes representing the junction current between the substrate and drain/source terminals. The transient gate, drain and source currents flowing into the device nodes are calculated using Eqn. 5-2 where $I_{i,DC}$(t) is the d.c. terminal current which depends on the bias condition. The second term, $(\delta Q_i/\delta V_j).(dV_j/dt)$ describes the displacement current showing the capacitances, C_{ij} explicitly [134][135].

$$I_i(t) = I_{i,DC}(t) + \sum_j \frac{\partial Q_i}{\partial V_j} \cdot \frac{dV_j}{dt} \tag{5-2}$$

5.4 35 nm Device Characterisation

In this section, we present results for the device characterisation of 35 nm gate length halo-doped MOSFETs developed using the 2-D process simulator, Sentaurus based closely on industrially relevant state-of-art physical MOSFETs. The devices were developed using Sentaurus Process TCAD tool which uses finite element mesh solver to solve the physical and analytical models that describe each

manufacturing process step [137]. The tool is used to replicate the doping profiles of a real 35 nm physical gate length *n*-MOSFET fabricated by Toshiba. The detailed description of the manufacturing process steps used to develop the devices themselves can be found in [88]. Next, the devices are simulated in Sentaurus device tool which uses a finite element solver to solve the semiconductor device equations coupling the Poisson, current density and continuity equations in determining its electrical properties [138]. The device characteristics obtained from this quasi-stationary and mixed mode simulations are fed into a parameter extraction tool to generate the BSIM4 compact model. The outcome of the device characterisation are the parameters of a BSIM4 compact model developed with the purpose of performing digital circuit simulations. Then, mixed-mode simulations are performed in the Sentaurus device tool to obtain transient responses of an inverter using the developed devices. In the mixed mode simulation, the semiconductor devices characteristics are calculated numerically and are combined with other circuit components, the time varying supply voltage and a constant capacitor using a similar model to the SPICE circuit simulation approach. The transient response from the TCAD simulation is then compared against the simulation performed in SPICE to test the accuracy of the compact models. Detailed description of the employed TCAD simulation methodology to produce the required data for comparison analysis presented in chapter can be obtained in [88].

5.4.1 Current-Voltage Characteristics

Each device is simulated in the Sentaurus TCAD tool to obtain 1) I_D-V_G characteristics at high and low drain biases with varying substrate/body biases; and 2) I_D-V_D characteristics at zero substrate bias with varying gate biases. About 100 points from each *I-V* curve are extracted from the TCAD simulation and given as input to the *Aurora* tool. Parameter extraction was performed using *Aurora*, a commercial general purpose optimisation software tool for fitting analytical models

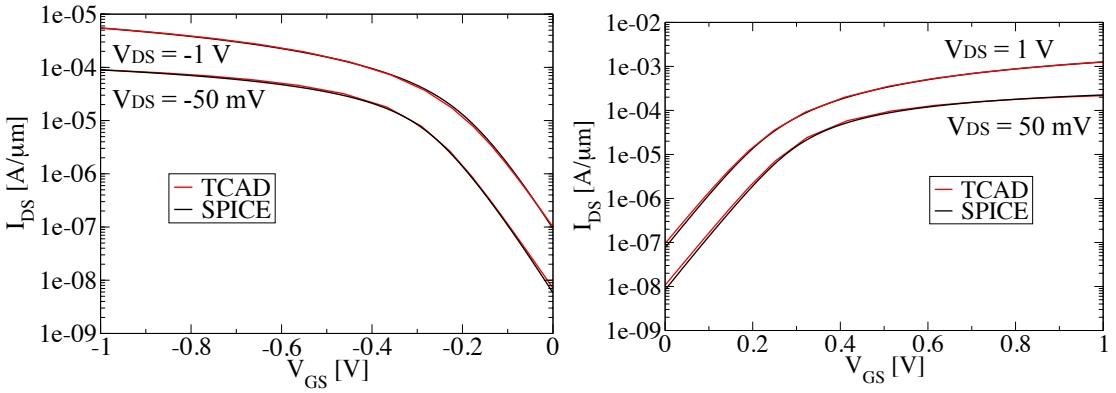


Figure 5-4 : Comparison of I_D-V_G characteristics of *p*-MOS (left) and *n*-MOS (right) between TCAD and SPICE simulation result.

to data [139]. After full extraction of the model parameters are completed, the current-voltage and capacitance-voltage characteristics of both the *n*-MOSFET and *p*-MOSFET devices were simulated using SPICE and then compared against the original TCAD simulation data. An overall *RMS error* comparing the SPICE and original TCAD results was calculated using Eqn. 5-3 where x_1 is the fitted data, x_2 is the actual data and n is the number of samples considered.

$$RMSerror = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}} \tag{5-3}$$

Fig. 5-4 shows the I_D-V_G characteristics of *p*-MOSFETs and *n*-MOSFETs biased at $|V_{DS}|$ = 50 mV and $|V_{DS}|$ = 1 V. Good agreement between the TCAD and SPICE simulation data is obtained for the drain current behaviour at low and high drain biases. The RMS error of the I_D-V_G characteristics fitting is shown in Table 5-1. The smaller fitting errors observed in *p*-MOSFET devices is due to the smaller absolute drain current values in *p*-MOS (approximately 2.3 times smaller than *n*-MOS drain currents). Overall, the normalised RMS error for I_D-V_G fitting for both *p*-MOS and *n*-MOS devices are in the range of 0.3-1.5% which is normalised by the span of the on- and off-current of the device at the measured terminal voltage conditions.

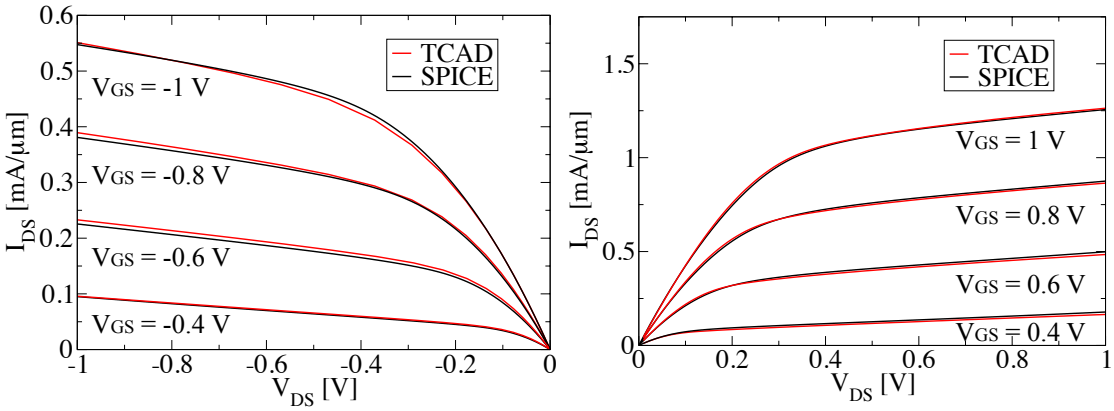


Figure 5-5 : Comparison of I_D-V_D characteristics of *p*-MOS (left) and *n*-MOS (right) between TCAD and SPICE simulation result.

TABLE 5-1
RMS error of the I_D-V_G curves at different applied drain biases.

	RMS error @|V_{DS}|=1V	RMS error @|V_{DS}|=0.05V
PMOS	2.315 E-06	1.181 E-06
NMOS	4.537 E-06	3.327 E-06

Fig. 5-5 shows the I_D-V_D characteristics of both *p*-MOSFETs and *n*-MOSFETs at different gate biases. Good agreement is also obtained in fitting the drain current biased at different gate voltage values to the TCAD data, with a normalised RMS error between 0.6-6.0% for both devices, which are normalised by the span of the on and off-current of the device at the measured terminal voltage conditions. Details are given in Table 5-2.

A good fit has been obtained for the d.c. characteristics of both the 35 nm gate length *p*-MOS and *n*-MOS devices at various applied biases. BSIM models are able to capture the drain current characteristics of these 35 nm gate length devices very well.

TABLE 5-2
RMS error of the I_D-V_D curves at different applied gate biases.

	RMS error @|V_{GS}|=1V	RMS error @|V_{GS}|=0.8V	RMS error @|V_{GS}|=0.6V	RMS error @|V_{GS}|=0.4V
PMOS	4.082 E-06	4.855 E-06	5.029 E-06	1.096 E-06
NMOS	7.520 E-06	8.343 E-06	9.509 E-06	9.204 E-06

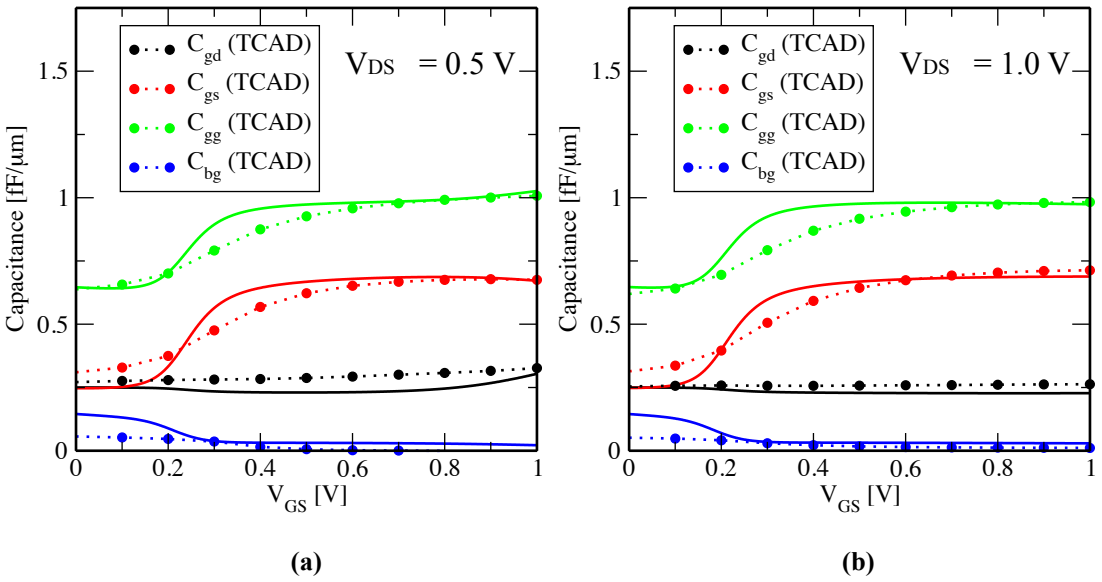


Figure 5-6 : C_G-V_G of *n*-MOSFET at different applied drain biases a) V_{DS} = 0.5 V b) V_{DS} = 1 V.

5.4.2 Capacitance-Voltage Characteristics

As discussed in the previous section, current-voltage characteristics only capture the d.c behaviour of a MOSFET device. In order to accurately predict the dynamic behaviour of the device in a circuit simulation, the capacitance-voltage characteristic must also be modelled accurately. In this subsection, fitting result of capacitance-voltage characteristics are presented. About 30 sample points of each capacitance-voltage curve with applied gate bias ranging from -1.5 V < V_{GS} < 1.5 V are extracted from the mixed-mode simulation in the Sentaurus Device tool and capacitance-voltage formulation in BSIM is fitted to the TCAD data. Our results showing the fitting of SPICE *C-V* simulations to the original TCAD *C-V* data is shown in Figs. 5-6 to 5-8.

It is clear that the capacitance-voltage characteristics simulated using the BSIM model do not deliver the same good match to the original TCAD data that were obtained in respect to d.c. characteristics. In addition the gate-related capacitance (C_{gg}, C_{gd}, C_{gs}, C_{bg}) at different applied biases also show a large deviation from the TCAD data particularly near the transition from weak to strong inversion regions as shown in Fig. 5-6. A slight deviation of the C_{gd} and C_{gs} curves from the

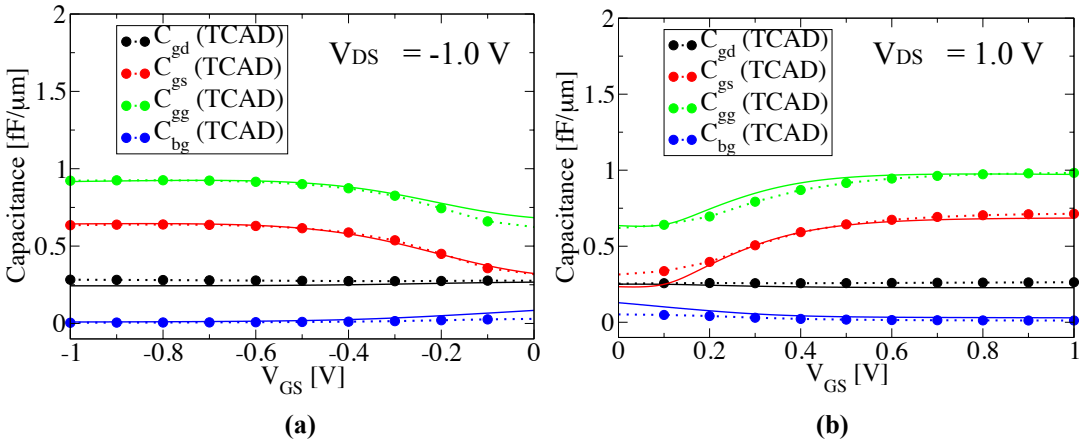


Figure 5-7 : Total gate-related capacitances comparison between TCAD and SPICE simulations a) *p*-MOS b) *n*-MOS.

TCAD, data particularly at V_{DS} = 0.5 V, is also observed. This is due to the charge partitioning scheme, which is done to evaluate the drain and source charges separately from the channel charge, Q_{inv} derivation and ensure charge conservation in the MOSFET model. Only the total substrate-to-drain/source capacitances, C_{bd} and C_{bs} fit well across the operating regime since the intrinsic bulk charge, Q_B is weakly dependent on the MOSFET threshold voltage, as can be observed in Fig. 5-8. The fitting of C_{bd} and C_{bs} will be discussed later in this section. However, due to the small value of the capacitances in the range of *femto*-Farad and the small difference between the minimum and maximum values, the normalised RMS error in terms of percentage is quite large when evaluating the capacitances fitting error especially the substrate-related capacitances.

Next, the gate-related capacitances, (C_{gg}, C_{gd}, C_{gs}, C_{bg}) are manually fitted using *acde*, *noff* and *moin*, BSIM fitting parameters to achieve better fitting near the transition from weak to strong inversion region. The source of fitting error near the transition may be due to a conflict in the fitting algorithm in Aurora tool since in the parameter extraction process, the lightly-doped drain (LDD) option is being disabled. In BSIM, the overlap region of a MOSFET is also modelled with bias-dependent component to account depletion effect in the LDD region during the MOSFET operation [28][140]. However, in the 35 nm gate length devices, the overlap region does not consist of lightly-doped drain structure, hence the LDD

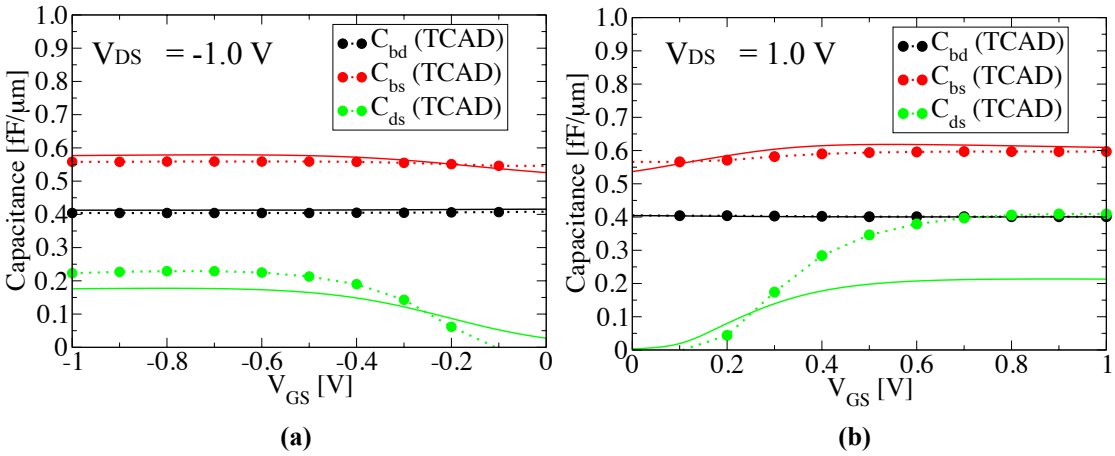


Figure 5-8 : Substrate-to-drain/source and drain-to-source capacitances obtained using TCAD and SPICE simulations a) *p*-MOS b) *n*-MOS.

option for capacitance fitting in Aurora tool is disabled. Fig. 5-7 shows total gate-related capacitances for *p*-MOS and *n*-MOS devices biased at V_{DS} = -1 V and 1 V respectively. Better agreement is achieved where the percentage error of the total gate capacitance, C_{gg} reduces from 15.31% to 6.5% after the refinement procedure. Table 5-3 displays the RMS error for the total gate-related capacitances fitting. Overall, the RMS error for every gate-related capacitance-voltage fitting of both devices is kept below 0.04 fF/μm per sample point.

TABLE 5-3
RMS error of the C_G-V_G curves.

	RMS error C_{gd}	RMS error C_{gs}	RMS error C_{gg}	RMS error C_{bg}
PMOS	2.948 E-17	0.896 E-17	2.691 E-17	2.286 E-17
NMOS	2.702 E-17	3.065 E-17	2.361 E-17	4.031 E-17

Fig. 5-8 shows the total substrate-to-drain/source capacitances, (C_{bd} and C_{bs}) of the 35 nm gate length *p*- and *n*-MOSFETs biased at $|V_{DS}|$ = 1 V where $C_{bd} = -dQ_B/dV_D + C_j$ and $C_{bs} = -dQ_B/dV_S + C_j$. The $-dQ_B/dV_D$ and $-dQ_B/dV_S$ terms are referring to the intrinsic-related capacitances and C_j is the sum of the three junction components at its respective terminal described in the introductory section earlier. The RMS error of the total substrate-to-drain/source capacitances are shown in Table 5-4 where the error is kept below 0.02 fF/μm per sample point.

All the 2-terminal capacitance components that form a network to represent a MOSFET in a transient analysis have been fitted to the TCAD data with accuracy of 0.04fF/μm per sample point or less. However, after inspecting the SPICE simulation result against the TCAD data, it is observed that a large deviation in the fitted data occurs at the total drain-to-source capacitance, C_{ds} across gate voltage, V_{GS} sweep as shown in Fig. 5-8. The issue is more prominent in the *n*-MOSFET where at V_{DS} = V_{GS} = 1 V, the C_{ds} value from the SPICE simulation is 1.91 times smaller than the C_{ds} value obtained in the TCAD simulation. While in *p*-MOSFET, the C_{ds} value biased at V_{DS} = V_{GS} = -1 V is 1.25 times smaller in comparison to the TCAD data. This is due to the formulation of the intrinsic charges, (Q_{inv} and Q_B) for transient simulation based on 1-dimension of Poisson equation which neglect several effects such as the mobility degradation in the channel [133]. Thus, the error between the TCAD and SPICE simulation is large in respect of the drain-to-source capacitance, C_{ds}. In SPICE, the C_{DS} (dQ_D/dV_S) is in function of other charges which preserves the charge conservation property as shown in Eqn. 5-4. Hence, the modelling the C_{DS} component separately will either introduce error to other drain-related capacitance components or void the charge conservation property.

$$\frac{dQ_D}{dV_D} = -(\frac{dQ_D}{dV_G} + \frac{dQ_D}{dV_S} + \frac{dQ_D}{dV_B}) \tag{5-4}$$

TABLE 5-4
RMS error of the *C*-V_G curves.

	RMS error C_{bd}	RMS error C_{bs}	RMS error C_{ds}
PMOS	0.853 E-17	1.660 E-17	4.558 E-17
NMOS	0.087 E-17	1.910 E-17	14.472 E-17

5.5 Transient Analysis of an Inverter

Next, the 35 nm gate length *p*- and *n*-MOSFETs are connected in series to form an inverter biased at the supply voltage, V_{DD} = 1 V. Fig. 5-9 (*a*) shows the

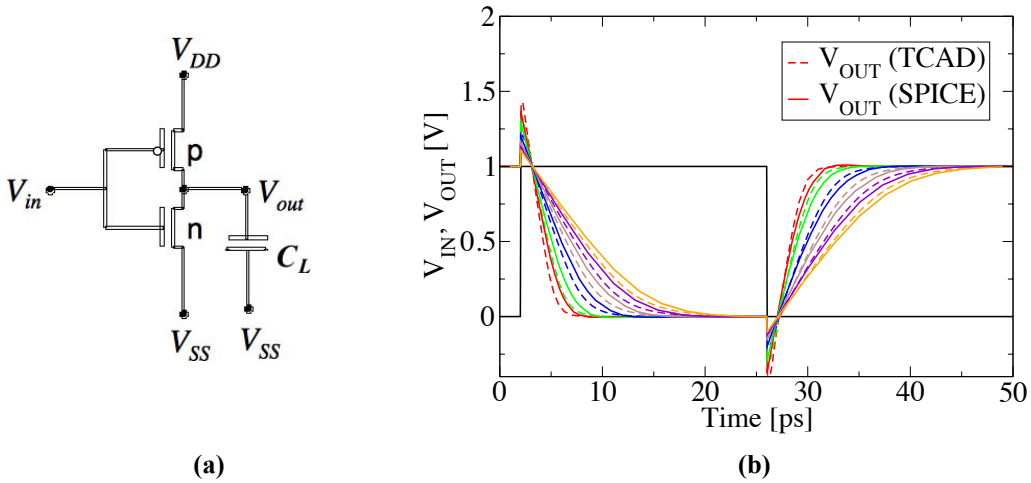


Figure 5-9 : a) Circuit schematic of an inverter implemented in SPICE and TCAD simulations. b) Transient response of the corresponding circuit in (*a*).

circuit schematic of the inverter with output connected to a fixed load capacitor, C_L. This load is varied from 1.08 fF to 10.8 fF with $C_L = n \times 1.08$ fF (n an integer) where 1.08 fF is equivalent to the total gate capacitance, C_{gg} of the simulated 35 nm x 1 µm *n*-MOS device in the linear regime. The applied input voltage is a pulse, linearly rising and falling between 0 V and 1 V, with transition time set to 0.05 ps. In order to match the on-current of the 1 µm width *n*-MOSFET (I_{DS} at $V_{DS} = V_{GS} = 1$ V), the width of the *p*-MOS device is chosen to be 2.3 µm. The 1 µm width *n*-MOSFET was simulated because it is the default device width value in the Sentaurus simulator.

Fig. 5-9 (*b*) shows the corresponding transient response of the inverter circuit shown in Fig. 5-9 (*a*) with the load, n = 1, 2, 4, 6, 8 and 10. The output transient of the inverter becomes longer due to charge/discharge of a larger load. The transient SPICE simulations are compared to the Sentaurus TCAD mixed-mode simulations – the similarly coloured dashed line in Fig. 5-9 (*b*). In every case the SPICE simulations show larger switching delay than the TCAD simulations. The inverter propagation delay of the falling-output transition (T_{DHL}) with C_L = 1.08 fF, obtained from the TCAD mixed-mode simulation, is 2.12 ps while in SPICE transient simulation the delay is 2.56 ps. The propagation delay of rising-output transition (T_{DLH}) for the same load, simulated in TCAD is 2.31 ps, while in SPICE it is 2.46 ps.

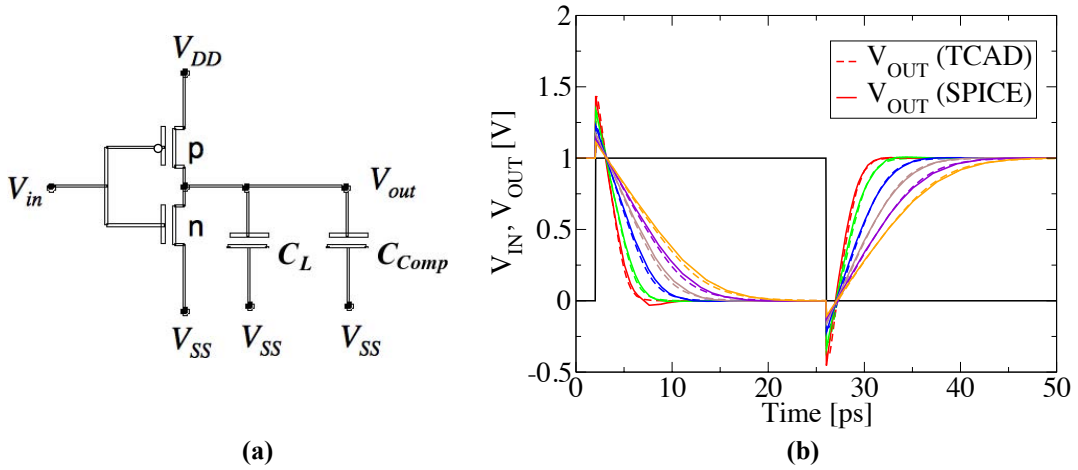


Figure 5-10 : a) Circuit schematic of an inverter implemented in SPICE in order to match TCAD simulations from Fig. 5-9 (*a*). b) Transient response of the corresponding circuits.

The difference in the inverter propagation delay between the TCAD and SPICE simulations is expected. It is the result of the large capacitance fitting error in the 35 nm gate length devices, particularly for the drain-to-source capacitance, C_{ds}, which itself results from the fact that BSIM does not account for the full 2-D physics of the devices, as discussed above. It is difficult to model such effects in BSIM while maintaining charge conservation and expecting fast and accurate circuit simulation.

The propagation delay of rising-output transition, T_{DLH} have a closer match to the TCAD data compared to the falling-output transition. This is consistent with the difference in the C_{ds} values seen above, where the C_{ds} error in *p*-MOSFETs are smaller than in *n*-MOSFETs. During the rising-output transition, the charging current is flowing through the *p*-MOSFET and hence the transient currents, calculated using the *p*-MOSFET models in the TCAD and SPICE simulations, are closer. The percentage errors in propagation delay are summarised in Fig. 5-11 (black curves).

In order to better match the TCAD simulation data, a compensation capacitor, C_{Comp} is connected in parallel to the load capacitor, C_L, as shown in Fig. 5-10 (*a*). The C_{Comp} value is varied to obtain the best fit to the TCAD propagation delay data. However, due to the differences in C_{ds} for *n*- and *p*-

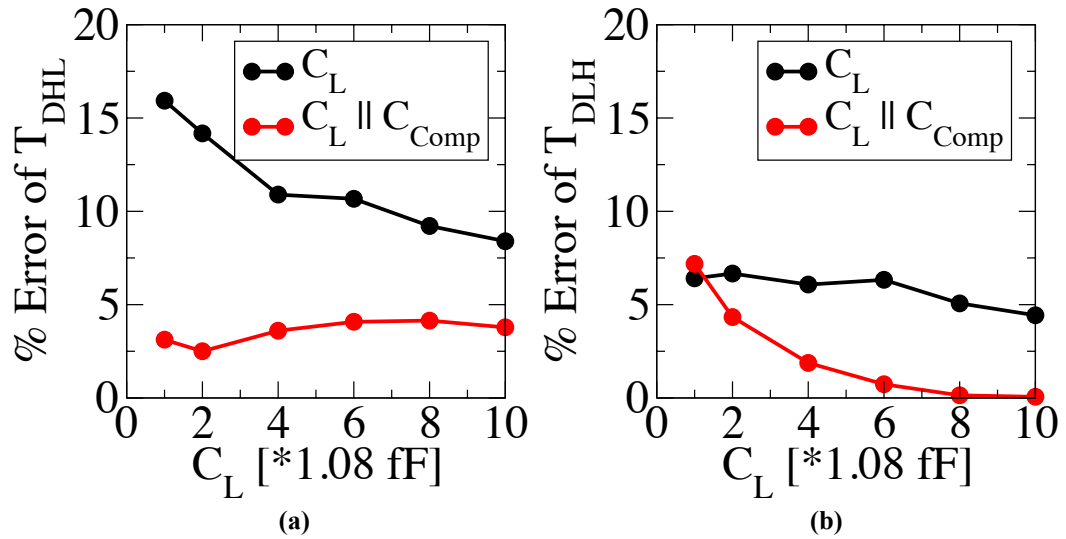


Figure 5-11 : Percentage error in the inverter propagation delay of a) falling-output transition, T_{DHL} b) rising-output transition, T_{DLH}.

MOSFETs, a single compensation value cannot match both propagation delays perfectly. Fig. 5-10 (*b*) shows the transient response of the inverter circuit with C_{Comp} fixed to 0.66 fF and the result compared with the TCAD data. A better agreement is indeed obtained with this compensation technique – however it should be reiterated that such compensation has no predictive power for different device sizes. Hence, when simulating a minimum size inverter of which both devices size are approximately 14 times smaller than the simulated inverter in this study, the compensation capacitor value may not scale by 14 times due to the increase in fringing effect not accounted in the BSIM model such as inner fringe and corner capacitances [28][197] which may dominate the transient response of the minimum size inverter obtained in TCAD. However, in order to obtain accurate magnitude of these effects, further 3D TCAD simulation and analysis are required which is beyond the scope of this study.

Fig. 5-11 shows the percentage error of the propagation delay for falling-output transition, T_{DHL} and rising-output transition, T_{DLH} for different applied circuit configurations simulated in SPICE. The percentage error of the falling-output transition, T_{DHL}, in the inverter without compensation varies between 8.5% to 16% (with the smaller error for larger loads). With compensation, the error in T_{DHL} varies between 2.5% and 4%. The percentage error of the propagation delay for the rising-

output transition, T_{DLH} in both circuit configurations decreases with increasing load, with the error in T_{DLH} rapidly decreasing with increasing load when compensated.

5.6 Summary

In this chapter, the accuracy of the BSIM4 compact model in capturing device characteristics and predicting circuit transient behaviour in SPICE simulation has been investigated. The compact models of the 35 nm physical gate length MOSFET were benchmarked against 2-D TCAD simulation. The BSIM4 compact model parameters were extracted over a range of device sizes and operating conditions using the compact model extraction tool, Aurora. The corresponding current-voltage and capacitance-voltage characteristics were compared against the current-voltage characteristics obtained from more *ab initio* TCAD simulations. The accuracy of the transient SPICE circuit simulation of an inverter using the extracted BSIM model of the 35 nm MOSFETs was evaluated against mixed-mode TCAD simulations. Excellent agreement between the TCAD and SPICE simulations are obtained for current-voltage characteristics of the MOSFET devices with normalised RMS error less than 6% for various applied gate and drain voltages. The main 5 BSIM model capacitors (C_{gd}, C_{gs}, C_{bs}, C_{bd}, C_{bs}) have been fitted accurately with fitting error below 0.04 fF/μm per sample point. Weaknesses in the BSIM capacitance model were discovered particularly in respect of the drain-to-source capacitance, C_{ds} at high drain bias for both *n*- and *p*-MOSFETs, found to be 1.91 and 1.25 times smaller than the capacitances obtained using TCAD physical device simulation. It was shown that these differences lead to inaccuracy in the transient simulation of the inverter where up to 16% larger falling-output propagation delay was obtained in SPICE simulation compared to the mixed-mode TCAD simulation. However, the percentage delay error reduces to 8.5% if a significant capacitive load (10 times higher than default) is connected at the output of the inverter. Compensation techniques were introduced to better match the SPICE simulated

propagation delay against the TCAD simulations leading to 4 times improvement in the SPICE propagation delay accuracy. Although these compensation techniques have little predictive power as devices scale, they will allow far more accurate transient BSIM simulation at any particular technology node, for a relatively small additional characterisation cost. The conclusion of this study is the BSIM4 compact model of the capacitive elements in advanced bulk-MOSFET must be revised in order to deliver greater predictive power in future scaled-devices resulting in accurate circuit simulations.

