A Novel Deep Submicron Bulk Planar Sizing Strategy For Low Energy Subthreshold Standard Cell Libraries

Jordan Morris

Newcastle University,

Newcastle upon Tyne, UK

A Thesis Submitted For The Degree Of

Doctor of Philosophy

November 2018

Abstract

This work investigates bulk planar deep submicron semiconductor physics in an attempt to improve standard cell libraries aimed at operation in the subthreshold regime and in Ultra Wide Dynamic Voltage Scaling schemes. The current state of research in the field is examined, with particular emphasis on how subthreshold physical effects degrade robustness, variability and performance. How prevalent these physical effects are in a commercial 65nm library is then investigated by extensive modeling of a BSIM4.5 compact model. Three distinct sizing strategies emerge, cells of each strategy are laid out and post-layout parasitically extracted models simulated to determine the advantages/disadvantages of each. Full custom ring oscillators are designed and manufactured. Measured results reveal a close correlation with the simulated results, with frequency improvements of up to 2.75X/2.43X observed for RVT/LVT devices respectively. The experiment provides the first silicon evidence of the improvement capability of the Inverse Narrow Width Effect over a wide supply voltage range, as well as a mechanism of additional temperature stability in the subthreshold regime.

A novel sizing strategy is proposed and pursued to determine whether it is able to produce a superior complex circuit design using a commercial digital synthesis flow. Two 128 bit AES cores are synthesized from the novel sizing strategy and compared against a third AES core synthesized from a state-of-the-art subthreshold standard cell library used by ARM. Results show improvements in energy-per-cycle of up to 27.3% and frequency improvements of up to 10.25X. The novel subthreshold sizing strategy proves superior over a temperature range of 0 °C to 85 °C with a nominal (20 °C) improvement in energy-per-cycle of 24% and frequency improvement of 8.65X.

A comparison to prior art is then performed. Valid cases are presented where the proposed sizing strategy would be a candidate to produce superior subthreshold circuits.

Acknowledgements

I wish to express my gratitude to my academic supervisors Alex Yakovlev and Andrei Mokhov, who provided me principally with the opportunity to conduct this research, several research areas of interest and supported me through paper submissions.

I also wish to express my gratitude to James Myers, Pranay Prabhat and the Applied Silicon Research Team at Arm. Their support and guidance helped tremendously, and their generous provision of silicon area on several tapeouts was tantamount to the success of the research herein presented.

Finally I wish to acknowledge the Engineering and Physical Science Research Council (EPSRC) and Arm Ltd for providing funding in the form of grants and studentships for this research to be conducted.

CONTENTS

List Of Tables

List of Acronyms

AES : Advanced Encryption Standard

ARM : Advanced RISC Machines

BEOL : Back End Of Line

BSIM : Berkley Short-Channel IGFET Model

CHE : Channel Hot Electron

CHISEL : Channel Initiated Secondary Electron Injection

CMOS : Complimentary Metal Oxide Semiconductor

DIBL : Drain Induced Barrier Lowering

DRC : Design Rules Check

DUT : Device under Test

EDA : Electronic Design Automation

EM : Electromigration

FO4 : Fan-Out of Four

HVT : High Threshold Voltage

IC : Integrated Circuit

INWE: Inverse Narrow Width Effect

LFD : Lithofriendly Design

LOCOS : Local Oxidization of Silicon

LOD : Length of Oxide Definition / Diffusion

LVS : Layout vs Schematic

LVT : Low Threshold Voltage

MOSFET : Metal Oxide Semiconductor Field Effect Transistor

NBTI : Negative Bias Temperature Instability

NWE : Narrow Width Effect

NMOS : N-Type Metal Oxide Semiconductor

PDK : Process Design Kit

PMOS : P-Type Metal Oxide Semiconductor

PVT : Process, Voltage, Temperature

RDF : Random Dopant Fluctuation

RET : Resolution Enhancement Techniques

RSCE: Reverse Short Channel Effect

RISC: Reduced Instruction Set Computing

RVT : Regular Threshold Voltage

SCE : Short Channel Effect

STI : Shallow Trench Isolation

Sub-Vt : Subthreshold

TCAD : Technology Computer Aided Design

TDDB : Time Dependent Dielectric Breakdown

TPD : Average Propagation Delay

TSMC : Taiwan Semiconductor Manufacturing Company

ULP: Ultra Low Power

VCD: Value Change Dump

VF: Vector Free

VHDL: VHSIC Hardware Description Language

VLSI : Very Large Scale Integration

VTH : Threshold Voltage

WPE : Well Proximity Effect

List Of Figures

Chapter 1: Introduction and Motivation

1.1 Motivation

Whilst the focus of increased computational performance remains prevalent today, in recent times a secondary avenue of research and development has emerged stimulated by the advent of mass consumable portable electronic devices, and emerging concepts like the Internet of Things (IoT). In these applications, the energy consumption becomes the market-driving factor, with degradation in computational performance accepted as the trade off for an increased functional lifetime or simply operability in a low power or low energy environment, as is the case for energy harvested applications.

1.2 Ultra Low Power and Subthreshold Design

Dennard's constant field scaling [1] has been the predominant method of device scaling used throughout the Moore's Law decades. Dennard proposed that a performance increase can be observed by scaling the minimum feature size. In order to alleviate the inevitable reliability issues and reduction in the functional lifetime of transistor devices scaled by a factor S, the geometry of the channel (Length and Width), oxide thickness, channel dopant density and operating voltage must also be scaled by the same factor, in order to maintain the same electric field. Additional benefits are observed from undertaking constant field scaling, such as the dynamic power consumption per device is scaled by a factor of S^2 and the energy consumption scales by a factor of S^3.

From the aforementioned discussion, it is clear that scaling benefits the VLSI designer by allowing the same design to be created on a smaller silicon area and consume less power, or a larger design may be created on the same silicon area for the same power budget. It is also evident that voltage scaling has been an integral part of the evolution of integrated circuit design ever since its inception. Figure 1 shows the progression in scaling of feature size over time [2].

Voltage scaling alone affects both the dynamic power consumption [3] and the leakage power consumption of standard CMOS based logic gates. In lowering the supply voltage, the switching speed and switching power dissipation on the output is reduced [4]. In order to mitigate this, Dennard stipulated that a proportional reduction in the threshold voltage should also be implemented.

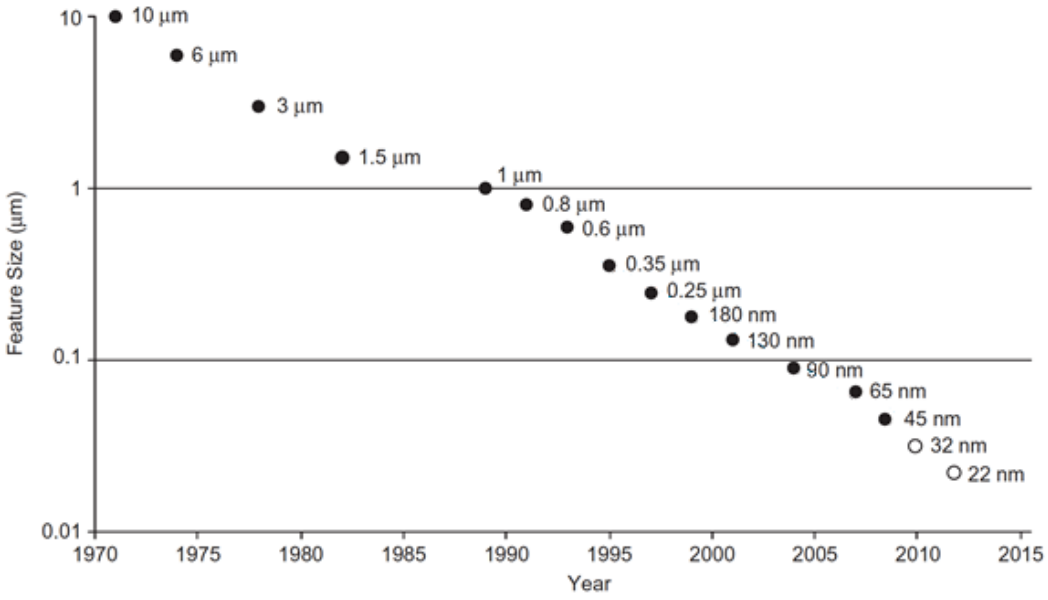


Figure 1: Minimum feature size scaling over time

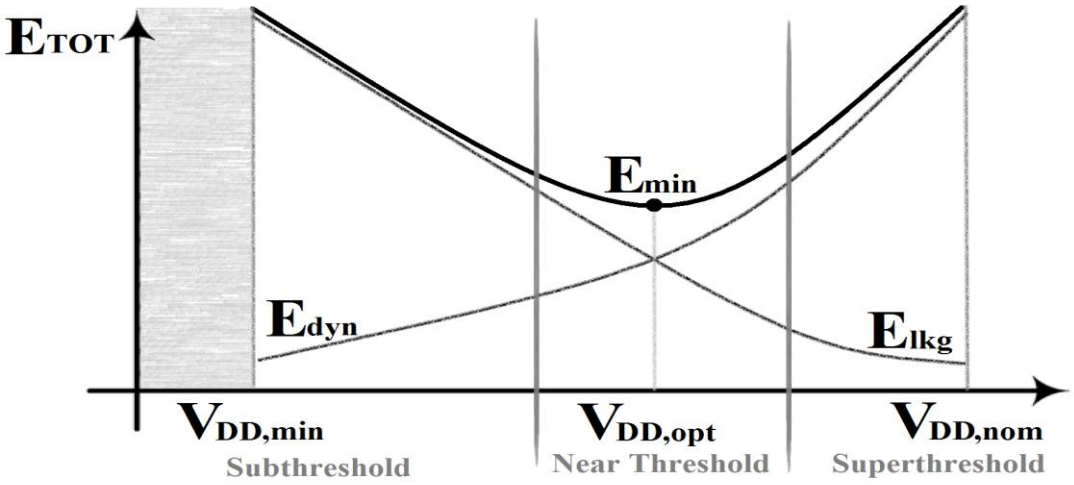


Figure 2: The Minimum Energy Point (MEP)

However, due to the electrostatic charge sharing between the gate and source-drain regions, a reduction in threshold voltage incurs a degradation of the subthreshold slope, leading to an increase in the leakage power consumption [5]. For this reason, the natural threshold voltage of the transistor may not be scaled at the same rate as the remaining characteristics of the transistor. This inevitably leads to the operation of devices below their nominal supply voltage in order to make further gains in power and energy consumption.

In mature deep submicron bulk planar technologies, dynamic energy consumption still dominates the energy per computation at nominal voltage, as shown in Figure 2. As the supply voltage is scaled towards the near threshold regime, the dynamic energy begins to fall, the device propagation speed and switching speeds begin to decrease, and the leakage energy begins to rise. As the supply voltage is scaled more aggressively towards the subthreshold regime, the degradation of the performance reaches such a level that the leakage energy begins to dominate. Critically, the overall effect of this is that an energy minimum is observed, often in the near to sub threshold regime, suggesting that true low power and low energy design occurs at a supply voltage less than that of the nominal voltage of the device. However, operating under such conditions introduces issues that must be overcome in order to achieve robust and cost efficient designs.

1.3 The Viability of Ultra Low Power Design

The three primary factors affecting the performance and consistent operation of digital logic circuits in any design scheme are process, voltage and temperature (PVT). Control of the processing procedure of integrated circuits is fundamental to ensuring that integrated circuits function as intended. In deep submicron bulk planar technologies, the largest contributor to process variation is random dopant fluctuation (RDF). This is a purely stochastic intradie variation, rendering it impossible to completely eradicate. The contribution of RDF to the overall process variation in deep submicron bulk planar technologies can reach as high as 70% [6] and consists of variation in the number and location of dopant atoms introduced into the silicon substrate in order to create the underlying device topologies. Whilst this variation is important in standard superthreshold operation, it is pivotal in the subthreshold regime where the transistor's driving (on) current, and therefore propagation delay characteristics become exponentially dependent on the threshold voltage (Vth) of the device.

The variability of devices due to temperature is also critical in the subthreshold regime as ULP devices are expected to function in a wide range of environments. In superthreshold operation, the primary mechanism of variation due to temperature is impairment to carrier mobility introduced by an increase in lattice scattering. This has a positive correlation with temperature. However, in the subthreshold regime, due to the exponential dependence of the device's current characteristics on the threshold voltage and the fact that the voltage has been scaled, the variation in threshold voltage due to temperature change contributes a higher proportionality to the overall variation. This has a negative correlation with temperature in the subthreshold regime [6]. Therefore temperature inversion is observed, designs outperform at high temperature and begin to fail at low temperatures.

Variations in the supply voltage are also exacerbated in the subthreshold regime. Whilst chip-wide supply voltage variation is generally considered diminished when the supply voltage is aggressively scaled, the subthreshold slope and therefore the noise margin is exponentially dependent on the supply voltage in the subthreshold regime. This can lead to circuit failure due to degradation in the voltage swing on the output of the logic gates and severely restricts the number of parallel devices that may be placed driving a single output in common logic topologies (e.g. Memory Arrays).

1.4 Commercial Integrated Circuit Design

Above all other concepts in the commercial production of integrated circuits, the most significant is that of design reuse. The evolution of cutting edge billion transistor designs has only been enabled by the fact that the IC designer is not obliged to design each transistor, each logic gate and each logic circuit for every chip to be manufactured. Every stage of the design is independently completed, and these completed stages may then be reused for many chip designs.

Firstly, the fabrication house designs a process from which integrated circuits may be lithographically constructed onto a silicon wafer. Technology Computer Aided Design (TCAD) simulations are undertaken to characterize the devices that may be constructed from that process. These are simplified into compact models to ease processing time on simulation. Logic gates are then constructed from the devices, optimized for their intended functionality and characterized using the technology models to produce

performance tables. These are grouped into a collection of logic gates with a similar purpose, known as a standard cell library.
A design may be described in a high-level hardware description language such as VHDL or Verilog. This design, along with the standard cell library is then passed into a synthesis tool. The synthesis tool constructs the design from the logic gates in the standard cell library and performs timing analysis on the design to ensure operability.

The objective of this thesis is to test two hypotheses, the first of which may now be posed:

Can a superior subthreshold standard cell library be created from devices that take advantage of the underlying subthreshold semiconductor physics?

Metrics of interest in the subthreshold regime include performance, variation, energy and robustness. The second hypothesis is presented at the end of Chapter 4 where several sizing strategies are compared and a novel strategy is proposed.

1.5 Outline

This thesis is presented as a logical progression of work undertaken to achieve the aforementioned objective. The chapters shall proceed as outlined below:

Chapter 2: Background and Literature Review
This chapter investigates the current state of academic research into this field alongside fundamental background knowledge. Techniques proposed by other researchers are delineated and evaluated in terms of possible contribution towards a superior subthreshold standard cell library.

Chapter 3: Subthreshold Bulk Planar Semiconductor Physics
Through electrical simulation and analytical methodologies, this chapter explores the effect on the electrical characteristics of devices from the underlying physics of deep submicron bulk planar technologies. Several physical effects described in Chapter 2 including the Reverse Short Channel Effect (RSCE) and Inverse Narrow Width Effect (INWE), as well as topological effects such as device stacking and parallelization are

evaluated and simulated to derive desirable characteristics for subthreshold standard cell design.

Chapter 4: Evaluation of Proposed Cell Sizing Strategy in Silicon
This chapter explores the implementation and measurement of ring oscillators committed to silicon in an industry standard (TSMC) 65nm bulk planar technology. The correlation between the physical effects explored in Chapters 2 and 3 and measured performance/leakage characteristics of the oscillators is evaluated. A small evaluation cell library is then characterized and the performance evaluated through a standard commercial EDA flow (Synopsys) via the synthesis of a block level digital circuit (32 bit multiplier).

Chapter 5: AES Core Synthesis Methodologies
A full low power standard cell library is constructed using the concepts of the previous chapters, characterized and three full AES cores synthesized. One core uses a full multi threshold range of cells from a novel library; One uses a restricted subset consisting only of Regular Threshold Voltage (RVT) cells from a novel library; One synthesized from ARM's current low power standard cell library as a baseline. These are then implemented in silicon.

Chapter 6: AES Core Results
The silicon cores are measured and compared. Terse evaluations are made pertaining to the results, pending full comparison in the following chapter.

Chapter 7: Discussion
The first hypothesis posed in the thesis is addressed with reference to the results provided in the following chapters and the design considerations raised in the literature review. The second hypothesis is addressed using the results from the AES core measurement. The contributions of the thesis are listed and a comparison to prior art presented.

Chapter 7: Conclusion
The work presented herein is then drawn to a conclusion and potential areas of future research outlined.

A historical overview of the nascent semiconductor research field is included in Appendix A.

1.5 Publication Listing

1.5.1 Conference

J. Morris, P. Prabhat, J. Myers and A. Yakovlev, "Unconventional Layout Techniques for a High Performance, Low Variability Subthreshold Standard Cell Library," *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, Bochum, 2017, pp. 19-24

A. Wheeldon, J. Morris, D. Sokolov and A. Yakovlev, "Power proportional adder design for Internet of Things in a 65 nm process," *2017 27th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, Thessaloniki, 2017, pp. 1-6.

1.5.2 Patent Applications

Application submitted to ARM PRC (Patent Review Committee) for Full Diffusion sizing strategy as outlined in Chapter 4.

Chapter 2: Background and Literature Review

2.1 Outline

For ease of understanding and navigation, this chapter is broken into several sections. Section 2.2 outlines the general operation and considerations of digital logic in the subthreshold regime. Sections 2.3, 2.4 and 2.5 then delineate individual effects that must be considered for digital subthreshold operation. These are separated into three broad categories depending on how they most effect digital subthreshold logic; reliability, variability and performance. Section 2.6 finally concludes with standard cell library strategies that other researchers have attempted in order to improve circuit operation in the subthreshold regime.

2.2 Digital Logic Overview

2.2.1 System Level

Due to the comparatively latent evolution of battery technology, academic and industrial attention has predominantly been focused on the optimization of energy and power consumption for the typical ultra low power integrated circuit application. Two methods by which these circuits may be powered are generally considered; Battery powered and energy harvested. For battery-powered systems, the key metric of battery lifetime may be derived as [7]:

$$T_{battery} = \frac{E_{battery}}{P_{avg}} \tag{1}$$

Whereby the battery lifetime is dependent upon the average power demanded from the IC and the total energy that the battery may deliver over its lifetime. The typical low energy application does not perform a continuous stream of computation, but rather performs a simple repetitive task for a short period of time then sleeps. In order to conserve power, a duty-cycled system architecture may be adopted. In this scheme of operation, a small selection of blocks that typically provide timing functionality are always on, but blocks required only during computation may be placed into a sleep mode when idle and activated only when required. The average power consumption of this type of system may be derived as [7]:

$$P_{avg} = P_{always-on} + p_{sleep} + \frac{E_{active}}{T_{wkup}} \quad (2)$$

Whereby $P_{always\text{-}on}$ is the average power consumption of the blocks that cannot be duty cycled in the given application, P_{sleep} is the average power consumption of the duty cycled blocks whilst in a non-active state and the final term is the energy consumption of the active blocks over the time period of activity. Depending on the scheme of operation, the duty cycle blocks may be clock gated, reducing the contribution of P_{sleep} purely to the leakage power of the blocks, power gated, reducing the contribution even further or even completely powered down, eliminating the contribution completely.

Energy scavenged systems principally follow battery powered systems, with the additional constraint that the instantaneous power available must be sufficient to power the device in its intended mode of operation.

From the above discussion it can clearly be seen that both power and energy must be minimized in order to extend battery lifetime or ensure sufficient instantaneous power is available.

2.2.2 Logic Level

In CMOS logic, there are three primary contributors to energy consumption; Short circuit, dynamic and leakage. Short circuit current is lost during a switching event when both the pull up and pull down transistor networks are simultaneously in a state of change. Due its sharp transitional characteristics, this is commonly such a small contribution in CMOS technology that it may be disregarded [7].

Dynamic energy consumption is the contribution of charging/discharging the capacitance of the following stage of logic (plus any parasitic capacitance) to one of the rail voltages through a MOS transistor network during a clock cycle. The dynamic energy per clock cycle may be calculated as [7]:

$$E_{Dyn} = C_{eff} {V_{DD}}^2 = \alpha_{sw} C_{TOT} {V_{DD}}^2 \quad (3)$$

Whereby C_{eff} is the effective capacitance, C_{TOT} is the total physical capacitance and α_{sw} is the switching factor. It may be observed from Equation 3 that a quadratic reduction in dynamic energy may be achieved by the scaling of the supply voltage. Hidden in

Equation 3 is the effect of voltage scaling on the capacitance, further described in Section 3.5.

Leakage energy consumption is the contribution of current flowing rail-to-rail through one or more transistors in the off state due to non-idealities in the devices. The leakage energy per clock cycle may be calculated as [7]:

$$E_{lkg} = V_{DD} I_{off} T_{ck} = V_{DD} I_{off} T_D \frac{LD}{X_{stack}} \quad (4)$$

Whereby I_{off} is the leakage current, T_{CK} is the clock period, T_D is the average propagation delay, LD is the logic depth and X_{stack} is the average stacking factor, further described in Section 3.4. One might discern from Equation 4 that a reduction in supply voltage produces a linear reduction in leakage energy. However, from the brief discussion in Section 1.3, it was shown that this is not the case as a reduction in supply voltage leads to an exponential increase in propagation delay. This increases leakage energy at almost an exponential rate. These relationships support the discussion of a minimum energy point as outlined in Section 1.3.

2.2.3 Device Level

2.2.3.1 General Subthreshold Device

The subthreshold current for an NMOS device may be derived as [8]:

$$I_{ST} = I_0 \frac{W}{L} e^{\frac{V_{GS} - V_{th}}{n.Vt}} (1 - e^{-V_{DS}/V_t}) \quad (5)$$

Whereby I_0 is the current at $V_{GS} = V_{th}$ for the given technology node, V_t is the thermal voltage equal to kT/q, W/L is the aspect ratio and n is the subthreshold factor (derived in Eq .27). The threshold voltage V_{th} may be derived as [7]:

$$V_{th} = V_{TH0} - L_{DS} V_{DS} - L_{BS} V_{BS} \quad (6)$$

Whereby V_{TH0} is the natural threshold voltage, L_{DS} is the DIBL coefficient, V_{DS} is the drain to source voltage, L_{BS} is the body effect coefficient and V_{BS} is the bulk to source voltage.

The natural threshold voltage may be derived as:

$$V_{TH0} = V_{fb} + W_s + E_{ox}T_{ox} \quad (7)$$

Whereby V_{fb} is the flat band voltage, W_s is the surface state potential at inversion (equivalent to $2kT/q(\ln(N_{sub}/N_i))$ [9], E_{ox} is the electric field in the gate oxide and T_{ox} is the oxide thickness.

2.2.3.2 Subthreshold Currents

An equation to describe the inherent strength of a subthreshold device may be derived as [7]:

$$B = I_0 . \frac{W}{L} . e^{-(V_{TH0} - \frac{L_{BS}V_{BS}}{n} . V_t)} \quad (8)$$

Using this metric, the on current may be derived as:

$$Ion = B . \frac{e^{V_{DD}}}{n} . V_t \ . [\frac{e^{L_{DS}V_{DD}}}{n} . vt(1 - e^{-V_{DD}/V_t})] \quad (9)$$

The off current may be derived as:

$$Ioff = B . \frac{e^{L_{DS}V_{DD}}}{n} . V_t \, (1 - e^{-V_{DD}/V_t}) \quad (10)$$

The Ion/Ioff ratio may be condensed to:

$$\frac{Ion}{Ioff} = \frac{e^{V_{DD}}}{n} . V_t \quad (11)$$

2.3 Robustness

2.3.1 Channel Hot Electron (CHE)

The first scientific description of hot carrier injection was a probabilistic analytical model termed the 'lucky electron model' delineated in 1979 [10]. The work proposed that an electron may penetrate the dielectric of a MOSFET if it gains sufficient energy by collision aversion and is redirected upwards in the channel via acoustic phonon scattering. The resultant accumulated trapped charge induces drain current degradation, transconductance reduction, threshold voltage shifting and ultimately device failure. The work postulated that the probability of an electron gaining sufficient energy is derived as:

$$P = e^{-d/\lambda} \tag{12}$$

Where d is the distance required to gain the sufficient energy for emission into the dielectric and λ is the mean free path. This was determined as 91 Angstrom for an electron at room temperature. The distance d may be determined as:

$$d = \theta B/E \tag{13}$$

Whereby θB is a factor of the potential-distance profile and effective barrier height of the point at which the redirection towards the dielectric occurs and E is the electric field in the direction of the length of the channel (longitudinal). The deflection angle must also be within 30 degrees of the orthogonal angle of the Si-SiO2 interface.

In accordance with the aforementioned relationships, the lucky electron model stipulated that the critical parameter for channel hot electron degradation to occur was a V_d-V_s drop of approximately 2.5V, agnostic of channel length and therefore Dennard scaling. As devices progressed into the submicron era, experimental evidence of channel hot electron persisted at voltages as low as 1.4V [11], subthreshold for the experiment's contemporary transistor. The lucky electron model was therefore augmented to account for additional sources of CHE, primarily Auger recombination, a process by which a hole and electron recombine, releasing energy to an adjacent electron allowing for the CHE effect to occur. The scientific community also conceded that degradation in MOSFET reliability occurred for carriers with insufficient energy to surmount the Si-SiO_2 interface but sufficient

energy to create interface states via collision with the interface. A negative correlation between CHE and temperature was also established for subthreshold operation, identified as a result of lower lattice scattering and therefore an increase in the mean free path due to reduction in collisions. The same principles were subsequently proven for holes [12].

2.3.2 Channel Initiated Substrate Electron Injection (CHISEL)

A secondary effect that contributed to carrier injection was identified, initially termed Secondary Ionization Induced Substrate Hot-Electron (SIHE) [13] and eventually renamed Channel Initiated Substrate Electron Injection (CHISEL) by the academic community [14]. Empirical evidence showed a positive correlation between the transverse electric field created by the potential difference between the gate and body and the magnitude of hot carrier degradation. [13] determined this to be related to secondary impact ionization in the drain depletion region. As the electrons are accelerated in this region, collisions occur generating electron-hole pairs. In the presence of the transverse field, the electrons drift towards the channel surface and holes towards the device body. These then generate additional electron-hole pairs in an avalanche multiplication effect. This secondary effect gives some electrons sufficient energy to surmount the SI-SiO_2 interface. This phenomenon has an impact on the use of body biasing [15].
A tertiary impact ionization effect was also discovered [14], extending the transverse field dependence to degradation caused by holes in PMOS devices and identifying alleviation of the phenomenon by drain engineering in the form of the lightly doped drain (LDD). A comprehensive study of carrier injection due to the transverse field using a novel charge pump technique was eventually conducted [16]. This showed that due to the low hole mobility in the gate oxide, hole oxide traps fill quicker, limiting the number of new hole traps over long periods of stress. Electron traps fill slower, the transverse field assists in detrapping charges and therefore no saturation point is observed.

2.3.3 Time Dependent Dielectric Breakdown (TDDB)

In a 65nm process, the gate oxide thickness across varying technologies is reported to be between 0.85nm to 2nm [17], merely several atomic layers thick [18]. At this thickness, any current passing through the oxide is a result of direct quantum-mechanical tunneling [6] in the form of Edge Direct Tunneling [19]. The effect of this percolation of carriers through the oxide is the generation of defect sites in the oxide layer [20]. Once a critical

density of defects is reached, the oxide breaks down. The breakdown of an oxide is defined as a sudden increase in conductance, usually accompanied by current noise [21]. The defects may be grouped into one of three groups: Paramagnetic interface defects [22], diamagnetic interface defects [23] and bulk electron traps [24]. Two long standing physical models exist to explain their generation; The Anode-Hole Injection (AHI) model [25] and E-model [26].

In order for breakdown to occur, a continuous chain of defects, known as a percolation path, must exist, spanning the thickness of the dielectric. Given that the defects are generated in a spatially (X,Y and Z) stochastic manner in the oxide, breakdown occurs at a critical defect density:

$$N^{BD} = \frac{1}{q}\int_0^{Q_{DB}} P_g dQ \tag{14}$$

Whereby: Q_{DB} is the total charge to breakdown, q is the electron charge and P_g is the defect generation per injected electron density, which may be derived as:

$$P_g = q.dN/dQ \tag{15}$$

This model suggests that the time to breakdown is dependent on the electrical stress placed on the oxide, the oxide thickness and the gate oxide area [27], the latter forming an inverse proportionality as a large oxide area equates to a higher probability that a single percolation path will exist and create a dielectric breakdown.

The defect generation rate obeys an approximate exponential decay below a gate voltage of around 5V. The dependence is actually on the electric field dropped across the gate oxide [28] and therefore any techniques employed to manipulate the magnitude of the electric field will have an effect on the defect generation rate. Experimental results show a measurable defect generation rate at voltages as low as 1.2V. Whilst subthreshold operation in a 65nm technology node is sub 375/325mV (PMOS/NMOS), any application of ultra wide dynamic voltage scaling up to the nominal 1.2V must be considered.

Assuming constant current, the operational lifetime may be derived as:

$$T_{BD} = Q_{BD}/J \tag{16}$$

Whereby J is the instantaneous current density through the oxide. This is proportional to the applied voltage in accordance with the predominant tunneling mechanism.

2.3.4 Back End Of Line (BEOL)

The 65nm technology node chosen to undertake this research utilizes a 6 layer dual damascene copper interconnect structure with a low-k dielectric. The back end of line structure poses it's own reliability issues. These are discussed in this sub-section.

2.3.4.1 Time Dependent Dielectric Breakdown (TDDB)

Analogous to TDDB in gate oxides, a similar effect is frequently observed in the BEOL structure. To reduce resistance-capacitance (RC) interconnect time delay, low-k dielectrics are utilized between interconnect layers and adjacent lines (Inter-layer and Intra-layer Dielectrics (ILD)). Due to the reduction in the dielectric constant, low-k dielectrics have an intrinsic breakdown strength that is weaker than the high-k dielectrics (SiO_2) typically used in gate oxides. The model therefore follows a SQRT(E) relation [29].

Unlike the causes of gate oxide TDDB discussed in Section 2.3.3, the primary cause of BEOL TDDB is metal ion diffusion into the dielectric. This process then follows an electron-fluence driven breakdown pattern, sensitive to the field-strength dropped across the dielectric and temperature. Breakdown occurs in a manner similar to the percolation path of gate oxide TDDB. As copper rapidly diffuses into most dielectrics, a liner or barrier is typically employed.

2.3.4.2 Electromigration

Electromigration may be split into two categories; process integration and geometric (layout). Whilst process integration reliability issues such as trench and via aspect ratios and liner deposition conditions are important, they are set by the fabrication house and are therefore beyond the scope of the library designer. Geometric considerations may be split into two subcategories; line depletion and via depletion.

2.3.4.2.1 Line Depletion

Line depletion electromigration is concerned with depletion of the copper interconnect along a single interconnect. The predominant model of isothermal electromigration is the Black model [30]:

$$t_{50} = AJ^{-N} exp\left(\frac{Ea}{kT}\right) \quad (17)$$

Whereby J is the current density in the interconnect path, Ea is the activation energy in electronvolts (0.9 - 1 for copper), k is the Boltsman constant, T is the temperature in Kelvin, A is a technology specific constant and n represents the failure behavior (typically 2). It is the designer's responsibility to ensure adequate interconnect capability for the current demands of the circuit under design. For self-heating circuits, a rise of only 5 degrees is enough to degrade the expected lifetime by 30%. This can be exacerbated by the use of low-K ILD [31]. For operation in the subthreshold regime, Joule heating is of negligible importance. However utilization in schemes such as ultra wide dynamic voltage scaling could place restrictions on interconnect width and length.

2.3.4.2.2 Via Depletion

Several geometric aspects affect via reliability in terms of electromigration. The two most prominent are via redundancy and interconnect overhang. Via redundancy utilizes several minimum aspect ratio vias to provide the role of a single interlayer connection. Minimum aspect ratio vias of a fixed size (90nm x 90nm in the adopted technology) are recommended, as via etching is area dependent on the quantity of etchant. Redundancy lowers the current density through each via, prolonging the lifetime and serving as a backup in the event of a single via breakdown.

Interconnect overhang is an extension in interconnect length beyond the via location. This is critical as it lowers the aspect ratio required during processing to fully line the via, increasing the liner quality and lowering the likelihood of copper migration into the dielectric [32]. Improper lining quickly leads to copper diffusion and via voiding, effectively destroying the connection.

2.3.5 Negative Bias Temperature Instability (NBTI)

Negative bias temperature instability is a physical degradation primarily affecting PMOS transistors [33]. The SiO_2 used to form the gate oxide is amorphous, differing from the monocrystalline silicon channel underneath. This creates a rough interface surface and leaves some silicon bonds dangling. These dangling bonds form interface traps. To alleviate this unwanted behavior, the interface is annealed in hydrogen ambient, which

diffuses into the gate oxide and passivates the dangling Si bonds, eliminating the interface traps.

For negatively biased devices (PMOS), holes at the channel surface interact with the Si-H bonds, weakening and eventually breaking them, to form the initial silicon dangling bond and a free hydrogen atom that can then diffuse back into the oxide. This process is exacerbated by strong bias conditions and elevated temperature, gradually building over time, reducing the reliability of the devices. The end result is a drift in device threshold voltage, an increase in parasitic capacitances and a lowering of carrier mobility resulting in a diminished drain current.

The currently accepted analytical model of NBTI is the Reaction-Diffusion model which primarily treats the physical effect as a chemical reaction such that [34]:

$$Si - H + h^+ \leftrightarrow Si \cdot + H^0 \tag{18}$$

Thus the interaction of a Si-H bond and hole form the dangling silicon bond and free hydrogen. The rate of increase in the interface trap density can therefore be described as the difference between the bond-breaking and bond-annealing processes:

$$\frac{dN_{it}}{dt} = k_f(N_0 - N_{it}) - k_r N_{IT} N_H{}^{(0)} \tag{19}$$

Whereby N_0 is the Si-H bond density before any NBTI degradation, Nit it the interface trap density, $NH^{(0)}$ is the density of hydrogen at the interface and k_f and k_r are the bond-breaking and bond-annealing rates respectively.

Initially, Nit and $NH^{(0)}$ are negligible and therefore the breaking process dominates. As the released hydrogen builds at the interface, the bond-annealing begins to occur and eventually the two processes reach an equilibrium point. If this process is left to persist, eventually the free hydrogen begins to diffuse away from the interface and into the oxide where it can no longer be re-annealed back into the Si-H bond, resulting in overall degradation. This rate may be determined by:

$$\frac{dN_H}{dt} = D_H \nabla^2 N_H \tag{20}$$

Whereby D_H is the diffusion constant of hydrogen. Crucially for the gate designer, ∇^2 is a geometry dependent term. [35] showed that as transistor channel width is reduced, the diffusion of the hydrogen released from the Si-H bonds is enhanced, resulting in an increase in the NBTI induced interface density. [36] shows the gate voltage and temperature relation of NBTI. These are well beyond the bounds of the non-self-heating deep subthreshold regime (over 3V at 300k). However it must be noted that the geometric impact of NBTI may be of concern for narrow width devices used at nominal voltages where self-heating will occur.

2.4 Variation

2.4.1 Mechanical Stress

In the 1980s a critical juncture was reached regarding device isolation. The contemporary process of Local Oxidization of Silicon (LOCOS) provided isolation in the form of a gradient known colloquially as the Bird's Beak [37]. In order to improve transistor density, a new technique known as Shallow Trench Isolation (STI) was introduced. This technique involves dry etching a trench into the silicon and introducing a gap-fill-oxide at 1000 °C. This structure is then densified in an O_2 ambient at the same temperature. As this process proceeds, the volume of the isolation increases, imparting a mechanical stress into the silicon. This additional stress impacts the front-end-of-line structures in a variety of ways including the dopant diffusion, carrier mobility and TDDB immunity.

Of primary concern in subthreshold operation is the effect of the stress on device threshold voltage and transconductance. Both of these parameters show significant deviation in presence of a mechanical stress. Interestingly, the STI-induced stress only travels through the silicon a fixed distance from the STI wall. This parameter is generally termed the Length of Oxide Diffusion/Definition (LOD). Due to this diminishing impact, the characteristics of a device are sensitive to the device layout pattern [38], and not just the conventionally modeled layout parameters such as length and width.

Both NMOS and PMOS devices show deviation from STI-induced stress. The drain engineering in NMOS devices suffer diffusion reduction in the presence of compressive stress. This results in an increase in the dopant at the edges of the channel, raising the threshold voltage and therefore diminishing the drain current capability. Biaxial stress has also been shown to degrade electron mobility through increased scattering [39] [40] [41]. In contrast, compressive stress in PMOS devices improves hole mobility, improving the

drain current capability. Therefore NMOS devices are impoverished and PMOS devices enhanced.

The meaningful deviation barrier of LOD has been shown to be around 2µm in bulk planar technologies [39]. An X1 inverter cell in a 65nm process is typically a total of 600nm wide (2.4 micron tall in a 12 track library). Therefore mechanical stress is unavoidable in a dense standard cell library at this technology node. Deviations in drain current of up to 13% have been reported [42] [43].

The same deviations also affect the leakage current, thereby improving (NMOS) and diminishing (PMOS) the device leakage characteristics.

Whilst this is an additional concern to the layout designer, the effect is modeled in BSIM models 4.5 and higher. Moreover, the contrarian impact on NMOS/PMOS devices actually brings the intrinsic design strengths closer to parity.

2.4.2 Lithofriendly Design (LFD) – Geometric Considerations

As device scaling progresses, physical effects once considered negligible begin to impart significant variation to device characteristics. These are generally dependent on the process and technology used. Layout design that takes these effects into consideration is know as Lithofriendly design. The geometric effects considered here are Polysilicon Pitch, Device Orientation, Polysilicon Neighborhood, Polysilicon Corner Rounding, Contact Density, Polysilicon Counter-Doping and Line-Edge Roughness (LER). Whilst technically a geometric consideration, Random Dopant Fluctuation is discussed in its own section due to its significant impact in the subthreshold regime (Section 2.4.3).

2.4.2.1 Polysilicon Pitch

Two effects govern variation imparted from polysilicon pitch; Optical proximity effects [44] and stress. Optical proximity effects arise from the processing technology used [45]. Resolution Enhancement Techniques (RET) are required to create feature sizes that are smaller that the lithographic wavelength. As there is no longer a 1:1 ratio on the aperture, the correct processing of features in close proximity becomes more difficult. [46] showed that variability in device performance becomes negligible for polysilicon gate pitches greater than $3\lambda/Na$, where λ is the lithographic wavelength and Na the numerical aperture of the process exposure system. For a typical 65nm process, these values are 193nm and 0.75 respectively. This determines a distance of interaction at 775nm. For a single polysilicon pitched gate (eg an inverter) that has a total cell width of 600nm in a 65nm

process, this might be on the bounds of negligibility. However, most logic gates contain more than a single polysilicon pitch (eg NAND gate) and therefore this must be taken into consideration. Variations of up to 5% for the weakest pitches have been reported [47]. Some deep submicron bulk planar technologies use stress engineering in the form of stress layers to impart a beneficial stress level into PMOS devices [48]. The stress layers are highly sensitive to the polysilicon density and therefore pitch of polysilicon on which they act. Drain current variation of up to 10% has been reported between reference and weak pitches [47].

2.4.2.2 Device Orientation

Modern CMOS technology utilizes the Czochralski method of silicon ingot formation [49]. This involves melting silicon in a silicate crucible, dipping a monocrystalline silicon seed into the silicon and slowly removing upwards to form the monocrystalline silicon ingot. This is then marked at one edge to identify the lattice orientation and processed into wafers. The lattice orientation used for modern CMOS processes has the miller index of [1,0,0] as this produces the lowest interface states and therefore surface sate potential during gate oxidization [50]. This is important as it reduces variation in the threshold voltage.

Theoretically, for a pure monocrystalline silicon structure, this orientation means that the intrinsic impedance imparted onto a current flow by the monocrystalline lattice should be constant provided that flow is in an orthogonal direction to the lattice structure. For a miller index of [1,0,0], this is in the vertical and horizontal dimensions from the viewpoint of the layout designer (looking vertically down onto the wafer). However, due to parametric gradients and other systematic (correlated) processing variations [51], extensive research shows this not to be the case. Therefore to reduce variability, polysilicon gate orientation should be restricted to a single dimension (typically vertical to produce a longitudinal gate orientation in the horizontal plane). Variations in drain current of up to 5% have been shown for orientation mismatch with weak polysilicon pitches [47].

2.4.2.3 Polysilicon Neighborhood

Polysilicon structures in advanced technology nodes use Reactive Ion Etching (RIE) to define their structures. Etching technologies, particularly chemically based, have variation based on the polysilicon density and structure. The simplest example is that of

minimum pitched, multiple polysilicon gates in a single orientation. During processing, the polysilicon lines may be viewed as a comb-like structure. Once the etching process begins, the ease of access to the outer polysilicon lines outweighs that of the polysilicon lines in the center of the structure. The result is that the outer polysilicon lines etch deeper than those in the center of the structure, creating variation. This is such a critical issue in analog design that sacrificial dummy polysilicon structures are placed on the outer most affected area. In standard cell design, where the overall logic is synthesized, keeping the polysilicon towards the center of the cell offers the best defense against this source of variation. Variations between crowded and sparse polysilicon can reach up to 10% [47].

2.4.2.4 Polysilicon Rounding

Due to the ever-decreasing dimensions in highly scaled processes, rounding errors become of greater significance [52]. Rounding errors are imperfections in the corners of structures with perpendicular abutments. A typical application would be the polysilicon routing close to the active gate area in order to abut two adjacent polysilicon gates. If the proximity of the abutment is such that the rounding encroaches onto the active gate area, this will cause a variation in the gate. To avoid this, routing on the polysilicon layer between gates should be kept at a distance from the active gate area. Variations of up to 7% in drain current have been reported [47].

2.4.2.5 Contact Density

Processing of vias can impart stress onto the front end of line structures and mutually between other vias in close proximity, creating variation in via conductivity. Best practice is therefore to avoid minimum via pitches. Variation of up to 5% in drain current has been reported between densely packed and sparse via arrangements [47].

2.4.2.6 Polysilicon Counter Doping

Intrinsic polysilicon has a comparatively high resistivity. Moreover it is beneficial to the process engineer to be able to alter the work function of PMOS and NMOS devices. For these reasons, polysilicon gate structures are doped with the same underlying dopant (Boron/Phosphorous) as the gate type. Due to the complexity of the process and the fact that alternate devices often share the same polysilicon structure, devices of a close proximity can often experience counter-doping, creating variation. PMOS devices suffer more with decreases in drain current up to 10% reported [47].

2.4.2.7 Line Edge Roughness

For highly scaled technology nodes with critical dimensions smaller than the lithographic wavelength and non-parity aperture ratios, even creating a straight line in polysilicon poses a challenge [53]. This phenomenon is known as line edge roughness and follows a statistical variation model proportional to the perimeter of the polysilicon structure [54]. Therefore upsizing gates dimensions reduces the impact of this form of variation.

2.4.2.8 Cost of Lithofriendly Design

The sources of variation outlined in this subsection can all be addressed by increasing structure to structure spaces or increasing device dimensions. Whilst this does alleviate variation, it is often prohibitively costly on silicon area. It is therefore left to the cell designer's judgment to accurately balance variation and area cost.

2.4.3 Random Dopant Fluctuation (RDF)

In bulk planar deep submicron technologies, the largest contributor of variation is random dopant fluctuation. This arises from the control of quantity [55] and location of dopant atoms implanted into the channel during processing. The dopant process used is ion implantation. The wafer is typically angled at 7 degrees to prevent deep dopant penetration during the implantation (known as channeling). This process is purely stochastic (uncorrelated) and follows a statistical relationship with the quadrature of the device such that [56]:

$$A_{vt} = \frac{1}{\sqrt{WL}} \tag{21}$$

Whereby W and L are the width and length of the device respectively.
If the dopants atoms penetrate deep into the channel, the dopant density at the surface of the channel is light. This lowers the flat-band voltage as the channel surface becomes easier to invert with applied voltage at the gate and therefore, the natural threshold voltage of the device is lowered, resulting in a device typically seen at the FF (fast) process corner. Alternatively if the dopant atoms remain closer to the surface of the channel, a higher voltage is required on the gate and the natural threshold voltage increases, resulting in a device typically seen at the SS (slow) process corner. This has

implications for both the active drain current (impacting performance) and inactive leakage current (impacting energy consumption).
This form of variation has been shown to account for up to 70% of the total variation in deep subthreshold operation [7]. As the variation is purely stochastic and is inversely proportional to the inverse of the square root of the quadrature, the only method to lower this source of variation is to upsize the device dimensions. This is a natural step for performance enhancement in subthreshold operation as outlined in Sections 2.5.1 and 2.5.5.

2.4.4 Well Proximity Effect (WPE)

Modern highly scaled bulk planar technologies utilize a retrograde well profile. This lowers the likelihood of bulk punch-through and aids in the prevention of latch-up [57]. Whilst this ion implantation process takes place, photoresist is used to mask off the well edges. As the ions hit the photoresist at the well edge, they can be laterally scattered into the well, increasing the dopant density close to the well edge [58].
If devices are placed within this region, the surface dopant density of the channel increases. This increases the threshold voltage. Moreover the additional dopant atoms increase the impurity scattering within the channel, lowering the effective mobility. Both of these effects result in a reduction of the drain current.
It has been shown that the distance between device and well edge where WPE begins to have a non-negligible effect is approximately 1µm [59]. At around 400nm a large proportion of the scattered ions reach the device channel. At a distance of 100nm, the variation in the drain current has been shown to be around 10% [60]. For a standard cell library in a 65nm technology node, these distances are of concern.

2.5 Performance

2.5.1 Reverse Short Channel Effect (RSCE)

Due to the significant impact on the academic contribution of this thesis, the Reverse Short Channel Effect (RSCE) shall be broken down into subsections; Superthreshold physical affects governing geometric length, RSCE and RSCE's effect on subthreshold characteristics.

2.5.2 Superthreshold Physical Effects

2.5.2.1 Superthreshold Current

Fabrication houses design technology processes with the continued avenue of performance enhancement as their primary motivation. At nominal voltage, a concise description of the drain current may be determined as:

$$I_D = K\frac{W}{L_{eff}}(V_{GS} - V_{th})^{\alpha} \tag{22}$$

Whereby W and L_{eff} are the width and effective length respectively, V_{gs} is the gate to source voltage, V_{th} is the threshold voltage and K and α are technology dependent parameters. It must be noted that L_{eff} is also a factor in determining V_{th}. Even so, it is evident that the drain current is at most quadratically proportional to the threshold voltage in superthreshold operation (and at least linearly proportional). The determining factor between these two options is mobility saturation.

2.5.2.2 Drain Induced Barrier Lowering (DIBL)

As the ion implantation of dopant atoms proceeds to create the source and drain regions, a depletion region and potential barrier is formed between the implant and substrate/well. In keeping with all other heterogeneous P-N junctions, this restricts the amount of current that may flow between the two regions. As a potential difference, V_{DS}, is placed across the drain and source, the depth of the depletion region increases, as does the field penetration at the surface of the device [61]. In long channel devices, the large distance between the two regions ensures that the source region is unaffected by this change in the drain voltage [62]. However, in short channel devices, the interaction of the drain and source at the surface of the channel increases and the potential barrier at the source region lowers. This creates a geometric effect whereby the threshold voltage of the device is lowered as the length of the device decreases.

2.5.2.3 Short Channel Effect (SCE)

Aforementioned was that fact the source and drain regions create a depletion region with the surrounding substrate/well. In self-aligned bulk planar technology nodes, this means that the depletion regions extend underneath the gate. As energy is no longer required at

the gate to invert this portion of the channel, the voltage required at the gate to invert the channel as a whole decreases (a phenomenon known as charge sharing [63]). As the device length is decreased, the threshold voltage may be seen to 'roll off' as the depletion regions begin to occupy a greater proportion of the overall channel. Therefore, this also creates an effect whereby the threshold voltage of the device is lowered as the length of the device decreases.

The above two physical effects are cumulatively considered as the Short Channel Effect (SCE) and account for the monotonic fall in drain current as the device length is increased in the superthreshold regime. This gives rise to the shortest channel length offering the highest geometric density and performance characteristics. However, the short channel effects pose problems to the process designer.

In superthrehsold operation, the aforementioned physical phenomena affect the off (leakage) current greater than the on (active) current. This leads to degradation in the Ion/Ioff ratio and consequently a reduction in the performance-to-energy metric. Moreover, DIBL and SCE derived from the depletion regions are severely affected by process variation resulting from deviation in the geometric length of the device. This translates into a larger variation in the threshold voltage. If the depletion regions extend far enough to meet in the channel center, carriers injected into the depletion region will be swept from drain to source, resulting in a substantial increase in the leakage current and loss of the control of the device. This condition is known as bulk punch-through.

To prevent these conditions from occurring, an additional implant step is added to increase the dopant density of the channel adjacent to the source/drain implants. These additional implants are known as HALO implants [64]. By increasing the dopant density in these regions, the depletion depths are reduced, mitigating the negative effect.

2.5.3 Subthreshold Physics

2.5.3.1 Reverse Short Channel Effect

In the subthreshold regime, due to the lower supply voltage, both DIBL and the depletion region SCE are reduced to insignificance. The non-uniform dopant profile of the channel created by the HALO implants is therefore left to induce another physical effect.

In short channel devices, the HALO implant dopants extend into the center of the channel, interacting to increase the dopant density. As the dopant density of the channel is

higher, it requires a greater voltage at the gate to invert it, thus the threshold voltage of the device is increased. As the channel length is increased, the HALO implant dopants are pulled further apart and their interaction in the center of the channel diminishes. This reduces the dopant density in the center of the channel and therefore a lower voltage is required at the gate to invert it, thus the threshold voltage of the device is lowered. Once the critical length at which the HALO implant dopants no longer interact is achieved, increasing the length merely serves to increase the impedance to carriers and the threshold voltage therefore begins to increase once more.
In contrast to the superthreshold regime, this suggests an optimal drain current for the device at a length greater than the process minimum.

2.5.3.2 Dopant Profile Optimization

Dopant profile optimization for subthreshold operation has been thoroughly explored. This typically involves removing the HALO dopants whilst altering the retrograde dopant profile to compensate. As the HALO implants increase the dopant density around the source/drain regions, the depletion depths are reduced and therefore the junction capacitances between these regions and the substrate/well increases. Removing the HALO implants therefore results in larger depletion regions and lower junction capacitances. This reduces both the switching power and delay times. It also results in a simplified processing technology.

The disadvantage to this technique is that it also increases the leakage current. This deleterious effect may be managed by altering the retrograde doping profile, which controls the gate depletion depth. This lowers the gate depletion capacitance, reducing the body effect coefficient and subthreshold swing and therefore increases the Ion-to-Ioff ratio.
Paul [65] utilized the above technique on a custom 50nm bulk planar technology node. An improvement of 7mV in the subthreshold swing and reduction of 35% in the junction capacitances was observed. This translated into a 44% reduction in inverter chain delay times and a 51% improvement in the Power Delay Product (PDP).
There are two caveats to this technique. This first is that [65] also observed that at a supply voltage of 800mV, the junction depletion regions converged in the center of the channel and bulk punch-through was observed, resulting in the complete loss of control

of the device. Therefore any form of adaptive voltage scaling such as Ultra Wide Dynamic Voltage scaling [66] is completely prohibited. The second is that the contemporary commercial VLSI designer may only use the technology nodes available from the fabrication houses. Therefore doping profile optimization is confined to academic research and the pragmatic cell library designer is left to derive novel techniques of using existing technology processes, which include HALO implants.

2.5.4 Effect on Substhreshold Characteristics

2.5.4.1 Capacitance

The capacitances of a NMOS device may be seen in Figure 3 [67]:

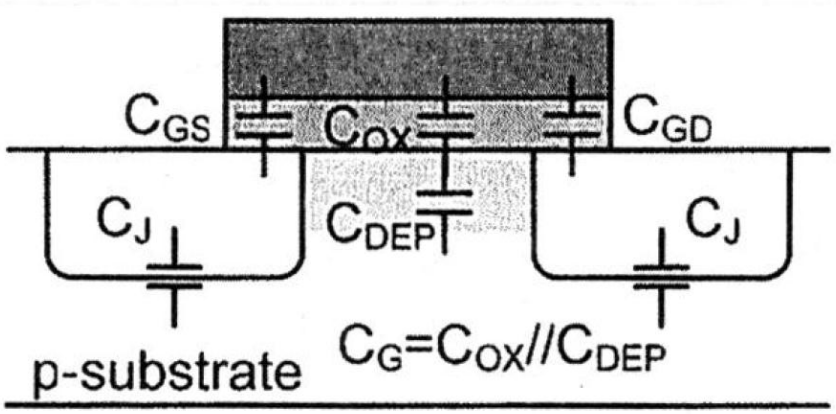


Figure 3: MOSFET capacitances [67]

These may be derived as [67]:

$$C_{DEP} = \frac{\in_{si}}{W_{DEP}} \quad , (23)$$

$$C_{OX} = \frac{\in_{OX}}{t_{OX}} \quad , (24)$$

$$C_{GD} = C_{GS} = WC_{OV} \quad , \text{and } (25)$$

$$C_J = WC_j + (2W + L_j)C_{jsw} \quad , (26)$$

whereby W_{dep} is the depletion width, t_{ox} is the oxide thickness, W is the device width, C_{ov} is the overlap capacitance per width, C_j is the junction capacitance per width, L_j is the junction length and C_{jsw} is the sidewall capacitance.

It may be seen that the capacitance of the device, and by extension it's delay and power consumption, are affected by any geometric alteration. Increasing the length has the effect of increasing the area of the gate, and therefore both C_{ox} and C_{dep} are greater. However, as RSCE lowers the threshold voltage of the device, the channel inverts deeper, increasing the depletion width and therefore lowering the depletion capacitance. The overall affect on the device capacitance when optimized for RSCE is therefore technology dependent.

2.5.4.2 Subthreshold Swing

In the subthreshold regime, the relationship between the drain current (I_{DS}) and applied gate voltage (V_{GS}) is known as the subthreshold swing and may be derived as:

$$S = m\frac{kT}{q}ln10\ (mV/dec) \tag{27}$$

Where the body effect coefficient may be derived as:

$$m = 1 + \frac{C_{DEP}}{C_{OX}} \tag{28}$$

As the depletion depth is increased when sizing is optimized for RSCE, the depletion capacitance is reduced. Critically, this reduces the subthreshold swing and therefore increases the Ion/Ioff ratio in the subthreshold regime. The knock on effect is that the performance-to-leakage current metric is improved.

2.5.5 Inverse Narrow Width Effect (INWE)

The Inverse Narrow Width Effect (INWE) shall be broken down into three subsections; Historical geometric width effects, INWE and INWE's effect on Subthreshold characteristics.

2.5.6 Historical Geometric Effects

2.5.6.1 Narrow Width Effect (NWE)

Historically CMOS technologies were isolated with (Local Oxidization of Silicon) LOCOS structures. This technique was semi-recessed, with isolation tapering in the form of the 'bird's beak' from the thin gate oxide into the thick field oxide. As a voltage was applied to the gate, the thin gate oxide depletion region was allowed to extend outwards and under the tapered isolation into the surrounding substrate/well. This created a parabolic depletion region that deviates away from the ideal depletion depth [68]. As the device width was reduced, a greater proportionality of the depletion extended underneath the tapered oxide, the deviation away from the ideal increased, resulting in a higher voltage required at the gate for the same level of channel inversion and therefore the threshold voltage of the device increased. This process is known as the Narrow Width Effect (NWE) [69].

2.5.7 Inverse Narrow Width Effect (INWE)

Supplementary to the aforementioned electrical characteristic with LOCOS, the device density restriction for CMOS was overwhelming by the 1980s, leading to transitional works in reducing the tapering [70] and eventually giving way to fully recessed isolation techniques such as Shallow Trench Isolation (STI). This provides isolation in the form of an orthogonal SiO_2 sidewall between the thin gate oxide and thick field oxide.

As the critical dimension (lambda) in a self aligned process is determined by the gate length, metal gave way to highly doped polysilicon as the material of choice for the gate contact. This is due to the fact that it is easier to geometrically control during processing. However, polysilicon is susceptible to 'claw-back'. This is the process of the polysilicon withdrawing towards the center of the channel away from the 'drawn' dimensions. To ensure adequate coverage of the gate, a minimum overlap between the gate and thick field oxide is enforced in the technology node DRM (Design Rule Manual).

These two processing methodologies combine to create a physical effect on the threshold voltage, which may be derived from first principles by applying the conservation of charge such that [71]:

$$Q_M + Q_F - (Q_N + Q_B) = 0 \quad (29)$$

Whereby Q_M is the charge on the gate, Q_F is the fixed charge in the dielectric, Q_N is the charge of the free carriers in the channel inversion layer and Q_B is the depletion region ionized impurity concentration. On application of Gauss's Law, the gate voltage for an n-type device may be derived as [71]:

$$V_G = V_{FB} + \psi_S + Q_B / C_G \quad (30)$$

Whereby V_{FB} is the flat band voltage, ψ_S is the surface state potential, and C_G the gate capacitance. At the point of inversion, the surface state potential is double the bulk potential. The threshold voltage may therefore be derived as [71]:

$$V_{TH} = V_{FB} + 2\phi_B + Q_B / C_G \quad (31)$$

Whereby the bulk potential ϕ_B may be derived as:

$$\phi_B = kT /q(\ln N_{SUB}/N_i) \quad (32)$$

Whereby N_{SUB} is the substrate dopant density and N_i the intrinsic dopant density. The third term of Equation 32 may be rewritten to represent the voltage dropped across the gate oxide:

$$V_{TH} = V_{FB} + 2\phi_B + E_{OX}T_{OX} \quad (33)$$

Whereby E_{OX} is the electric field dropped across the gate oxide and T_{OX} is the gate oxide thickness.

Where the polysilicon overhangs the thick field oxide, the electric field still penetrates into the channel through the STI. This is known as the fringing field. The fringing field's impact on the threshold voltage may be compacted into the concept of the fringing factor [72]:

$$F = (4 T_{OX} /\pi)\ln(2T_{FIELD}/T_{OX}) \quad (34)$$

Whereby T_{FIELD} is the field oxide thickness. This may then be added to derive the fringing field impacted threshold voltage:

$$V_{TH} = V_{FB} + 2\phi_B + E_{OX}\frac{T_{OX}}{1 + F/W} \quad (35)$$

Whereby W is the device width. It may be observed both analytically [73] and anecdotally that as the width is decreased, the proportion of the channel influenced by the fringing field increases. This provides a deeper depletion at the channel edges and therefore results in a lower threshold voltage, increasing the active current. As lowering the threshold voltage also increases the leakage current [74], the impact on the Ion/Ioff ratio must be closely monitored.

2.5.8 Effect on Subthreshold Characteristics

2.5.8.1 Subthreshold Swing and Capacitances

As the depletion depth is increased by INWE, the depletion capacitance and therefore gate capacitance decreases in the same manner as for RSCE. Moreover, as the optimal width provides the highest drain current, the area of the gate is minimized. This means that the gate capacitance is decreased further.

2.5.8.2 Exploiting Parallelism – Fingering

The logical conclusion drawn thus far in the field is that if the minimum width provides the biggest current to capacitance benefit, then gate strengths greater than that of a single minimum width device should be constructed from parallelized minimum width devices, a technique that has come to be known as fingering. Whilst almost all research that has been conducted in subthreshold standard cell libraries uses this technique, its efficacy over the full voltage range is explored empirically for the first time in this thesis.

2.5.9 Stack Forcing

As outlined earlier, the current of a subthreshold device may be derived as:

$$I = I_0 \frac{W}{L} e^{\frac{V_{GS} - V_{th}}{mV_T}} (1 - e^{\frac{V_{DS}}{V_T}}) \quad (36)$$

Whereby I_0 is the technology dependent subthreshold current for $V_{GS} = V_{th}$, W and L are the width and length respectively, V_{th} is the threshold voltage, m is the technology dependent parameter derived in Equation 28 and V_T is the thermal voltage = kT/q.

The currents of a single device and composite device consisting of multiple, serially stacked devices may therefore be approximated as [75]:

$$I_{single} = I_0 \frac{W_{single}}{L} e^{\frac{V_{GS} - V_{th\ single}}{mV_T}} (1 - e^{\frac{V_{DS}}{V_T}}) \quad (37)$$

$$I_{stacked} \approx \frac{1}{N} I_0 \frac{W_{stacked}}{L} e^{\frac{V_{GS} - V_{th\ stacked}}{mV_T}} (1 - e^{\frac{V_{DS}}{V_T}}) \quad (38)$$

Whereby N is the number of stacked devices. In order to determine the width upscaling required for the stacked network to match the current of the single device, the following equation may be solved:

$$\frac{W_{stacking}}{W_{single}} = N e^{\frac{V_{t\ stacked} - V_{t\ single}}{mVt}} = N e^{\frac{\Delta V_t}{mVt}} \quad (39)$$

In a 65nm technology node [75], ΔV_t and mV_t are approximately 45mV and 35mV respectively for N = 3 transistors. This gives rise to a width upscaling factor of around 10. This is simply too large of an area overhead to be feasible. Failing to perform this upsizing will have severe ramifications on the noise margin to the network of stacked devices. Moreover, stacking in subthreshold has a greater effect on the active drive current than the leakage current, therefore the Ion/Ioff ratio is also degraded by stacking. For these reasons, transistor stacks greater than two are typically prohibited in the subthreshold regime.

Conversely, this relationship may also be utilized. By stacking two devices where one would suffice, one device will be pushed into supercutoff (V_{GS} less than 0) thereby reducing the leakage. The disadvantage to this is that the composite device is slower than the single device. This technique is known as stack forcing [76].

2.6 Subthreshold Library Design Studies

2.6.1 RSCE Singular Implementation

Kim [67] performed RSCE optimization on the International Symposium on Circuits And Systems (ISCAS) benchmark circuits in a commercial 120nm bulk planar technology node. A 10.4% improvement in delay was observed due to the reduction in gate capacitance. This also created a power saving of up to 39% and a subsequent energy reduction of 41.2%. An improvement in the subthreshold swing of 16mV/dec was observed, translating to a 30% reduction in off-current iso on-current. The overall benchmark results therefore showed a larger improvement in leakage power reduction than in dynamic power reduction. As RSCE increases the gate area, an improvement in delay and power consumption variability was also observed, with improvements in the sigma/mu ratios of 37.5% and 70% respectively. [77] performed inverter chain analysis in a 40nm commercial technology node, utilizing HVT transistors. Delay reduction of up to 53% were observed translating into a 71% improvement in active energy consumption and a 68% reduction in standby power.

2.6.2 RSCE/INWE Combination (Minimum Width Sizing)

Liao [78] identified the major challenges of subthreshold cell design as robustness under variation and Ion/Ioff ratio optimization and attempts subthreshold cell improvement by combining both RSCE and INWE in a 180nm bulk planar technology.
The methodology employed geometric sweep simulation using a compact transistor model to determine optimal threshold voltage. The key observations made were:

1. No impact on threshold voltage for transistor fingering. This is logical from the discussion presented in this chapter, as fingering is simply repetition of the same underlying device.
2. The threshold voltage increased between minimum width and 3X minimum width. This is characteristic of the INWE.