Chapter 6

Inverter Performance Variability Due To Random Discrete Dopants

6.1 Introduction

In Chapter 4 we have investigated the digital fault associated with statistical variability and the associated restrictions on the supply voltage. While two of the other manifestations of *statistical variability* (SV) at circuit and system level which are timing and power variability will be investigated in this chapter. At the 45 nm technology generation, intrinsic variability already accounts for more than 50% of the total variability seen experimentally, and is expected to become more dominant at the 32 nm technology generation and beyond [141]. Thus, understanding the impact of SV on digital circuit performance is crucial because it likely to become a limiting factor in future circuit and system design.

In conventional physical implementation flows, process variability has been handled using corner analysis. However, with advances in technology, more sources of variability and the possibility of correlations between variability sources, there are too many corners to be considered in the design process. This makes the worst and best case validation technique before sign-off very pessimistic. Accordingly, statistical design techniques have been put forward for the purpose of reaching a more optimal design before real tape-out. A statistical approach will provide

designers with a better understanding of how well the circuits behave when subject to statistical variation.

Attempts have been made to investigate the effects of random discrete dopants (RDD) on delay and power variations by generating circuit models [142] [143][144] using estimated fluctuations in the main electrical parameters such as threshold voltage, on-current, off-current and sub-threshold slope; with respect to the probability density function of overall doping concentration [145][78]. However, the analytical expressions developed in this methodology to predict device electrical parameters assume an ideal, uniformly doped substrate. Realistic, modern deca-nanometer device has highly non-uniform doping profiles (i.e. employ retrograde and halo doping) to suppress short-channel effects (SCE) in bulk-FETs [85][21]. Thus, these analytical formulae are not robust and scaleable, and have no predictive power for succeeding device generations.

In the following sections, as a step towards developing a methodology for the investigation of general digital circuits subject to the effects of intrinsic parameter fluctuations in real devices, we investigate foundational CMOS inverter circuits. These circuits are analysed subject to differing fan-in and fan-out conditions (with realistic form of the input and output signals established by embedding the inverters under test in an inverter chain, as shown in Figure 6-1) and using transistor models subject to RDD which closely match ITRS guidelines and present industry practice.

In section **6.2 Circuit Configurations**, we first describe why the inverter chain configuration is used as our testbench, comparing an idealised slew input to the inverter under test with more realistic input signal supplied by an inverter chain. The concepts of fan-in and fan-out are also discussed. Next, section **6.3 Inverter Switching Paths and Trajectories**, introduces the concept of the *dynamic noise margin* followed by a discussion of dynamic noise margins and inverter switching trajectories obtained from circuit simulation under differing drive and load conditions. The different definitions of drive current which are used in inverter delay

approximations using the *CV/I* metric are also assessed. Then, in section **6.4 Inverter Timing subject to Variability**, inverter delay distributions under different *FO/FI* conditions for 35 nm, 25 nm, 18 nm and 13 nm devices are investigated, and compared with analytic results using the definitions of the drive current discussed in section 6.3. Then, concept of logic depth, L_d is introduced and a critical delay path through a circuit is modelled as L_d stages of inverters fulfilling the maximum delay, T_{MAX} requirement. The impact of RDD on the logic depth, critical path delay and optimisation strategy to overcome the impact of RDD in the critical path subject to device scaling are investigated. In section **6.5 Inverter Power Dissipation subject to Variability** the impact of increasing the logic gate size on power dissipation is discussed, and in the last section a chapter summary is made.

6.2 Circuit Configurations

6.2.1 Inverter Chain

A chain of inverters as shown in Fig. 6-1 is used in this study. Inverters 1 to 4 (and inverters 5 to 8) are each of nominally identical width. Only INV4 and INV5 are selected randomly from the statistical ensemble of model cards (and thus exhibit statistical variability) whilst the other inverters are modelled from continuously doped devices. Their role is to provide realistic input/output transient shapes for the inverters under test.

In reality an inverter does not have infinite transconductance and will never deliver an ideal square output signal even if its input signal is an ideal square wave. In fact, if an ideal input waveform (square waveform) is applied directly at the input gate of the inverter under test connected to a very small load capacitor, a very high and unphysical voltage overshoot can be observed at the output of the inverter. In order to generate a realistic waveform shape for the device under test, inverters are connected in a chain. It is observed that the delay and shape of the waveform are consistent after passing through only 3 inverter stages even with an unphysical,

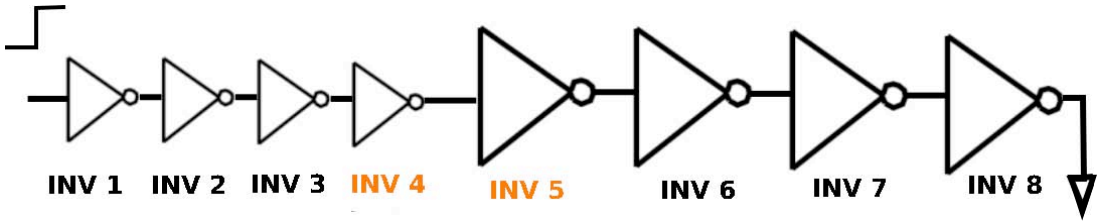


Figure 6-1 : Simplified inverter chain circuit diagram with *FO/FI=8*.

idealised input waveform is applied at the first inverter. Similarly the output of INV5 is connected to an inverter chain of 3 stages to represent a realistic load.

In this study, the minimum sized, or unit *n*-MOSFET device in each inverter has a width (*W*) of twice the gate length (*L*) values of 35 nm, 25 nm, 18 nm and 13 nm. *p*-MOSFET devices of double the width of the *n*-MOSFET devices are employed in order to match the *n*- and *p*-MOS drive currents in the CMOS inverters. (It has been reported that with the introduction of strain engineering in state-of-the-art devices, the effective mobility of holes can approach the electron mobility in the scaled-devices [146]. However, in this study, such mobility enhancement is not introduced in the test bed transistors. Including such mobility enhancement would lead to differing *p*-MOS to *n*-MOS sizing to match drive currents and thus may affect the delay variation results presented below.) A supply voltage of 1 V is assumed. Interconnect resistance and capacitance are neglected, since the aim of this investigation is to study the limiting impact of device variability. It should be noticed that short range interconnect variability may start to play important role in future technology generations.

6.2.2 Fan-in and Fan-out Concepts

Before considering circuit configurations which consist of more complex gates, the interaction between 2 inverters in the presence of RDD is investigated. In general, fan-in is a term used to describe a number of logic gate connected to an input node of a cell (i.e. the cell could be inverter, NAND, NOR logic gates) while fan-out is used to describe the number of subsequent input logic gate connected to

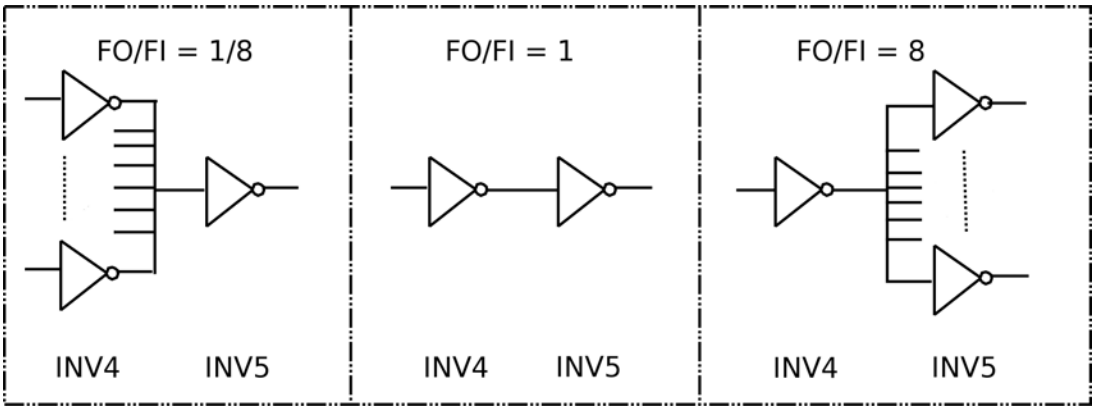


Figure 6-2 : Fan-out and fan-in configurations.

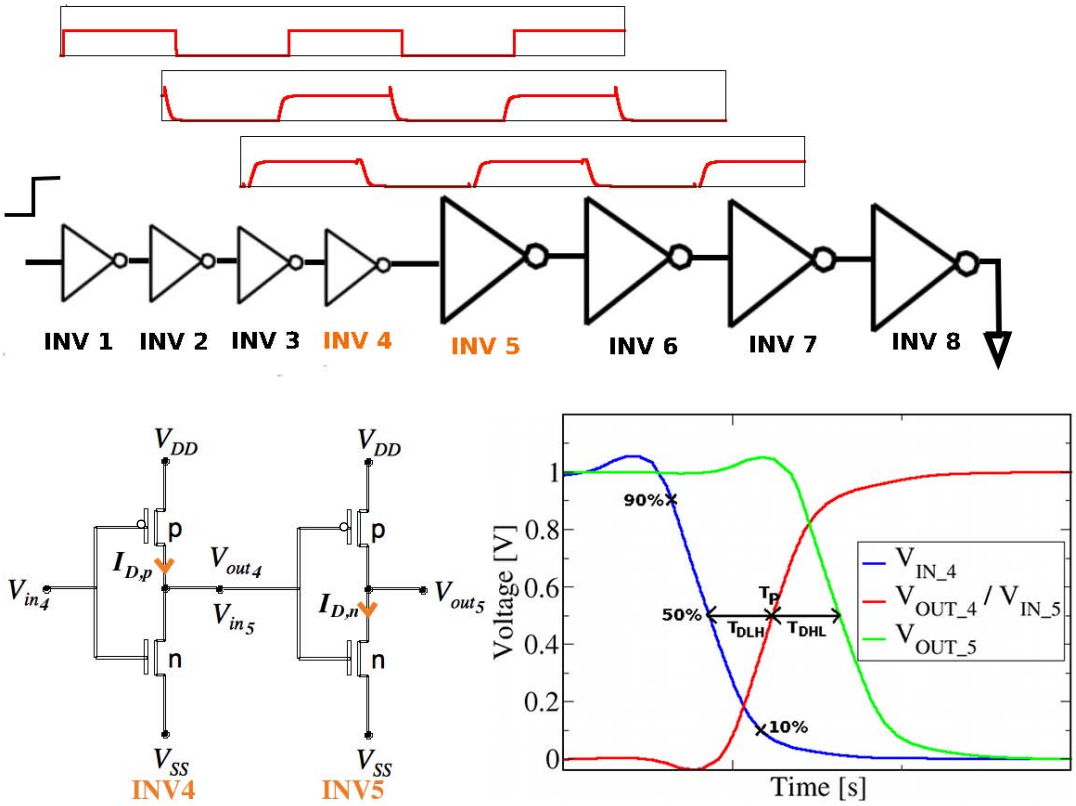
the output node of the cell. For these inverters, fan-in (*FI*) measures the width/ strength/current drive of the inverter driving the one under consideration (where that width is measured as a multiple of the width of the inverter under consideration). Fan-out (*FO*) measures the width of the inverter being driven by the output node of the inverter under consideration, or the number of identical inverters being driven.

Throughout this chapter, the term (*FO/FI*) refers to a certain configurations of INV4 and INV5 from the inverter chain of Fig. 6-1. Fig. 6-2 shows the circuit configurations that refer to *FO/FI = 1/8, 1* or *8*; which represent a large inverter driving a small inverter, balanced driver and load inverters; and small inverter driving a large inverter. The test vehicle is chosen from a recent study based on a chain of inverters shown in Fig. 6-1 with different *FO/FI* conditions [147]. The test configurations are suitable to study the effect of loading and input transition time on the propagation delays of inverters in realistic circuit simulations. The study in [147] demonstrated that both linear and saturation drive current has to be considered in the dynamic behaviour of an inverter. However that analysis does not include the impact of RDD, and thus cannot give a full understanding of the magnitude of timing and power *variability* at different *FO/FI* conditions.

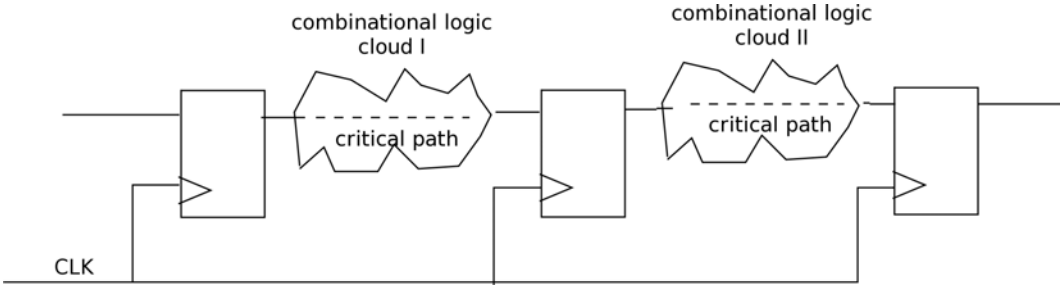
6.3 Switching Paths and Trajectories

6.3.1 Noise Margin Concept

Before discussing the results of this chapter, we first introduce more advanced concepts regarding noise margins and transient simulation. In undergraduate textbooks the static noise margin is discussed and understood. However, discussion of the *dynamic noise margin* is rarely encountered. The dynamic behaviour of a logic gate (INV, NAND, NOR, *etc*.) is best illustrated using a transient curve which plots the varying input or output voltages against time as shown in Fig. 6-3 (*a*). However, this transient curve does not clearly illustrate the noise margin that can be withstood by the logic gate during operation, and which is crucial when analysing and designing a circuit using sub-micron technologies. This is because in more advanced integrated circuit (IC) fabrication technologies, an increase in transistor density per unit area and reduction in interconnect layer thickness may introduce greater cross-talk noise originating from the increased capacitive coupling between the interconnect layers in the circuit. This capacitively coupled noise can affect gate delay (which occurs in transient mode, if the noise is in the form of short pulses during a switching event) or in the worst case scenario upsetting the function of a logic gate (which can occur either in transient or static mode, if the noise pulse width is infinite). In addition, supply voltage scaling also reduces logic gate noise margins, making them more susceptible to functional errors or delays. In real systems, critical path timing specifications must be met by a design at all times. In the local clock-enabled circuit shown in Fig. 6-3 (*b*), the logic state at the end of each critical path must be stable before it is being latched to another cloud of combinational logic. Thus, even if noise does not cause an overt functional failure at a particular logic gate, it may cause functional failure of the system if the overall delay in a path causes late arrival of a signal with respect to a clock edge. The mechanisms on how cross-talk noise produces circuit delay, using a classical victim/ aggressor model, is described in [148][149][150].



(a)



(b)

Figure 6-3 : (a) Inverter chain with its timing diagrams from INV1 to INV3. Transistor level circuit diagram of INV4 and INV5 showing the voltages and drain currents used in this study. Transient responses of INV4 and INV5 showing the timing definitions which will be used later in this thesis. (b) Local clock-enabled circuit showing critical path in combinational logic clouds.

Fig. 6-3 (*a*) also shows the timing definitions used in this study in section 6.3 and below. The propagation delays of an inverter are T_{DHL} and T_{DLH}, the delays during falling-output transition and rising-output transition respectively, measured from the 50% V_{DD} points of both the input and output voltage traces. T_P is the total propagation delay through two successive inverters (in this study, INV4 and INV5)

and the input slew rate, $SR = 1/T_T$, where T_T, the input transition time, is measured from 90% to 10% of V_{DD} or *vice versa*.

Let us now differentiate between the static noise margin and dynamic noise margin by using an inverter as an example. From the literature, the definition of the static noise margin of an inverter is clearly defined from a static voltage-transfer curve based on unity-gain point concept [151][152]. In a static voltage-transfer curve, the output voltage of an inverter is plotted against its input voltage taken from DC simulation of an inverter. Static noise margin indicates the DC noise amplitude that must occur at the gate of a long chain of inverters to cause an upset in the logic states after a very large number of inverter stages [153][154]. Of course in a normal design, it is rare to have infinite or large number of inverter stages, but this circuit topology is equivalent to two inverters connected in such a way that input node of the first inverter is connected to the output node of the second inverter and *vice versa* (also known as cross-coupled inverter pair) as shown in Fig. 6-4. This kind of circuit topology can be observed in flip-flops, latches and SRAMs. By using the cross-coupled inverter pair, the DC noise amplitude that, if occurring at the input of each inverter, will upset the logic state, can be clearly observed when the static voltage-transfer characteristics of both inverters are plotted on the same graph, as shown in Fig. 6-4(*c*). The DC noise amplitude is the size of the 'eye' in this 'eye diagram'.

In contrast to the static noise margin, a dynamic noise margin cannot be directly obtained from dynamic voltage-transfer curves (in the static noise margin case, it has a clear definition from the static voltage-transfer curve based on unity-gain points), neither is it trivial to calculate. This is because the dynamic noise margin does not only take into account the noise amplitude but also the noise pulse duration, which means the analysis of dynamic noise margin depends on the shape of the noise during a transient event [153][151]. The dynamic noise margin is best illustrated using a noise immunity curve which plots the noise pulse amplitude as a

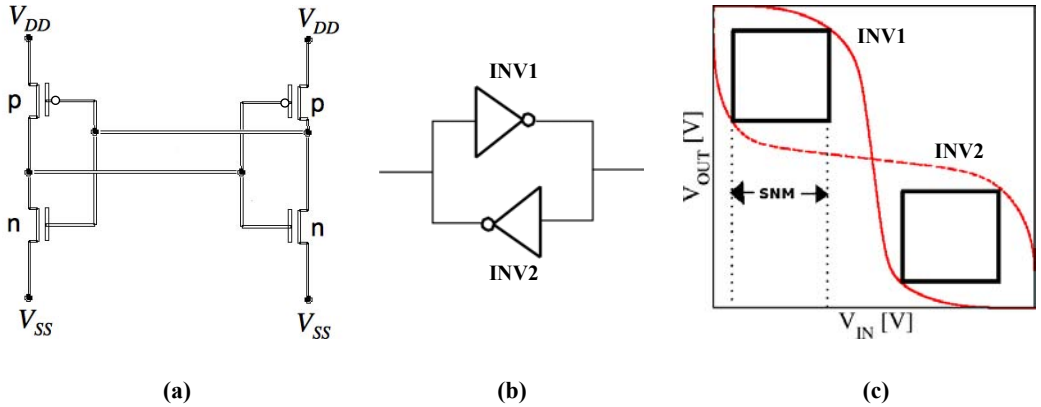


Figure 6-4 : (a) Transistor level circuit diagram cross-coupled inverter pair (b) logic level circuit diagram of cross-coupled inverter pair (c) static voltage-transfer characteristics of the cross-coupled inverter pair.

function of noise pulse duration. However, the noise immunity curve approach does not produce a single number and it is difficult to compare for different applications.

An attempt to define the dynamic noise margin by using a family of dynamic voltage-transfer curves (also referenced as *switching paths* in this chapter) has been made in [155]. The dynamic voltage-transfer curve is a plot of the output voltage against the input voltage of an inverter obtained from a transient simulation of which the applied input voltage and response at the output node of the inverter are varying with time. The author in [155] obtained the maximum square between normal and mirrored voltage transfer curves as the method of determining the static and dynamic noise margins as shown in Fig. 6-5. Fig. 6-5 (*a*) shows three transfer characteristics of an inverter where the curve in the middle is obtained from a DC simulation of an inverter and the other two curves are obtained from transient simulations of an inverter with the same applied input transition time, T_T and fixed load capacitor, C_L. The curves only differ in the output transition direction, where the right-hand side curve is plotted during falling-output transition while the curve on the left-hand side is obtained during a rising-output transition. Fig. 6-5 (*b*) shows the maximum square method applied by the author of [155] to obtain dynamic noise margin of an inverter during the rising-output transition. The inverter dynamic transfer curve during the rising-output transition is mirrored on $y = -x + V_{DD}$ axis, and the maximum square that can be fitted between the dynamic transfer curve and

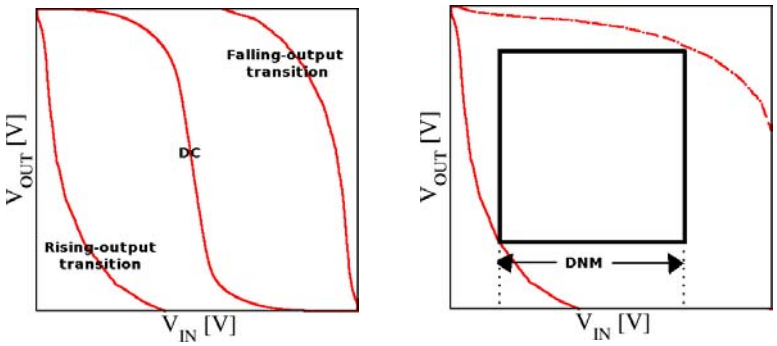


Figure 6-5 : (a) Static voltage-transfer curve and dynamic voltage-transfer curves of an inverter plotted on the same axes (b) dynamic noise margin obtained by using maximum square method used in [155].

its mirror is the size of the dynamic noise margin. However, the dynamic noise margin obtained by this method ignores the contribution of noise pulse duration and the dynamic noise margin is observed to be far larger than the static noise margin. As discussed by Loh Stroh in [153], the dynamic noise amplitude is allowed to be higher than the static noise margin because the dynamic noise margin is also dependent on the noise pulse duration (thus, a short pulse width with high noise amplitude may not cause a functional error). Using the approach in [155] it is sufficient to obtain relative comparisons between the dynamic noise margins of an inverter for different loadings and input slew rate conditions for some given, consistent applied noise shape. Smaller noise margins will indicate that the logic is more susceptible to functional error.

The dynamic voltage-transfer characteristics represent the relationship between output and input voltages of an inverter is shown. Depending on the properties of the switching transistors, *on* → *off* or *vice versa* in the inverter, input slew rate and output load conditions, the dynamic voltage-transfer curve (switching path) may vary for the same inverter as discussed in detail in the next section. From voltage-transfer characteristics as shown in Fig. 6-5 (*b*), the point (input voltage) at which the output voltage of the inverter begins to switch can be observed. The output voltage of an inverter is defined as the potential difference between the *drain* terminals of both *p*- and *n*-MOSFETs in the inverter; and the ground (V_{SS}) as shown

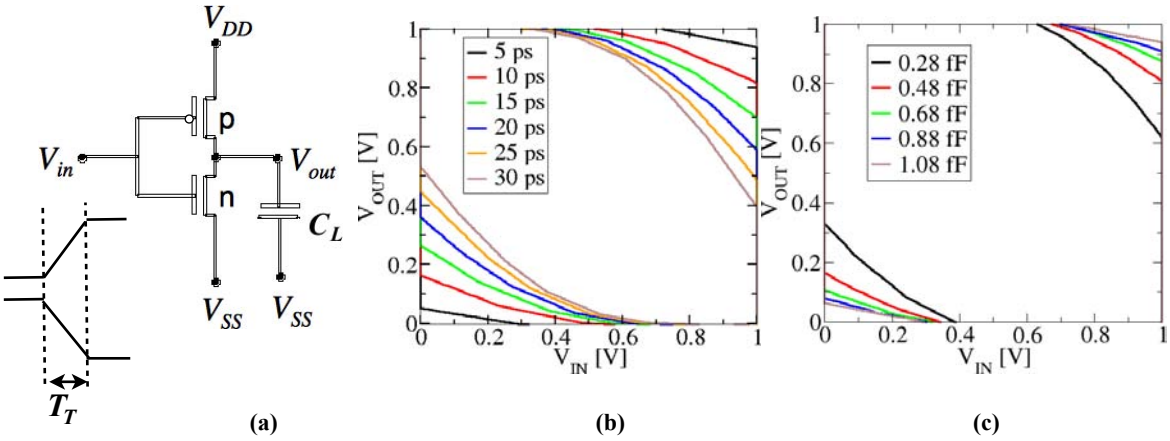


Figure 6-6 : (a) Schematic of a single inverter simulation where the input transition time, T_T and fixed load capacitor, C_L are the variables (b) dynamic transfer curves of an inverter with 1.08 fF fixed output load and T_T is varied (c) dynamic transfer curves of an inverter with 5 ps T_T and C_L is varied.

in Fig. 6-3 (*a*). The input voltage of the inverter is obtained from the potential difference between the *gate* terminals of both *p*- and *n*-MOSFETs; and the ground (V_{SS}). The same definitions of input and output voltages are also applied in plotting the static voltage-transfer curves.

In the following section, dynamic voltage-transfer curves (switching paths) and switching trajectories obtained from statistical SPICE simulations are presented. The aim of the study is to comparatively analyse the effects of RDD on the transient behaviour of an inverter. The dynamic noise margin discussed in the following section is obtained by using the maximum square method as described in [155] and is sufficient for this purpose. However, in order to quantify the dynamic noise margin for a specific circuit configuration, capacitive environment and specific noise pulses, further simulations are needed which are not covered in this study.

6.3.2 Inverter Switching Paths

As discussed in the previous section, the dynamic voltage transfer characteristics vary not only due to the different properties of the switching devices in an inverter (falling- or rising-output transition) but also due to different applied input transition time, T_T and fixed output load capacitor, C_L. Fig. 6-6 illustrates the effect of varying the input transition time, T_T and fixed output load capacitor, C_L on

the dynamic transfer characteristics of an inverter. The simulation was carried out using the compact models of the 35 nm gate length with uniform doping devices. *p*- to *n*-MOS ratio of 2 is selected and *n*-MOS width is chosen to be twice the gate length of the device. T_T of 5 ps and C_L of 0.28 fF intervals are chosen because 5 ps is approximately equivalent to half of intrinsic delay, τ of the simulated inverter while 0.28 fF is equivalent to 3.5 times of the total gate capacitance, C_{GG} of the simulated *n*-MOS transistor in the linear regime. The values of T_T and C_L are varied such that they cover the inverter simulation that ranges from fast to slow transient events. Fig. 6-6 (*b*) shows that the DNM decreases with increasing of input transition time while Fig. 6-6 (*c*) illustrates the increase in DNM with the increase in output load.

Next, the same size inverter as discussed above is simulated following the schematic diagram illustrated in Fig. 6-1 and three circuit configurations as shown in Fig. 6-2 are analysed. INV4 and INV5 are simulated using the 'atomistic' compact models for 35 nm to 13 nm gate length bulk-MOSFET devices subjected to RDD while the other inverters are simulated using their compact models of uniformly doped devices. As a result, variation in the input voltage with respect to time of INV4 and the output load of INV5 are not taken into account in the simulations. Fig. 6-7 shows the switching paths of INV4 during rising-output transitions, and INV5 during falling-output transition, for an ensemble of 200 circuits subject to RDD for different *FO/FI* cases. Switching paths of 35 nm and smaller gate length devices, at the same *FO/FI* values, are also illustrated in Fig. 6-7. By observing the switching paths of INV4, the influence of loading effect during each rising-output transition can be investigated. Inverters with higher output loads (*FO/FI=8*) stretch the switching path towards the bottom-left of the axis in the voltage-transfer characteristic thus maximising the dynamic noise margin. In the 35 nm gate length devices, the dynamic noise margin for inverter with *FO* of 8 increases to 0.89 V from 0.69 V for inverter with *FO* of 1. The relative increase in the dynamic noise margin of a minimum-sized inverter with 8 times increase in the load size is approximately 1.28 - 1.36 times for 35 nm - 13 nm gate length devices. Larger

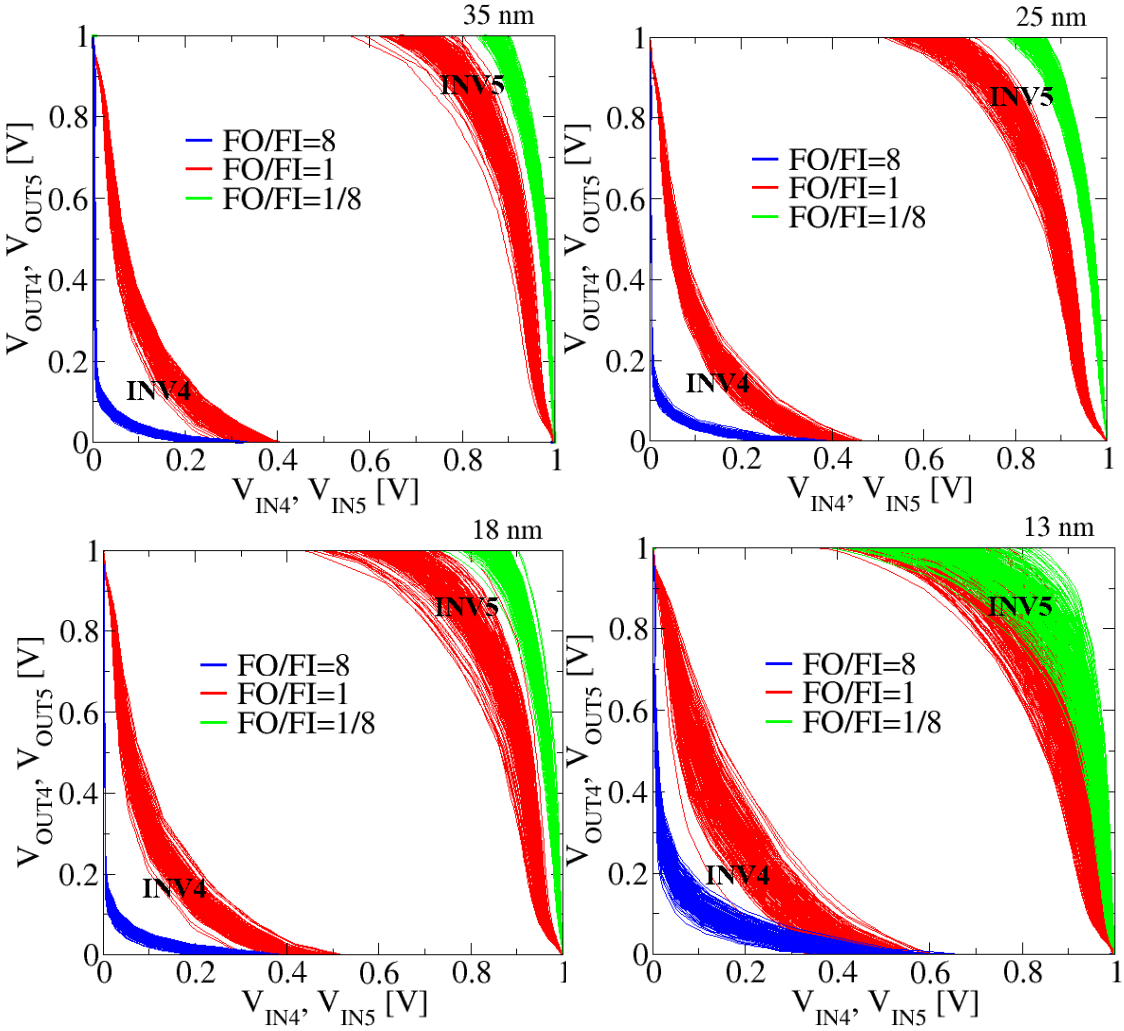


Figure 6-7 : Switching paths for INV4 (during rising-output transition)and INV5 (during falling-output transition) plotted on the same graph. INV4 and INV5 are subject to RDD and applied for different *FO/FI* cases. The switching paths are also plotted for devices with gate length of 35nm, 25 nm, 18 nm and 13 nm.

dynamic noise margin indicates that a higher amplitude of noise or larger noise pulse duration at the output node is needed to cause an upset in the logic state during the transient switching of a heavily loaded inverter (*FO/FI=8*). In the case of coupling noise, a larger coupling capacitance is needed to introduce a higher amplitude of noise pulse which can cause functional errors in the heavily loaded (*FO/FI=8*) compared to the lightly loaded (*FO/FI=1*) inverters. This is deduced from a simple model for crosstalk prediction described in [150] where the relationship of the noise pulse in function of coupling capacitance between the aggressor and victim, C_{12} and the capacitance at the victim interconnect to ground, C_{victim} is expressed using Eqn. 6-1. In this study, the C_{victim} is referring to the total capacitance of the two inverters

(INV4 and INV5) since the interconnect capacitance between the transistors is assumed zero. From Eqn. 6-1, a smaller amplitude of noise pulse is expected with higher load capacitance at the victim's interconnect, C_{victim}. Thus, in order to produce a higher amplitude of noise pulse, a larger coupling capacitance, C_{12} or smaller capacitance at the victim's plane, C_{victim} is needed.

$$\Delta V = V_{DD} \cdot \frac{C_x}{1 + C_x} \quad ; \text{ where } C_x = \frac{C_{12}}{C_{victim}} \qquad (6\text{-}1)$$

The effect of input slew rate from the switching paths of INV5 during the falling-output transition is also shown in Fig. 6-7. An inverter with larger fan-in (*FO/FI=1/8*) has a smaller input transition time (higher input slew rate) because of the larger capacity of drive current from the pre-driver inverter (in this case INV4) to charge/discharge its small load (INV5) quickly. The dynamic noise margin increases by 0.16 V for the 35 nm gate length inverter with larger *FI* from 0.71 V, the dynamic noise margin for INV5 with *FO/FI =1*. From Fig. 6-7, INV5 with *FI* of 8 for 35 nm, 25 nm, 18 nm and 13 nm devices show a relative increase of 1.22, 1.28, 1.29 and 1.13 in the dynamic noise margin respectively. This indicates that an inverter with higher input slew rate requires higher noise amplitude to cause a functional error at its output.

In the presence of RDD, the dynamic noise margin of an inverter with the same input slew rate or load conditions varies due to the variation in the electrical parameters introduced by random dopants. The relative variation, σ/μ of the dynamic noise margin for 35 nm gate length INV4 with *FO=8* is 0.7% which is smaller in comparison to 2.6% for INV4 with *FO=1*. While the relative variation σ/μ of the dynamic noise margin for 35 nm gate length inverter with *FI=8* is 1.5%. Due to the larger widths of the *p*-MOSFETs in INV4, the switching paths during the rising-output transition have smaller fluctuations compared with the switching paths during the falling-output transition in INV5 from Fig. 6-7. Thus, higher relative variation of the dynamic noise margin is expected in the inverter with falling-output transition than its rising-output transition.

In smaller gate length devices, where the variation in the electrical parameters introduced by RDD becomes more pronounced, the dynamic noise margin variation is also expected to increase. In Fig. 6-7 (*FO/FI=1*), the dynamic noise margin for INV4 is observed to decrease by the rate of 10% with device scaling. Not only that, its dynamic noise margin variation, σ also increases by 9%, 21% and 57% with device scaling as expected. This leads to an increase in the relative variation (σ/μ) of the dynamic noise margin in smaller devices. The reduction in the noise margin of scaled-devices is due to the reduction of intrinsic gate capacitance of a transistor, from geometry scaling. Thus it reflects the smaller load seen at the output gate of the scaled-inverter. Based on the previous discussion, the dynamic noise margin is shown to decrease with a smaller load size and the reduction of dynamic noise margin with device scaling is as expected. On the other hand, in order to maintain at least the same coupling noise amplitude at a reduced gate capacitance in smaller gate length inverters, the coupling capacitance, C_{12} between interconnects needs to be reduced when advancing to the next technology generation. This is because without the reduction in the coupling capacitance, C_{12} (for example, constant dielectric material of the the interconnect or thickness between interconnect layers) the coupling-noise amplitude is expected to increase in the circuit using smaller devices at the same applied supply voltage. This will certainly impose greater dangers to the signal integrity and logic functionality in circuits of which the logic gates have smaller dynamic noise margins. In the scaled devices where the effect of RDD becomes more prominent, when determining the maximum coupling capacitance based on the information from dynamic noise margin, variation in the dynamic noise margin induced by RDD must also be taken into consideration.

The dynamic noise margin and the effects of coupling noise generated by coupling capacitance between interconnects have been discussed. Even though there are other types of noise that can appear in circuits, such as supply and ground bounce noise [156] which could affect the transient behaviour of a logic gate, these

types of noise are not discussed further here, as we regard the maximum square method [155] sufficient to compare the relative susceptibility of circuits subject to RDD and scaling, using the dynamic noise margin under different *FO/FI* conditions. In summary, we have shown that scaling lowers the dynamic noise margins and increases their variability; while higher load and slew rates improve the noise margins and noise margin variability.

6.3.3 Inverter Switching Trajectories

The switching trajectories presented in Fig. 6-8 and 6-9 are the traces of switching current obtained from the drain terminal of the *p*-MOSFET in INV4 and *n*-MOSFET in INV5 plotted against output voltage during the rising-output transition of INV4, and during the falling-output transition of INV5, respectively. Fig. 6-8 and 6-9 demonstrate the variation in active switching profiles for 35 nm gate length devices. In these figures, the operating-point trajectories of inverters in ensembles under three different *FO/FI* conditions, and subject to RDD, are superimposed on I_{DS}-V_{DS} characteristics of uniform doping *p*- and *n*-MOS transistors respectively.

From Fig. 6-8, the switching current of the *p*-MOSFET in INV4 during rising-output transitions with *FO* of 8 reaches saturation at an early stage of switching $V_{OUT} \leq 0.9V_{DD}$, whilst for a *FO* of 1 the switching current reaches saturation when the output voltage has switched to somewhat over 50% of V_{DD}. This shows that the switching current flowing through the *p*-MOSFET of INV4 with smaller fan-out condition spends lesser time in saturation regime during the rising-output transition. This is because with large fan-out, larger current is being charged into the large load (which is the INV5). On the other hand, for lightly loaded INV4 (*FO/FI=1/8*), the switching current of the *p*-MOSFET in INV4 during the rising-output transition barely reaches saturation during switching. This is because the larger size inverter (INV4) produces a larger drive current, easily charging a small capacitance load (the gates of the transistors in INV5) resulted in a very fast rising-

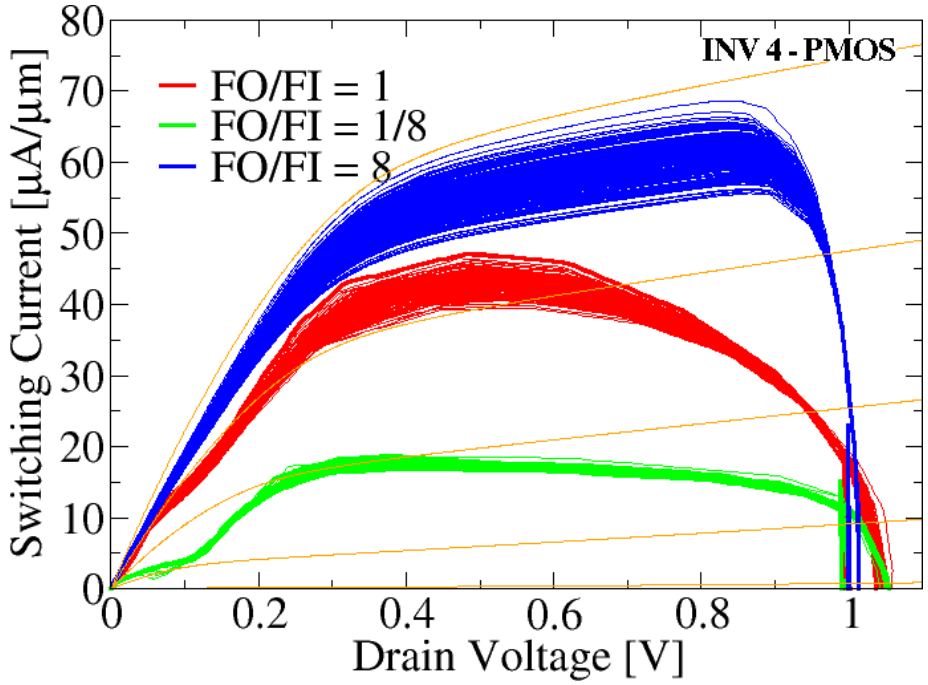


Figure 6-8 : Switching current of *p*-MOSFET in INV4 during rising-output transition as function of *FO/FI* ratio with variability in 35 nm gate length device.

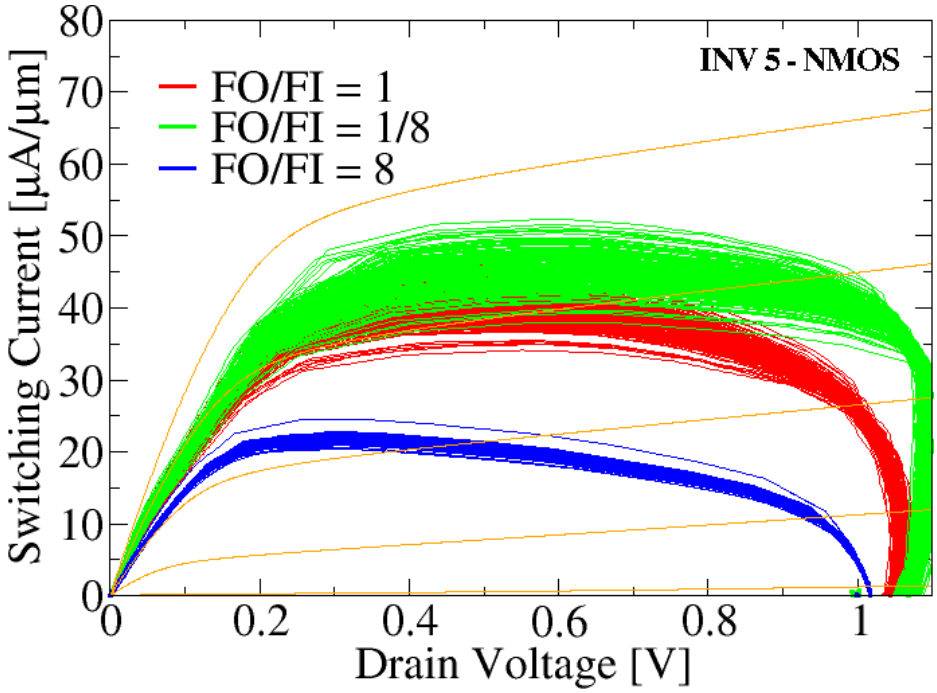


Figure 6-9 : Switching current of *n*-MOSFET in INV5 during falling-output transition as function of *FO/FI* ratio with variability in 35 nm gate length device.

output transition. In the presence of RDD, the largest variation in the switching profile is observed at INV4 with *FO=8* when the current trajectories are in saturation mode. The smallest variation in the overall current trajectory during rising-output transition is shown in the switching current of *p*-MOSFET in INV4 with *FO/FI=1/8* as illustrated in Fig. 6-8.

Fig. 6-9 depicts the voltage overshoot phenomenon, where the output voltage becomes larger than the V_{DD}, occurs at the beginning of the switching trajectories of *n*-MOSFET in INV5 during the falling-output transition. In Fig. 6-9, the switching current of *n*-MOS in INV5 for *FI=8* condition (which shows the highest magnitude of output voltage overshoot due to the smallest input transition time) reaches the highest saturation current in the middle of the trajectory instead of at the beginning of output voltage switching. INV5 with higher input slew rate (*FI=8*) applied at its input results in higher saturation current achievable during the falling-output transition. Higher voltage overshoot is also observed, with larger current flowing through the *n*-MOSFET in INV5 at the beginning of the switching trajectory. In the presence of RDD in INV5, the highest input slew rate (*FI = 8*) applied at its input shows the largest variation not only in the switching current achieved in saturation regime but also in the overshoot current during the falling-output transition as shown in Fig. 6-9. On the other hand, INV5 with the smallest input slew rate applied at its input (*FO/FI=8*), shows the smallest variation in the switching current flowing through its *n*-MOSFET during the falling-output transition.

Referring to Fig. 6-8 and 6-9, it can be observed that even with *p*-MOSFETs which are four times wider than the minimum transistor size (and thus statistically are expected to have half the maximum expected magnitude of statistical variations at this technology generation), the impact of RDD on charging current through the *p*-MOSFET of INV4 (during rising-output transition) can be still very large. The σ of the charging on-current is up to 3-4% of the mean charging on-current. As expected, due to the smaller *n*-MOS transistor width implemented in the minimum-sized inverter in this study, the variations in the discharge current through the *n*-

MOSFETs of INV5 (during falling-output transition) are larger that of their PMOS counterparts in INV4 during rising-output transition, the σ of the discharge on-current being in the range of 5-6% of the mean value.

In the conventional *CV/I* metric, I_{ON} ($I_{D \mid VGS=VDS=VDD}$) is used to estimate the intrinsic delay of an inverter. The intrinsic delay is defined as the delay of an inverter driving an identical inverter (*FO=1*) with no interconnect parasitics [157]. However, during inverter switching in ultra-scaled devices, the switching current, as shown in Fig. 6-8 and 6-9, never reaches I_{ON}. Hence, I_{ON} is unlikely to accurately represent the intrinsic delay of an inverter in scaled bulk-MOSFETs. When considering inverters with sub-micron CMOS feature lengths, I_{EFF} has been shown to more accurately capture the delay behaviour of an inverter and it has been used as an important metric to improve device performance [158][159][160]. The effective current I_{EFF} is defined as the average of drain currents I_{D_H} (measured at $V_{GS} = V_{DD}$ and $V_{DS} = V_{DD}/2$) and I_{D_L} (measured at $V_{GS} = V_{DD}/2$ and $V_{DS} = V_{DD}$) [158].

TABLE 6-1
Relative variation of I_{ON} and I_{EFF} of the 35 nm gate length *n*-MOSFET from 1000 I_{DS}-V_{DS} characteristics for $W \geq 2L$.

L x *W*	σ/μ [%] I_{ON} [$I_{D \mid VGS = VDD}$]	σ/μ [%] I_{EFF} [$(I_{D\_H} + I_{D\_L})/2$]
35nm x 35nm	8.4	11.1
35nm x 2(35nm)	5.8	7.7
35nm x 4(35nm)	4.5	5.5

The relative variations (σ/μ) in I_{ON} and I_{EFF} for *n*-MOSFETs with gate lengths of 35 nm, for various device widths, are tabulated in Table 6-1 for comparison. The mean and standard deviation of I_{ON} and I_{EFF} values for minimum-size transistors (35 nm x 35 nm gate area) are extracted from the I_{DS}-V_{DS} characteristics of 200 devices simulated using the Glasgow Atomistic simulator. The mean and standard deviation of I_{ON} and I_{EFF} values for a transistor larger than its minimum-size ($W = n.L$, where n is a positive integer) are extracted from the I_{DS}-V_{DS}

characteristics of 1000 devices simulated using statistical SPICE simulations using the methodology described in Chapter 3. In the presence of RDD, I_{EFF} shows larger variation than I_{ON} by about 30-35%. This is because variation in I_{EFF} is affected by the lightly screened Coloumbic potential fluctuations in weak-inversion ($V_{th} \leq V_{GS} \leq V_{DD}$) whereas variation in I_{ON} is smoothed by the screening from the higher inversion layer carrier fluctuation at high V_{GS}. As a result, variations in inverter intrinsic delay will be larger if I_{EFF} is used instead of I_{ON} in the *CV/I* delay calculation for an inverter. This will be shown to be the case in section 6.3.2 below.

In this section, the switching current trajectories of an ensemble of inverters made of MOSFETs subject to RDD have been presented. Three different *FO/FI* conditions were investigated. The inverters have different switching trajectories depending on the load and input slew rate conditions. It was also shown that the variability of the switching characteristics of an inverter depend on the different *FO/FI* conditions. In the presence of RDD, the relative variation of I_{EFF} is higher than the relative variation of I_{ON}.

6.4 Inverter Timing Subject to Variability

6.4.1 Delay Distribution in 35 nm Devices

We have shown above that the linear regime of the transistor operation plays a significant role in determining the intrinsic delay of an inverter (*FO/FI=1*). Real sub-micron *n*- and *p*-MOS devices (as discussed in Chapter 5) may exhibit different transition from linear regime of operation to saturation. Thus, when designing an inverter, perfectly matching the on-current of both devices will not guarantee a perfect match in the effective drive of the pull-up PMOS and pull-down NMOS transistors, and inverter delays will be different depending on whether the output is transitioning from logic *0* → *1* or *vice versa*. Variations due to RDD will affect this matching, and the statistics will be further complicated by the fact that the PMOS transistors usually exhibit less variation due to their relatively larger width (due to

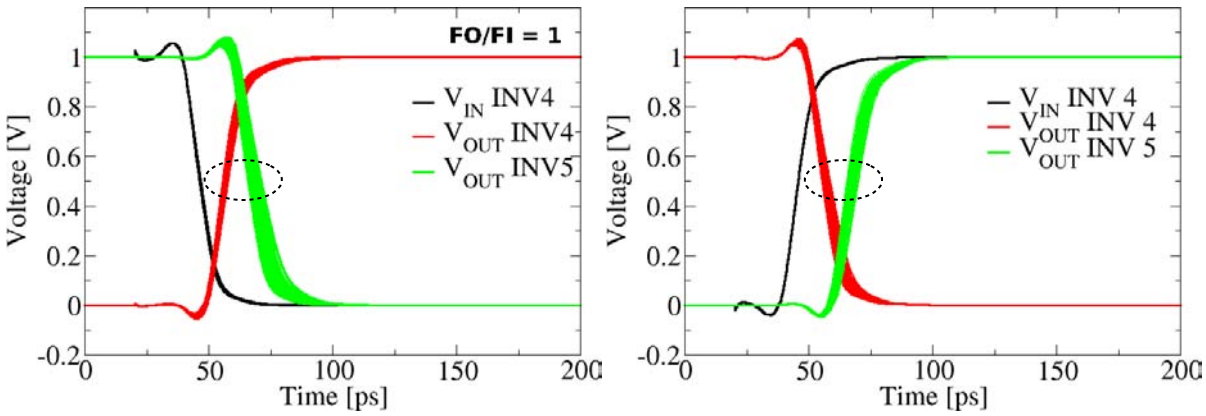


Figure 6-10 : Transient simulation of inverters with *FO/FI* ratio of 1 with falling-input (left) and rising-input (right) transitions applied at the input of INV4.

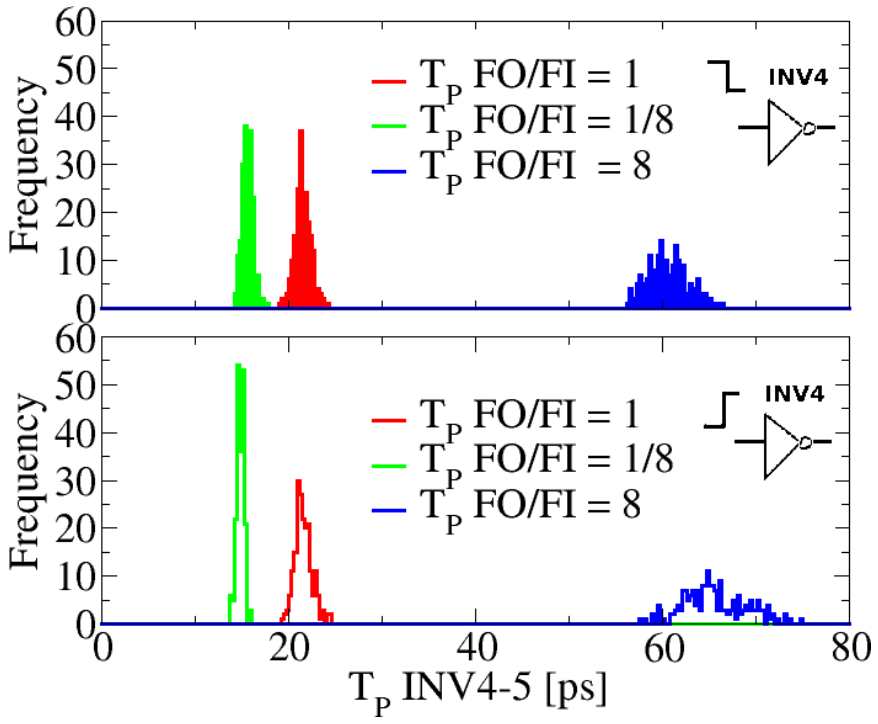


Figure 6-11 : Propagation delay distribution of two subsequent inverters, (from the input of INV4 to the output of INV5) subject to RDD variation during falling-input (above) and rising-input (below) transitions applied at the input of INV4.

the differences in effective mobility of holes and electrons in the active channel region). In order to explore these effects, the propagation delays for rising-output and falling-output transitions at each INV4 and INV5 for ensembles of inverter chains are investigated. To simplify this study, simulated *p*- and *n*-MOS devices are

assumed symmetric allowing the effective drive current of the uniform devices to be easily matched.

Fig. 6-10 shows the transient response of the inverters under observation with *FO/FI* = 1. Although for INV4, the output transit characteristics involved with NMOS discharge during falling-output transition (right figure) will exhibit more variation than its PMOS counterpart during rising-output transition (left figure), the variations in the final output characteristics at INV5 are dominated by the INV5 stage itself. This results in the output voltage of INV5 with logic 0 showing greater variations as seen on the left figure of Fig. 6-10, even though the input transition of INV5 has smaller variations. The analysis of this example emphasises that whole circuits must be considered, with all their interactions, rather than naively considering only separate stages in isolation.

Fig. 6-11 shows the distribution of the total propagation delay, T_P, of the input voltage of INV4 to the output of INV5 with respect to *FO/FI* ratio. As expected, for *FO/FI=1*, the mean value of delay for both *0* → *1* and *1* → *0* output transitions are similar since the devices in INV4 and INV5 are nominally matched. However, the spread (σ) of the total propagation delay distribution is found to be more than 10% larger in the case of the *1* → *0* output transition (top figure in Fig. 6-11), because T_P is dominated by INV5 during falling-output transition (*n*-MOS is discharging) as explained during the discussion of Fig. 6-10. The same observation also helps explain the results for *FO/FI* = 1/8, where the variation (σ) of total propagation delay is also dominated by INV5 (top figure in Fig. 6-11). However, for *FO/FI* = 8, a *smaller* spread (σ) of T_P distribution for the *1* → *0* output transition (top figure in Fig. 6-11) is obtained. In this case the σ of total propagation delay is dominated by INV4 during its *n*-MOS switching. Worst case variation happens for the *FO/FI=8* configuration and σ is 3.3 ps, which is around 5% of mean delay value.

Fig. 6-12 and 6-13 show the distribution of the propagation delays, T_{DHL} and T_{DLH}, of the input voltage of INV4 to the output of INV4 (measured at the 50% points) and input voltage of INV5 to the output of INV5 as a function of *FO/FI*

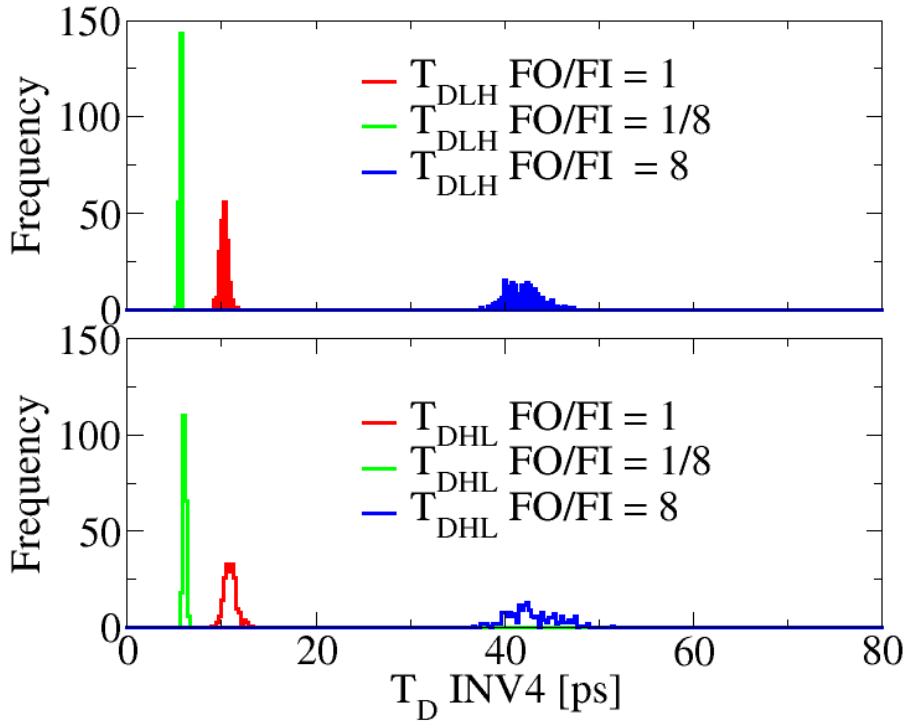


Figure 6-12 : Propagation delay distribution of INV4 during rising-output transition (above) and falling-output transition (below) subject to RDD variation.

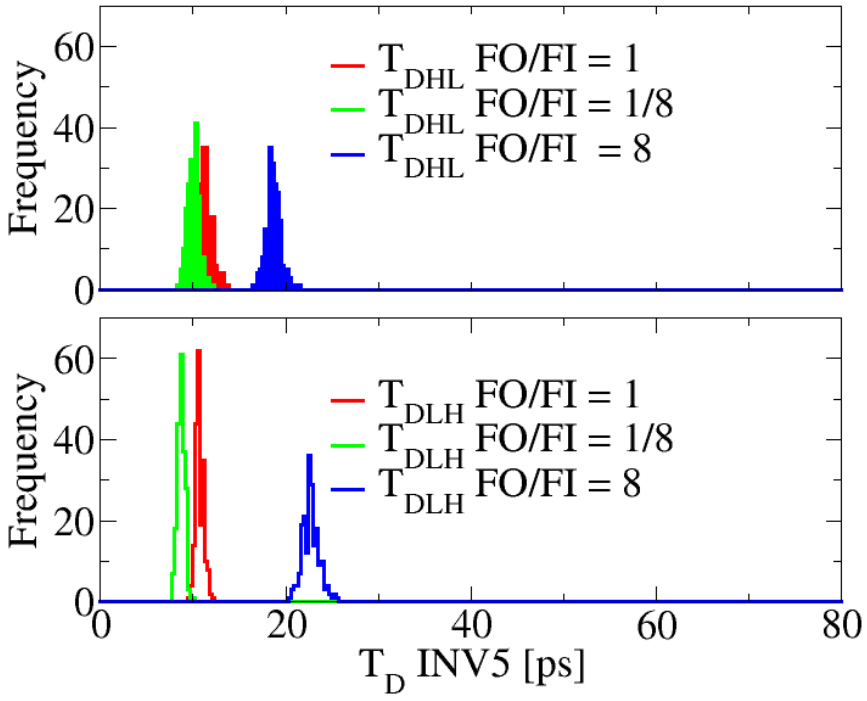


Figure 6-13 : Propagation delay distribution of INV5 during falling-output transition (above) and rising-output transition (below) subject to RDD variation.

ratio. In both figures, the delay variation for T_{DHL} is larger than T_{DLH} because the variations in discharge current during falling-output transition is larger than that for the charging current during rising-output transition.

6.4.2 Delay Variation Approximation

Various models of inverter delay have been proposed in order to capture delay behaviour from the current-voltage characteristics and SPICE simulations [160][161][147][159][158][162][163][164][165][166][167]. In general, the intrinsic delay, τ of an inverter is represented by a *CV/I* metric as shown in Eqn. 6-2 where C_L is the capacitive load, V_{DD} is the supply voltage and I is the drive current in the inverter. In the traditional approximation this drive current is the on-current, I_{ON} but it has been proposed that in sub-micron technologies this drive current should be substituted by using effective-current, I_{EFF} as defined earlier. We also have seen from the previous discussion that by varying the load size (in this study, by varying *FO/FI*), we vary not only the total propagation delay of the inverter but also the propagation delay of the subsequent inverter by changing its input transition time / slew rate. In order to consider this effect, the total propagation delay of an inverter is normally represented by Eqn. 6-3 [196], where the total propagation delay of an inverter, T_{PROP} is the result of addition of intrinsic delay, τ and input transition time, T_T.

$$\tau = \frac{C_L.V_{DD}}{2I} \tag{6-2}$$

$$T_{PROP} = T_T + \tau \tag{6-3}$$

The propagation delay variations with respect to device scaling for different *FO/FI* cases in INV4 during rising-output transition and INV5 during falling-output transition are summarised in this section. Relative variation (σ/u) of the propagation delay (T_{DLH}, T_{DHL}) which is extracted from 1000 inverter chain simulations for different *FO/FI* cases are plotted against the device gate length in the graphs shown in black symbols and line in Fig. 6-14 (*a-f*). The aim of this study is to compare the

relative variation of inverter propagation delays obtained from statistical SPICE simulation with the same results calculated from the relative variation of I_{ON} (shown in red symbols and line in Fig. 6-14) and I_{EFF} (shown in green symbols and line in Fig. 6-14) extracted directly from transistor I_{DS}-V_{DS} characteristics. Since it has been shown in [158] that the intrinsic delay of an inverter (*FO/FI=1*) not subject to intrinsic parameter fluctuations can be best calculated using I_{EFF}, we would like to check if I_{EFF} is also useful when calculating inverter delay variation in the presence of RDD. In addition, inverter delay variation behaviour will be observed and recorded for a number of different *FO/FI* conditions.

I_{EFF} and I_{ON} are extracted from 1000 transistor I_{DS}-V_{DS} characteristics, for each of the transistor widths that are used in the inverter chain under test. Results are obtained for circuits using 35 nm, 25 nm, 18 nm and 13 nm gate length devices. Relative variations (σ/u) of the drive currents are plotted against device gate length, L. As described by Eqn. 6-2, the propagation delay of an inverter is inversely proportional to its drive current to a first approximation. Assuming constant effective capacitance and supply voltage, the variation in propagation delay, σ_τ will be reflected by the variation seen in the drive current of the inverter as shown in Eqn. 6-4 obtained from [168].

$$\sigma_\tau = \frac{\delta\tau}{\delta I}.\sigma_I = -\tau.\frac{\sigma_I}{I} \tag{6-4}$$

Fig. 6-14 (*a*, *c* and *e*) show the relative variation (σ/u) in the delays and drive currents from INV4, when the *p*-MOSFET is switching on, resulting in an output change from 0 → 1. For *FO/FI =1* from (*a*), it can be seen that relative variations of I_{EFF} (green line) overestimates by 14 - 26%, while I_{ON} (red line) underestimates by 4 - 16%, the variation of T_{DLH} (black line) of INV4. Based on this graph, I_{ON} variation reflects the variation in T_{DLH} of INV4 better than the I_{EFF}, even though, as shown in Fig. 6-8 above, the switching current does not spend most of its switching trajectory in saturation.