3. The threshold voltage decreased from minimum to 10X minimum length. This is characteristic of the RSCE. However, 10X minimum length is a overly large limit for the underlying physical effect caused by the HALO implants.
4. PMOS was affected less than NMOS for INWE. This may be technologically dependent.
5. RSCE variation and threshold voltage variation decrease with increasing gate area. This follows the known relationship with decreasing RDF.

Geometric sweeping was also performed to determine optimal drain current per gate capacitance. The technology used exhibited optimal current to capacitance at the minimum geometry and no degradation in this figure-of-merit through the process of fingering. The study therefore proceeds by constructing cells of parallel minimum (220nm Width/ 180nm Length) devices.

The study performs secondary optimization in the form of parametric analysis to determine the optimum number of fingers for an X1 Inverter, NAND2 and NOR2 gate. This analysis is constrained by the author to at least two fingers per network under the observation that single finger networks would display prohibitive variation. These are then simulated in an FO4 test bench at 400mV and compared against conventionally sized equivalent gates. The subthreshold geometrically optimized gates exhibit a reduction in energy delay product of 72%. Whilst this study is demonstrative, all simulation is performed pre-layout.

Pons [79] performed a similar study in a 180nm technology node. Geometric sweeping is performed and a reduction in rise times of 3X is observed by upsizing the device length from 180nm to 400nm due to the RSCE. The minimum width is shown to produce the best drain current to gate capacitance and therefore cells are constructed from minimum width fingering. An existing low power library is modified by dropping cells with inputs larger than three and the aforementioned sizing alteration to produce a reduced subthreshold optimized standard cell library of 45 cells. These are then characterized at 400mV and 1V, over three process corners (SS/TT/FF) and over three temperatures (-40C, 25C, 125C) using Liberate. Two test circuits are then synthesized and routed using Encounter RTL and Sock Encounter respectively.

No meaningful delay/power/energy advantage could be gleamed from the study simply due to the failure of the standard library at 400mV although similar improvement as in [78] would have been expected. Area increases of between 1.5X and 1.9X were observed.

2.6.3 Constant Yield Library

Kwong [80] postulated the critical factor in subthreshold library design is yield. In order to meet a fixed yield the study identifies two primary metrics of concern; failure from insufficient output swing and current variability. The study concedes that delay variability is also of utmost importance, however both current variability and delay are lognormally distributed with the same standard deviation (σ). Delay variability is therefore a result of current variability.

The study presents a methodology of using butterfly plots of back to back gates in order to determine failure by observing the magnitude of inscribed squares, analogous to determining static noise margins in 6T RAM cells. To determine the output low voltage of a gate (VOL), the gate may be placed back to back with the largest stacked NOR gate in the library, as this has the most stringent input low voltage level (VIL) due to the stacking of NMOS devices in the pull down network and parallel PMOS device in the pull-up network. By the same reasoning VIL may be determined using the largest stacked NAND gate. These test benches are then simulated in 5k-monte carlo runs at the worst-case temperature corner. The threshold voltage of the transistors is then randomized with both local and global variations. This is performed across a voltage sweep and the failure rates determined.

In the 65nm technology node chosen in the study, the trends observed from the above test showed a decrease in failure rate of approximately 4X in a standard inverter simply by upsizing the gates from minimum width to 1.66X minimum. The trends also show that in a 5k run, failure may be completely eradicated by simply upsizing the device widths or increasing the supply voltage.

The study analytically derives the coefficient of active current in the subthreshold regime as:

$$\frac{\sigma I_{sub}}{\mu I_{sub}} = \sqrt{e^{\left(\frac{\sigma V_T}{nV_{th}}\right)^2}} - 1 \qquad (40)$$

Whereby σV_T is the normally distributed standard deviation of the threshold voltage, n is the subthreshold swing factor and V_T is thermal voltage. The study highlights the fact that as V_{DS} is lowered, the subthreshold swing factor decreases. As V_{DD} is shared in stacked devices they are less susceptible to current variation. Increasing width also has the same effect. This is simulated on 1/2/3 stacked NMOS/PMOS devices and demonstrated within the study.

The study therefore postulates a subthreshold sizing strategy of defining a fixed yield, and based on the two aforementioned metrics, sizing the width based on the minimum size required to meet the yield target. The minimum sizing optimization guarantees that this will offer the greatest energy efficiency whilst meeting the yield criterion. The stack effect displays higher variation in output swing but lower current variation. Given the former could generate functional errors, it is prioritized over the timing uncertainty introduced by the latter.

To test the strategy, a 32-bit Kogge-Stone adder is synthesized. A 3-Sigma failure rate of 0.13% is targeted and the supply voltage swept. The devices are then upsized in the gates until the yield target is met. The results show that even for a simple adder, the minimum sizing strategy only meets the yield target down to approximately 340mV. Beyond this critical point, the device widths must be upsized dramatically, with a VDD of 250mV requiring at least a width 1.5X times the minimum to meet the yield target.

2.7 Chapter Summary

This chapter explored the field for the physical considerations required when creating a subthreshold bulk planar standard cell library. These were broken down into robustness, variability and performance considerations and the magnitudes of their impact on library design expressed from studies conducted by other researchers.

The chapter then presented subthreshold library design studies, highlighting what the corresponding researcher attempted to achieve, how successful they were at achieving their aim and how their results compared with other studies.

Chapter 3: Subthreshold Bulk Planar Semiconductor Physics

3.1 Outline

The work undertaken throughout this chapter was to apply the geometric (RSCE/INWE), stacking and parallelization concepts outlined in the previous chapter to a commercial deep submicron bulk planar technology node. The concepts were first applied to the device level via extensive simulations on BSIM4.5 compact models. The fundamental parameters of drive current, leakage current and the Ion/Ioff ratio were determined over SS/TT/FF process corners. The impact of stack forcing on these parameters was then examined. Parallelization comparisons were made. The simulated devices were placed into an inverter gate configuration. Various SPICE level test benches were performed to determine minimum operating voltage, gate capacitances, junction capacitances and propagation delays. Sample inverters for the entire design space in a 12 track library were then laid out, DRC (Design Rule Check) and LVS (Layout Vs Schematic) validated, parasitically extracted and simulated for propagation delay, leakage current and variation via Monte Carlo analysis.

3.2 Device Level Simulation

All simulation work was completed using release 14 of the TSMC 65lp Process Design Kit (PDK). This includes a BSIM4.5 compact model, which models WPE (Well Proximity Effect) and LOD (Length of Oxide Definition). The manufacturing grid in the process node is 5nm. The minimum device length is 60nm. The minimum device width is 120nm. Synopsys Hspice was used to simulate the spice deck test benches.

3.2.1 Nominal (Superthreshold) Simulation

To provide validation on the models and methodology, a test bench was created to simulate the geometric physics of the model under nominal conditions (1.2V / TT / 25°C). The test bench included a single transistor from the TSMC BSIM model kit with all four terminals (VDD/VSS/GATE/BODY) connected to the relevant VDD/GND supplies. The perimeter and area values of the BSIM transistor model (PD/PS/PS/AS) were mathematically calculated by the test bench from the provided lengths and widths. An RVT NMOS device was swept over lengths 60nm to 1000nm and widths 120nm to 1000nm. The drain current (I_d) was then simulated under each geometric test case. Figure 4 shows the results.

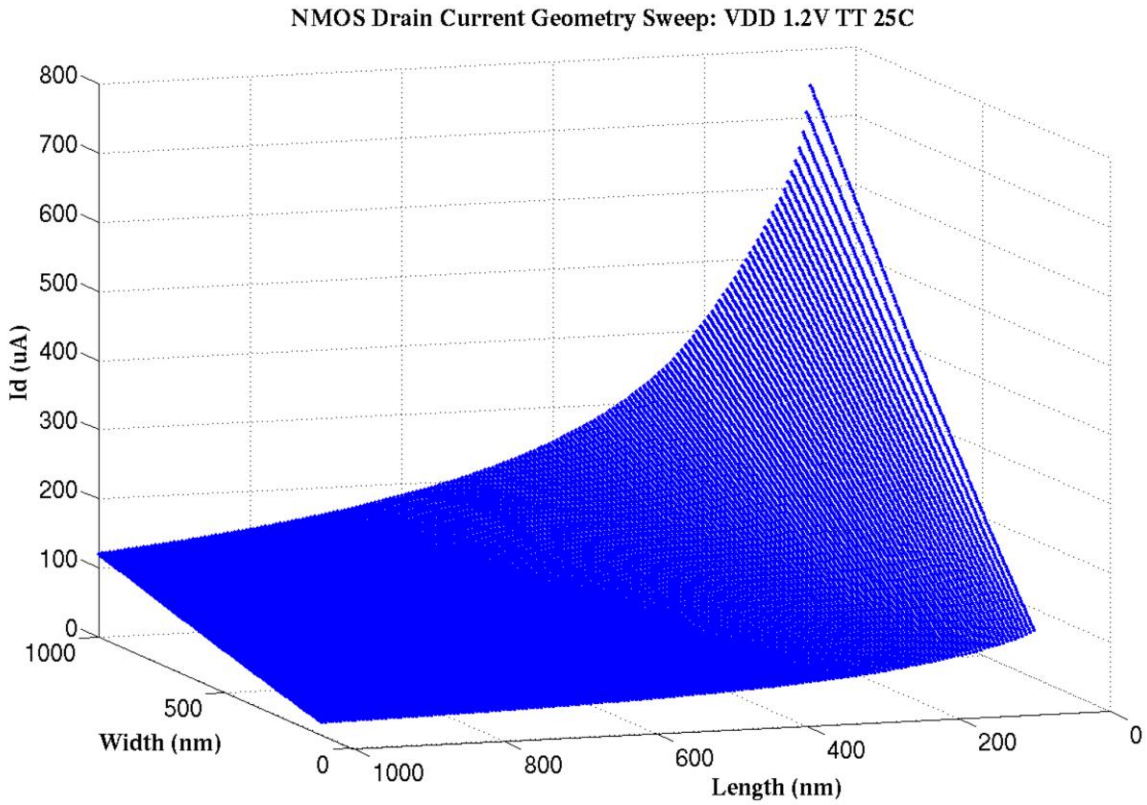


Figure 4: Nominal voltage active current sweep

The monotonic fall in drain current from minimum length is characteristic of the short channel effect (Section 2.5.2.3), the dominant length effect in submicron bulk planar technologies under nominal operation. The linear rise in drain current over width is also accurate, verifying the test bench and model under these operating conditions. From the figure, the conventional sizing choice of a minimum length device in order to maximize current capability and transistor density is verified. The conventional use of aspect ratio (Length/Width) in nominally operated sizing strategies for different device strengths is also clear owing to the simplicity of the linear rise in drain current according to device width.

3.2.2 NMOS Subthreshold RVT

The same test bench was then used to perform the same geometric sweep with the supply voltage lowered to 250mV (VT = 402mV). Test cases were performed for SS/TT/FF process corners. The test bench was then modified to connect the gate terminal to ground and the same sweeps performed to determine the leakage current. Finally the data of both test benches was used to calculate the Ion/Ioff ratio for the sweeps. This data can be seen in Figures 5, 6 and 7.

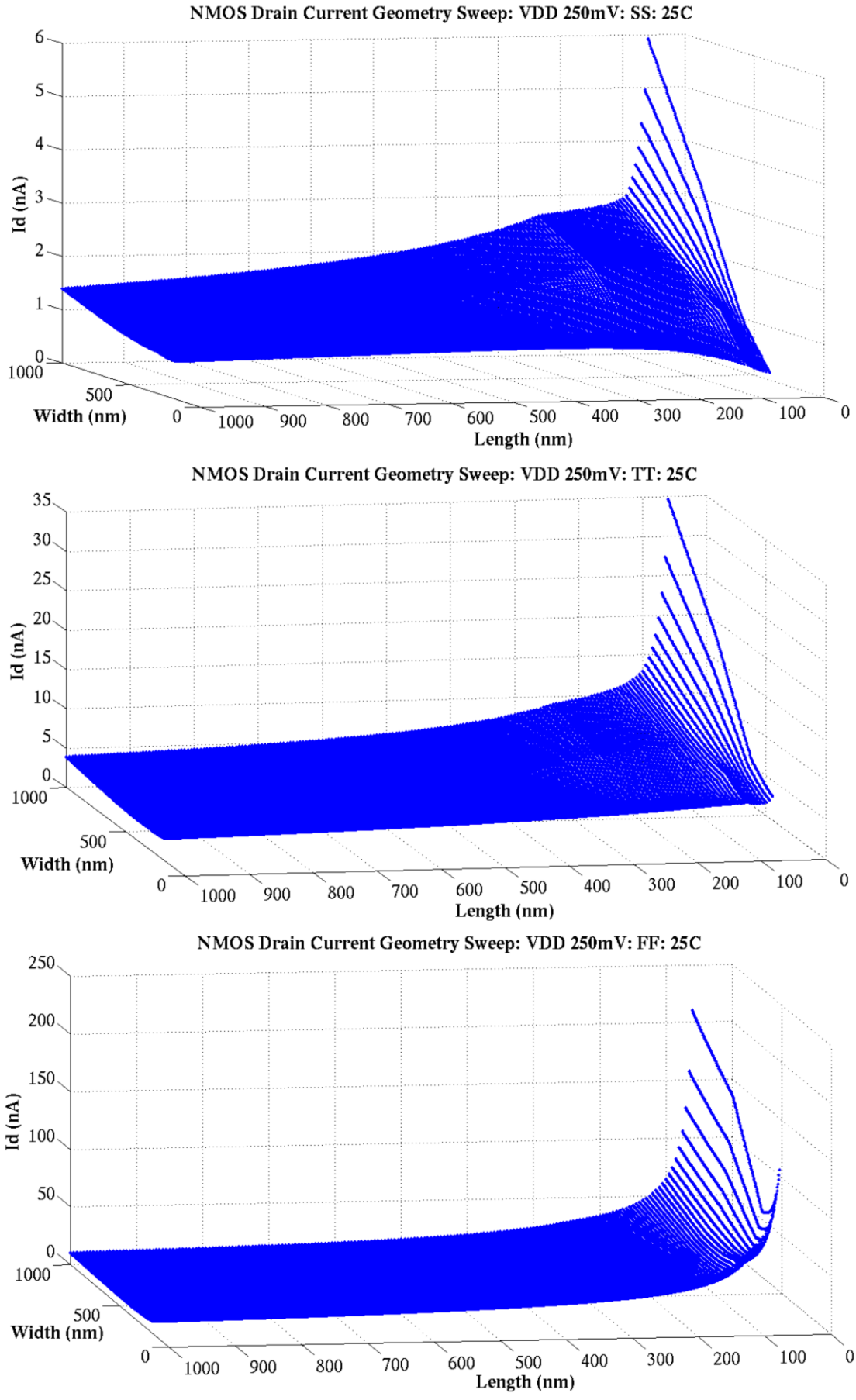


Figure 5: Subthreshold NMOS RVT active current sweep

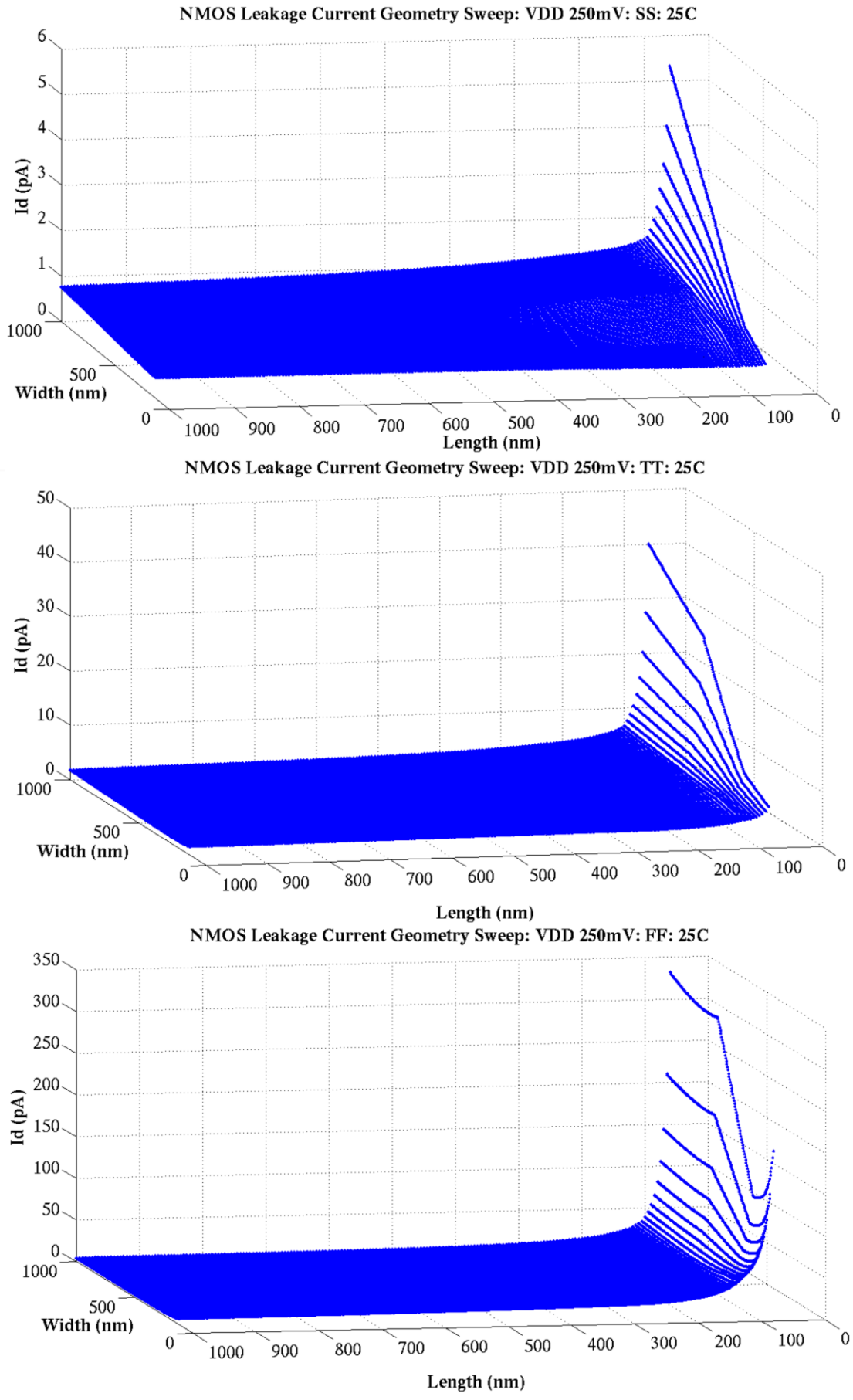


Figure 6: Subthreshold NMOS RVT leakage current sweep

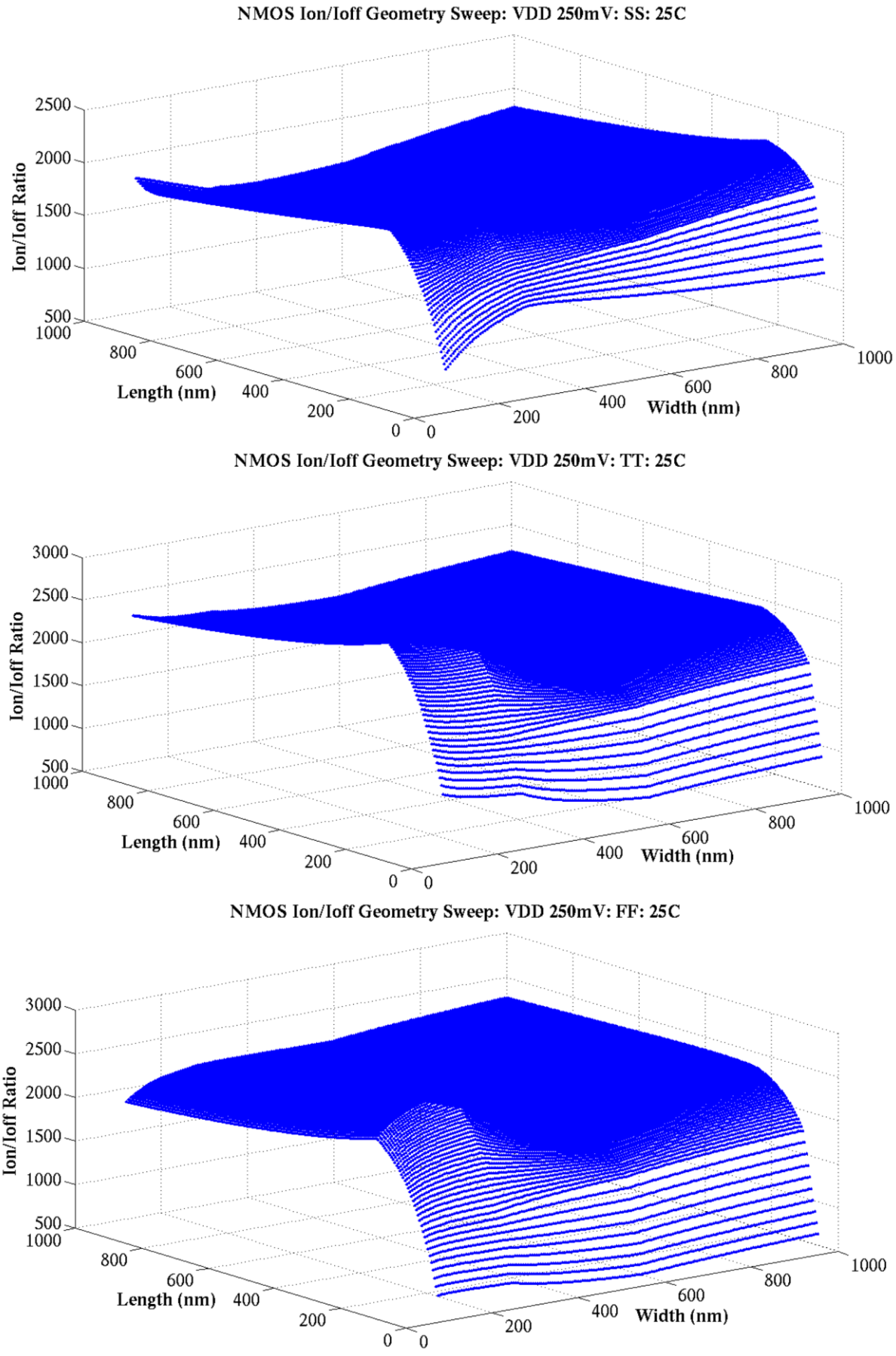


Figure 7: Subthreshold NMOS RVT Ion/Ioff sweep

Figure 5 shows the vestiges of the short channel effect still manifest in the drain currents across all process corners. In a self-aligned deep submicron bulk planar process, high dopant density source and drain regions are implanted via an ion implantation step, as described in the previous chapter. Due to the [100] orientation of the silicon monocrystalline lattice structure (used to minimize threshold variation by reducing variations in surface state potential), the wafer is angled at around 7% during ion implantation to prevent channeling (dopants travelling deep into the substrate due to collision avoidance with the lattice structure). Ion implantation is also used to create the retrograde dopant profile of the channel in order to prevent bulk latch up and set up the characteristics of the device. In terms of global variation, the variation in channel dopant depth and overall dopant density in the channel from the above processing step is one of the differences between the SS/TT/FF process corners.

In the SS corner where the channel dopant has remained relatively close to the surface of the device, the dopant density at the channel surface is greater. This requires more voltage at the device's gate to invert it. At minimum width, the SS sweep in Figure 5 shows a typical RSCE, with the drain current increase in response to an increase in length, contrary to superthreshold operation. The higher dopant density means that the INWE effect becomes less observable, but is still present in the width response, defying the linear relationship by offering a greater drain current per unit width as the width tends towards the minimum. Bends and quirks are apparent in the SS and FF plots where smooth continuation would be expected. This is simply a result of the approximation of the actual physics by compact equations in the compact BSIM models. The apparent observable discontinuation is simply a result of the model transitioning from one equation to another.

In the TT sweep where the dopant has successfully created the intended channel profile, a balance between the opposing SCE and RSCE physical effects is observed at minimum width. Deviation in the linearity of the width response is more observable, with a larger increase in drain current per unit width as the width tends towards the minimum.

In the FF sweep where the dopant has channeled deeper into the substrate, the dopant density at the surface of the device is less, requiring less voltage at the gate to invert the channel. The SCE dominates due to the diminishing influence of the HALO dopants that create the RSCE. The monotonic rise in drain current as the length tends towards minimum length therefore resumes. However, the lower dopant density at the surface of the channel means that the additional fillip provided by the INWE is greater, resulting in

not only a deviation in the linear current-width proportionality, but an increase in drain current as the device width tends towards minimum.

Figure 6 shows the leakage current sweeps across the three process corners. Due to the underlying physics, similar features are observed in the leakage current characteristics as are observed in the drive current characteristics. In the SS corner, the HALO implants have less impact on the channel current when the device is off. The balance between the SCE and RSCE is therefore favored more towards SCE. More impact is observed from the INWE with the deviation in the linear width-leakage current relationship deviating greater than in the drive current sweep. In the TT sweep, the SCE has already begun to dominate the length response, with little sign of the RSCE. The INWE is more prominent than in the drive current sweep. Finally in the FF sweep, SCE is completely dominant with a clearly observable monotonic rise in leakage current as the length tends towards the minimum. The INWE is also prominent, with a substantial rise in leakage current as device width tends towards minimum.

By comparing the results from both figures, several overall trends are observed. Firstly, the RSCE is most prominent in the slow (SS) corner and least prominent in the fast (FF) corner. This is logical, given that the underlying physical effect is caused by increased dopant density via HALO implants (Section 2.5.2.3). Should these additional dopants channel deeper into the substrate or have fewer dopants (both occur as process tends to FF) then the impact of the RSCE will begin to diminish. Secondly, the INWE is most prominent in the fast corner and least prominent in the slow corner. This again is logical, given that the underlying physical effect is a fillip analogous to a thinning of the gate oxide at the orthogonal sidewalls (Section 2.5.7). If the dopant density at the surface of the channel is lowered, a deeper depletion and inversion is induced from the additional electric field strength dropped across the gate oxide translating into an increase in current. The typical (TT) corner is therefore a balance between these two physical effects. Thirdly, the impact of the effects are slightly different in the on and off currents. This therefore warranted further processing in the form of Ion/Ioff ratio sweeps, an important metric when determining performance/leakage or performance/energy.

Figure 7 shows similar trends in the Ion/Ioff ratio across all three-process corners. The SS corner sweep shows that Ion/Ioff ratio rolls off as the device length approaches minimum.

There is also a distinct increase in the Ion/Ioff ratio as the width approaches minimum width, after the peak at around 230nm in length (Where the RSCE diminishes).
The TT corner sweep shows that the Ion/Ioff ratio falls off as the device length approaches minimum, however this is deeper than in the SS corner. No increase is observed at the width approaches minimum as was observed in the SS corner.
The FF corner sweep shows the same fall off in Ion/Ioff ratio as the length approaches the device minimum but is even deeper still. This time a decrease in Ion/Ioff ratio is observed as the width approaches minimum after the RSCE peak at around 180nm.
These trends suggest several relationships. The first is that the SCE has a greater impact on leakage current than on the drive current. Therefore this impoverishes the Ion/Ioff ratio as the length tends towards minimum. Given that the SCE is least prominent in the slow corner, the fall off in the Ion/Ioff ratio is least at this corner and greatest in the fast corner.
The second relationship observed is that the INWE also has a greater impact on leakage current than on the drive current. This is shown by the increase to decrease trend in the Ion/Ioff ratio at minimum width as the process corners tend from SS to FF.

3.2.3 NMOS Subthreshold LVT

The same test bench was then used to perform the same geometric sweeps for an NMOS LVT device at 250mV (VT = 314mV). This data may be seen in Figures 8, 9 and 10.

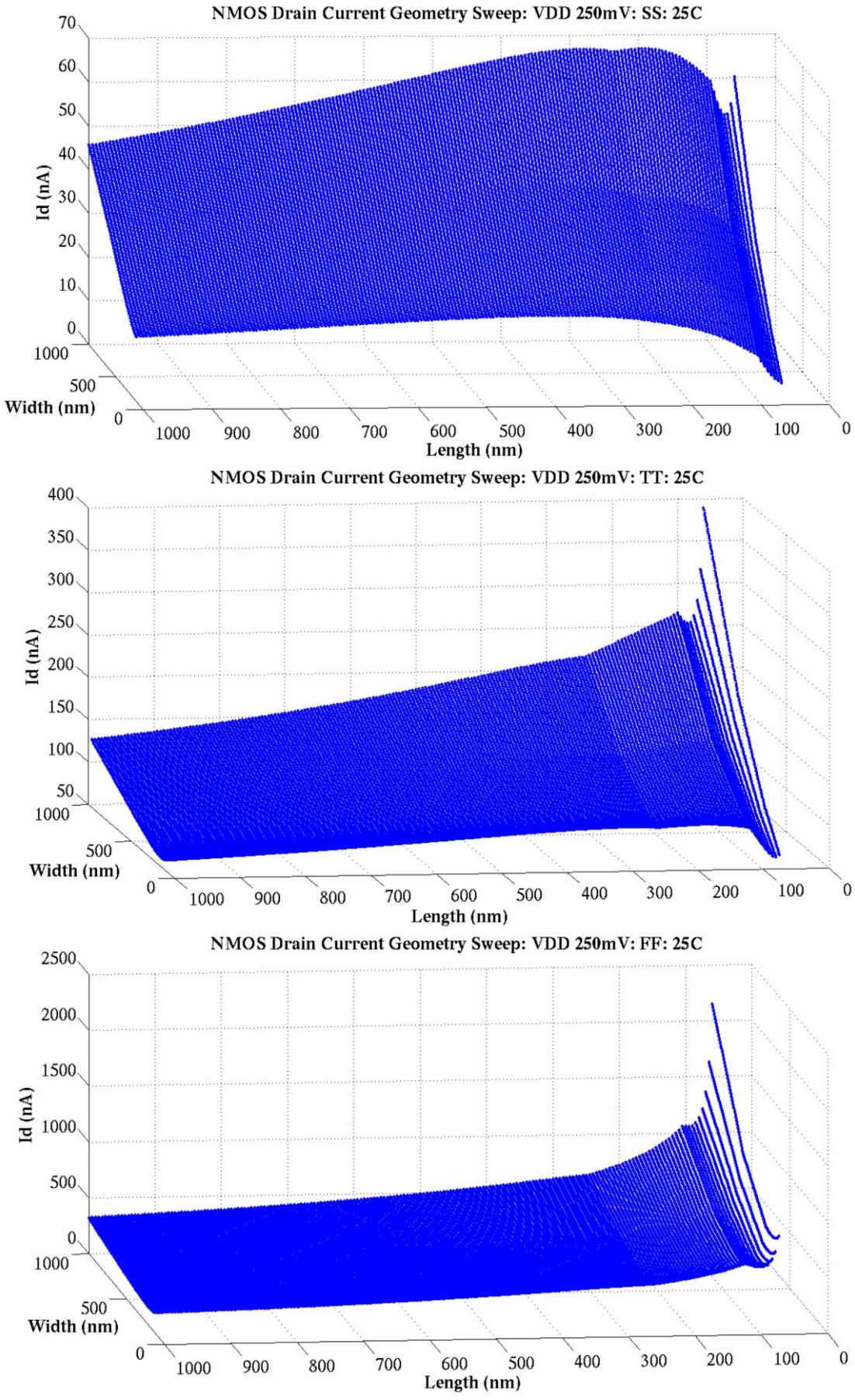


Figure 8: Subthreshold NMOS LVT active current sweep

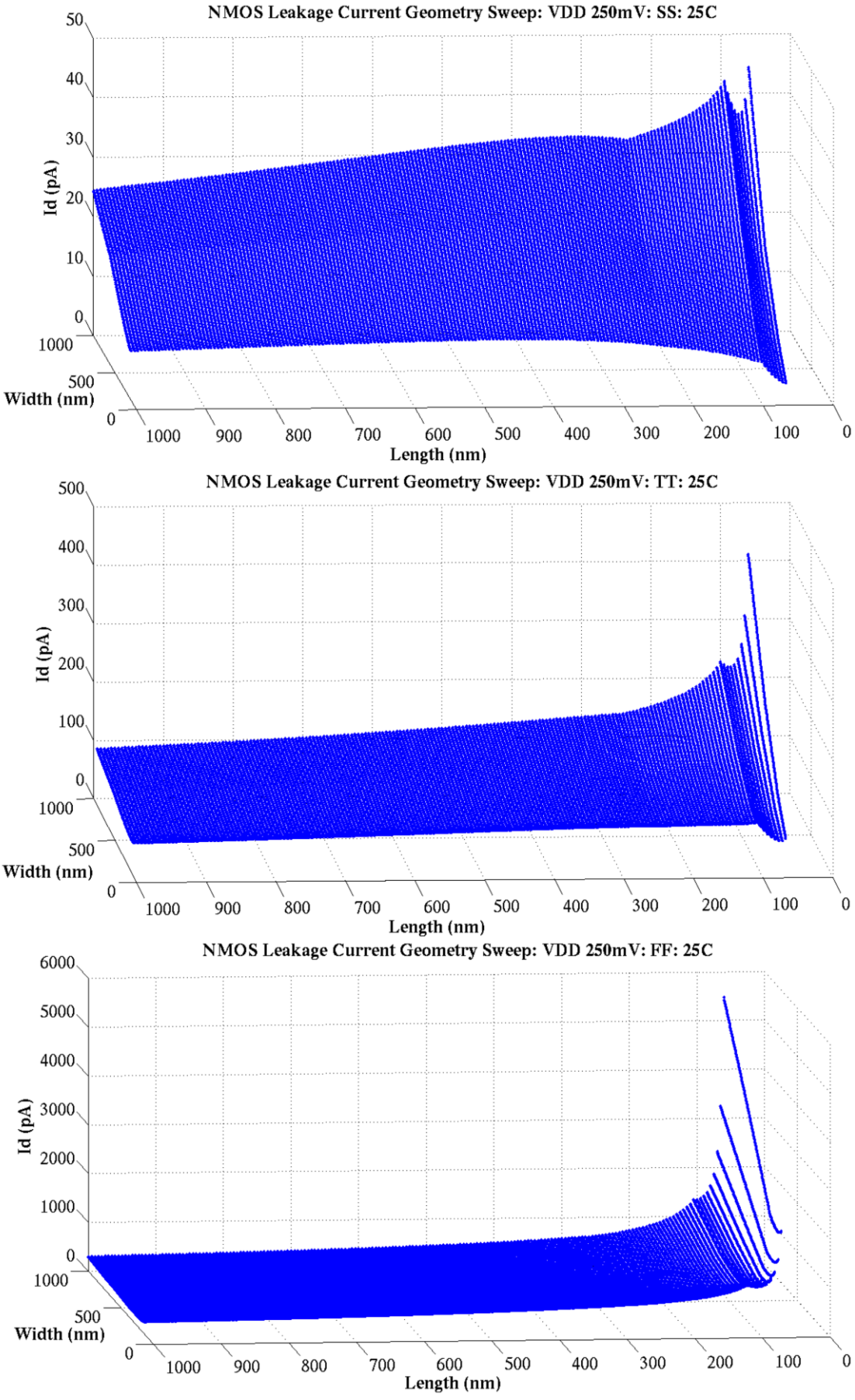


Figure 9: Subthreshold NMOS LVT leakage current sweep

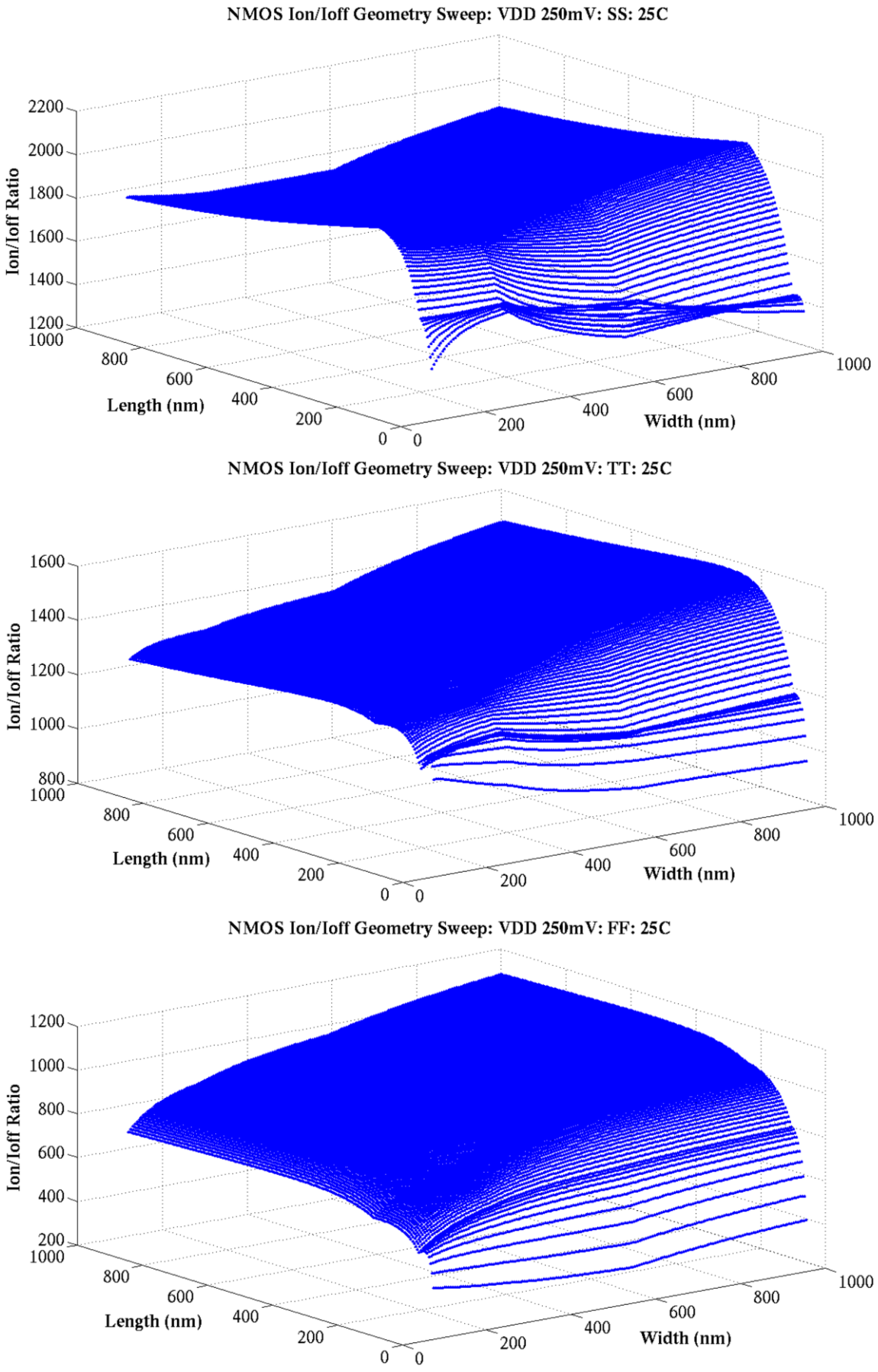


Figure 10: Subthreshold NMOS LVT Ion/Ioff sweep

Figure 8 shows that the dopant density increase imparted by the HALO implants, in comparison to the background retrograde doping profile of the channel, is greater in the LVT NMOS device than in the RVT device. For this reason, the RSCE is comparatively stronger than the SCE and therefore has a greater influence on the geometric drain current response in the LVT device. In the SS corner, the SCE is all but suppressed, with the drain current increasing as the device length increases, eventually peaking and diminishing when the channel center reaches the background channel dopant density. The INWE provides a similar relationship as in the RVT device, with the effect relatively suppressed in the SS corner due to the added effort required to invert the channel.
In the TT corner, the SCE is slightly stronger, but the RSCE still dominates. There is a sharp rise in drain current from minimum to 100nm, then a gradual but clear fall off as the channel center reaches the background channel dopant density. The INWE is more pronounced, with an observable increase in drain current as the device width tends towards minimum.
In the FF corner, the RSCE is still present, but the SCE has started to dominate. The INWE effect is clearly pronounced with an appreciable increase in drain current as the device width tends towards minimum.
Figure 9 shows the LVT leakage characteristics across the three process corners. The SS corner shows a greatly subdued RSCE response and SCE beginning to dominate at larger channel width. The INWE is rather negligible like the drive current response.
The TT corner shows the SCE starting to dominate more of the width. The devices around the minimum width however are still dominated by the RSCE. The INWE is more prevalent with greater deviation in the linearity of the width response and slight increases in leakage current around the minimum width.
The FF corner shows vestiges of the RSCE, but the SCE has started to dominate the entire width. The INWE is clearly visible with considerable increase in leakage current as the width tends towards minimum.
Similar trends across corners are observed in the LVT as were observed in the RVT device. The RSCE is most prominent in the SS corner and least prominent in the FF corner. The INWE is most prominent in the FF corner and least prominent in the SS corner. The TT corner offers the balance between these two trends.
By comparison to the RVT NMOS device, the RSCE is more prominent in the LVT device owing to the increase in proportionality of the HALO dopant density to that of the background channel. Conversely the INWE is more subdued. This may be down to

several processing parameters. However the most likely culprit is change in the surface state potential and flat-band voltage (Section 2.7.3), both of which are derived from the dopant densities. Figure 10 shows the data processed into the geometric Ion/Ioff sweeps. The SS corner shows the same overall trend as the RVT device. The greater dominance of the SCE in the leakage current characteristic has the effect of lowering the Ion/Ioff ratio as the device length tends towards minimum. However, unlike the corresponding SS RVT sweep, the dominance of the RSCE in the drive current interrupts this trend considerably, producing a kink in the response until the INWE dips into the minimum length/width device. The INWE effect is less prominent as the width increases, as would be expected. The TT corner shows a similar trend due to the comparative influence the SCE has on the leakage current over the drive current. There is less interruption from the RSCE as its influence is less than in the SS corner. The INWE effect is more prevalent as the lift visible at minimum width in the SS corner is missing in the TT corner.
Finally in the FF corner, the influence of the RSCE has all but faded, leaving only the characteristic roll-off of due to the SCE. The effect of the INWE is greatest with a clear dip in the Ion/Ioff ratio at minimum width owing to its greater influence on the leakage current. The same overall trends are observed as in the NMOS device. The SCE affects the leakage current greater than the on current, degrading the Ion/Ioff ratio towards minimum length. However in the LVT case there is noticeable interruption due to the influence of the RSCE on the on current. The process dependence is the same, with a higher proportional degradation in the fast corner. The INWE has a similar greater influence on leakage current, resulting in degradation in the Ion/Ioff ratio at minimum device width and is most prominent in the fast corner. Important to note is the absolute magnitudes of the Ion/Ioff ratio in comparison to the RVT device and process corner to corner. The lower dopant density in the channel affects the leakage current greater than the on current. The result is that the Ion/Ioff ratio of a TT RVT device is around 2000 for an RSCE aware length and minimum width, whilst the same device in LVT has an Ion/Ioff ratio of only 1300. This effect is clearly visible as the process corners progress towards the fast corner with the lowest channel dopant density, with the Ion/Ioff ratio falling off to around 700. This indicates a design consisting purely of LVT devices may perform faster than one consisting purely of RVT devices, but that the performance-to-leakage ratio and therefore overall energy consumption will be considerably worse. Moreover this would have a significant impact on circuit topology, restricting designs with parallel transistors.

3.2.4 PMOS Subthreshold RVT

Figure 11 shows the same test bench with an RVT PMOS device with supply voltage at 250mV (Vt = 475mV). The SS corner shows RSCE domination at low widths and SCE domination at high widths. This response is similar to the RVT NMOS device. There is little evidence of the INWE in the width response. The TT corner shows some vestiges of the RSCE but SCE dominant at all device widths. The INWE starts to manifest with deviation in the linear current response as the width tends towards minimum. The FF corner shows SCE dominant across all device widths and the INWE evident by the increase in drain current as the device width tends towards minimum.

Figure 12 shows the leakage current test bench sweeps for the same RVT PMOS device. Cursory observation across the process corners shows little influence of the RSCE, with its impact on the leakage current response diminishing towards the fast corner. The impact of the INWE follows the established trend increasing as the corners progress towards the fast corner.

Figure 13 shows the processed Ion/Ioff ratios for the PMOS RVT device across the three process corners. The results show similar trends to those observed in the RVT NMOS device. The greater propensity of the leakage current responses to the SCE results in the characteristic roll-off in Ion/Ioff ratio as the length tends towards the minimum. This roll-off increases as the corners progress towards the fast corner where the SCE is most prominent. However, unlike the NMOS device, the INWE effect does not degrade the Ion/Ioff ratio. This degradation is offset by the increase in Ion/Ioff ratio as the process corner tends towards the fast corner, opposite to the trend in the NMOS device. There are many physical effects that could cause this, but it is most likely due to the balance of dopant stages from the initial substrate P type doping and subsequent N type NWELL doping. As fabrication houses only provide process models and not direct parameters to protect their intellectual property, the exact cause of this may not be determined.

Overall the characteristics of the NMOS and PMOS RVT devices are very similar. Even though the absolute magnitudes are not the same, the relative trends match very well with local maxima occurring at similar sizes. This is promising as identical sized PMOS/NMOS devices lower the amount of effort required in the layout phase and limit the amount of silicon area wasted in cell design should one device require a substantially different length than the other. It is likely that the fabrication house attempted to match these parameters at full voltage with the HALO implant processing at minimum length and fortuitously, this creates an NMOS and PMOS device with similar RSCE responses.

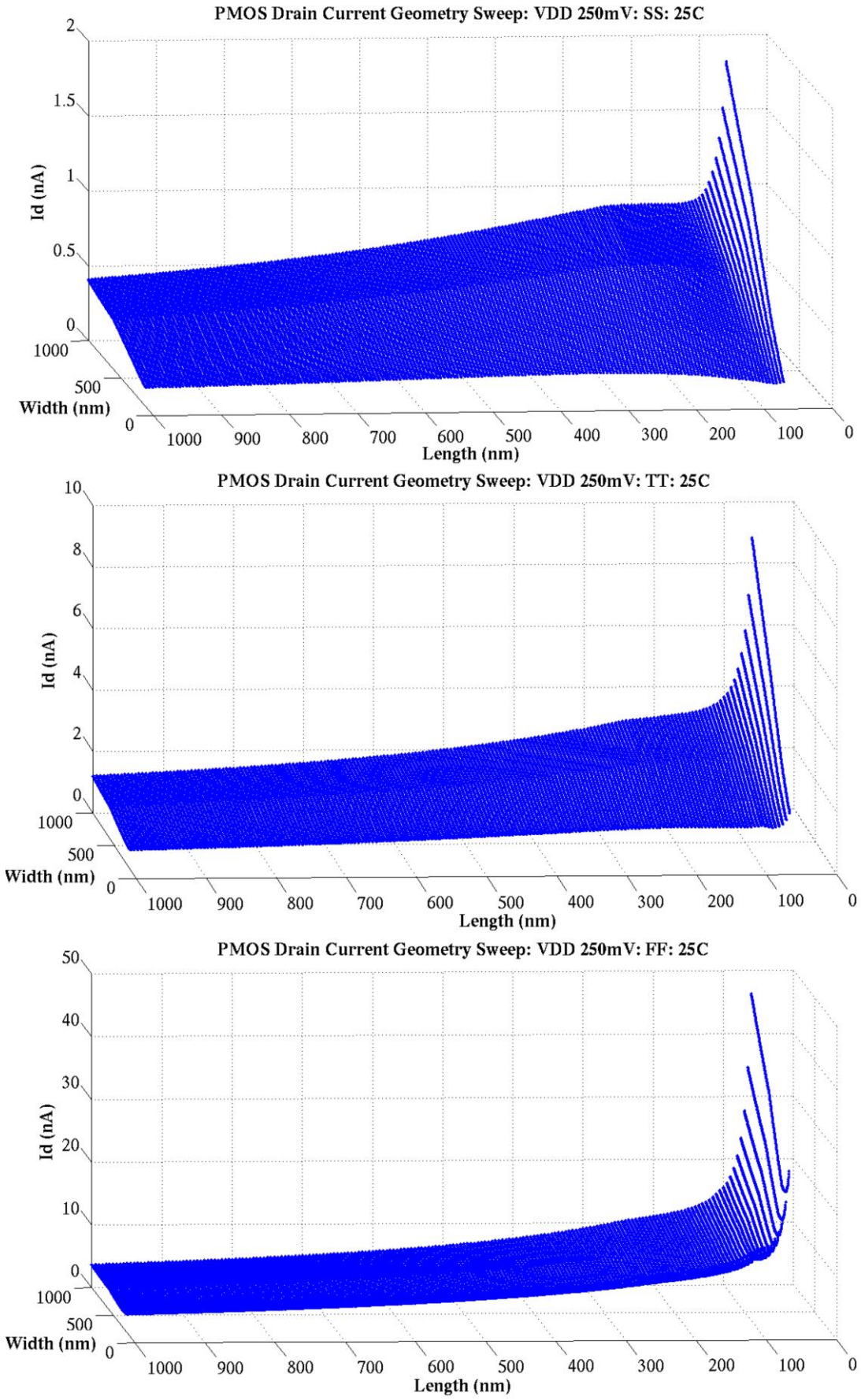


Figure 11: Subthreshold PMOS RVT active current sweep

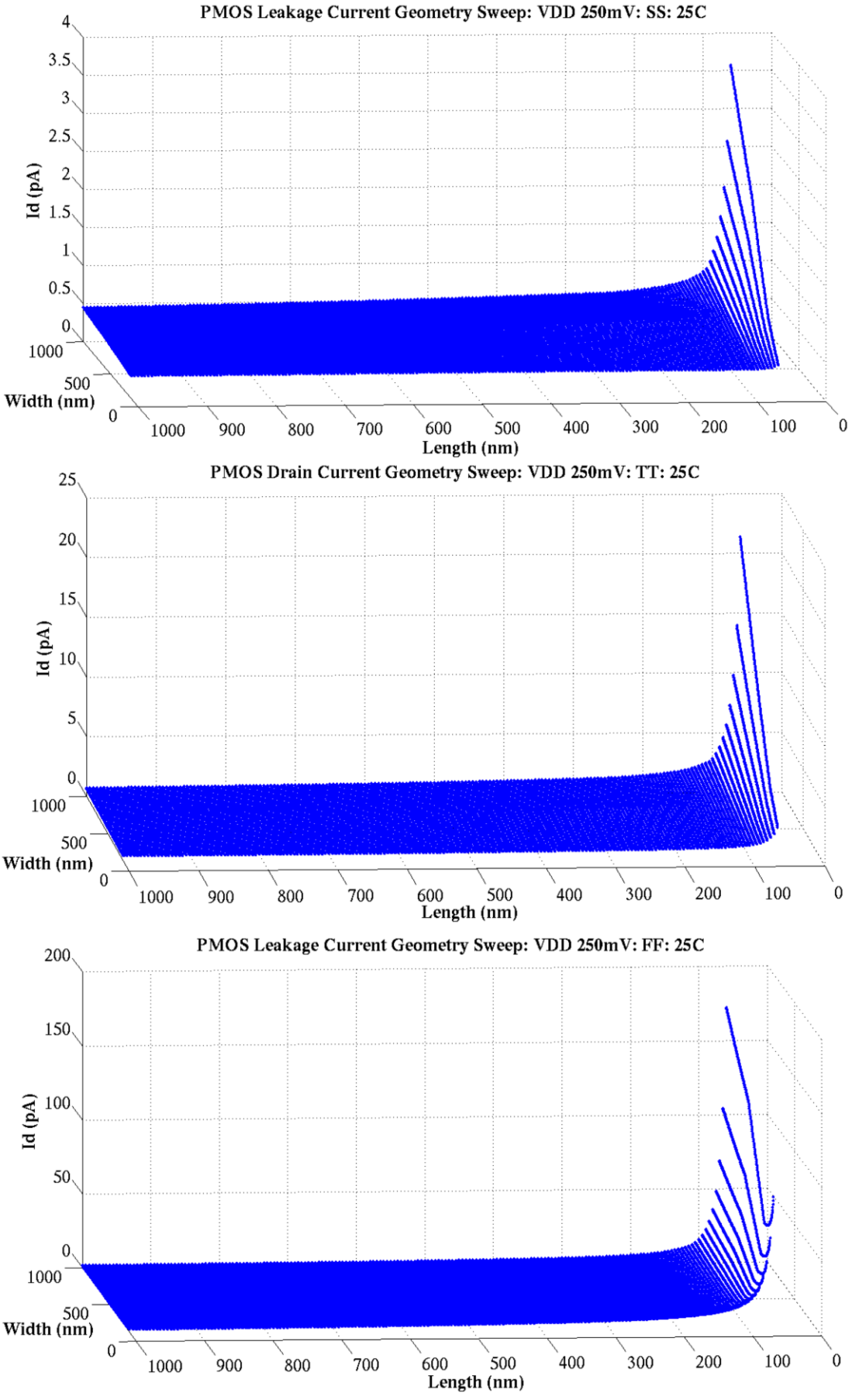


Figure 12: Subthreshold PMOS RVT leakage current sweep

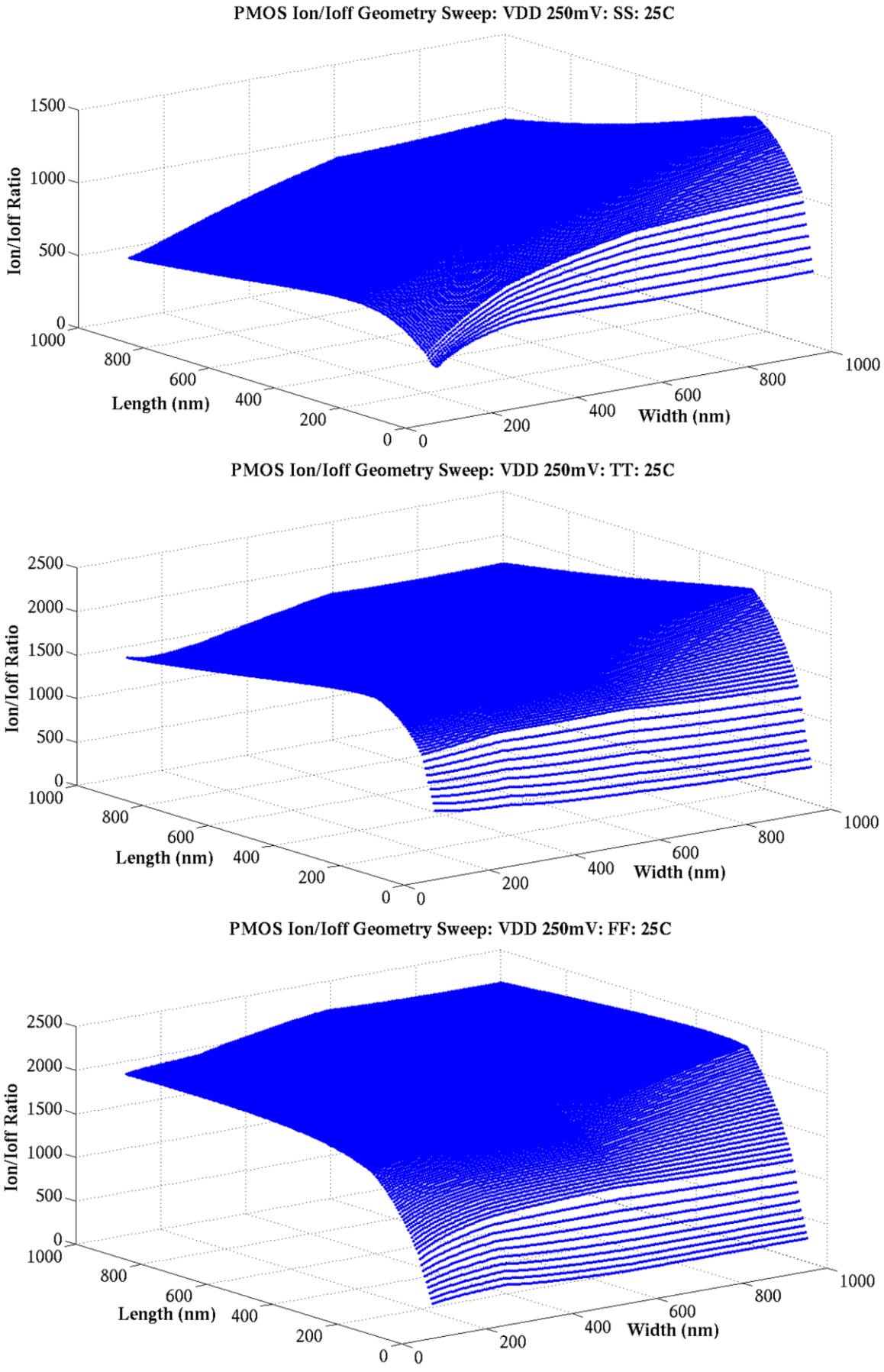


Figure 13: Subthreshold PMOS RVT Ion/Ioff sweep

3.2.5 PMOS Subthreshold LVT

Figure 14 shows the PMOS LVT device in the same test bench with the supply voltage set to 250mV (Vt = 409mV). The SS sweep response shows complete dominance of the RSCE. There is also almost no evidence of INWE, the linearity of the drain current maintained. Interestingly the response shows little current reduction or gain loss in current per unit width after RSCE optimal length. It is possible that there is little impedance from ionic scattering due to a lack of dopant atoms or alternatively that we have simply reached the limits of the compact model. The TT corner shows RSCE dominance but SCE creeping back into the length response. There is a reduction in the linearity of the width response indicating INWE. As in the SS response, there is little current degradation after the RSCE optimal length. The FF corner shows SCE dominance and the remnants of the RSCE. There is greater deviation in the width response at minimum width indicating INWE.

Figure 15 shows the PMOS LVT device in the leakage test bench. The SS sweep shows both SCE at the minimum length and RSCE as the length increases. The same lack of current degradation is observed as in the drive current sweep. Fortunately this means if we have reached the limits of the compact model, this will not affect the Ion/Ioff ratio analysis.

The TT sweep shows dominance of the SCE with some degradation in the width response due to INWE. The FF sweep shows no trace of the RSCE and clear evidence of the INWE.

The same trends observed in other devices are present, RSCE most prominent in the SS corner, INWE most prominent in the FF corner.

Figure 16 shows the processed Ion/Ioff ratios. The same trends are observed as in the other devices. The SCE's greater impact on the leakage current has the effect of rolling off the Ion/Ioff ratio as the device length tends towards minimum. The INWE effect moves in keeping with the rest of the Ion/Ioff shift in response to global process variability and therefore minimal degradation penalty is paid for its use. Due to the anomalous width response in the model, the Ion/Ioff response falls off slightly as the width is increased towards 1um. This may or may not be a limitation of the model. A similar trend to the NMOS device is observed whereby the absolute magnitudes of the Ion/Ioff values degrade as the process tends towards the fast corner. The optimal length values are similar to the NMOS LVT device, again suggesting that full voltage length optimization has probably resulted in similar characteristics in the subthreshold regime.

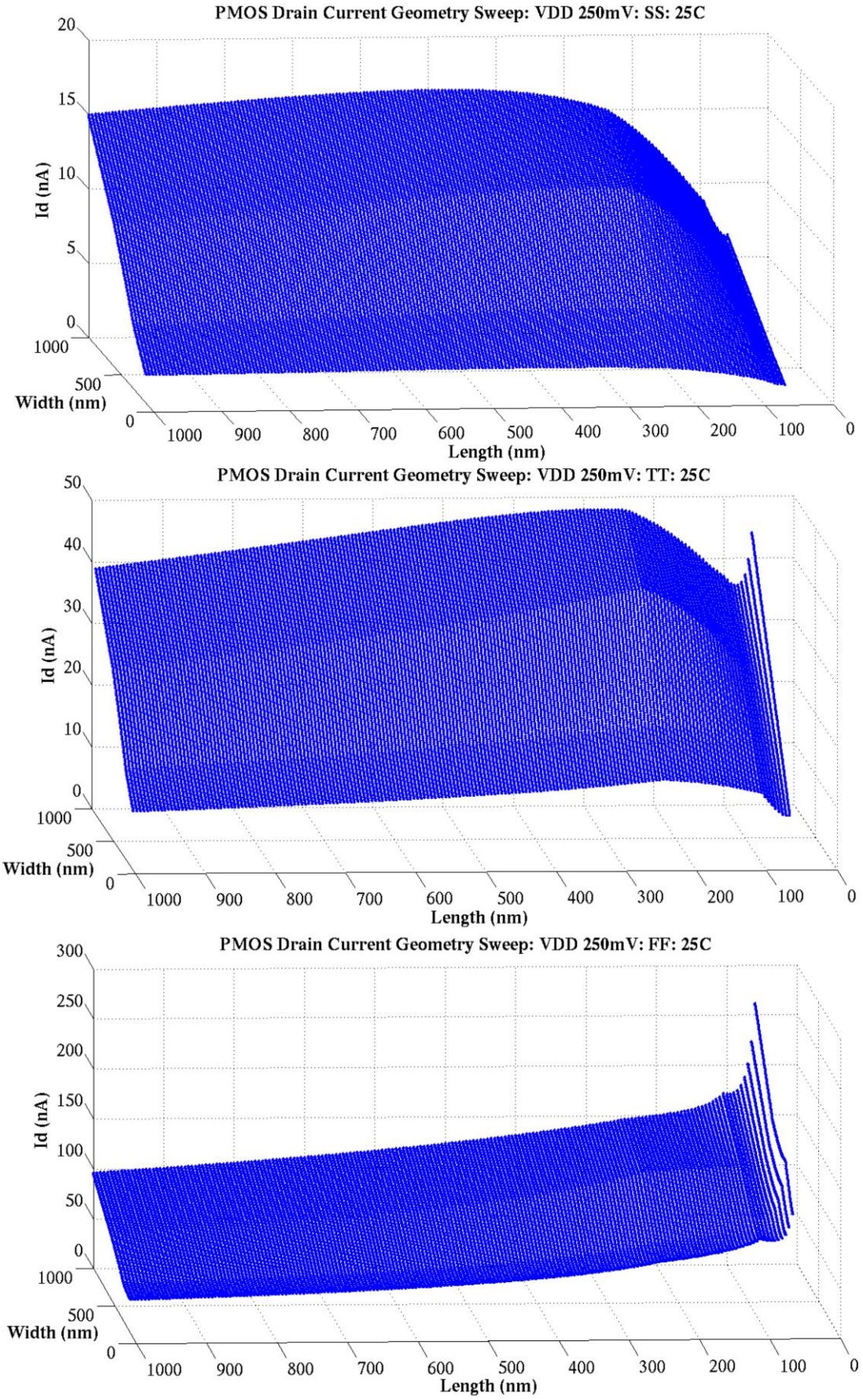


Figure 14: Subthreshold PMOS LVT active current sweep

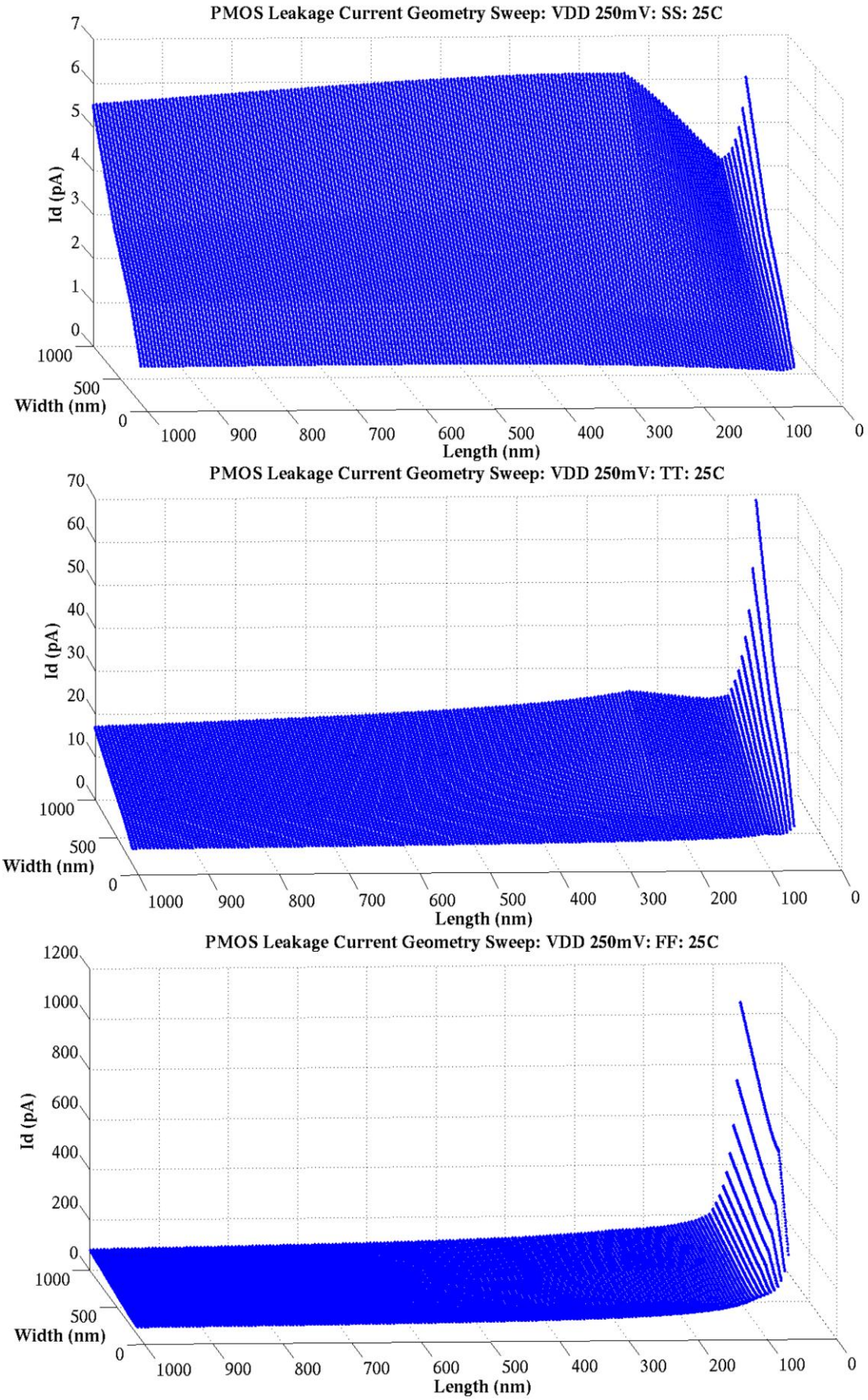


Figure 15: Subthreshold PMOS LVT leakage current sweep

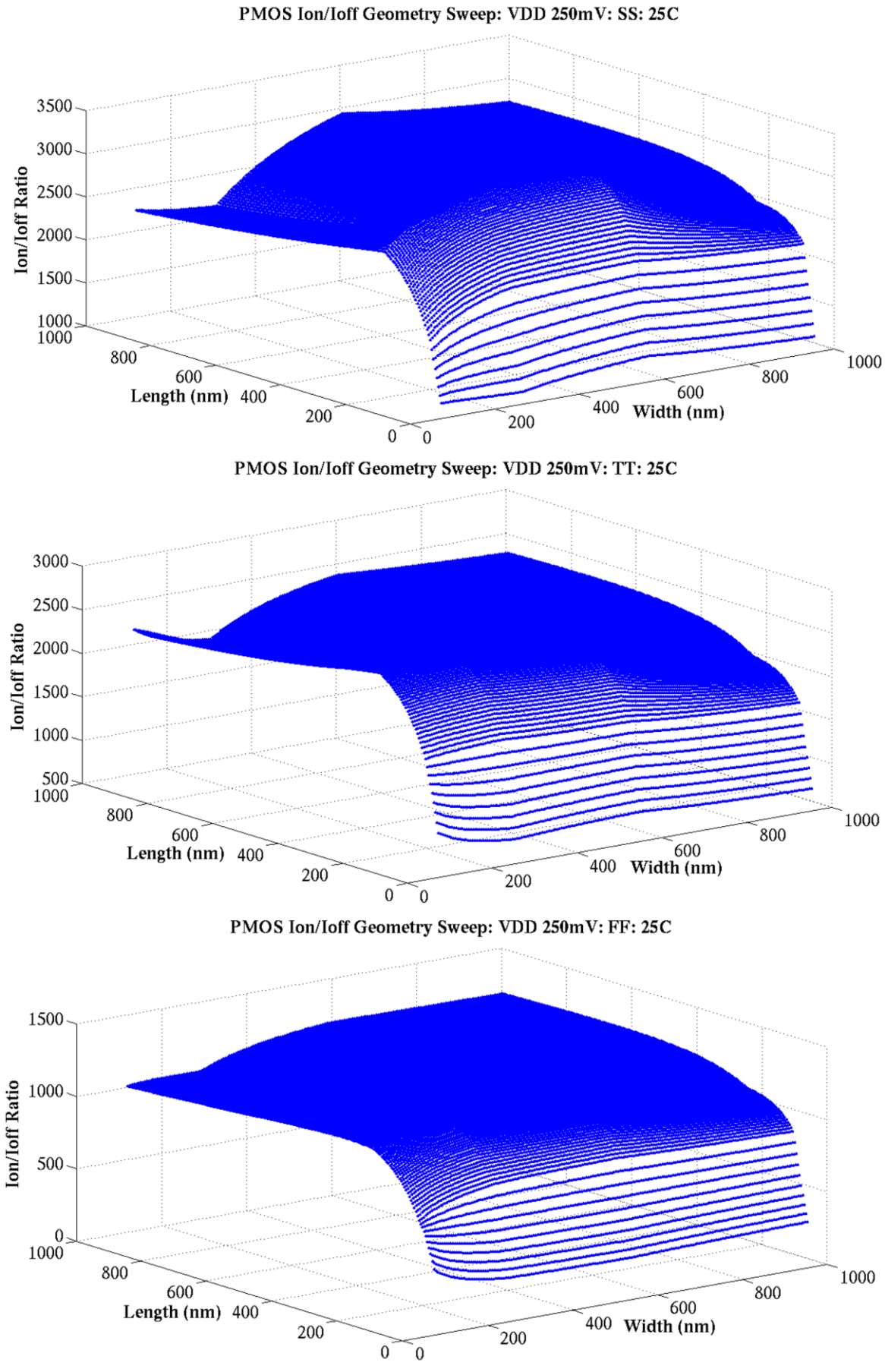


Figure 16: Subthreshold PMOS RVT Ion/Ioff sweep

3.2.6 Single Device Overview

There are several important points to address from the single device current geometric sweeps. The first is that the consensus in the field is that optimal device length sizing in the subthreshold regime is driven by the current increase as a result of the RSCE. These simulations prove that this notion is not accurate. There are several factors that impact the optimal subthreshold device length sizing; the current increase from the RSCE is only one. Given that the objective of subthreshold operation is power and energy minimization, the SCE also play a pivotal role in device length optimization. The simulations showed considerable roll-off in the Ion/Ioff ratio for all devices at all corners as the length approaches minimum. This precludes a choice of minimum length devices, as there would be considerable leakage penalty for any increase in performance at these lengths. The overall energy consumption would therefore increase due to the additional contribution of the leakage energy.

The second point is that to some degree, the INWE appears in all devices. The effect on the Ion/Ioff ratio for each device is different. In some cases the ratio is degraded, in others the ratio remains unaffected. This level of degradation is minimal, and therefore the increase in drive current would justify an INWE aware sizing approach for the width.

In summation, in terms of raw current capability, for this particular technology node, devices should be sized larger than minimum length and with widths that take advantage of the INWE. However, average propagation delay is also dependent on device capacitance. This changes in response to RSCE and INWE and therefore this would also need to be taken into consideration. This relationship is explored later in Section 3.5.

3.3 Device Parallelization (Fingering)

The description of the physics in Section 2.5.5 and single device sweeps show that the INWE exerts greatest impact in the current response when the device width is minimized. The INWE approach explored in the field is to exchange a single, large width device with a parallel stack of minimum width fingers.

To determine the advantage offered by the INWE, the fairest comparison is to determine the current increase possible, iso-area, matching the number of minimum width fingers with the total silicon area required of a single large device. In the given technology node, the minimum device width is 120nm. Therefore each finger will be set to this width. The minimum spacer distance (the STI width between fingers) is 110nm. Therefore a device of two fingers (2 x 120nm + 110nm) is compared against a single device of 350nm. A

device of three fingers (3 x 120nm + 2 x 110nm) is compared against a single device of 580nm. Test benches were designed to perform this comparison. Figures 17 and 18 show the results.

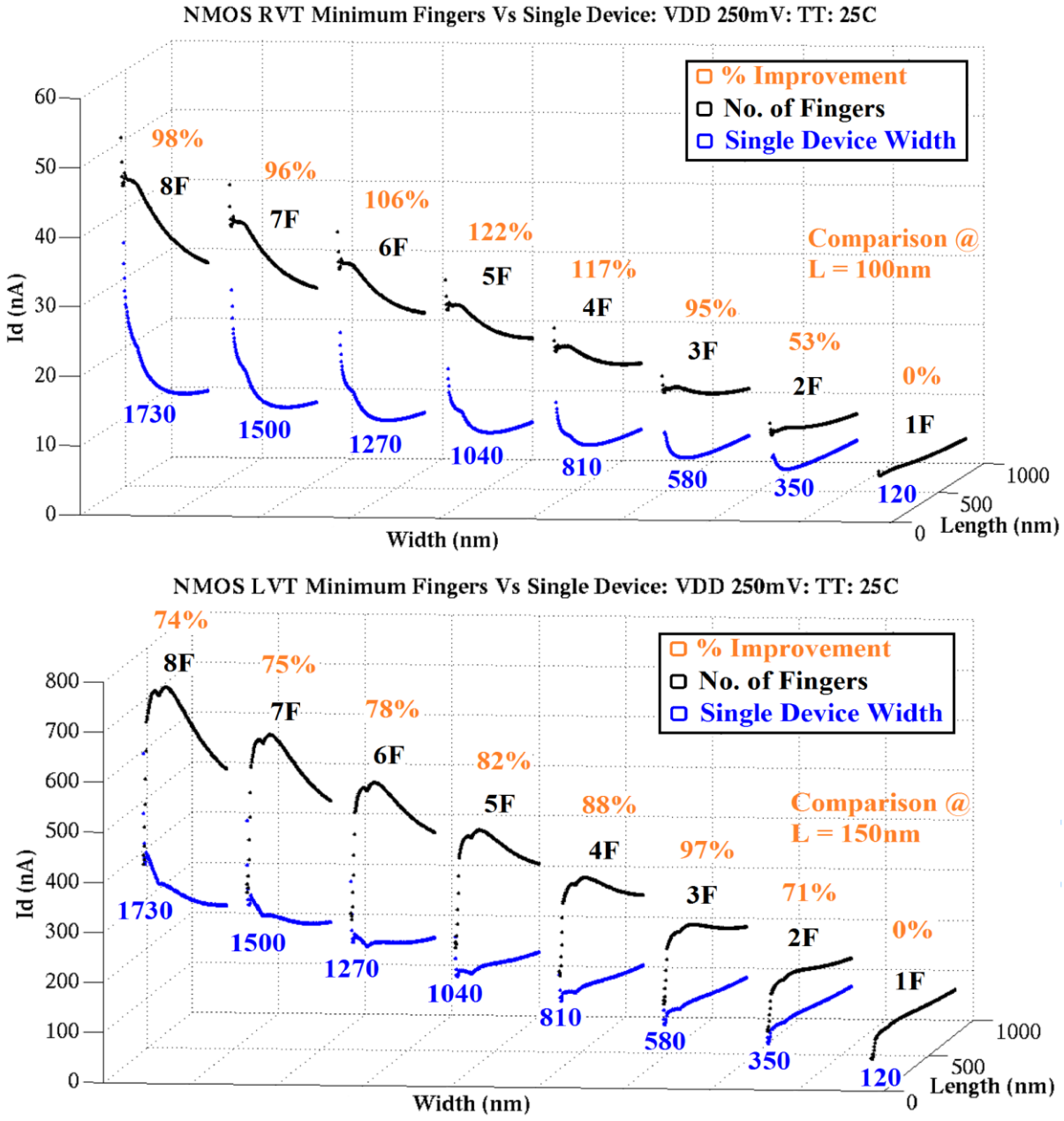


Figure 17: Subthreshold Iso-Area parallelization comparison: NMOS TT

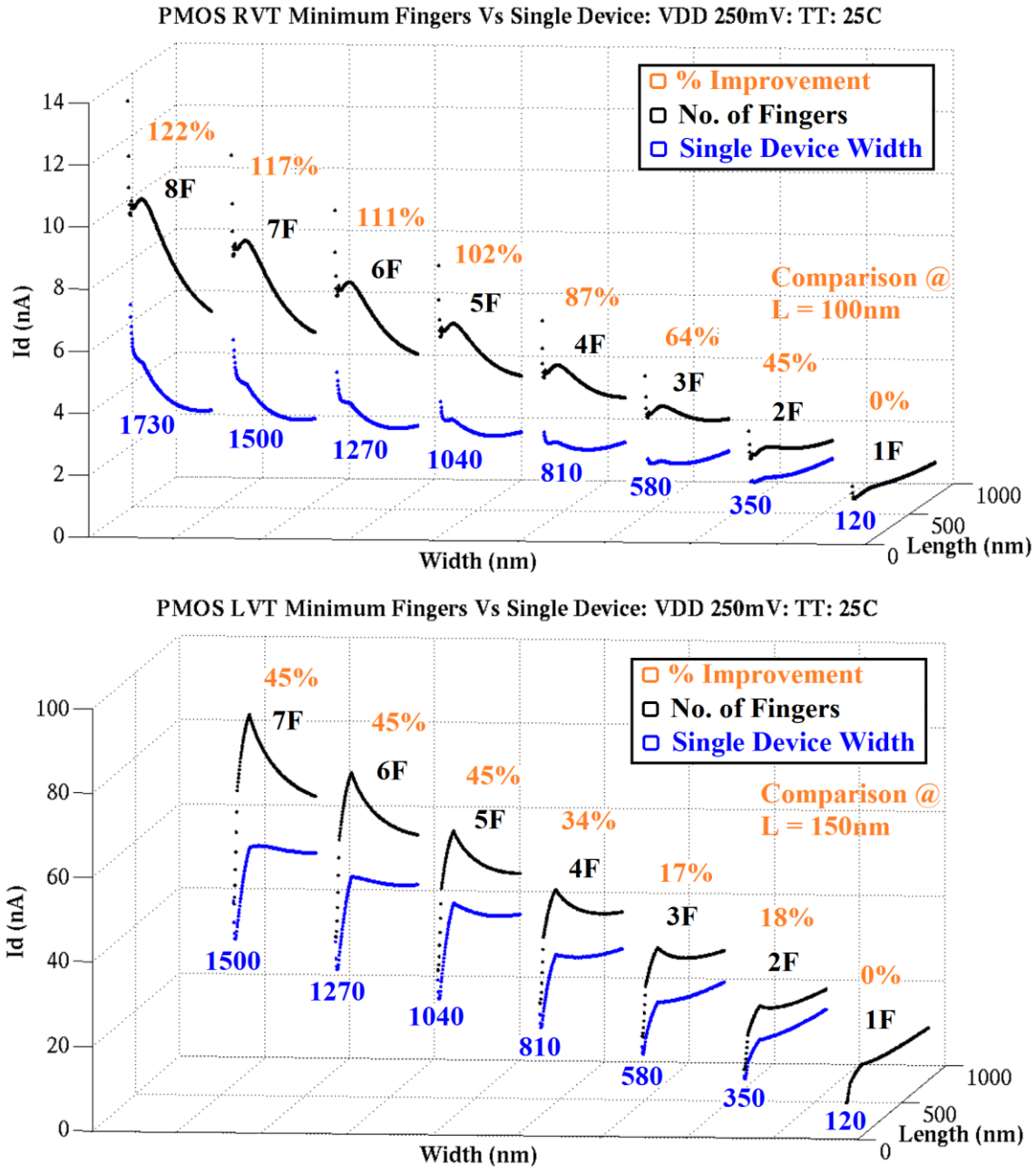


Figure 18: Subthreshold Iso-Area parallelization comparison: PMOS TT

Figure 17 shows the NMOS device comparison at a supply voltage of 250mV, typical process corner and temperature of 25°C. Length sweeps were performed at each width/finger point to ensure the relationship does not change over the device length. An interesting feature immediately obvious from the figure is that the RSCE optimal length in the parallelized (fingered) composite devices does not change. This means the comparison may be performed at the same length for each case. The RVT device has an RSCE optimal length of 100nm. The first composite device (2 Fingers) shows an immediate iso-area current improvement of 53%. The improvement then increases, peaking at the 5 fingered device and waning towards the 8 finger device. This is due to the optimal RSCE length shifting in response to width in the single device baseline beyond 1040nm. Assuming equal P to N sizing, the maximum width afforded in a 12 track library in the selected technology node is 840nm. This means composite devices up 4 minimum width fingers are viable. The maximum ideal current increase (before the addition of parasitics) is therefore 117% for the NMOS RVT device.

The LVT device shows similar features. The RSCE optimal length in this case is 150nm with the improvement trend peaking at the 3 fingered device. The maximum ideal current increase is 97%.

Figure 18 shows the PMOS device comparison at a supply voltage of 250mV, typical process corner and temperature of 25°C. In the RVT PMOS device, the increase in improvement does not wax and wane. This is due to the fact the RSCE optimal length does not vary greatly in response to an increase in width in the baseline single device. This is therefore a good representation of the continuing improvement possible from this technique. The maximum ideal current increase is 87% for a device viable in a 12 track library for this technology node (max 4F/840nm).

Finally the LVT PMOS device also shows constant improvement until the 5 finger device. At this point the improvement peaks. The absolute magnitude of the current improvement is also less for the LVT PMOS. This is due to the superior current sensitivity on the width of this device. The maximum ideal current increase is 34% for a viable cell. The 8 finger variant is omitted only due to corruption of the underlying data.

These simulations highlight a number of important points. The first is that the INWE always improves the drive current capability over the single baseline device in both NMOS and PMOS devices at both threshold voltages (RVT/LVT). This means that even at the lowest width that INWE can be applied (2 fingers = 350nm) there is a gain achievable by it's application.

The second point of significance is the potential INWE gain appears greater in the NMOS device than in the PMOS devices. According to the TSMC 65LP DRM, the nominal oxide thickness of the NMOS devices is 26 Angstrom whilst the PMOS device is 28 Angstrom. This is important from the INWE discussion in Section 2.5.7 where it stipulates a thinning of the gate oxide results in a deeper depletion region. The naturally thinner oxide of the NMOS device would therefore be more susceptible to INWE.
The third point of significance is that the potential INWE gain appears greater in the RVT devices than for the LVT devices. There are many reasons this may be true, some of which were discussed in the previous chapter (differing dopant densities etc.).
From the above points, the device offering the highest INWE gain is the NMOS RVT device and the device offering the lowest INWE gain is the PMOS LVT device.

3.4 Stack Forcing and Leakage Current

As described in Section 2.5.9, one of the techniques explored in subthreshold standard cell design is stack forcing. This involves forcing additional devices into single device paths in the pull up/pull down networks of CMOS cells. To enable the use of this technique, it was imperative to determine its interaction with the INWE. The test bench was therefore modified to include two devices in a series stack. The INWE was then applied and measured against a single device of iso-area analogous to the test in the previous section. The leakage current was also simulated in order to determine the effect of the INWE on device leakage. Figures 19 and 20 show the results.

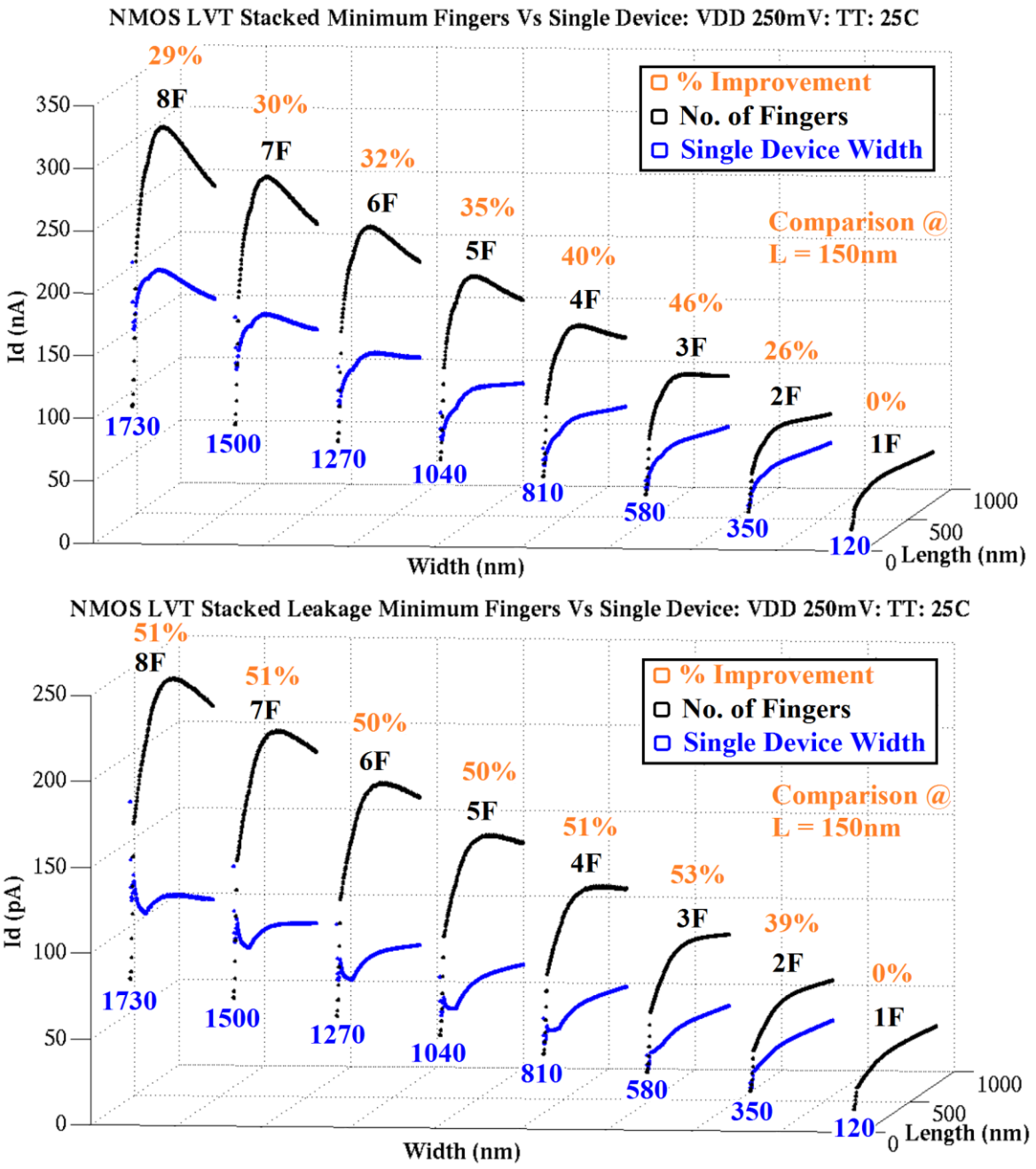


Figure 19: Stacked Iso-Area parallelization comparison: NMOS TT

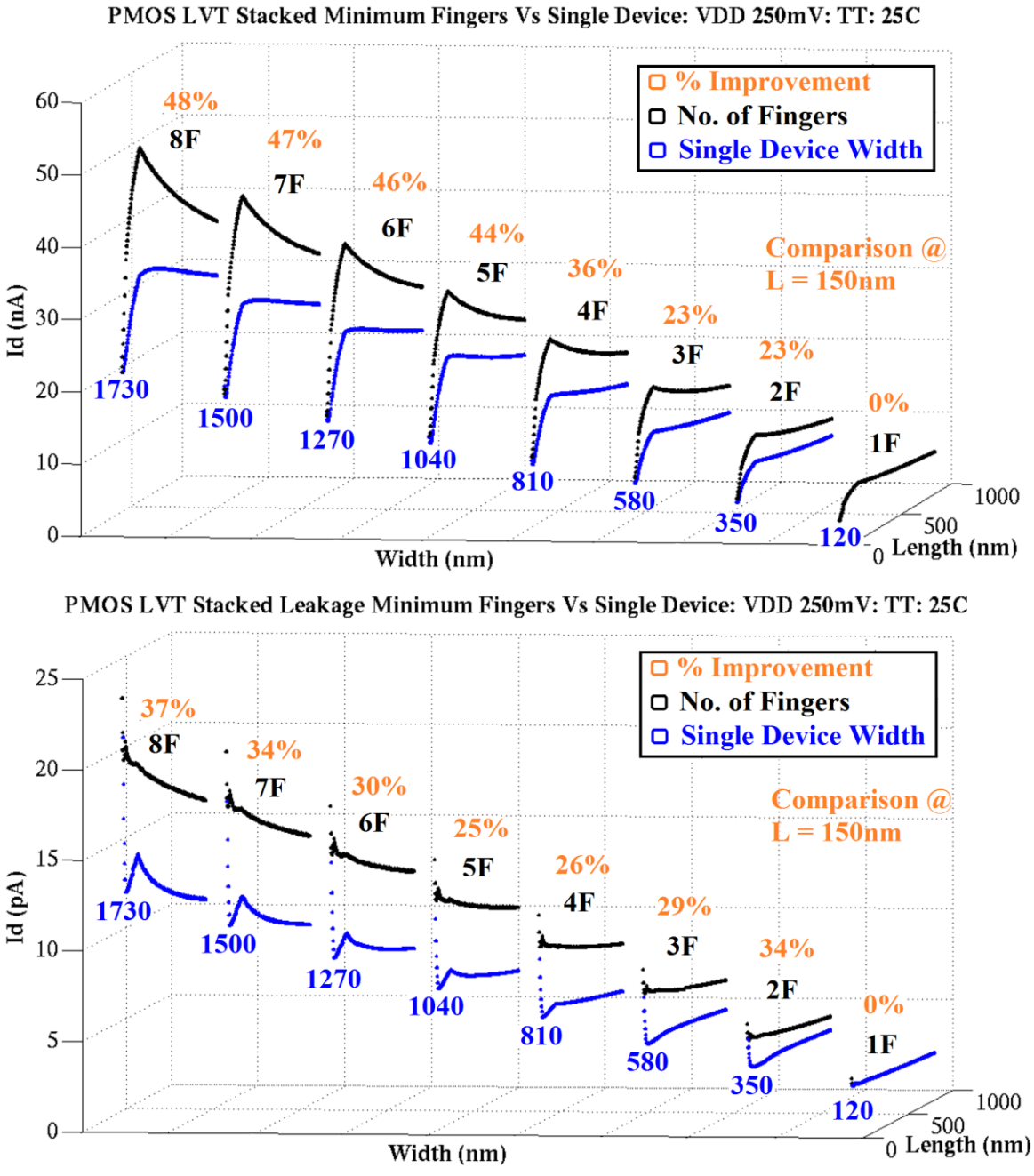


Figure 20: Stacked Iso-Area parallelization comparison: PMOS TT

Figure 19 shows the comparison for the NMOS LVT device at a supply voltage of 250mV, typical process corner and temperature of 25°C. The same trends exhibited by the single device comparisons remain the same for the stacked devices. The NMOS LVT stacked devices show an immediate improvement in the drive current as soon as INWE is introduced at 2 fingers / 350nm (26%). The absolute magnitude of this improvement is diminished by the effect of stack forcing, but the magnitude trend is preserved, peaking at the 3 finger device and waning afterwards in response to the shift in the RSCE optimum length. The RSCE response is also preserved, indicating that the optimal length is not affected by the stack forcing. The absolute drive current values are diminished by a factor of between 2x and 3x by introducing an identical device in series.

The NMOS LVT leakage response shows an increase in leakage on the introduction of the INWE effect at 2 fingers / 350nm (39%). The leakage increase peaks at the 3 finger device and reduces afterwards, following the same trend as the drive current. The comparative increase however is larger (39% vs 26%). This indicates that the INWE increases the leakage current proportionally greater than it increases the drive current for the NMOS LVT device, affecting the Ion/Ioff ratio as discussed in the previous subsection and possibly having a negative impact on the energy efficiency.

Figure 20 shows the comparison for the PMOS LVT device at a supply voltage of 250mV, typical process corner and temperature of 25°C. The PMOS stack also exhibits the same drive current trend as the single PMOS device, continually improving as the number of minimum parallelized fingers is increased. Introducing INWE at 350nm / 2 fingers immediately improves the drive current by 23%. In this instance, the comparative increase is larger than in the single device (23% vs 18%) due to greater degradation in the baseline single device via stack forcing. Moreover, stack forcing the PMOS device reduces the absolute drive currents by a factor of between 1.8x and 2x, lower than that of the NMOS device. This would result in NMOS/PMOS complementary device strengths closer than single devices.

The PMOS leakage response also shows an immediate increase in leakage current as the INWE is introduced at 350nm / 2 fingers (34%). As is the case in the NMOS device, the comparative increase in leakage after introducing the INWE is greater than the drive current (34% vs 23%). Therefore, like the NMOS device, introducing the INWE affects the Ion/Ioff ratio, possibly having a negative effect on the energy efficiency.

From the information presented in Figures 19 and 20 it is clear that stack forcing does not introduce massive deviation in the application of the INWE. The INWE still always

increases the current capability, iso-area, over a single device, even if those devices are stacked in series.

However, stack forcing does affect the NMOS/PMOS balance in terms of relative improvement over a single device. Whilst the improvement is still greater for the NMOS devices when stacked, the relative improvement is closer than for the single devices, improving the parity of the absolute magnitudes of the NMOS and PMOS devices.

Interestingly the underlying device characteristics described in the geometric sweeps are still viable when more complex topologies are introduced to the devices. The greater proportionality of leakage increase to drive current increase remains firm, eliminating any possibility of super-cut off leakage reduction (as outlined in Section 2.5.9.).

3.5 Device Capacitance

3.5.1 Gate Capacitance

One of the key metrics for ultra low power applications is performance per energy consumption. Maintaining energy consumption at a fixed rate and increasing performance is an equally worthwhile endeavor to maintaining performance and reducing energy consumption. The inherent relationship between them usually means an improvement in one may be traded off for the other via voltage scaling or a variety other techniques.

Performance is inherently tied to average cell propagation delay. Propagation delay is essentially the time required for a cell to source/sink sufficient charge to impart the desired state on the following stage's inputs. Therefore, it is critical not only to determine the effect of any geometric sizing strategies on current characteristics, but also on gate capacitance.

This is somewhat more complicated than determining the drive and leakage currents for several reasons. The switching gate capacitance (effective gate capacitance) is not the same as the static gate capacitance due to the complementary switching of the gate and drain. This means that the capacitance is dependent on switching speed. It is important to ensure the actual capacitance measured is accurate as it directly affects the subthreshold factor as discussed in Section 2.5.4.2 and therefore minimum operating voltage (further discussed in Section 3.7). The logical solution to determine a representative gate capacitance is therefore to observe the method of logical effort [81] and undertake this measurement using an FO4 test bench. The test bench used is taken from the de facto CMOS textbook, Harris and West [82]. Figure 21 shows a schematic for the test bench.

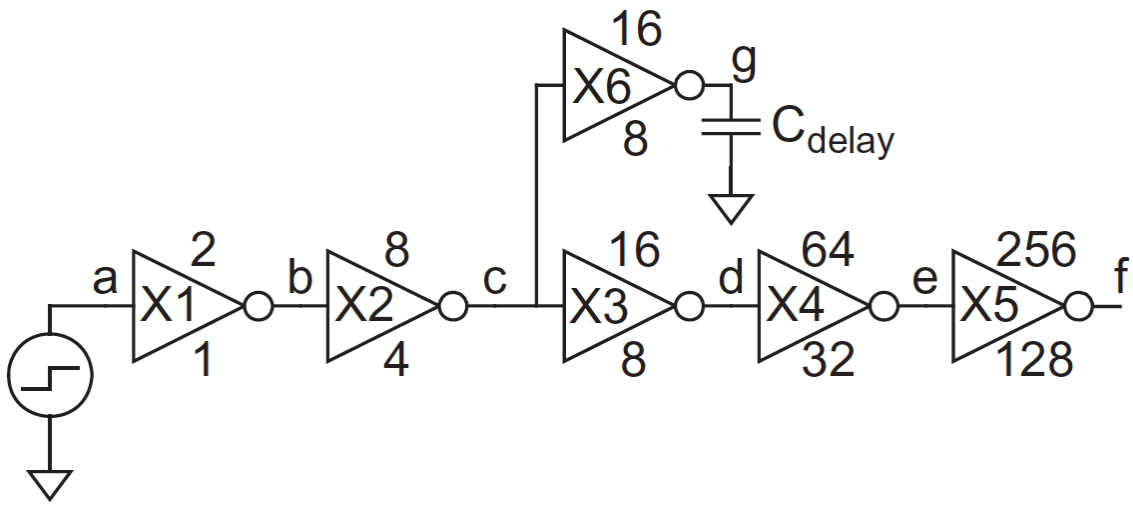


Figure 21: Gate capacitance SPICE test bench [82]

Inverters X1 and X2 are input forming stages to eliminate the step function dependence on the input of the device under test (DUT). X3 is the device under test. X4 and X5 are load stages. The test bench measures the rise and fall delays between nodes c and d (DUT) and c and g. Hspice's optimizer is used to iteratively vary C_{delay} until the rise and fall delays match between the two paths. At this point, C_{delay} matches the gate capacitance of X4 and the gate capacitance has been found. This can then be used to determine a 'Capacitance-per-micron-width' metric for an instantaneous value of width independent capacitance. The whole test bench is then geometrically swept like the current test benches in the previous subsections. Each geometric sweep is performed for both RVT and LVT devices across the three process corners SS/TT/FF.

Sweeping length from 60n to 1µm and width from 120nm to 1µm on the 5nm manufacturing grid creates a pool of 33453 test cases. Tight convergence constraints for the optimizer at each point and quirks in the range transitioning in the model result in sporadic areas of failed test cases. Post processing using MATLAB was therefore required. Failed points were brushed from the data and Delaunay triangulation used to create an underlying grid of the remaining successful aperiodic data. Trisurf 3D plots were then created to form the surface plots of the capacitance response. These were specifically chosen over interpolated shaded surface plots as the included and approximated data points from the processing are evident from the triangulated mesh lines. All datasets consist of 85% or greater successful test cases.

Figures 22 and 23 show the results.

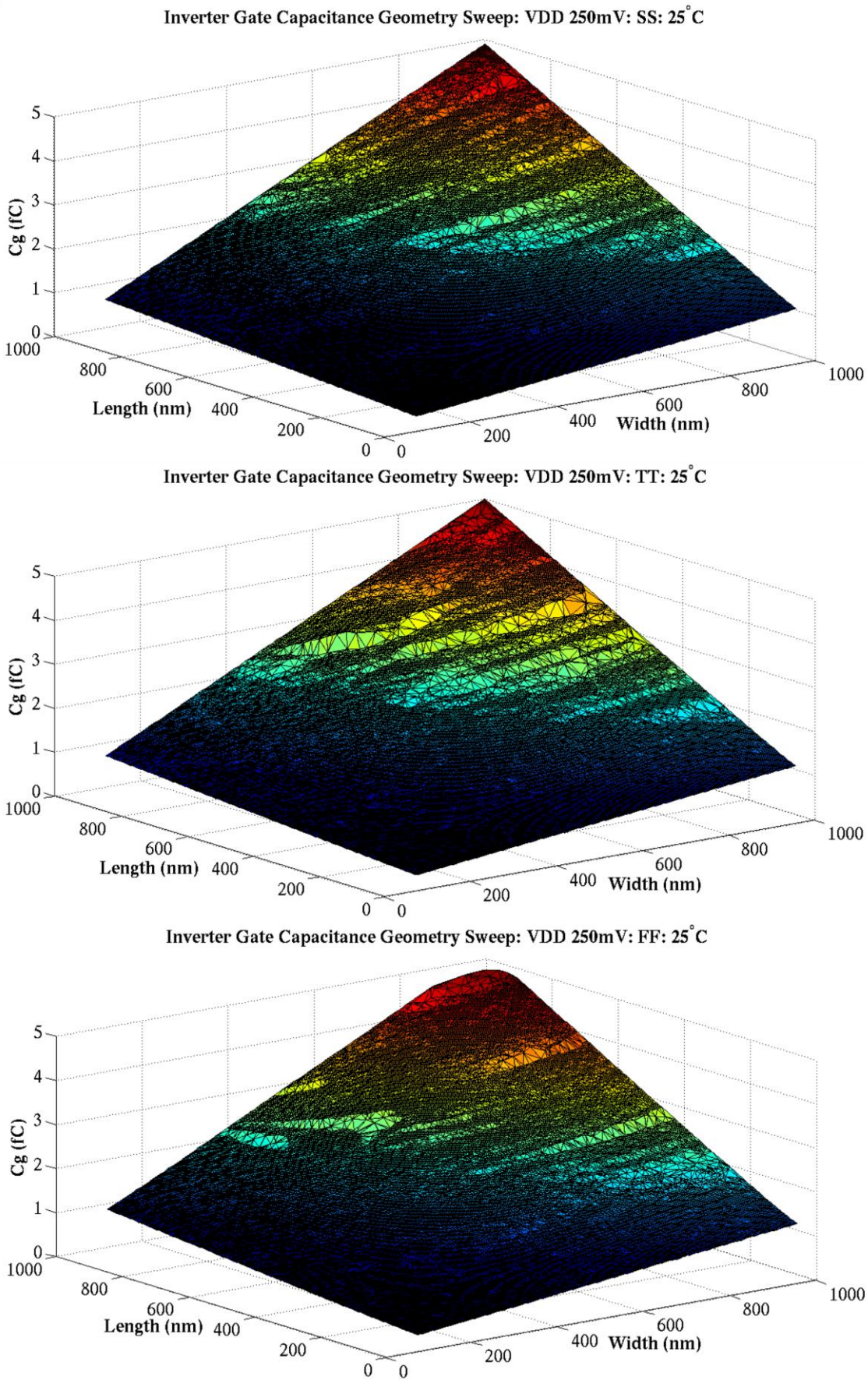


Figure 22: RVT gate capacitance sweeps

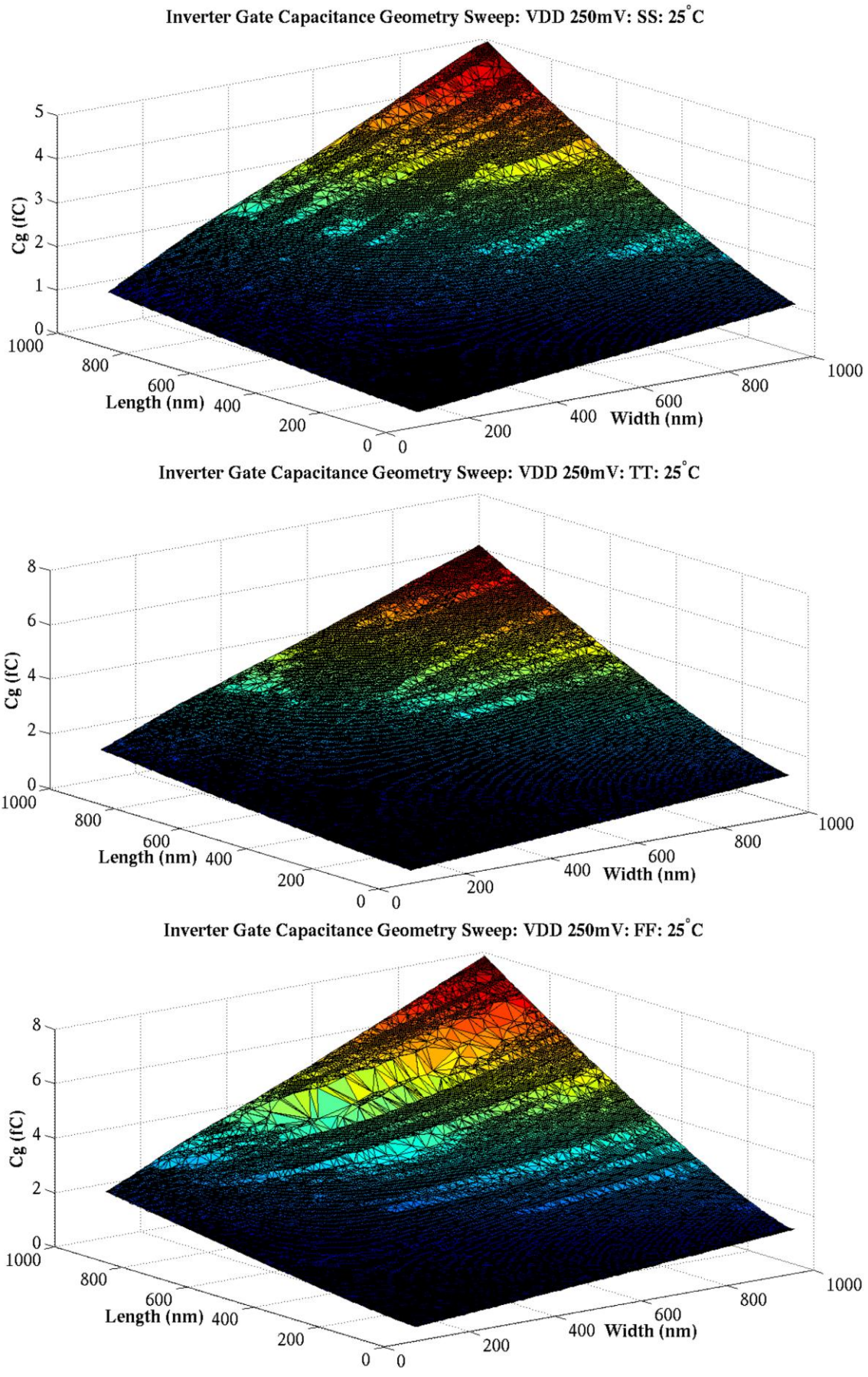


Figure 23: LVT gate capacitance sweeps

Figure 22 shows the RVT device gate capacitance sweeps conducted at a temperature of 25°C and supply voltage of 250mV. All process corners show almost ideal linear capacitance responses to length and width variation, leading to the expected relationship to the quadrature. Although the sporadic failed test case areas differ between the corners, the underlying geometric gate capacitance relationships are almost identical, indicating global process variation plays little part in determining gate capacitance. The discussions from section 2.5.4.1 describe the gate capacitance as the sum of the oxide capacitance and depletion capacitance. Whilst any difference in oxide capacitance is not accounted for in the corner analysis, changes in the channel dopant density and therefore depletion capacitance are. This would indicate that the change in depletion depth over the entire global variation space is marginal and therefore, so is the impact in depletion capacitance and gate capacitance. The deviation in the FF corner at maximum length/width is merely due to a test case failure area occurring at that location and insufficient successful test cases available to construct the correct response. There is no reason to suggest this roll-off actually occurs.

Whilst the RVT gate capacitance responses are almost ideal, there is very slight deviation at minimum length and minimum width. This shall be addressed in the following LVT response discussion.

Figure 23 shows the LVT device gate capacitance sweeps conducted at a temperature of 25°C and supply voltage of 250mV. All process corners again follow the almost ideal linear geometric relationship with the quadrature, however the global process variation does have an influence on the absolute magnitudes of the gate capacitance, with the value increasing as the global variation tends towards the fast corner. This suggested the LVT devices are more sensitive to dopant density variation than the RVT devices.
The deviation in the linear geometric relationship is also greater for the LVT devices, with slight increases in gate capacitances as device length and width tends towards the minimum. Discussion from sections 2.5.1 and 2.5.5 highlight the expected relationships. The RSCE discussion suggests that a disproportionately higher gate capacitance would be expressed at minimum length due to the increase in dopant density in the center of the device channel from the overlapping HALO dopants, resulting in a smaller depletion width and therefore a higher depletion capacitance. As the length is increased this effect diminishes until the background channel dopant density is reached and the linear

relationship resumes. This phenomenon is observed in the simulated results. The INWE discussion suggests that a disproportionately lower gate capacitance would be expressed at minimum width due to the additional fringing field increasing the channel depletion depth and therefore lowering the gate capacitance. This phenomenon is not manifest in the results. In order to determine the validity of the claims expressed in the field, the actual width independent gate 'capacitance-per-micron-width' was therefore plotted to highlight any observable alterations in width independent capacitance. Figures 24 and 25 show the geometric sweeps.

Figure 24 shows the RVT device width independent metric geometrically swept at a temperature of 25°C and a supply voltage of 250mV. The lengths in these sweeps show the linear dependence expected from the methodology and therefore hold little value. The deviation in length at maximum length/width in the fast corner is merely the deviation expressed in the previously discussed figure where the test case failure area occurs at this location. The deviation in gate capacitance at minimum width is clearer when expressed in this metric. The figure shows a measurable increase in width independent gate capacitance as the width tends towards minimum. This increase is affected by global variation, increasing as the process tends towards the fast corner. This trend is consistent with the increase in the influence of the INWE as expressed in the current geometric sweeps described earlier. The polarity however is opposite, with a decrease in gate capacitance expected instead of the increase observed. It may be that the INWE does indeed lower gate capacitance as suggested in the physics discussed in Section 2.5.5 but that a separate physical effect has a greater influence. This could possibly be alterations in junction capacitance or overlap capacitances also affected by device geometry.
Figure 25 shows the LVT device width independent metric geometrically swept at a temperature of 25°C and a supply voltage of 250mV. Similar trends are observed with width independent gate capacitance increasing towards the fast corner, however the increase in magnitude over global variation is greater than in the RVT device. This again suggests that the LVT device is more sensitive to dopant density variation and to the INWE, as was highlighted by the current geometric sweeps.

It is important to note from these figures that the width independent metric exaggerates the capacitance increases and trends in comparison to the actual measurements shown in Figures 22 and 23 by making the assumption that the entire gate area has a uniform

capacitance. Whilst this is a good tool to highlight the underlying physics, it must be noted that the peak deviation in the results occurs at minimum width where their affect on the actual gate capacitance is minimized due to the 'per-micron' nature of the metric. The increase therefore has little manifestation in the actual gate capacitance at minimum width. This is important, as INWE optimal sizing at minimum width would offer a greater improvement in current than increase in capacitance of the following stage, decreasing propagation delay whilst having little effect on dynamic energy consumption.

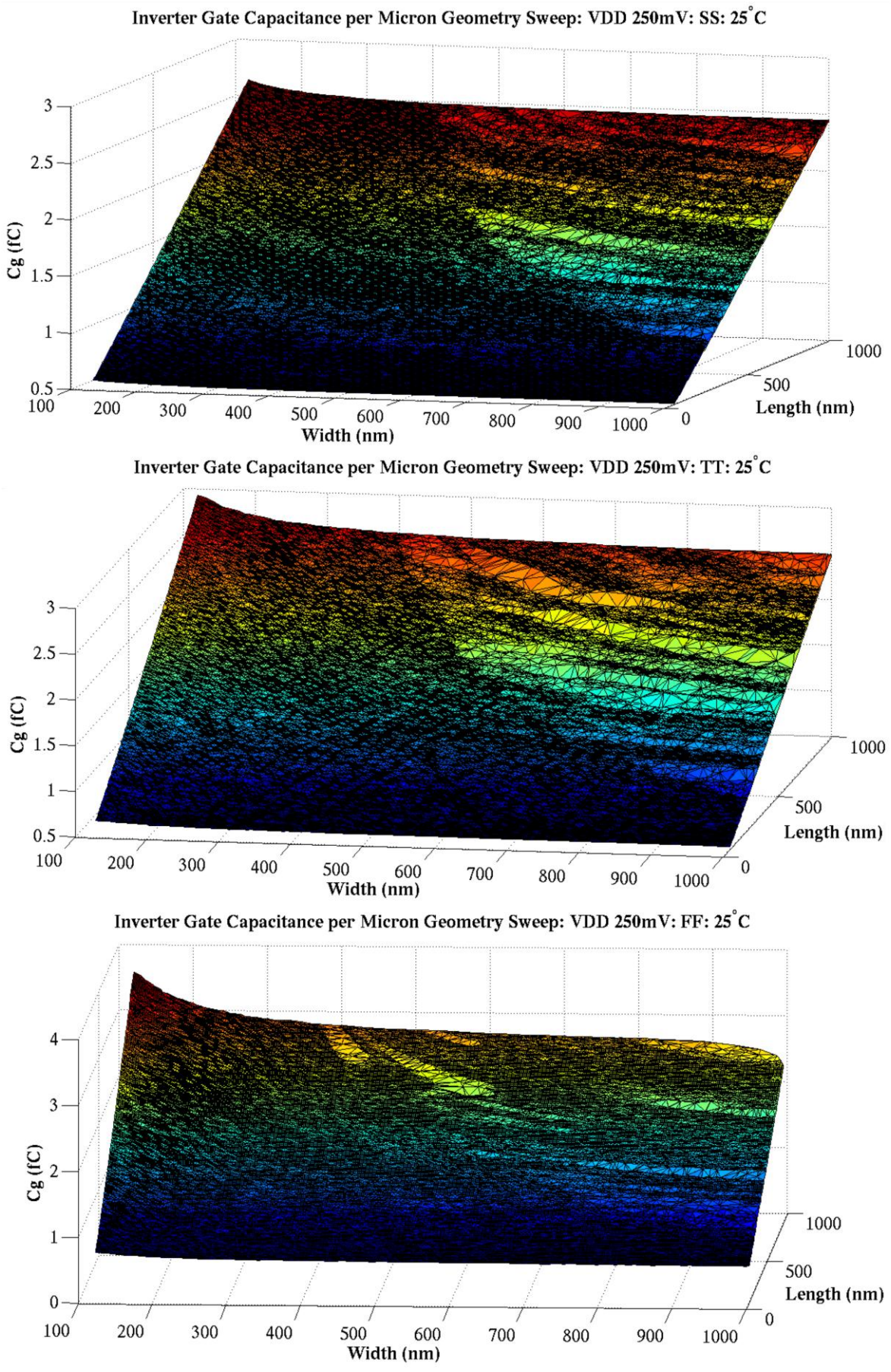


Figure 24: RVT gate capacitance per micron sweeps

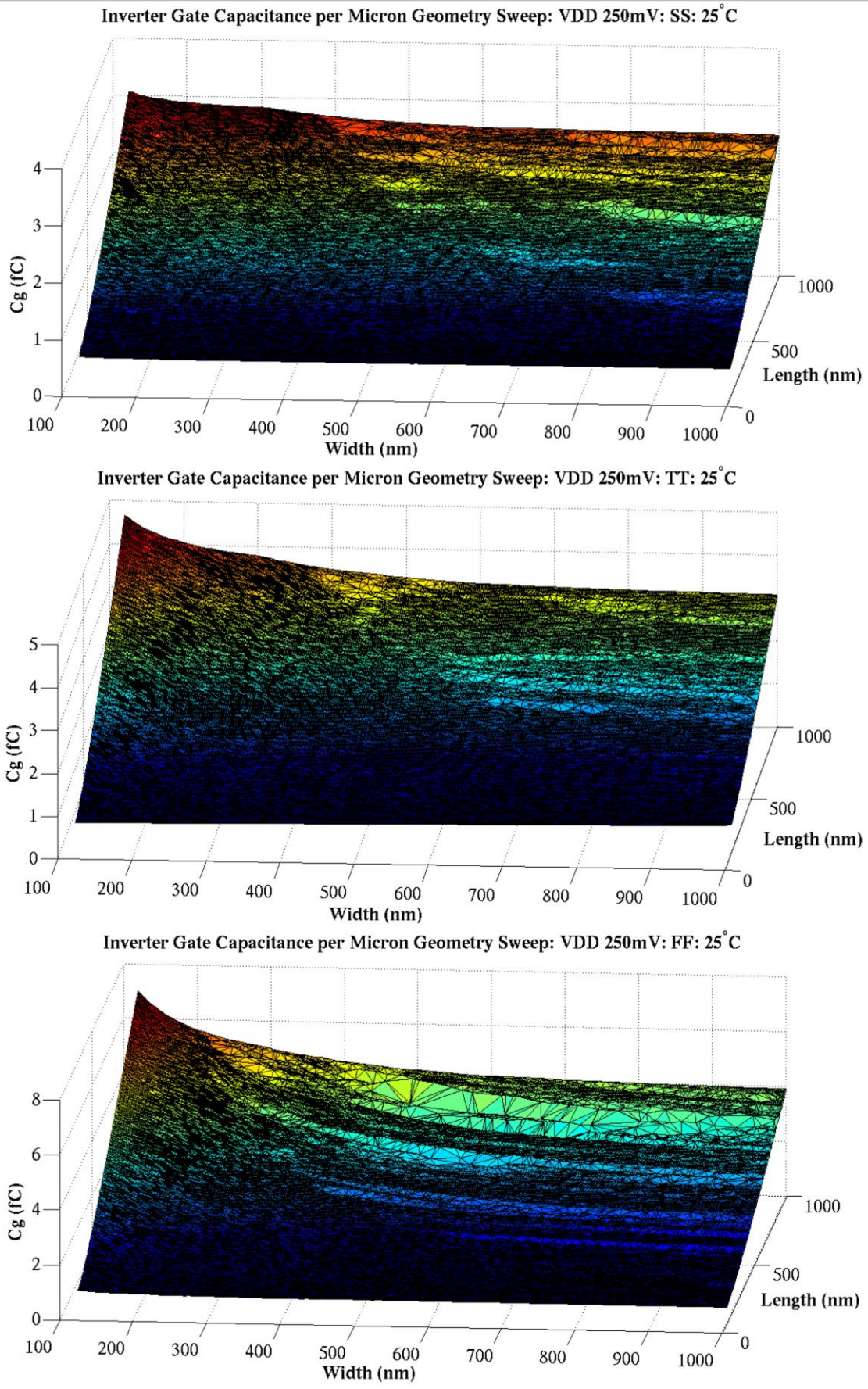


Figure 25: LVT gate capacitance per micron sweeps

Dynamic energy consumption itself is the charging and discharging of all switching capacitances required to achieve the correct state during operation. Whilst gate capacitance is often considered the largest, junction capacitances at the drain are also charged/discharged in response to a state change. A similar test bench was therefore created to measure this. Once again this was based upon an FO4 spice deck presented in Harris and Weste [82]. Figure 26 shows a schematic for the test bench. Once again the HSpice optimizer is used to match the rise and fall delays between nodes c and d (DUT) and c and e (dependent variable C_{delay}). However in this test bench the DUT is loaded with the drain junction capacitance of a single device. The parameters AD and PD are BSIM4.5 transistor model parameters for the drain area and drain perimeter and are specifically calculated in the deck from the length and width at that particular test point in the geometric sweep. This is to zero out any overlap between the device gate and drain, removing any implicit overlap capacitance that would influence the results of the test bench. As this test bench simulates a single device's junction capacitance, each geometric sweep was performed on NMOS/PMOS devices, at both LVT/RVT thresholds and across all three global process corners, SS/TT/FF. The same post simulation processing as the gate capacitance was performed. Figures 27 and 28 show the junction capacitance and width independent capacitance for the RVT NMOS device.

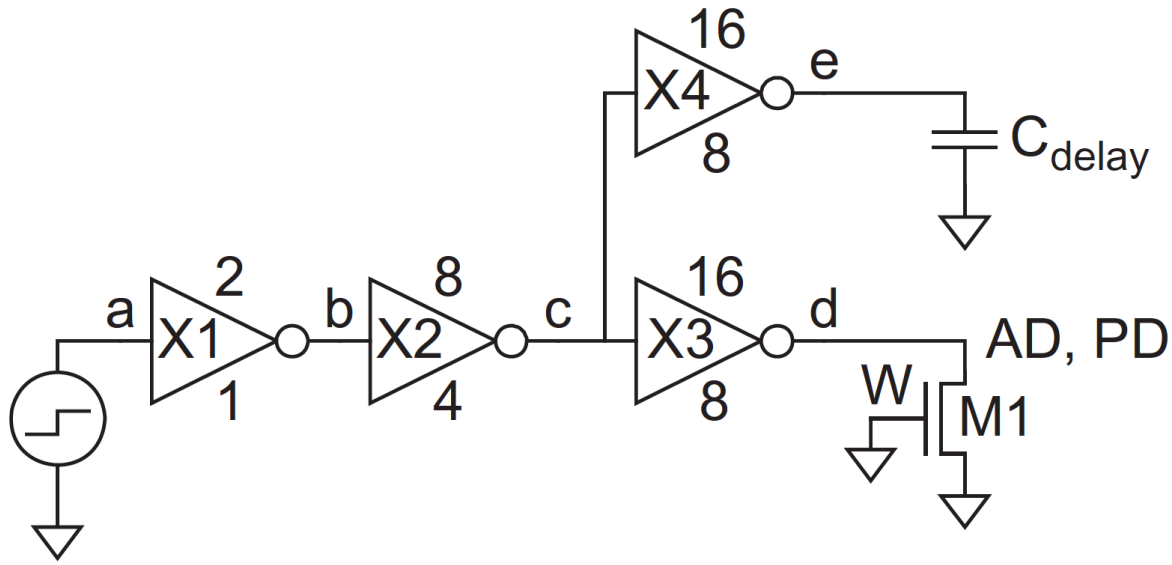


Figure 26: Device junction capacitance SPICE test bench

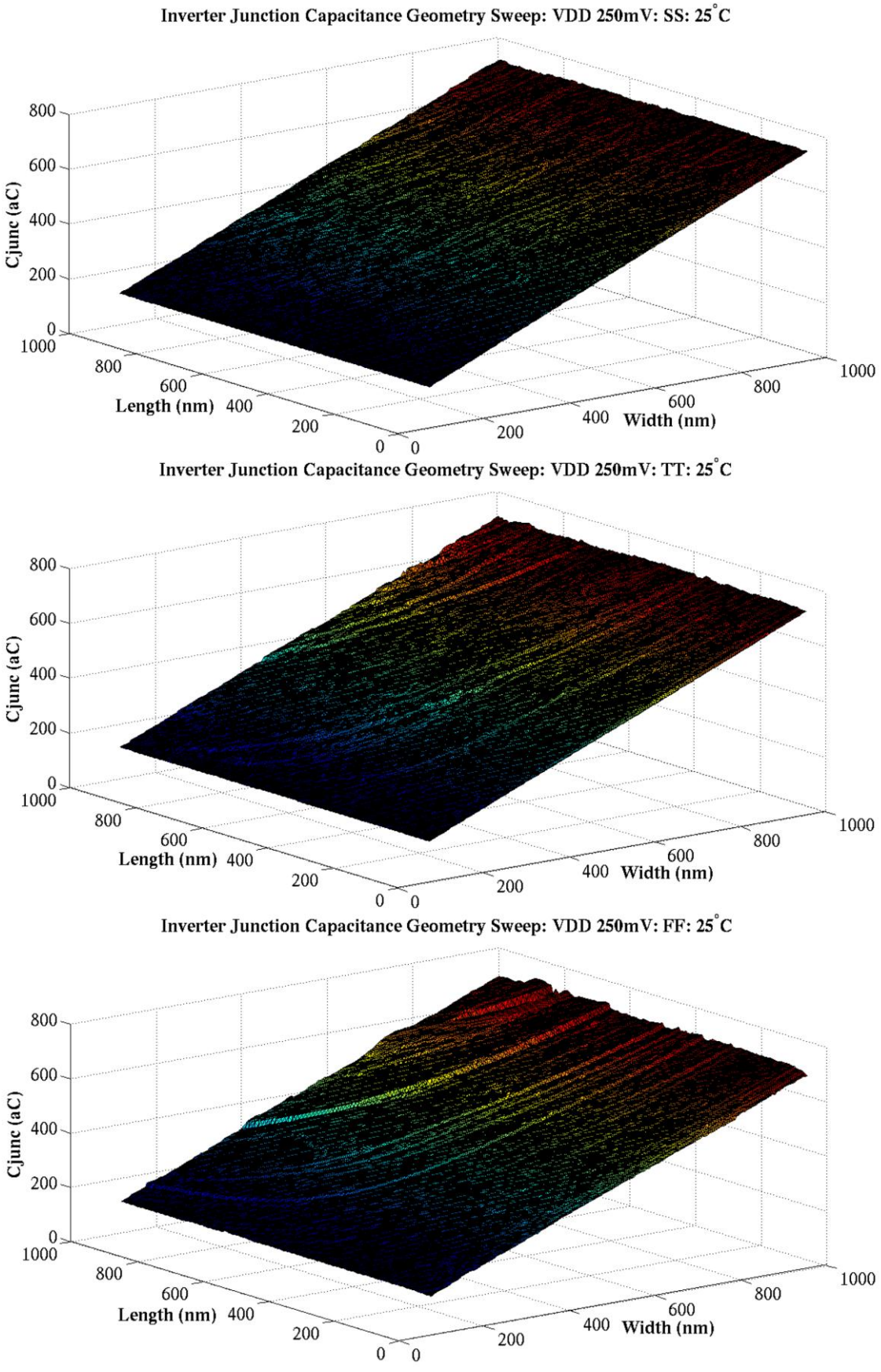


Figure 27: Junction capacitance sweeps: NMOS RVT

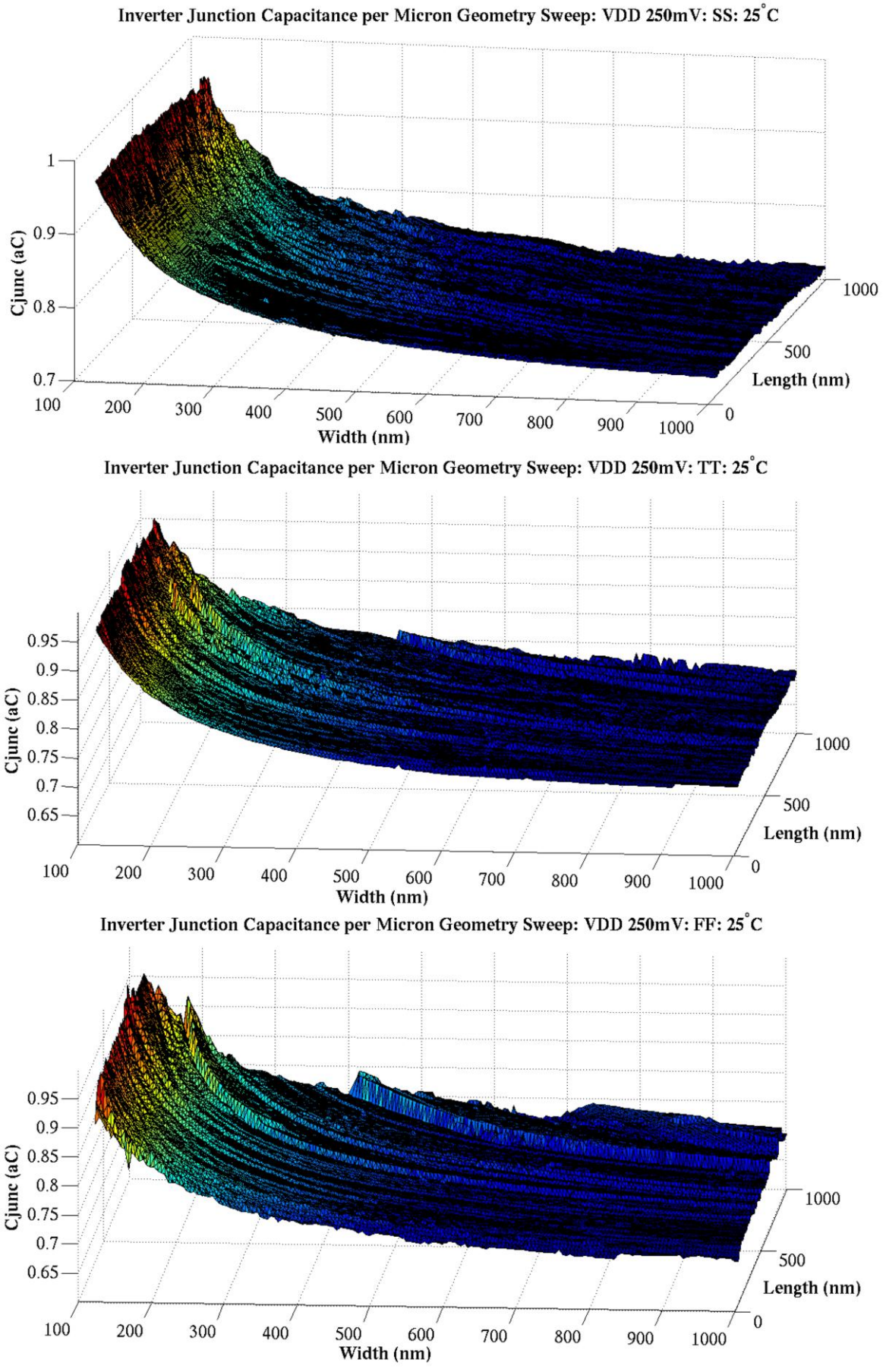


Figure 28: Junction capacitance per micron sweeps: NMOS RVT

Figure 27 shows the junction capacitance geometry sweeps for the RVT NMOS device at a temperature of 25°C and a supply voltage of 250mV. The topology of MOS devices suggests that an alteration in device length does not directly affect the drain/source junction and therefore this capacitance is independent of device length. This relationship is proven in the geometric sweep. Given that the channel dopant density and dopant location varies globally, this would theoretically change the junction depletion depths and capacitance. However, the junction capacitance response shows little alteration over global variation. There may be several explanations for this. The first is that the magnitude in dopant density alteration over global variation translates into a change in depletion depth so small is has little effect on the junction capacitance. The second is that the global variation primarily affects channel dopant density, however HALO implants or drain engineering may pose such a significant impact that the effect of global variation on the channel is lost. The third is that the compact model simply doesn't contain the relevant accuracy to model this effect. With only the information in the model provided in the PDK it is not clear which of these cases are true.