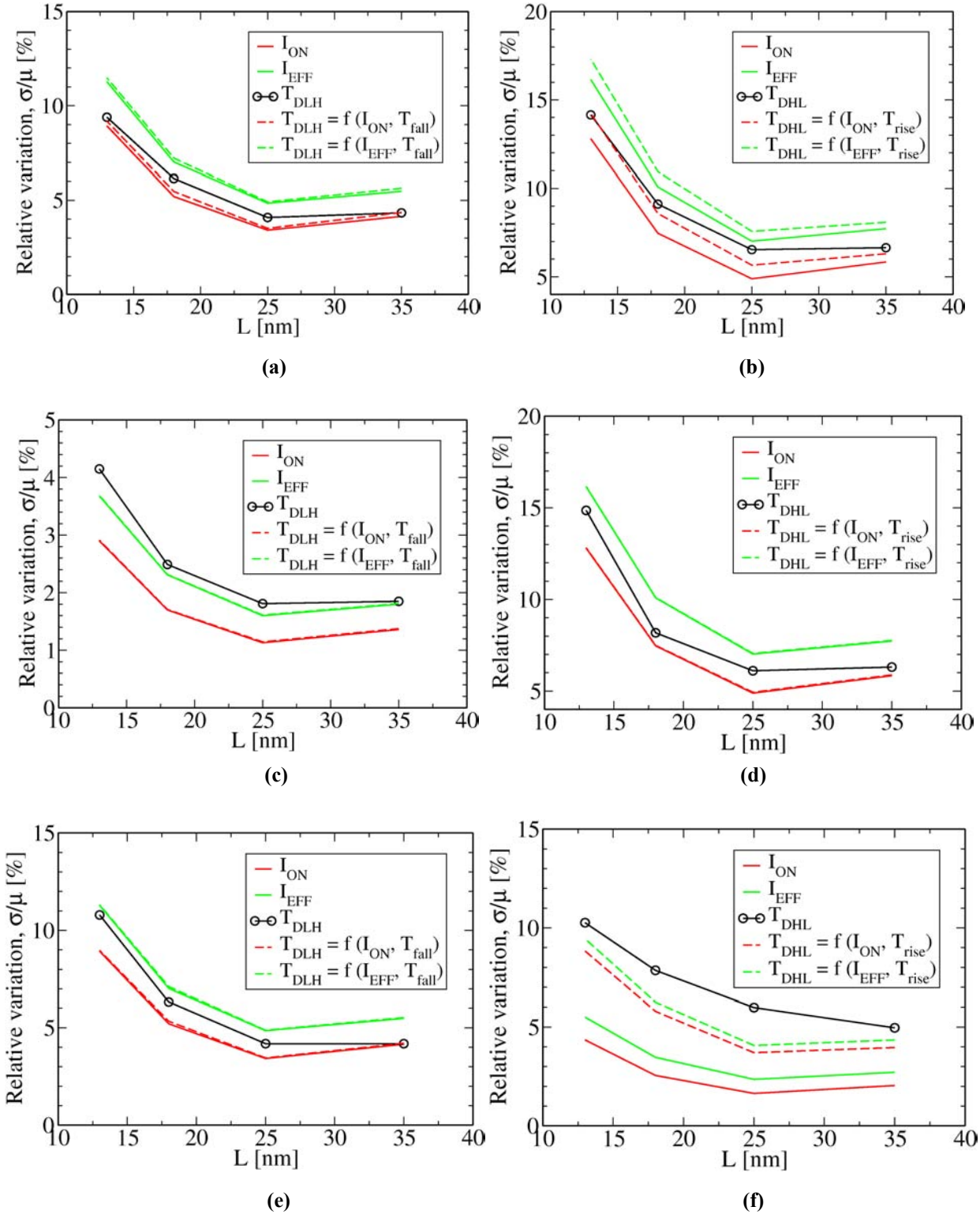


Figure 6-14 : Relative variations of the propagation delay (extracted from simulation and calculated based on Eqn. 6-5) wrt device scaling for INV4 (a, c and e) during rising-output transition and INV5 (b, d and f) during falling-output transition with different FO/FI conditions. (a and b) for FO/FI=1, (c and d) for FO/FI=1/8 and (e and f) for FO/FI=8 configurations.

Note that the relative variation magnitude on the *y*-axis in the Fig. 6-14 (*c*) is about 3 times smaller than in Fig. 6-14 (*a* and *e*). This is because the *p*-MOSFET size implemented in the INV4 is 8 times larger than the other INV4 with *FO/FI* cases. For *FO/FI =1/8* from Fig. 6-14 (*c*), the relative variations of I_{EFF} (green line) underestimates by 2 - 12%, while I_{ON} (red line) underestimates by 26 - 37%, the percentage error of T_{DLH} (black line). In contrast to *FO/FI* = 1, for a large inverter driving a smaller inverter, I_{EFF} variation best captures the variation in T_{DLH} of INV4 (during rising-output transition). Recall that in Fig. 6-8 it was shown that the trajectory of INV4 does not reach saturation under these load conditions.

On the other hand, for a heavily loaded inverter (*FO/FI = 8*) the results of Fig. 6-14 (*e*), indicate that the relative variations of I_{EFF} (green line) overestimate the T_{DLH} (black line) of INV4 by 8.7 - 31%, and I_{ON} (red line) underestimate by 0.7 - 18%. The errors in Fig. 6-14 (*e*) change with device scaling; as gate lengths are scaled below 35 nm, inverter propagation delay variation gradually moves from being close to the I_{ON} curve, towards I_{EFF} being the most accurate estimate for relative variations. This might be due to the contribution of increasing effective capacitance variation seen at the output of inverter with respect to device scaling, which needs to be considered in determining its propagation delay variation. It is common practice to obtain early estimates of MOSFET threshold voltage from *C-V* characteristics [169]. Numerical studies using 3-D simulations [170][171] have shown the effect of RDD on *C-V* characteristics, and the variation seen during the transition from weak to strong inversion is expected to increase as geometries scale. The relative variation of intrinsic gate capacitance will, of course, rise as the intrinsic gate capacitance magnitude reduces with device scaling. In the presence of RDD, not only the effective drive current is subject to variations, but also in the effective gate capacitance seen at the output of an inverter. Hence, its correlation needs to be included in determining variations in circuit propagation delay, especially in the absence of large interconnect components.

The impact of input transition time (slew rate) variation on the total propagation delay variation is now investigated from the INV5 simulations. Input transition time is extracted from the 1000 inverter chain simulations and it is defined by the time taken to switch from 10% to 90% points (or *vice versa*) of the input switching voltage. Assuming the input transition time, T_T is uncorrelated with the intrinsic delay, τ, from Eqn. 6-3, the relative variation in the total propagation delay of an inverter can be represented as shown in Eqn. 6-5.

$$\sigma_{T_{PROP}} = \sqrt{\sigma_{T_T}^2 + \sigma_\tau^2} \tag{6-5}$$

Fig. 6-14 (*b*, *d* and *f*) show the relative variations of delays and drive currents from INV5, when the *n*-MOSFET is switching on with output changing from $1 \rightarrow 0$. Due to RDD, devices will not have identical switching times, and the time it takes to fully charge/discharge their load capacitors will be different. The load capacitors seen at the output of INV4 are themselves not constant, as the input gates of INV5 are subject to variations which can be seen in their gate capacitances. These factors all contribute to a larger variation in the input transition time observed in INV5 in comparison to INV4.

For *FO/FI = 1* from Fig. 6-14 (*b*) the variations in delay calculated from I_{ON} (red line) underestimate the actual T_{DHL} (black line) of INV4 during *n*-MOSFET switching by 9 - 25%. The calculated delay variation as a function of I_{ON} and T_T from Eqn. 6-4 (shown as red dashed line), underestimates T_{DHL} by 0.6 - 6%. This shows that in this case the relative variation of the propagation delay in a balanced inverter can be better estimated by the relative variations of I_{ON} and input transition time, T_T rather than relying only on the relative variation of I_{ON} for inverter with *FO/FI=1*.

For INV5 with *FO/FI = 1/8*, the condition which has the smallest input transition time variation, the calculated relative variation of both T_{DHL} as a function of I_{ON} and $T_{DHL} = f(I_{ON}, T_T)$ show errors of around 7 - 19% as illustrated in Fig. 6-14 (*d*). The observed large deviation of the calculated variation from the extracted propagation delay variation may be due to the overshoot voltage contribution in

determining the variation in the total propagation delay as shown in Fig. 6-9. The overshoot voltage happens because the input switching time is shorter than the transit time of mobile charges in the devices forming the inversion layer in *n*-MOSFET (and to form accumulation layer in *p*-MOSFET) causing the gate-drain capacitances of the inverter which are constant, to couple the change in voltage at its input directly to its output nodes [172][173][174]. In the presence of RDD, the overshoot current during voltage overshoot is subject to variation as discussed in the previous section, thus needs to be considered in calculating the delay variation.

On the other hand, for an inverter with very slow input transition (*FO/FI = 8*), from Fig. 6-14 (*f*), the relative variations of T_{DHL} calculated as either a function of (I_{ON}, T_T) or (I_{EFF}, T_T) show percentage errors of 14 - 38% and 7 - 31% respectively. Large deviations in calculated T_{DHL} variation may be due to the contribution of the short circuit current variations in determining the total propagation delay variation. During slow input switching, there is a direct current path flowing from the supply voltage to the ground through the inverter and the magnitude of this current is directly proportional to the input transition time [165] [166]. This short circuit current prevents the maximum charging/discharging current from flowing through the on-transistor, thus increasing the switching delay. In the presence of RDD, the short circuit current is subject to variation and it cannot be ignored in the calculation of the inverter delay variation.

From this study, a better estimate is obtained for the variation in the intrinsic delay of an inverter (subject to RDD), by considering both I_{ON} and T_T variations. It is shown that I_{EFF} considerably overestimates the delay variation. Under different *FO/FI* conditions, assumptions of one device switching at a time during the rising-output or falling-output transition of an inverter, neglecting the voltage overshoot impact, and load variation effects may introduce larger errors into the estimates of delay variation for circuits composed of deep sub-micron devices.

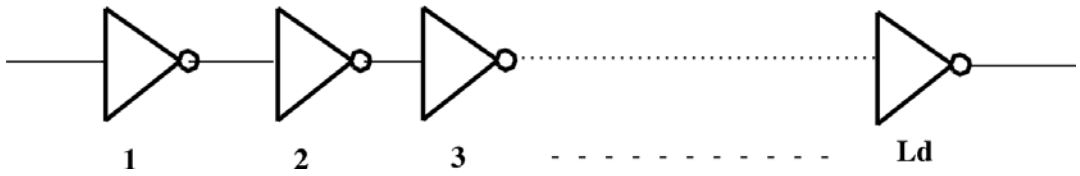


Figure 6-15 : Circuit diagram of a critical path with L_d number stages of inverter. L_d is the logic depth in a critical path.

6.4.3 Critical Delay Variation

Actual digital circuits are not only designed with inverter gates but also with more complex logic gates such as NAND, XOR, etc. These gates are connected to perform logic functions in such a way that the maximum delay in the critical path cannot exceed the maximum period in a given clock cycle specified by a local clock frequency of the chip ($T_{MAX} = 1/f$). It is important to understand that the delay in the critical path, T_{CRIT} of the combinational logic within the combinational logic cloud must not exceed this maximum clock period, T_{MAX} and this requirement ($\boldsymbol{T_{CRIT} \leq T_{MAX}}$) must be met at all times.

Logical effort based design, which calculates the delay inherent in the circuit topology necessary to implement a logical function [175] is often used in designing circuits. This approach is normally used early in design, when access to well characterised standard cells is not available. In this approach, the delay of every primitive gate is assigned a logical effort value which is relative to τ, the intrinsic delay of an inverter driving another inverter in the same technology, in the absence of interconnect parasitics. The depth of any logic path is the delay of that path measured in units of τ, and can be obtained from the logical effort values of each gate in the path. Modern synchronous CMOS systems are designed using register transfer methods, where information is launched from data registers on a rising clock edge, processed or transferred by chains of combinatorial logic to be stored in receiving registers on the next rising edge of the system clock. The logical depths of paths through the combinatorial chains are crucial to the speed of the digital system, and the maximum possible logical depth of such a path (L_d) is found by dividing the system clock period , T_{MAX} by the intrinsic delay, τ [176]. The path which has the

longest delay between two sets of registers in an array of combinatorial logic is the critical path through that combinatorial logic (the longest path/highest number of combinatorial logic gates does not necessarily determine the critical path).

TABLE 6-2
Projection of maximum on-chip local clock for high-performance MOSFETs devices from ITRS 2007.

L [nm]	On-Chip Local Clock [GHz]
35	9.3
25	15.0
18	23.0
13	39.7

In this subsection, the impact of RDD on critical paths will be investigated by considering logical path depths and using the SPICE statistical simulations detailed in Chapter 3. This work is an extension of the *FO/FI* simulations discussed previously. Table 6-2 states the projected maximum clock frequencies for designated technology nodes obtained from the 2007 ITRS (Note: the devices used in this study are designed to follow this scaling trend [177]). Based on this information, the maximum possible logical depth of the critical path in a system with gate length of 35 nm, 25 nm, 18 nm and 13 nm devices are calculated by using Eqn. 6-6, where L_d must be an integer and $T_{CRIT} \leq T_{MAX}$. T_{CRIT} is the delay measured from the 50% of V_{DD} at the input of the first stage inverter to the 50% of V_{DD} at the output of the L_d stage inverter. The inverter intrinsic delay, τ is obtained from an inverter chain simulation with *FO/FI =1* regardless of the inverter size.

$$L_d = \lfloor \frac{T_{MAX}}{\tau} \rfloor = \frac{T_{CRIT}}{\tau} \tag{6-6}$$

Based on the projected maximum logical depth for each technology node, an inverter chain, as shown in Fig. 6-15, is constructed with L_d inverter stages to model such a critical path. The inverter chain is simulated twice: first using minimum-sized inverters (1xINV), and then using inverters 8 times the width of the minimum-sized

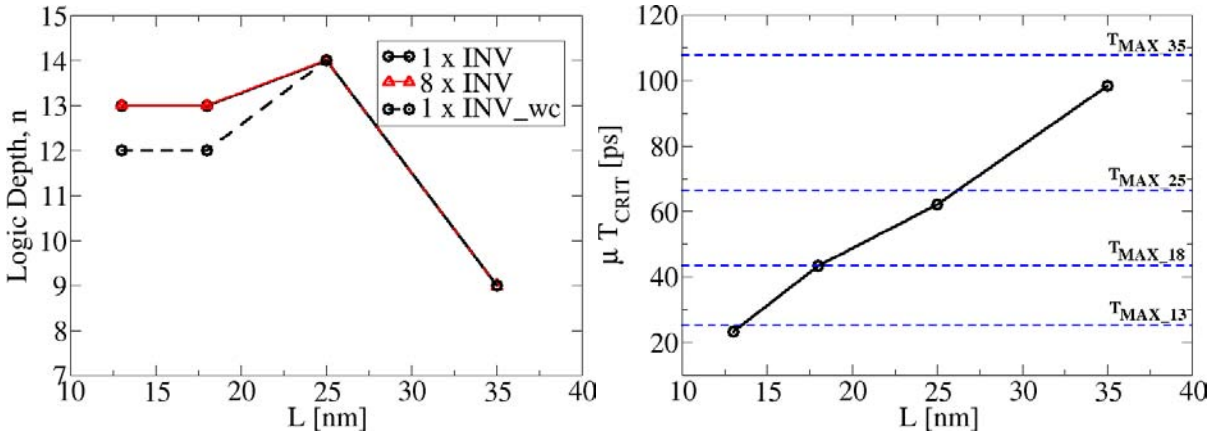


Figure 6-16 : (a) Projection of logic depth, *n* for minimum-sized inverter (1xINV), larger-sized inverter (8xINV) and 3σ worst-case design for 1xINV. (b) Critical delay in a critical path simulated in a chain consists of L_d stages of 1xINV inverter predicted from the left figure and T_{MAX} are also shown for each technology node.

inverters (8xINV). Because the strength of the inverters in each chain match, mean delay through each chain should be approximately identical. However variations in the propagation delays in the two chains will differ due to reduced transistor variation in the wider devices.

Fig. 6-16 (*a*) shows the calculated maximum logical depth in each technology generation from 35 nm to 13 nm gate length. From the figure, the predicted maximum logic depth for minimum-sized inverters and larger-sized inverters (8xINV) are indeed identical. These two results assume identical performance from each of the transistors in the system. With geometry scaling, the logical depth from 35 nm to 25 nm gate length increases from 9 to 14. From 25 nm to 18 nm it decreases by 1, and from 18 nm to 13 nm the logical depth stays constant. Ideally, a constant maximum logical depth in the critical path is desirable when moving from one technology node to another. This is because changes in the maximum logical depth at a new technology node will result in a lengthened design cycle and increased design costs since the design re-use strategy and design optimisations must be re-calibrated, and the logic gates in any possible critical path need to be redesigned at an architectural level to ensure timing specifications continue to be met [178][179].

The average critical delay, T_{CRIT} for each technology node is extracted from 1000 critical path simulations, where all the inverters simulated in the critical path are subject to RDD variation. The extracted mean T_{CRIT} is plotted against the gate length of the devices in Fig. 6-16 (*b*) and the maximum clock period, T_{MAX} for each technology node, as obtained from the 2007 ITRS is also being marked on the graph. From the figure, as we can observe that with geometry scaling, the local clock frequency on the chip increases (T_{MAX} decreases) thus imposing more stringent requirements on the timing specification of high-speed logic. Fig. 6-16 (*b*) also shows the mean of the T_{CRIT} fulfils the $T_{CRIT} \leq T_{MAX}$ requirement in each technology node based on the projected logic depth, L_d for minimum-sized inverter (1xINV) from Fig. 6-16 (*a*). The difference T_{MAX} - T_{CRIT} is essentially a random discretisation effect, but of course the bounds of T_{MAX} - T_{CRIT} will decrease as T_{MAX} decreases.

In the presence of RDD where variations are random across gates in a critical path, and assuming that the distribution of the inverter delay follows the Gaussian distribution, the standard deviation of the critical path, σ_{TCRIT} can be obtained from Eqn. 6-7 obtained from [180].

$$\sigma_{T_{CRIT}} = \sqrt{\sigma^2_{T_{DHL,1}} + \sigma^2_{T_{DLH,2}} + ... + \sigma^2_{T_{DHL/LH,L_d}}} \qquad (6\text{-}7)$$

Based on the calculated standard deviation of the critical path using Eqn. 6-7, the maximum logic depth, L_d for 3-sigma worst-case design ($u_{TCRIT} + 3\sigma_{TCRIT}$) is projected for minimum-size inverter with respect to device scaling. In 3-sigma worst-case design, at least 99.7% of all the critical delay, T_{CRIT} is guaranteed to fulfil the timing requirement. Fig. 6-16 (*a*) shows that for the L_d result of 3σ worst-case design, the projected L_d of 18 nm and 13 nm devices decrease by 1 logic count from the maximum logic depth projected for its nominal design.

Now, we investigate the distribution of these critical delays. This will be done using normal probability plots such as those of Fig. 6-17. Normal probability plot is a graphical method used to quickly assess whether collected samples follow a normal distribution. The *y*-axis of the normal probability plot indicates the probability of finding a sample of the value recorded on the *x*-axis. A straight line

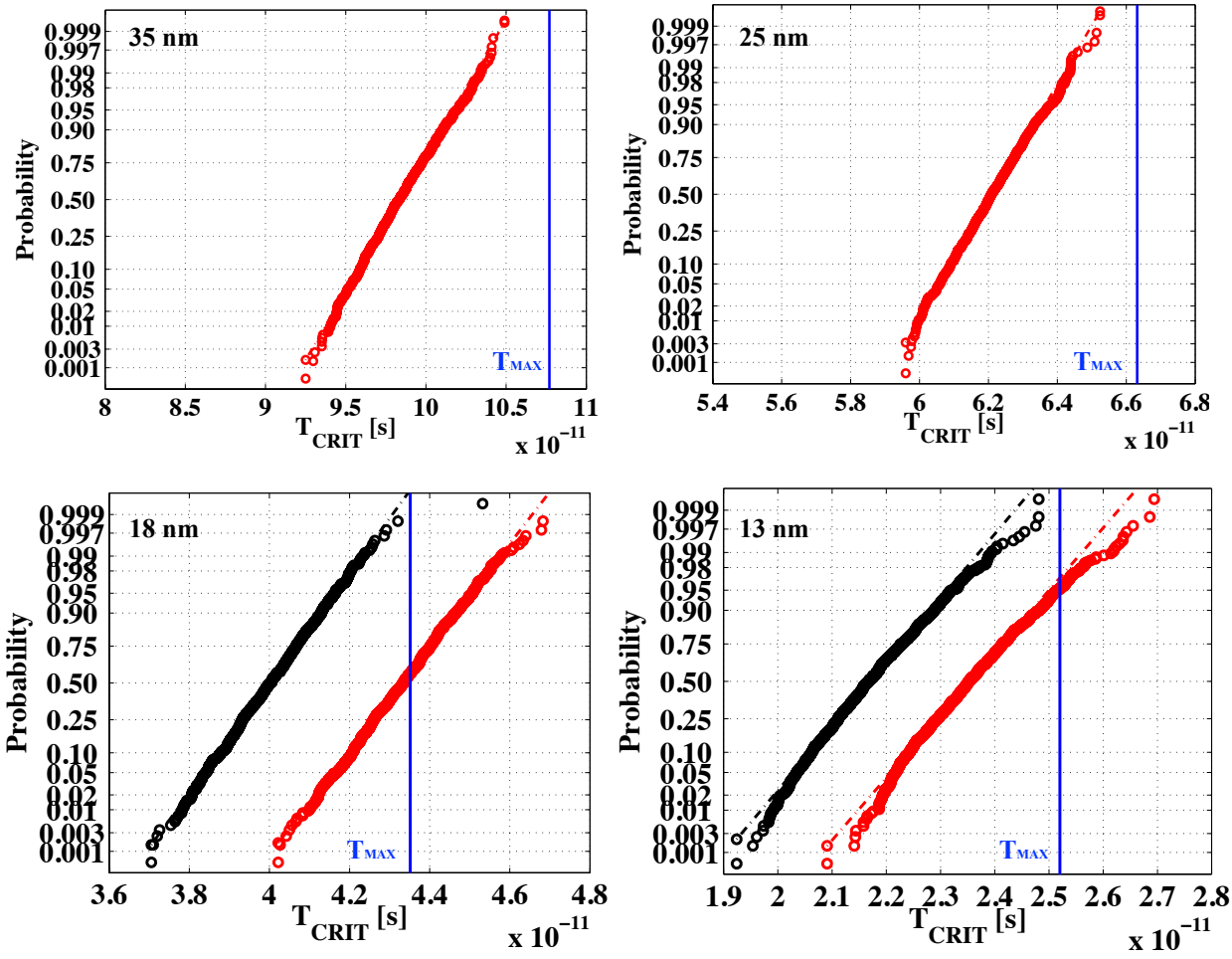


Figure 6-17 : Normal probability plots showing the critical delay distribution in a critical path consists of L_d stages of inverter which are projected from Fig. 6-16 (*a*) for minimum-sized inverter (1xINV) showed by red symbol and black symbol shows critical delay distribution of the L_d -1 stages of inverter when considering 3σ delay variation induced by RDD in the nominal design for 1xINV.

drawn in the normal probability plot indicates a normal distribution, with the gradient of the line proportional to σ. In this study all the normal probability plots are generated using MATLAB.

Fig. 6-17 shows the normal probability plot of the critical delay for chains of minimum-sized inverters in nominal design with L_d stages of inverter (shown in red symbol) and when considering 3σ delay variation induced by RDD in nominal design with L_d -1 stages of inverter (shown in black symbol) for 35 nm, 25 nm, 18 nm and 13 nm gate length devices. T_{MAX} for each technology generation is also marked on the plot. From Fig. 6-17, the mean (probability of 50%) of the critical

delay for minimum-sized inverter (red symbol) in every technology generation is observed to fulfil its timing requirement. For the specified L_d stages of inverter in a nominal design, the *delay margin* which is the delay difference between the mean of the critical delay and T_{MAX} is approximately 9.3 ps, 4.2 ps, 0.2 ps and 1.6 ps for 35 nm, 25 nm, 18 nm and 13 nm devices respectively. In the case of the 18 nm device, even though the average delay for 13 stages of inverter fulfils the $T_{CRIT} \leq T_{MAX}$ specification the *delay margin* is very close to zero. This will impose a great disadvantage to the 18 nm device in the optimisation process of this critical delay at later stage of design cycle and more importantly, this critical path is very susceptible to timing violation in the presence of any type of noise/parameter variation that will lead to the increase in the timing margin. In the presence of RDD, timing violation is observed in the 18 nm (as expected from the previous discussion) and 13 nm devices where only 56.75% and 95.25% of the critical delay lies below the T_{MAX} respectively. Note here that non-normal distribution is observed with the tail of the critical delay distribution deviating from the straight line on the probability plot. By assuming a Gaussian distribution, the estimated critical delay for 13 nm gate length at 3σ value, is 26.0 ps ($T_{CRIT@3\sigma}$ is observed at probability of 99.7% from the normal probability plot). However, in the actual distribution of the critical delay, it shows 0.5 ps larger value for the 13 nm devices.

In the case of 18 nm devices, the best design strategy in ensuring $T_{CRIT} \leq T_{MAX}$ specification can be met in the presence of RDD by reducing the logic depth count by 1 which makes $L_d = 12$. By reducing L_d, the *delay margin* increases by 3.3 ps. Let us assume that there is an area design constraint in the 13 nm devices and thus, to ensure the timing requirement is being met in the presence of RDD the logic depth is decreased by 1 inverter count. The distribution of critical delays at reduced logic depth for 18 nm and 13 nm devices are shown in Fig. 6-17, in black. From Fig. 6-17 of 18 nm and 13 nm gate length devices, the 3σ critical delay $T_{CRIT@3\sigma}$ of the actual and Gaussian distributions did not violate the T_{MAX}.

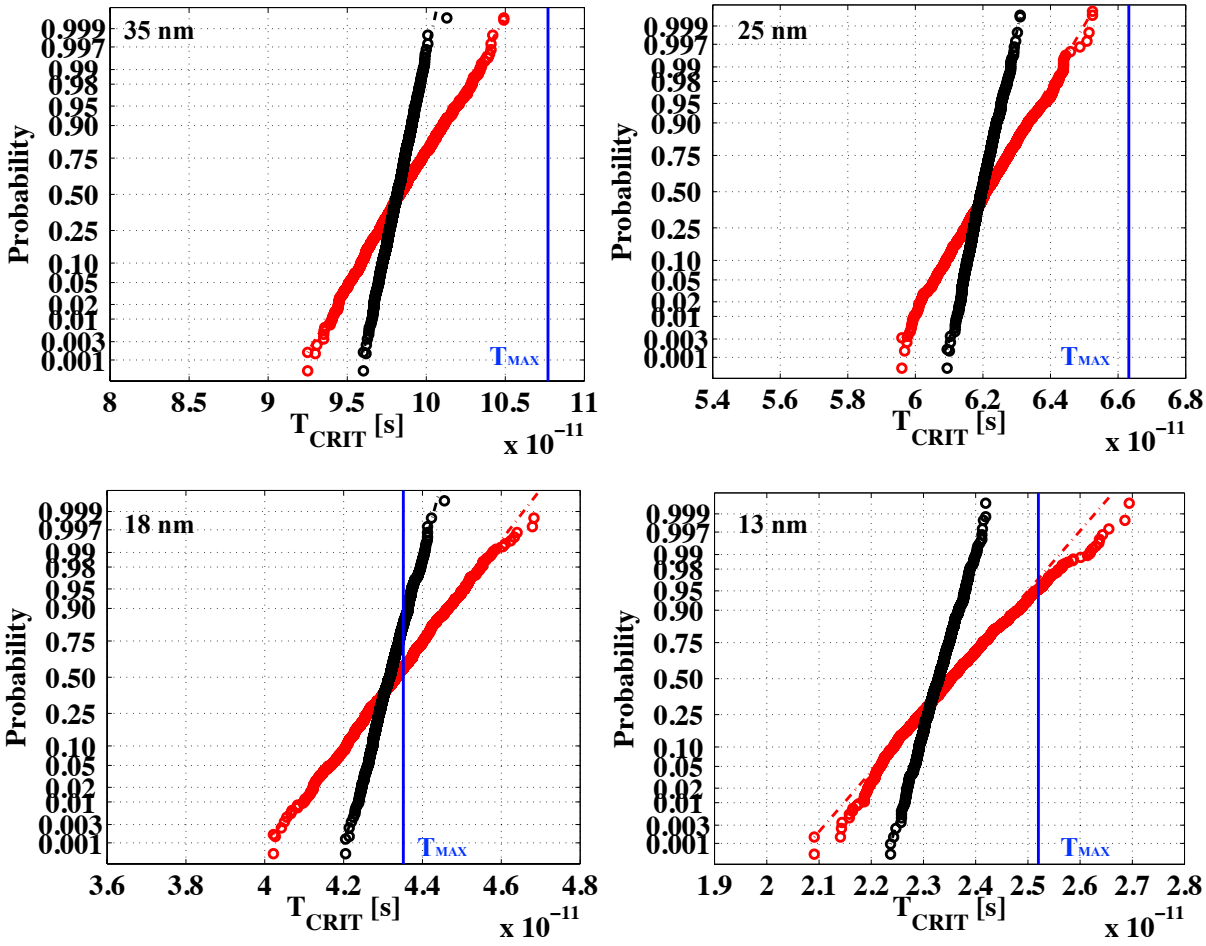


Figure 6-18 : Normal probability plots showing the critical delay distribution in a logic path consists of L_d stages of inverter which are projected from Fig. 6-16 (*a*) for minimum-sized inverter (1xINV) showed by red symbol and larger-sized inverter (8xINV) showed by black symbol.

Optimisation of T_{CRIT} by reducing the critical delay variation subject to RDD can be achieved by increasing the width size of the inverter by 8 times. From Pelgrom's law, the threshold voltage variation subject to RDD is inversely proportional to $\sqrt{W.L}$, thus by increasing the width size of each transistor in the inverter by 8 times, the threshold voltage variation, σ_{Vth} is approximately reduced by 2.8 times for both *n*-MOS and *p*-MOS devices. Fig. 6-18 illustrates the normal probability plot of the critical path for minimum-sized inverter (1xINV) and larger-sized inverter (8xINV) for 35 nm, 25 nm, 18 nm and 13 nm gate length devices. From Fig. 6-18, the mean of the critical delay of larger-sized inverter is approximately 0.15 - 0.29 ps smaller than the mean of the critical delay of the

minimum-sized inverter. This is due to the relative increase in the drive current is slightly unequal to the relative increase in the output load of each wider-sized inverter in the critical path. Even though the percentage difference of the critical delay is approximately 0.2-1%, because of the large reduction in the critical delay variation, σ_{TCRIT} the impact of the small difference in the T_{CRIT} on the critical delay verification may become large. In the presence of RDD, increasing the inverter size by 8 times reduces the critical delay variation, σ_{TCRIT_8xINV} by 2.8, 2.8, 2.9 and 3.1 times the critical delay variation of the minimum-sized inverter, σ_{TCRIT_1xINV} for 35 nm, 25 nm, 18 nm and 13 nm devices respectively. There is a slightly smaller value of $\sigma_{TCRIT8xINV}/\sigma_{TCRIT\_1xINV}$ for the 18 nm and 13 nm devices. This may be due to the contribution of the output load variation, σ_{CL} in the minimum-sized inverter which increases the inverter delay variation, σ_{DHL/LH_1xINV} when subject to RDD in the 18 nm and 13 nm devices as discussed in the previous section, thus directly affecting the variation in the critical delay of the minimum-sized inverter, σ_{TCRIT_1xINV}. In the wider-sized inverter, the contribution of the output load variation becomes smaller and thus, the critical delay variation is dominated by the variation in the drive current when subject to RDD.

In Fig. 6-18 (*18nm*), the timing violation reduces from a 43.25% to 16.11% failure rate in meeting the timing requirement in the critical path when increasing the inverter size by a factor of 8. In the case of the 18 nm design, the inverter size in the critical path needs to be increased further in order to guarantee 100% timing yield, while for 13 nm devices, increasing the inverter size by a factor of 8 guarantees all the devices that are subjected to RDD fulfil their timing requirement. Because there is a 1.01 ps margin between the 100% probability of T_{CRIT} and the T_{MAX}, the inverter size of the 13 nm gate length devices can be reduced to further optimise the design.

In this subsection, we have shown that variability reduces the logic depth count and the critical delay distribution of the minimum-size inverter (1xINV) is non-normal when subject to device scaling. We also have shown that delay

optimisation can be performed by increasing the width of the inverter, which eventually could preserve the logic depth count in the critical path.

6.5 Inverter Power Dissipation Subject to Variability

To complete the analysis of inverter performance, in this section inverter leakage and average power variation will be briefly discussed. The simulation is performed by using minimum-sized (1xINV) and wider-sized (8xINV) inverters for 35 nm, 25 nm 18 nm and 13 nm gate length devices. During the operation of a CMOS inverter, there are 3 sources that contribute to the total power consumption, which are dynamic power, P_{DYN}, leakage power, P_{LEAK} and short circuit power, P_{SCC}. Dynamic power is the power dissipated during charging/discharging of its output load. It is dependent on the total capacitance, C, supply voltage, V_{DD}, switching frequency, f and activity factor, α as shown in Eqn. 6-8. Leakage power is the power dissipated during static mode (no switching activity) and short-circuited power is the power dissipated when there is a direct current flowing from the supply voltage to ground rails during inverter switching.

$$P_{DYN} = C.V_{DD}^2.\alpha.f \tag{6.8}$$

The leakage current obtained from this study considers only the subthreshold leakage current. In small device geometry, there are other mechanisms of leakage current such as gate tunnelling current [181][182] which results from the thinning of gate oxide as a function of scaling, and band-to-band tunnelling which results from abrupt doping profiles in the channel/drain. Both sources can contribute to the total leakage current of an inverter in static operation. However, in this simulation study, the gate tunnelling current and the band-to-band tunnelling current are not being considered.

Fig. 6-19 shows the relative variation, σ/μ of the leakage power for minimum-sized (1xINV) and wider-sized (8xINV) inverters with respect to device gate length while the inset shows average leakage power. The average leakage

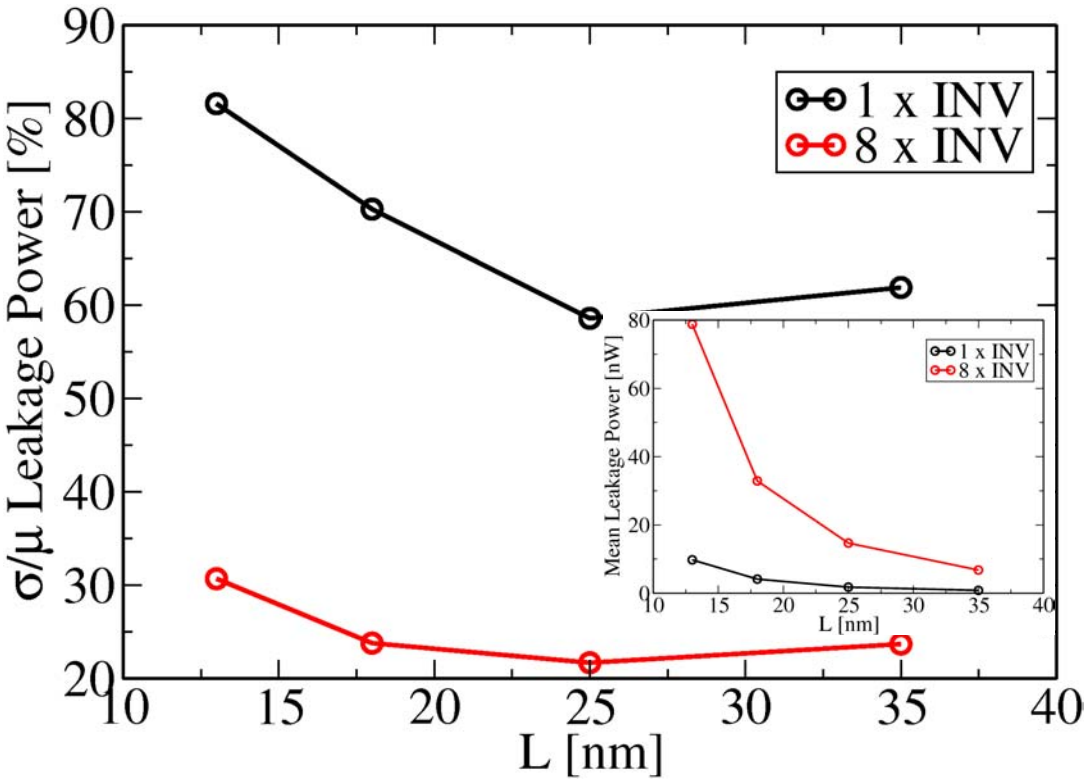


Figure 6-19 : Relative variation of leakage power for different inverter sizes wrt device scaling. Inset showing the mean values of leakage power.

power in both inverters doubles with each successive device scaling. This is due to the increase in doping concentration in the channel to control short-channel effect and decrease in threshold voltage, V_{th} to maintain good voltage overdrive, V_g-V_{th} in the transistor at lower supply voltage values. By increasing the width size by a factor of 8, the mean leakage power of the inverter also increases by approximately 8 times for all gate length devices. This is because the subthreshold current is directly proportional to the gate width of the transistors. In the presence of RDD, the relative variation of the leakage power increases with reduction in gate length as expected. The relative increase is due to the increase in the threshold voltage variation in smaller devices. The relative variation of the leakage power is reduced by approximately a factor of 2 when increasing the *W/L* ratio by 8 times, as expected.

The average power, P_{AVG} is obtained by integrating the power supply current flowing into/out of a minimum-sized inverter with *FO* of 1 and 8 for a full cycle, T = 400 ps as shown in Eqn. 6-9. The calculated average power includes all the

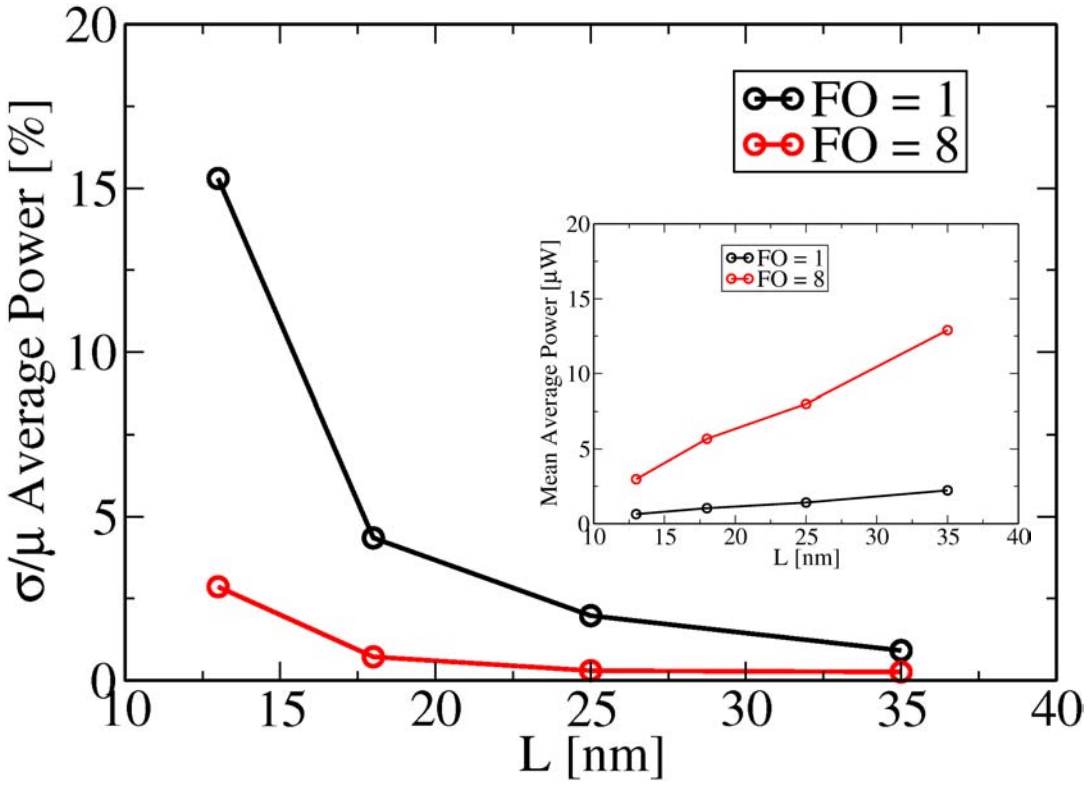


Figure 6-20 : Relative variation of average power for different load sizes wrt device scaling. Inset showing the mean values of average power.

power dissipation sources in the transient operation of an inverter discussed above. In this simulation study, the load size (*FO*) is varied in order to investigate its impact on the average power dissipation of the minimum-sized inverter in the presence of RDD.

$$P_{AVG} = \frac{\int_0^T V_{DD}.I(t)dt}{T} \qquad (6\text{-}9)$$

Fig. 6-20 shows the relative variation of the average power for a minimum-sized inverter with *FO* of 1 and 8 for 35 nm, 25 nm, 18 nm and 13 nm gate length devices. In the presence of RDD, the relative variation of the average power increases with successive device scaling. The relative variation of the average power in the inverter with *FO* = 1 is larger than that of an inverter with *FO* = 8. This is because with larger load size, the transistors in the inverter have to supply/withdraw higher current in order to charge/discharge the load. Inset of Fig. 6-20 shows 5-6 times larger average power dissipation for the inverter with *FO* = 8. In contrast to

the mean leakage power, the mean average power of an inverter with FO = 1 decreases by approximately 1.4-1.7 times when moving to smaller technology nodes due to smaller gate capacitances obtained as a result of geometry scaling. However, in the presence of interconnect components of which does not scale very well in comparison to device scaling [183][184], a larger mean value of the average power is expected at smaller technology nodes.

In this section, leakage power and average power dissipation of an inverter have been discussed. Increasing the inverter width by 8 times, increases approximately 8 times the average leakage power and reduces by half its relative variation in comparison with a minimum-sized inverter. While for an inverter driving 8 times size of load the average power dissipation is 5-6 times higher.

6.6 Summary

In this chapter, the effect of statistical variability introduced by random discrete dopants on the dynamic behaviour of an inverter employing the well scaled 35, 25, 18 and 13 nm gate length bulk MOSFET is presented. The dynamic noise margins, delays and power dissipation of inverters subject to RDD was extensively investigated using three differing fan-out/fan-in conditions which are used to establish realistic input signals and loads in circuits made of the scaled devices. In the first part of this chapter, the dynamic noise margin (DNM) as a measure of the inverter's susceptibility to noise during transients is studied. There is no a standard way of evaluating the DNM consistently while noise immunity curves do not produce a single DNM value therefore it is difficult to compare the DNM for different technologies. In this study, the DNM is obtained by following the maximum square method described in [155] assuming consistent applied noise shape. We showed that scaling lowers the dynamic noise margins by approximately 10% in subsequent technology generations and in the presence of RDD, increases dynamic noise margin variability by 9%, 21% and 57% when scaling from 65 nm to

45 nm, 45 nm to 32 nm and 32 nm to 22 nm technology nodes respectively. Higher output loads and input slew rates improve the noise margins, thus making inverters less susceptible to functional error or delay uncertainty issues caused by the presence of circuit noise. For example, the dynamic noise margin for the 35 nm gate length inverter with *FO* of 8 increases by 28% from the dynamic noise margin for an inverter with *FO* of 1. The relative variation (σ/μ) of the dynamic noise margin of the 35 nm gate length with *FO* of 8 is 0.7% which is 1.9% smaller than the relative variation for the inverter with *FO* of 1. Reduction in the DNM of smaller gate length devices certainly will impose greater danger to the signal integrity and logic functionality of circuits. This is exacerbated by the increase in the variation magnitude induced by RDD in the scaled devices. Although statistical variability can affect the susceptibility of circuits to noise, the effect can be reduced by increasing the output load or input slew rate of the circuit.

The switching trajectories of inverters constructed from 35 nm gate length transistors, under different fan-in and fan-out (*FO/FI)* conditions were simulated and these results used to study the distributions of inverter delay under different conditions of *FO/FI,* load and input slew rate. The FO of 8 inverter with high load has a trajectory that reaches saturation regime at an early stage of active switching, while the introduction of a high slew rate results in a large overshoot at the beginning of the active switching. The inverter with *FO/FI* of 1 has a trajectory that does not spend most of the switching in saturation regime. The distribution of the switching trajectory of the inverters subject to RDD also differs at every switching stage depending on the load and slew rate conditions. This indicates that the load and input slew rate must be evaluated when formulating the statistical delay models. In an inverter chain with *FO/FI* = 1, a reduction of approximately 30% in the rising-output propagation delay variation is obtained in comparison to its falling-output propagation delay as a result of the averaging effect of wider *p*-MOSFET. We have investigated the relative variation in the propagation delay of an inverter against the standard *CV/I* intrinsic delay metric, considering two drive current definitions, I_{ON}

and I_{EFF}. Counterintuitively, we have found that the best estimate of the delay variation in the intrinsic delay of an inverter (*FO/FI=1*) subject to RDD is obtained when using values of I_{ON} and input transition time, T_T variations, rather than using I_{EFF}. This is because the extracted I_{EFF} have higher variability in comparison to I_{ON}. Our estimate gives errors in the range of 0.6-6% for the well-scaled 35 nm, 25 nm, 18 nm and 13 nm devices, a useful practical result for developing statistical delay models that could immediately be incorporated into statistical timing analysis tools.

We also investigated delay variation in more complex circuits ensembles from 35 nm, 25 nm, 18 nm and 13 nm gate length devices subject to RDD. The delay of a circuit critical path modelled by L_d inverter stages is simulated. Depending on the clock system requirement and the intrinsic speed of the inverter, the possible logic depth, L_d is determined. In the presence of RDD, the critical path constructed from minimum-sized inverters shows an increase in the critical delay distribution from 35 nm to 13 nm devices. Large critical delay distribution is observed in 18 nm and 13 nm devices resulting in failure to fulfil 100% its timing requirement. In order to maintain the 18 nm and 13 nm circuit performance, circuit adaptation can be made by increasing the inverter size. However, this results in an increase in circuit size with scaling at the expense of larger power dissipation. Our results also indicate that the adopted statistical simulation tools in this study can quantitatively predict the loss in maximum possible logic depth due to IPFs for any given system and target clock frequency, and that the critical delay distribution of a minimum-size inverter (1xINV) is non-normal when subject to device scaling. Our methodology to predict maximum logic depth, opens the possibility for the development of more accurate delay optimisation tools. The prediction of the distinct non-normality of the critical delay distribution calls into question some simplifying assumptions in present commercial statistical timing analysis toolsets.

Lastly, we have investigated the impact of increasing logic gate size on power dissipation and found that when dynamic and leakage power were taken into account, together with the optimisations required due to component variability, then

increasing the width of an inverter by 8 times increases the average leakage power by approximately 8 times and the average power dissipation by 5-6 times.

Chapter 7

Accuracy Of Standard Cell Characterisation Techniques

7.1 Introduction

Device scaling continues to increase the component count of modern digital circuits and systems. Static timing analysis (STA) has become the common approach to verify timing constraints in full-chip timing analysis with the necessary computational efficiency. Delay calculations based on non-linear delay model (NLDM) look-up tables are widely used in STA approaches. NLDM look-up tables require considerable prior simulation characterisation using tools such as LIBERTY and are based on circuit simulators like HSPICE, ELDO, SPECTRE etc. In NLDM methods gates are characterised based on their load capacitance and input signal slew rate, where the single slew rate / slope parameter is used to capture the influence of complex input waveform shape on the gate delay. Accurately capturing the shape of signal waveforms by using such a single slope (input slew rate) approach is becoming increasingly difficult in the decananometer regime.

In this chapter we study the impact of the slew rate definition on the accuracy of timing characterisation in NLDM format of an inverter, the simplest possible example of a standard cell. Section 7-2 to 7-5 provide on introduction to the subject. In section 7-2 a standard cell is described. In section 7-3, the switching

waveforms which represents the important aspect in determining the accuracy of the cell timing characterisation and abstraction process are discussed and the timing arc and slew definitions are further detailed. Section 7-4 discusses the interconnect load and how it is represented at different stages of design cycle. Section 7-5 describes NLDM and the details of how information is used in a static timing analysis tool to calculate the delay. Section 7-6 presents a delay comparison study between characterised and ramp input waveforms shape of an inverter using different slew rate definitions.

7.2 Standard Cell

A *standard cell* is a basic VLSI building block which implements a logic function, and might be provided to the logic designer by the silicon foundry, or created in-house. A database of cells contains the information (such as functionality, contact geometries and cell sizes) which allows the design process to take place, so that logic functions can be mapped onto a silicon surface. Often cell logic functions are as simple as NAND, NOR, OR-AND-INVERT, *etc.* (although larger standard cells representing, for example fixed width adders, registers or SRAM memory are possible). Cells are arranged and connected to create the complex functionality of a chip. Physically, standard cells have a fixed height, to allow for regular power grids across a chip, but vary in width. Fig. 7-1 (*a,b*) shows the transistor circuit schematic of an inverter and its corresponding layout in a standard cell format at the 65 nm technology node. In the sub-nanometer range, layout design rules have evolved from simple fixed rules into extremely complex sets of fixed and recommended rules [185]. In these recommended rules, layout implementations are recommended in order to guarantee higher yield and reliability after chip fabrication. The standard cell layout for a given logic function at the 65 nm technology node may be considerably different from the layout in an older technology generation, for example at the 0.25 μm technology node. Note that the inverter layout shown in Fig. 7-1 (*b*), consists of 2 poly-silicon tracks indicated by red rectangles overlapping the

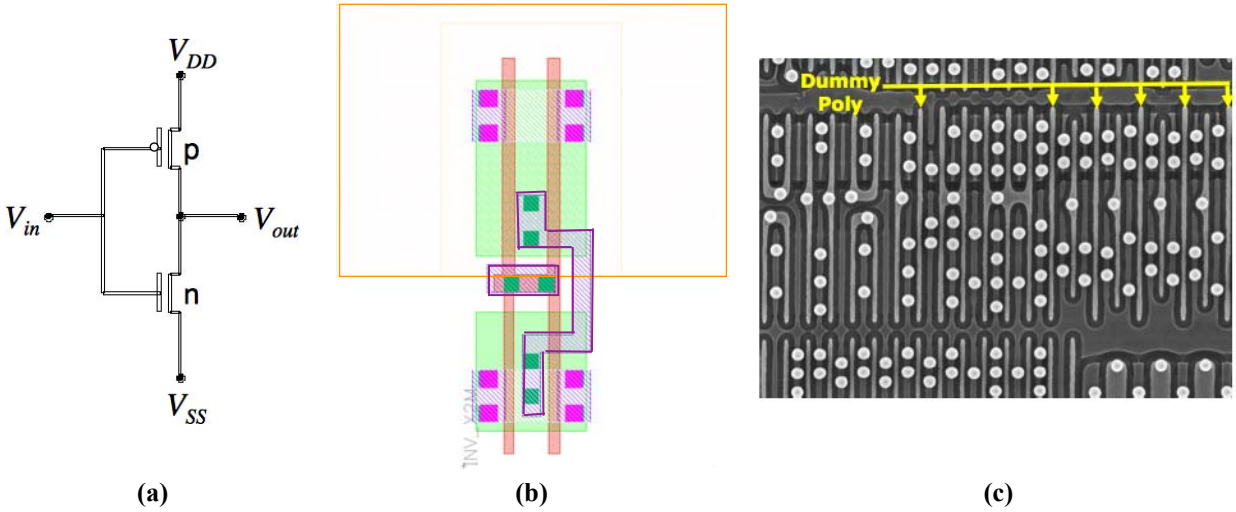


Figure 7-1 : (a) Transistor circuit schematic (b) standard cell of an inverter (c) Logic area in 65 nm AMD Athlon after [186].

active diffusion area to form the *n*- and *p*-MOSFETs. At smaller technology nodes dummy poly-silicon 'gates' are often included, as shown in Fig. 7-1 (*c*), in order to retain a highly regular structure that makes physical fabrication more feasible and to reduce lithography and strain systematic variability [186]. The active *n*+/*p*+ diffusion area is highlighted in Fig. 7-1 (*b*) in light green while the *n*-well which isolates the *p*-MOS from the *n*-MOS transistors in the standard cell is highlighted in orange. Contacts to the diffusion area are highlighted in pink while the contacts to metal1 track are highlighted in dark green. Note that double contacts are applied at every line end enclosure in the design. This is done in order to avoid high *RC* parasitics at the contact which may be exacerbated by manufacturing defects. The metal1 track is highlighted in purple.

In addition to the functional and geometrical information, a standard cell description also includes timing and power estimates for the specified logic function. Timing and power information is based on exhaustively pre-characterised transistor and passive component models and is performed using SPICE circuit simulation. The information is stored as look-up tables in a format that is readable by timing/ power analysis tools. This format may be in non-linear model (NLDM) format, where the timing information is characterised by varying the input slew rate and the

output load capacitance and then stored in a 2-D look-up table; or in a more advanced format such as composite current source (CCS). In CCS, the look-up table stores characterised cell output current-voltage characteristics and cell input load capacitance parameterisations, and the timing information is calculated by the timing analysis tool based on this information for each standard cell interconnection. Whether NLDM, CCS, or any other format is employed, the timing analysis tools then uses the extracted timing information to verify the maximum or minimum delays of logical paths in the chip and flags notifications in an ASCII format timing report if any violations are found. Timing analysis is performed in an incremental manner in the design cycle and depending on the design phase (gate-level simulation, pre-layout simulation, post-layout simulation, *etc*.), the timing information is refined based on circuit information at each stage, and assumptions on the interconnect and clock conditions. There are 2 types of power information stored in a standard cell: leakage and the internal switching power of its specified logic function. Leakage power is the power dissipated when there is no switching activity in a logic cell and the sources of leakage power can be the subthreshold leakage current or tunnelling current through the gate oxide. Internal switching power is related to the internal energy dissipated per transition when there is a switching activity occurring at the input or output nodes of a logic cell. Note that this is not the output switching power, which is related to the output capacitive load, switching frequency and power supply voltage.

Each standard cell in a library is also specified at different operating conditions: typical, fast and slow corners. For the typical corner, the operating temperature of the logic cell is nominal (e.g. 25℃) and the supply voltage is also nominal (e.g. 1 V). While for the fast corner, the temperature is the lowest (e.g. -40 ℃) and the supply voltage is the highest (e.g. 1 V + 10%). At the other extreme, for the slow corner, the temperature is the highest (e.g. 125℃) and the supply voltage is the lowest (e.g. 1 V - 10%). Not only the physical quantities like temperature or voltage are considered in determining the corners, but also process conditions

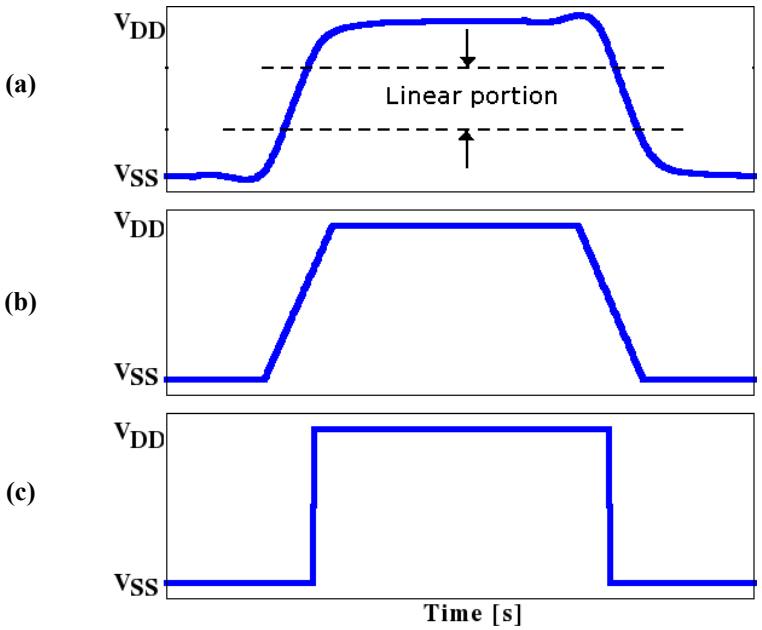


Figure 7-2 : CMOS transient waveforms (a) actual waveform from SPICE circuit simulation (b) approximate waveform used in timing analysis (c) ideal waveform used in timing analysis at higher level of abstractions.

related to manufacturing-induced variations. However, the more advanced IC fabrication becomes, the more factors become important in determining cell timing, the number of process corner increases. This becomes one of the biggest challenges in standard cell characterisation for a single operating point condition.

7.3 Switching Waveform

7.3.1 Timing Arc

Fig. 7-2 (*a*) shows the transient response at the output of a CMOS cell calculated using SPICE circuit simulation. In the switching waveform, the over/ undershoot voltage phenomenon where the waveforms exceed the minimum V_{SS} and maximum V_{DD} values can be clearly seen. A linear portion can also be observed in the middle of the transition waveform. Fig. 7-2 (*b*), shows an approximation to the waveform with a transition time from one logic state to the other. The approximate waveform is represented as a linear ramp during the transition period. Fig. 7-2 (*c*)

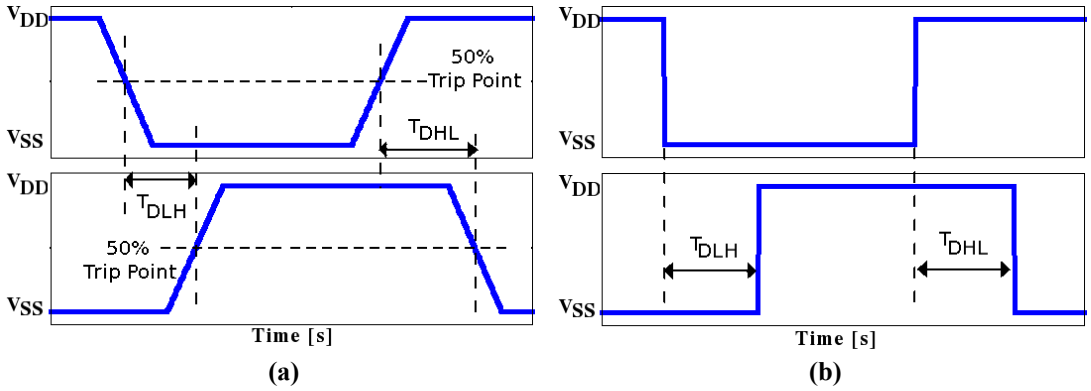


Figure 7-3 : Propagation delay measured at the input to the output transitions (a) using approximate waveforms (b) ideal waveforms.

shows the same waveform using a transition time of 0, that is, a completely idealised waveform.

The propagation delay of a logic cell is determined by measurement from a specific point from the input switching waveform to and equivalent switching level at its output nodes. Fig. 7-3 shows the propagation delay definition for an inverter using approximate waveforms and completely idealised waveforms. In Fig. 7-3 (*a*), the propagation delay of the inverter is defined as the delay measured with respect to 50% of V_{DD} trip points from the input waveform to the output waveform. T_{DLH} is the delay related to the output-rising edge transition from logic-0 to logic-1 while T_{DHL} is the delay related to the output-falling edge transition from logic-1 to logic-0. Fig. 7-3 (*b*) shows the propagation delays measured using the ideal waveforms, where the propagation delay is the delay between the two edges.

The idealised waveform is usually used in higher abstraction levels of design during a timing analysis such as in the gate-level simulation. In digital design, higher levels of abstraction are required to achieve quick timing closure and sign-off. Because delay calculations are critical for timing closure and sign-off throughout the design flow, it is important to generate an accurate library model and use a consistent delay calculation.

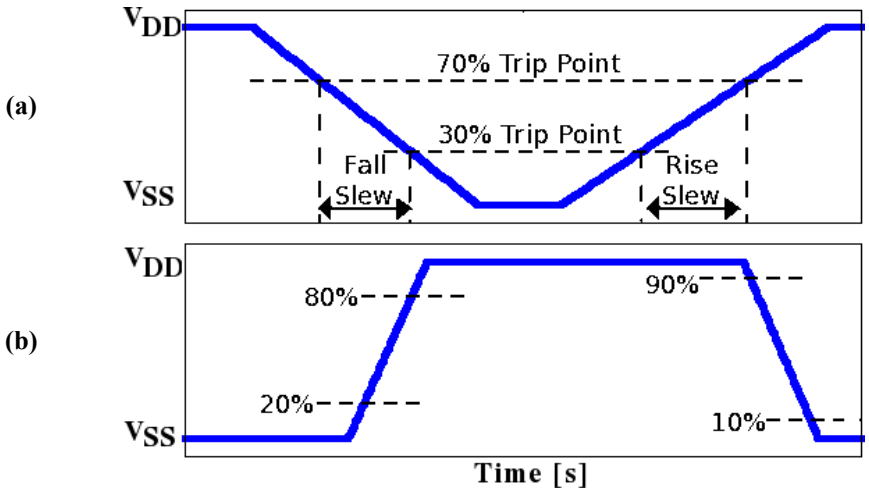


Figure 7-4 : (a) Fall and rise transition times measured at 70% V_{DD} to 30% V_{DD} trip points (b) another examples of slew measurements at 80%-20% and 90%-10% trip points.

7.3.2 Slew

The slew rate is defined as the rate of change in the voltage transition of logic-0 to logic-1 or *vice versa* and is typically measured in terms of transition time. (The transition time is actually inverse of the slew rate.) Different slew rates result in different delay characteristics for a given logic cell.

Fig. 7-4 (*a*) illustrates again approximations to the actual waveform from a logic cell, showing how the slew rate is calculated. As shown in Fig. 7-3 (a), the actual waveform is non-linear at the start and end points, and a choice must be made when extracting the slew as to whether the 'trip points' for measurement are taken at 70% and 30% of V_{DD}, or as shown in Fig. 7-4 (b), at 2080 (20% to 80%) on the rising edge or 9010 (90% to 10%) on the falling edge. Throughout this chapter 1090 or 9010 are used interchangeably, indicating the same trip points but differing in the transition directions.