The junction capacitance relation to width is linear, again as is suggested by the underlying MOS topology. The width independent metrics were also plotted to determine any deviation in the linearity.

Figure 28 shows the width independent junction capacitance geometry sweeps for the RVT NMOS device at a temperature of 25°C and a supply voltage of 250mV. The length independent nature of the junction capacitance is maintained under the closer scrutiny of the width independent metric across all global process corners. The width shows a marked increase in width independent junction capacitance as the width tends towards the minimum. There are two possible causes of this. The first is that the 'lengths' of the high dopant density drain region form a higher proportion of the perimeter and therefore contribute a greater influence on the junction capacitance at minimum width. The second is that the abrupt nature of the model validity at minimum width (120nm) simply leads to an over estimation of the junction capacitance. The overall trends are firm across all global process corners.

The absolute magnitudes of the deviation must be considered. The total deviation represents less than three orders of magnitude in the actual junction capacitance and is therefore within the bounds of statistical insignificance.

The results for the LVT NMOS device are shown in Figures 29 and 30.

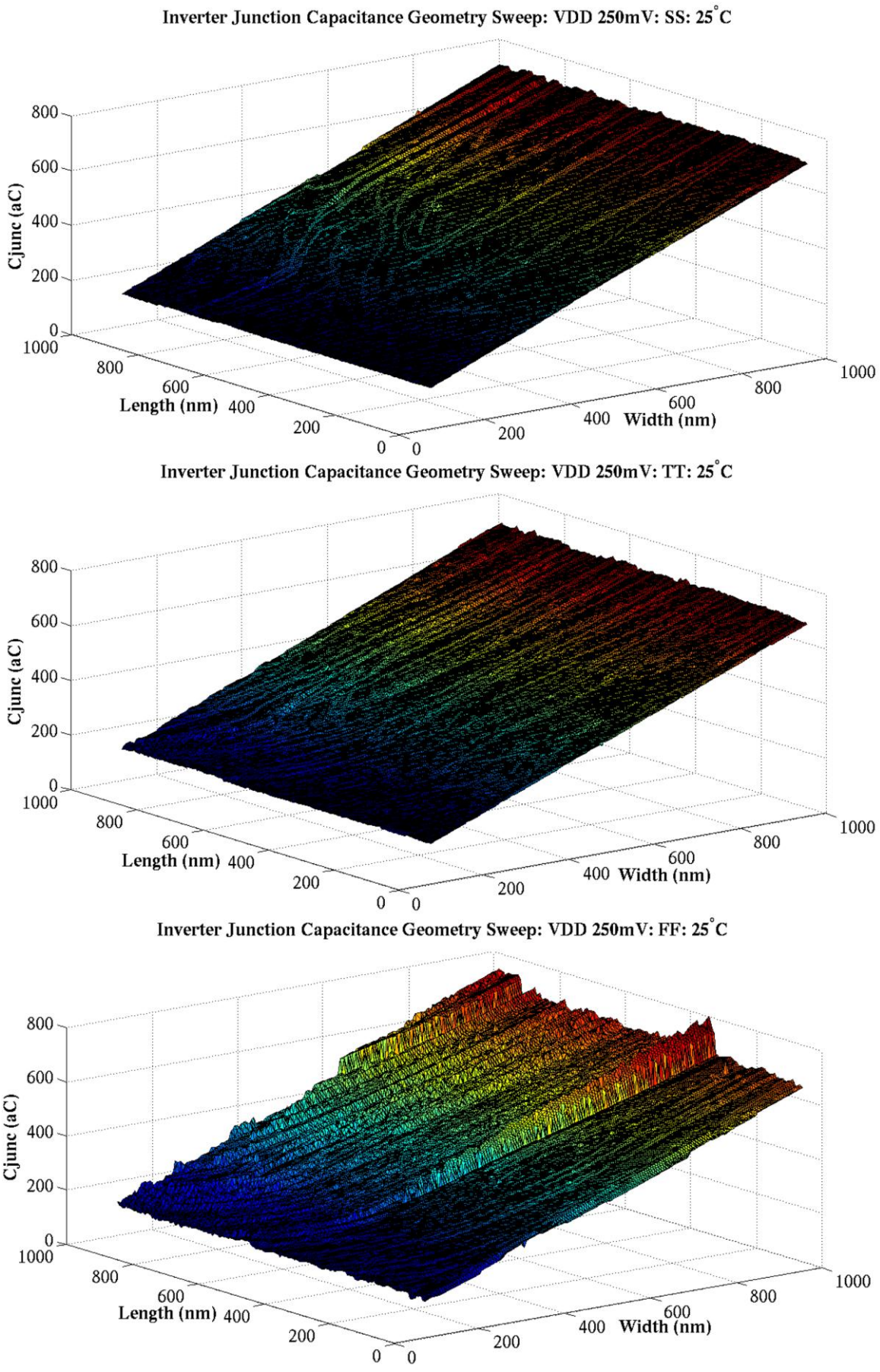


Figure 29: Junction capacitance sweeps: NMOS LVT

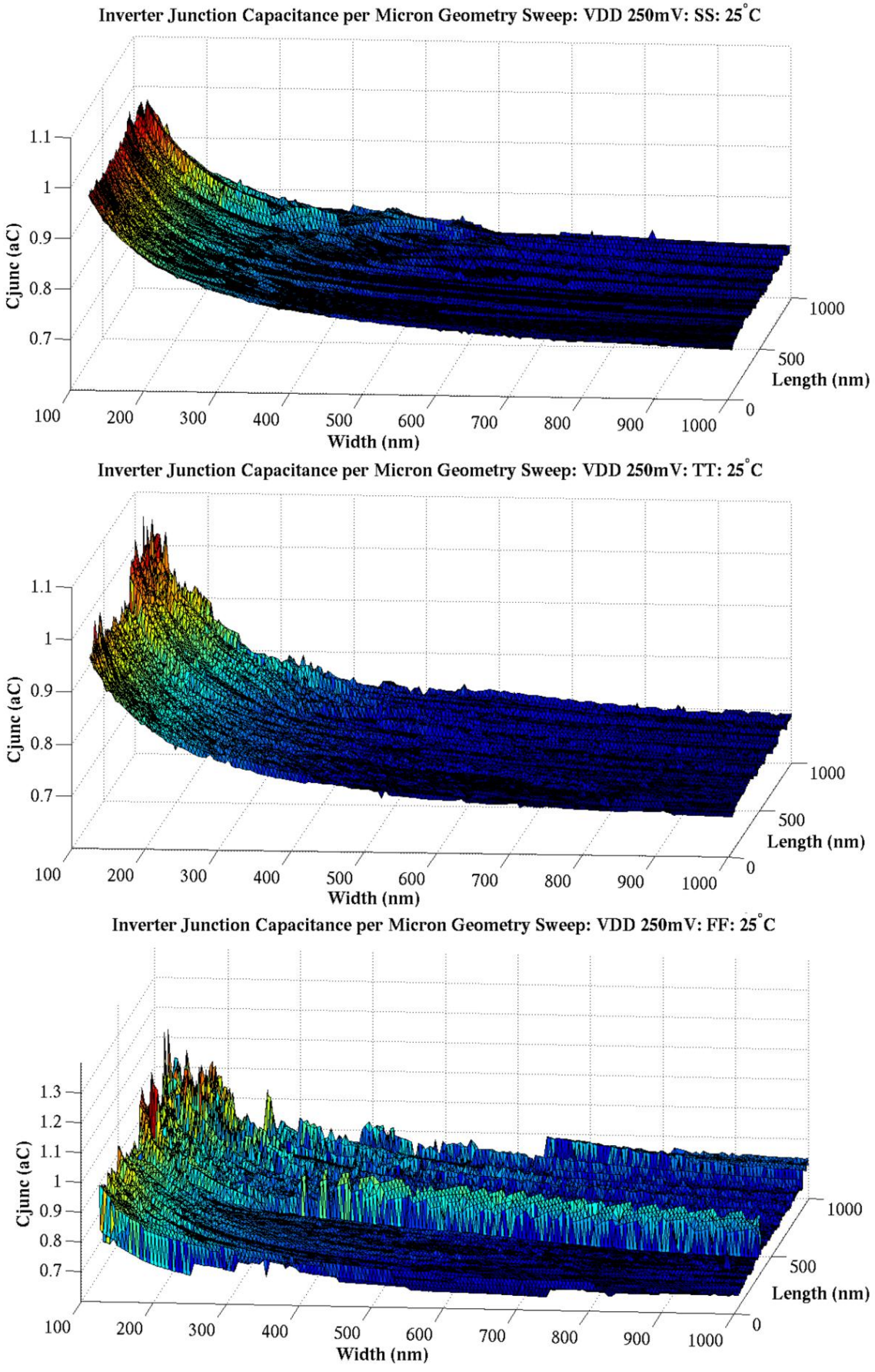


Figure 30: Junction capacitance per micron sweeps: NMOS LVT

Figure 29 shows the junction capacitance geometry sweeps for the LVT NMOS device at a temperature of 25°C and a supply voltage of 250mV. The same trends are observed as in the RVT device. The responses remain firmly independent of global variation and the absolute magnitudes of the capacitance response are the same as the RVT device. The three hypotheses outlined in the RVT device discussion therefore remain possible. There is a large amount of noise in the FF response, suggesting the model is more susceptible to inaccuracy as the operating conditions deviate further away from the nominal.
Figure 30 shows the width independent junction capacitance geometry sweeps for the LVT NMOS device at a temperature of 25°C and a supply voltage of 250mV. The same overall trends are observed as in the RVT device although the absolute magnitudes are marginally higher. The fast corner once again exhibits significant noise. The two hypotheses outlined in the RVT device discussion are still possible.
Figures 31 and 32 show the junction capacitance geometry sweeps for the RVT/LVT PMOS devices at a temperature of 25°C and a supply voltage of 250mV. The trends established in the NMOS devices are observed. The absolute magnitudes are equivalent, indicating little change due to the reverse polarities of the semiconductor types and majority carrier type. Figures 33 and 34 show the width independent junction capacitance geometry sweeps for the RVT/LVT PMOS devices at a temperature of 25°C and a supply voltage of 250mV. The same trends and magnitudes established in the NMOS devices are observed.

From the results of the simulations presented, it is clear that there is no indication of any of the subthreshold physical effects discussed in Chapter 2 affecting the junction capacitances. Moreover the results show a difference in significance of around an order of magnitude between the gate capacitance and junction capacitance for the chosen technology node. This contradicts evidence commonly cited in subthreshold literature that an increase in junction capacitance as a result of lower bias due to voltage scaling can offset the gains made in gate capacitance [7]. Whilst a notable increase in both gate capacitance and junction capacitance is observed in the width independent metric, the absolute magnitude is not influential in the overall capacitance response. Therefore the possible increase in capacitance as a result in minimum width sizing should not be a hindrance in INWE sizing.

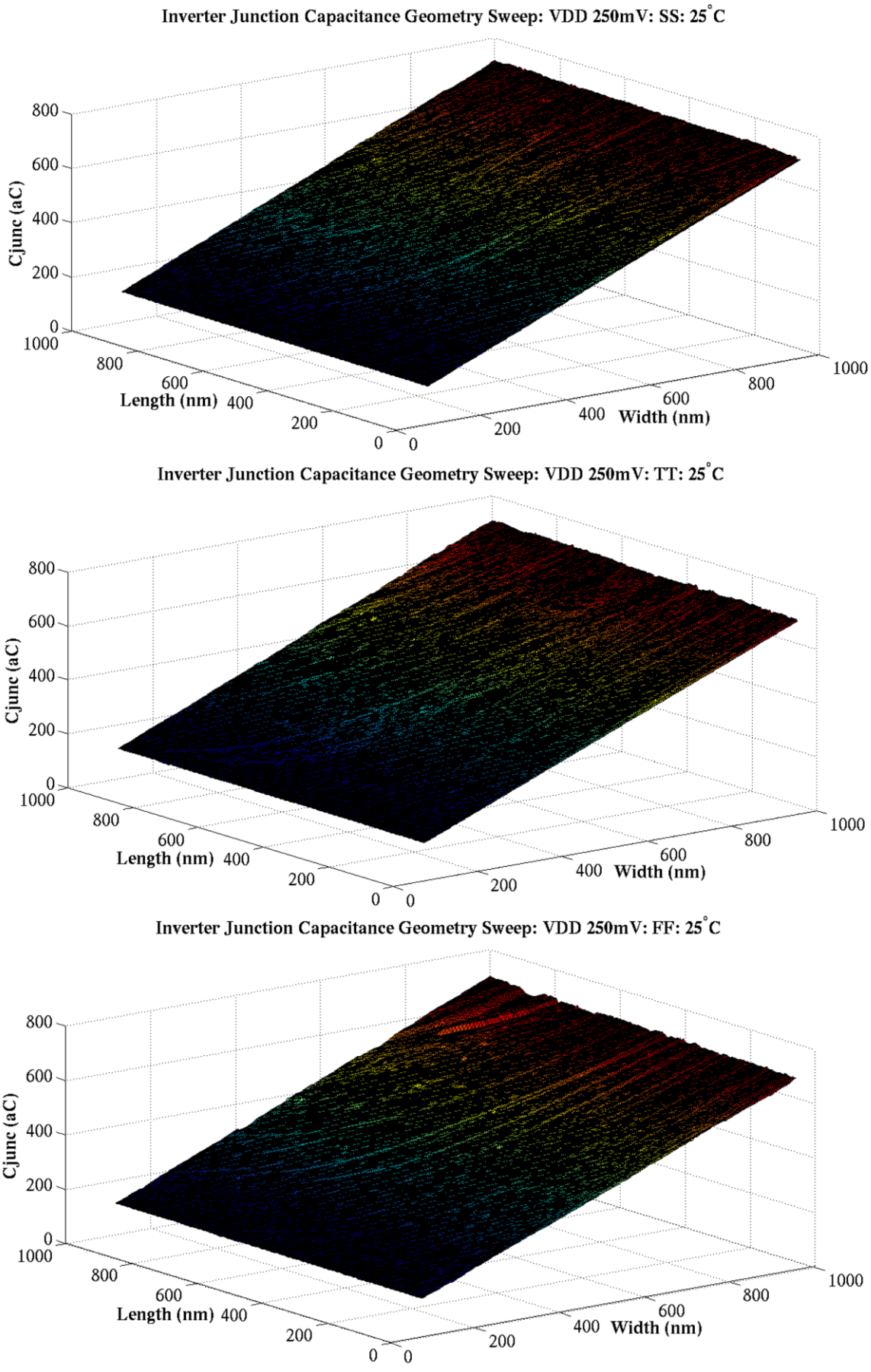


Figure 31: Junction capacitance sweeps: PMOS RVT

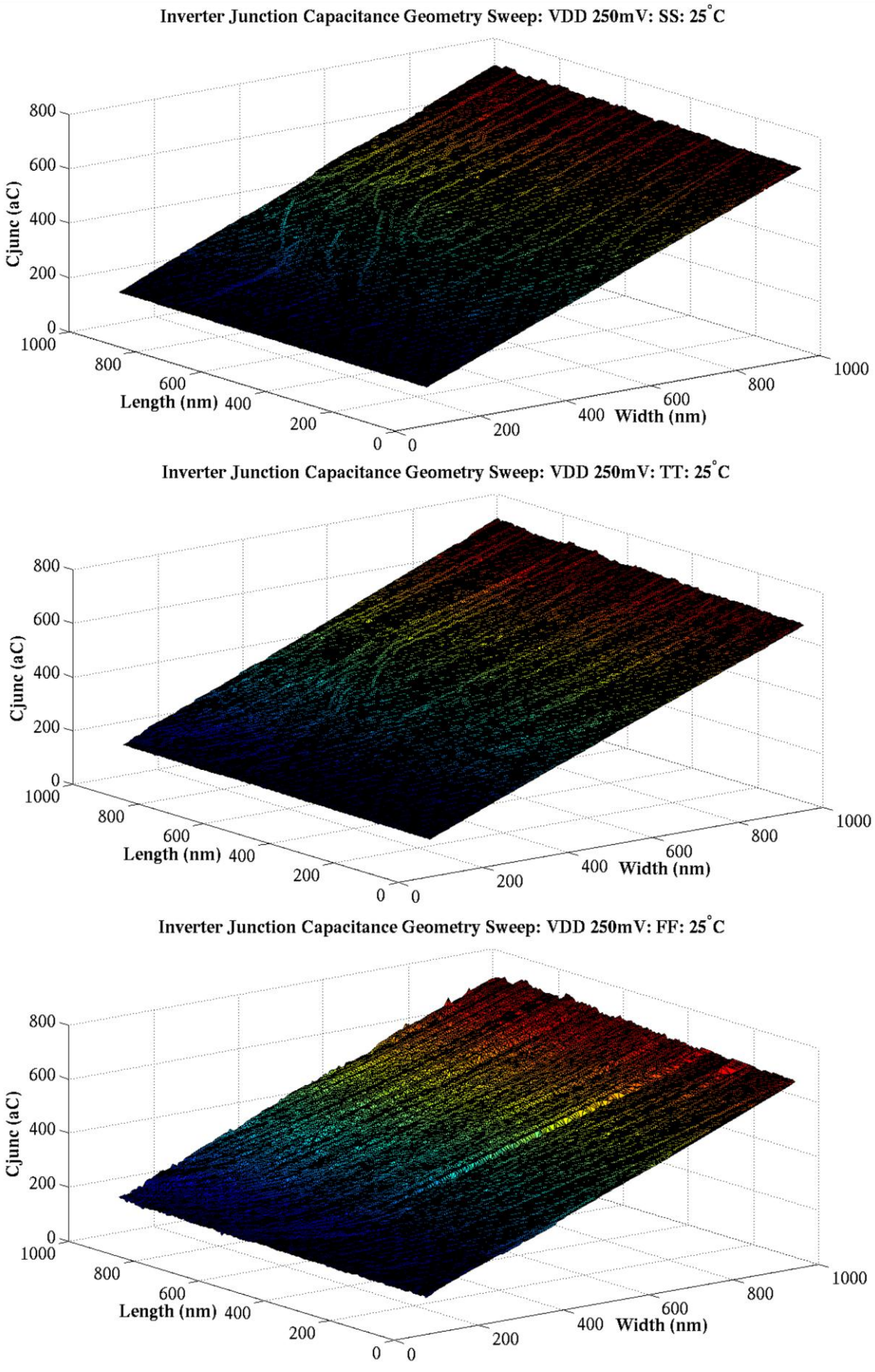


Figure 32: Junction capacitance sweeps: PMOS LVT

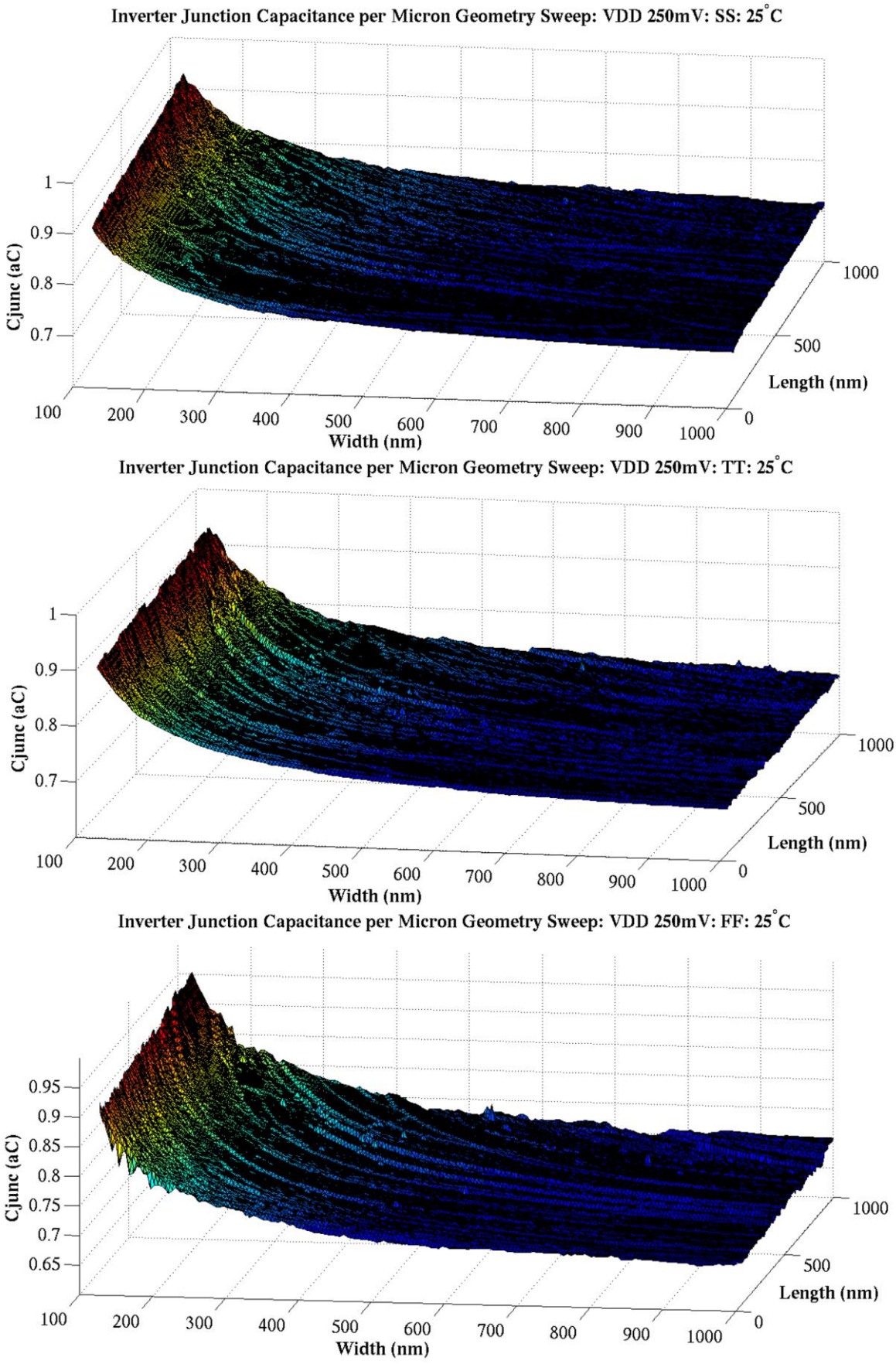


Figure 33: Junction capacitance per micron sweeps: PMOS RVT

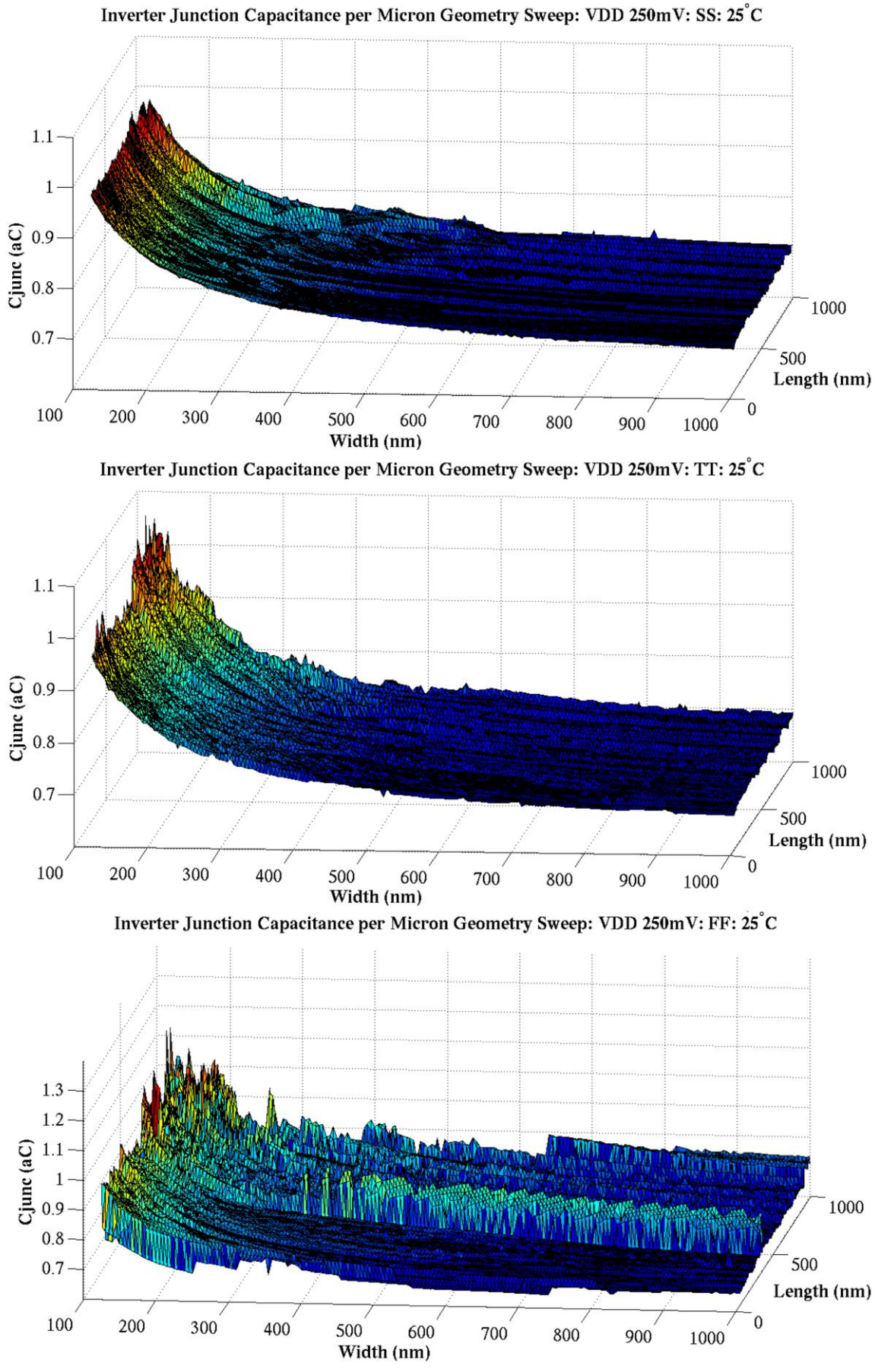


Figure 34: Junction capacitance per micron sweeps: PMOS LVT

3.6 Propagation Delay

As discussed in the previous subsection, performance is inherently dependent on device propagation delay. This is influenced by both the device's ability to sink/source current and device capacitances. Whilst both of these metrics have been independently explored based on their variation under subthreshold physical effects, the actual propagation delay may be directly simulated using a simple FO4 inverter test bench. Once again this test bench is based upon a sample test bench outlined in Harris and Weste [82]. Figure 35 shows a schematic for the test bench.

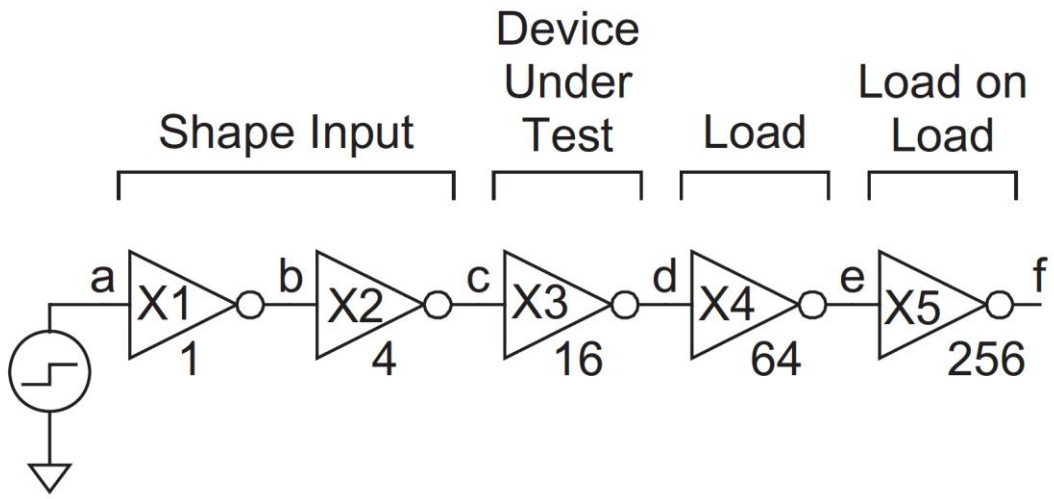


Figure 35: Propagation delay SPICE test bench

The two input stages prior to the DUT shape the input transition to the DUT, reducing the dependence of the DUT's transition characteristics on the source step function. Two stages follow the DUT in order to provide a realistic load.

To determine a reliable evaluation on propagation delay, rising and falling delays are measured and the two cases averaged to generate an average propagation delay. The measurement is triggered for a rising propagation delay from node c crossing the 50% VDD boundary and measured as node d crosses the 50% VDD boundary on the corresponding fall transition on the output. The falling propagation delay is measured for the complementary transitions.

Both the P and N devices are geometrically swept in an identical simultaneous fashion, resulting in a constant P-to-N ratio of 1:1 for all test cases. Sweeps were created for both LVT and RVT devices at all three process corners. Figure 36 shows the results for the LVT devices.

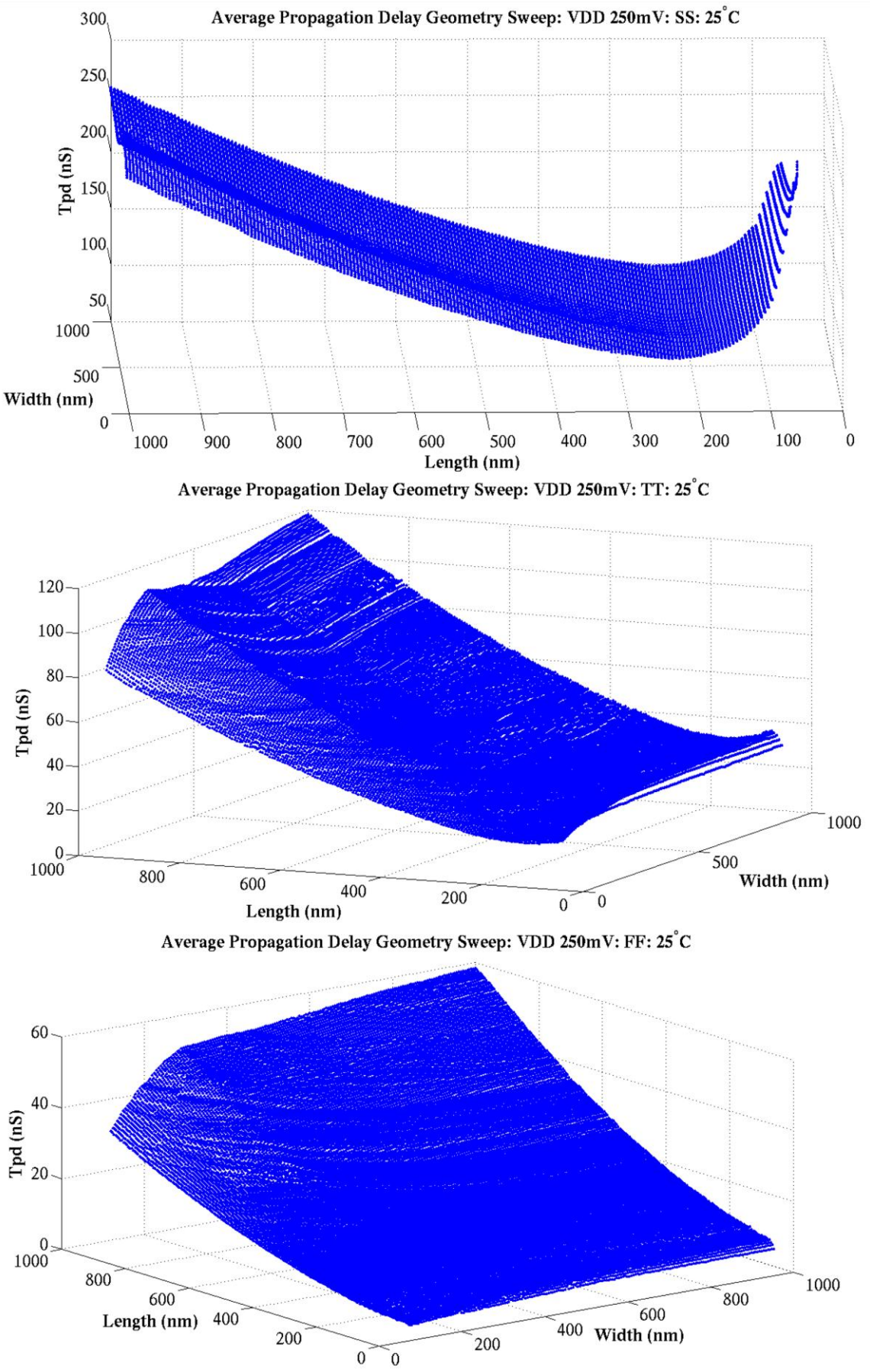


Figure 36: Average propagation delay sweeps: LVT

Figure 36 shows the average propagation delay geometry sweep for the LVT SS device at a temperature of 25°C and a supply voltage of 250mV. Both the PMOS and NMOS SS device drive current responses showed a significant and clear increase from minimum length due to the RSCE. The limited effect of the INWE showed a strong current response in response to increasing width. The gate capacitance response shows a linear proportionality to the quadrature of the device. These effects combine to show a clear decrease in propagation delay from minimum length, with the minimum delay at 230nm. With the exception of the minimum length, the response also shows a slight decrease as the device width tends to minimum. The optimal device sizing for minimum delay is therefore 230nm length, minimum width devices. Interestingly, due to the linear capacitance relationship with width and the increasing width related drive current response, the gain in current is offset by the additional capacitance of the load and the propagation delay remains fairly flat over device width.

A minimum width device therefore provides the same performance as a large width device, with the additional benefits of smaller silicon area and dynamic power reduction due to lower gate capacitance.

Figure 36 also shows the average propagation delay geometry sweep for the LVT TT device at a temperature of 25°C and a supply voltage of 250mV. Both the PMOS and NMOS TT device drive current responses showed a mixture of SCE and RSCE. This is reflected in the propagation delay response where there is a sight inflection at minimum length, but ultimately the RSCE dominated, with the lowest propagation delay occurring at 150nm. The INWE current deviation in the width for both underlying devices was more apparent, and is therefore more apparent in the propagation delay response, with a clear decrease as device width tends towards minimum. Again, the optimal device sizing is therefore larger than minimum length and minimum width.

Finally the figure also shows the geometry sweep for the FF corner at the same PVT parameters. In the underlying device physics, the SCE completely dominated the drive current responses. This is shown in the propagation delay response, where the minimum propagation delay actually occurs at minimum length. The underlying physics also showed the fast corner to be most influenced by the INWE. This again is evident in the propagation delay response, where the delay reduces greatly as the width tends towards minimum. The optimal device sizing is therefore minimum length, minimum width. In reality, in the subthreshold regime, the variation would likely see these devices upsized

from minimum length anyway. This shall be discussed later in the section on device variation.

The same test bench was used to perform the same geometry sweeps for the RVT devices. Figure 37 shows the results of these simulations.

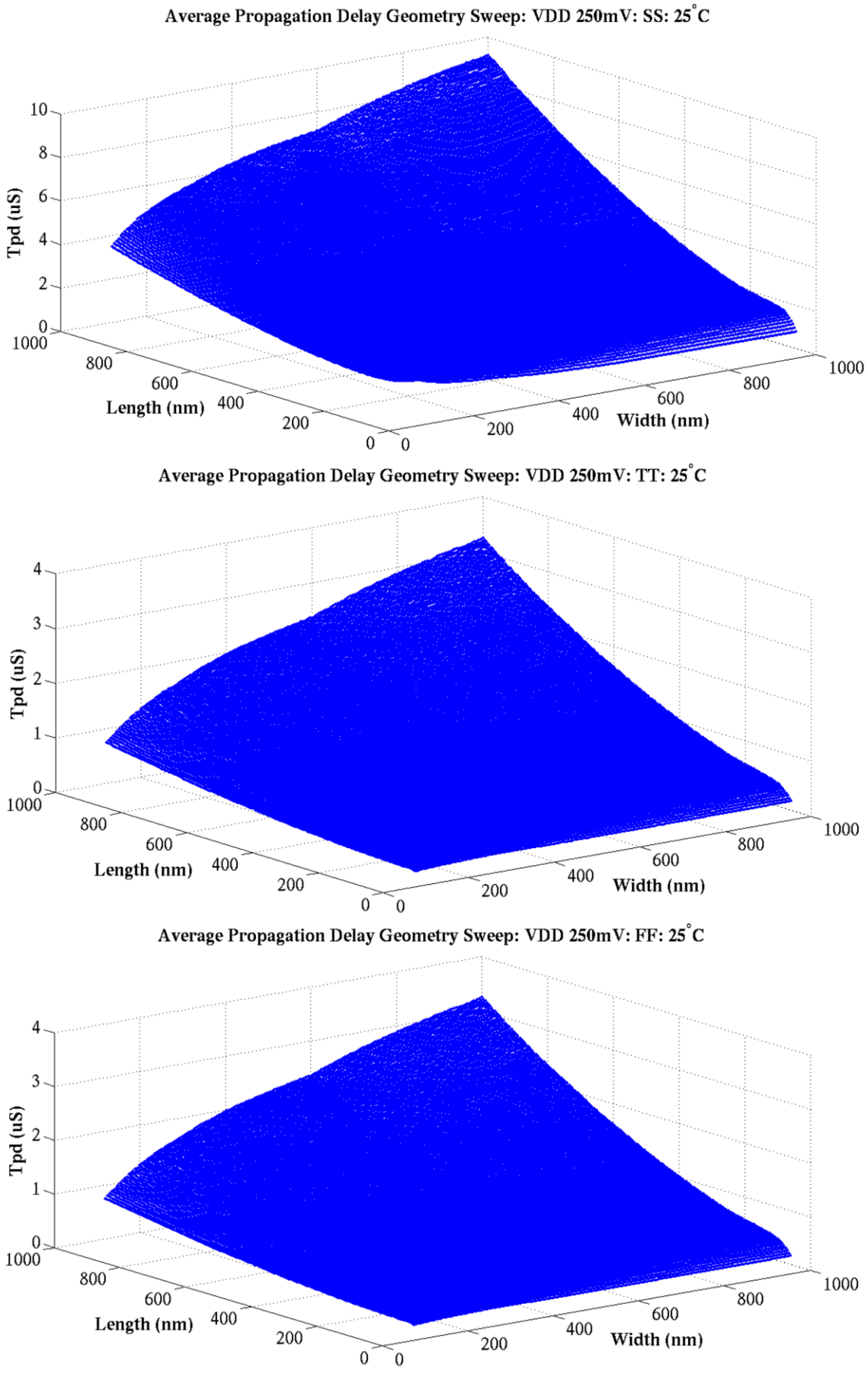


Figure 37: Average propagation delay sweeps: RVT

Figure 37 shows the average propagation delay geometry sweep for the RVT SS device at a temperature of 25°C and a supply voltage of 250mV. Both NMOS and PMOS devices at this PVT point showed width dependent response to length change in their drive current responses, i.e. the dominance of SCE/RSCE depended on the width. At minimum width, the RSCE dominated. As width increased, the SCE began to dominate. This is reflected in the propagation delay response. At minimum width, the optimal length is 200nm. As the width is increased, the optimal length decreases, eventually reaching minimum length. Due to the complexnature of the RSCE/SCE interaction, the INWE is interrupted at minimum width by the dominance of the RSCE. Other than the minimum width, minimum length point, the INWE dominates and the propagation delay decreases as width tends towards minimum. The optimal device sizing for this gate is therefore variant. The highest current per unit area and therefore performance is achieved at minimum width, 200nm length. Changes in device width would require a corresponding reduction in length to maintain the optimal value.

The figure also shows the RVT TT device at the same PVT point. Both the underlying devices show muted RSCE and slight dominance of the SCE. This is reflected in the propagation delay response, where the length response is fairly monotonic, decreasing towards minimum length. There is a slight RSCE behavior at minimum width, giving a propagation delay at minimum at 100nm length for minimum width. As the width increases, the optimal length quickly tends towards the device minimum. The INWE affects both underlying devices and is therefore apparent in the propagation delay response, giving a minimum propagation delay at minimum width for all lengths above the 100nm optimal point. Accepting the variation limitation in length, the optimum sizing for this device is therefore minimum width, 100nm length. Should the width be increased, the length would have to be decreased towards minimum length as much as variation would allow.

Finally the figure also shows the RVT FF device at the same PVT point. The SCE dominates the length response of the underlying devices and therefore, the propagation delay is lowest at minimum length for all widths. The INWE dominates the width response of the underlying devices at minimum width. The propagation delay therefore falls off monotonically towards minimum width. The optimal sizing strategy for this device is therefore minimum width and as close to minimum length as variation allows.

When the results are examined in their entirety, there are a few trends that may be observed. The first is that the optimal sizing strategy varies between the LVT and RVT devices. The LVT devices are more affected by RSCE. This translates into an increase their optimal device length. As the device length corresponds to a possible increase in cell width, the LVT devices will inherently generate larger cell footprints and therefore occupy a larger silicon area than designs synthesized from the RVT devices.
The second trend is that the process corner affects the RSCE, and therefore the slower the process corner, the larger the optimal length and corresponding cell footprint. Whilst the underlying silicon received from the fabrication house is beyond the control of the designer, the choice of process targeting strategy is. This is discussed later in the design sections.
The third trend is that the INWE affects the propagation delay in all devices and corners. For devices above 90nm in length (the minimum length considered safe for variation) the minimum propagation delay always occurs at minimum width. Utilizing the parallelization techniques outlined earlier, the current per unit width and therefore performance will always be maximized using minimum width devices. The cost of this strategy however is the related increase in leakage current.

3.7 Minimum Operating Voltage (Robustness)

The functionality of CMOS logic is dependent on the concept of a gate's ability to successfully convey state based upon the voltage it receives on it's input/s to that produced on it's output. The simplest mapping of input to output voltages of an inverter is known as the voltage transfer characteristic. From this, two input voltage thresholds are derived. The first is Voltage Input High (VIH), which defines the minimum voltage that the input will deterministically recognize as a valid state of '1'. The second is Voltage Input Low (VIL), which defines the maximum voltage that the input will deterministically recognize as a valid state of '0'. Between these voltages, the combined impact of the NMOS/PMOS devices on the drain are indeterminate, and therefore the ability to deterministically convey state has been lost and the logic cannot be deemed operational. From the two input thresholds, two output levels may be derived. The Voltage Output Low (VOL) is derived as the voltage on the output when the voltage on the input is VIH. The Voltage Output High (VOH) is derived as the voltage on the output when the voltage on the input is VIL.

In order for a gate to be able to drive itself, VOH must be greater than VIH for a state of '1' to be successfully conveyed from one stage to the next. The difference between VOH and VIH is known as the Noise Margin (High), NMH. VOL must also be less than VIL to successfully convey a state of '0'. The difference between VIL and VOL is known as the Noise Margin (Low), NML. The theoretical operation of CMOS may therefore be defined as valid when both NMH and NML are greater than zero. As the number of gates in a design increases, the likelihood of a gate failing to meet this criteria increases, and therefore the MOV increases [83].

Two primary issues are posed by operation in the subthreshold regime. The first is that as supply voltage is aggressively scaled, MOS devices stop behaving like voltage controlled current sources and start behaving in an ohmic fashion. The voltage dropped across them therefore begins to account for a greater proportionality of the rail to drain voltage, degrading the output swing of the gate. Degradation of the output levels degrades the noise margins.

The second issue is that the inherent strengths of NMOS/PMOS devices are degraded at different rates under aggressive voltage scaling. Typically in bulk planar deep submicron technology nodes, the PMOS device degrades more than the NMOS device. This results in a reduction in VOH and a corresponding degradation of the NMH. As such, aggressive voltage scaling will eventually lead to gate failure. It is therefore imperative to determine the impact of the limits of operability on device sizing.

A test bench was created to simulate these limits from the BSIM compact model. This test contains a single inverter with NMOS and PMOS devices of matched width and length, giving a 1:1 P/N sizing ratio. A supply voltage is applied, and the input voltage to the inverter swept from ground to the supply voltage to generate the voltage transfer characteristic. VIL and VIH are determined as the input voltage levels which generate first order derivatives equal to -1, a common methodological practice used in the field. VOH and VOL are then determined as the output voltages corresponding to the input voltages VIL and VIH. The noise margins NMH and NML are calculated as absolute values and also as a percentage of the supply voltage. The supply voltage is then lowered and the whole process repeated. The practical minimum operating voltage is then determined as the point at which either NMH or NML reaches 10% of the supply voltage, again a common methodological practice in the field. Finally, the whole test bench is geometrically swept over length and width for both LVT and RVT variants of the devices. Figure 38 shows the results.

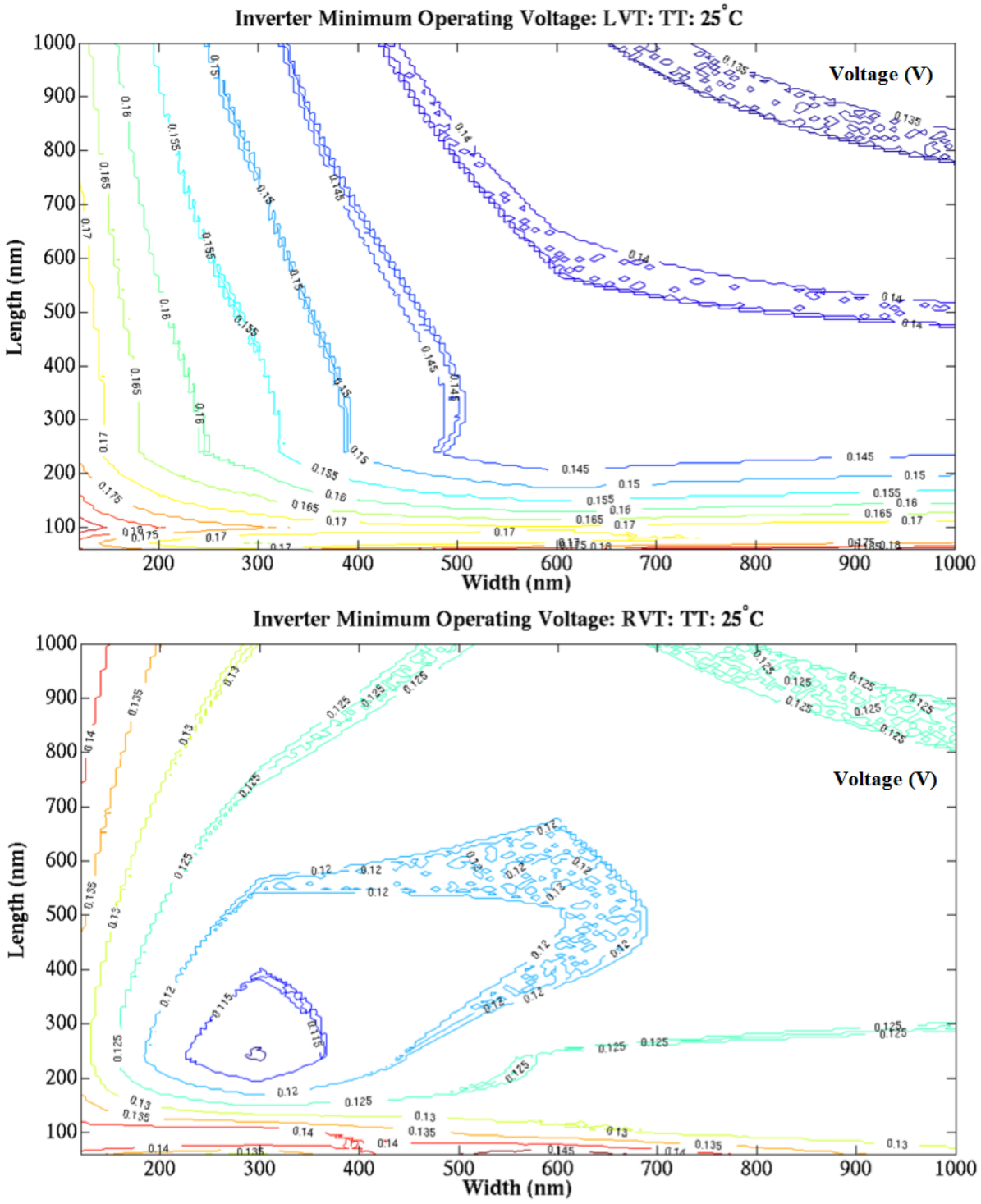


Figure 38: Minimum operating voltage sweeps

Figure 38 shows the practical minimum operating voltage of the LVT device inverter at the TT process corner and a temperature of 25°C. The trend shows that increasing the device length and width has the effect of decreasing the minimum operating voltage. Both the drive current sweeps for the NMOS and PMOS devices showed positive current trends for length and width. For the purposes of propagation delay, gate capacitance grew faster than drive current and therefore propagation delay was minimal at around 150nm length and minimum width. For noise margin, the results suggest the relationship is somewhat more complicated. The overall result is likely a combination of many factors, including Ion/Ioff ratio, P/N optimal sizing, threshold voltage variation in response to sizing, drive current, inherent ohmic nature of the devices etc. Interestingly, large gains in minimum operating voltage can be made by upsizing slightly from minimum dimensions. The gains then swiftly diminish. For a length sized at 150nm for minimum propagation delay, a 25mV improvement in minimum operating voltage can be obtained by upsizing the width to 350nm. The lowest minimum operating voltage for the analysis was 133mV for lengths/widths greater than 950mV. The highest was 190mV, covering the whole width of the minimum length devices.

Figure 38 also shows the practical minimum operating voltage of the RVT device inverter at the TT process corner and a temperature of 25°C. This is quite different to the LVT response, with a local minimum centered on a length of 240nm and width of 300nm. Increasing beyond these values increases the minimum operating voltage. Again the response is likely an amalgamation of many factors. The lowest minimum operating voltage in the response is 109mV at the 240nm/300nm minimum focal point outlined above. The highest was 149mV at minimum length.

Several conclusions may be drawn for the above analysis. Firstly, the optimal sizing for minimum propagation delay and minimum operating voltage are not the same. A design decision between performance and yield is therefore presented to the library designer. The second trend observed is that the worst minimum operating voltage occurs at minimum length for both device variants. The standard superthreshold practice of sizing at minimum length would therefore have significant impact on yield. A third observation is that the absolute values of the RVT variant are less than the LVT variant, which is initially somewhat counterintuitive, given that the natural threshold voltage of the RVT device is higher. The likely reasoning behind this is the greater Ion/Ioff ratio of the RVT device. As outlined earlier, MOS devices begin to display ohmic behavior at aggressively scaled voltages, resulting in logical behavior analogous to ratioed logic. In the LVT

device, the off device in the inverter will leak a larger amount of current, placing a greater contention on the drain, degrading the noise margin and therefore minimum operating voltage.

3.8 Chapter Summary

This chapter performed simulations on a BSIM 4.5 compact model of the chosen technology library. Geometric sweeps were performed to measure active and leakage currents for both PMOS and NMOS devices. These were shown to be highly dependent on LVT/RVT variant and process corner due to their impact on the manifestation of the RSCE and INWE. The Ion/Ioff ratios were then calculated. All responses showed the ratio degraded towards minimum length due to the SCE. The rest of the responses were explained using the underlying physics outlined in Chapter 2.

Minimum-fingered topologies were then explored and their active and leakage currents compared to a single iso-area device. All minimum width composite devices showed some level of improvement over a single iso-area device. The absolute magnitude of this improvement was highly device dependent.

Geometric sweeps to simulate gate capacitance were then performed. Little deviation resulting from RSCE or INWE was observed, thus supplementary simulations to determine junction capacitance were performed. The responses showed dominance of the direct correlation of gate capacitance and area.

Propagation delay was then simulated to determine the impact of the previously simulated current and capacitance metrics on device performance. This showed the optimal sizing was highly dependent on LVT/RVT variant and process corner. The INWE dominated for all lengths above the 90nm length variation cutoff.

Finally, the minimum operating voltages were simulated to determine the impact of device sizing on robustness. These showed little correlation to performance optimal sizing.

Chapter 4: Evaluation of Proposed Cell Sizing Strategy in Silicon

4.1 Chapter Outline

This chapter presents three subthreshold sizing strategies. Two strategies have previously been explored in the field and one is a novel contribution from this work. X1 inverter cell ranges were created from each strategy using an identical silicon cell area. These were laid out in the chosen technology node. The cells were then verified using commercial DRC (Design Rule Check) and LVS (Layout-Vs-Schematic) tools and accurate SPICE models were parasitically extracted using a commercial extraction tool. Each cell was then simulated to determine average propagation delay and leakage. Monte Carlo analysis was then performed to determine variability. Finally, the three sizing strategies were then compared.

Ring oscillators were then designed from the novel sizing strategy, signed off and committed to silicon. The dies were swept over supply voltage and temperature and the performance and leakage characteristics measured using a source meter.

4.2 Inverter Design Space

In order to determine the best method to create a standard cell for subthreshold operation, the boundaries encompassing all permutations of possible design strategies must first be defined. Two of these boundaries have previously been explored in the forms of the regular superthreshold sizing strategy and minimum width sizing strategy outlined in Chapter 2 and simulated in Chapter 3. The novel sizing strategy provides the final boundary.

The simplest cell included in a practical standard cell library is the X1 inverter. This was therefore chosen as the test case. A 12 track (2.4µm height) implementation was chosen, as this is the form factor utilized by the current ARM R&D library. Considering minimum poly overlap rules, P&R boundary distances and assuming equal P and N type diffusion areas, this affords a maximum of 840nm diffusion width for each device (maximum device width).

4.2.1 Regular (Superthreshold) Sizing Strategy

The regular sizing strategy suggests that increasing the width of the NMOS and PMOS devices increases their current producing capability, and therefore produces faster cells. Inverters of matched NMOS/PMOS width of 120, 210, 420, 630, 840nm were created.

Conventional (based on superthreshold operation) submicron physics suggests that increasing the width increases the performance and also the gate capacitance, but that the former increases at a greater rate than the latter, leading to an improvement in the propagation delay. The leakage current for the cell would also increase, as this is also proportional to width. Figure 39 shows a sample of these cells laid out in the target technology node.

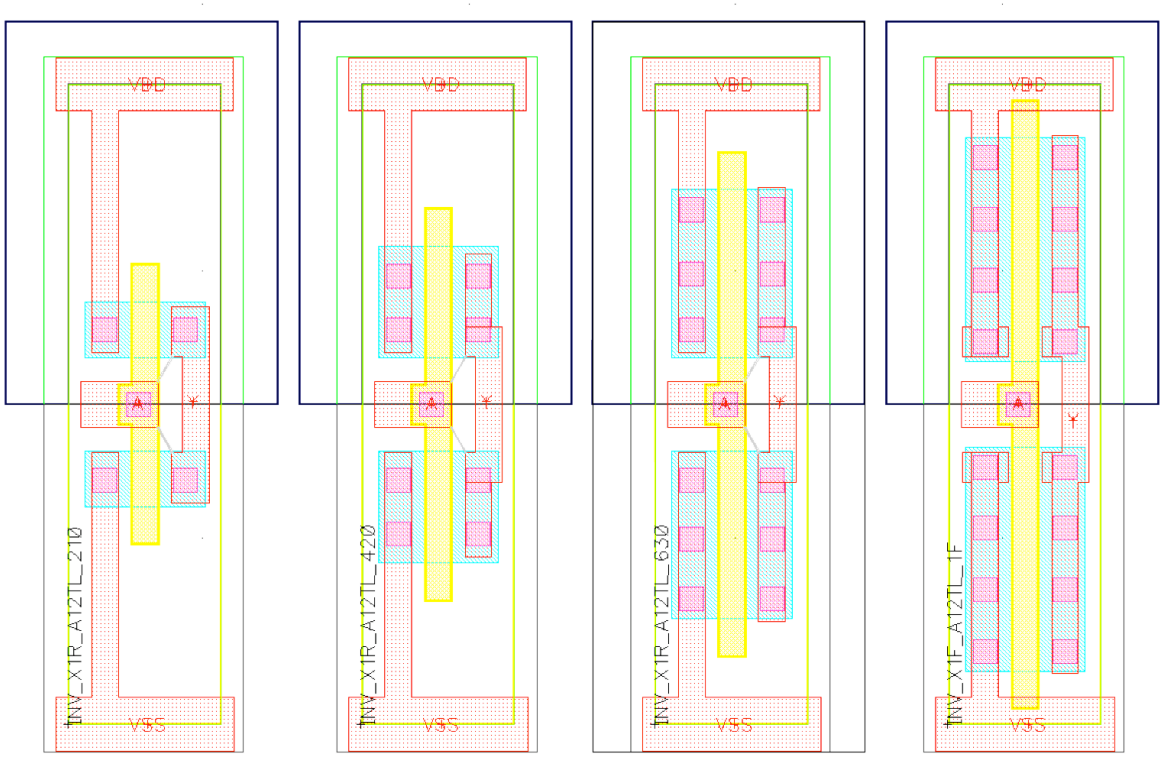


Figure 39: Regular sizing strategy inverter cell layouts

4.2.2 Minimum Width Sizing Strategy

The minimum width sizing strategy suggests that to maximize gate control and thereby the Ion/Ioff ratio, minimum width devices should be chosen as this maximizes the advantage produced by the INWE. In the chosen process the minimum width is 120nm. The minimum STI spacer between diffusion regions is 110nm. To create the sizing variations, multiple fingers of 120nm are stacked in parallel. For the chosen cell size, this allows up to four devices each for NMOS/PMOS ((4 x 120nm) + (3 x 110nm) = 810nm). Inverters with matched finger numbers of 1/2/3/4 x 120nm were created. The subthreshold physics discussed in the previous chapters suggests the addition of each device will proportionately increase both the performance and leakage of the device. Comparatively, according to the parallelization and capacitance results presented in

Chapter 3, the drive current produced should be greater and the gate capacitance smaller than the regular sizing strategy, leading to superior performance. Figure 40 shows a sample of these cells laid out in the target technology.

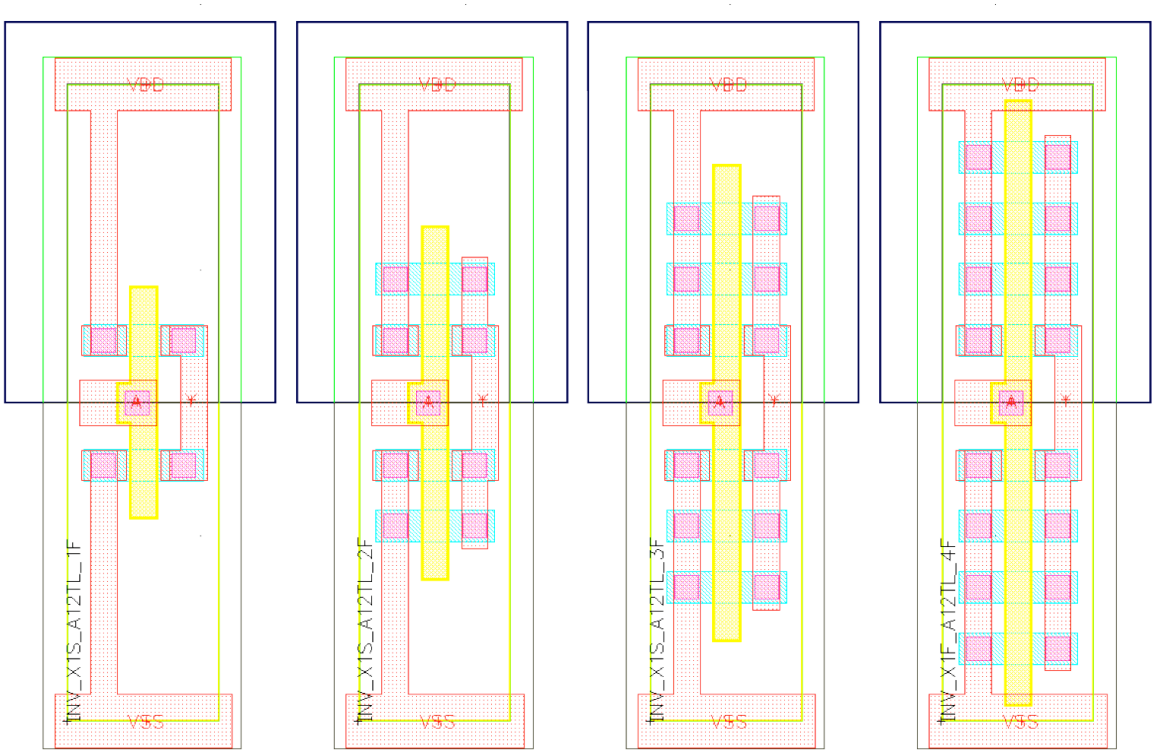


Figure 40: Minimum width sizing strategy inverter cell layouts

4.2.3 Full Diffusion Sizing Strategy

Due to the determination of previous researchers in the field to maximize the performance of the underlying transistors, the focus of proposed standard cell design strategies has always be drawn to minimum width devices. The discussion of the physics of the INWE from Chapter 2 suggests that anywhere an STI sidewall is encountered by the diffusion region, the INWE will be experienced. It is therefore possible to invert the form. Instead of assuming a blank silicon area where diffusion area may either be extended or occupied by an increasing number of quantized optimal devices, the silicon area may be assumed to be full and STI spacers added to create a geometric sizing methodology.