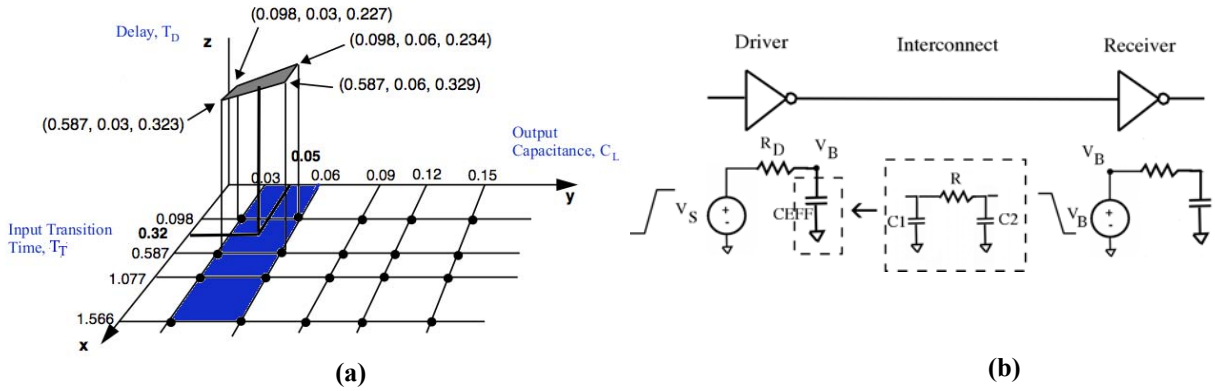


Figure 7-5 : a) Non-Linear Delay Model and interpolation example. b) Illustration of C_{EFF} in the presence of π interconnect model and circuit equivalent model for NLDM timing library implemented in static timing analysis tool.

7.4 Load

The presence of interconnect in a design introduces passive resistance (*R*), capacitance (*C*) and inductance (*L*). The resistance (*R*) component is introduced in the interconnect between the output node of a logic cell to input node of the fanout cells. The capacitive (*C*) component consists of capacitance from the interconnect to the ground, and capacitance between neighbouring interconnect layers. The inductive (*L*) component arises due to current loops and can typically be ignored. This inductive component is important only when considering packaging and board level analysis [187][188].

In the real implementation of a design, accurate interconnect information can only be obtained after the routing process has been completed. An extraction tool is used to extract the detailed parasitics (*RC*) from a routed design. In the absence of physical information related to placement at logical design phase, ideal interconnect can be assumed where *RC* is assumed to be 0. Before placement and interconnect routing it is most useful to identify the logic gates that will contribute to the worst path delays. A wireload model can be applied during pre-layout design stage which provides the estimated *RC* value for an estimated length of interconnect. In this technique, the wireload model provides estimated wire length as a function of cell fanout [189][190].

7.5 Non-Linear Delay Model

In a NLDM cell characterisation process, the propagation delay is not only characterised by varying the input slew but also by varying the output load capacitance. In NLDM, the delay can be interpolated or extrapolated for and specified load capacitance and slew rate from a look-up table. Fig. 7-5 (*a*) shows the graphical representation of the non-linear delay model. The delay (*z*-axis) is shown to be sampled at a few input slew and output capacitance points (*x*- and *y*-axis). The interpolation process is also shown in Fig. 7-5 (*a*) where the cell's delay is obtained from the nearest 4 neighbouring delay points in the table.

However, because the characterised output load in the NLDM is purely capacitive ($R=0$), during the static timing analysis in the presence of a resistive component, an effective capacitance value is estimated in order to consider the effect of resistance on delay. The effective capacitance is found by finding a single capacitance value that is equivalent to the delay of a cell connected to the total *RC* load as shown in Fig. 7-5 (*b*) bounded by dashed-line rectangles. The effective capacitance is then matched to the characterised output load values in the cell library to obtain the cell delay. There are various methods of calculating this effective capacitance during the timing analysis: moment-matching techniques such as Asymptotic Wave Evaluation (AWE) [191], or iteration technique [192][193]. In the iteration technique, the cell's output impedance is estimated and the delay is obtained from the cell's look-up table. Based on these 3 values (input slew, estimated impedance and corresponding cell delay), the charge transferred at the cell's output when using the actual *RC* load is matched with the charge transferred when using the effective capacitance. The iteration continues until the effective capacitance converges in the iteration process [192]. Once the total delay has been obtained from the 2-D look-up table, the input slew of the receiver cell is then approximated. In Fig. 7-5 (*b*), an equivalent circuit model for the driver cell is shown where R_D is the pull-up/pull-down resistance of the standard cell. V_S and V_B are voltage sources with a ramp signal for driver and receiver cells respectively.

Thevenin's theorem is applied to obtain the falling/rising rate of the effective capacitance voltage, V_B by fitting $R_D(T_T, C_L)$ to a polynomial approximation which is then matched to the input transition time of the receiver [194].

In the next section, we study the effect of input slew rates on propagation delays of realistically loaded inverters using HSPICE simulation. The focus of the study is more on the accuracy of tabulating the delay for a single cell (in this case, an inverter) for ultra-scaled devices in a real environment rather than the accuracy of delay calculation in determining the *arrival time* which has been addressed in [194]. The arrival time of a signal is the time elapsed for a signal to arrive at a certain point. Because the accuracy of the arrival time calculation is heavily dependent on the gate delay characterised in the 2-D table, it is important to study the accuracy of the gate delay characterisation process. These simulations are based on 35 nm gate length bulk-MOSFETs (halo-doped) with performance matching the published state-of-the-art 45 nm technology generation, and MOSFETs which are further scaled to 25 nm channel length.

7.6 Inverter Timing Characterisation

In this section, we will present a propagation delay comparison study between inverters subject to realistic transient input signals, and the same inverters subject to ramp input waveforms with slew rates calculated from 9010, 8020, 7030 and 6040 trip point values. The realistic transient input signals will give timing accuracies representative of industrial CCS timing models (in CCS format, the input signal can be of any shape), whereas ramp input signals are used in the industrial characterisation of NLDM propagation delays. 35 nm and 25 nm gate length devices are investigated.

A CMOS inverter with *p*- to *n*-MOSFET gate width ratio of 2:1 and *n*-MOSFET gate width to length ratio of 2:1 is simulated. In order to model realistic input/output conditions, the inverter/cell under test (CUT) is simulated in a 7-stage inverter chain as shown in Fig. 7-6 (*a*). Input voltage and drain current waveforms at

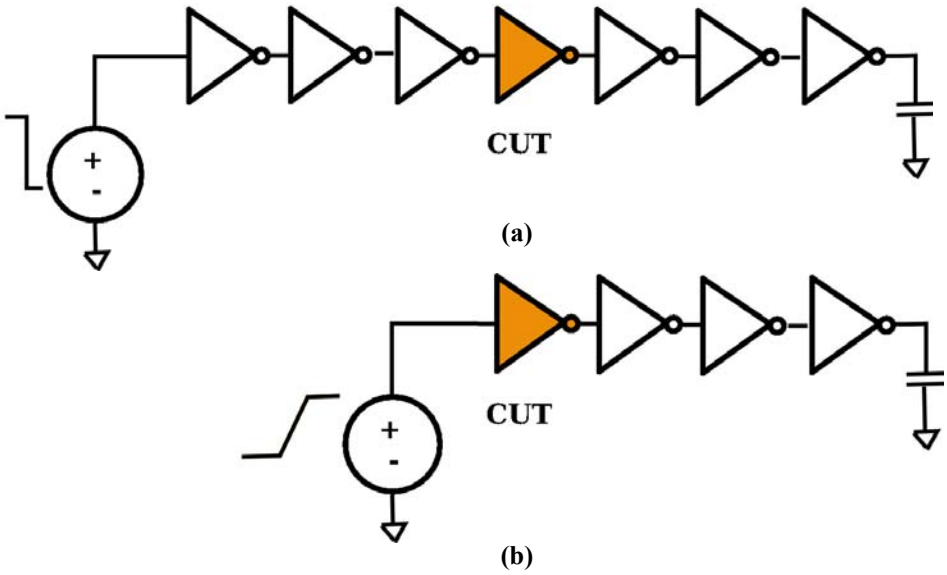


Figure 7-6 : Circuit configurations. (a) 7-stages of inverter chain and the CUT (cell under test) is in the middle of the chain and (b) the CUT is directly connected to a voltage source.

the test inverter are recorded and are referred to as the characteristic waveform throughout this chapter. A 4-stage inverter chain (as shown in Fig. 7-6 (*b*)) with an idealised/linearised input signal is then used to investigate the impact of ramp input signals using various slew rate definitions on the inverter characteristics. Slew rates are calculated using the 90%-10% (9010), 80% - 20% (8020), 70% - 30% (7030), and 60% - 40% (6040) of the supply voltage in the characteristic input waveform, as shown in Fig. 7-7. The propagation delay of the CUT is measured as the time between the input and output waveforms crossing $V_{DD}/2$ with V_{DD} fixed at 1 V. Simulations are performed with balanced inverter drivers and load with both fan out (*FO*) and fan in (*FI*) of 1, a weakly driven, heavily loaded CUT (*FO* = 8, *FI* = 1) and heavily driven weakly loaded CUT (*FI* = 1, *FO* = 8).

The shape of the characteristic inverter waveform is shown in Fig. 7-7. The different values of calculated slew rate extracted using the different slew rate definitions from the previous section are given in Table 7-1. As expected, the 9010 trip points definition results in a smaller slew rate compared to the 6040 trip point definition. For the heavily driven CUT, the input waveform is close to linear at the 7030 and 6040 trip points and the corresponding slew rates differ by only 0.05%.

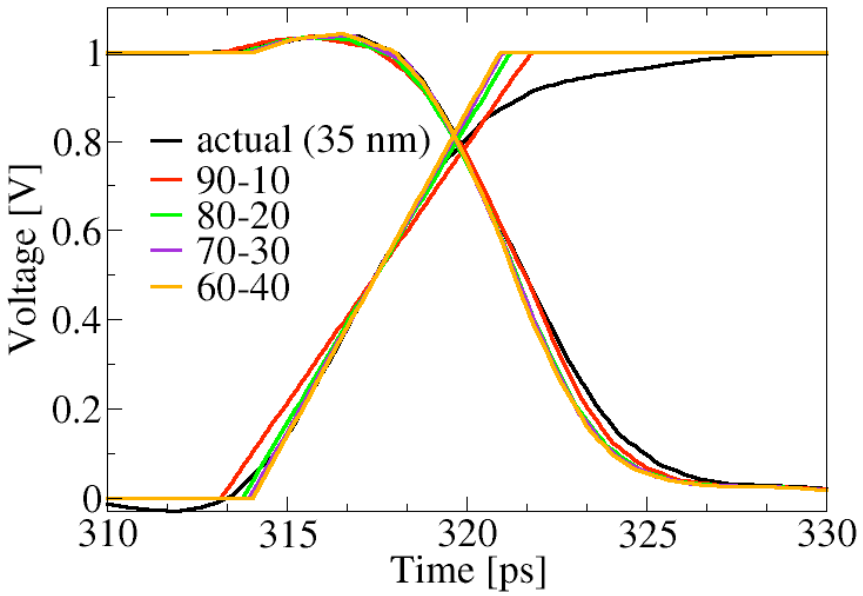


Figure 7-7 : Transient response of an inverter (of 35 nm devices) with balanced driver and load (FO/FI = 1) during falling-output transition.

The difference between the 7030 and 6040 trip points for the heavily loaded CUT is the largest in comparison to the other inverter configurations, due to the large non-linearity in the corresponding characteristics. It should be noticed that the slew rates for the heavily loaded CUT are larger than those of the well-balanced device. This perhaps counterintuitive result is due to the dynamic nature of the loads experienced by these CUTs, and demonstrates the importance of modelling such loads accurately.

TABLE 7-1
Slew rates (*V/ps*) for CUT with 35 nm gate length devices for different trip point cases.

Trip Point	FI = 8	Balanced	FO = 8
9010	0.2004	0.1158	0.1289
8020	0.2289	0.1339	0.1429
7030	0.2455	0.1419	0.1462
6040	0.2456	0.1456	0.1515

Fig. 7-7 shows the transient response during a rising input / falling output transition of a CUT in a balanced inverter chain, and with linearised input signals applied to the 4-stage inverter chain simulation. The linearised input traces are

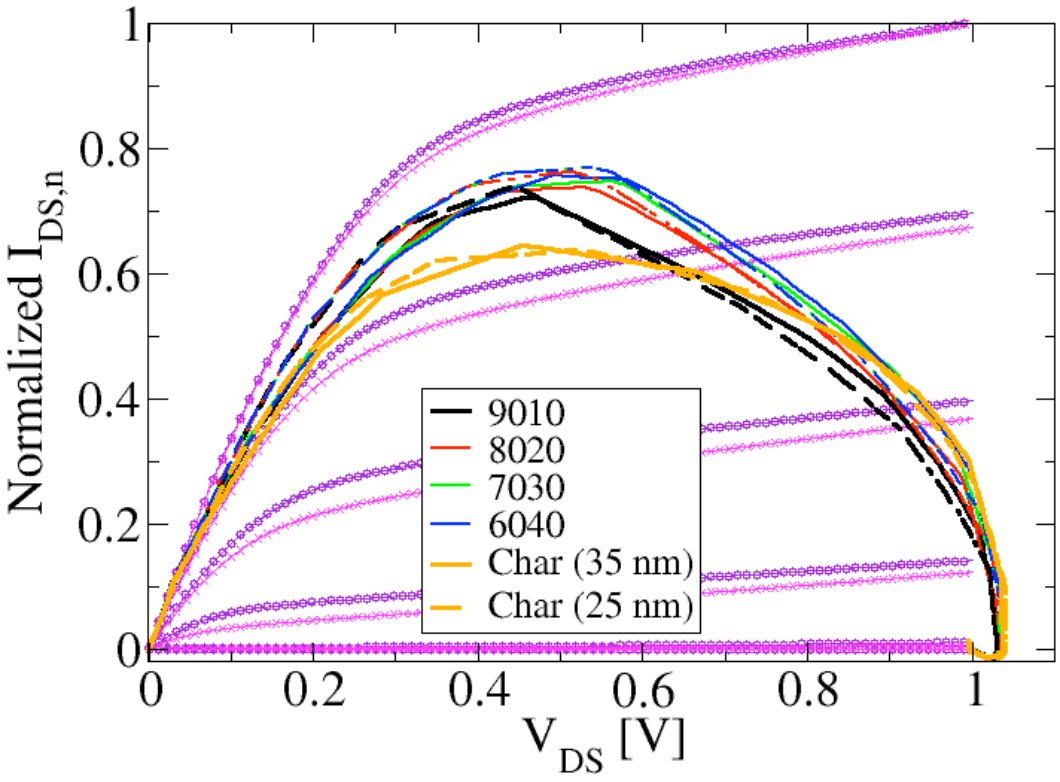


Figure 7-8 : Switching trajectories of an inverter with balanced driver and load (*FO/FI = 1*) during falling-output transition. Also shown are the normalized I_D-V_D curves of the 35 nm (circle symbol) and 25 nm (x symbol) *n*-MOSFET devices.

shifted so that their $V_{DD}/2$ points match the characteristic input waveform. Higher slew rates lead to shorter propagation delays, as can be observed in Fig. 7-7 and supported by the propagation delay, T_{DHL} values in Table 7-2.

TABLE 7-2
Propagation delay, T_{DHL} (falling output transition) of inverter with 35 nm gate length devices.

	FI = 8	Balanced	FO = 8
Char	3.630 ps	4.138 ps	13.251 ps
9010	3.512 ps	4.126 ps	12.759 ps
8020	3.426 ps	3.907 ps	12.753 ps
7030	3.390 ps	3.872 ps	12.656 ps
6040	3.390 ps	3.835 ps	12.616 ps

Fig. 7-8 shows switching trajectories for 35 nm (solid line) and 25 nm (dashed line) transistors with balanced driver and load of *FO/FI = 1* during a high-to-low output transition. The trajectory resulting from the 9010 ramp input waveform underestimates the magnitude of switching current in comparison to the characteristic input waveform from the start point of the trajectory (when V_{DS} = 1 V). At V_{DS} ~ 0.65 V, this becomes an overestimation of the drain current when compared to the characteristic trajectory. Increasing the input slew rate increases the overestimation and reduces the underestimation of the drain current. After the point where the *p*-MOSFET is effectively off, the switching current for all slew rates and the characteristic input waveforms converge to approximately the same values.

A detailed inspection of the transient response of Fig. 7-7 shows that the *p*-MOSFET is effectively turned off (we assume at V_{GS} = 0.9 V) at higher V_{DS} values for the ramp trajectories in comparison to the accurate characteristic trajectories. This explains the smaller propagation delay observed for the inverter at high slew rates and highlights the sensitivity of propagation delay estimations to small changes in the chosen input slew rate, and thus in the trip points chosen to define the slew.

The trajectory shapes and normalised peak drain current values for the 25 nm inverters exhibit approximately the same trends as those found in 35 nm inverters. Table 7-3 shows the propagation delays extracted for inverters using 25 nm devices.

TABLE 7-3
Propagation delay, T_{DHL} (falling-output transition) of inverter with 25 nm gate length devices.

	FI = 8	Balanced	FO = 8
Char	2.855 ps	3.345 ps	10.828 ps
9010	2.752 ps	3.288 ps	10.350 ps
8020	2.681 ps	3.135 ps	10.280 ps
7030	2.660 ps	3.071 ps	10.220 ps
6040	2.589 ps	3.070 ps	10.210 ps

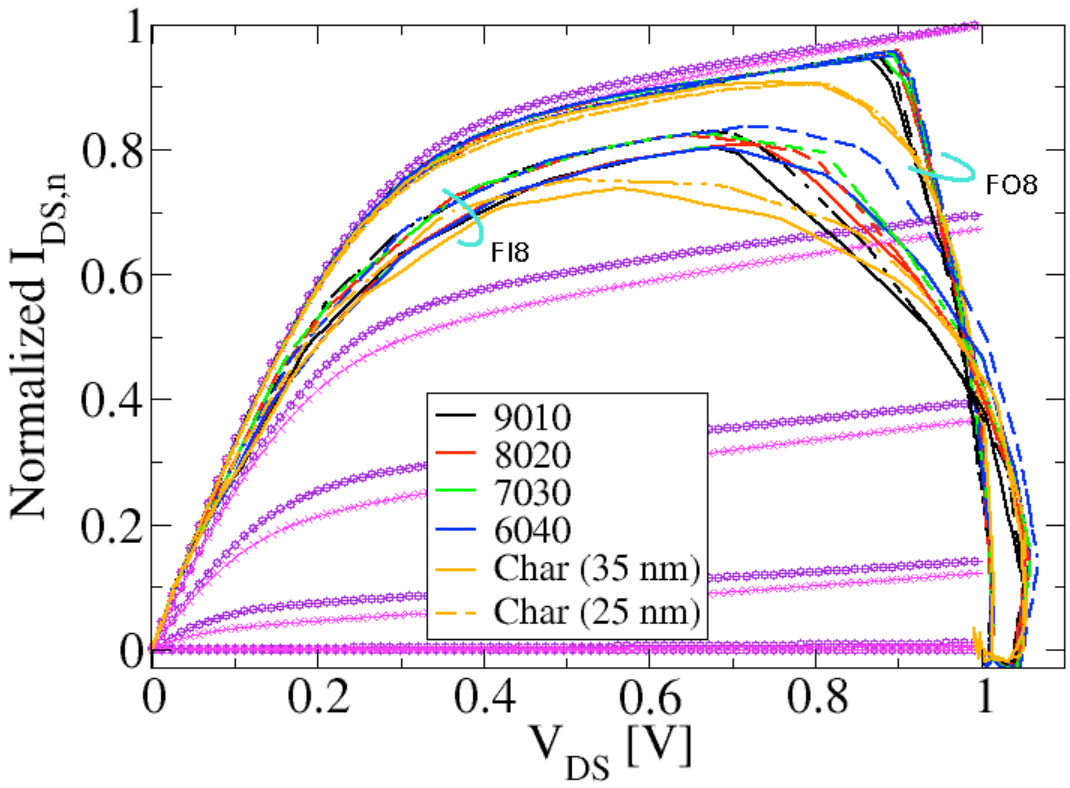


Figure 7-9 : Comparison of switching trajectories of an inverter with unbalanced driver or load (*FO/FI* = 8 and 1/8) during falling-output transition. It is mapped onto the normalized I_D-V_D curves of 35 nm (circle symbol) and 25 nm (x symbol) of *n*-MOSFET devices.

Fig. 7-9 shows the inverter trajectory of the falling-output transition for unbalanced inverter chains. Heavily loaded (*FO=8*) inverters show the highest peak of the drain current, occurring at the beginning of the trajectory (when V_{DS} = 1 V), due to large load sizes. They reach a higher peak current than for balanced inverters. Strongly driven (*FI*=8) inverters have an increased slew rate compared to the balanced inverter chain. Thus, higher switching currents are observed at V_{DS} = 0.2 V compared with the balanced inverter characteristic trajectories. This leads to the shorter propagation delays in heavily driven inverters shown in Table 7-2.

Fig. 7-10 shows the percentage error in propagation delay, T_{DHL} as a result of different definition of the slew rate approximating the CUT input signal using different trip points, and simulating the CUT in a 4-stage inverter chain. The error is calculated in comparison with the characteristic waveforms extracted from a full 7-stage inverter simulation. The error is in the range of 10% and in general, higher

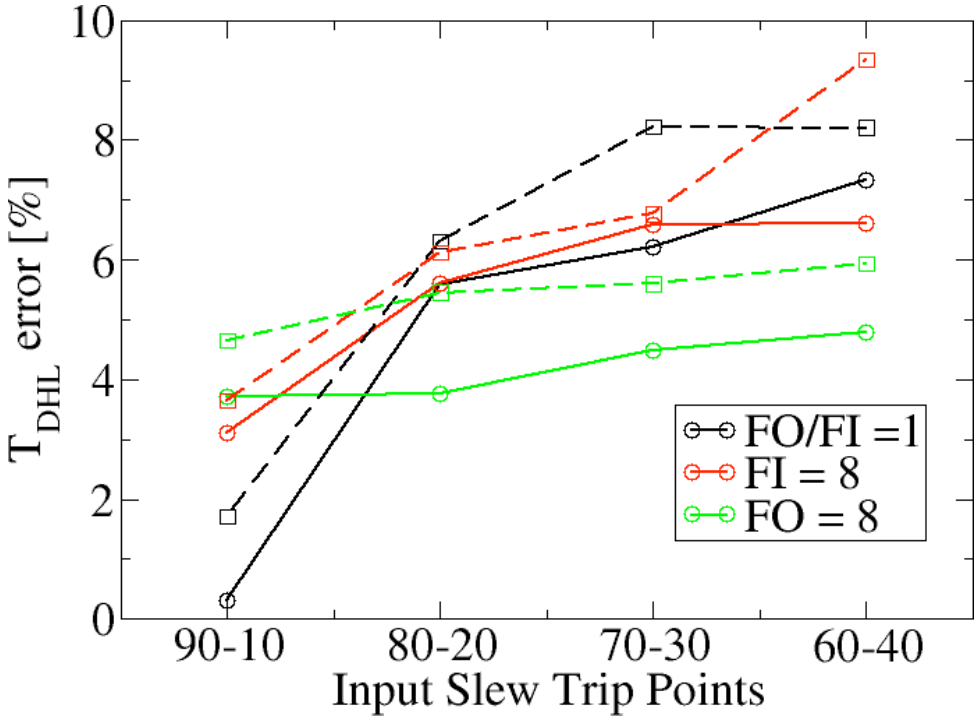


Figure 7-10 : Percentage error of propagation delay, T_{DHL} with respect to input slew trip points. Solid line represents the 35 nm device data and dashed line represents the 25 nm data.

input slew rates produce larger percentage errors due to overestimation of the switching currents as described above. Modelling the characteristic input waveform using an approximated waveform with slew rate equivalent to a linear line tripped at 9010 of the actual waveform did not capture the linear region of the actual waveform accurately, however it gives the smallest percentage error in terms of the propagation delay. This is because the non-linear portion of the actual waveform constitutes of a significant large portion in the voltage swing particularly at the 'tail' as can be observed from Fig. 7-7. Thus, the propagation delay with slew rate at 9010 trip point which samples a proportion of the non-linear region but underestimates the linear region, gives the smallest error due to the errors of the overestimate and underestimate current during the voltage swing cancelling each other out. However, this still leaves the question: What is the best criteria for choosing the ramp during the cell characterisation in order to represent the most accurate delay value in the look-up table.

Fig. 7-10 also shows higher percentage error of the propagation delay in the inverter with larger fan-in or fan-out. This shows that the same trip points to characterise the input waveform for different loading or slew rate conditions cannot be applied at the same inverter. This is because the shape of the voltage swing changes with different fan-in or fan-out conditions as clarified in Table 7-1. Hence the error between the over- and underestimate currents must be re-calculated in order to obtain the trip point value which gives the smallest delay error. This will introduce a 'fudge factor' in the calculation of the arrival time in order to obtain an accurate path delay based on the 2-D delay look-up table characterised by this technique. The fudge factor is required because of the different trip point definitions used to characterise the same inverter at different slew rate and load conditions in the same 2-D table. The inverter with *FO=8* introduces the largest percentage error of T_{DHL} when characterised with a linear ramp taken from the 9010 trip points. However, the percentage error is observed to be less sensitive to the other trip point definitions shown by the smallest increase rate in the percentage error from fig 7-10. This is because the *n*-MOS switching current of the inverter with *FO=8* only starts to change when it has reached the saturation region as shown in Fig. 7-9. Hence, slew rates with different trip point definitions which aim to best capture the linear region of the characteristic waveform, play a smaller role in determining the final inverter delay with large fan-out.

We can also observe the same trend in the percentage error of the propagation delay with scaled devices from fig 7-10 where it increases at every trip point definition. This is due to the different in the I_D-V_D characteristics of the 35 nm and 25 nm gate length devices as shown in Fig. 7-8 and 7-9.

Due to the sensitivities of the gate delay to the shape of the input waveform, characterising the standard cell delay using a single ramp waveform proves to be successively less accurate as scaling proceeds. Also, due to the tighter timing requirements with device scaling, the need for delay accuracy becomes more

important because it is used in the verification of the critical delay of a digital design before the *sign-off* process.

7.7 Summary

In Chapter 7, we examined the accuracy of the standard non-linear delay model (NLDM) for standard cell characterisation of deca-nanometer transistor technologies. In practice, when NLDMs are used, extracted cell propagation times were found to be highly dependant on the definition of the cell input slew rates (for example, whether these are defined from the 10%-90% transition points, or 20%-80% points). For inverters using 35 nm gate length transistors, a 1.77 ps difference in the defined input transition time was found to result in up to an 8% propagation delay error. Sensitivity to the input slew rate value was found to decrease with higher cell load, when the output transition dominates the total propagation delay of the inverter. Cells employing 25 nm gate length devices show up to 4% higher percentage errors compared to their 35 nm counterparts. Due to high sensitivity of the characterised delay to the shape of the input signal of the circuit, we suggest that the NLDM is not suitable for characterising standard cell library of 45 nm technology node and below. This is not only because of the increasing error of the tabulated delay but also due to the deficiency in characterising the delay distribution subject to statistical variability which is critical in ensuring a successful tape-out beyond the 45 nm technology generation.

Chapter 8

Conclusions And Future Work

The aim of the research carried out in this thesis was to study the impact of statistical variability on the statistical analysis of digital circuits. A detailed, predictive study of the impact of variability on foundational CMOS circuits has been carried out, considering devices with gate lengths from 35 nm down to 13 nm. We have investigated ultimate supply voltage limits to circuit operation, circuit noise susceptibility, and the statistical behaviour of timing and power dissipation of these circuits using statistical SPICE simulation. In order to carry out these analyses we have developed statistical simulation and characterisation methodology which can be applied to any small-to-medium scale circuit, and form the foundation of a statistical variability toolkit for statistical timing/power analysis. The tools and methodologies adopted in this study can be easily interfaced with the current industry tools as a result of our use of industry standard compact models in our study.

In Chapter 2, the CMOS scaling and its major bottlenecks were discussed. The device scaling bottleneck of most interest to this work – intrinsic parameter fluctuations (IPFs) caused by random discrete dopants, line edge roughness and oxide thickness variation – was described. IPFs complicate the design and verification processes used to achieve optimum circuit performance and necessitate quantitative timing / power / yield design trade-offs. We described how traditional

methodologies to optimise circuit performance using static timing analysis become less effective post the 65 nm technology node, and showed that techniques which can adequately cope with *statistical* variability in devices are required. The immaturity of present statistical design tools was shown to be an impetus to the aim of this work; to study the impact of statistical variability on digital circuits and develop tools and methodologies to understand this impact.

In Chapter 3, the statistical circuit simulation methodology adopted in this study was described, including: the 35 nm physical gate length devices and simulation tools calibrated and used to provide foundational, predictive device parameters for the tool-chain and the BSIM compact models employed. The template devices are based on state-of-the art 35 nm gate length MOSFET with electrical characteristics that have been calibrated against published data [36][85]. The scaling includes strain-engineered devices and follows the ITRS prescriptions. Using this approach based on calibrated device, gives confidence that the statistical data obtained from the Glasgow Atomistic Device Simulator closely reproduce the actual statistical data of the prototyped devices. The scaled set of transistors were the closest devices that could be publicly used by the group based on close relationship with industrial/research partners which reflect currently manufactured devices in the semiconductor industries and the predicted future-scaled devices beyond the year 2007 - when this research began. Several devices have been used previously in the literature which were unrealistic in terms of their doping profile and structure; and obsolete in terms of technology nodes [79][80]. This has resulted in results that are significantly more realistic than any other work in the field.

The key properties of the 3-D Glasgow Atomistic Simulator also have been discussed, including use of density gradient quantum corrections [92], an essential feature in predicting the correct behaviour of decananometer MOSFETs where quantum effects start to play important role. This simulator captures well the subthreshold regime and threshold voltage of the simulated transistors but underestimates the on current and its variation [90]. This is because the drift-

diffusion method cannot capture non-equilibrium transport effects. The Monte Carlo method is needed in order to capture the real transport behaviour in the decananometer scale transistors. However, simulation of one semiconductor device in order to obtain one current-voltage point takes approximately 2 weeks of simulation time and it is computationally prohibitive for statistical variability studies. There are several device modelling groups which are developing Monte Carlo simulation methods [198][199] but none has successfully applied it for statistical variability studies. At the University of Glasgow some progress have been made in using Monte Carlo simulation for statistical variability studies [91][200] [201] however it is still immature for large scale production simulations. Whilst the augmented drift-diffusion technique we employ does not capture the on-current as well as full Monte Carlo simulation, it is the most accurate and practical technique presently published in the literature.

Next, generation of BSIM ‘atomistic’ compact models was carried out using a 2-stage extraction strategy where in the first stage, a full set of BSIM parameters are extracted based on the uniform device characteristics. In the second stage, 7 parameters are chosen to encapsulate the variation in the electrical characteristics observed in the microscopically different devices subject to statistical variability. In the literature, several attempts have been made to study the impact of statistical variability on circuits by varying parameters in the compact model. However, the approaches are either making an assumption that the distribution of a chosen parameter, e.g. threshold voltage, is Gaussian [142][143][144] or neglect correlations between the chosen device parameters to reflect the underlying physics of statistical variability [78]. Therefore, our approach produces more accurate and predictive result for the aimed technology node as each of the compact model is fitted to 3-D device simulation result subject to statistical variability.