In the chosen technology node for a 12-track library, the full diffusion region for both NMOS/PMOS is 840nm each. The addition of a single STI spacer into this diffusion region forms two equal devices of 365nm width ((365 x 2) + (110 x 1) = 840nm). The addition of two STI spacers forms three equal devices of 205nm and three STI spacers form four equal devices of 120nm. In accordance with the INWE argument from Chapter

2, as each spacer is added, the influence of the INWE should increase and therefore, so should the performance and leakage of the cells. Figure 41 shows a sample of these cells.

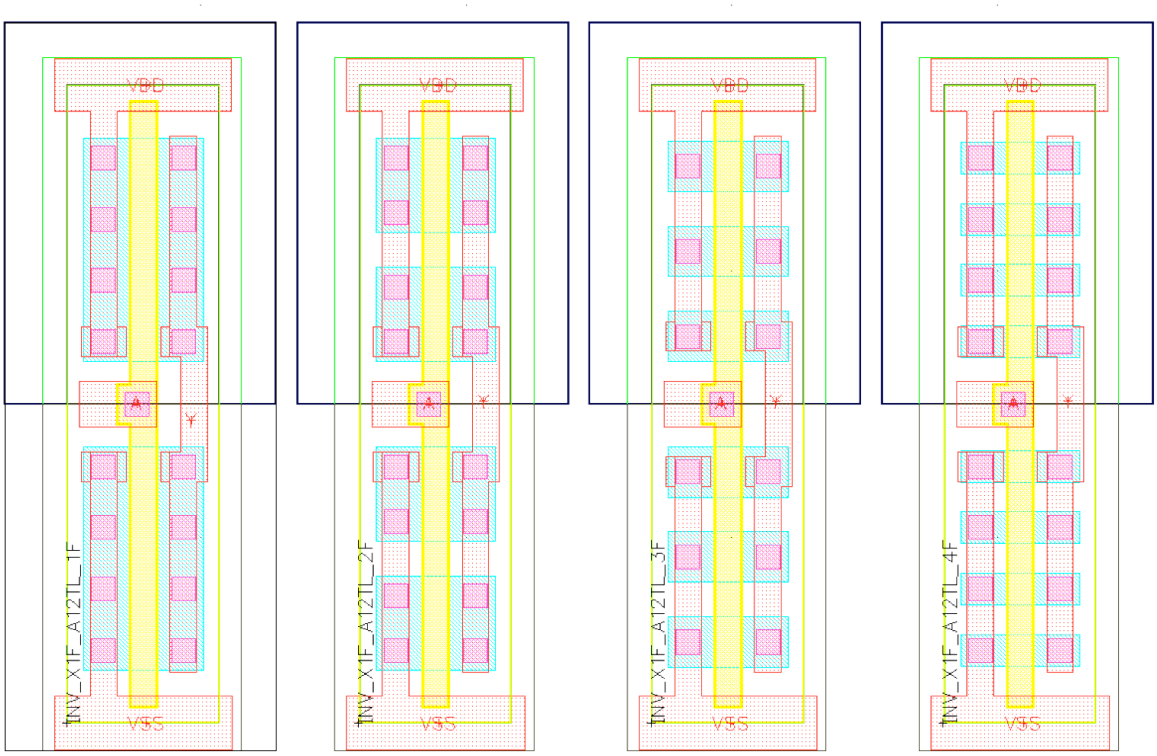


Figure 41: Full diffusion sizing strategy inverter cell layouts

4.3 Performance and Leakage Comparison (LVT Test Case)

To give a more accurate comparison of the sizing strategies, X1 inverters using all three strategies were laid out in Cadence Virtuoso from the LVT devices. To provide a fair iso-area comparison for each of the sizing strategies, the overall optimal length value of 100nm was chosen for all cells in both threshold voltages. DRC and LVS verification of the layouts were individually performed using Calibre to ensure valid compliance to the DRM for the chosen technology. Calibre was then used to perform parasitic extraction on the cells using a 3D process methodology [84]. This generated a new SPICE model, creating the model of the composite topology of one or more devices in each cell with the addition of the parasitic resistance and capacitances introduced by the vias and metal connections. This SPICE model was then re-introduced into the HSPICE test benches used in Chapter 3. The FO4 propagation delay test bench from Section 3.6 was used to simulate the propagation delay of each individual cell using the same methodology.

A second test bench was created from the FO4 test bench. The second test bench removed the step function voltage source and replaced this with a fixed source of either VDD or GND. The device under test was also removed from the global supplies and a separate supply of the same VDD was applied. The current was then measured between the supply rails of this supply whilst the DUT was static. This measured the leakage current for one state (input = high/low). The input source was then flipped to the alternate state (VDD/GND) and the static current measured again. These two measurements were then averaged to give the average leakage current of the DUT. Figure 42 shows the results for the LVT devices.

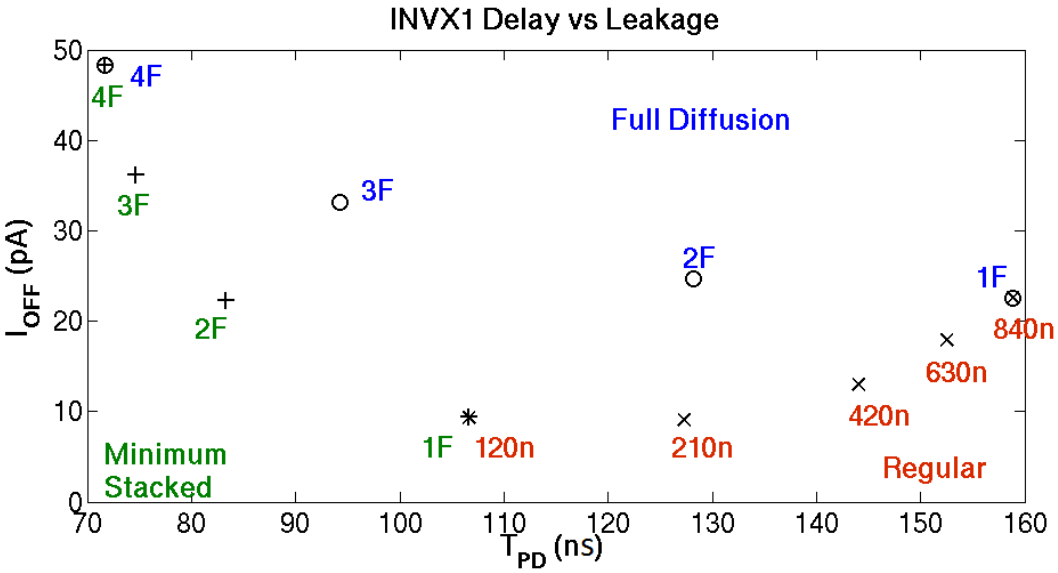


Figure 42: LVT inverter cell design space

The figure represents the entire design space for an X1 inverter at the typical process corner, nominal temperature of 25°C and a supply voltage of 250mV. Any X1 inverter of the same silicon area has to be a permutation of one or more of the three design strategies presented in the previous sections.

The regular (superthreshold) design strategy did not behave in the same manner as expected in the subthreshold regime. The lack of improvement in propagation delay in response to width increase at this PVT point (as shown in Chapter 3) was superseded by the linear increase in gate capacitance (on the load). Therefore as the width increased, the capacitance on the load grew faster than the device current driving it and the gates got slower. Moreover, as the leakage current still maintains a positive relationship with the

device width, as the width was increased, the leakage current increased. Therefore, somewhat counter intuitively, the optimal device in the regular sizing strategy was actually the device with the lowest width (120nm). In accordance with the variation discussion in Section 2.4, this is also the device with the lowest gate area and therefore highest variability. This is explored further in the next subsection.

The minimum stacked sizing strategy exhibited the best performance to leakage ratio, as was expected given the devices have the highest INWE influence and therefore highest gate controllability and performance-to-leakage ratio. As the number of minimum width fingers increased, the inverters got faster and leakier. The correct proportionality was therefore observed. However, the same variation argument applies to the lesser number of stacked devices as applied to the regular sizing strategy (Lower number of fingers equates to higher variability).

The novel full diffusion strategy exhibited the largest range in propagation delays, spanning the full gamut of the two other strategies combined. The cost of this range was the additional leakage arising from the maximal usage of the area and therefore greatest overall width of each cell. This however should be a benefit in terms of variation, which is inversely proportional to area as discussed in Section 2.4.2.7.

4.4 Stack Forcing and VT Generalization

The above design methodology can be equally applied to both LVT and RVT devices and both device VT's can be stack forced. This creates 4 permutations. All sizing strategies were therefore laid out in all four permutations, parasitically extracted and simulated in the same test benches from the previous subsection. Figure 43 shows the results.

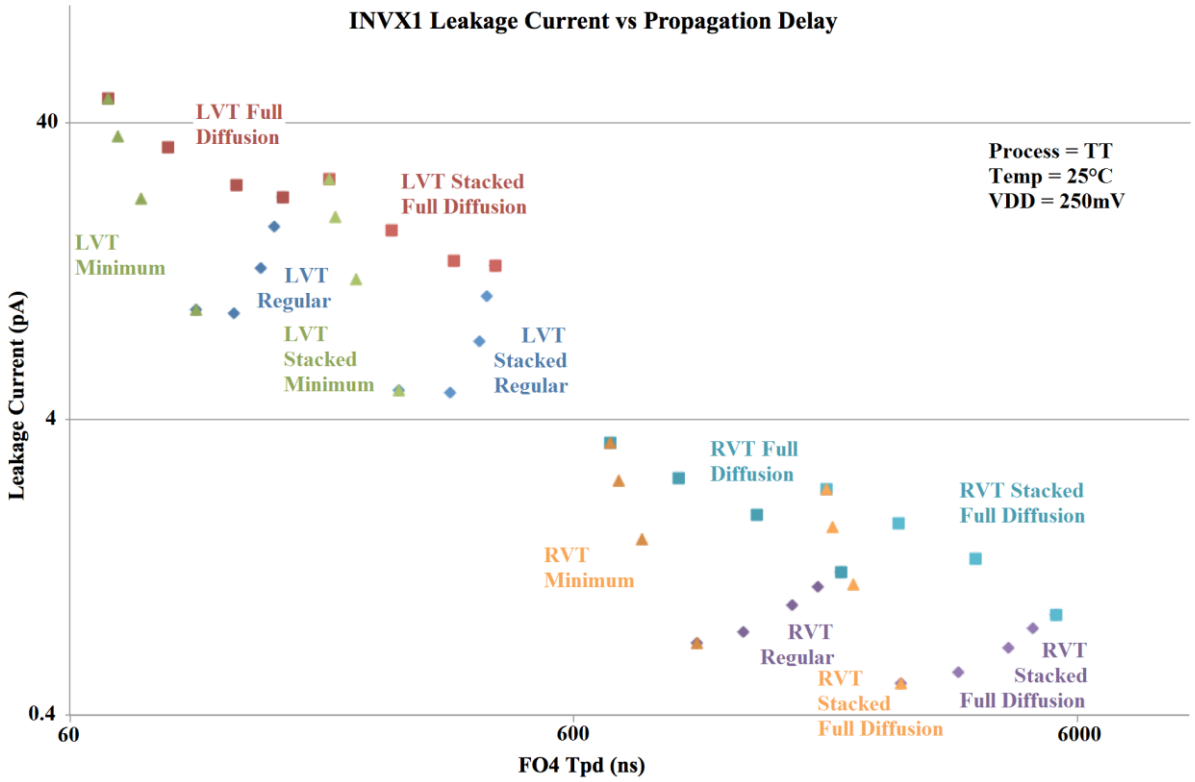


Figure 43: Full multi-Vt inverter cell design space

The results reveal several important features of the cell strategies. The first is that, whilst the absolute performance/leakage values were not directly transposed between LVT and RVT, the underlying relationships remain fairly constant and therefore the advantages/disadvantages posed by the LVT test case in the previous subsection apply equally to RVT.

The second feature is that the strategies respond equally to stack forcing. As the cells were stack forced, the additional device created additional resistance in the pull up /pull down paths. This has the effect of reducing leakage but increasing propagation delay. However it must be noted that stack forcing proportionally increased the leakage current greater than it lowered the propagation delay. This means that the supercutoff effect was not present or is superseded by another physical effect in the chosen technology library. This correlates with the results of the stack forcing experiment of Section 3.4. Moreover, stack forcing doubles the gate area and therefore gate capacitance of the cells. This would substantially increase the dynamic power consumption, a side effect that must be considered before the liberal use of stack forcing in any design.

4.5 Monte Carlo Analysis (Variation)

The FO4 propagation delay test bench, complete with parasitically extracted models for the three sizing strategies, was then modified to include the Monte carol BSIM models for local variation analysis. Each inverter cell, in each sizing strategy, for each of the four permutations was simulated for 1000 iterations. The data was collated, and the mean (μ) and standard deviation (σ) calculated. The critical variation metric sigma/mu (σ/μ) was then determined for each inverter cell. Table 1 presents the results in a tabular from. Figure 44 presents the result as column graphs.

LVT

T_{PD} (ns)	Mean (μ)	Std. Dev. (σ)	σ/μ
Regular			
120nm	118.24	32.28	0.273
210nm	139.73	35.82	0.256
420nm	154.98	30.91	0.199
630nm	159.24	26.27	0.165
840nm	164.28	23.7	0.144
Min. Stacked			
1F	118.24	32.28	0.273
2F	83.57	14.97	0.179
3F	71.51	10.38	0.145
4F	68.73	8.5	0.124
Full Diffusion			
1F	164.28	23.7	0.144
2F	127.95	18.96	0.148
3F	89.86	12.79	0.142
4F	68.73	8.5	0.124

LVT Stacked

T_{PD} (ns)	Mean (μ)	Std. Dev. (σ)	σ/μ
Regular			
120nm	306.16	57.71	0.189
210nm	381.91	70.07	0.183
420nm	421.10	61.55	0.146
630nm	426.47	51.85	0.122
840nm	436.65	46.38	0.106
Min. Stacked			
1F	306.15	57.71	0.188
2F	236.20	29.69	0.126
3F	208.77	21.11	0.101
4F	202.80	18.83	0.093
Full Diffusion			
1F	436.65	46.38	0.106
2F	363.64	39.78	0.109
3F	270.03	28.48	0.105
4F	202.80	18.83	0.093

RVT

T_{PD} (ns)	Mean (μ)	Std. Dev. (σ)	σ/μ
Regular			
120nm	1452.8	731.18	0.503
210nm	1607.4	602.37	0.375
420nm	1818.0	465.90	0.256
630nm	1974.0	407.10	0.206
840nm	2157.8	380.54	0.176
Min. Stacked			
1F	1452.8	731.18	0.503
2F	836.58	283.88	0.339
3F	654.30	174.74	0.267
4F	609.00	139.57	0.229
Full Diffusion			
1F	2157.8	380.54	0.176
2F	1381.9	273.23	0.197
3F	877.06	191.54	0.218
4F	609.00	139.57	0.229

RVT Stacked

T_{PD} (ns)	Mean (μ)	Std. Dev. (σ)	σ/μ
Regular			
120nm	3811.0	1390.3	0.365
210nm	3399.6	1215.1	0.276
420nm	4958.8	939.69	0.190
630nm	5322.3	809.75	0.152
840nm	5783.4	751.01	0.130
Min. Stacked			
1F	3811.0	1390.3	0.365
2F	2544.9	622.91	0.245
3F	2143.3	419.20	0.196
4F	2047.0	367.70	0.180
Full Diffusion			
1F	5783.4	751.01	0.130
2F	4038.9	476.50	0.143
3F	2821.0	448.15	0.159
4F	2047.0	367.60	0.180

Table 1: Sizing strategies Monte Carlo (variation) analysis

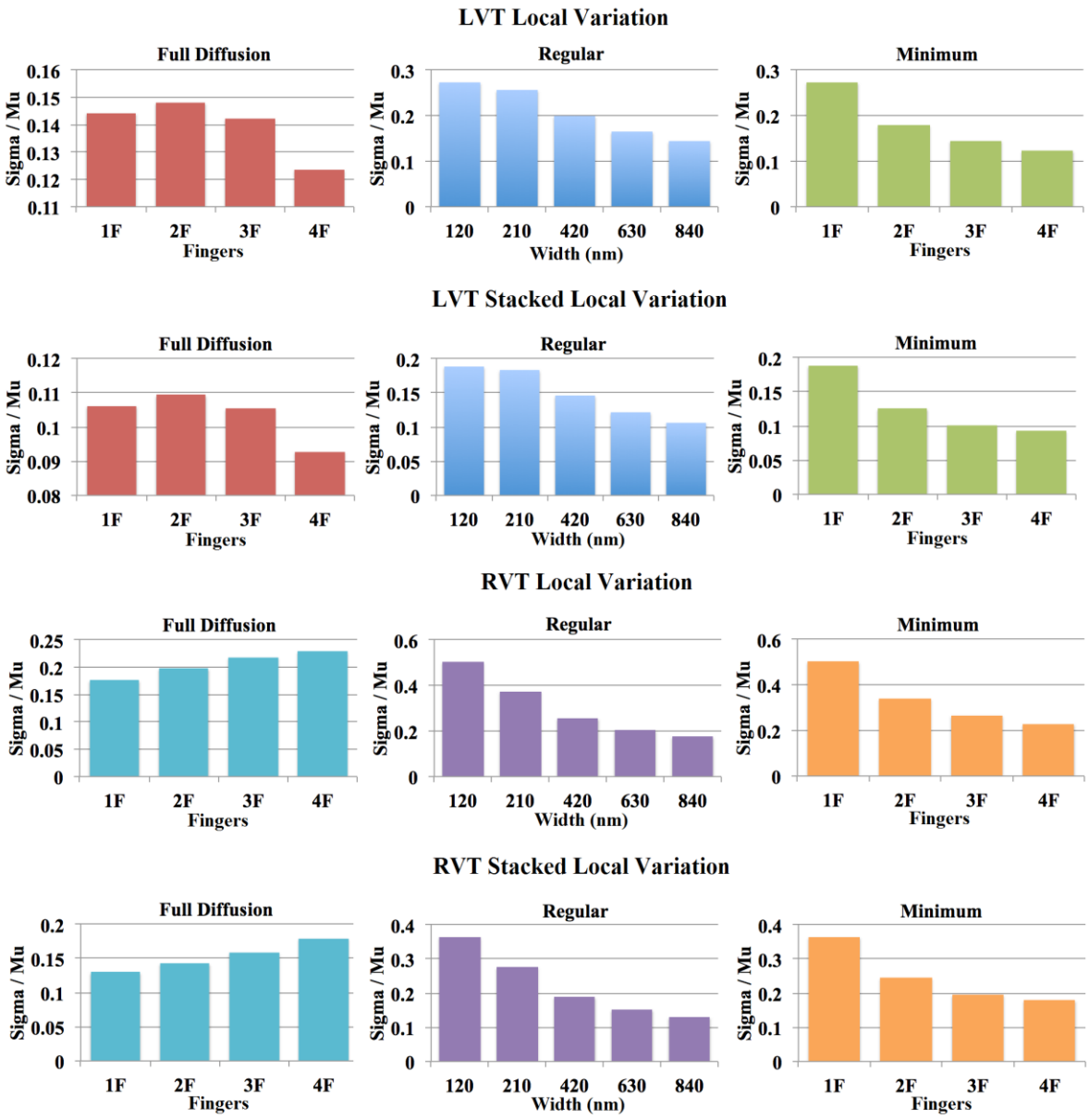


Figure 44: Sizing strategies Monte Carlo (variation) analysis

The table and figure show the variation characteristics of average propagation delay for each inverter cell at the typical process corner, supply voltage of 250mV and operating temperature of 25°C. For all four permutations, the variation of the regular sizing strategy follows the inverse-quadrature proportionality outlined in the variation discussion of Section 2.4.3 (i.e. variation reduces in proportion to the square root of the gate area). The mean average propagation delay increases as the width is increased, confirming the solitary test benches performed earlier. On the other hand the standard deviation reduces as the width is increased. This results in a reduction of sigma/mu as the gate width increases.

The variation of the minimum sizing strategy follows the Gaussian averaging proportionality. The metrics that govern local variations are typically of a Gaussian distribution. In the minimum sizing strategy, all fingers are of the same (minimum) width. Therefore the Gaussian distribution for the variation of all fingers is the same. By increasing from one to two fingers, the effect is that of taking two samples from the same underlying distribution and summing their characteristics to form a composite device. As the number of fingers tends to infinity, the characteristic of the composite device tends to the mean of the distribution. Therefore as the number of fingers increases, the variation decreases. This is the behavior observed for all four permutations of the sizing strategy.

The variation of the novel full diffusion sizing strategy is an amalgamation of the two aforementioned relationships. As each successive STI spacer is added to create more fingers from the same diffusion area, the overall amount of 'utilized width' decreases. In accordance with the inverse-quadrature proportionality, the variation therefore increases. However, adding the spacer increases the number of equally sized fingers, therefore the variation decreases. These opposing relationships counteract to form a unique variation proportionality where the amount of variation change depends on the relative strength of the contribution of the two underlying variation functions. These are discussed later.

For the LVT permutation of the sizing strategies, the range of variation of the regular sizing strategy was 90% (from the most variable (120nm) to the least variable (840nm)). The range of variation for the minimum sizing strategy was 2.2x (1F to 4F). The range of variation for the full diffusion sizing strategy was only 19% (2F to 4F), locked to the least variable device in the entire design space (4F).

For the LVT stacked permutation of the sizing strategies, the range of variation of the regular sizing strategy was 78%. The range of variation for the minimum sizing strategy

was 2.27X. The range of variation for the full diffusion sizing strategy was only 17%, again locked to the least variable device in the entire design space.
The RVT permutation follows a similar trend, with the range of the variation for the regular sizing strategy at 2.86X, the variation range for the minimum sizing strategy is 2.2X and the variation range of the full diffusion sizing strategy is 28%.
Finally the RVT stacked permutation had variation ranges of 2.81X, 2.03X and 38.5% for the regular, minimum and full diffusion sizing strategies respectively.

An overview of all permutations reveals interesting trends. Firstly, for the three permutations of interest to most designers (LVT, LVT stacked, RVT), the sizing strategy with the highest range of variability is actually the minimum width sizing strategy, which is the most frequently used in the subthreshold regime [50] [38] [85] [86] [79] [78] [75] [87]. In accordance with the underlying driving force behind this strategy, the performance increase gained from the INWE, this sizing strategy contained the fastest cell in the whole design space in each permutation, the 4 x 120nm fingered device. There is no other way to design a faster X1 inverter in a 12 track library than using this cell. Interestingly, in the LVT permutations, this is also the cell with the lowest variation.
Alternatively, for the three permutations of interest to most designers (LVT, LVT stacked, RVT), the sizing strategy with the lowest range of variability is the proposed full diffusion sizing strategy. The largest variation from the lowest variability cell was never more than 19% in LVT and 28% in RVT. This has a direct impact on the usability in terms of minimum VDD and yield. In the RVT permutation, the cell with the lowest variability was actually the 1F full diffusion / 840nm cell in the regular sizing strategy. This means that the full diffusion sizing strategy is the only strategy to always contain the cell with the lowest variability across all permutations of interest.
The difference in LVT and RVT trends for the full diffusion strategy are also interesting. In the LVT permutations, the Gaussian averaging effect is greater than the inverse-quadrature relationship. Therefore as the number of fingers increase, the variation of the cells generally decrease. In the RVT permutation, the opposite is true. The loss of area outweighs the averaging effect and the cell variability increases as the number of fingers increases.
Quick observation of the results shows that the in terms of absolute values, the LVT devices are inherently less variable than the RVT devices and stack forcing always lowers the performance variability of the design. Both of these concepts follow the consensus

established in the field, where stack forcing is often used to achieve a reduction in variability. However, when combined with the sizing strategies, the range of variation in each strategy benefitted only slightly.

4.6 Physical Design Overview

The important underlying physical characteristics of this technology node have now been studied. It is important to establish the trade offs between the choices of library design so that the designer can match the benefits of the library to the intended circuit.

The analysis in the previous subsection revealed that the fastest cell is always the cell with the highest INWE fillip. This will always be the cell with the highest number of minimum width fingers in accordance with the physics of Section 2.5.7. In the case of the 12 track library, this is the 4F cell shared between the minimum and full diffusion sizing strategies. Therefore, if the objective of intended circuit is speed increase, the designer should choose one of these strategies.
The analysis also revealed that the cell with the lowest variability is dependent on the underlying physics and is therefore not always the same cell. For the chosen technology node, the 4F cell shared between the minimum and full diffusion sizing strategies exhibits the lowest variability when the LVT devices are used. However, the 840nm/1F cell shared between the regular and full diffusion sizing strategies exhibits the lowest variability when then RVT devices are used. Therefore the best design choice for circuits where variability is the key design metric is the full diffusion strategy.
Other than the reduction in non-recurring engineering cost associated with the reuse of a regular sized library, there seems no logical benefit to using the regular sizing strategy in the subthreshold regime.

From the results presented thus far, it would seem that the consensus in the field of using the minimum sizing strategy is correct, as the characteristics in the subthreshold regime are most advantageous. It provides the highest performance-to-leakage ratio and would therefore reduce leakage energy. It also has the lowest gate area and therefore gate capacitance, lowering dynamic energy. There is however a problem. The amount of variation introduced by removing fingers in this topology introduces too much variation for it to be feasibly used. In reality, any design under 3 fingers in the minimum sizing strategy is simply too variable, as is any design under 630nm in the regular sizing

strategy. This argument was raised in the constant yield study presented in Section 2.6.3. Both of these strategies are therefore limited in range.
This is not true for the proposed full diffusion sizing strategy. The entire range of cells satisfies this variation watershed and may therefore be included in the library. The cost of this reduction in variation is an increase in leakage, and therefore leakage energy. The gate area and therefore gate capacitance falls between the minimum and regular sizing strategies. This means that the full diffusion strategy has an inherent dynamic energy consumption better than regular but poorer than minimum. As the full diffusion strategy shares the same 4F cell as the minimum sizing strategy, the maximum circuit speed achievable is also shared between the two and therefore a circuit constructed from the minimum sizing strategy can be no faster than one from the full diffusion sizing strategy. The advantage differences between the two are that the minimum sizing strategy has lower leakage and dynamic energy, but the full diffusion has lower variation and greater range.

For circuits where static timing analysis shows very similar path times for all paths, the full diffusion sizing strategy offers no benefit. If the timing constraint and dynamic energy consumption target of the circuit can be met with the regular sizing strategy, this should be used as it offers the lowest leakage energy contribution. If either of these two constraints cannot be met, the minimum sizing strategy should be used.

However, most circuits are very complex, and static timing analysis reveals path times of varying lengths. The primary objective of the synthesis tool is to meet timing. Given a range of cells, it will always introduce the fastest cell to ensure this timing objective is met. The synthesis tool may then perform a leakage recovery step, which iteratively introduces slower, less leaky cells until reaching the point of invalidating the timing constraint. If the timing requirement of the faster paths is a large distance away from the critical path, inefficiencies will inevitably occur with the use of the minimum sizing strategy as the synthesis tool will have no option but to use cells faster and leakier than necessary to construct the path.
A good way to visualize this is to imagine a circuit where 5% of the paths require the fastest possible cells. From the discussion in Section 4.4, this is the LVT permutation of the 4F cell shared by the minimum width and full diffusion libraries. Now envisage that the remaining 95% of the paths are 10-15 times slower. The discussion from Section 4.4

would suggest that the best cells would be chosen from the RVT permutation. In the superthreshold regime, this problem is solved exactly in this manner, via multi-Vt synthesis. However, the limits in the ranges of the minimum and regular sizing strategies preclude multi-Vt synthesis in the subthreshold regime.
To put this into context, for the 840nm X1 inverter cell in the chosen technology node, the LVT permutation offers a 10% performance increase for a 1.69X leakage increase in the superthreshold regime. The same cell in the subthreshold regime offers a 9.61X performance increase for a 1.78X leakage increase. The performance difference of the two VT devices is simply irreconcilable by the synthesis tool.
This is where the proposed full diffusion sizing strategy has the advantage because the full range can be used. By using the LVT stacked devices sparingly to introduce a bridge between the LVT and RVT permutations, this gives the synthesis tool a full range of performance spreading from LVT right into RVT, allowing multi-Vt synthesis to be safely re-introduced into the subthreshold regime. As the design can now spread safely back into the RVT permutation, levels of leakage recovery can be reached that would otherwise be prohibited.

The second hypothesis of the thesis may now be presented:

Can the Full Diffusion sizing strategy be used, along with any necessary stacked interstitial libraries, to create a more energy efficient design by the increase in granularity of multi-vt synthesis in the subthreshold regime?

The remainder of this work aims to test this hypothesis by proving the results from the full diffusion simulations match the physics in silicon, and then constructing circuits of a reasonably complex nature (Full 32-bit AES cores, Chapter 6) and comparing the energy-performance profile against a state-of-the-art subthreshold library constructed from the regular sizing strategy (The ARM LE (Low Energy) Library [88]).

4.7 Extension to Other Gate Types

As few useful circuits can be constructed from inverters alone, it was important to establish that the sizing strategy could be successfully extended to other combinational logic gates. At this stage, the decision was taken to create NAND2, NOR2 and AOI22, the latter used extensively in subthreshold logic for multiplexors. These basic gates, along

with the INV gate from the previous subsection, are enough to test the commercial methodology and synthesize small circuits. Figures 45 and 46 show the NAND2 gates and AOI22 gates respectively

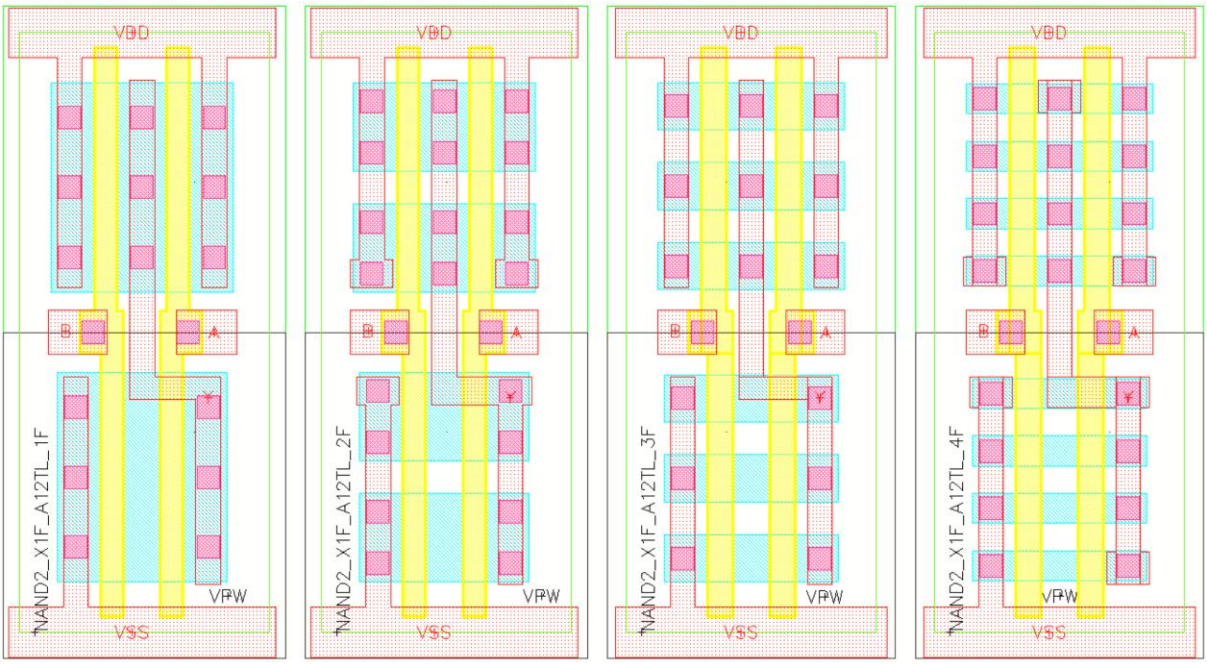


Figure 45: Full Diffusion NAND2 cell layouts

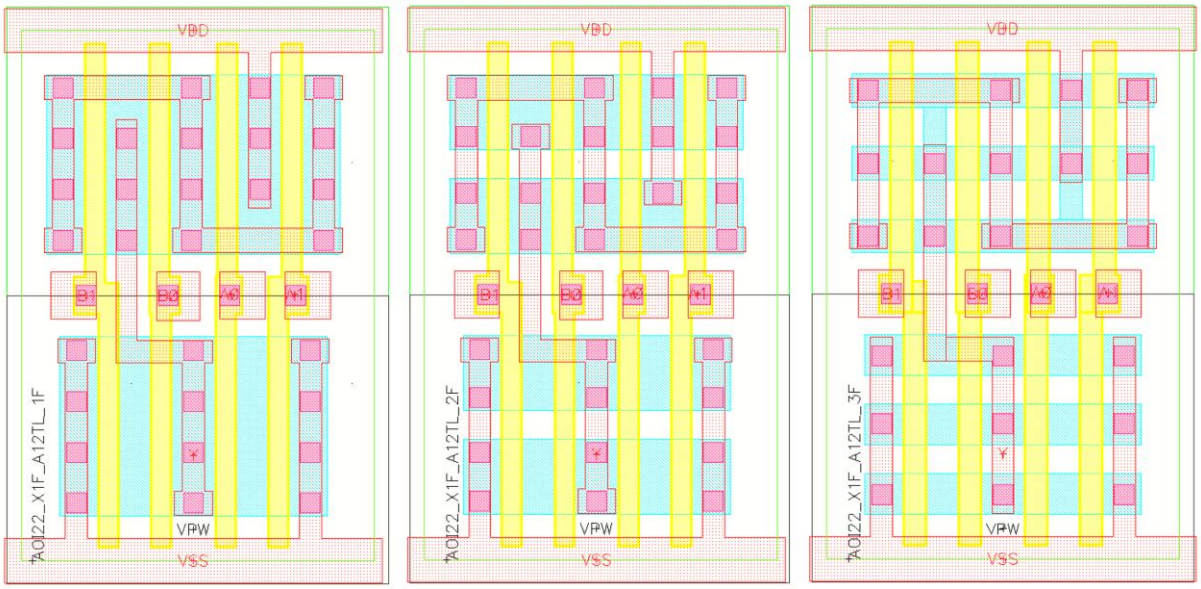


Figure 46: Full Diffusion AOI22 cell layouts

Figure 45 shows that DRC clean NAND2 gates can be constructed for all four finger permutations without serious issues arising from the design rules. Other than 'T' section ends on metal routing to obey VIA extension rules (for the electromigration issues outlined in Section 2.3.4.2.2), no additional layout design effort was required above and beyond that applied to the inverters of the previous chapter. The NOR2 gates are simply NAND2 gates with the pull up and pull down network routing flipped. Hence they are omitted for the sake of brevity.

Figure 46 shows the full diffusion sizing strategy applied to AOI22 gates. These gates are more complex due to the requirement of four inputs. This has the effect of crowding the center of the cell and lowering the allowable diffusion area. Moreover, there is significant crowding on the metal layer in the pull up network. This means that vias cannot reach all diffusion fingers to connect to the appropriate net. Routing is therefore completed on the diffusion layer. For the given technology, metal routing has a resistance of 0.1 Ohms per square. Diffusion routing has a resistance of 15 Ohms per square. This type of design is therefore typically precluded in superthreshold cell design, where RC delays are on a similar order of magnitude to switching delays. However, in subthreshold operation, switching delays are several orders of magnitude greater than RC delays, and therefore introducing this additional resistance is perfectly acceptable to complete the cell [89]. The additional strut between diffusion fingers in the final (3 finger) cell changes the spacing rules from the DRM. Instead of the spacing bounded only on two sides like the other cells, the strut forms a C based spacer, which the process demands a greater minimum distance for (110nm for two sides, 180nm for C shape). This additional spacer size accounts for the inability to create a 4 finger variant in the available height of a 12 track library.

A new HSpice test bench was created based on the propagation delay test bench from the previous chapter. All possible input combinations for the three gates were sensitized to the output by tying off unused inputs to the VDD or GND rails. For NAND2/NOR2 this created A-to-Y / B-to-Y / Both-to-Y combinations. For AOI22 all singular inputs were sensitized, along with test cases for Both-A's-to Y / Both-B's-to-Y / All-to-Y. To simplify the test bench, the FO4 aspect was dropped, with each of the five stages simply driving a single copy of itself. The average propagation delay was calculated from all test cases for each gate.

A second test bench was then adapted to simulate leakage current (also outlined in the previous chapter). The leakage current for each cell was simulated for all permutations of the inputs to generate leakage current values for all possible input states. These were then averaged to calculate an average leakage current.

Given that the performance of a circuit is determined by the slowest (critical) path, the propagation delay of focus was the slowest transition for each gate. For both the NAND2 and NOR2 gates, this is the B-to-Y transition. For the AOI22 gate, the slowest transition was A0-to-Y. Figure 47 shows the average propagation delays and leakage currents for these transitions.

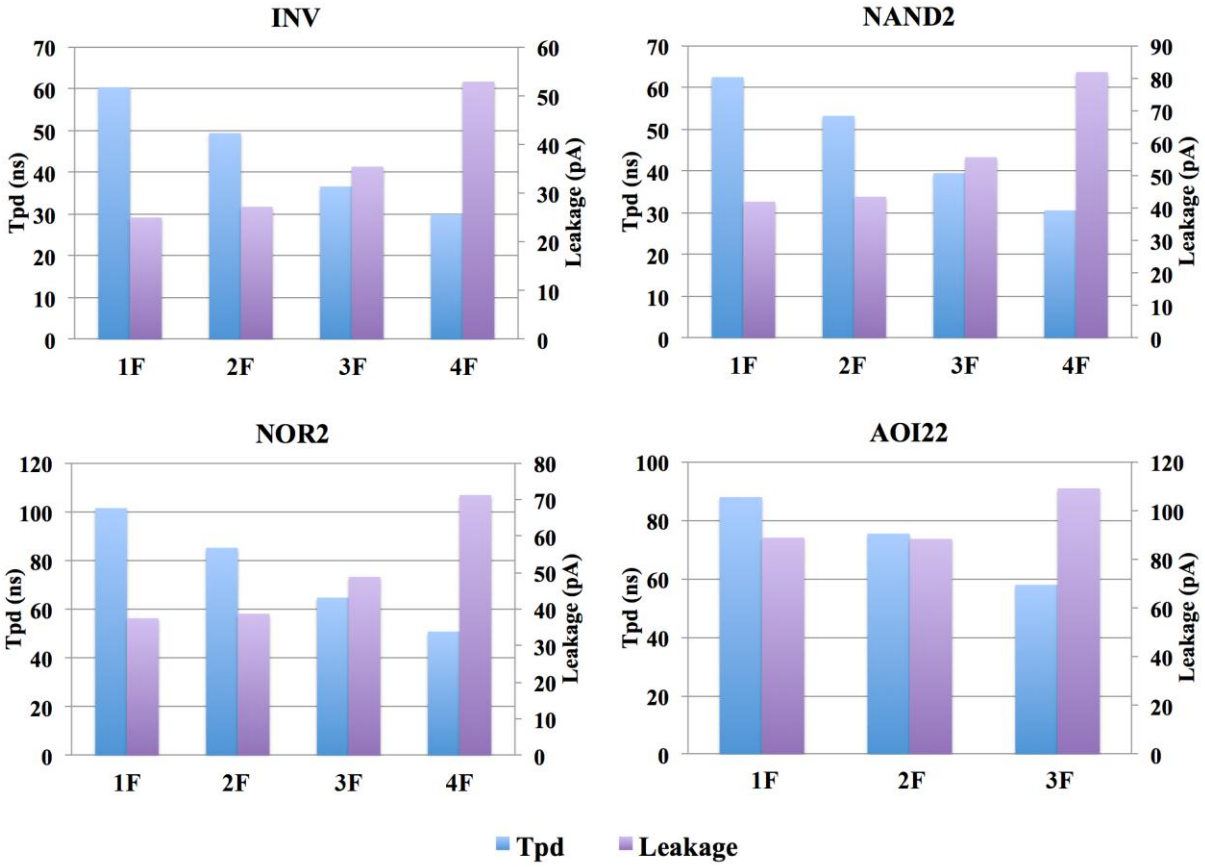


Figure 47: Combinational gates characterization

Figure 47 shows the average propagation delay and leakage current characteristics for the 4 gate types at the typical process corner, temperature of 25°C and supply voltage of 250mV. The figures shown are for gates using the LVT variants of the devices. All four gates follow the same trends established for the inverters thus far presented. As the

number of fingers increases, the impact of the INWE increases, lowering the propagation delay and increasing the leakage current. For the INV/NAND2/NOR2 gates, a decrease of around 2X can be achieved in the propagation delay at the penalty of around 2.1X increase in leakage current. For the AOI22 gate, the decrease in propagation delay is a modest 30% for 15% increase in leakage. This analysis shows that for the simpler combinational logic gates, the same benefits of the full diffusion sizing strategy are experienced. For more complex gates, a decrease in propagation delay is still observed, but the magnitude of improvement is less. The same trends are observed in the RVT devices and therefore, in aid to brevity, more detailed results are omitted here and presented in the following chapter.

4.8 Ring Oscillator Design

All of the work presented thus far has been based on simulations from the BSIM4.5 compact SPICE models provided in the PDK from the fabrication house (TSMC). Whilst this is a mature process node and the models have seen several iterations and revisions since initial release, the full diffusion sizing strategy uses these models in a way that they have not intentionally been designed for. The vast majority of VLSI designers using the process node are targeting superthreshold operation. Therefore the focus on the accuracy of the models is always at minimum length where superthreshold devices will be invariably sized. Even if subthreshold operation is a supplementary focus of model accuracy, the consensus in the field is the minimum width sizing strategy; therefore the INWE may only be accurately modeled at minimum widths. It is imperative to ensure that the results generated from the models are an accurate representation of the device behavior before the cells are used to commit any complex circuitry to silicon. The performance and leakage characteristics of a range of cells therefore required silicon validation.

It is difficult to measure the leakage current of a single cell on chip with practical lab bench equipment, given that it's leakage current in the given technology is likely to be measured in the picoamp range. It is even more difficult to measure transition times of individual cells. Moreover, given that RDF (which is purely stochastic) dominates device variation, the measured characteristics of a single device might not be representative of the average characteristics. Ring oscillators solve these issues through averaging whilst maintaining the correlated effect of the INWE [90]. Figure 48 shows the gate schematics.

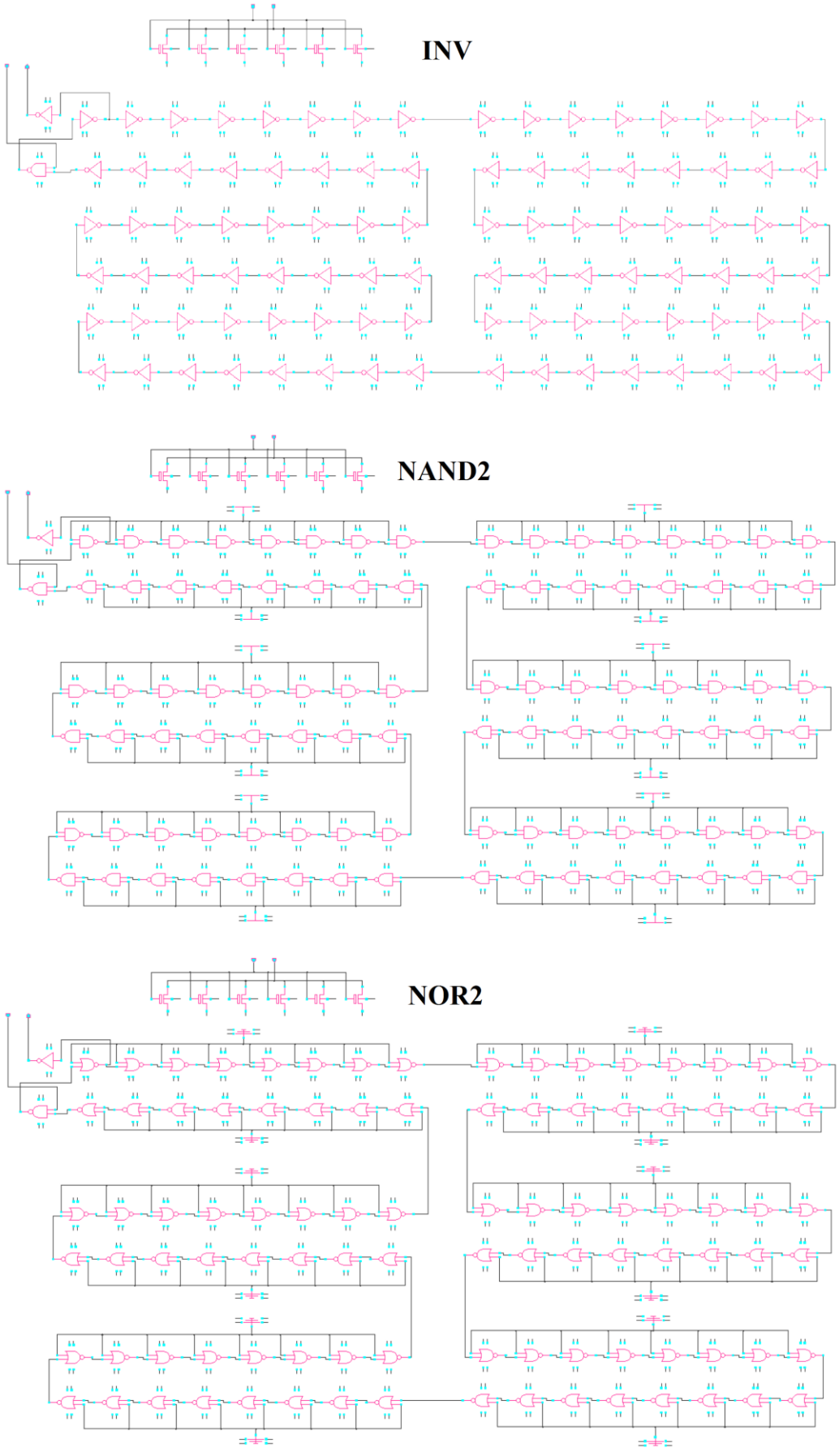


Figure 48: Ring oscillator schematics

Each ring oscillator consisted of 96 cells of identical gate types and a NAND2 gate for the oscillator enable. All 97 gates were the same number of full diffusion fingers. Repetition of the same gate increases the leakage current to a practically measurable value and helps to average out the random nature of the variation. The same principle also applies to the propagation delays that cumulatively average to form the frequency of oscillation. A 4F inverter was used to disconnect any meaningful capacitance between the oscillator and the on chip frequency measurement circuitry (connected via a level shifter). Six Thick Gate Oxide (TGO) footers were used for power gating each oscillator from the power rails to allow multiple oscillators to be connected to the same power pin. For the NAND2 and NOR2 gates, the transition of interest was the slowest transition as determined in the previous subsection. The unwanted pins (pin A) were therefore tied off to VDD and GND respectively to sensitize the B-to-Y transition. Figure 49 shows a sample oscillator layout.

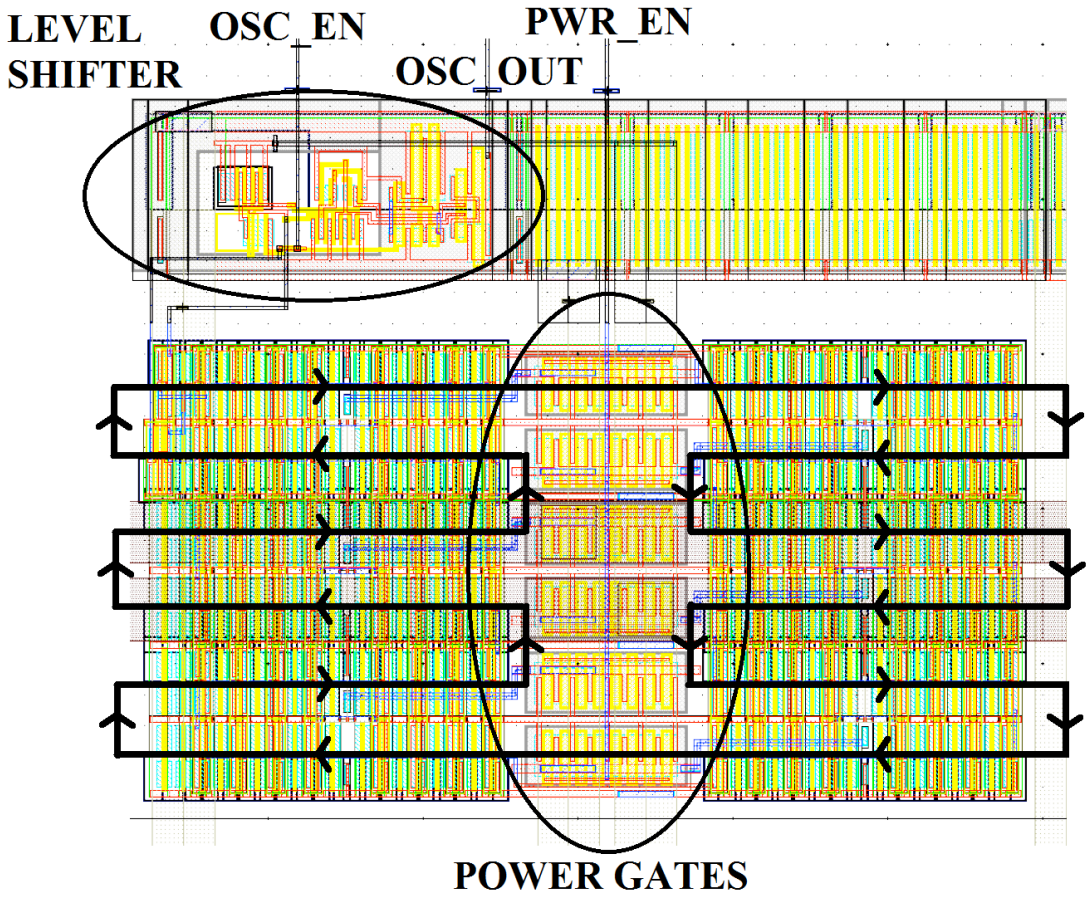


Figure 49: Ring oscillator sample layout

Each oscillator was manually laid out by first creating a power rail template placing the six TGO power gates in the center of the rails. The gates under test were then populated by hand with each gate under test interspersed with a filler cell of the same width. This gave a realistic area utilization of around 50% and also created realistic wire delays whilst introducing a level of determinism in the overall RC wire delay that would otherwise be lost in automated place and route. Tie cells were placed in the center of each row to provide the necessary VDD/GND tie voltage for the unused pins without connecting them directly to the rails. This prevents transients caused by large switching / ElectroStatic Discharge (ESD) events damaging the gate oxide of the devices tied off. The output of the oscillator sampling inverter was then routed through a level shifter to provide adequate voltage levels for the on chip frequency counter. The power rails of this structure were then decoupled using fill cap cells.

The gate types included were the 1F/2F/3F/4F full diffusion finger variants of cells INV1X, NAND2 and NOR2. All gates were also created in both LVT and RVT variants, to give a total of 24 individual oscillators. The power rails of these oscillators were routed together to a single power pin on a larger chip design. On chip control logic was used to multiplex the oscillator and power enable input pins of each oscillator to control registers and to multiplex the outputs of each oscillator to the on chip frequency counter. Calibre was used to DRC and LVS check the entire design and the design was submitted to TSMC for fabrication. Calibre was also used to create parasitically extracted models of each of the 24 oscillators so that the model and silicon could be accurately compared.

4.8 First Oscillator Test Methodology

Upon successful reception of the test chips from TSMC, the oscillators were tested on two separate occasions using two different methodologies. This subsection describes the first test methodology. Test chips were placed into a bespoke test board. The board contained two separate USB connections; one for power and one for serial communication via python programming to an mbed processor contained on the board. This processor then drove the on chip logic in the test chip. The board also contained several additional peripherals, including an on board temperature sensor situated around 5cm from the test chip and an on chip voltage regulator capable of remotely being programmed. Finally a configurable on chip ADC was able to directly measure the voltage on the oscillator rails. Figure 50 shows a flowchart representation of the python program used to sample and read the oscillator frequencies.

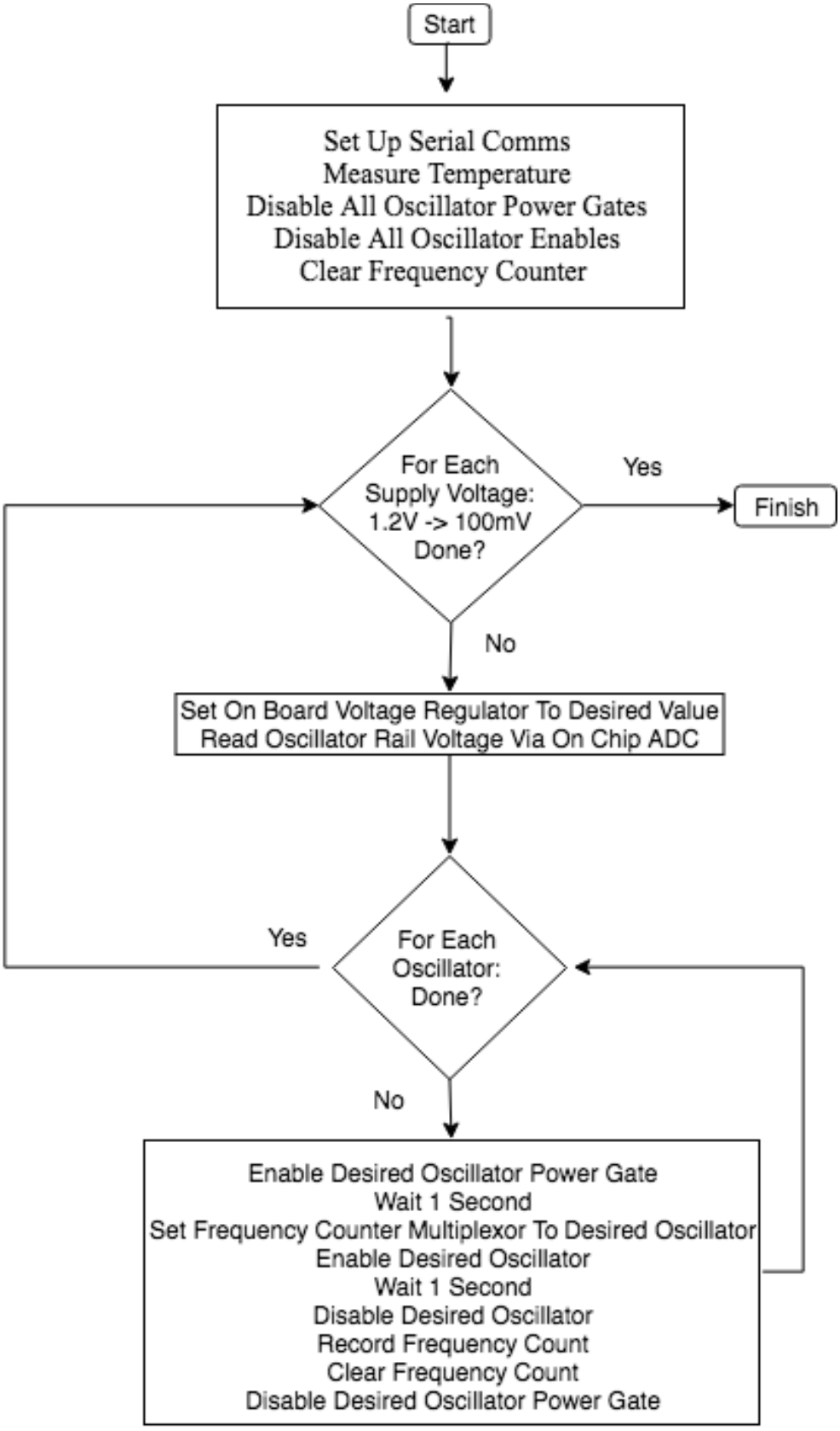


Figure 50: First ring oscillator test methodology

The temperature was set using an external thermal lamp. The proximity of the lamp was fixed, and the temperature perpetually read using the on board temperature chip. This configuration was consistently adjusted until the temperature settled to close to a temperature of interest and long enough to perform a reading without temperature change. The program was then executed.

Initially the program set up communication with the desktop lab machine then read and recorded the on board temperature sensor. The program then disabled the power gate and oscillators enable lines to all oscillators and cleared the frequency counter register. This initialized the board for the test. It then cycled down through the supply voltages, starting from 1.2V and finishing at 100mV in 100mV steps. The supply rail voltage was set by programming the on board programmable voltage regulator. To account for any on board loss of voltage, the on chip ADC sampled the actual on chip power rail and returned the voltage. This was the supply voltage recorded for the results. The program then cycled through each oscillator, powering up the oscillator by turning on its power gates, waiting a second for the voltage levels to normalize, enabling the oscillator, waiting a second, disabling the oscillator and recording the value of the frequency counter. The frequency counter was then cleared and the same measurement performed for the next oscillator. The temperature points successfully recorded during the experiment were 23.8, 33.1, 44.3, 53 and 63.5 °C. Figure 51 shows the LVT results for a nominal temperature of 23.8 °C.

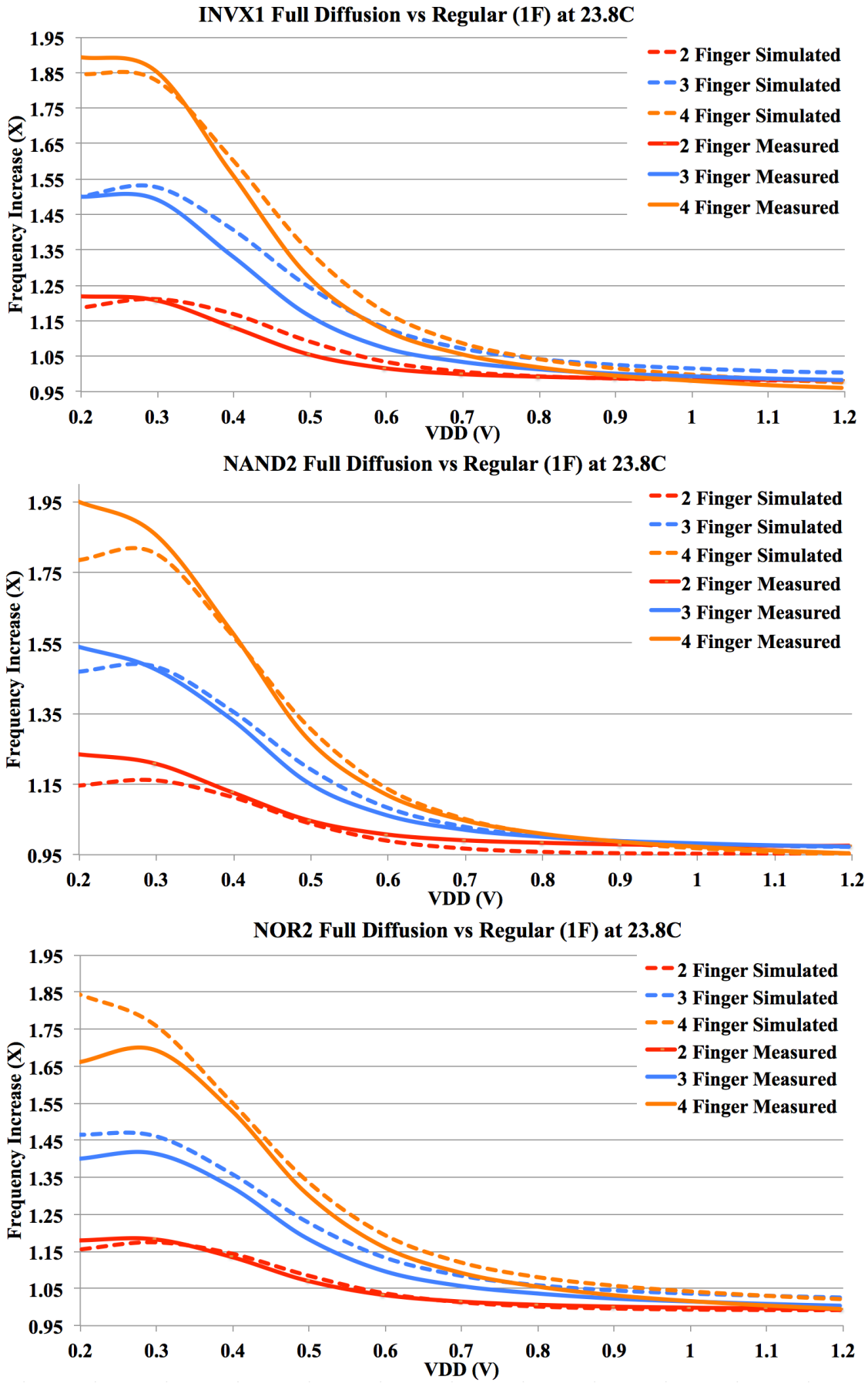


Figure 51: Ring oscillator comparative frequency increases: LVT

Figure 51 shows the comparative frequency increase for the 2F/3F/4F LVT full diffusion cells over the single finger/regular designed cell. HSpice simulations of the full, parasitically extracted oscillator are provided along side the measured results.
The results for the INV oscillators showed excellent overall matching between measured results and the simulations. The comparative reduction in propagation delay and henceforth increase in oscillation frequency matched well to the BSIM models. This is true for the full supply voltage range and for all fingers and henceforth, underlying device widths. This indicated that the INWE is modeled accurately in the compact models. All 2F/3F/4F finger variations showed an increase in frequency below 800mV. Optimal frequency improvements were observed below 300mV and constituted improvements of 1.23X/1.5X/1.9X for the 2/3/4 finger designs respectively. The contributions of current increase to capacitance reduction for these figures break down to approximately 94%/6%, 83%/17% and 86%/14% respectively.
This means that for the LVT inverter cell, the full diffusion sizing strategy was able to provide a substantial range of devices up to 1.9X faster than the regular subthreshold sizing methodology. Moreover, the worst recorded comparative measurement was 0.95X at nominal voltage, indicating a 5% performance degradation is the highest possible penalty experienced.

The results for the NAND2 gates also showed a close correlation between measured and simulated results until the 200mV reading, which showed slight deviation. Given that the inverter results shows close correlation, it is likely that the underlying modeling of the INWE was not the culprit behind this deviation. The models appear to slightly under estimate the improvement at deep subthreshold voltages, with measured results of 1.23X/1.54X/1.95X recorded for the 2F/3F/4F finger variations. This could be due to the models over-pessimistic evaluation of the stacking factor, non-typical chip selection, non-idealities of the tie cells holding tied off inputs at diminished voltages which effect transition values or even simple quantization errors of the on chip frequency counter. Even with the deviation considered, the full diffusion sizing strategy still provides a proven range of cell performance for NAND cells. As with the inverter cells, all NAND cells exhibited performance increase below 800mV and the maximum recorded performance penalty was 5%.

The results for the NOR cells also showed a close correlation between measured and simulated results down to the supply voltage of around 200mV, where a deviation was observed. In this instance the simulations over estimated the performance increase of the cells. Measured improvements of 1.19X/1.42X/1.71X were observed, peaking at the 300mV supply voltage point. The deviation below this point could be attributable to many factors, as outlined in the NAND discussion above. All cells show performance increase below 900mV and the maximum recorded performance penalty was 2% at nominal voltage. Figure 52 shows the results for the RVT gates.

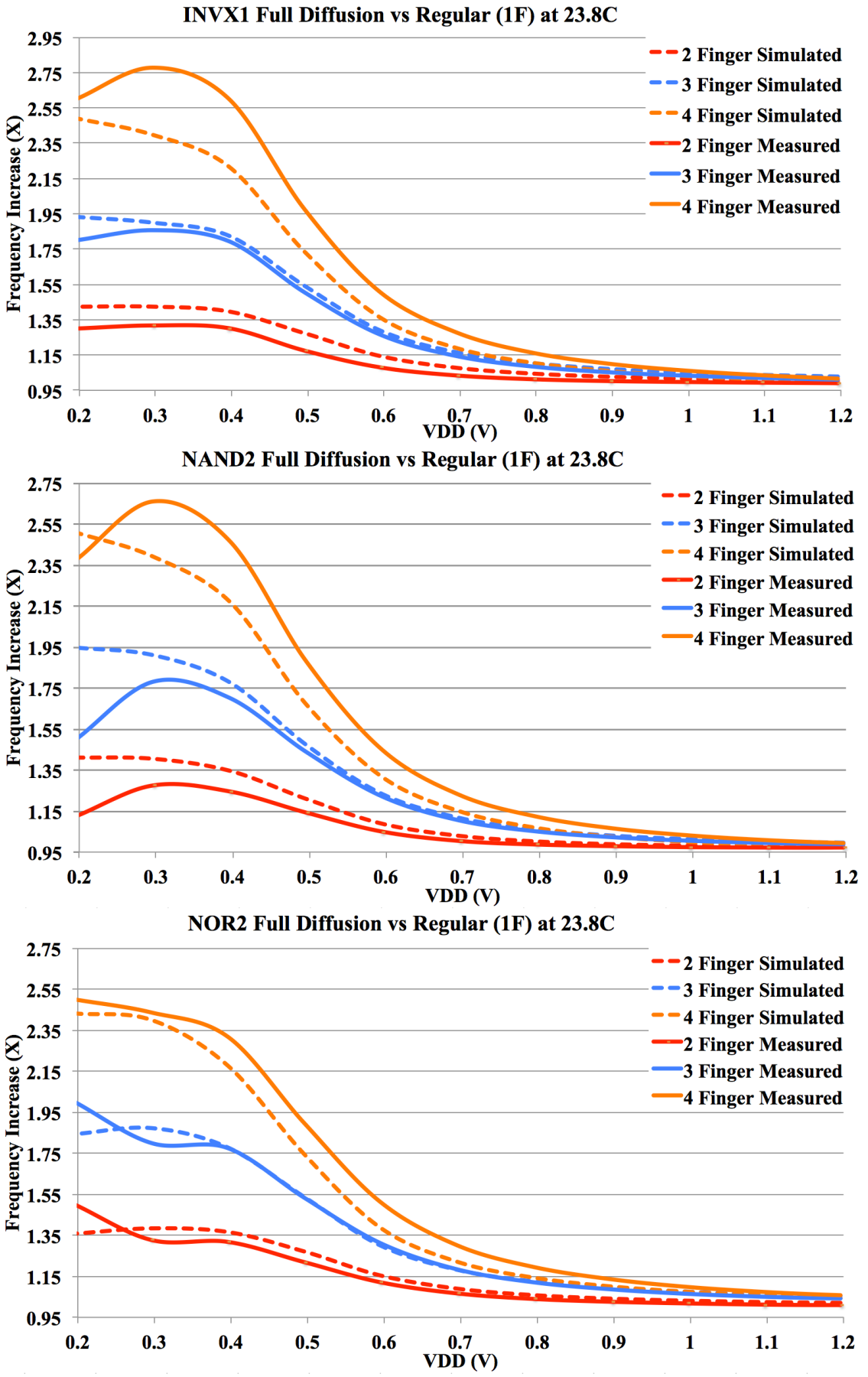


Figure 52: Ring oscillator comparative frequency increases: RVT

Figure 52 shows the measured and simulated comparative frequency results for the RVT oscillators. The inverter showed greater gains in the absolute values in comparison to the LVT variants. Performance increases of up to 1.33X/1.9X/2.77X were observed for the 2F/3F/4F finger variations, indicating a greater range of performance is created by the full diffusion methodology for the RVT inverter gates. The contributions of current increase to capacitance reduction for these figures break down to approximately 95%/5%, 87%/13% and 89%/11% respectively.

There is noticeable deviation between the BSIM models and measured chip results across the full voltage range. The model over estimates the increase for the 2F/3F gates and under estimates the increase for the 4F gate. This could indicate that the model incorrectly models INWE over different device widths. It could also be accounted for by the simple global variation of a single sampled chip from the test run. All devices show improvement below 1V.

The NAND2 gates also showed greater gains in absolute values than the LVT variants with performance increases of up to 1.3X/1.82X/2.68X observed for the 2F/3F/4F finger variations. Interestingly the same pattern of simulation under/over estimation is observed as in the inverter results. As in the LVT variant, the maximum deviation from the measured results occurs at the 200mV point, indicating the same culprit is likely responsible. All variants showed improvement below 800mV and the maximum performance penalty observed was 4%.

Finally, the NOR2 gates also showed greater gains in absolute value than the LVT variants. Performance increases of up to 1.51X/2X/2.49X were observed. The NOR2 gates seem to closely match the simulated results. Given that the critical switching transition in the NOR gate is the stacked PMOS devices in the pull up network, this may indicate that the deviation observed in the inverter and NAND2 gates lies within the RVT NMOS devices, where these devices have a greater impact. Improvement in frequency was observed below 1.1V and the maximum performance penalty observed was 2% at nominal voltage.

Observing the results in their entirety, several important points must be addressed. The first is that, although the correlation between the simulations and device under test varied slightly from gate to gate and LVT to RVT, the hypothesis that the full diffusion methodology can be used to successfully invoke the INWE at device widths greater than minimum was successfully established. Moreover, the methodology can be used to create

the performance range in the basic gate set required to synthesize combinational logic. The gates always displayed a progressive and logical increase in performance as the number of fingers was increased across the entire supply voltage range of increased performance, indicating a deterministic level of controllability in the process of creating the performance range.

A second important point is the wide range of supply voltages that the full diffusion methodology gave an increase in performance. This is one of the subtler novel contributions of this work. The study of the INWE has been limited thus far in the field to subthreshold operation. It was therefore surprising that across all gates tested, an increase in performance is observed right up to 800mV. Moreover, the maximum penalty observed across all gates for operation at the nominal (1.2V) voltage was 5%. This means that this methodology is especially suited to ultra wide dynamic voltage scaling schemes, where typical operation is in the subthreshold regime but the supply voltage may be temporarily increased to give faster operation during periods of increased computational requirement. The deterministic nature of increasing performance as number of fingers increased, across the full supply voltage range, indicates that the same circuitry may be driven at different voltages without the introduction of timing errors by the disproportionate increase in propagation delay over supply voltage.
A third important point that must be addressed is the level of deviation between simulations and the actual measured results. It is important to express the results presented were derived from a single randomly selected chip from a run of 80. The ring oscillators were designed to average out local variation in the propagation delays that are induced in a stochastic nature by RDF. However, there is nothing to account for systematic interdie (global) variation such as gate oxide or other parametric gradients. Given that the technology node chosen is a decade old, is regularly used by other researchers and shown to consistently return silicon around the typical global process corner, there is a scale around which this assumption can be judged to be true. The simulations were performed with typical process corner models and did correlate well with the sample chip, indicating this likely to be the case. The practice of matching absolute values however must be discouraged due to the deterministic nature of the simulation models and the probabilistic nature of the testing. For deterministic variation testing, matching arrays rather than ring oscillators should be used [91] [92]. The purpose of the measurement was to determine whether the comparative gains from the INWE

based full diffusion sizing strategy displayed in the models accurately conveyed the behavior of the silicon. On this point, the testing can be viewed as successful.

The test methodology itself turned out to be flawed for two main reasons. During the design of the sample chip, it was assumed from gathered information that an oscillator of 96 gates would increase the leakage current to a level measurable by on chip circuitry. This assumption turned out to be incorrect. The resolution of the on chip sampler was simply too large to accurately measure the small amount of current produced by the leaking gates. As a back up, the power rails were given a dedicated pin on the chip so that the on board programmable voltage regulator could be abandoned in favor of directly powering the rails from a bench power supply. This provided the opportunity to measure the leakage on the rails directly. However, the amount of current leaking through an individual oscillator again was simply too small for a standard 6½ digit multimeter to measure. In designing the control logic for the power gates and oscillator enables, it had been foreseen that states should exist in the control register to control the oscillator macro such that one oscillator was on and all other were off. However, this functionality was not one-hot encoded. This meant that it was also impossible to power all oscillators on, take a base line reading, turn off the oscillator in question and then subtract the measured current reading against the baseline and determine the leakage current. Using this particular methodology, the author was resigned to accepting that the leakage current could not be measured.
The second issue with this methodology was the accurate control and measurement of temperature. Whilst on board temperature sensing was provided, the sensor was 5cm from the test chip and the heat source used was a heat lamp. As the temperature measured was not that of the ambient temperature surrounding the board and the temperature sensor was some distance from the test chip, the level of accuracy of the temperature measurement could only be asserted as +/- 3 °C. Moreover, the temperatures finally accepted were those consistently stable within the limitations of the set up and were therefore aperiodic (23.8, 33.1, 44.3, 53 and 63.5 °C).

To address the two issues outlined above, a second testing methodology was devised.

4.9 Second Oscillator Test Methodology

To overcome the limitations of the first test methodology, additional equipment was acquired. To enable successful measurement of the leakage characteristics of the oscillators, an Agilent B2911A source meter was rented. The on chip programmable voltage regulator was disconnected from the oscillator rails and the source meter used to directly provide the supply voltage via the available chip pin. As the oscillator macro was not the only circuitry powered by this pin, additional steps were required to accurately measure the leakage current of each oscillator. At each voltage step, all circuitry connected to the pin was turned off, a baseline reading was taken to determine the amount of current drawn by everything, then the power gates to the oscillator of interest were enabled. The difference of these two measurements was calculated as the oscillator leakage current.

Whilst this solves the leakage current issue of the previous methodology, the on chip ADC required to measure the supply rails directly on chip assumes usage of the on board programmable voltage regulator. This means that a degree of inaccuracy is introduced into the results as any voltage dropped between the current source and rail cannot be measured. The results from the previous methodology suggested drops of up to 5mV were experienced. Whilst this is insignificant at full voltage, it accounts for a 2.5% inaccuracy at 200mV, which translates into an indeterminate level of discrepancy in propagation delay and leakage current. For this reason, the measured results presented in this methodology were not simulated in the BSIM models. The results are therefore presented as is and comparison made to the relative performance increases of previous subsections.

To enable a higher level of controllability and accuracy over the temperature testing, a Cincinnati Sub Zero EZT430i temperature chamber was procured. This was able to sweep from 0 to 100 °C, accurately maintaining the ambient temperature to within 0.1 degree. The on board temperature sensor was then used to indicate the successful adoption of the board to the ambient temperature after a sufficient period of time had elapsed.

The python control program was modified to issue Virtual Instrument Software Architecture (VISA) commands to the source meter and temperature chamber to automate the process of setting voltage and temperature values and taking and recording measurements. Figure 53 shows the modified program in flowchart format.

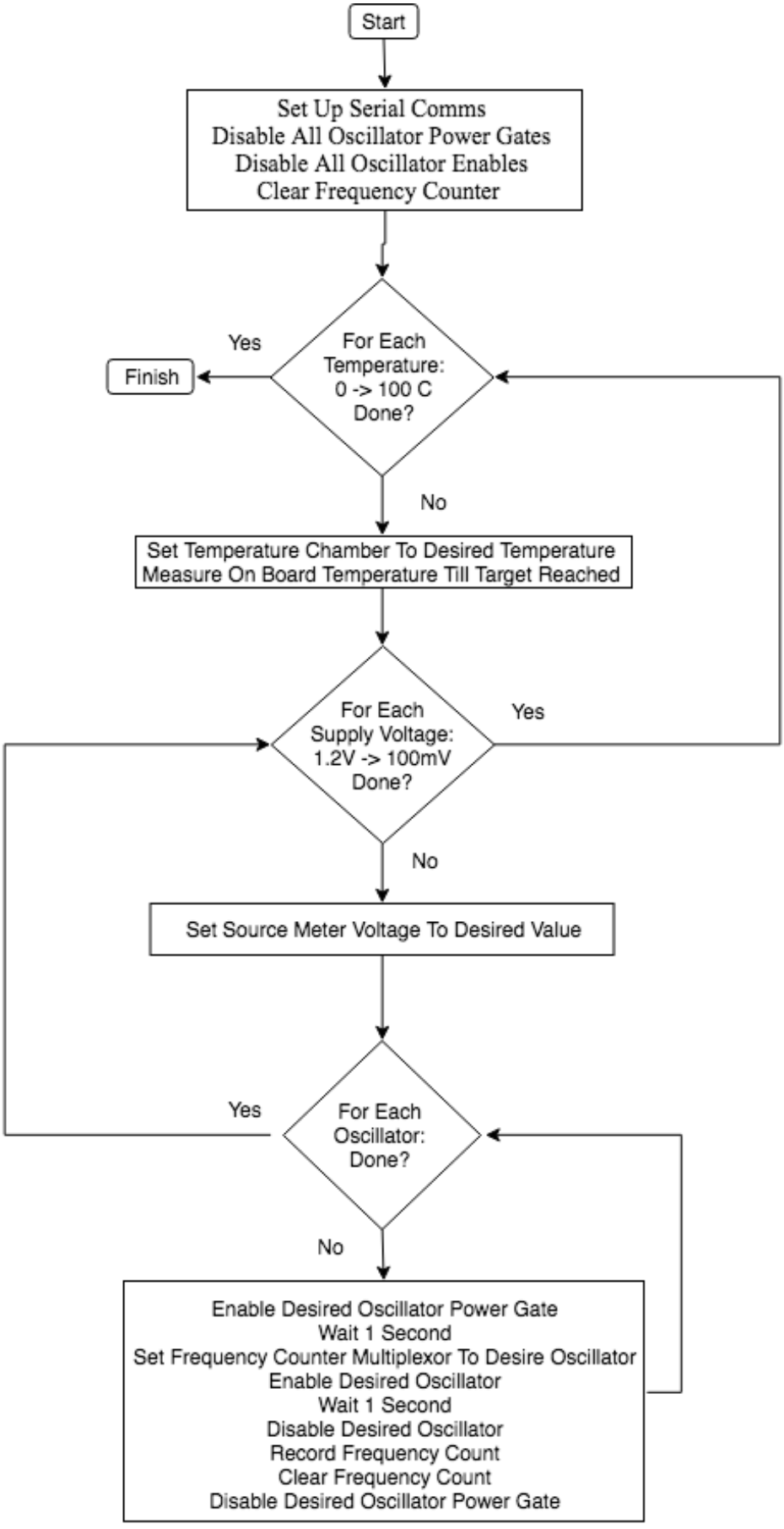


Figure 53: Second ring oscillator test methodology

All oscillators were powered down and disabled and the frequency counter cleared. The temperature chamber was then instructed via VISA to set a particular temperature in the chamber. Due to the exponential nature of temperature adjustment, an over shoot algorithm was devised to set the temperature 5 degrees beyond the required temperature. The on board temperate sensor was then periodically sampled. As the board temperature neared the desired temperature, the temperature chamber was instructed to remove the 5 degree overshoot and meet the board at the desired temperature. This proved highly successful in shortening the duration of time required to change the board temperature between points. The board was then polled periodically. After a fixed number of periods returned the desired temperature, the program proceeded to the voltage step. The first temperature measured was 0 °C and this was swept at 10 degree intervals to the final measurement of 100 °C.
For each temperature setting, the source meter was instructed via VISA to set a particular voltage. Each oscillator was then measured in a method similar to the first test methodology. The voltage was then swept from 1.2V to 100mV in 100mV steps, performing the oscillator measurements at each supply voltage until all measurements had been performed.

Current measurement was provided by an additional function in the program. Figure 54 shows the procedure in flowchart format. To save time and to provide more accurate results, the current measurement was performed during the same temperature and voltage runs as the oscillator frequencies. The additional functionality disables all oscillator power gates and takes a baseline current reading. The oscillators are then systematically enabled, the current measured and disabled. The difference calculated between the baseline and oscillator measurement is recorded as the leakage current.

Figure 55 shows the results for the LVT variants.

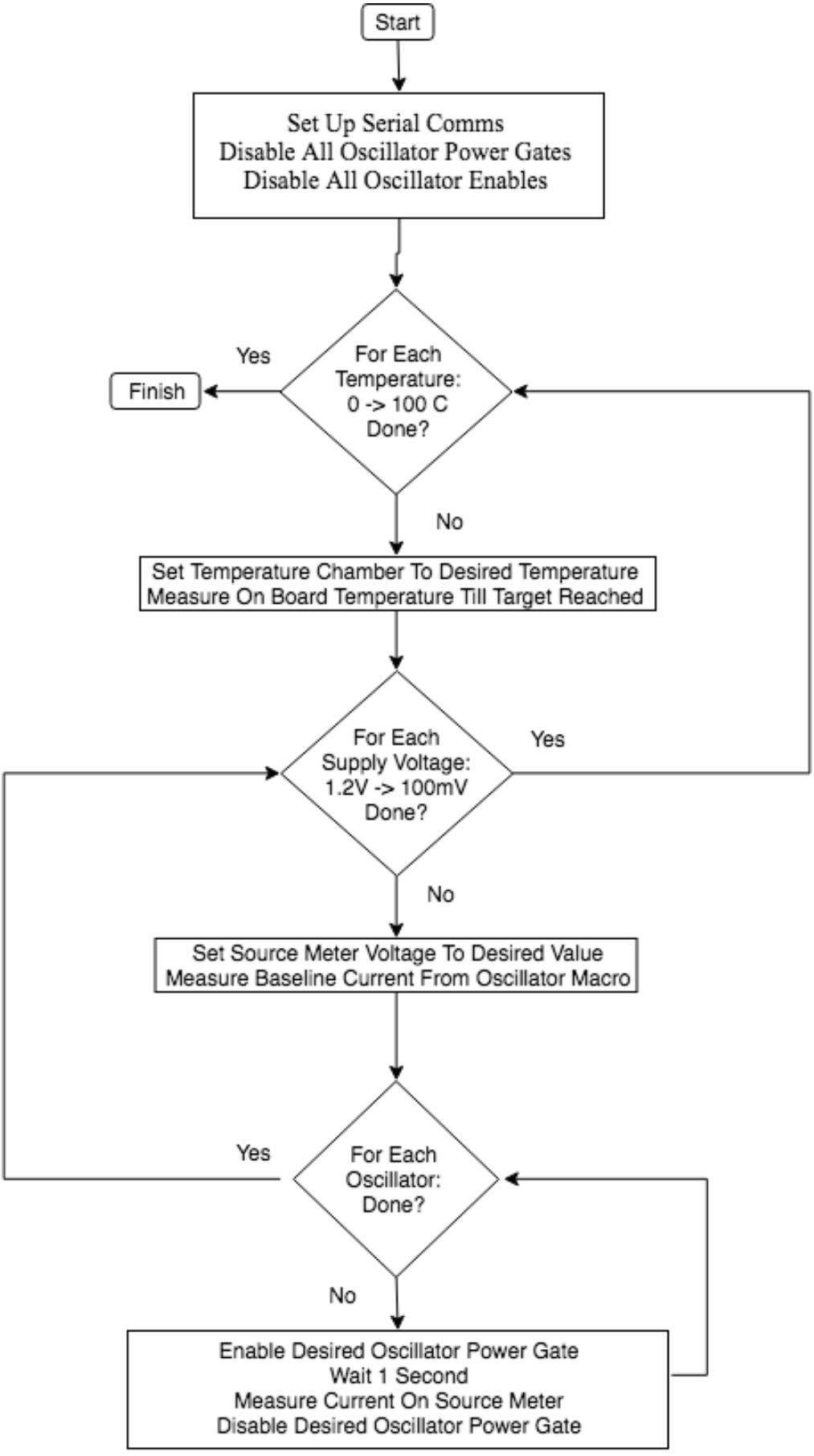


Figure 54: Ring oscillator leakage current test methodology

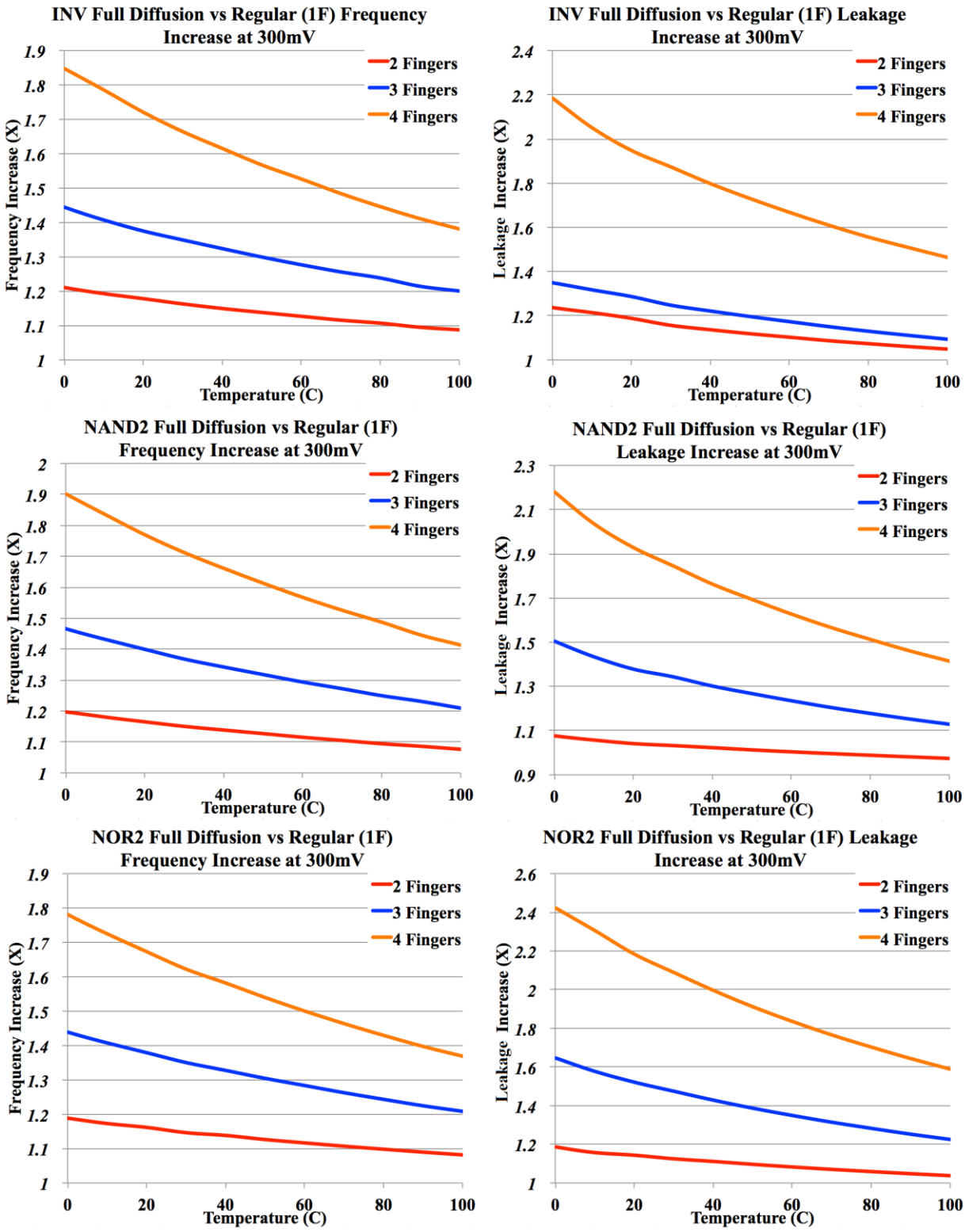


Figure 55: Ring oscillator temperature analysis: LVT

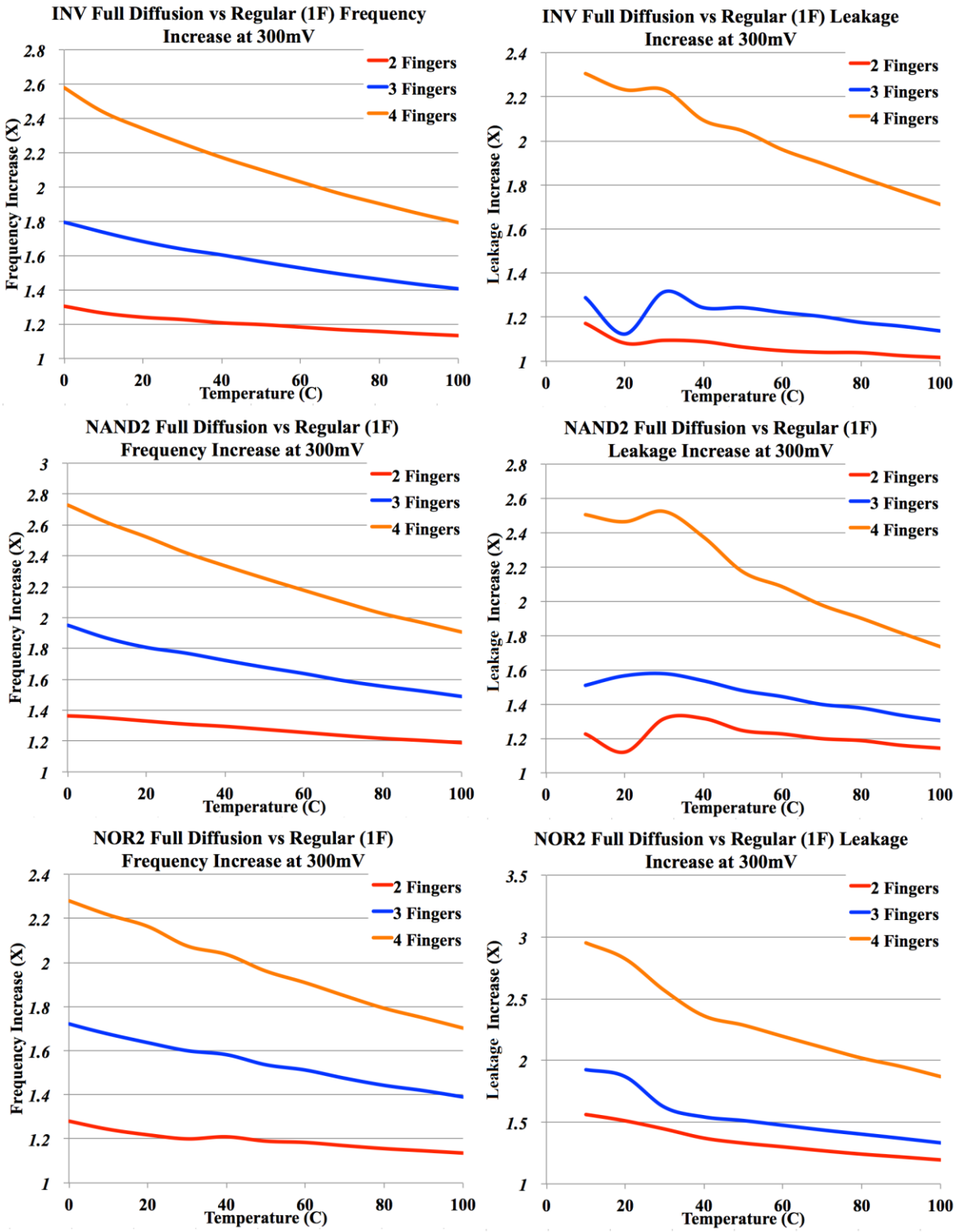


Figure 56: Ring oscillator temperature analysis: RVT

Figure 55 shows the frequency and leakage increase of the LVT gate variants at 300mV swept over temperature. The inverter frequency improvement showed that the consistency in the comparative frequency increase demonstrated in the previous subsection over supply voltage was also true over temperature. As the number of fingers increased, the impact of the INWE effect also increased and the performance improved. The inverter leakage results showed that the leakage current also increased in response to an increase in the number of fingers and therefore the INWE. Both 2F and 4F inverter devices showed a comparable increase in leakage greater than that observed in frequency. Interestingly, the 3F device appeared to have a frequency increase greater than the corresponding increase in leakage current. Therefore for this gate, the technique applied actually improved the performance to leakage ratio.

The NAND2 gate showed similar trends to the inverter gate. However, on this gate, the 2F device experienced the performance to leakage improvement and the 3F and 4F devices showed performance to leakage degradation. The NOR2 gate shows similar results.

Figure 56 shows the frequency and leakage increase of the RVT gate variants at 300mV swept over temperature. All three gates showed similar trends to the LVT devices. The absolute value of the frequency improvements were greater than the LVT values, corroborating the result from the first test methodology, indicating that the RVT devices are impacted greater by the INWE and therefore experience a greater comparative increase in performance leading to an increase in the performance range. There was more noticeable deviation in the frequency trends as the temperature approaches 0 °C. This is clearly noticeable in the current measurements.

In contrast to the LVT devices, the performance to leakage ratios of the RVT inverter and NAND2 gates generally improved in response to an increase in fingers. The same metric in the NOR2 degraded uniformly. Conspicuous by their absence are the 0 °C leakage figures. Whilst data for this point was obtained, it strayed beyond the range of feasibility and was therefore omitted as likely erroneous.

There are several important points to address from the results presented. Another subtle novelty of this contribution was the examination of the INWE over temperature. Consistent across all gate types and both LVT and RVT variants, the INWE fillip is greater at lower temperatures and diminishes as the temperature increases. This is due to the temperature dependence of the threshold voltage outlined in Section 2.5.7. As the temperature increased, the additional kick in provided by the INWE was superseded by

the drop in threshold voltage (known as temperature inversion) and the comparative increases in performance and leakage current diminished. This is actually advantageous. Due to temperature inversion, subthreshold logic speeds up as temperature increases, and slows down as temperature decreases. The INWE based full diffusion sizing strategy could therefore provide an additional level of stability in the subthreshold regime. As the circuit cools, the additional performance increase provided offsets the performance lost by the increase in the natural threshold voltage of the devices. This should extend the temperature range of operation at lower temperatures.

Another point of importance is the variation in functionality at lower temperatures is different between the LVT and RVT variants. The frequency results showed that no oscillator failed during the entire test methodology. However, the results suggest that the LVT devices coped better at lower temperatures than the RVT devices. In reality, this may or may no be true. Due to temperature inversion, the absolute current values decrease as temperature decreases. As the absolute current values of the LVT devices are larger, this would suggest a point of failure lower than the RVT devices. However, it may also be that due to the lower currents of the RVT devices, they were simply too small to be reliably measured by the current source.

4.10 Combinational Logic Library

After confirmation of the results in silicon, it was important to prove that the advantages displayed in the individual gates extend to synthesized circuitry when using an industry standard workflow. The first step in this process was creating a library of cells capable of synthesizing combinational logic. The 4 simple gates types already laid out earlier in the chapter were sufficient for this purpose. Commercial synthesis tools are exceptionally adept at constructing alternate functions from this basic set of gates. Buffers can be constructed from back-to-back INV/NAND2/NOR2 gates, and NAND2/NOR2 are universal gates types with the ability to construct other basic functions such as XOR/XNOR. AOI22 is convenient for multiplexors. Table 2 lists the cells in the library.

Gate Type	Finger Variants
INV	1/2/3/4
NAND2	1/2/3/4
NOR2	1/2/3/4
AOI22	1/2/3

Table 2: Combinational library cells

Schematics for all gate types were created in Cadence Schematics XL and then laid out in Cadence Layout XL. Mentor Graphics Calibre was used to perform DRC and LVS verification. The gates were then ready for characterization.

4.11 Combinational Library Characterization

Characterization is the process of running SPICE level simulation on cells to provide functional and delay information that can be used in commercial digital design workflows. This usually takes the form of compiled liberty tables, from which, digital synthesis tools are able to estimate propagation delay for each cell given load capacitance and slew rates. Static timing analysis may then determine path delays and feed this into clock tree synthesis. The exact process is described fully in the next chapter for a full subthreshold library.

For the small combinational library, a basic 'catch-all' TCL workflow was provided by ARM called CellBuilder. This primarily uses the Synopsys toolchain. Cadence Virtuoso is used to generate the GDSII and CDL (Circuit Design Language) files. Synopsys Hercules is then used for DRC and LVS checks. Mentor Graphics Calibre performs a secondary physical validation check. The output of the Hercules LVS checks are then used to create parasitically extracted SPICE models by Synopsys StarRCXT. These models are then passed into Synopsys SiliconSmart, which takes a 'catch-all' fall/rise time set and capacitance range and produces Liberty Tables for given PVT values. These are then compiled to provide the library inputs for the digital synthesis workflow. Table 3 lists the PVT corners created. Targeted supply voltages follow ARM's standard methodology of +/- 5% for the Typical corner for the Fast/Slow corners respectively.

Process Corner	Voltage (mV)	Temperature
SS	142.5	25°C
TT	**150**	**25°C**
FF	157.5	25°C
SS	190	25°C
TT	**200**	**25°C**
FF	210	25°C
SS	237.5	25°C
TT	**250**	**25°C**
FF	262.5	25°C
SS	285	25°C
TT	**300**	**25°C**
FF	315	25°C

Table 3: Combinational library characterization corners

To provide a fair comparison, the same cells listed in the previous subsection were characterized from ARM's Low Energy cell library using the identical workflow above.

4.12 Multiplier Synthesis

Synthesis was performed using Synopsys Design Compiler. A simple TCL workflow was used to synthesize 32 bit multipliers from each library at 4 different PVT corners. Both LVT and RVT variants were synthesized. The tool was provided as a simple Verilog behavioral description. This allowed the synthesis tool free will to chose the most efficient multiplier design available to meet the timing constraints specified. As the tool is not designed to iteratively search for the highest frequency of operation possible, this was performed manually by relaxing the timing constraint and synthesizing. The path delays were then checked to determine if any failed to meet the timing constraints specified. If all paths met the timing criterion, the path delay was decreased and the synthesis rerun. This iterative process continued until one or more paths failed. The synthesis run with the lowest time constraint and all paths passing was then determined to be the critical path delay time. The easiest way to report leakage power in such a succinct synthesis workflow is to include a leakage power corner. All synthesis runs were signed off at the typical corner as the geometric sizing for the full diffusion library was optimized for the typical corner physical characteristics. Results for the critical path delays and leakage power reports are shown in tables 4 and 5 respectively.

Critical Path Delay (ns)			
RVT	**Full Diffusion**	**LE Library**	**% Decrease**
300 mV	7634	13907	182.17
250 mV	26170	48655	185.92
200 mV	82813	213480	257.79
150 mV	1741070	3302650	189.69
LVT			
300 mV	880	1485	168.75
250 mV	2720	4785	175.92
200 mV	8856	16020	180.89
150 mV	27410	50035	182.54

Table 4: 32 Bit multiplier critical path delays

Leakage Power (nW)			
RVT	**Full Diffusion**	**LE Library**	**% Increase**
300mv	25.02	13.04	191.87
250mv	20.12	9.781	205.70
200mv	14.43	7.464	193.33
150mv	9.798	5.029	194.83
LVT			
300mv	493.9	196.9	250.84
250mv	406.6	152	267.50
200mv	291.7	112.4	259.52
150mv	202.9	79.55	255.06

Table 5: 32 bit multiplier leakage powers

The results from Table 4 show a marked improvement in critical path delay across all corners synthesized. The results from Table 5 show a comparable increase in leakage power. The RVT corners showed a proportional increase in leakage power with the 300/250/150 mV corners showing a slightly higher increase in leakage current than delay improvement. The 200 mV RVT corner stands alone in all corners tested to show a delay improvement greater than the increase in leakage power. Whilst all multipliers synthesized by the tool were Wallace Tree Multipliers, it is likely at this corner that the synthesis tool found a more efficient permutation of generating the 32 bit multiplier than it did in the other corners. The LVT corners showed a comparatively larger increase in

leakage power than they did improvement in critical path delay. This shows that the underlying physical characteristics of the gates propagate into the circuits they create.
The intent of this experiment was to determine if the full diffusion library cells could be successfully used in a digital synthesis flow and how the tool used throughout the process would react. As suspected, Design Compiler accepted the library without issue. There was however limitations in the design flow. Due to the simplistic nature of the TCL workflow, the only manner to derive leakage power data for the synthesized circuit without altering the underlying results was to include a separate corner in an MCMM (Multi-Corner Multi-Mode) design. The limitation of this is that leakage recovery cannot be performed on the typical sign off corner. This means that no leakage current optimization was performed and the synthesis workflow can only report the leakage power of the typical sign off corner. This corner has sole focus on meeting the timing constraint. Investigation therefore showed that it consistently favored the gates with the highest number of fingers (4F for INV/NAND2/NOR2 3F for AOI22), which are the fastest but leakiest gates. The leakage results presented are therefore not representative of a full digital synthesis workflow that would exchange fast leaky gates for slower less leaky gates once the timing constraint was met, but are representative of a worst case leakage scenario and the minimum width sizing strategy. A by-product of this issue was the discovery that the synthesis tool was smart enough to discriminate between the range in performance and leakage characteristics provided by the full diffusion sizing methodology for gates of the same type.
The synthesis methodology was also extremely limited in it's execution. As no sequential elements were provided, the path delays had to be reported instead of the clock period that would typically be used in a full synthesis workflow. Moreover, as only the circuit synthesis step was performed, the delays are actually location aware estimates provided by Design Compiler. In a full synthesis flow, floor planning, cell placement, clock tree synthesis, routing and static timing analysis would be required to provide accurate critical path timing.
To address the aforementioned issues, a full practical subthreshold cell library complete with sequential elements and clock gates was required to synthesize a larger circuit using a complete digital synthesis workflow. This is the focus of the following chapter.

4.13 Chapter Summary

This chapter explored the design space for the chosen technology node by designing X1 strength inverters from three distinct design strategies; two explored within the field and one novel. Post layout simulation was then performed to determine FO4 average propagation delay and leakage current. The advantages and disadvantages of each strategy were then discussed.

Monte Carlo simulation was then performed to determine parametric variation of average propagation delay for the Full Diffusion sizing strategy. The result showed that the LVT devices were inherently less variable than the RVT devices. The least variable device was not the same for both VT variations; therefore the Full Diffusion sizing strategy was the only sizing strategy to always contain the least variable cell.

Ring oscillators of basic combinational logic gates were then manually laid out, taped out and measured. The results showed frequency improvement up to 800 mV for all gates and an additional stability mechanism at low temperature as a result of the INWE.

Finally 32 bit multipliers were synthesized using a simplified digital synthesis flow. Results showed that the performance improvements measured in the ring oscillators extended into synthesized digital circuits.

Chapter 5: AES Core Synthesis Methodologies

5.1 Chapter Outline

The chapter proceeds by revisiting the optimal length discussion in Chapter 3. The sequential elements and multiple strength cells required for a practical library are then discussed. The characterization process is outlined and the synthesis of 32 bit AES cores described. These are then signed off for tape out.

5.2 Optimal Length Revisited

In order to accurately determine the accuracy of the BSIM model's expression for the INWE, the decision was taken to design the oscillators in the previous chapter with cells all designed with the same length (100nm). The discussion in Section 3.6 showed that the optimal length for propagation delay minimization is actually determined by the process corner targeted and the width of the device. Therefore, a library accurately reflecting the full capabilities of the full diffusion strategy would size the device length in each cell depending on how many fingers the cell has (determining the device widths) and the process corner. As the maximum number of fingers allowed for a 12 track library in the chosen technology node is 4, this is the number of width permutations that must be determined from the results of Section 3.6.

In a standard digital synthesis workflow, the slow corner is typically chosen as the signoff corner in order that the resultant circuits may be sold with a guarantee of meeting a fixed target frequency. However, this poses a problem as the silicon invariably received for the technology node chosen is around the typical process corner. The advantages and disadvantages of selecting the slow or typical corner to sign off in the digital synthesis flow is presented later in the chapter. As this is a designer decision, both libraries were created to provide the designer with the choice. Tables 6, 7, 8 and 9 present the optimal lengths as derived from the results of Section 3.6 for the SS LVT, SS RVT, TT LVT and TT RVT cells respectively.

SS LVT		
Fingers	**Width**	**Optimal Length**
1F	840nm	230nm
2F	355nm	230nm
3F	205nm	230nm
4F	120nm	230nm

Table 6: Optimal lengths: SS LVT

SS RVT		
Fingers	**Width**	**Optimal Length**
1F	840nm	90nm
2F	355nm	90nm
3F	205nm	190nm
4F	120nm	200nm

Table 7: Optimal lengths: SS RVT

TT LVT		
Fingers	**Width**	**Optimal Length**
1F	840nm	150nm
2F	355nm	150nm
3F	205nm	150nm
4F	120nm	150nm

Table 8: Optimal lengths: TT LVT

TT RVT		
Fingers	**Width**	**Optimal Length**
1F	840nm	90nm
2F	355nm	90nm
3F	205nm	100nm
4F	120nm	100nm

Table 9: Optimal lengths: TT RVT

Table 6 shows that the optimal length for minimum average propagation delay in the SS corner LVT cells is consistently 230nm across all device widths. This is consistent with the RSCE dominated response in Section 3.6.
Table 7 shows that this is far from the case for the SS corner RVT device. The width dependent RSCE dominates at lower widths but is superseded at higher widths by the SCE. Both the 1F/2F designs express minimum propagation delay at minimum length (60nm in chosen technology node) however, identical in the comparison (ARM LE Library), the minimum length is capped at 90nm to avoid undue variation. Beyond a width of 300nm, the RSCE begins to dominate and the optimal length increases, giving lengths of 190nm for the 3F devices and 200nm for the 4F device.
Table 8 shows that the TT LVT response is similar to the SS LVT response in that the RSCE consistently dominates across the full width of the device, giving identical optimal lengths across all finger permutations. However, the TT process corner reduces this value from 230nm to 150nm.
Finally Table 9 shows optimal length values similar to the SS RVT response. However this time the RSCE dominance is diminished, with its maximum influence at minimum width resulting in an optimal length of 100nm for the 3F/4F devices. Beyond 300nm, the SCE once again dominates and the optimal lengths tend towards minimum length. These are therefore capped at 90nm.

All cells presented thus far were redesigned such that they conformed to the length values presented in the preceding tables in order to optimize for minimum average propagation delay. A few cells were reduced in length, thereby negatively impacting the minimum operating voltage as discussed in Section 3.7. Given that the length reduction was 10nm at most, the impact was minimal. Conversely, most cell lengths were increased, some by as much as 2.3X. The discussion in Section 3.7 suggests the minimum operating voltage for these cells will improve. The resultant change in minimum operating voltage is therefore dependent on the choice of signoff corner, but more importantly, the cell chosen from the range by the synthesis tool.

5.3 Sequential Elements

Crucial to the concept of synchronous digital circuit design is the premise that periodically, for a designated period of the clock cycle, all signals are valid. To store the valid signals, sequential elements are required. These generally fall into one of two categories depending on how they transition from their opaque to transparent states, or put more simplistically, their sensitivity to the clock signal. Latches are level sensitive, whereas flip-flops are edge sensitive. The 32 bit AES target design required positive clock edge triggered sequential elements, therefore a flip-flop was chosen for the library. To provide an additional avenue of debugging, a scanable flip-flop was chosen so that the contents of paths providing consistent errors could be traced during testing. The target design also required asynchronous resetting.

The design chosen was taken from ARM's LE Library for fair comparison. The typical design for such a cell typically contains a 3-stack series in the feedback paths. This is prohibited in the subthreshold regime due to excessive degradation to the Ion/Ioff ratio. The cell chosen had already solved this issue by pushing one of the transistors from the 3-stack into the forward path. For the purposes of corporate privacy, the schematic for the design is not presented. However, the layouts of both the original cell from ARM's LE Library and the full diffusion variant are presented in Figure 57.

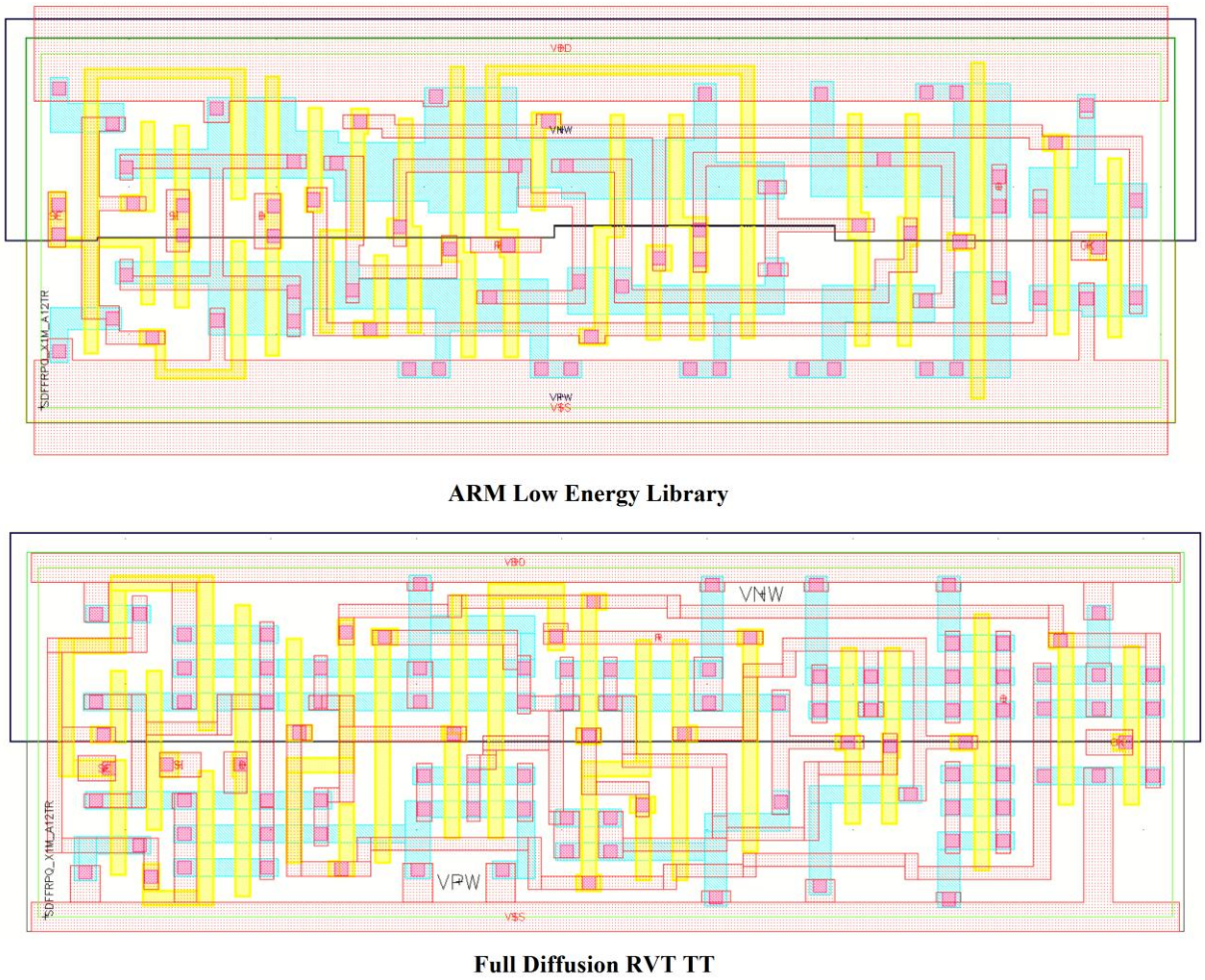
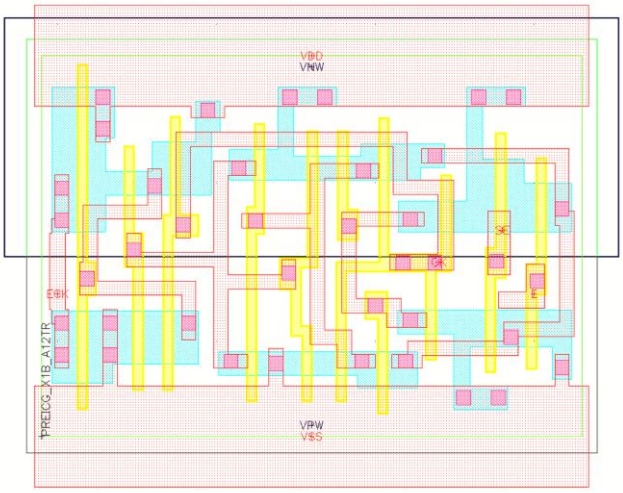


Figure 57: Full Diffusion scanable flip-flip layout

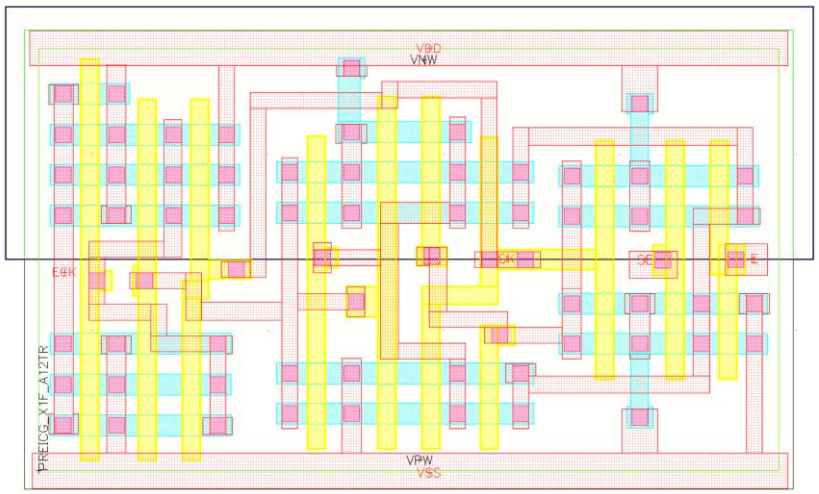
The original ARM LE cell was auto generated from a tool provided by ProlifIC for superthreshold operation. Nominal voltage operation requires larger power rails to provide adequate current carrying capability. For subthreshold operation, these rail sizes are not required and are therefore lowered for this cell and all other cells in the full diffusion library. This affords the greater number of finger permutations in the full diffusion sizing strategy, but would affect interconnect electromigration (Section 2.4.3.2.1) if used at higher voltages without due care. Narrowing the rails also provides additional routing space for dense via placement required by full diffusion sizing.
The flip-flop is designed to have gates of different strengths within a single cell. In the original design, the regular sizing strategy is adopted where increasing transistor widths provides greater strength. The discussion in Section 4.3 showed this assumption erroneous when operating in the subthreshold regime. The full diffusion sizing strategy creates different strength devices by using multiple fingers. To maintain the originally

intended strength differences of the design, the widths were quantized to the nearest number of fingers available. Due to the routing density of the cell, the maximum number of finger permutations permitted in the design was three. The routing density prohibited some nodes from being connected by interconnect on the metal layer. These were therefore connected on the polysilicon layer at the cost of the C spacer rule described in Section 4.7.

In order to reduce dynamic energy consumption, the synthesis methodology allows for clock gating. This is a commonly used technique in low power digital design. An integrated clock gate cell was therefore chosen from ARM's LE library and redesigned using the full diffusion sizing strategy. Figure 58 shows both cell layouts.



ARM LE Library



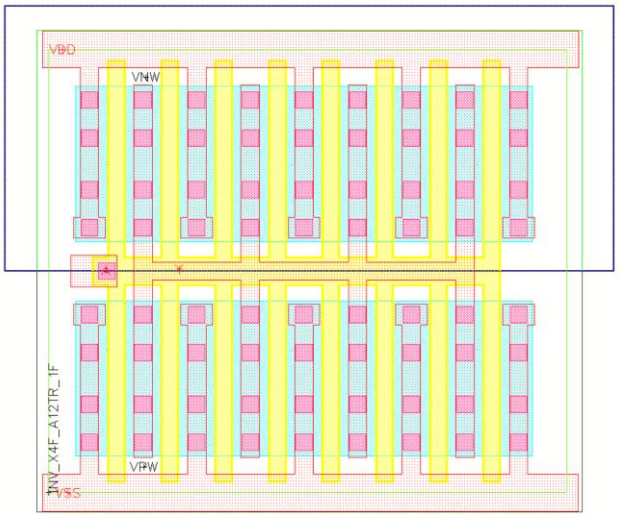
Full Diffusion RVT TT

Figure 58: Full Diffusion pre-integrated clock gate circuit

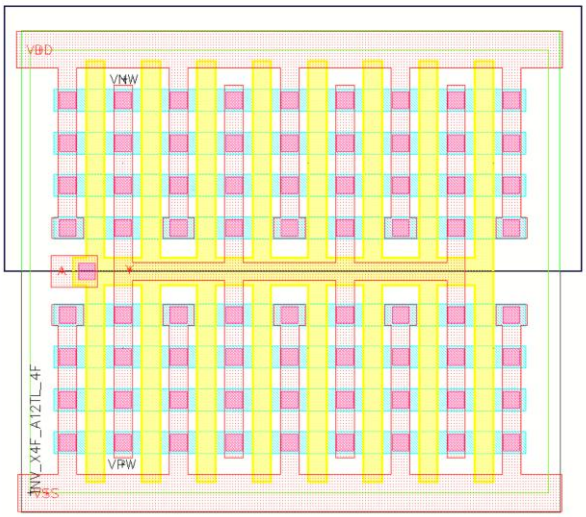
The rails were narrowed as in the flip-flop cell. The via density and finger quantization required additional spacing between devices on the polysilicon layer. The width of the full diffusion device is therefore greater than the original cell. However, this additional space and lack of horizontal routing requirements on the metal 1 layer meant devices of 4 fingers were permissible in the design. The four length permutations required for the LVT/RVT SS/TT libraries were created for both sequential elements. Mentor Graphics Calibre was used to perform DRC and LVS validation to ensure manufacturability in the chosen technology node.

5.4 Additional Strength Cells (X2/X4/X8)

To provide adequate drive current to switch long interconnects or high fan out logic, cell libraries contain devices of variable strength. Basic combinational logic cells (i.e. NAND2/NOR2) etc. are usually provided with a strength range. The strength increase is typically provided in the form of a weaker combination cell (i.e. X1) followed by a stronger inverter or buffer to provide current drive. Other than a minor improvement in silicon area utilization, there is little advantage to the additional design effort of creating these cells as compared to simply creating the strength variant inverters and buffers and allowing the digital synthesis tool to correctly match these when the need arises. Inverters of X2/X4/X8 strength were created by strapping multiple instances of the X1 inverter together, interleaving the horizontal orientation of the initial design to share source and drain diffusion contacts. As well as saving silicon area, this also has the effect of sharing junction capacitances, lowing the overall capacitance of the cell to less than the multiple of the constituent parts. The finger permutations and length optimizations were observed for all four full diffusion libraries. Figure 59 shows X8 inverter designs for the RVT TT 1F and 4F permutations.



Full Diffusion RVT TT 1F



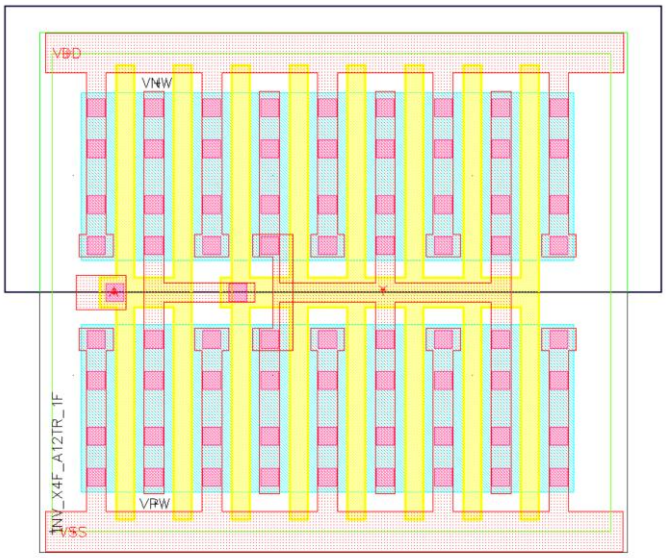
Full Diffusion RVT TT 4F

Figure 59: Four finger X8 inverter layouts

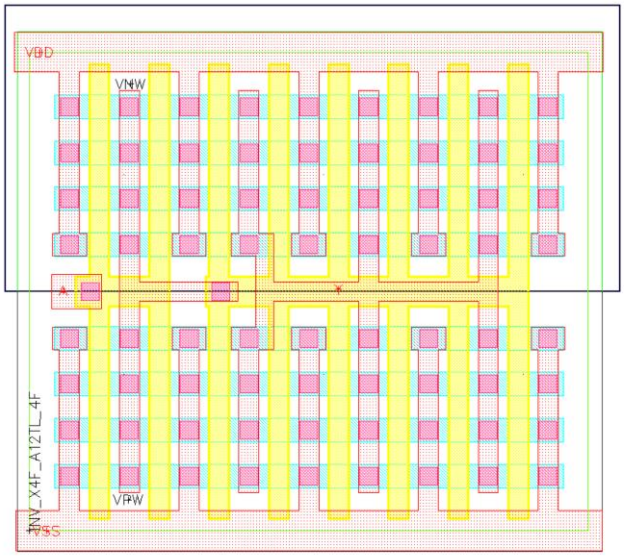
Whilst the via density is extremely high, the regularity of the design means that no additional consideration is required in the layout and therefore there is no penalty in silicon area for using the full diffusion sizing strategy.

Buffer cells were also created. As well as providing non-inverting drive current increases as previously outlined, they are also imperative for adequate clock tree synthesis. Buffers are simply designed using two stages of inversion in the same cell. A typical buffer configuration is to have one quarter of the strength of the cell invert the signal on the input and three quarters of the cell strength drive the output. As the cell internally fixes the fan out to three, this configuration provides adequate drive for the second stage and

better output drive than having two inversion stages of equal size. As this was the configuration chosen in the comparative library (ARM LE) this was matched in the full diffusion libraries. Strengths of X4/X8 were designed. Smaller strength variation is provided in the clock tree by the inverters. The same design principles were followed as outlined in the inverter design. Fig [60] shows X8 buffer designs for the RVT TT 1F and 4F permutations.



Full Diffusion RVT TT 1F



Full Diffusion RVT TT 4F

Figure 60: Four finger X8 buffer layouts

Similar to the inverter, the regularity of the design means that no additional silicon area is required for the full diffusion implementation.

5.5 Full Libraries

All combinational and sequential cells included within the libraries have now been presented in one variation or another. Additionally included in the combinational logic cells is the OAI22 cell (simply a vertical flipping of the AOI22 cell already presented). For a fully functional library, several ancillary cells are required. The first is a well tie. This cell simply ties the N wells and substrates to the VDD and GND rails respectively. As the cells are contiguous after place and route, the N wells and substrate areas are simply merged. The well ties are simply placed at intervals during synthesis by the synthesis tool between the synthesized logic. This interval can be provided to the tool to ensure the distance between the well tie and furthest cell is sufficiently small to eradicate any voltage drop lost due to the resistance of the semiconductor material (N well or substrate). Tie high and tie low cells are provided for the same reasons outlined the ring oscillator design.

Filler cells are required to allow the synthesis tool to assign all die area a cell affixed to the unit cell size. For the chosen technology, the unit cell size is 200nm. This means the widths of all cells in the library must be a multiple of this value. Sizes X1 (200nm) and X4 (800nm) are included in the library. The X1 cell contains no front end of line structures other than what is required to join the N well and rails of adjacent cells together. The X4 includes dummy patterning on the diffusion and polysilicon layers. If large areas are left void on these layers, the layers above have a tendency to bow into the void during processing, a phenomenon known as dishing [93]. The design rules therefore stipulate structure density on the diffusion and polysilicon layers. The dummy structures provide this [94].

Fill cap cells are also provided. These use the gate capacitances to form a fixed capacitance between the rails and are primarily used for decoupling.

All four library permutations (LVT/RVT SS/TT) contained the exact same cells. Table 10 shows the full cell list.

Combinational Cells	Strengths	Fingers
AOI22	X1	1F/2F/3F
Buffer	X4/X8	1F/2F/3F/4F
Inverter	X1/X2/X4/X8	1F/2F/3F/4F
NAND2	X1	1F/2F/3F/4F
NOR2	X1	1F/2F/3F/4F
OAI22	X1	1F/2F/3F
Sequential Cells	**Strengths**	**Fingers**
Preintegrated Clock Gate	X1	4F
Scanable Flip-Flip w/ Asynchronous Reset	X1	4F
Ancillary Cells	**Strengths**	**Fingers**
Well Tie	X1	1F
Tie High	X1	1F
Tie Low	X1	1F
Filler	X1/X4	1F
Fill Capacitance	X4/X16	1F

Table 10: Full Diffusion library cell list

To ensure the final cells produced the characteristic range expected of the full diffusion strategy after the length refactoring, all X1 strength combinational cells were manually re-characterized using the FO4 average propagation delay and leakage current test benches from the previous chapter. Additionally the gate capacitance was measured using the test bench from Chapter 3. Figure 61 shows the results for the TT corner at a nominal temperature of 25 °C and supply voltage of 250mV. The results show that the trends observed previously are maintained. As a final check, all cells in all libraries were manually passed through DRC and LVS using Mentor Graphics Calibre.