Lastly, the statistical circuit simulation employed in this study has been described. An ensemble of compact models which are macroscopically identical but microscopically different are randomly chosen to be used for the individual

transistor instances in circuit. A practical difficulty with this approach, the generation of wider-sized transistors was discussed and a solution is described. Having the capability to run circuit simulations with the generated model cards, this work enables the transition to a higher level of abstraction which is the characterisation of statistical standard cells. Whilst there are more mature system analysis tools reported in the literature to analyse systems subject to device variability from IMEC [202] the results of this work presently provide the only practical systems analysis methodology to give device accuracy of better than 2% accuracy.

The work described above forms the foundation for the novel results of this thesis.

In Chapter 4, using statistical SPICE simulations, the impact of statistical variability on power supply voltage scaling in digital circuits was investigated. Statistical simulations were performed using the integrated 'atomistic' compact models of well scaled 35, 25, 18 and 13 nm gate length bulk MOSFETs, applying supply voltage levels prescribed by the ITRS. The minimum power supply voltage was evaluated for the ideal case and taking into consideration the safety margins and noise margin. An analytical model for the statistical variability of a CMOS inverter based on a simple model for the saturation current in decananometer scale MOSFETs was presented. The model was validated with respect to statistical circuit simulations of inverters with 35 nm, 25 nm, 18 nm and 13 nm physical gate lengths MOSFETs. The analytical model relates directly the inverter variability to the threshold voltage variability of the underlying MOSFETs. Results of comprehensive physical simulations of the threshold voltage variability of the scaled transistors were used to estimate the gate length dependence of the minimum supply voltage determined by hard logical failures of inverters at chosen design margins. Random Discrete Dopants (RDD), Line Edge Roughness (LER) and Poly Silicon Granularity (PSG) were considered as statistical variability sources in this study. In the

simulations, two scenarios were explored with respect to LER scaling. In the first scenario the LER was scaled according to the requirements of the 2005 edition of the International Technology Roadmap for Semiconductors (ITRS). In the second scenario LER was kept at the present level [110]. For 6σ design margin of a minimum sized inverter, the minimum gate length which allows supply voltages below 1 V is in the neighbourhood of 15 nm, depending on the LER scaling scenario. For larger *W/L* ratios, the supply voltage floor is lower, moving the 1 V floor level to gate lengths of around 10 nm in a scenario which assumes continued LER scaling, and to 14 nm in a scenario which assumes that LER stays the same. Restriction in the supply voltage scaling of future-scaled bulk CMOS devices due to the presence of statistical variability will counteract the advantage of geometry scaling as the dynamic power cannot be scaled any further. The restriction results from the circuit failing to function, in this case, the inverter is unable to invert its input logic level in the presence of statistical variability - not because of manufacturing defects which creates topological changes in the manufactured circuit. Although statistical variability can affect the actual operation of minimum size CMOS devices, this effect can be ameliorated simply by increasing the *W/L* ratio of the logic. However, this technique will reduce the advantages from the scaling in terms of increasing the circuit density. It also increases the output load capacitance and subthreshold leakage current in circuits of which contributes to larger dynamic and static components of power dissipation respectively. In modern digital electronic, especially mobile electronics, circuits not only have to operate correctly, but operate within a timing and power constraints to be commercially viable. The results of this chapter give the circuit designer a simple first order analytical technique to make informed choices balancing device width (and thus circuit size and silicon area) against reliability which can give first order results with minimal computational effort. This is a novel result of this work.

In Chapter 5, the accuracy of the BSIM4 compact model in capturing device characteristics and predicting circuit transient behaviour in SPICE simulation has

been investigated. The compact models of the 35 nm physical gate length MOSFET were benchmarked against 2-D TCAD simulation. The BSIM4 compact model parameters were extracted over a range of device sizes and operating conditions using the compact model extraction tool, Aurora. The corresponding current-voltage and capacitance-voltage characteristics were compared against the current-voltage characteristics obtained from more *ab initio* TCAD simulations. The accuracy of the transient SPICE circuit simulation of an inverter using the extracted BSIM model of the 35 nm MOSFETs was evaluated against mixed-mode TCAD simulations. Excellent agreement between the TCAD and SPICE simulations are obtained for current-voltage characteristics of the MOSFET devices with normalised RMS error less than 6% for various applied gate and drain voltages. The main 5 BSIM model capacitors (C_{gd}, C_{gs}, C_{bs}, C_{bd}, C_{bs}) have been fitted accurately with fitting error below 0.04 fF/μm per sample point. Weaknesses in the BSIM capacitance model were discovered particularly in respect of the drain-to-source capacitance, C_{ds} at high drain bias for both *n*- and *p*-MOSFETs, found to be 1.91 and 1.25 times smaller than the capacitances obtained using TCAD physical device simulation. It was shown that these differences lead to inaccuracy in the transient simulation of the inverter where up to 16% larger falling-output propagation delay was obtained in SPICE simulation compared to the mixed-mode TCAD simulation. However, the percentage delay error reduces to 8.5% if a significant capacitive load (10 times higher than default) is connected at the output of the inverter. Compensation techniques were introduced to better match the SPICE simulated propagation delay against the TCAD simulations leading to 4 times improvement in the SPICE propagation delay accuracy. Although these compensation techniques have little predictive power as devices scale, they will allow far more accurate transient BSIM simulation at any particular technology node, for a relatively small additional characterisation cost. The conclusion of this study is the BSIM4 compact model of the capacitive elements in advanced bulk-MOSFET must be revised in order to

deliver greater predictive power in future scaled-devices resulting in accurate circuit simulations.

In Chapter 6, the effect of statistical variability introduced by random discrete dopants on the dynamic behaviour of an inverter employing the well scaled 35, 25, 18 and 13 nm gate length bulk MOSFET is presented. The dynamic noise margins, delays and power dissipation of inverters subject to RDD was extensively investigated using three differing fan-out/fan-in conditions which are used to establish realistic input signals and loads in circuits made of the scaled devices. In the first part of this chapter, the dynamic noise margin (DNM) as a measure of the inverter's susceptibility to noise during transients is studied. There is no a standard way of evaluating the DNM consistently while noise immunity curves do not produce a single DNM value therefore it is difficult to compare the DNM for different technologies. In this study, the DNM is obtained by following the maximum square method described in [155] assuming consistent applied noise shape. We showed that scaling lowers the dynamic noise margins by approximately 10% in subsequent technology generations and in the presence of RDD, increases dynamic noise margin variability by 9%, 21% and 57% when scaling from 65 nm to 45 nm, 45 nm to 32 nm and 32 nm to 22 nm technology nodes respectively. Higher output loads and input slew rates improve the noise margins, thus making inverters less susceptible to functional error or delay uncertainty issues caused by the presence of circuit noise. For example, the dynamic noise margin for the 35 nm gate length inverter with *FO* of 8 increases by 28% from the dynamic noise margin for an inverter with *FO* of 1. The relative variation (σ/μ) of the dynamic noise margin of the 35 nm gate length with *FO* of 8 is 0.7% which is 1.9% smaller than the relative variation for the inverter with *FO* of 1. Reduction in the DNM of smaller gate length devices certainly will impose greater danger to the signal integrity and logic functionality of circuits. This is exacerbated by the increase in the variation magnitude induced by RDD in the scaled devices. Although statistical variability can

affect the susceptibility of circuits to noise, the effect can be reduced by increasing the output load or input slew rate of the circuit.

The switching trajectories of inverters constructed from 35 nm gate length transistors, under different fan-in and fan-out (*FO/FI)* conditions were simulated and these results used to study the distributions of inverter delay under different conditions of *FO/FI,* load and input slew rate. The FO of 8 inverter with high load has a trajectory that reaches saturation regime at an early stage of active switching, while the introduction of a high slew rate results in a large overshoot at the beginning of the active switching. The inverter with *FO/FI* of 1 has a trajectory that does not spend most of the switching in saturation regime. The distribution of the switching trajectory of the inverters subject to RDD also differs at every switching stage depending on the load and slew rate conditions. This indicates that the load and input slew rate must be evaluated when formulating the statistical delay models. In an inverter chain with *FO/FI* = 1, a reduction of approximately 30% in the rising-output propagation delay variation is obtained in comparison to its falling-output propagation delay as a result of the averaging effect of wider *p*-MOSFET. We have investigated the relative variation in the propagation delay of an inverter against the standard *CV/I* intrinsic delay metric, considering two drive current definitions, I_{ON} and I_{EFF}. Counterintuitively, we have found that the best estimate of the delay variation in the intrinsic delay of an inverter (*FO/FI=1*) subject to RDD is obtained when using values of I_{ON} and input transition time, T_T variations, rather than using I_{EFF}. This is because the extracted I_{EFF} have higher variability in comparison to I_{ON}. Our estimate gives errors in the range of 0.6-6% for the well-scaled 35 nm, 25 nm, 18 nm and 13 nm devices, a useful practical result for developing statistical delay models that could immediately be incorporated into statistical timing analysis tools.

We also investigated delay variation in more complex circuits ensembles from 35 nm, 25 nm, 18 nm and 13 nm gate length devices subject to RDD. The delay of a circuit critical path modelled by L_d inverter stages is simulated. Depending on the clock system requirement and the intrinsic speed of the inverter,

the possible logic depth, L_d is determined. In the presence of RDD, the critical path constructed from minimum-sized inverters shows an increase in the critical delay distribution from 35 nm to 13 nm devices. Large critical delay distribution is observed in 18 nm and 13 nm devices resulting in failure to fulfil 100% its timing requirement. In order to maintain the 18 nm and 13 nm circuit performance, circuit adaptation can be made by increasing the inverter size. However, this results in an increase in circuit size with scaling at the expense of larger power dissipation. Our results also indicate that the adopted statistical simulation tools in this study can quantitatively predict the loss in maximum possible logic depth due to IPFs for any given system and target clock frequency, and that the critical delay distribution of a minimum-size inverter (1xINV) is non-normal when subject to device scaling. Our methodology to predict maximum logic depth, opens the possibility for the development of more accurate delay optimisation tools. The prediction of the distinct non-normality of the critical delay distribution calls into question some simplifying assumptions in present commercial statistical timing analysis toolsets.

Lastly, we have investigated the impact of increasing logic gate size on power dissipation and found that when dynamic and leakage power were taken into account, together with the optimisations required due to component variability, then increasing the width of an inverter by 8 times increases the average leakage power by approximately 8 times and the average power dissipation by 5-6 times.

In Chapter 7, we examined the accuracy of the standard non-linear delay model (NLDM) for standard cell characterisation of deca-nanometer transistor technologies. In practice, when NLDMs are used, extracted cell propagation times were found to be highly dependant on the definition of the cell input slew rates (for example, whether these are defined from the 10%-90% transition points, or 20%-80% points). For inverters using 35 nm gate length transistors, a 1.77 ps difference in the defined input transition time was found to result in up to an 8% propagation delay error. Sensitivity to the input slew rate value was found to decrease with higher cell load, when the output transition dominates the total

propagation delay of the inverter. Cells employing 25 nm gate length devices show up to 4% higher percentage errors compared to their 35 nm counterparts. Due to high sensitivity of the characterised delay to the shape of the input signal of the circuit, we suggest that the NLDM is not suitable for characterising standard cell library of 45 nm technology node and below. This is not only because of the increasing error of the tabulated delay but also due to the deficiency in characterising the delay distribution subject to statistical variability which is critical in ensuring a successful tape-out beyond the 45 nm technology generation.

8.1 Future Work

In the short term there are several lines of research arising from this work which should immediately be followed. First is in 'atomistic' compact model development. In our current approach, wider-sized transistors are represented by square-sized devices connected in parallel. Implementation of width-dependent 'atomistic' compact models directly into the statistical SPICE simulator would be valuable because: 1) device widths of fraction value can be incorporated for design evaluation including statistical variability, 2) there would be a significant reduction in the number of compact device models generated to describe each system, leading to significantly faster SPICE simulation time, and the ability to simulate larger systems.

A second area of research is in the statistical timing and power development tool. From Chapter 6, the distribution of small-scaled devices when subject to statistical variability is shown to be non-Gaussian. Hence, development of non-Gaussian statistical delay and power models should be pursued and implemented in statistical analysis tools to 1) enable incremental statistical timing/power analysis capability 2) obtain faster simulation results of which could save up several CPU hours for large-scale circuit in Monte Carlo simulation approach.

In the long term, the results we have obtained indicate that the industry, over the next 5-10 years should put into place a concerted effort to manage statistical variability because it is becoming a dominant source of variability of circuit performance. This include appropriate training for circuit engineers in mastering statistical design techniques in achieving optimum performance and high yield. Accurate tool development to assess such requirements is needed in order to gain the confidence of the industry to employ it in their design flow.

Appendix A

A.1 Log-Normal Distribution

In probability theory, a log-normal distribution is a probability distribution of a random variable whose logarithm is normally distributed. For example, if Y is a random variable with a normal distribution, the $X = \exp(Y)$ has a log-normal distribution; likewise, if X is log-normally distributed then $Y = \log(X)$ is normally distributed. The mean, E[X] and standard deviation, STD[X] of the lognormal distribution can be derived from the mean, μ and standard deviation, σ values from its natural logarithm as shown in Eqn. A-1 [195].

$$E[X] = e^{\mu+\frac{1}{2}\sigma^2} \text{ and } STD[X] = e^{\mu+\frac{1}{2}\sigma^2}\sqrt{e^{\sigma^2}-1} \quad \text{(A-1)}$$

In Chapter 6, where the discussion of leakage power was made, mean, μ and standard deviation, σ values are presented in Fig. 6-15. These values can be used to calculate its corresponding expected and standard deviation by using Eqn. A-1. In a MOSFET, the threshold voltage is an exponential function of the subthreshold current. Thus a linear variation in the threshold voltage results in an exponential change in subthreshold current. The leakage power distribution is therefore expected to follow a lognormal distribution. Fig. A-1 shows the lognormal probability plot of the leakage power for minimum-sized and wider-sized inverters based on 25 nm devices. The plot verifies that the distribution of leakage power follows a lognormal distribution, although the tail of the leakage power distribution for minimum-sized inverter deviates from a lognormal distribution somewhat.

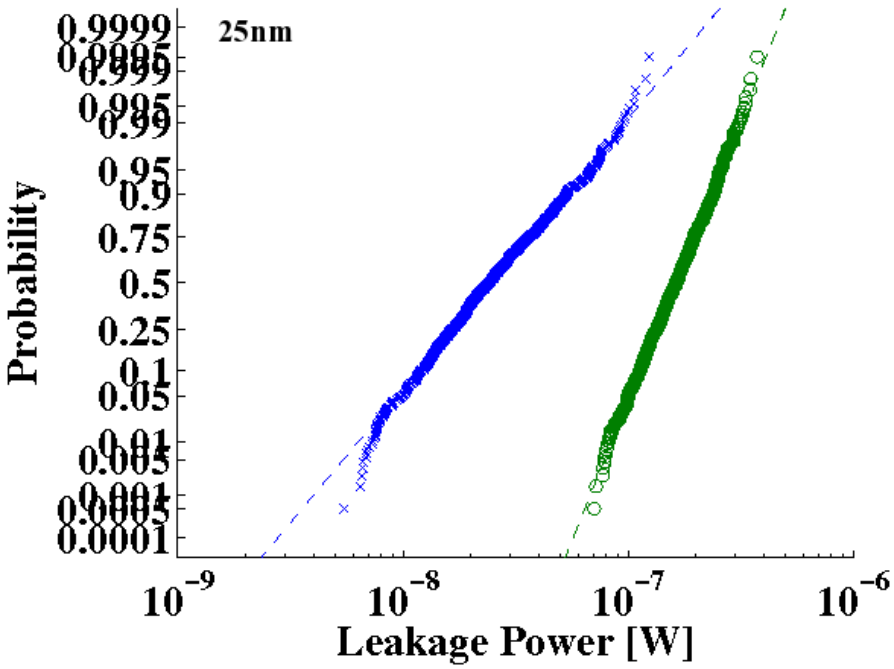


Figure A-1 : Lognormal probability plot of leakage power for minimum-sized (blue symbol) and wider-sized (green symbol) inverter for 25 nm devices.

A.2 Variability Block in HSPICE

In HSPICE, *monte* statement is used to invoke Monte Carlo simulation by varying selected model parameters using Gaussian or uniform distribution. The skewed parameters can be defined with a distribution independently to model global or local variation in circuit. In HSPICE, the global variation is simulated by using common shared model parameters for all the circuit components in a single simulation while in local variation simulation, the model parameters are selected randomly. However, the selected model parameters 1) is not parameterized to tailor the distribution in each device depending on its size and 2) are generated randomly without considering the correlations between the selected parameters.

Bibliography

[1] R. W. Keyes, “Effect of randomness in the distribution of impurity ions on FET thresholds in integrated electronics,” in *IEEE Journal of Solid-State Circuits*, vol. 10, no. 4, pp. 245–247, Aug. 1975.

[2] T. Mizuno, J. Okumtura, and A. Toriumi, “Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET’s,” in *IEEE Transactions on Electron Devices*, vol. 41, no. 11, pp. 2216–2221, Nov. 1994.

[3] K. Takeuchi, T. Fukai, T. Tsunomura, A. T. Putra, A. Nishida, S. Kamohara, and T. Hiramoto, “Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies,” in *IEEE International Electron Devices Meeting, 2007*, pp. 467–470.

[4] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET's: A 3-D “atomistic” simulation study," in *IEEE Transactions on Electron Devices*, vol. 45, no. 12, pp. 2505-2513, Dec. 1998.

[5] 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,” in *IEEE Transactions on Electron Devices*, vol. 53, no. 12, pp. 3063–3070, Dec. 2006.

[6] A. R. Brown, J. R. Watling and A. Asenov, “Intrinsic parameter fluctuations due to random grain orientations in high-κ gate stacks,” in *Journal of Computational Electronics*, vol. 5, no. 4, pp. 333-336, Dec. 2006.

[7] B. Cheng, S. Roy, G. Roy, F. Adamu-Lema, and A. Asenov, “Impact of intrinsic parameter fluctuations in decanano MOSFETs on yield and functionality of SRAM cells,” in *Solid-State Electronics*, vol. 49, pp. 740–746, May 2005.

[8] H. Yamauchi, “A discussion on SRAM circuit design trend in deeper nanometer-scale technologies,” in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 5, pp. 763–774, May 2010.

[9] S. R. Nassif, “Process variability at the 65nm node and beyond,” in *IEEE Custom Integrated Circuits Conference, CICC 2008*, pp. 1–8.

[10] S. R. Nassif, N. Mehta, and Y. Cao, “A resilience roadmap,” Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010, pp. 1011–1016.

[11] S. R. Nassif, A. J. Strojwas, and S. W. Director, “A methodology for worst-case analysis of integrated circuits,” in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 5, no. 1, pp. 104–113, Jan. 1986.

[12] “Altera’s Strategy for Delivering the Benefits of the 65-nm Semiconductor Process,” Altera Corporation, 2006.

[13] G. E. Moore, “Cramming more components onto integrated circuits,” in *Electronics*, vol. 38, no. 8, Apr. 1965.

[14] D. J. Frank, “Power-constrained CMOS scaling limits,” in *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 235–244, Mar. 2002.

[15] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H. J. C. Wann, S. J. Wind, and H.-S. Wong, “CMOS scaling into the nanometer regime,” in *Proceedings of the IEEE*, vol. 85, no. 4, pp. 486–504, Apr. 1997.

[16] E. J. Nowak, “Maintaining the benefits of CMOS scaling when scaling bogs down,” in *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 169–180, March 2002.

[17] S. Thompson, P. Packan, T. Ghani, M. Stettler, M. Alavi, I. Post, S. Tyagi, S. Ahmed, S. Yang, and M. Bohr, “Source/drain extension scaling for 0.1 μm and below channel length MOSFETs,” in *Digest of Technical Papers Symposium on VLSI Technology, 1998*, pp. 132–133.

[18] Kelin J. Kuhn, 2nd International CMOS Variability Conference Lecture, “Variation in 45 nm and Implications for 32 nm and Beyond,” London, 2009.

[19] J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," in *Proceedings of the IEEE*, vol.89, no.3, pp. 259-288, Mar. 2001.

[20] M. A. Quevedo-Lopez, S. A. Krishnan, D. Kirsch, C. H. J. Li, J. H. Sim, C. Huffman, J. J. Peterson, B. H. Lee, G. Pant, B. E. Gnade, M. J. Kim, R. M. Wallace, D. Guo, H. Bu, and T. P. Ma, "High performance gate first hfsion dielectric satisfying 45nm node requirements," in *IEDM Technical Digest, IEEE International Electron Devices Meeting, 2005,* pp. 4 pp.–428.

[21] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS design considerations," in *IEDM Tech. Dig. Papers*, 1998, pp. 789–792.

[22] C. L. Alexander, G. Roy, and A. Asenov, "Random impurity scattering induced variability in conventional nano-scaled mosfets: Ab initio impurity scattering monte carlo simulation study," in *International Electron Devices Meeting, 2006,* pp. 1–4.

[23] J. Welser, J.L. Hoyt, S. Takagi, and J.F. Gibbons, "Strain dependence of the performance enhancement in strained-Si *n*-MOSFETs," in *IEDM Tech. Dig.*, pp.373-376, 1994.

[24] K. Rim, J. Welser, J.L. Hoyt, and J.F. Gibbons, "Enhancement hole mobilities in surface-channel strained-Si *p*-MOSFETs," in *IEDM Tech. Dig.*, pp.517-520, 1995.

[25] O. Semenov, A. Pradzynski, and M. Sachdev, "Impact of gate induced drain leakage on overall leakage of submicrometer CMOS VLSI circuits," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, no. 1, pp. 9–18, Feb. 2002.

[26] T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, "The impact of gate-induced drain leakage current on mosfet scaling," *Electron Devices Meeting, 1987 International*, vol. 33, pp. 718–721, 1987.

[27] T. C. Chen, G. W. Liao, and Y. W. Chang, "Predictive formulae for opc with applications to lithography-friendly routing," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 1, pp. 40–50, Jan. 2010.

[28] Xuemei (Jane) Xi, Mohan Dunga, Jin He, Weidong Liu, Kanyu M. Cao, Xiaodong Jin, Jeff J. Ou, Mansun Chan, Ali M. Niknejad and Chenming Hu, "BSIM4.3.0 MOSFET Model - User's Manual," University of California, Berkeley, 2003.

[29] T. Kanamoto, Y. Ogasahara, K. Natsume, K. Yamaguchi, H. Amishiro, T. Watanabe, and M. Hashimoto, "Impact of well edge proximity effect on timing," in *37th European Solid State Device Research Conference, 2007*, pp. 115–118.

[30] A. Asenov, A.R. Brown, J.H. Davies, S. Kaya, G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs" in *IEEE Transactions on Electron Devices*, vol. 50, no. 9, pp. 1837-1852, Sep. 2003.

[31] J. N. Randall and A. Tritchkov, "Optically induced mask critical dimension error magnification in 248 nm lithography," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 16, no. 6, pp. 3606–3611, Nov 1998.

[32] E. Morifuji, H. Aikawa, H. Yoshimura, A. Sakata, M. Ohta, M. Iwai, and F. Matsuoka, "Layout dependence modeling for 45-nm CMOS with stress-enhanced technique," in *IEEE Transactions on Electron Devices*, vol. 56, no. 9, pp. 1991–1998, Sept. 2009.

[33] B. Smith, "Under Water," *SPIE's OEMagazine*, pp. 22–25, July 2004.

[34] Th. Zell, "Present and future of 193 nm lithography," in *Microelectronic Engineering*, Vol. 83, Issues 4-9, pp. 624-633, Apr-Sep. 2006.

[35] C. Auth, et al., "45nm High-κ + Metal gate Strain-Enhanced Transistors," in *Symp. VLSI Technology*, pp. 128-129, Jun. 2008.

[36] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C. H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz, B. Mclntyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D. Parsons, C. Prasad, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, and K. Zawadzki, "A 45nm logic technology with high-k+metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging," in *IEDM Tech. Dig. Papers,* 2007, pp. 247–250.

[37] S. E. Thompson, M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, T. Ghani, G. Glass, T. Hoffman, C. H. Jan, C. Kenyon, J. Klaus, K. Kuhn, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, B. Obradovic, R. Nagisetty, P. Nguyen, S. Sivakumar, R. Shaheed, L. Shifren, B. Tufts, S. Tyagi, M. Bohr, and Y. El-Mansy, "A 90-nm logic technology featuring strained-silicon," in *IEEE Transactions on Electron Devices*, vol. 51, no. 11, pp. 1790–1797, Nov. 2004.

[38] S. E. Thompson, "Strained Si and the future direction of CMOS," in *Proceedings. Fifth International Workshop on System-on-Chip for Real-Time Applications*, pp. 14–16, 20-24 July 2005.

[39] R.A. Bianchi, G. Bouche, O. Roux-dit-Buisson, “Accurate modelling of trench isolation induced mechanical stress effects on MOSFET electrical performance,” in *IEDM Tech. Dig.*, pp.117-120, 2002.

[40] H. Aikawa, T. Sanuki, A. Sakata, E. Morifuji, H. Yoshimura, T. Asami, H. Otani, and H. Oyamatsu, “Compact model for layout dependent variability,” in *2009 IEEE International Electron Devices Meeting (IEDM)*, pp. 1–4, 7-9 Dec. 2009.

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

[42] Y. Cheng, K. Chen, K. Imai, and C. Hu, “A unified MOSFET channel charge model for device modeling in circuit simulation,”in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 8, pp. 641–644, Aug. 1998.

[43] K. J. Kuhn, “Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS,” in *IEDM Tech. Dig. Papers,* 2007, pp. 471–474.

[44] C.-C. Liu, P. F. Nealey, Y.-H. Ting, and A. E. Wendt, “Pattern transfer using poly(styrene-block-methyl methacrylate) copolymer films and reactive ion etching,” in *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 25, no. 6, pp. 1963–1968, Nov. 2007.

[45] 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,” in *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 433–449, July 2006.

[46] M. Gotoh, K. Sudoh, H. Itoh, and K. Kawamoto, “Analysis of SiO2/Si(001) interface roughness for thin gate oxides by scanning tunneling microscopy,” in *Applied Physics Letters*, vol. 81, p. 430, 2002.

[47] D. Buchanan, “Scaling the gate dielectric: materials, integration and reliability,” IBM Journal Research & Development, vol. 43, p. 245, 1999.

[48] A. Asenov, S. Kaya and J. H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," in *IEEE Transactions on Electron Devices*, vol. 49, pp. 112–119, 2002.

[49] S. Markov, S. Roy, and A. Asenov, “Direct tunnelling gate leakage variability in nano-cmos transistors,” in *IEEE Transactions on Electron Devices*, vol. 57, no. 11, pp. 3106–3114, Nov. 2010.

[50] W.E. Taylor, N.H. Odell, and H.Y. Fan, “Grain boundary barriers in Germanium,” *Phys. Rev.*, Vol. 88, No. 4, pp.867-875, November 15, 1952.

[51] John Y.W. Seto, “The electrical properties of polycrystalline silicon films,” in *Journal of Applied Physics*, Vol. 46, No. 12, pp. 5247-5254, Dec. 1975.

[52] A. R. Brown, G. Roy, and A. Asenov, “Poly-Si-Gate-related variability in decananometer MOSFETs with conventional architecture,” *IEEE Transactions on Electron Devices*, vol. 54, no. 11, pp. 3056–3063, Nov. 2007.

[53] A. R. Brown, N. M. Idris, J. R. Watling, and A. Asenov, “Impact of metal gate granularity on threshold voltage variability: A full-scale three-dimensional statistical simulation study,” in *IEEE Electron Device Letters*, vol. 31, no. 11, pp. 1199–1201, Nov. 2010.

[54] Y. Nakajima, K. Sasaki, T. Hanajiri, T. Toyabe, T. Morikawa, and T. Sugano, “Confirmation of electric properties of traps at silicon-on-insulator (SOI)/buried oxide (BOX) interface by three-dimensional device simulation,” *Physica E: Low-dimensional Systems and Nanostructures*, vol. 24, pp. 92–95, Aug. 2004.

[55] S. Masui, T. Nakajima, K. Kawamura, T. Yano, I. Hamaguchi, and M. Tachimori, “Evaluation of fixed charge and interface trap densities in SIMOX wafers and their effects on device characteristics,” *IEICE Transactions on Electronics*, vol. 78, no. 9, pp. 1263–1272, Sep. 1995.

[56] P. C. Yang, H. S. Chen, and S. S. Li, “Measurements of interface state density in partially- and fully-depleted silicon-on-insulator MOSFETs by a high-low-frequency transconductance method,” *Solid-State Electronics*, vol. 35, pp. 1031–1035, Aug. 1992.

[57] H. Morris, E. Cumberbatch, V. Tyree, and H. Abebe, “Analytical results for the I-V characteristics of a fully depleted SOI-MOSFET,” *IEE Proceedings Circuits, Devices and Systems,* pp. 630–632, Dec. 2005.

[58] T. Ushiki, K. Kotani, T. Funaki, K. Kawai, and T. Ohmi, "New aspects and mechanism of kink effect in static back-gate transconductance characteristics in fully-depleted SOI MOSFETs on high-dose SIMOX wafers," *IEEE Transactions on Electron Devices*, vol. 47, no. 2, pp. 360-366, Feb. 2000.

[59] Ying-Che Tseng; Huang, W.M.; Ilderem, V.; Woo, J.C.S.; , "Floating body induced pre-kink excess low-frequency noise in submicron SOI CMOSFET technology," *IEEE Electron Device Letters*, vol. 20, no. 9, pp. 484-486, Sep. 1999.

[60] J.P. Colinge, “Silicon-On-Insulator Technology: Materials to VLSI, Second Edition,” Kluwer Academic Publishers, 1997, Chapters 4 & 5.

[61] Y.-C. Tseng, W. M. Huang, C. Hwang, and J. C. S. Woo, "Ac floating body effects in partially depleted floating body SOI nMOS operated at elevated temperature: an analog circuit prospective," *IEEE Electron Device Letters,* vol. 21, no. 10, pp. 494–496, Oct. 2000.

[62] S. C. Lin and J. B. Kuo, "Temperature-dependent kink effect model for partially-depleted SOI NMOS devices," *IEEE Transactions on Electron Devices*, vol. 46, no. 1, pp. 254–258, Jan. 1999.

[63] B. H. Calhoun and A. P. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," *IEEE Journal of Solid-State Circuits,* vol. 41, no. 7, pp. 1673–1679, July 2006.

[64] R. Heald and P. Wang, "Variability in sub-100nm SRAM designs," in *Proc. IEEE/ACM International Conference on Computer Aided Design*, 2004, pp. 347–352.

[65] Burnett, K. Erington, C. Subramanian, and K. Baker, "Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits," in *Symp. VLSI Tech. Dig. Tech. Papers*, 1994, pp. 15–16.

[66] M. Khellah, Y. Ye, N. S. Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb, and V. De, "Wordline & bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65nm CMOS designs," in *Digest of Technical Papers Symposium on VLSI Circuits,* 2006, pp. 9–10.

[67] P. Liu, J. Wang, M. Phan, M. Garg, R. Zhang, A. Cassier, L. Chua-Eoan, B. Andreev, S. Weyland, S. Ekbote, M. Han, J. Fischer, G. C. F. Yeap, P.-W. Wang, Q. Li, C. S. Hou, S. B. Lee, Y. F. Wang, S. S. Lin, M. Cao, and Y. J. Mii, "A dual core oxide 8T SRAM cell with low Vccmin and dual voltage supplies in 45nm triple gate oxide and multi Vt CMOS for very high performance yet low leakage mobile SOC applications," in *Symposium on VLSI Technology (VLSIT),* 2010, pp. 135–136.