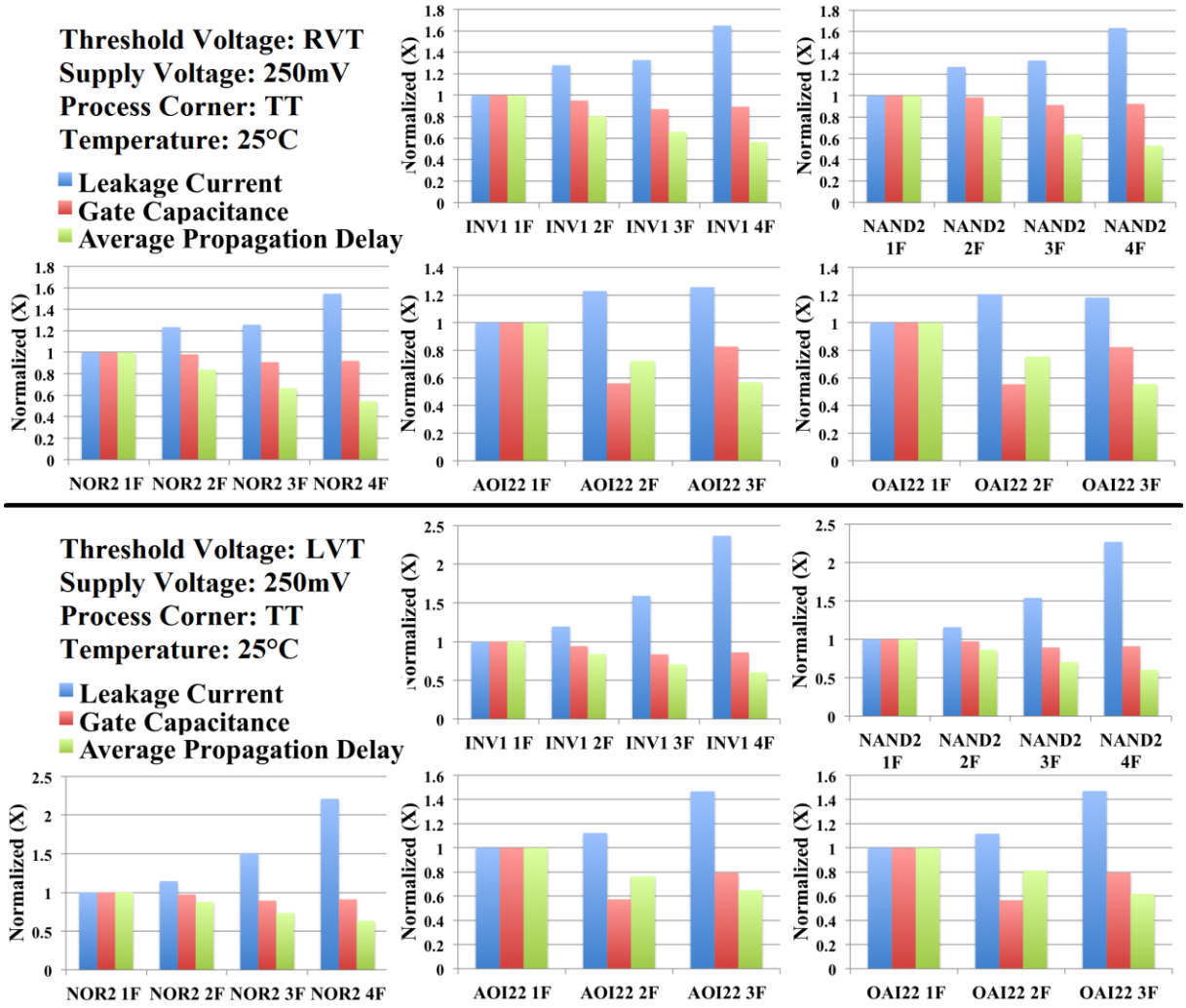


Figure 61: RSCE length optimized cell characterizations

In addition to the four full libraries, four corresponding stacked libraries were created to augment the ranges provided by the full diffusion sizing strategy and provide performance stepping stones between the LVT and RVT ranges should the synthesis tool be unable to synthesize the target AES core without them. The AOI and OAI cells already contain series stacks of two devices and therefore are already in a stacked formation. Given that the buffer cells are used to either improve the drive strength of combinational cells or in the clock tree, there is little reward in generating stack forced versions. The Inverter, NAND2 and NOR2 cells were stack forced. For the inverter cell, all stacks are required. However, in the NAND2 and NOR2 cells, the pull down/pull up network respectively are simply the same stack. As the major drawback of stack forcing is the increase in the dynamic energy consumption due to the increase in overall gate

capacitance, these duplicate stacks were removed. Figure 62 shows the 4F cells in the RVT TT stacked library. Table 11 lists the cells in the stacked libraries.

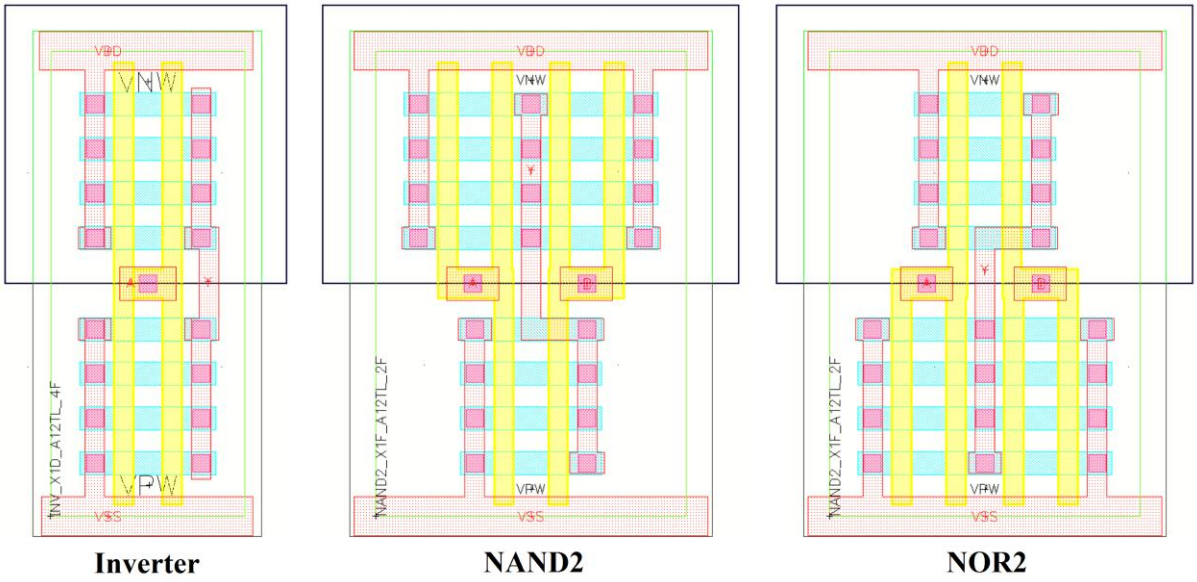


Figure 62: Optimized stack library cell layouts

Stacked Combinational Cells	Strengths	Fingers
Inverter	X1	1F/2F/3F/4F
NAND2	X1	1F/2F/3F/4F
NOR2	X1	1F/2F/3F/4F

Table 11: Stacked library cell list

As a final check, all cells in all libraries were manually passed through DRC and LVS using Mentor Graphics Calibre.

5.6 Characterization

Characterization for the final libraries was performed using ARM's Cellbuilder workflow. This time the workflow was customized for the each full diffusion library. The steps using Cadence Virtuoso to generate the GDSII and CDL, Mentor Graphics Calibre and Synopsys Hercules for DRC and LVS validation and Synopsys StarRCXT for generating parasitically extracted SPICE models remain the same. The next step in the workflow to generate the liberty tables using Synopsys SiliconSmart required customization.

SilconSmart functions by taking the parasitically extracted SPICE models generated in the previous steps, a list of PVT parameters and slew rate/load capacitance ranges and simulates propagation delays for all permutations of the inputs. These are placed into tables for each possible pin transition and the tables amalgamated into a file called a liberty file. These files are then the input for the digital synthesis tools, which estimates path delays based on the tabular information and makes synthesis decisions based on those delays.

It is common commercial practice to produce tables of 7 input slew rates and 7 load capacitances for a total table size of 49 delay estimates. This is small enough that the characterization process is manageable and the efficiency of the digital synthesis tool does not suffer from choice paralysis. It is also large enough to provide adequately accurate results.

SiliconSmart has the functionality to generate the best 7-value range from only three values; minimum slew, typical slew and maximum slew. For thousand cell superthreshold libraries, ARM has a several step process for generating these values from advanced analytical methods and modeling. For small subthreshold libraries, the approved methodology is simpler and requires only three test benches.

A test bench of a single X2 strength inverter driving no load generates the minimum slew. The supply voltage/process corner/temperature are set to the desired values and a step function generator is used to drive the inverter with an ideal input transition. The output slew of the inverter is then measured transitioning between 30% and 70%. The step function is then switched to the alternate transition and the output slew of the inverter is measured between 70% and 30%. The lowest slew value is used as the minimum slew.

The maximum slew test bench consisted of a single X1 inverter with a load of a 45 identical X1 inverters to generate a fan out of 45. The minimum slew as previously determined is then used to drive the input of the single inverter. Again the 30/%70% and 70%/30% slew rate measurements are taken on the output of the single inverter. The largest of these slew rates is used as the maximum slew. The Typical slew rate is simply the FO4 slew rate as generated from the test bench used throughout the previous chapters.

A script was written to automate these tests for 4 subthreshold supply voltage target points. The inverters used were the X1 inverters of the corresponding libraries. Tables 12, 13, 14 and 15 show the results for the TT length optimized full diffusion cells and ARM's Low Energy libraries.

ARM LE Library LVT			Slew (ns)		
Voltage	**Process**	**Temperature**	**Minimum**	**Typical**	**Maximum**
142.5 mV	SS	25°C	2934.60	12620.00	116240.00
150 mV	TT	25°C	178.91	849.61	7936.80
157.5 mV	FF	25°C	20.13	48.60	547.74
190 mV	SS	25°C	653.30	2643.00	37113.00
200 mV	TT	25°C	44.90	184.61	2587.60
210 mV	FF	25°C	5.04	12.81	182.66
237.5 mV	SS	25°C	163.45	605.46	10327.00
250 mV	TT	25°C	20.34	44.11	744.47
262.5 mV	FF	25°C	1.50	5.51	62.87
285 mV	SS	25°C	45.25	149.07	2774.90
300 mV	TT	25°C	5.51	12.64	224.66
315 mV	FF	25°C	0.41	2.39	25.22

Table 12: Liberty table slew rates: ARM LE LVT

Full Diffusion TT LVT			Slew (ns)		
Voltage	**Process**	**Temperature**	**Minimum**	**Typical**	**Maximum**
142.5 mV	SS	25°C	370.44	1352.80	6702.70
150 mV	TT	25°C	49.16	183.01	955.88
157.5 mV	FF	25°C	5.54	25.91	108.10
190 mV	SS	25°C	111.05	353.34	1628.60
200 mV	TT	25°C	13.82	48.45	238.07
210 mV	FF	25°C	1.68	6.98	30.60
237.5 mV	SS	25°C	33.00	100.60	458.22
250 mV	TT	25°C	4.14	23.90	64.77
262.5 mV	FF	25°C	0.57	2.06	9.58
285 mV	SS	25°C	9.60	28.62	119.58
300 mV	TT	25°C	1.27	5.80	18.85
315 mV	FF	25°C	0.21	1.08	7.44

Table 13: Liberty table slew rates: Full Diffusion LVT

ARM LE Library RVT			Slew (ns)		
Voltage	**Process**	**Temperature**	**Minimum**	**Typical**	**Maximum**
142.5 mV	SS	25°C	8473.90	53352.00	584690.00
150 mV	TT	25°C	1499.10	8625.50	110320.00
157.5 mV	FF	25°C	206.68	1147.50	15823.00
190 mV	SS	25°C	2508.30	14085.00	180050.00
200 mV	TT	25°C	425.21	2252.00	32709.00
210 mV	FF	25°C	66.34	347.74	5346.40
237.5 mV	SS	25°C	764.92	3845.40	53603.00
250 mV	TT	25°C	125.89	615.99	9523.70
262.5 mV	FF	25°C	34.08	95.78	1571.30
285 mV	SS	25°C	233.52	1071.60	15722.00
300 mV	TT	25°C	42.90	172.52	2744.30
315 mV	FF	25°C	9.40	37.04	467.14

Table 14: Liberty table slew rates: ARM LE RVT

Full Diffusion TT RVT			Slew (ns)		
Voltage	**Process**	**Temperature**	**Minimum**	**Typical**	**Maximum**
142.5 mV	SS	25°C	17524.00	86935.00	958160.00
150 mV	TT	25°C	1268.60	5713.10	77942.00
157.5 mV	FF	25°C	65.66	322.39	4471.20
190 mV	SS	25°C	4680.50	19801.00	288100.00
200 mV	TT	25°C	337.31	1362.50	21782.00
210 mV	FF	25°C	28.82	85.49	1345.50
237.5 mV	SS	25°C	1267.00	4978.50	82229.00
250 mV	TT	25°C	87.00	344.76	5956.20
262.5 mV	FF	25°C	7.22	47.75	421.42
285 mV	SS	25°C	348.90	1299.40	22556.00
300 mV	TT	25°C	36.06	92.65	1599.10
315 mV	FF	25°C	2.51	31.42	132.51

Table 15: Liberty table slew rates: Full Diffusion RVT

The results show that the slew rates for the full diffusion LVT inverters are less than the ARM LE comparative library across all voltages and corners. The full diffusion RVT inverters display lower slew rates in the typical and fast corners, but the slew times for the slow corner are up to 40% greater than the comparative library. Whilst this appears as a disadvantage, the critical metric is actually propagation delay. As the test bench results are slew rates from driving cells of the same library, little comparative knowledge can be gained.

These results were used to generate accurate liberty files for all libraries, ready for the digital synthesis tools in the next step.

5.7 Digital Synthesis

To provide a fair comparison between libraries, a single identical test circuit was synthesized from all libraries under test. The circuit chosen was a 128-bit AES core with 32-bit datapath and Logical Built In Self-Test (LBIST). This circuit was chosen for several reasons. It is sufficiently large to demonstrate that commercial digital synthesis workflows can adequately convey the benefits of full diffusion cells into complex digital circuits. This is important as it proves the commercial feasibility of the technique.
The circuit represents a good mix of the available gates being combinatorial and sequential in nature. This allows utilization of the full cell list from each library, which helps gauge the benefits of the sizing strategy for the different gate types.
The circuit also provides a simple method of determining failure. As the circuit is able to encrypt and decrypt a provided test vector, then determine if the correct vector has been returned via the LBIST, a simple pass/fail result can be returned for any operating point the circuit is asked to perform under. Optimal operating points may then be derived using simple search algorithms to speed up testing.

Finally, the RTL for the design was already available from ARM. The advantage to this is that the design has been tried and tested by previous researchers and found to be robust and functional. This alleviates any errors that may be introduced at design time by the author to invalidate the results. The disadvantage to this is that the exact functioning of the circuit was not designed by the author and therefore not customizable without invalidating some of the advantages. The simplistic nature of the LBIST makes the

integration of the circuit into a larger chip relatively simple due to the reduction in I/O lines.

The commercial tool chain chosen to perform the digital synthesis was the Synopsys tool chain. This is widely considered within the industry to be the most professional and offer the most control over synthesis. However it is also the most expensive to license.

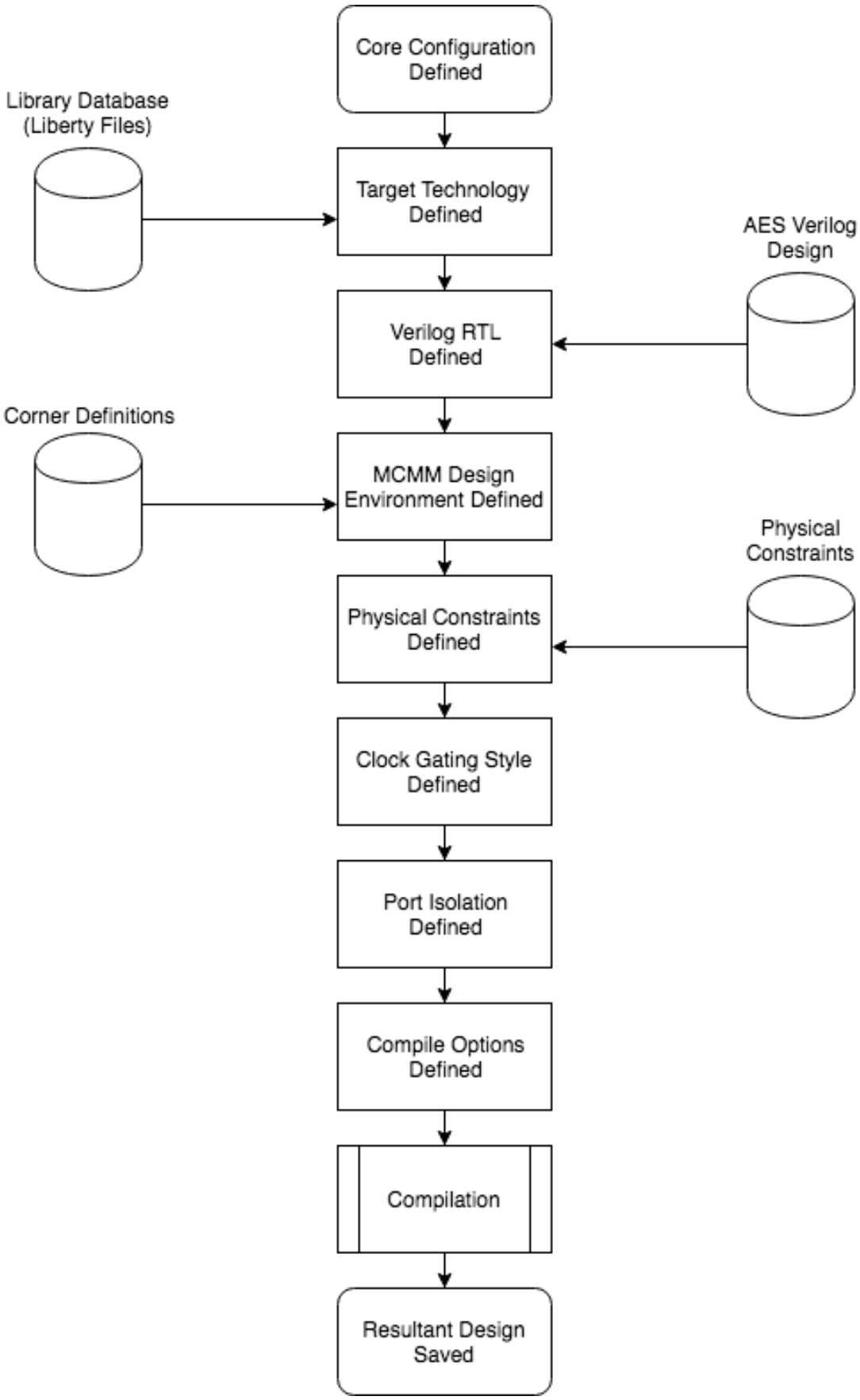


Figure 63: Design compiler digital synthesis work flow

Four tools from this tool chain were used within the digital synthesis workflow; Design Compiler, IC Compiler, PrimeTime and VCS.

5.7.1 Circuit Synthesis (Design Compiler)

Design Compiler is Synopsys's digital synthesis tool. Figure 63 shows the stages of the synthesis flow. The core configuration is first described. This defines the design's name and ports, namely the scan enable, clock and reset ports for the AES core.

The next stage in the flow defines the target technology node. The compiled liberty files for all PVT corners generated from the characterization of the cell libraries are read and analyzed by the synthesis tool. Physical parameters of the target technology are set. This stage defines the expected input cell and output capacitance, from which the circuit can determine the input slew rate and output load capacitance to expect from the environment. For the full diffusion libraries, this is X1 strength 4F inverter from the library under synthesis. If the synthesis is a multi-Vt run, this is the inverter cell from the fastest library (LVT). For the ARM's LE Library, this is the equivalent X1 strength inverter.

The clock parameters are then defined. These include the hold margins, derate factors, uncertainty, critical range, insertion latency and ideal transition times. Most importantly, this is where the target clock period is defined.

Routing parameters are then defined including the maximum metal layer allowed for cell routing, as well as the maximum allowed routing multipliers for the clock tree. The maximum fan out (set to 32) is also set.

Following this are the cell definitions where the types of cell (Ties, fills, fill caps) are matched to the corresponding cells in the libraries. The maximum cell distance to a well tie is also defined to prevent the drift of well voltages due to accumulation of resistance.

The synthesis tools then reads in and analyzes the AES RTL Verilog design. The design is then elaborated. This involves the synthesis tool scanning the Verilog description looking for known generic technology cells and transforming the design into a netlist, which can then be optimized based on the timing information mapped from the previous technology stage.

The corner definitions are then defined. Design Vision is capable of performing Multi-Corner Multi-Mode (MCMM) synthesis to ensure the functionality of a synthesized design under different operating conditions. These are termed environments within the tool. Five main environments were defined in the flow corresponding to the SS/TT/FF

process corners at supply voltages of 250mV, 400mV, 500mV, 600mV and 1.2V (centered around the typical process corner).

The physical constraints for the design are then defined. Two types of physical constraints were used during synthesis for two different processing runs described in later in the chapter. The first defines aspect ratio and area utilization. This allows the tool to simply expand into whatever area is required. This has the advantage of allowing flexibility of area in the synthesis decision. The second is a strict area boundary. This is the preferred method for silicon signoff as it guarantees a final design fits within the silicon area allocated in the floor plan of the die. In this case it is the area utilization that is flexible.

Next the clock gating is set up. This defines the clock gate cell as well as the logic types that drive it and the enable signal. This is followed by the final set of options set for the compiler and isolation settings for the ports. All synthesis runs performed used inverter isolation to provide accurate time model creation of the internal circuit design.

The tool then compiles the design. It makes several iterative runs until the timing is met or the lowest Worst Case Negative Slack (WNS) is determined. The determination of the timing figures are based on assumptions made about the probable relative geometry of cells in the final design. Therefore the reported WNS at this stage is only an estimate.

The design is then saved to a Synopsys DDC database file so that the rest of the tool chain can read it.

5.7.2 Cell Placement (IC Compiler)

IC Compiler is Synopsys's automated place and route tool. Figure 64 shows the stages of the cell placement flow. The first two stages are identical to the circuit synthesis flow in Design Compiler, namely the core configuration is defined and the target technology node mapped. The next stage imports the synthesized design from Design Compiler in the form of the DDC database file. A check is the performed to ensure the MCMM corners specified in the Design Compiler flow have been correctly imported into IC Compiler. The placement optimization parameters are then set. These include which cells should/shouldn't be used and where tie cells should be placed (ensuring they are placed in each row of the digital logic). The setting of the routing parameters follows, defining the width and spacing rules, routing layers and routing specification of the clock tree so that the clock tree synthesis of the flowing stage may be approximated to ease it's eventual implementation.

The initial placement is then performed. The tool then iteratively performs timing and DRC optimization stages until a DRC clean placement with either all paths with timing met or a placement with the lowest WNS is determined. The design is then saved for clock tree synthesis to be performed.

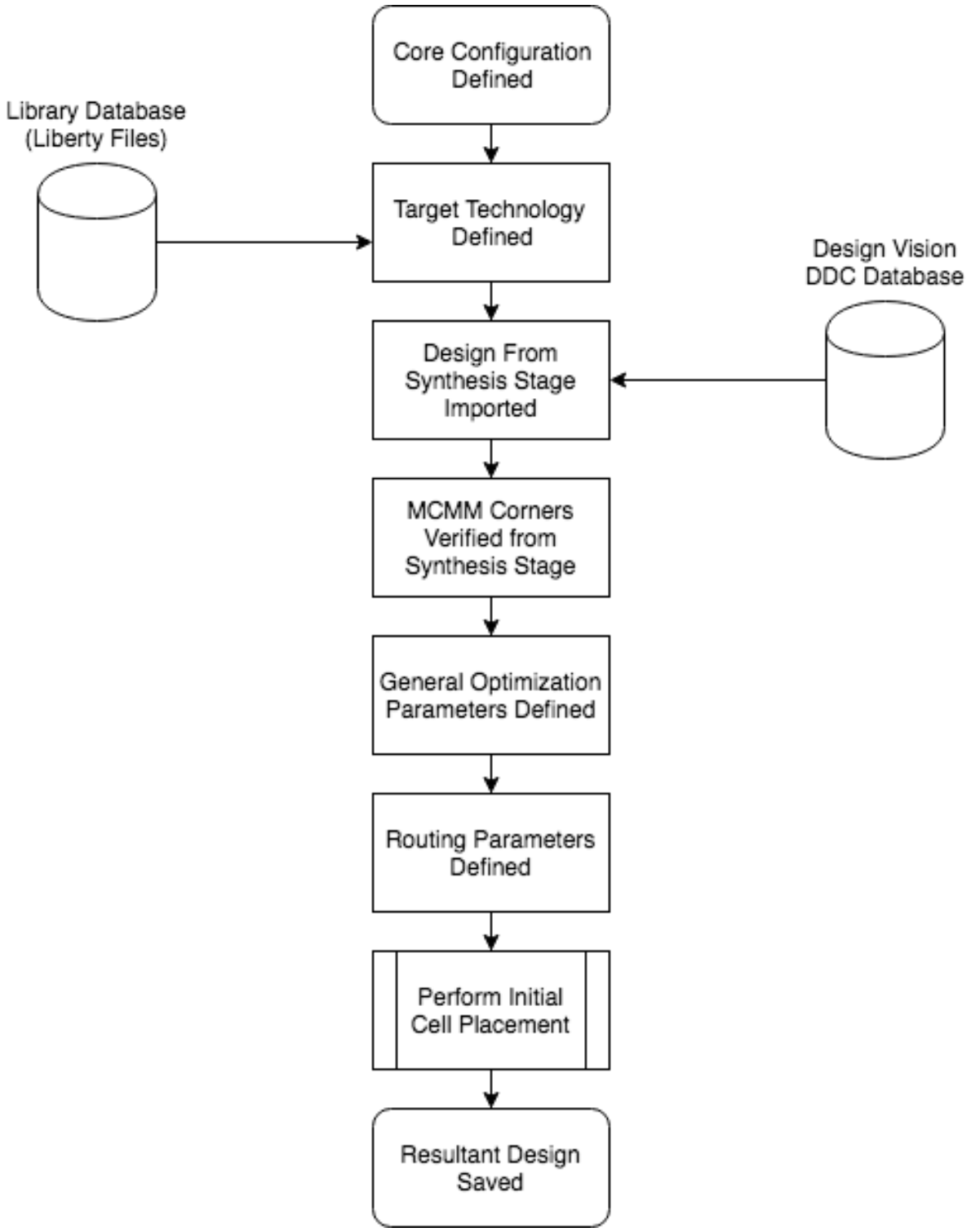


Figure 64: IC compiler cell placement work flow

5.7.3 Clock Tree Synthesis (IC Compiler)

The clock tree synthesis is also performed in IC Compiler. Figure 65 shows the design flow. The first five stages are identical to the automated placement range. The clock tree options are then defined. The maximum transition time, routing layer and maximum fan out (32) are set. The logic level balance is set to false, allowing the faster transition to dominate the clock tree. Hold fixing is set to prefer buffer insertion as the solution. The derate factors are set then the clock tree is synthesized, optimizing for power and re-ordering the scan chain. The clock tree synthesis process iterates until all hold violations are fixed. The design is then saved for the final routing stage.

5.7.4 Routing (IC Compiler)

The final stage in the digital synthesis flow is the routing stage. This is also completed using IC Compiler. Figure 66 shows the design flow. Again the first five stages are identical to the previous IC Compiler workflows. The next stage is defining the routing parameters. The routing optimizer is set to minimize delta delay using total slew transition times. Crosstalk prevention is enabled with a threshold of 25%. Static noise analysis is also enabled. The routing optimizer strategy is set to interactively repair any failing routes up to a maximum of 20 times. Synopsys's multicore ZRoute technology is then enabled.

The first of three routing stages is then performed. This is a simple initial route to make all interconnections in the design with minimal optimization. The success of this stage is then reported. A second routing stage follows with crosstalk and power optimization. A second report is then generated. Finally a third routing stage is performed to incrementally eradicate all hold time violations.

Once the final routing stage has completed, the final process of adding filler cells into the unused design space is performed. The final design is then saved ready for export and analysis.

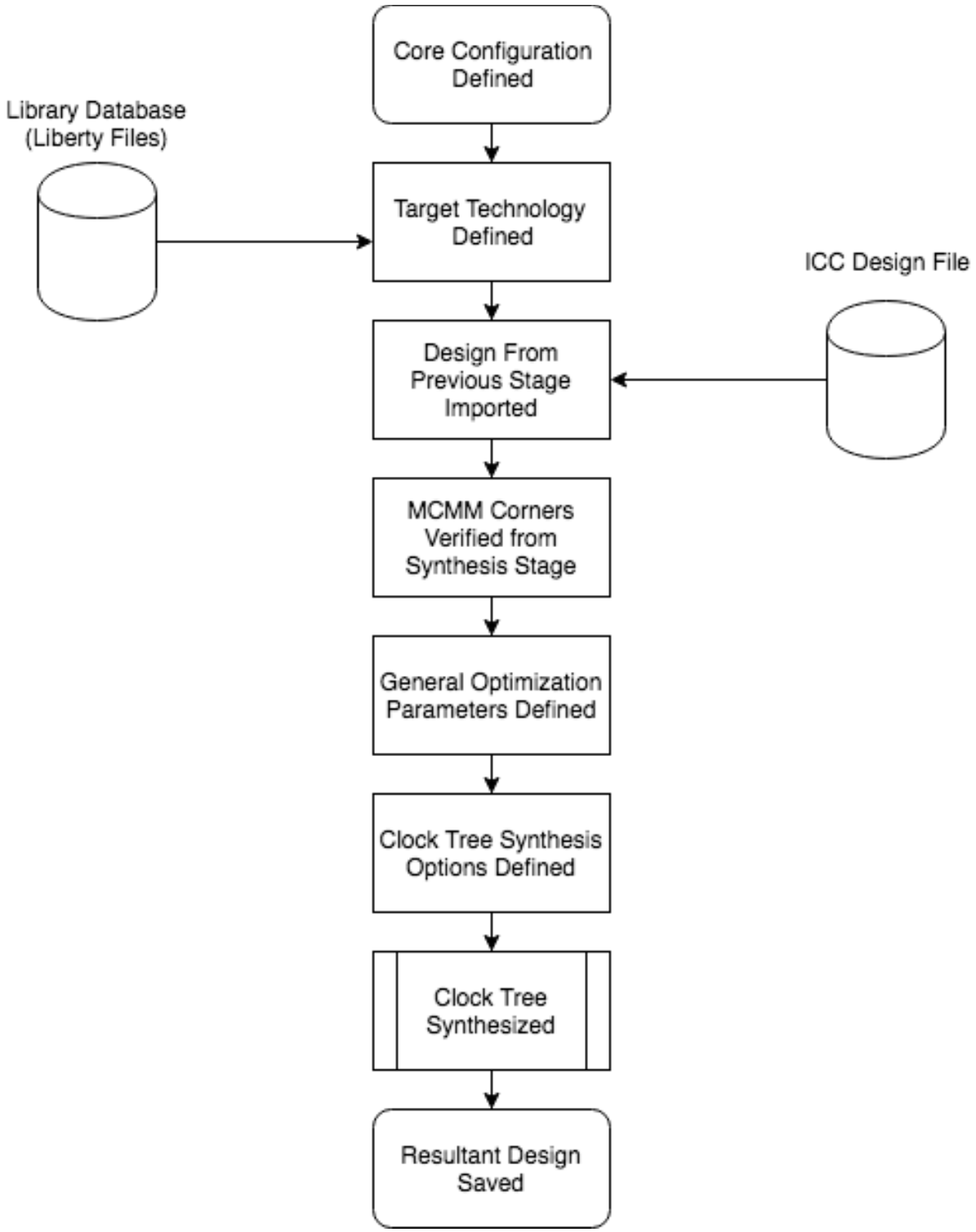


Figure 65: IC compiler clock tree synthesis work flow

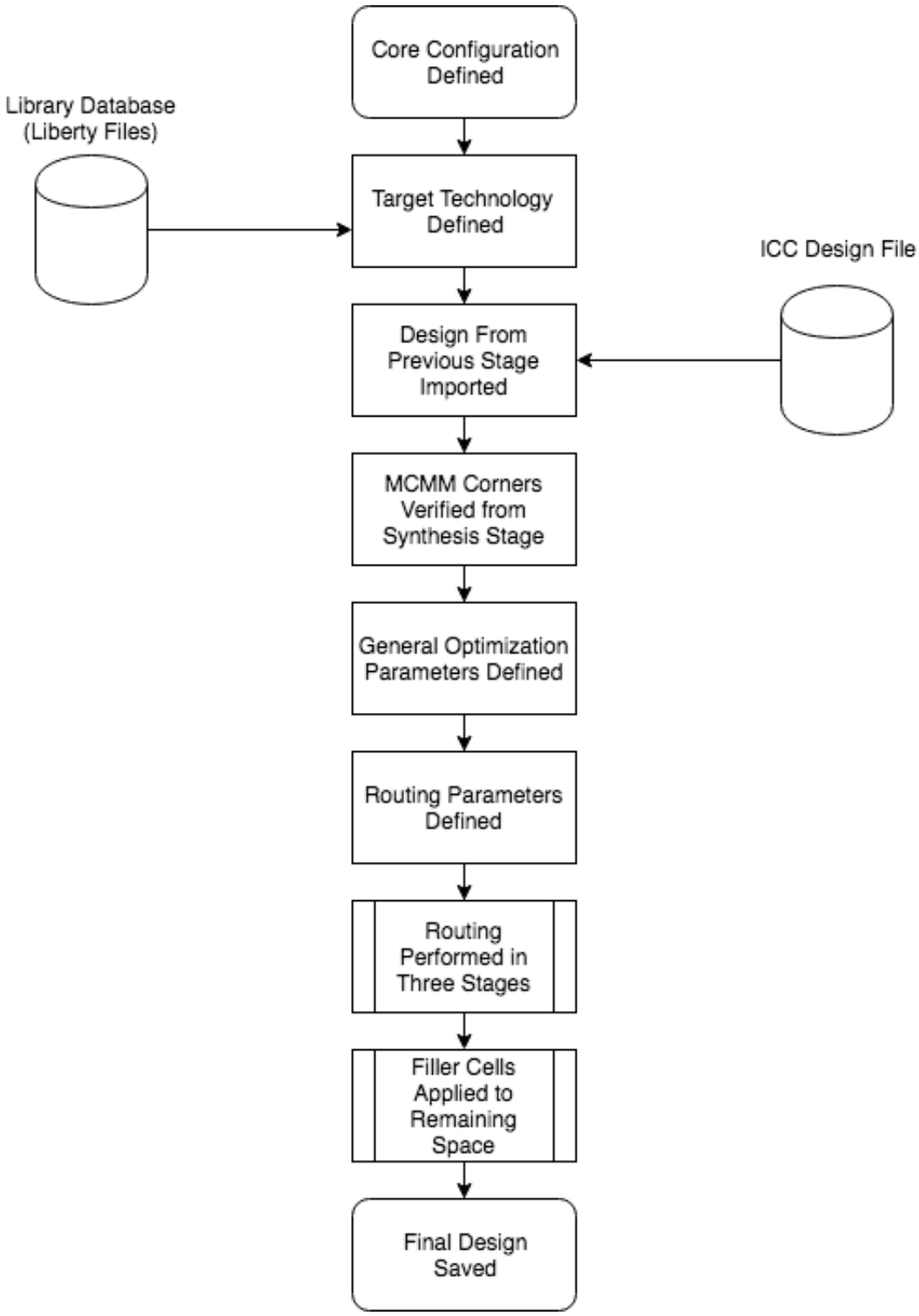


Figure 66: IC compiler automated routing work flow

5.7.5 Design Export

Once the design is complete, it must be exported such that it may be further analyzed or used by other tools. This is also performed in IC Compiler. Figure 67 shows the workflow.

The core configuration, target technology and design import are identical to the previous stages. An LVS stage is then completed to ensure the final design performs the intended behavior described by the initial Verilog model. This is followed by a DRC check of the routing to ensure no design rule violations have inadvertently been introduced during the routing phase. If the design passes these validation checks, two export stages proceed. The first is the generation of an FRAM cell view. This is the view of choice for IC Compiler's die level floor planning layout methodology. The second is a GDSII stream. This is the industry standard design stream.

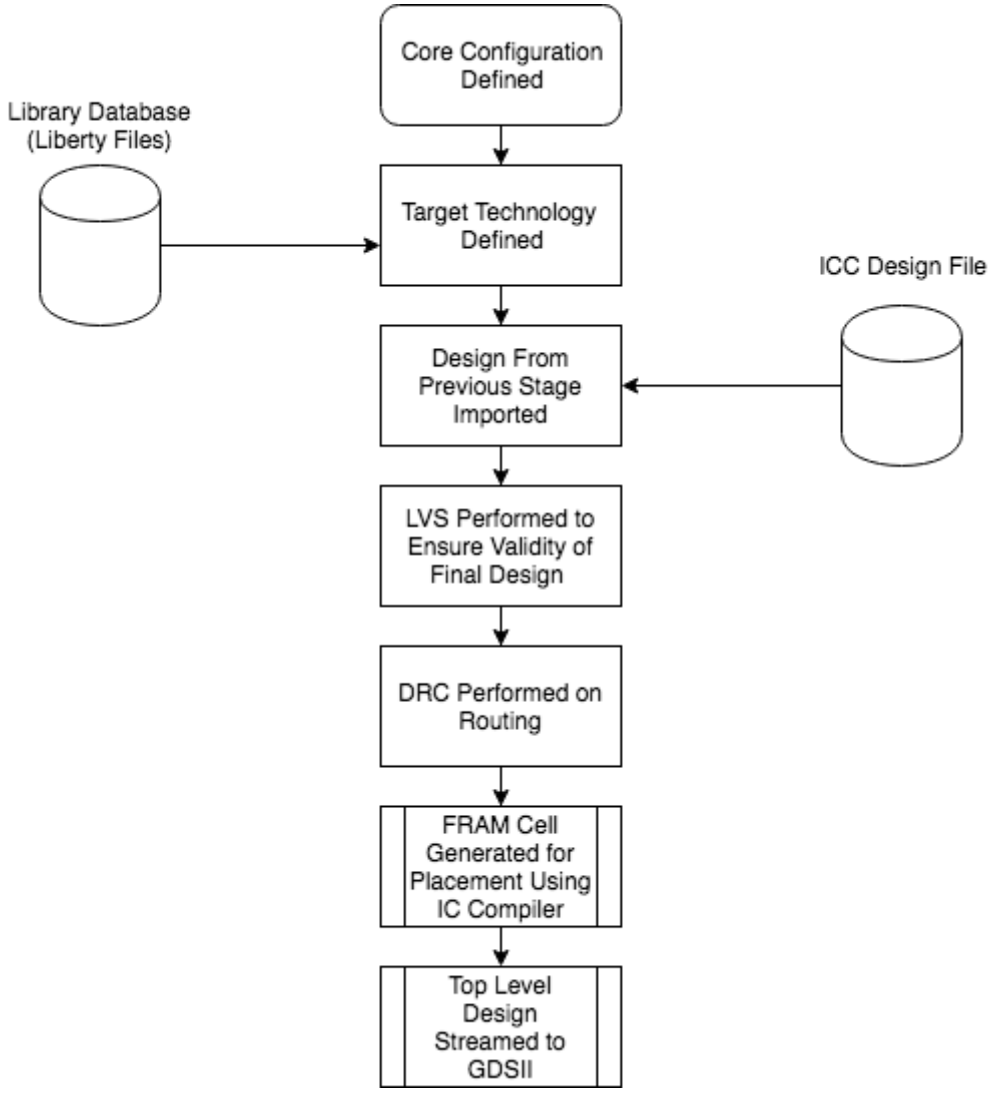


Figure 67: IC compiler design export work flow

5.8 Choice of Libraries and Corner Choice

The individual stages in synthesizing the AES cores have now been defined. A decision on which libraries should be implemented was then made. Limitation of area on the target die restricted the number of cores that could be implemented in silicon to three. For any kind of meaningful comparison, one core had to be synthesized from ARM's state-of-the-art Low Energy Library. This library does not provide the performance/leakage characteristic range that the full diffusion library offers. As such, ARM has taken the view that multi-Vt synthesis in the subthreshold regime is simply too impractical due to the 10X difference in performance between the LVT and RVT devices. As the primary focus of subthreshold operation is the reduction in energy consumption, ARM's LE library consists purely of RVT devices, as the Ion/Ioff ratio at the typical corner is almost double that of the equivalent LVT device.

To provide a direct equivalent to this, the decision was taken to synthesize the second core purely from a full diffusion RVT only library. This provides the full 1F to 4F range in the cells described earlier in Section 5.5. No cells were included from the stacked library.

Several simple synthesis runs were completed to determine the libraries used for the third core. The result showed that the range provided by the full diffusion LVT and RVT libraries was sufficient to reliably synthesize the design with only a performance cost of 1.2%. The interstitial stacked libraries were therefore omitted from the design to ensure their increased capacitance could not negatively impact the dynamic energy consumption of the final design. Should the reliability of the design have been compromised by this decision, the interstitial libraries would be mandatory. The final core was therefore a multi-Vt implementation consisting of the full diffusion LVT and RVT libraries.

The full diffusion libraries were length optimized during design for use at the SS and TT corners. A decision over which library should be used was therefore weighed. The typical digital synthesis flow is signed off at the SS corner because the resultant circuit and therefore product can be guaranteed to function at a particular clock frequency. This is commercially advantageous as it means a higher yield only impacted by logically non-functional dies. Conversely, in some instances designs are signed off at a nominal (usually typical) corner and then a process known as frequency binning is performed. Here, the resultant dies are tested over the full range of global variation and allocated into frequency bins. Those only functioning at a clock speed less then nominal are de-rated

and sold at a discount. Those showing operation at clock speeds higher than nominal are sold at a premium.

There are advantages and disadvantages to both approaches in terms of the intended outcome of the experiment. The advantage of signoff at the SS corner is that this is the commercially accepted approach and guarantees the functionality of the circuit a particular frequency. The disadvantage to this approach is that the technology node chosen has been consistently shown to return TT silicon. Therefore, signing off at the SS corner would unnecessarily restrict the results with the limitations imposed by the digital synthesis flow and cell library by attempting to account for the worst case scenario. Conversely, the advantage of signing off at the TT corner is that there is no restriction imposed upon the design by the synthesis flow. Moreover, the underlying physics of the silicon matches the sizing optimization of the cell library. The disadvantage of this approach is that this is not the commercially accepted methodology of digital synthesis. In order to demonstrate a more accurate representation of the abilities of the full diffusion library, the decision was made to sign the design off at the TT corner. This option also gives ARM's LE library a fairer comparison, as the device length use throughout their library is 90nm, yet the optimal length results presented at the beginning of the chapter showed that this can reach as high as 200nm in the SS corner, but only 100nm in the TT corner for the RVT devices.

5.9 Determination of the Maximum Frequency of Operation

Two additional stages are required to perform analysis on the final design; parasitic extraction and static timing analysis. The former is performed using Synopsys StarRCXT. This is the same tool used in the cell characterization workflow and is performed on the exported design. This gives greater accuracy to the static timing analysis.

The static timing analysis is performed using Synopsys PrimeTime. Figure 68 shows the workflow. The first two stages of defining the core configuration and defining the target technology node are the same as previous workflows. The next stage imports the post layout Verilog netlist from the synthesis stage. The netlist is then annotated with the RC parasitics extracted during the parasitic extraction workflow. Three corners are generated during the extraction workflow; RC Best, RC Worst and Typical. All three corners are run during static timing analysis. The RC worst corner was used for sign off to ensure functionality [95].

The timing parameters are then defined including the derate values and analysis options. The static timing analysis is then performed. Once completed, the generated timing data is written out in the IEEE defined Standard Delay Format (SDF) to be used during post-layout simulation workflows. Finally, human readable reports are generated. These include reports for skew, latency and transitions times, as well as overall coverage and any bottlenecks. Most importantly, reports are generated detailing the fastest 10 paths in the design and the slowest 10 paths in the design. The latter report is used to determine whether all paths met the timing criterion and if not, determine the Worst Case Negative Slack (WNS) of the design.

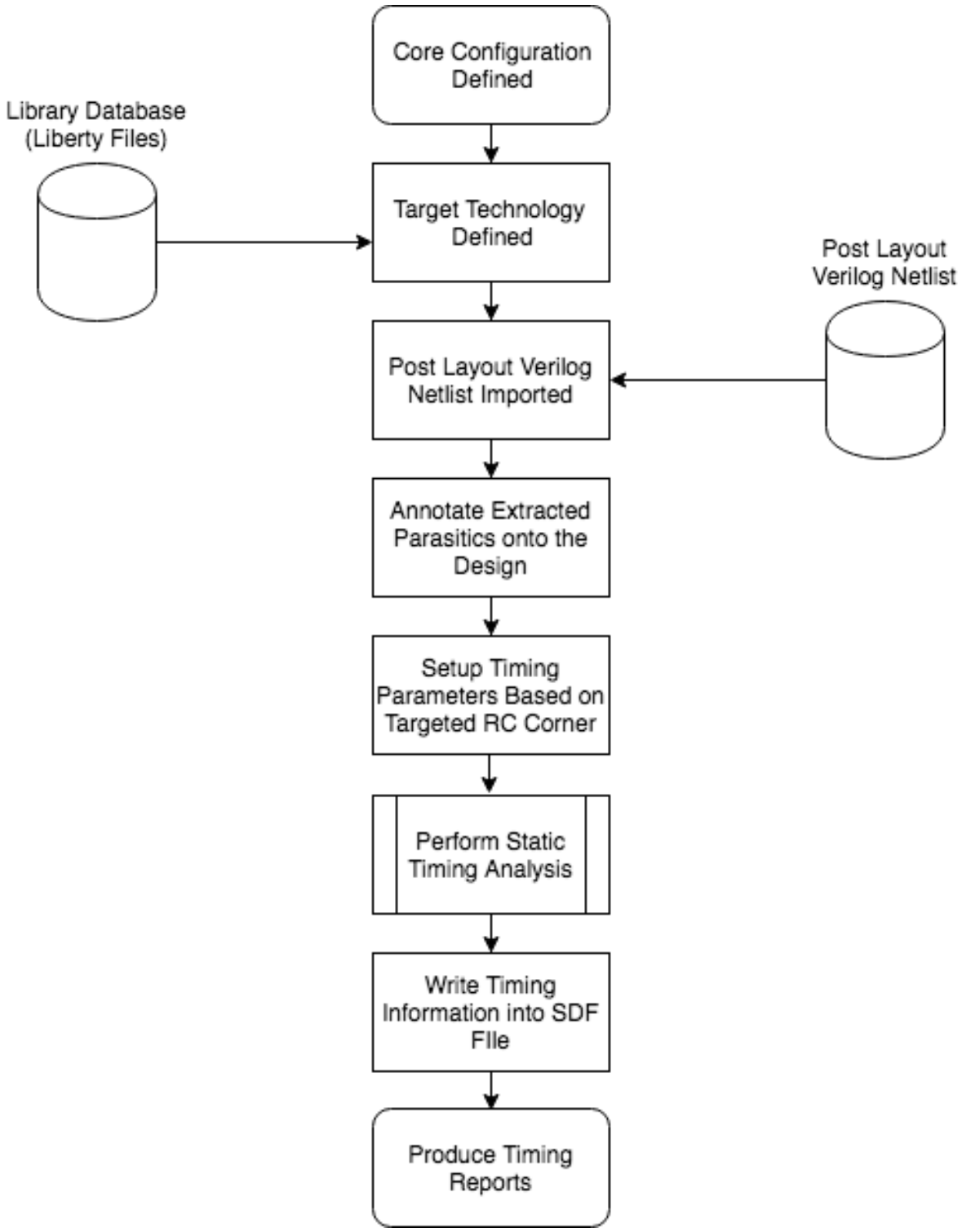


Figure 68: PrimeTime static timing analysis work flow

Thus far, no inherent methodology of determining the maximum frequency (minimum clock period) for a given PVT corner has been described. The reason for this is that the tools are simply not designed to perform in this manner. When commercially designing a digital circuit, the designer is invariably provided with a target frequency and modifies the design to meet this constraint.

An iterative methodology was therefore designed to perform this task. Figure 69 shows the methodology. Initially, an extremely relaxed clock period for the design is provided. The design is then synthesized using the MCMM Design Vision synthesis workflow outlined in Section 5.7.1. The cells are then placed using the IC Compiler automated placement workflow outlined in Section 5.7.2. Clock tree synthesis is then performed using the CTS workflow outlined in Section 5.7.3. The Routing workflow in Section 5.7.4 finalizes the design stages. The design is then exported using the workflow outlined in Section 5.7.5. Parasitic extraction is performed using the workflow from Section 5.9 and finally static timing analysis performed using the workflow from Section 5.9. The WNS for the critical path is determined for the entire run from the timing analysis reports. If all paths meet the provided clock period or the WNS is less than 5% of the provided clock period, the clock period is reduced and the whole methodology iterated. If the WNS is 5% of clock period, this period is determined as the maximum frequency of operation. The 5% watershed was an arbitrary value chosen for the experiment. The primary reason for this value is that the synthesis tools are notorious for non-deterministic performance. As the tools are designed simply to meet a provided clock constraint, once this target is met, little optimization is performed afterwards other than leakage recovery. If the same design is provided with a more stringent timing target, the tools work harder to meet that target. Therefore, there is no linear relationship between provided clock period constraints and WNS. Moreover, this is compounded by the tools freedom of design choice. It is completely feasible that for a chosen clock period, a synthesized circuit type fails to meet the target, only for a tighter constraint to be applied, a new circuit type option to become available and the clock target to be met. The 5% value allows enough room in the synthesis runs to allow the synthesis tools the freedom of design, whilst obtaining a high value of certainty that no other synthesis choice will provide a faster design. Moreover, for a signed-off circuit, 5% is a marginal amount to increase the clock frequency by in order to attain stable functionality. Whilst figure 69 shows that the methodology stops at 5%, for the aforementioned reasons, several runs were performed beyond the 5% watershed to ensure no superior design could be synthesized beyond this point.

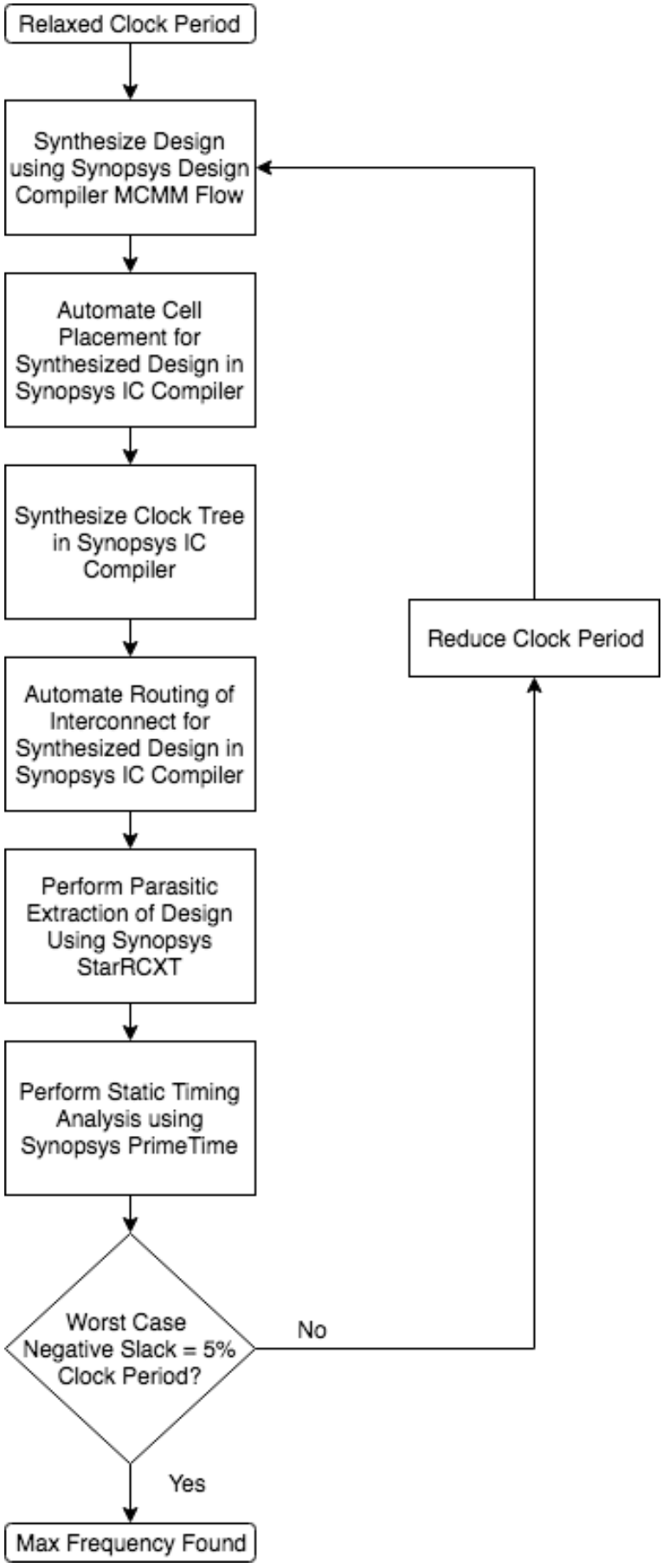


Figure 69: AES core maximum frequency determination methodology

The methodology was performed for all three libraries. Figure 70 shows the results. The ARM Low Energy library produced a maximum frequency of 23.5 KHz from a clock period of 42550 ns. At this point, the WNS was 4.5% of the clock period. Beyond this point, no run produced a value less than 5%. The individual points show the non-conformity in the WNS over frequency, most notably the 22.72 KHz point produces a WNS less than the previous run of 22.22 KHz. It is likely that at this frequency point, the tool is able to synthesize a circuit variation it is incapable of synthesizing at the lower frequency point.

The Full Diffusion RVT library produced a maximum frequency of 42.88 KHz from a clock period of 23320 ns. At this point, the WNS was 4.2%. The next run period of 23310 ns produced a WNS of 8%. The difference in WNS between two clock constraints of such proximity again suggests the synthesis tool is only able to produce certain circuit choices under certain clock constraints.

The Full Diffusion multi-Vt library produced a maximum frequency of 414.94 KHz from a clock period of 2410 ns. The WNS for this run was 5%. The 2400 ns run produced a WNS of 5.6%, again expressing the non-conformity in the synthesis process.

There are a few important points of consideration from the experiment. The first is that on a 'like-for-like' basis, the maximum frequency of the Full Diffusion RVT library was 1.83X faster than ARM's Low Energy library. This means that even for circuits with complexity running at around 7000 gates, the potential benefit of the full diffusion sizing strategy measured on a cell for cell basis is still observable.

The second is that the Full Diffusion multi-Vt run had a maximum frequency 9.63X faster than the equivalent than the Full Diffusion RVT only run and 17.66X faster than the ARM library. Unsurprisingly, the multi-Vt run is therefore clearly the choice for circuit speed. The underlying information from the run provides more details of interest. When pushed to the frequency limits, the synthesis tool chose to implement the design using 94.45% of the cells from the LVT library and the remaining 5.55% from the RVT library. This shows that the synthesis tool is displaying the correct behavior for frequency critical multi-vt synthesis.

Whilst this methodology proved successful at determining the maximum frequency, it was time consuming and only accurate to within 5%. Faster and more accurate methodologies would provide an interesting basis for future work.

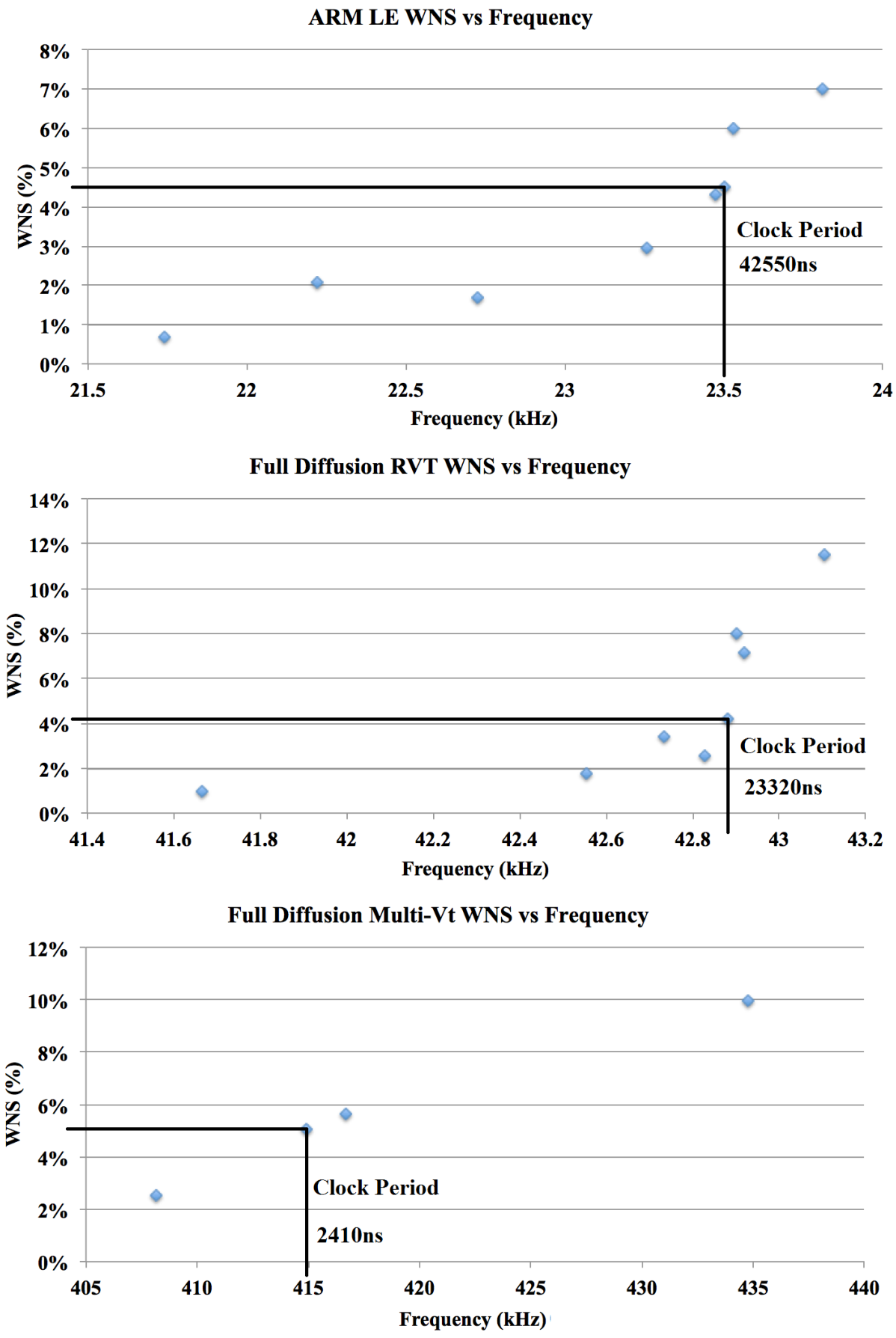


Figure 70: AES core maximum frequencies

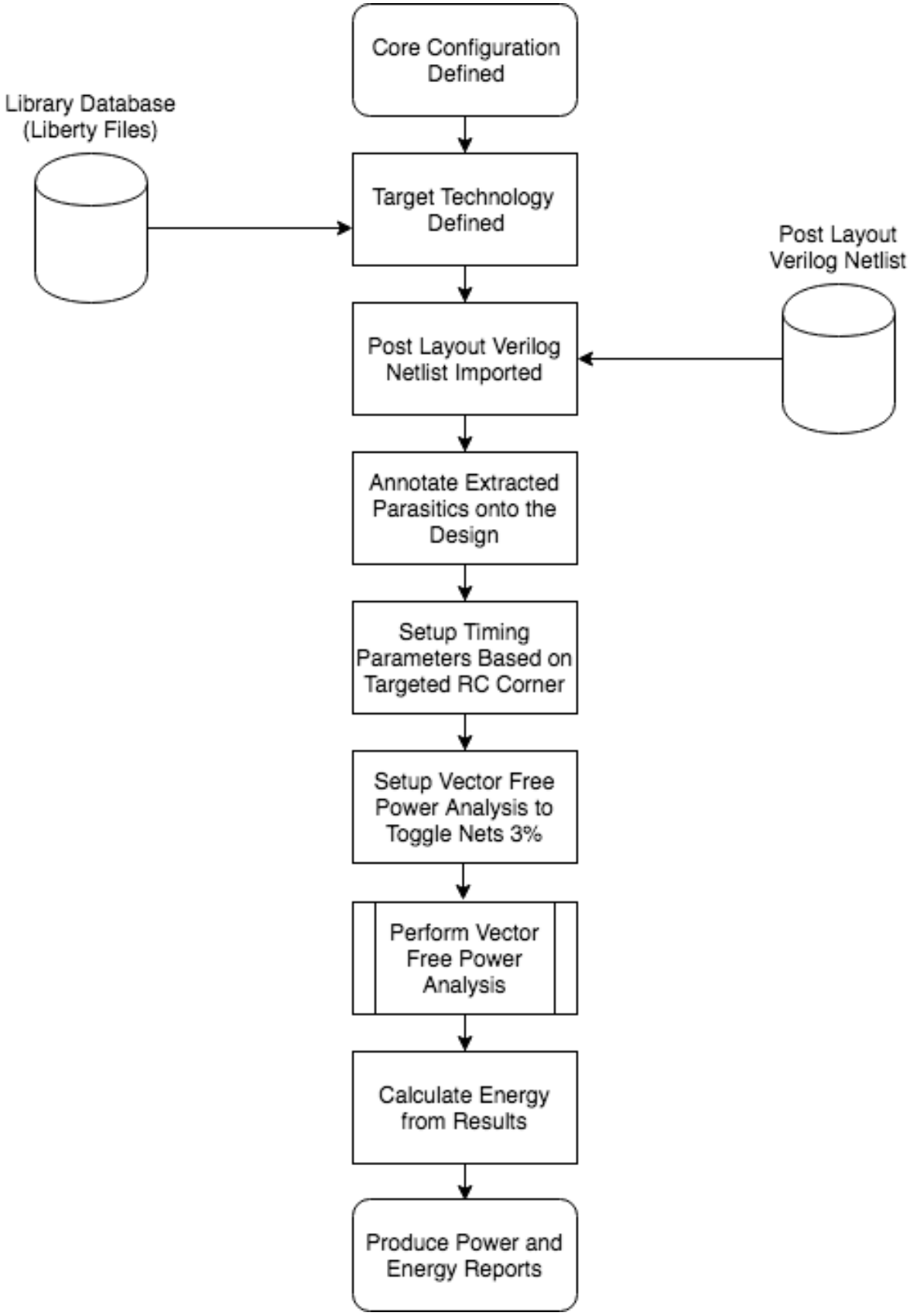


Figure 71: Vector free power analysis work flow

5.10 Vector Free Power Analysis

To ensure correct MCMM synthesis, a simple post-layout power analysis stage was performed across all supply voltage corners. This was performed in Synopsys PrimeTime. Figure 71 shows the work flow.

The first five stages of the workflow are identical to the static timing analysis workflow described in Section 5.9. The next stage sets up the power analysis parameters. The simplest form of power analysis is vector free analysis, which involves simply instructing the tool to toggle all nets at a fixed activity factor. This is adequate to test the MCMM flow is functioning as expected. The chosen activity factor was 3%. The power analysis was then performed. A final stage converts the results from the power analysis into energy results and reports for both power and energy are produced.

This stage was performed at each supply voltage corner in each library for the clock period determined to produce the maximum frequency at 250mV. Figure 72 shows the results. The characteristic energy sweep (dynamic energy reducing and leakage energy increasing as supply voltage tends to zero) is clear from all three libraries. Inconsistencies in the graphs (total energy dipping below dynamic energy and leakage energy dipping below zero etc.) are purely a function of the smoothing algorithm implemented to show the correct response from a small number of points.