[68] C. V. Ramamoorthy and H. F. Li, "Pipeline Architecture" in *ACM Computing Survey,* vol. 9, pp. 61-102, Mar. 1977.

[69] K. A. Bowman, X. Tang, J. C. Eble, and J. D. Meindl, "Impact of extrinsic and intrinsic parameter fluctuations on CMOS circuit performance," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 8, pp. 1186–1193, Aug. 2000.

[70] M. B. Srivastava, A. P. Chandrakasan, and R. W. Brodersen, "Predictive system shutdown and other architectural techniques for energy efficient programmable computation," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 4, no. 1, pp. 42–55, Mar. 1996.

[71] D. M. Brooks, P. Bose, S. E. Schuster, H. Jacobson, P. N. Kudva, A. Buyuktosunoglu, J. Wellman, V. Zyuban, M. Gupta, and P. W. Cook, "Power-aware microarchitecture: design and modeling challenges for next-generation microprocessors," in *IEEE Micro,* vol. 20, no. 6, pp. 26–44, Nov. 2000.

[72] P. Watson, "Good Timing: Effective Current Source Modeling is the Future," in *IQ Magazine Online*, pp. 44-46, 2007, www.iqmagazineonline.com/IQ/IQ21/pdfs/IQ21_pgs44-46.pdf

[73] 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 *Proceedings of the European Solid-State Device Research Conference (ESSDERC), 2010*, pp. 130–133.

[74] D. Blaauw, K. Chopra, A. Srivastava, and L. Scheffer, "Statistical timing analysis: From basic principles to state of the art," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 4, pp. 589–607, Apr. 2008.

[75] J. Jaffari and M. Anis, "On efficient monte carlo-based statistical static timing analysis of digital circuits," in *IEEE/ACM International Conference on Computer-Aided Design, 2008. ICCAD 2008*, pp. 196–203.

[76] A. Srivastava, K. Chopra, S. Shah, D. Sylvester, and D. Blaauw, "A novel approach to perform gate-level yield analysis and optimization considering correlated variations in power and performance," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 27, no. 2, pp. 272–285, Feb. 2008.

[77] Harry Veendrick "Chapter 6 : Memories" in *Nanometer CMOS ICs From Basics to ASICs*, Springer, 2008, pp. 306.

[78] X. Tang, V. K. De, and J. D. Meindl, "Intrinsic MOSFET parameter fluctuations due to random dopant placement," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 5, no. 4, pp. 369–376, Dec. 1997.

[79] A. Bhavnagarwala, X. Tang, and J. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, Apr. 2001.

[80] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaassen, "Modeling statistical dopant fluctuations in MOS transistors," *IEEE Transactions on Electron Devices*, vol. 45, no. 9, pp. 1960–1971, Sep. 1998.

[81] P. A. Stolk and D. B. M. Klaassen, "The effect of statistical dopant fluctuations on MOS device performance," in *International Electron Devices Meeting, 1996,* pp. 627–630.

[82] A. Asenov and S. Saini, "Polisilicon gate enhancement of the random dopant induced threshold voltage fluctuations in sub 100 nm MOSFETs with ultrathin gate oxides," in *IEEE Trans. Electron Devices*, vol. 47, pp. 805–812, Apr. 2000.

[83] 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," in *IEEE Design & Test of Computers,* vol. 27, no. 2, pp. 26–35, Mar. 2010.

[84] U. Kovac, D. Dideban, B. Cheng, N. Moezi, G. Roy, and A. Asenov, "A novel approach to the statistical generation of non-normal distributed PSP compact model parameters using a nonlinear power method," in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2010*, pp. 125–128.

[85] S. Inaba, K. Okano, S. Matsuda, M. Fujiwara, A. Hokazono, K. Adachi, K. Ohuchi, H. Suto, H. Fukui, T. Shimizu, S. Mori, H. Oguma, A. Murakoshi, T. Itani, T. Iinuma, T. Kudo, H. Shibata, S. Taniguchi, M. Takayanagi, A. Azuma, H. Oyamatsu, K. Suguro, Y. Katsumata, Y. Toyoshima, and H. Ishiuchi, "High performance 35 nm gate length CMOS with NO oxynitride gate dielectric and Ni salicide," in *IEEE Transactions on Electron Devices*, vol. 49, no. 12, pp. 2263–2270, Dec. 2002.

[86] Fikru Adamu-Lema, "*Scaling and Intrinsic Parameter Fluctuations in nano-CMOS Devices*, " PhD. thesis, University of Glasgow, 2005.

[87] G. L. Vick and K. M. Whittle, "Solid solubility and diffusion coefficients of boron in silicon," in *Journal of The Electrochemical Society*, vol. 116, pp. 1142–1144, Aug. 1969.

[88] Xingsheng Wang, "*Simulation study of scaling design, performance characterization, statistical variability and reliability of decananometer MOSFETs*" PhD. thesis, University of Glasgow, 2010.

[89] U. Ravaioli, "Hierarchy of simulation approaches for hot carrier transport in deep submicron devices," in *Semiconductor Science and Technology*, vol. 13, no. 1, 1998.

[90] C. L. 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," in

IEEE Transactions on Electron Devices, vol. 55, no. 11, pp. 3251–3258, Nov. 2008.

[91] U. Kovac, C. Alexander, G. Roy, C. Riddet, B. Cheng, and A. Asenov, "Hierarchical simulation of statistical variability: From 3-d MC with "ab initio" ionized impurity scattering to statistical compact models," in *IEEE Transactions on Electron Devices,* vol. 57, no. 10, pp. 2418–2426, Oct. 2010.

[92] A. Asenov, G. Slavcheva, A.R. Brown, J.H. Davies, S. Saini, "Increase in the random dopant induced threshold fluctuations and lowering in sub-100 nm MOSFETs due to quantum effects: A 3-D density-gradient simulation study, " in *IEEE Transactions on Electron Devices,* vol. 48, no. 4, pp. 722-729, Apr. 2001.

[93] D. Frank, Y. Taur, M. Ieong, and H.-S. Wong, "Monte Carlo modelling of threshold variation due to dopant fluctuations," in *Digest of Technical Papers Symposium on VLSI Technology*, 1999, p. 169.

[94] T. Yoshinobu, A. Iwamoto, K. Sudoh, and H. Iwasaki, "Scaling of Si-SiO2 interface roughness," in *Journal of Vacuum Science and Technology*, vol. 13, p. 1630, 1995.

[95] S. Goodnick, D. Ferry, and C. Wilmsen, "Surface roughness at the Si(100)-SiO2 interface," *Physical Review B*, vol. 32, p. 8171, 1985.

[96] G. Declerk, "A look into the future of nanoelectronics," in *Symposium on VLSI Technology, Digest of Technical Papers*, 2005, pp. 6–10.

[97] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," in *IEEE Journal of Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, Feb. 2002.

[98] B. Cheng, S. Roy, G. Roy, A. R. Brown, and A. Asenov, "Impact of random dopant fluctuation on bulk CMOSs 6-T SRAM scaling," in *Proc. of 36th European Solid-State Device Research Conference*, 2006, pp. 258–261.

[99] S. Saxena, C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky, and M. Quarantelli, "Variation in transistor performance and leakage in nanometer-scale technologies," *IEEE Transactions on Electron Devices,* vol. 55, no. 1, pp. 131–144, Jan. 2008.

[100] X. Tang, V. K. De, and J. D. Meindl, "Intrinsic MOSFET parameter fluctuations due to random dopant placement," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 5, no. 4, pp. 369–376, Dec. 1997.

[101] 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*, vol. 40, no. 9, pp. 1787–1796, Sep. 2005.

[102] M.-C. Chang, C.-S. Chang, C.-P. Chao, K.-I. Goto, M. Ieong, L.-C. Lu, and C. H. Diaz, "Transistor-and circuit-design optimization for low-power cmos," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 84–95, Jan. 2008.

[103] E. Morifuji, T. Yoshida, M. Kanda, S. Matsuda, S. Yamada, and F. Matsuoka, "Supply and threshold-voltage trends for scaled logic and sram mosfets," *IEEE Transactions on Electron Devices*, vol. 53, no. 6, pp. 1427–1432, Jun. 2006.

[104] S.-W. Sun and P. G. Y. Tsui, "Limitation of CMOS supply-voltage scaling by MOSFET threshold-voltage variation," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 8, pp. 947–949, Aug 1995.

[105] A. Forestier and M. R. Stan, "Limits to voltage scaling from the low power perspective," in *Proc. 13th Symposium on Integrated Circuits and Systems Design,* 2000*,* pp. 365–370.

[106] 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*, vol. 50, no. 5, pp. 1254–1260, May 2003.

[107] M. Lundstrom and Z. Ren, "Essential physics of carrier transport in nanoscale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 49, no. 1, pp. 133–141, Jan. 2002.

[108] A. Asenov, M. Jaraiz, S. Roy, G. Roy, F. Adamu-Lema, A. R. Brown, V. Moroz, and R. Gafiteanu, "Integrated atomistic process and device simulation of decananometre MOSFETs," in *International Conference on Simulation of Semiconductor Processes and Devices,* 2002, pp. 87–90.

[109] G. Roy, F. Adamu-Lema, A. R. Brown, S. Roy, and A. Asenov, "Intrinsic parameter fluctuations in conventional mosfets until the end of the ITRS: A statistical simulation study," *Journal of Physics: Conference Series, vol. 38*, no. 1, pp. 188–191, 2006.

[110] J. Thiault, J. Foucher, J. H. Tortai, O. Joubert, S. Landis, and S. Pauliac, "Line edge roughness characterization with a three-dimensional atomic force microscope: Transfer during gate patterning processes," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 23, no. 6, pp. 3075–3079, Nov. 2005.

[111] M. Nagase, H. Namatsu, K. Kurihara, K. Iwadate, K. Murase, and T. Makino, “Nano-scale fluctuations in electron beam resist pattern evaluated by atomic force microscopy,” *Microelectronic Engineering*, vol. 30, pp. 419-422, Jan. 1996.

[112] W. D. Hinsberg, F. A. Houle, M. I. Sanchez, J. A. Hoffnagle, G. M. Wallraff, D. R. Medeiros, G. M. Gallatin, and J. L. Cobb, “Extendibility of chemically amplified resists: another brick wall?,” in *Advances in Resist Technology and Processing XX*, vol. 5039, (Santa Clara, CA, USA), pp. 1–14, SPIE, Jul. 2003.

[113] 2005 International Technology Roadmap for Semiconductors, http://public.itrs.net.

[114] A. V.-Y. Thean, Z.-H. Shi, L. Mathew, T. Stephens, H. Desjardin, C. Parker, T. White, M. Stoker, L. Prabhu, R. Garcia, B.-Y. Nguyen, S. Murphy, R. Rai, J. Conner, B. E. White, and S. Venkatesan, “Performance and variability comparisons between multi-gate FETs and planar SOI transistors,” in *IEDM Tech. Dig. Papers*, 2006, pp. 1–4.

[115] R. Vaddi, S. Dasgupta, and R. P. Agarwal, “Device and circuit co-design robustness studies in the subthreshold logic for ultralow-power applications for 32 nm CMOS,” *IEEE Transactions on Electron Devices*, vol. 57, no. 3, pp. 654–664, Mar. 2010.

[116] N. Sugii, R. Tsuchiya, T. Ishigaki, Y. Morita, H. Yoshimoto, and S. Kimura, “Local V_{TH} variability and scalability in silicon-on-thin-box (SOTB) CMOS with small random-dopant fluctuation,” *IEEE Transactions on Electron Devices*, vol. 57, no. 4, pp. 835–845, Apr. 2010.

[117] T. Ohtou, N. Sugii, T. Hiramoto, "Impact of Parameter Variations and Random Dopant Fluctuations on Short-Channel Fully Depleted SOI MOSFETs With Extremely Thin BOX," *IEEE Electron Device Letters,* vol. 28, no. 8, pp. 740-742, Aug. 2007.

[118] R. Tanabe, Y. Ashizawa, and H. Oka"Investigation of SNM with Random Dopant Fluctuations for FD SGSOI and FinFET 6T SOI SRAM Cell by Three-dimensional Device Simulation," in *International Conference on Simulation of Semiconductor Processes and Devices,* 2006, pp. 103-106.

[119] A. Asenov, A. R. Brown, G. Roy, B. Cheng, C. Alexander, C. Riddet, U. Kovac, A. Martinez, N. Seoane, S. Roy, “Simulation of statistical variability in nano-CMOS transistors using drift-diffusion, Monte Carlo and non-equilibrium Green’s function techniques,” *Journal of Computational Electronics,* vol. 8, no. 3, pp. 349-373, Oct. 2009.

[120] D. S. Kung, and R. Puri, “Optimal P/N width ratio selection for standard cell libraries,” in *IEEE/ACM international Conference on Computer-Aided Design*, 1999, pp. 178-184.

[121] “TSMC 0.18μm Process 1.8-Volt SAGE-XTM Standard Cell Library Databook,” Artisan Components, Inc.,pp. 110 Oct. 2001.

[122] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, “Matching properties of MOS transistors,” *IEEE Journal of Solid-State Circuits,* vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[123] “Agilent 85190A IC-CAP 2006 Nonlinear Device Model Manual,” Agilent Technologies, Vol.1, 2007.

[124] S. Lee and H. K. Yu, “A semianalytical parameter extraction of a spice BSIM3v3 for RF MOSFET’s using S-parameters,” in *IEEE Transactions on Microwave Theory and Techniques*, vol. 48, no. 3, pp. 412–416, Mar. 2000.

[125] A. J. Scholten, G. D. J. Smit, B. A. De Vries, L. F. Tiemeijer, J. A. Croon, D. B. M. Klaassen, R. van Langevelde, X. Li, W. Wu, and G. Gildenblat, “The new CMC standard compact MOS model PSP: Advantages for RF applications,” in *IEEE Journal of Solid-State Circuits,* vol. 44, no. 5, pp. 1415–1424, May 2009.

[126] M. Chan, X. Xi, J. He, and C. Hu, “Approaches and options for modeling sub-0.1μm CMOS devices,” in *IEEE Electron Devices Meeting, 2002,* pp. 79–82.

[127] J. Watts, C. McAndrew, C. Enz, C. Galup-Montoro, G. Gildenblat, C. Hu, R. van Langevelde, M. Miura-Mattausch, R. Rios, and C.-T. Sah, “Advanced compact models for MOSFETs,” in *Proc. Tech. WCM*, 2005, pp. 3–12.

[128] J. E. Meyer, “MOS models and circuit simulation,” *RCA Rev.*, vol. 32, pp. 42-63, Mar. 1971.

[129] Y. Cheng, M.-C. Jeng, Z. Liu, J. Huang, M. Chan, K. Chen, P. K. Ko, and C. Hu, “A physical and scalable I-V model in BSIM3v3 for analog/digital circuit simulation,”in *IEEE Transactions on Electron Devices*, vol. 44, no. 2, pp. 277–287, Feb. 1997.

[130] K.-W. Chai and J. J. Paulos, “Comparison of quasi-static and non-quasi-static capacitance models for the four-terminal MOSFET,” in *IEEE Electron Device Letters,* vol. 8, no. 9, pp. 377–379, Sep. 1987.

[131] H. J. Park, P. K. Ko, and C. Hu, “A charge-conserving non-quasistatic MOSFET model for SPICE transient analysis,” in *Technical Digest International Electron Devices Meeting,* 1988*,* pp. 110–113.

[132] D. E. Ward and R. W. Dutton, “A charge-oriented model for MOS transistor capacitances,” in *IEEE Journal of Solid-State Circuits*, vol. 13, no. 5, pp. 703–708, Oct. 1978.

[133] W. Liu, X. Jin, Y. King, and C. Hu, “An efficient and accurate compact model for thin-oxide-MOSFET intrinsic capacitance considering the finite charge layer thickness,” in *IEEE Transactions on Electron Devices,* vol. 46, no. 5, pp. 1070–1072, May 1999.

[134] M. Chan, K. Y. Hui, C. Hu, and P. K. Ko, “A robust and physical BSIM3 non-quasi-static transient and AC small-signal model for circuit simulation,” in *IEEE Transactions on Electron Devices,* vol. 45, no. 4, pp. 834–841, Apr. 1998.

[135] S.-Y. Oh, D. E. Ward, and R. W. Dutton, “Transient analysis of mos transistors,” in *IEEE Journal of Solid-State Circuits,* vol. 15, no. 4, pp. 636–643, Aug. 1980.

[136] “HSPICE® MOSFET Models Manual,” Synopsys, 2004.

[137] “Sentaurus Process User Guide” Synopsys, 2006.

[138] “Sentaurus Device User Guide” Synopsys, 2006.

[139] “Aurora Reference Guide,” Synopsys, 2006.

[140] N. Wakita and N. Shigyo, “Verification of overlap and fringing capacitance models for MOSFETs,” in *Solid-State Electronics*, vol. 44, pp. 1105–1109, June 2000.

[141] H. Aikawa, E. Morifuji, T. Sanuki, T. Sawada, S. Kyoh, A. Sakata, M. Ohta, H. Yoshimura, T. Nakayama, M. Iwai, and F. Matsuoka, “Variability aware modeling and characterization in standard cell in 45 nm CMOS with stress enhancement technique,” in *Symposium on VLSI Technology,* 2008, pp. 90–91.

[142] B. L. Austin, K. A. Bowman, X. Tang, and J. D. Meindl, “A low power transregional MOSFET model for complete power-delay analysis of CMOS gigascale integration (GSI),” in *Proc. Eleventh Annual IEEE International ASIC Conference,* 1998, pp. 125–129.

[143] K. A. Bowman, X. Tang, J. C. Eble, and J. D. Meindl, “Impact of extrinsic and intrinsic parameter variations on CMOS system on a chip performance,” in *Proc. Twelfth Annual IEEE International ASIC/SOC Conference,* 1999, pp. 267–271.

[144] X. Tang, K. A. Bowman, J. C. Eble, V. K. De, and J. D. Meindl, "Impact of random dopant placement on CMOS delay and power dissipation," in *Proc. of the 29th European Solid-State Device Research Conference,* 1999, vol. 1, pp. 184–187.

[145] X. Tang, V. K. De, and J. D. Meindl, "Effects of random MOSFET parameter fluctuations on total power consumption," in *Proc. International Symposium on Low Power Electronics and Design,* 1996, pp. 233–236.

[146] P. Packan, S. Akbar, M. Armstrong, D. Bergstrom, M. Brazier, H. Deshpande, K. Dev, G. Ding, T. Ghani, O. Golonzka, W. Han, J. He, R. Heussner, R. James, J. Jopling, C. Kenyon, S. H. Lee, M. Liu, S. Lodha, B. Mattis, A. Murthy, L. Neiberg, J. Neirynck, S. Pae, C. Parker, L. Pipes, J. Sebastian, J. Seiple, B. Sell, A. Sharma, S. Sivakumar, B. Song, A. St. Amour, K. Tone, T. Troeger, C. Weber, K. Zhang, Y. Luo, and S. Natarajan, "High performance 32nm logic technology featuring 2 generation high-k + metal gate transistors," in *IEEE International Electron Devices Meeting (IEDM),* 2009, pp. 1–4.

[147] R. Gwoziecki, S. Kohler, and F. Arnaud, "32nm device architecture optimization for critical path speed improvement," in *Symposium on VLSI Technology,* 2008, pp. 180–181.

[148] A. B. Kahng, S. Muddu, D. Vidhani, "Noise and delay uncertainty studies for coupled RC interconnects," in *IEEE International ASIC/SOC Conference*, 1999, pp. 3–8.

[149] K. T. Tang, E. G. Friedman, "Delay and noise estimation of CMOS logic gates driving coupled resistive-capacitive interconnections," *VLSI Journal of Integration*, vol. 29, pp. 131–165, 2000.

[150] F. Caignet, S. Delmas-Bendhia, E. Sicard, "The challenge of signal integrity in deep-submicrometer CMOS technology," *Proceedings of the IEEE*, vol. 89, no. 4, pp. 556–573, Apr. 2001.

[151] J. S. Yuan, L. Yang, "Teaching digital noise and noise margin issues in engineering education," *IEEE Transactions on Education*, vol. 48, no. 1, pp. 162–168, Feb. 2005.

[152] J. M. Rabaey, "*Digital Integrated Circuits: A Design Perspective*," Englewood Cliffs, NJ: Prentice-Hall, 1996.

[153] J. Lohstroh, "Static and dynamic noise margins of logic circuits," *IEEE Journal of Solid-State Circuits*, vol. 14, no. 3, pp. 591–598, June 1979.

[154] J. R. Hauser, "Noise margin criteria for digital logic circuits," *IEEE Transactions on Education*, vol. 36, no. 4, pp. 363–368, Nov. 1993.

[155] L. Ding, P. Mazumder, "Dynamic noise margin: definitions and model," in *Proceedings of 17th International Conference on VLSI Design,* 2004, pp. 1001–1006.

[156] K. L. Shepard, V. Narayanan, "Noise in deep submicron digital design," in *Digest of Technical Papers IEEE/ACM International Conference on Computer-Aided Design*, 1996, pp. 524–531.

[157] E. J. Nowak, "Ultimate cmos ulsi performance," in *International Electron Devices Meeting (IEDM) Technical Digest,* 1993*,* pp. 115–118.

[158] M. H. Na, E. J. Nowak, W. Haensch, and J. Cai, "The effective drive current in cmos inverters," in *International Electron Devices Meeting,* 2002, pp. 121–124.

[159] K. von Arnim, C. Pacha, K. Hofmann, T. Schulz, K. Schriifer, and J. Berthold, "An effective switching current methodology to predict the performance of complex digital circuits," in *IEEE International Electron Devices Meeting (IEDM)* 2007, pp. 483–486.

[160] K. von Arnim, K. Schruefer, T. Baumann, K. Hofmann, T. Schulz, C. Pacha, and J. Berthold, "A voltage scaling model for performance evaluation in digital CMOS circuits," in *IEEE International Electron Devices Meeting (IEDM),* 2009, pp. 1–4.

[161] A. Khakifirooz and D. A. Antoniadis, "MOSFET performance scaling-part I : Historical trends," *IEEE Transactions on Electron Devices*, vol. 55, no. 6, pp. 1391–1400, June 2008.

[162] A. I. Kayssi, K. A. Sakallah, and T. M. Burks, "Analytical transient response of CMOS inverters," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 39, no. 1, pp. 42–45, Jan. 1992.

[163] T. Sakurai and A. R. Newton, "Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas," *IEEE Journal of Solid-State Circuits,* vol. 25, no. 2, pp. 584–594, Apr. 1990.

[164] A. Nabavi-Lishi and N. C. Rumin, "Inverter models of cmos gates for supply current and delay evaluation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 13, no. 10, pp. 1271–1279, Oct. 1994.

[165] L. Bisdounis, S. Nikolaidis, and O. Loufopavlou, "Propagation delay and short-circuit power dissipation modeling of the cmos inverter," in *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 45, no. 3, pp. 259–270, Mar. 1998.

[166] S. Nikolaidis and A. Chatzigeorgiou, “Modeling the transistor chain operation in CMOS gates for short channel devices,” in *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 46, no. 10, pp. 1191–1202, Oct. 1999.

[167] Y. Wang and M. Zwolinski, “Analytical transient response and propagation delay model for nanoscale CMOS inverter,” in *IEEE International Symposium on Circuits and Systems (ISCAS),* 2009, pp. 2998–3001.

[168] M. H. Abu-Rahma and M. Anis, “A statistical design-oriented delay variation model accounting for within-die variations,” in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 11, pp. 1983–1995, Nov. 2008.

[169] “Gate Dielectric Capacitance-Voltage Characterization Using the Model 4200 Semiconductor Characterization System,” in *Keithley Application Note Series*, 2006.

[170] A. Brown and A. Asenov, “Capacitance fluctuations in bulk MOSFETs due to random discrete dopants,” in *Journal of Computational Electronics*, vol. 7, pp. 115–118, Sept. 2008.

[171] Y. Li, C.-H. Hwang, and T.-Y. Li, “Random-dopant-induced variability in nano-CMOS devices and digital circuits,” in *IEEE Transactions on Electron Devices*, vol. 56, no. 8, pp. 1588–1597, Aug. 2009.

[172] R. K. Cavin III and V. V. Zhirnov, "Future Devices for Information Processing," in *Proc. of 31st European Solid-State Circuits Conference,* 2005, pp. 7-12.

[173] J. M. Rabaey, A. Chandrakasan and B. Nikolic, in *Digital Integrated Circuits*, Prentice Hall Higher Education, 2008.

[174] K. O. Jeppson, “Modeling the influence of the transistor gain ratio and the input-to-output coupling capacitance on the CMOS inverter delay,” in *IEEE Journal of Solid-State Circuits,* vol. 29, no. 6, pp. 646–654, June 1994.

[175] I. Sutherland, B. Sproull, D. Harris, “ The method of logical effort,” in *Logical Effort: Designing Fast CMOS Circuits*, Morgan Kaufmann Publishers, 1999, pp. 1-26.

[176] R. Gonzalez, B. M. Gordon, and M. A. Horowitz, “Supply and threshold voltage scaling for low power CMOS,” in *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, Aug. 1997.

[177] Fikru Adamu-Lema, “Scaling and Intrinsic Parameter Fluctuations in nano-CMOS Devices” PhD. thesis, University of Glasgow, 2006.

[178] 2009 International Technology Roadmap for Semiconductors, http://public.itrs.net.

[179] F. Arnaud, A. Thean, M. Eller, M. Lipinski, Y. W. Teh, M. Ostermayr, K. Kang, N. S. Kim, K. Ohuchi, J. P. Han, D. R. Nair, J. Lian, S. Uchimura, S. Kohler, S. Miyaki, P. Ferreira, J. H. Park, M. Hamaguchi, K. Miyashita, R. Augur, Q. Zhang, K. Strahrenberg, S. ElGhouli, J. Bonnouvrier, F. Matsuoka, R. Lindsay, J. Sudijono, F. S. Johnson, J. H. Ku, M. Sekine, A. Steegen, and R. Sampson, "Competitive and cost effective high-k based 28nm CMOS technology for low power applications," in *IEEE International Electron Devices Meeting (IEDM),* 2009, pp. 1–4.

[180] K. A. Bowman, X. Tang, J. C. Eble, and J. D. Menldl, "Impact of extrinsic and intrinsic parameter fluctuations on cmos circuit performance," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 8, pp. 1186–1193, Aug. 2000.

[181] Stanislav Markov, "Gate Leakage Variability in Nano-CMOS Transistors" PhD. thesis, University of Glasgow, 2009.

[182] N. Yang, W. K. Henson, and J. J. Wortman, "A comparative study of gate direct tunneling and drain leakage currents in n-MOSFET's with sub-2 nm gate oxides," *IEEE Transactions on Electron Devices*, vol. 47, no. 8, pp. 1636–1644, Aug. 2000.

[183] M. T. Bohr, "Interconnect scaling-the real limiter to high performance ULSI," *International Electron Devices Meeting, 1995*, pp. 241–244.

[184] S. Bothra, B. Rogers, M. Kellam, and C. M. Osburn, "Analysis of the effects of scaling on interconnect delay in ulsi circuits," *IEEE Transactions on Electron Devices,* vol. 40, no. 3, pp. 591–597, Mar. 1993.

[185] D. N. Maynard, S. L. Runyon, B. B. Reuter, "Yield enhancement using recommended ground rules," in *IEEE Conference and Workshop on Advanced Semiconductor Manufacturing (ASMC), 2004*, pp. 98–104.

[186] D. James, "Design-for-manufacturing features in nanometer logic processes - a reverse engineering perspective," in *IEEE Custom Integrated Circuits Conference (CICC), 2009*, pp. 207–210.

[187] E.-P. Li, X.-C. Wei, A. C. Cangellaris, E.-X. Liu, Y.-J. Zhang, M. D'Amore, J. Kim, T. Sudo, "Progress review of electromagnetic compatibility analysis technologies for packages, printed circuit boards, and novel interconnects," in *IEEE Transactions on Electromagnetic Compatibility*, vol. 52, no. 2, pp. 248–265, May 2010.

[188] M. S. Zhang, Y. S. Li, C. Jia, L. P. Li, "Signal integrity analysis of the traces in electromagnetic-bandgap structure in high-speed printed circuit boards

and packages," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 55, no. 5, pp. 1054–1062, May 2007.

[189] D. Sylvester, C. Wu, "Analytical modeling and characterization of deep-submicrometer interconnect," *Proceedings of the IEEE*, vol. 89, no. 5, pp. 634–664, May 2001.

[190] P. Gopalakrishnan, A. Odabasioglu, L. Pileggi, S. Raje, "An analysis of the wire-load model uncertainty problem," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 1, pp. 23–31, Jan. 2002.

[191] P. K. Chan, "Comments on 'Asymptotic Waveform Evaluation for timing analysis'," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,*, vol. 10, no. 8, pp. 1078–1079, Aug. 1991.

[192] J. Qian, S. Pullela, L. Pillage, "Modeling the "effective capacitance" for the RC interconnect of CMOS gates," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 13, no. 12, pp. 1526–1535, Dec. 1994.

[193] M. Hafed, M. Oulmane, and N. C. Rumin, "Delay and current estimation in a CMOS inverter with an RC load," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 20, no. 1, pp. 80–89, Jan. 2001.

[194] F. Dartu, N. Menezes, and L. T. Pileggi, "Performance computation for pre-characterized CMOS gates with RC loads," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 15, no. 5, pp. 544–553, May 1996.

[195] E. L. Crow and K. Shimizu, in Lognormal Distributions: Theory and Applications, New York: Dekker, 1988.

[196] "CMOS nonlinear delay model calculation, in: Library Compiler User Guide", vol. 2, Synopsys, 1999.

[197] L. Wei, F. Boeuf, T. Skotnicki, and H. S. P. Wong, "CMOS technology roadmap projection including parasitic effects," in International Symposium on VLSI Technology, Systems, and Applications, 2009. VLSI-TSA '09, pp. 78–79.

[198] W. J. Gross, D. Vasileska and D. K. Ferry,"Three-dimensional Simulations of ultra small metal-oxide-semiconductor-field-effect-transistors: The role of the discrete impurities on the device terminal characteristics," *J. Appl. Phys.*, vol. 91, pp. 3737– 3740, 2002.

[199] C. J. Wordelman and U. Ravaioli, "Integration of a Particle-Particle-Particle-Mesh Algorithm with the Ensemble Monte Carlo Method for the Simulation of Ultra- Small Semiconductor Devices," *IEEE Trans. Elec. Dev.*, vol. 47, pp. 410–416, 2000.

[200] Craig L. Alexander, "*Ab initio Scattering From Random Discrete Charges and its Impact on the Intrinsic Parameter Fluctuations in Nano-CMOS Devices*," PhD. thesis, University of Glasgow, 2005.

[201] Urban Kovac, "3D Drift Diffusion and 3D Monte Carlo Simulation of on-current Variability due to Random Dopants," PhD. thesis, University of Glasgow, 2010.

[202] B. Dierickx, M. Miranda, P. Dobrovolny, F. Kutscherauer, A. Papanikolaou, and P. Marchal, "Propagating variability from technology to system level," in *International Workshop on Physics of Semiconductor Devices, 2007. IWPSD 2007*, pp. 74–79, 16-20 Dec. 2007.