The ARM LE library shows a minimum total energy of less than 1 pJ per cycle around 470 mV. A similar result is shown for the full diffusion RVT library, although the minimum total energy is slightly higher. The multi-Vt full diffusion run shows a minimum energy greater than 1 pJ per cycle.

The overall sweeps indicate that the MCMM flow is functioning as expected and therefore sanity checking using vector free power analysis is an acceptable method of validation. Little further analysis is provided for these results for several reasons. This first is that no meaningful design simply toggles its nets for a fixed activity factor. Therefore the test is only designed to provide an indication of the underlying power, not absolute results. Moreover, for the AES core that operates several stages of the design simultaneously, the activity factor is considerably greater than 3%.

Most importantly the methodology used to synthesize the design was constructed to determine the maximum frequency possible for the design from the libraries, not the design displaying minimum energy. In order to do this, a different synthesis methodology and more accurate power analysis stage is required. This is the focus of the next section.

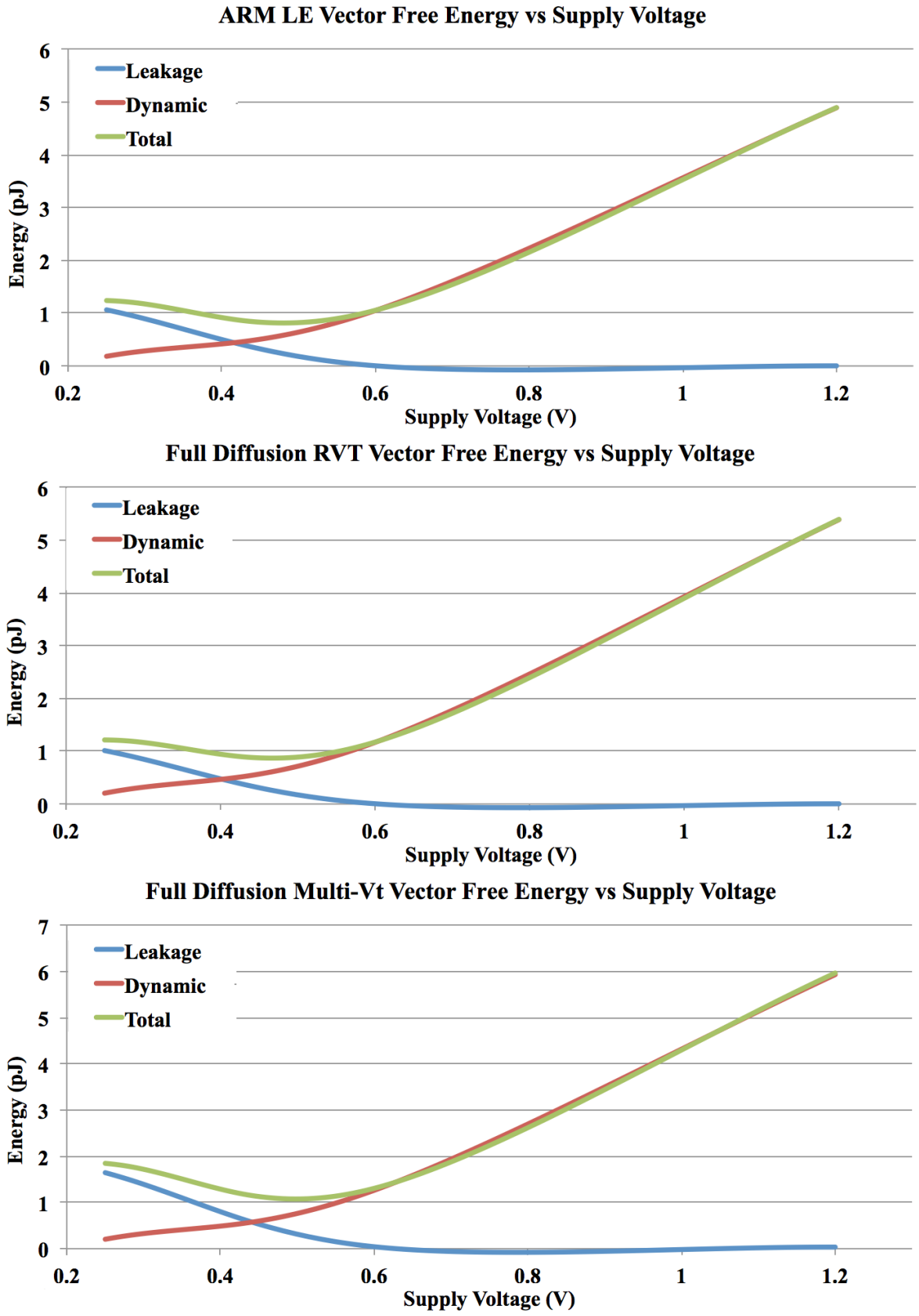


Figure 72: AES core vector free minimum energy points

5.11 Determination of Minimum Energy Implementation

Synthesis of the maximum frequency design from each library showed that the speed improvement of the underlying full diffusion cells propagate correctly all the way up to complex designs. However, the purpose of voltage scaling is to lower dynamic energy consumption and find the minimum energy point of operation. The synthesis decisions for the design are different for maximum frequency and minimum energy. For maximum frequency, the fastest cells available are almost exclusively chosen. The drawback to this is that these cells are also the leakiest and in subthreshold design, the leakage energy often dominates the total energy consumption. Lowering the target frequency gives the synthesis flow more freedom to provide leakage recovery. This is the process of exchanging fast, leaky cells to slower, less leaky cells in fast timing paths without violating the overall time constraints.

To systematically perform this type of synthesis, a second methodology was performed. This required two additional analysis stages to improve the power and thereby energy analysis of the final design.

The first stage creates a test bench from the design using the interface to the LBIST and the SDF timing data files exported during the PrimeTime static analysis stage. This is compiled using Synopsys's VCS verification tool. The test bench is designed to provide a test vector to the LBIST and run the AES core for 20 BIST cycles. As the timing comes from the SDF files, which are generated after the export and parasitic extraction of the final design, this type of simulation is very accurate. The test bench is then simulated using Synopsys VCS. The simulation is designed to record exactly which nets of the final design toggle during an actual encryption/decryption LBIST operation. Once the simulation is completed, the toggle data is exported as an industry standard Value Change Dump (VCD) file.

The second additional analysis stage imports this back into the PrimeTime power analysis workflow, allowing the tool to provide VCD power analysis based on the actual nets that toggle during the function of the design. This is far more accurate than the vector free workflow previously described.

Figure 73 shows the new synthesis methodology. Initially the clock constraints are set to those determined for the highest frequency point found in the previous methodology. The same MCMM synthesis workflow is performed in Design Vision with leakage recovery enabled. As the objective of the methodology is to relax the timing constraints to find the minimum energy point, the buffering required to meet these constraints should lower,

lowering the overall cell count and therefore silicon area. The area of the maximum frequency implementation is therefore used to provide a hard boundary in an additional floor planning stage. This allows this area to be allocated on the die floor plan. The three automated place and route stages are then performed in IC compiler. At this point the design is exported and extracted as in the previous methodology. These stages are omitted from Figure 73 for the sake of brevity. Static timing analysis is then performed to generate the parasitic annotated SDF timing data files. The design test bench is then compiled and simulated in Synopsys VCS to generate the VCD vector file. Finally VCD power analysis is performed in PrimeTime and the power and energy reports generated for the design.

As the objective is to determine a local minimum, a fixed number of 10 iterations were performed at 10% frequency intervals for the ARM LE and Full Diffusion RVT libraries and 2X intervals for the Full Diffusion Multi-Vt library. This gives a final frequency test point of 2X slower for the RVT only libraries and 20X for the Multi-Vt library. Figure 74 shows the results.

Quick observation shows that for the AES core, the 250mV TT PVT point was dominated by the dynamic energy consumption for all libraries. This is likely due to the high activity factor the design. All libraries showed that the total energy per cycle peaks at the maximum frequency point determined from the first methodology. This is no surprise given the additional buffering in the design required to meet that frequency. The total energy per cycle then diminished as the frequency was relaxed.

The minimum energy implementation for the ARM LE library was 40% slower than the maximum frequency implementation with a clock period of 59570 ns equating to a frequency of 16.79 KHz. At this point, the total energy per cycle was 3.27 pJ. After this point, the total energy began to rise to a peak, fall and then began to steadily rise again. This discrepancy is likely due to the inability to synthesize a particular circuit structure around this frequency, indicated by the rise in dynamic energy consumption but little variation in leakage energy consumption. The leakage energy fell off sharply from the maximum frequency point and then began to slowly rise. Whilst this may seem counter intuitive as the level of leakage recovery increases, leakage recovery with a choice of only one cell variant requires restructuring of the circuit. Moreover, as the frequency reduces, the paths leak for a longer period of time. The only explanation for the latent rise in dynamic energy per cycle is a local maximum created due to an inability to synthesize

a circuit topology available at higher frequencies. This should be expected to reduce should the number of runs be extended.

The minimum energy implementation of the Full Diffusion RVT library was 70% slower than the maximum frequency implementation with a clock period of 39644 ns equating to a frequency of 25.22 KHz. This means that the comparative frequency gain over the ARM library has been diminished. Moreover, at the minimum energy point, the total energy per cycle was 3.81 pJ. This is an energy consumption 16.5% higher than ARM's library.

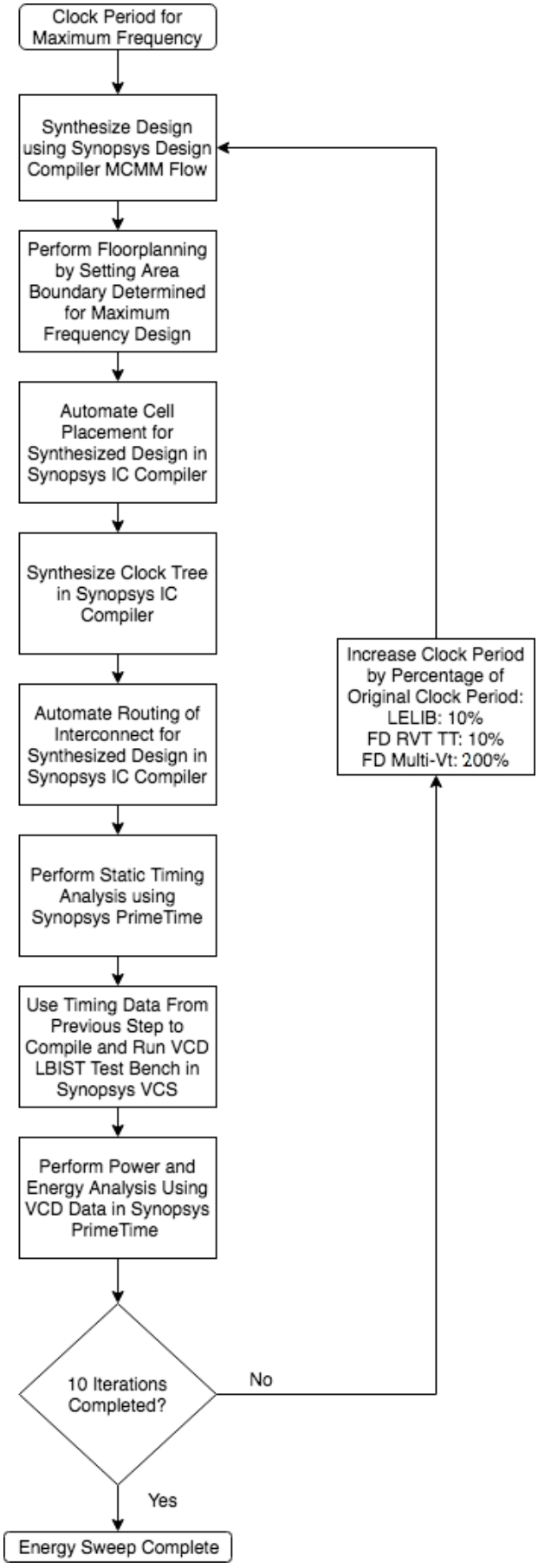


Figure 73: AES core minimum energy methodology

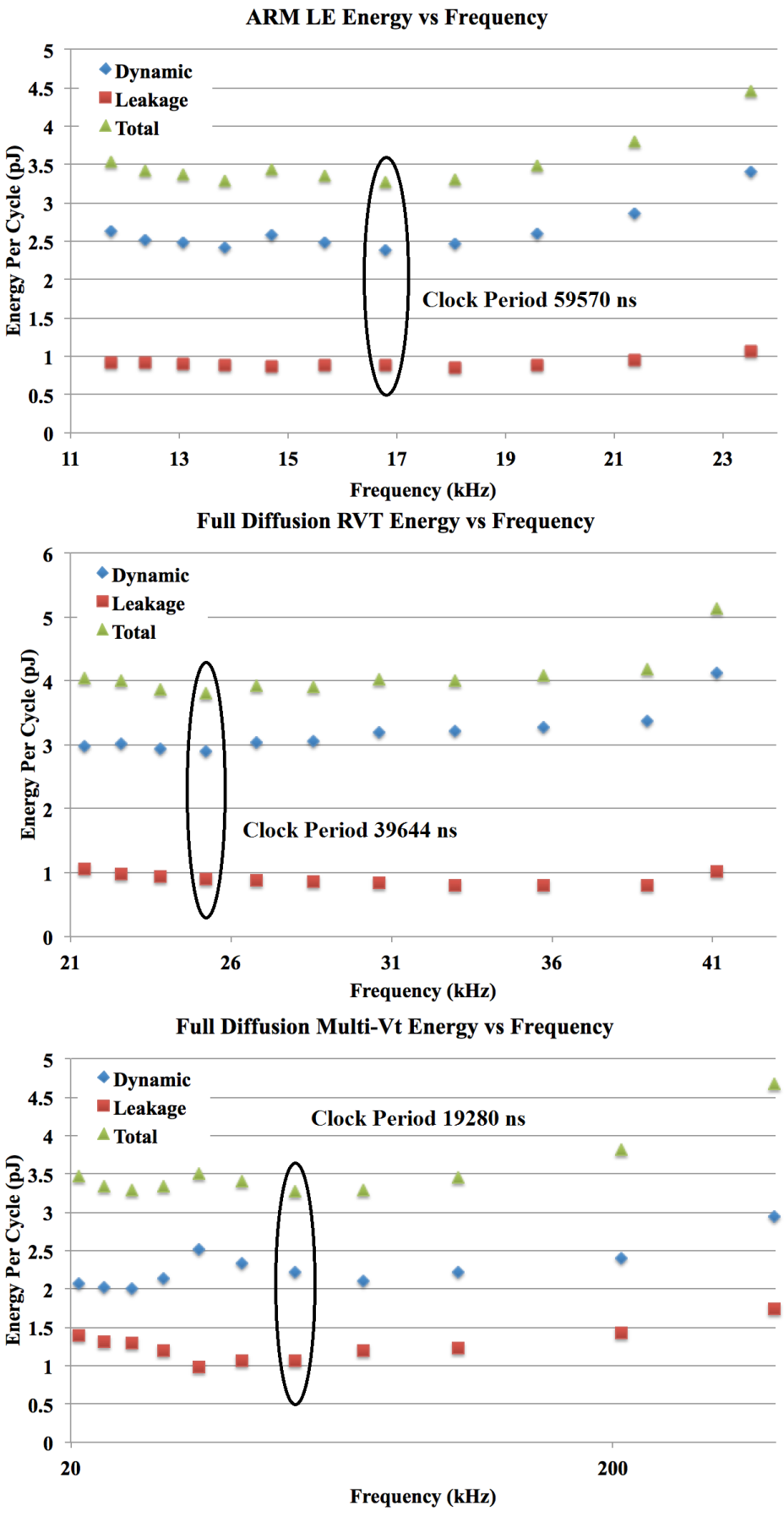


Figure 74: AES core minimum energy points

The behavior over the timing relaxation methodology follows the established principles. As the frequency was relaxed from maximum frequency, both dynamic and leakage energy reduced as the number of cells in the design reduced. The dynamic energy continued to reduce throughout the entire process. The leakage energy diminished as leakage recovery introduced slower less leaky cells into the faster paths. Eventually when this process was exhausted, the leakage energy began to rise as the time period cells leaked for increased.

The minimum energy implementation of the Full Diffusion Multi-Vt library was 8X slower than the maximum frequency implementation with a clock period of 19280 ns equating to a frequency of 51.87 KHz. The total energy per cycle at the minimum energy point was 3.28 pJ. This is almost identical to the ARM LE library. The dynamic and leakage energies followed the same pattern as the Full Diffusion RVT library with the exception of a local peak at 34.6 KHz. The underlying data showed an interesting difference between the maximum frequency and minimum energy implementations for the multi-Vt synthesis. The minimum energy implementation was constructed from only 4% LVT devices and 96% RVT devices. This is almost the exact opposite to the maximum frequency design.

A snapshot of the underlying data is shown in Table 16. These are the final designs signed off for tapeout. The underlying physics outlined in Chapter 2 and the preliminary simulation in Chapter 3 suggested that the full diffusion designs should have faired better against ARM's library. This leads to several possibilities. The first is that all aspects of the INWE and RSCE effects are not captured in the compact BSIM models. This is a possibility, as the physics suggest a greater deviation in the capacitance than was displayed in the geometric sweeps provided in Chapter 3. A second possibility is that information is lost during the characterization process of transforming the cell designs into liberty tables. This is also possible as the process depends on many different user defined variables including the capacitance and slew ranges. A final possibility is simply that the PrimeTime power analysis stage is unable to provide accurate power analysis for the cells.

Alternatively, the literature in the field and prior simulation could prove inaccurate. Accurate testing of a silicon device is the only way to discern which of the above is true. This is the goal of the next chapter.

Circuit Data	ARM LE	Full Diffusion RVT	Full Diffusion Multi-Vt
Composition	RVT (100%)	RVT (100%)	LVT (4%) / RVT (96%)
Clock Period	59570 ns	39644 ns	19280 ns
Clock Frequency	16.79 kHz	25.22 kHz	51.87 kHz
Total Energy Per Cycle	3.27 pJ	3.81 pJ	3.28 pJ
Area (μm^2)	45742	49812	49315
Total Cells	5626	7318	7244
Combinational	5094	6784	6715
Sequential	437	437	437
Clock Tree	95	97	92

Table 16: AES core synthesis overview

5.12 Chapter Summary

This chapter revisited the RSCE length optimization of the cells for the final libraries to ensure maximum advantage of the RSCE and INWE are achieved. These proved to be highly dependent on VT variants and process corner, varying from the variation limited 90 nm minimum up to a maximum of 230 nm. Sequential and supporting cells were created using the same design methodologies and each library characterized in a customer characterization flow using calculated liberty table values from representative SPICE test benches.

A 128 bit AES core with 32 bit datapath and LBIST was chosen as the test circuit based on ease of testing, proven robustness from prior implementation and RTL availability. Three library permutations were chosen based on the silicon area limitations on the tapeout chip; ARM Low Energy, Full Diffusion RVT and Full Diffusion Multi-Vt.

A full digital synthesis methodology was designed to determine the maximum frequency achievable for each library. Full static analysis was performed to produce reliable simulations of the critical paths. The Full Diffusion RVT/Multi-Vt runs showed maximum frequency improvements of around 1.8X and 20X respectively.

A second digital synthesis methodology was then designed to determine the minimum energy point. Additional stages to provide accurate switching activity by circuit simulation via the LBIST were added. VCD power analysis showed comparative minimum energy increases of 16.5% and 0.3% for the Full Diffusion RVT and Multi-Vt respectively for comparative speed improvements of 52% and 309% respectively

Chapter 6: AES Cores Results

6.1 Chapter Outline

The chapter proceeds by detailing several test methodologies used to measure the performance, energy and leakage of each core under different operating conditions. The first methodology describes measuring the nominal performance of a single sample chip at room temperature (20 °C). Results are then presented and the libraries compared. This is followed by temperature analysis at 0 °C and 85 °C. These results are then presented and the libraries compared. Finally a multi-chip variation study is performed with a random sample of 10 chips. The results of this analysis are then presented.

6.2 Nominal Operation Measurement

The same test board used to measure the ring oscillator experiment in Chapter 4 is also used to measure the AES cores. Figure 75 shows the test board with mbed control processor.

The same VESA controlled source meter and temperature chamber were also used.

An on-chip clock generator was used to drive the AES cores. This proved a limiting factor in the measurement. Although the clock generator was capable of providing a clock period up to 100 MHz, a single on chip interconnect was used to connect the clock to the AES cores. Clocked design with such a wide frequency range typically requires active design on interconnects to prevent unwanted ringing. Sadly, this manifested in the final design. Figure 76 shows oscilloscope measurements of the clock trace as the frequency is increased. At low frequencies, no ringing is visible. As the frequency is increased, overshoot and undershoot becomes visible, but the waveform is still functional as a clock signal. Eventually at high frequencies, the clock signal becomes so distorted it is not fit for purpose. This limited the measurement of the cores to below a VDD of 500 mV where the operating frequencies were small enough to allow functional operation with a high degree of confidence in the results. Fortunately, this region of the supply voltage is the subthreshold region where the Minimum Energy Point (MEP) resides. The quantization of the clock is 1 kHz. Whilst this only proves a factor for the slower cores at extremely low VDD, it must still be considered a source of minor inaccuracy.

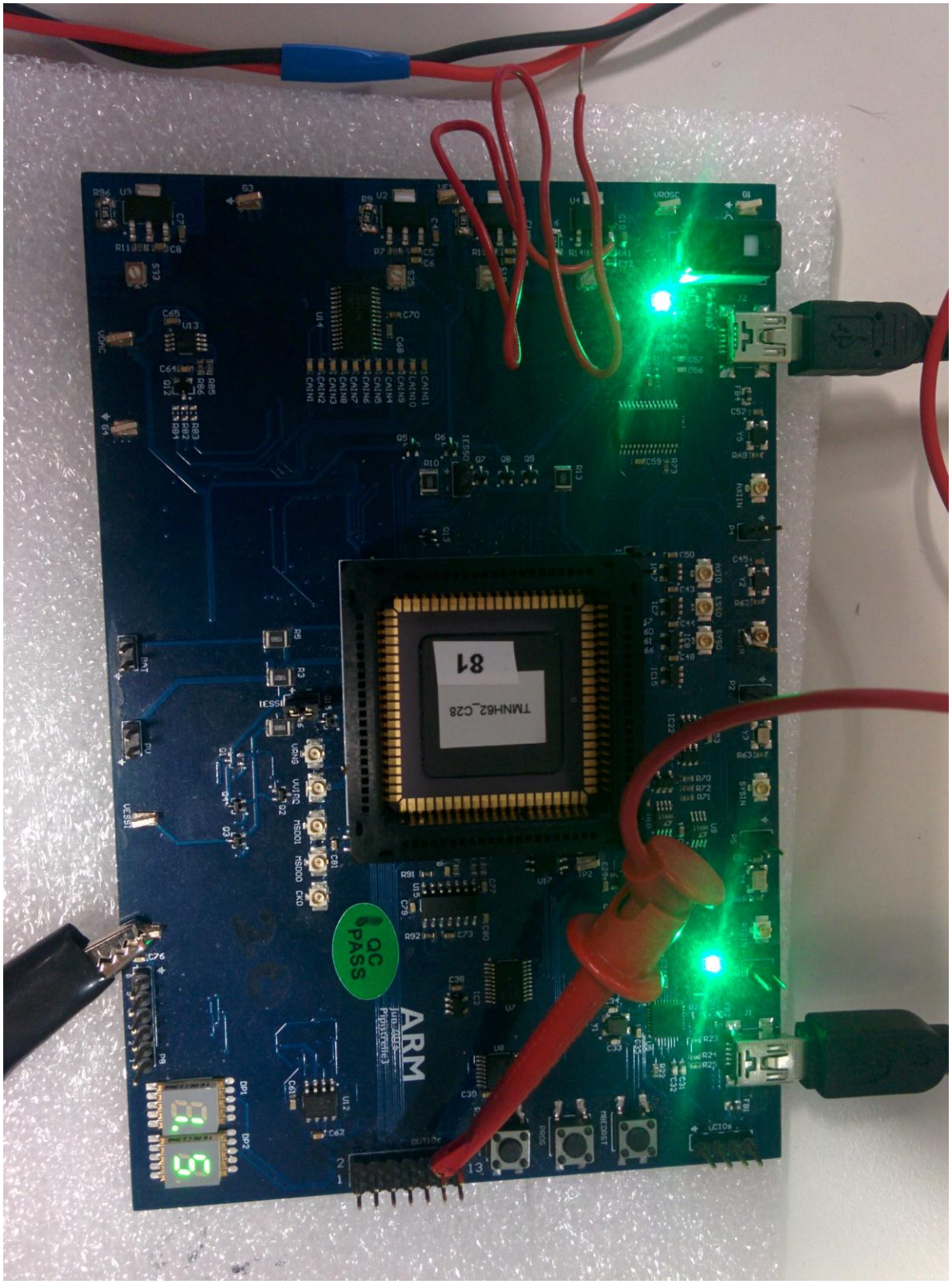


Figure 75: AES core test board

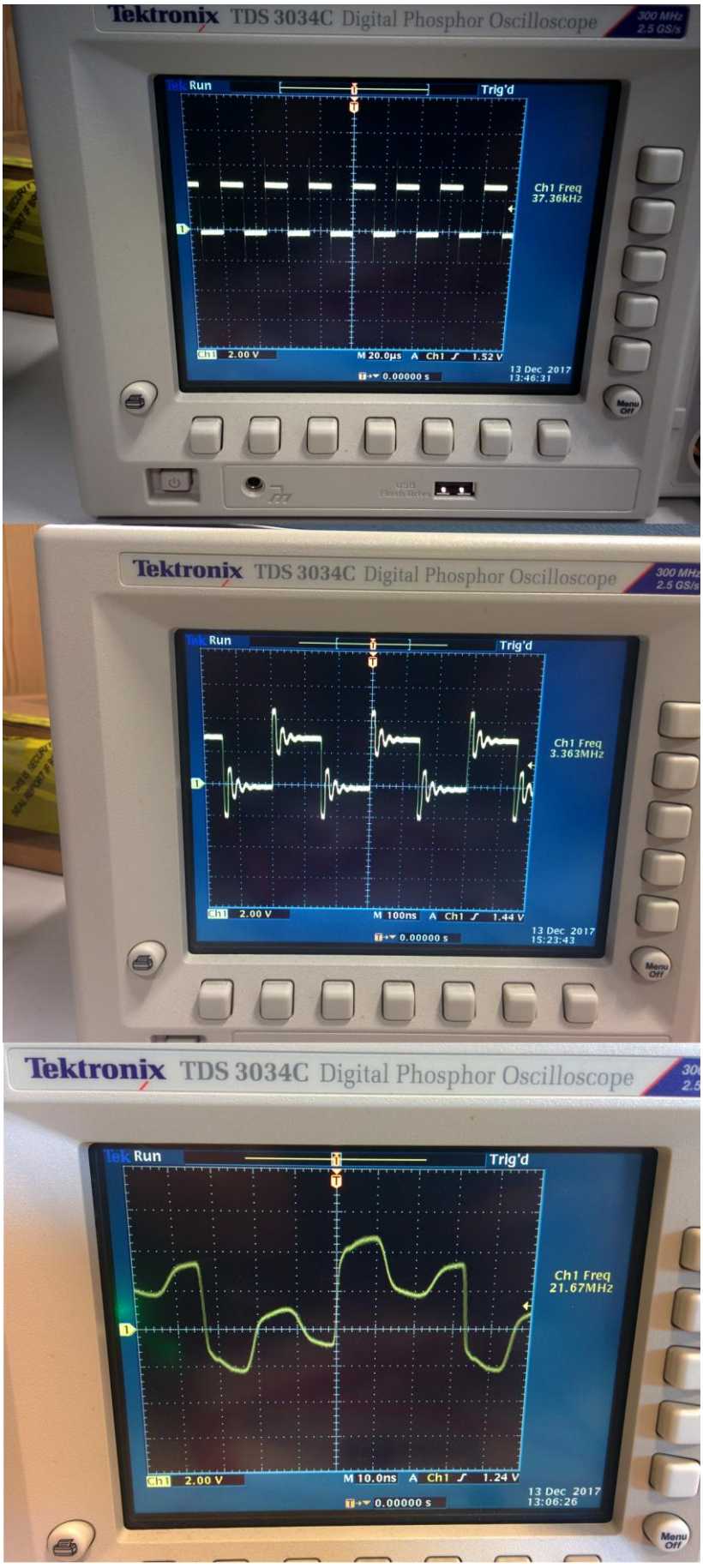


Figure 76: AES core clock signal degradation

Python programs were written to control the test chip via the mbed control processor similar to the ring oscillator experiment. The first program was designed to determine the maximum operating frequency of each core at each supply voltage point using a binary search algorithm. Figure 77 shows the sequence of operation. For the nominal test runs, the board was placed into the temperature chamber and the temperature set to 20 °C. The on board temperature sensor was polled until the temperature stabilized.

Initially, the communication protocols to the board were set up. Initial values were then given to the program parameters. The maximum and minimum frequencies were defined as 66 MHz and 1 kHz respectively. These set the upper and lower bounds between which the binary search algorithm attempts to find the maximum operating frequency. The first test frequency was set to the midway point between these limits. The maximum and minimum supply voltages were defined as 500 mV and 150 mV respectively. These defined the supply voltage range over which the cores were be swept in 10 mV steps, starting with the highest voltage.

The first VDD was set via the source meter. The clock generator was then set up using the initial frequency. The on chip logic was then set to the desired state. The core under test was selected and the scan enable line disabled. The LBIST provides two modes of operation, continuous and single run. For determining the maximum frequency, the single run mode was selected. A test vector was then loaded into the LBIST. The test vector 0x3BCD was used for all runs. The core was then held in reset by lowering the reset line. The reset was then disabled and the single encryption/decryption cycle run. The encryption/decryption process required around 130 clock cycles. A wait time of 1 second was therefore adequate before polling the result. The LBIST handled checking the decrypted vector to see whether it matched the initial vector and provided a single pass or fail result.

If the run passed, the algorithm eliminates the half of the frequency range below the tested frequency by setting the minimum frequency to the test frequency. If the run failed, the algorithm eliminates the half of the frequency range above the test frequency by setting the maximum frequency to the test frequency. The new mid point was then calculated and the test performed at the new frequency.

This iterated until one of two conditions were determined. The first was that a successful run falls within a 1% convergence of either the maximum or minimum frequency values. At this point, the maximum frequency of operation had been found. This introduced a

convergence error of maximum 1% deviation from the true value, but drastically reduced test time.
The second condition was a failure at the initial minimum frequency value (1 kHz). If the run failed at this value, the core simply didn't function at this supply voltage point or below and therefore the run for that core was ended.
Once the maximum frequency of operation was determined, a second test run was performed at this frequency. The on chip logic was set to select the same core and keep the scan line disabled. This time the run mode was set to continuous, which simply performed the encryption/decryption cycle perpetually. The same test vector was written and the core held in reset. The reset was then released. As the core repeatedly performed the operation, the source meter was instructed to read the average current draw. The program used this to calculate the power and energy consumption.
Finally the program checks if all supply voltage points have been performed. If not, the 10 mV step was removed from the supply voltage and the binary search algorithm began anew. If so, the run terminated.

A second program was then used to perform the leakage measurement. These programs were separated, as the leakage measurement had no need to know the maximum operating frequency. Figure 78 shows the sequence of operation.
The maximum and minimum supply voltage parameters were initialized with the same values as the previous program. The core VDD was set using the source meter. The on chip logic was then set to select the core under test and disable the scan line. The run mode was set to single cycle, the test vector written and the reset line disabled such that the core was positioned to perform the encryption/decryption cycle. However, for this test the clock signal was disabled. This ensures the current measurement was performed on representative inactive logic.
A small wait time of a second was performed to allow the voltage values within the logic to equalize. The current measurement was then performed. This process was then iterated across the full supply voltage range.

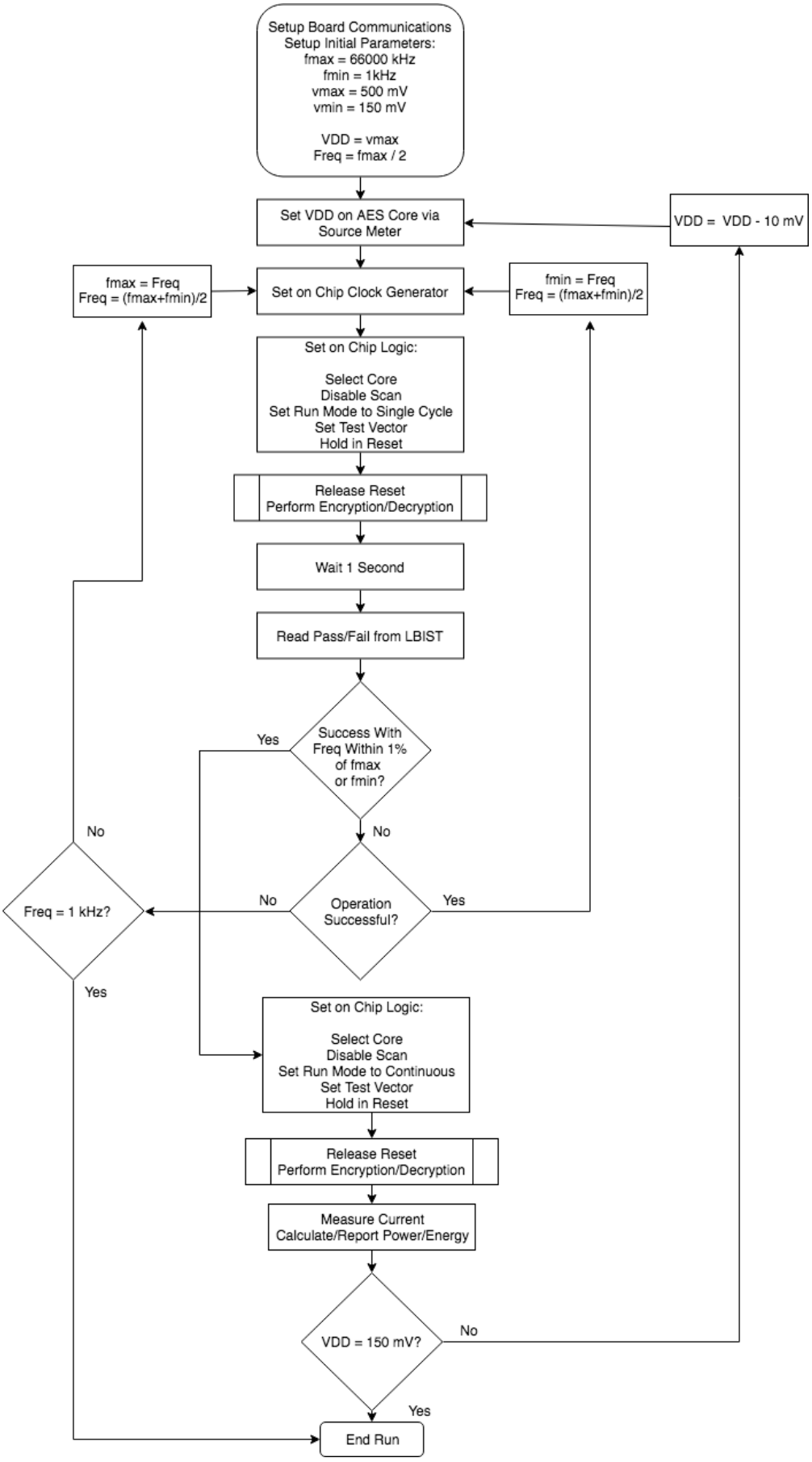


Figure 77: AES core maximum operating frequency test methodology

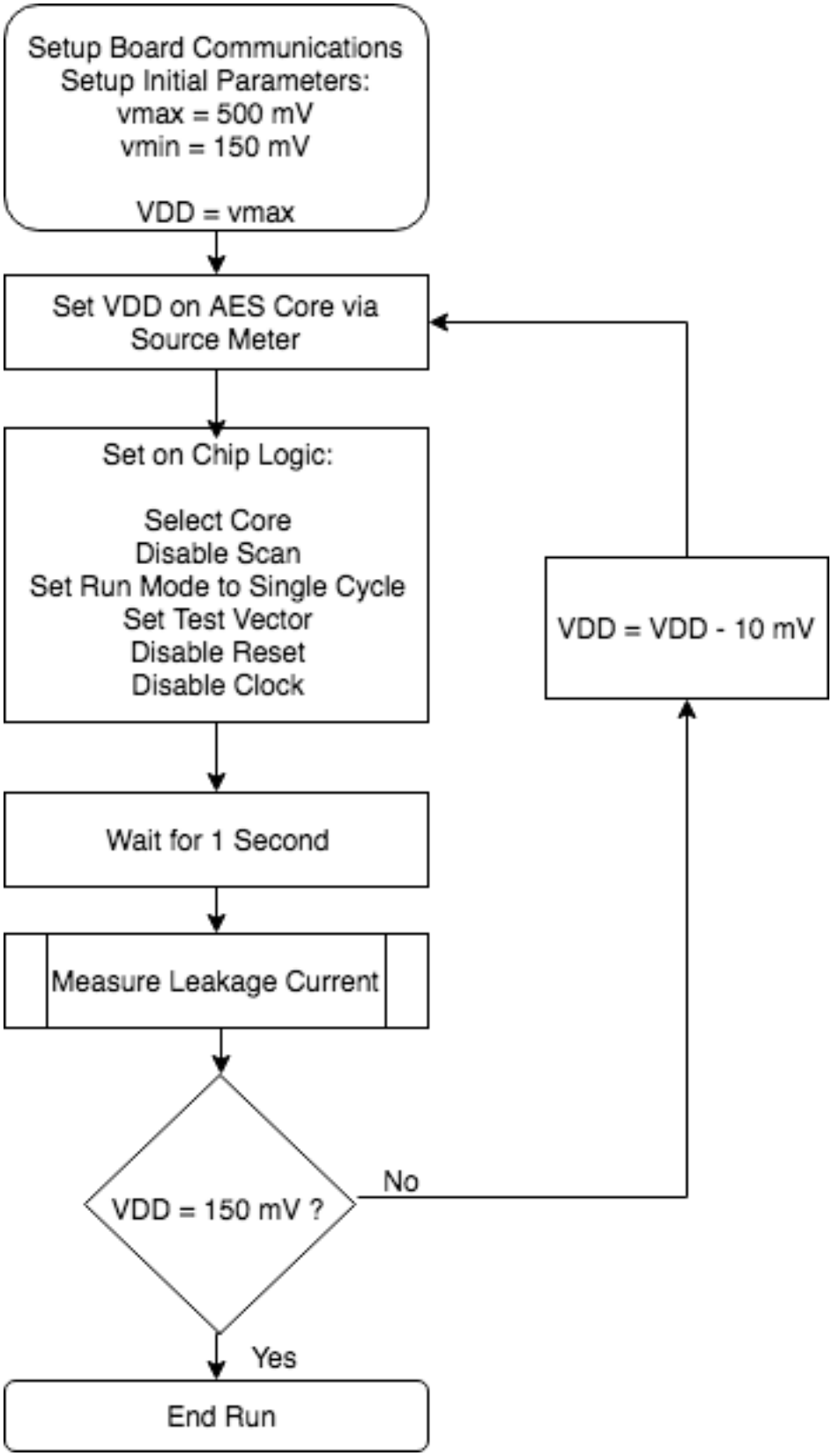


Figure 78: AES core leakage test methodology

6.3 Nominal Operation Results

Figure 79 shows the nominal operation results. Initial observation shows that the VCD power analysis performed in the previous chapter underestimated the energy consumption of the ARM Low Energy Library. However, it was surprisingly accurate at estimating the Full Diffusion Libraries. Conversely, the static timing analysis performed in the previous chapter proved surprisingly accurate of the ARM LE library, but massively underestimated the attainable frequencies of the Full Diffusion libraries. These two effects combined created analysis that underestimated the potential of the Full Diffusion sizing strategy overall.

At 20 °C, the Full Diffusion RVT library had an energy-per-cycle of 3.84 pJ, a 7% saving on the ARM LE library. The Full Diffusion Multi-Vt library had an energy-per-cycle of 3.11 pJ, a 24% saving on the ARM library. The MEPs of the Full Diffusion libraries were also observed at higher supply voltages. There are several advantages to this. The first is that variation is voltage dependent and therefore operating at a higher supply voltage helps counteract negative effects induced by variation. The second is that when implementing Ultra Wide Dynamic Voltage Scaling, the energy savings made in voltage scaling to the MEP are often offset by losses in the DC to DC conversion, which can have voltage conversion efficiencies as low as 70-80% [96]. These efficiencies are a factor of the distance between the DC voltages being converted. Therefore if the MEP is situated at a higher supply voltage, this helps reduce losses in the voltage conversion by aiding efficient DC-to-DC conversion.

The frequencies of the Full Diffusion RVT and Multi-Vt libraries were 34 kHz and 147 kHz respectively, representing speed advantages of 2X and 8.65X. Therefore, not only were the Full Diffusion libraries more energy efficient, they were also considerably faster.

The energy per frequency figures for the three cores were therefore 0.893 pJ/kHz, 0.459 pJ/kHz and 0.07 pJ/kHz for the ARM LE, Full Diffusion RVT and Full Diffusion Multi-Vt libraries respectively. This represents a reduction of 49% and 92% respectively.

It is very interesting that exchanging only 4% of the gates for LVT variants results in such a drastic change in the energy and frequency profiles of the design. This highlights the advantage of the characteristic range provided by the Full Diffusion sizing strategy. The discussion from Chapter 3 showed that the Full Diffusion sizing strategy was the leakiest strategy of the three presented. Moreover, the LVT device variants have Ion/Ioff ratios almost half that of their RVT counterparts (approximately 1500 to 3000). This was

shown in the energy profile as the MEP pushed higher up the supply voltage range as the leakage contribution increased. However, as the design only used these sparingly, the ability to improve the frequency by a factor of 8.65X offsets those losses to form an advantage. Understanding how the synthesis tools makes these choices during synthesis enabled a superior design to be constructed from technically inferior cells. This is analogous to losing the design battle of the individual cells to win the war of the overall synthesis of the circuit by trading the speed benefit of the cells for an energy benefit in the synthesized circuit.

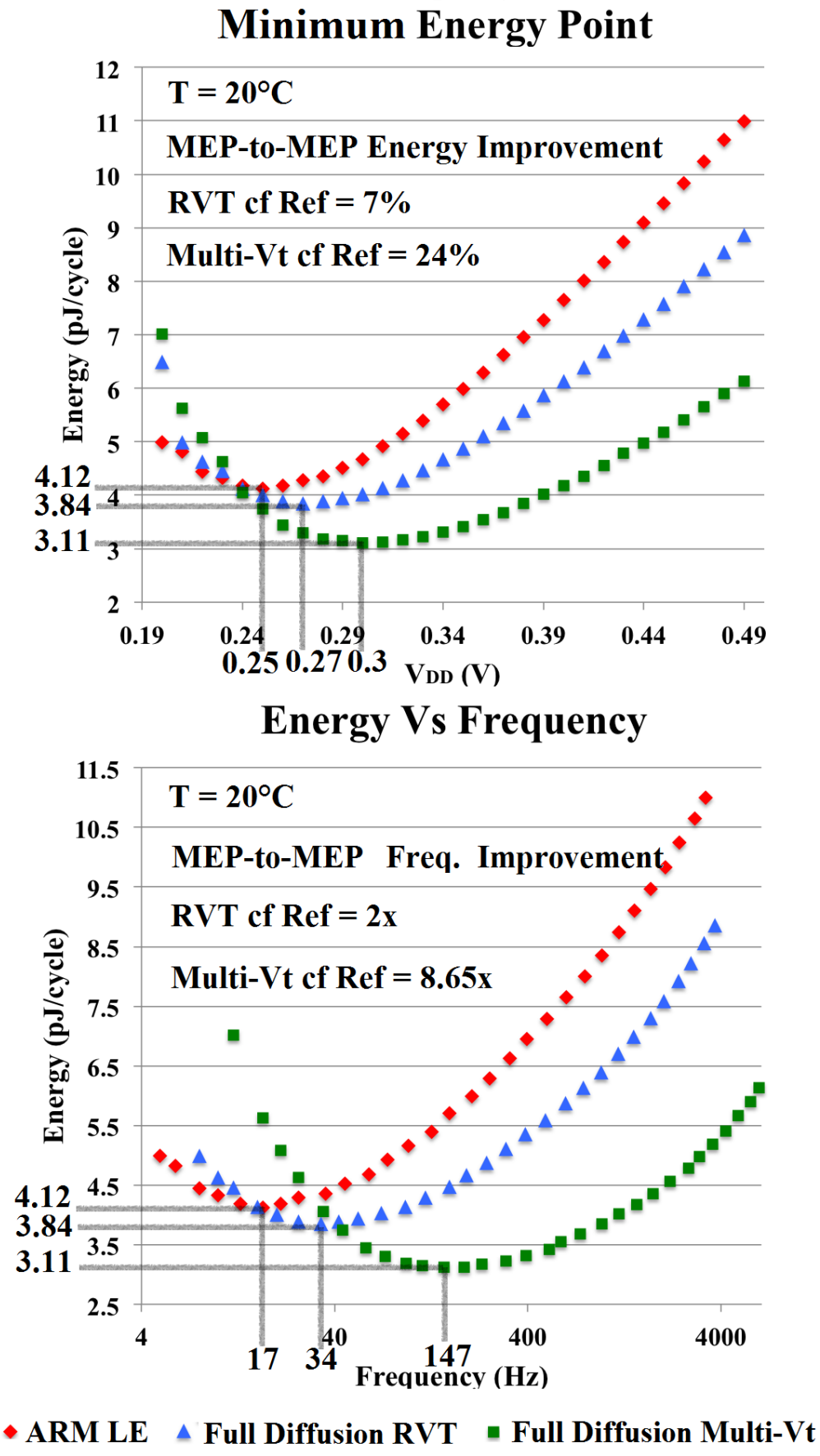


Figure 79: AES core nominal operation results

6.4 Temperature Analysis

The same two test procedures were performed again in the temperature chamber at 0 °C and 85 °C. These values were chosen as they represent the extremes of what the test board can handle without permanent damage. The purpose of the experiment was to determine whether the benefits of the Full Diffusion sizing strategy exhibited in the nominal case extend across a large temperature range. Figure 80 shows the results. Initial observation shows that the same trends were displayed at both temperature extremes. The Full Diffusion RVT library had an MEP consistently lower in energy-per-cycle and higher in frequency than ARM's Low Energy library. The Full Diffusion Multi-Vt library consistently had the lowest energy per cycle at MEP and the highest frequency. All libraries display temperature inversion (operating faster with a higher energy consumption at the higher temperature point) due to subthreshold operation.

Across the entire temperature range, the maximum comparative frequency gain was 10.25X between the Full Diffusion Multi-Vt library and ARM's LE library, MEP to MEP, at 0 °C. This suggests the additional temperature stability at low temperatures that was demonstrated in the ring oscillator discussion extends to more complex circuits. The maximum comparative energy-per-cycle gain was 27.3% between the Full Diffusion Multi-Vt library and ARM's LE library, MEP to MEP, at 85 °C. These equate to savings in energy per frequency of 92% and 80% respectively.

All of these gains are considerable given that no additional effort is required during synthesis beyond switching out the library data.

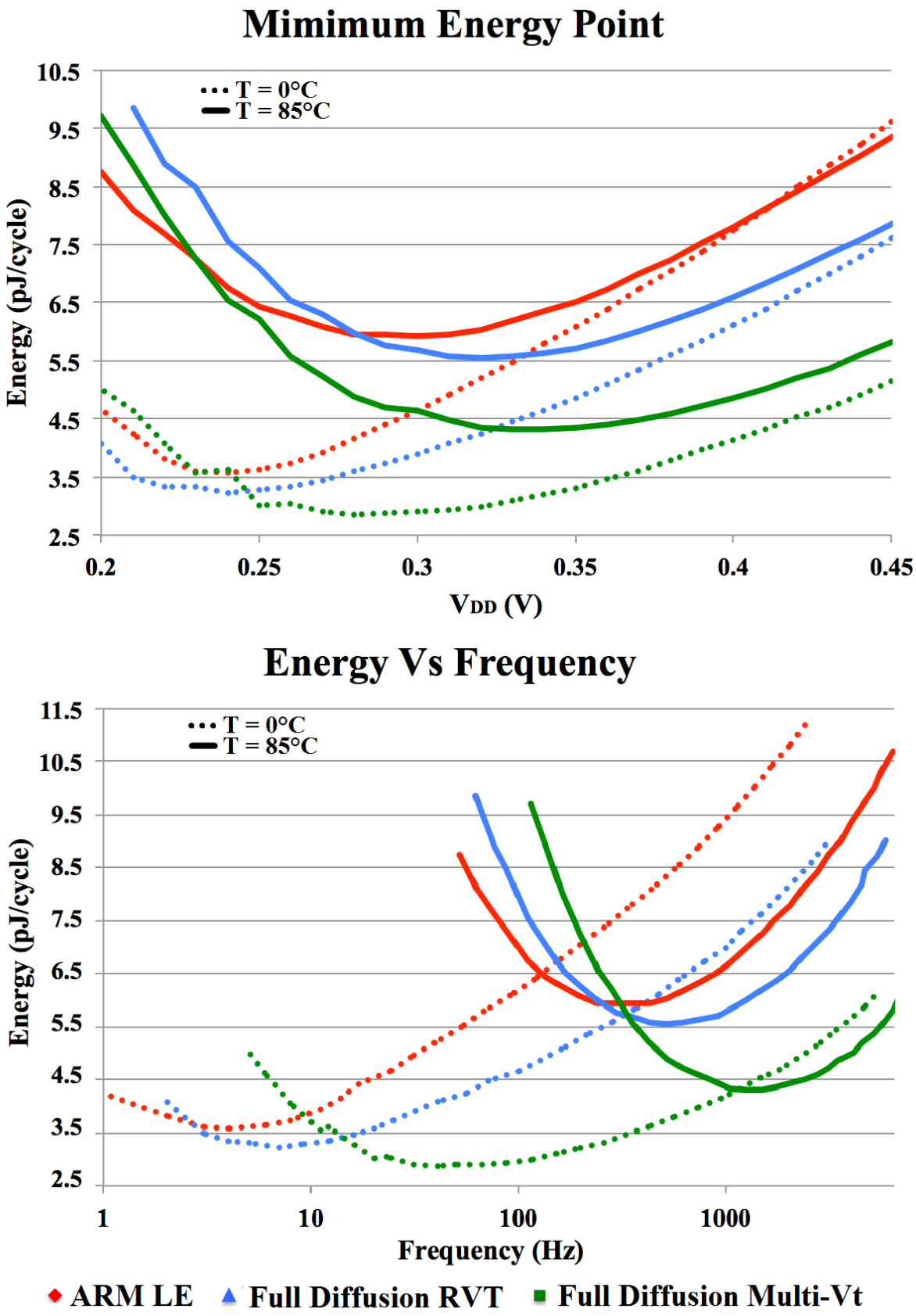


Figure 80: AES core temperature analysis results

6.5 Variation Analysis

A random sample of 10 test chips was performed using the nominal test methodology (chamber temperature of 20 °C) to determine the impact of inter-die variations on the cores. Figure 81 shows the results. All trends showed tighter groupings of frequency and energy characteristics that began to spread as the frequency and therefore supply voltage was reduced. This followed the correct variation trend. Importantly, the results showed that the nominal and temperature analysis of the previous subsections were representative of the overall fabrication run.

Table 17 shows the measured statistical metrics from the experiment, along with a recap of the key data extracted for all three experiments . The results showed a comparative increase in frequency variation of 53% and 32% for the Full Diffusion RVT and Multi-Vt libraries respectively. The discussion in Chapter 4 indicated that an increase in variation on the RVT cell level of up to 28% could be expected. This accounts for some of the additional variation. The remaining additional variation is likely a result of the logical decomposition of the larger ARM LE cells into combinations of smaller Full Diffusion cells as the synthesis tool attempted to recover leakage from the Full Diffusion synthesis runs. The comparatively lower variation of the Full Diffusion Multi-Vt run was likely a combination of the inherently less variable LVT cells and the lower activity factor of the design. The 4 finger Full Diffusion cell was determined to be the least variable of all LVT designs in the analysis of Chapter 4.

Interestingly, the maximum deviation in frequency from the mean does not follow the statistical sigma/mu metric. The worst case deviation from the mean for ARM's LE library was 7.7% whilst this was only 5.9% for the Full Diffusion RVT library. This suggests that although the statistical probability of a circuit performance close to the mean is higher for the ARM LE library, the performance of the outlier circuits varies greater, suggesting a deviation of the Gaussian distribution [97]. The maximum deviation from mean of the Full Diffusion Multi-Vt library was larger at 9.1%. As the LVT and RVT implantation stages are performed separately during fabrication, there is no correlation between global systematic variation. This is likely the source of this increase. Comparatively, the statistical variation of the leakage current was 29% larger for the Full Diffusion RVT library but 21% smaller for the Multi-Vt library. These values are likely due to the parametric RDF variations in the underlying cells.

Finally, the comparative statistical variation in energy-per-cycle was 115% and 54% higher for the Full Diffusion RVT and Multi-Vt libraries respectively. These results show

that the performance and leakage variations outlined above combine to create quite a large deviation in the energy per cycle. This is the largest trade off for the mean lower energy per cycle value.

A final observation must be made regarding this experiment. It was speculated in Chapters 2, 3 and 4 that the Full Diffusion sizing strategy may negatively affect the minimum operating voltage. Individual device sizing and the P to N ratio primarily affect this metric. As the Full Diffusion sizing strategy pushes these away from the ideal value, it was speculated that the minimum operating voltage would be higher. This observation proved to be accurate as technically the Full Diffusion RVT and Multi-Vt runs stopped functioning at higher supply voltages. However, this metric must be taken in context. Both Full Diffusion libraries also pushed the MEP higher up the supply voltage range. Therefore, the actual response cannot be regarded as a degradation in the minimum operating voltage, but actually as a 'slide-to-the-right' of the entire energy curve. Important to note is that no core tested during this experiment failed to operate at its minimum energy point.

6.6 Chapter Summary

This chapter explained the test procedure for measuring the nominal operation of each AES core. MEP to MEP, the Full Diffusion RVT library proved 7% more energy efficient and 2X faster than the baseline ARM library. The Full Diffusion Multi-Vt library proved 24% more energy efficient and 8.65X faster.

The test procedure for the temperature analysis was then described. The analysis showed the Full Diffusion libraries were always faster and more energy efficient over the full temperature range tested. The highest comparative energy efficiency measured was 27.3% at 85 °C for the Full Diffusion Multi-Vt library. The highest comparative frequency gain measured was 10.25X for the Full Diffusion Multi-Vt library.

Finally, the test procedure for the multi-chip variation analysis was described. The results from a 10 chip sample pool showed that the variation for the Full Diffusion libraries was greater, but that the outlier results did not necessarily correlate with the Gaussian distribution.

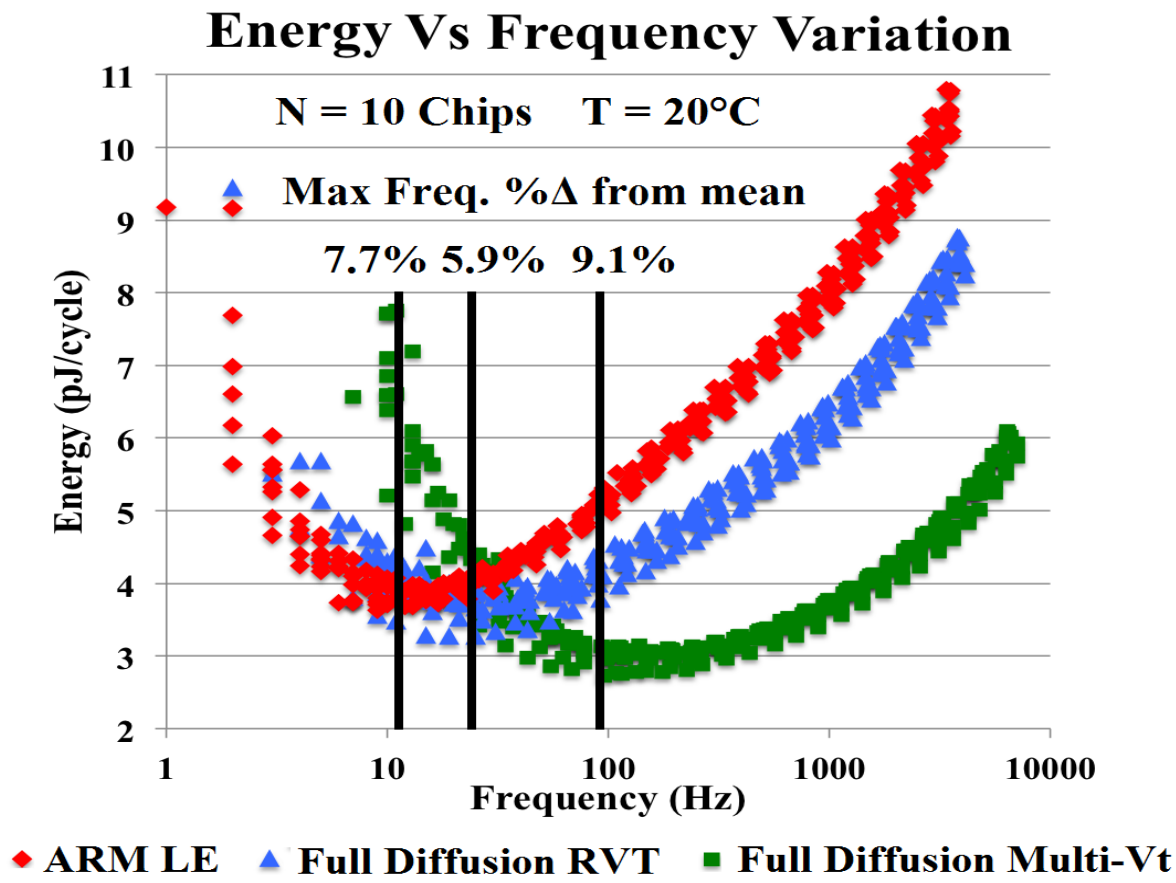


Figure 81: AES core variation analysis results

	Reference (RVT)	Proposed RVT	Proposed Multi-Vt
Circuit Data			
Composition	REF RVT (100%)	RVT (100%)	LVT (4%) / RVT (96%)
Area (μM²)	45742	49812	49315
Total Cells	5626	7318	7244
Combinational	5094	6784	6715
Sequential	437	437	437
Clock Tree	95	97	92
Measured Data			
@ 0°C			
Minimum Energy Point	240mV	240mV	280mV
Frequency @ MEP	4kHz	7kHz (1.75X Faster)	41kHz (10.25X Faster)
Energy @ MEP	3.57 pJ/cycle	3.21 pJ/Cycle (10% Less)	2.85 pJ/cycle (20% Less)
Energy per Frequency	0.893 pJ/kHz	0.459 pJ/kHz (49% Less)	0.07 pJ/kHz (92% Less)
@ 20°C			
Minimum Energy Point	250mV	270mV	300mV
Frequency @ MEP	17kHz	34kHz (2X Faster)	147kHz (8.65X Faster)
Energy @ MEP	4.12 pJ/cycle	3.84 pJ/cycle (7% Less)	3.11 pJ/cycle (24% Less)
Energy per Frequency	0.242 pJ/kHz	0.113 pJ/kHz (53% Less)	0.021 pJ/kHz (91% Less)
@ 85°C			
Minimum Energy Point	300mV	320mV	330mV
Frequency @ MEP	350kHz	517kHz (1.48X Faster)	1.26 MHz (3.6X Faster)
Energy @ MEP	5.93 pJ/cycle	5.55 pJ/cycle (6.4% Less)	4.31 pJ/cycle (27.3% Less)
Energy per Frequency	0.017 pJ/kHz	0.011 pJ/kHz (35% Less)	0.0034 pJ/kHz (80% Less)

Table 17: AES core results summary

Chapter 7: Discussions

7.1 Chapter Outline

The chapter proceeds by addressing the two hypotheses postulated in the thesis. The first hypothesis presented in Chapter 1 is addressed by drawing on the criteria outlined in Chapter 2. Namely, did the proposed Full Diffusion sizing strategy utilize the underlying subthreshold physics to create a superior subthreshold standard cell library based on robustness, variability and performance. The second hypothesis presented in Chapter 4 is addressed by revisiting the AES results presented in Chapter 6. The chapter then provides a comparison to prior art and concludes with a discussion on the constructive and limiting aspects of the work.

7.2 Hypothesis One

The first hypothesis presented in the thesis posed:

Can a superior subthreshold standard cell library be created from devices that take advantage of the underlying subthreshold semiconductor physics?

Superiority infers that one thing may be compared against another, using one or more metric/s of interest, to determine which holds an advantage for a given application. Chapters 3 and 4 explored the entire design space for how standard cells can be designed for a given library size (in this case 12 Track) and technology node (in this case 65nm bulk planar). Three distinct sizing strategies emerged. Two had previously been explored in the field; minimum sizing and regular (superthreshold) sizing. The third was a novel contribution named the Full Diffusion sizing strategy.
Chapter 2 systematically presented a set of criteria against which these sizing strategies may be compared. This comparison is therefore presented to determine how the novel subthreshold sizing strategy eventually pursued in the thesis may or may not be deemed superior to the other two sizing strategies.

7.3 Robustness

7.3.1 Carrier Injection

Section 2.3.1 described several forms of carrier injection mechanisms. The first was carrier hot electron (CHE), a mechanism by which electrons accelerated by the

longitudinal electric field in the device are deflected upwards into the oxide by scattering mechanisms associated with the crystalline lattice structure. CHE is therefore proportional to V_{DS} (which generates the longitudinal electric field) and temperature (which affects the lattice scattering). It could be argued that correct RSCE-aware length sizing in the subthreshold regime reduces ionic scattering as the dopant density in the center of the channel is reduced, therefore reducing the probability of a collision, increasing the probability of electrons attaining the required momentum for CHE and decreasing robustness. However, the studies presented in Chapter 2 showed negligible CHE below a V_{DS} of 1.4V. Therefore the performance gain from RSCE outweighs the potential increase in lifetime degradation. All of the three sizing strategies, if used within the subthreshold regime, should be sized at the RSCE optimal length.
None of the three sizing strategies presented have any effect on the longitudinal electric field. Therefore, no sizing strategy can be deemed superior to another based on this form of robustness.

7.3.2 CHISEL

Section 2.3.2 described a secondary carrier hot injection mechanism called Channel Initiated Substrate Electron Injection (CHISEL). This degradation mechanism was found to be proportional to the transverse field (gate to body) of the device, and caused by carriers resulting from impact ionization gaining sufficient momentum in the vertical device dimension to surmount the Si-SiO_2 interface. As discussed in the INWE derivation of Section 2.5.5, the INWE takes advantage of the thick gate oxide overlap requirement to increase the transverse field by the intentional introduction of the fringing field. Both the minimum sizing strategy and full diffusion sizing strategy take advantage of the INWE to produce cells of either outright performance superiority or characteristic ranges of cells. It could therefore be argued that based on this metric, the regular sizing strategy is superior, followed by the full diffusion sizing strategy and finally the minimum sizing strategy. It must be noted however, that CHISEL produces almost negligible degradation in the subthreshold domain. Moreover, the full diffusion sizing strategy contains a single cell variant for all combinational cells that contain no INWE modification. Therefore, should the full diffusion library be chosen, the synthesis tool could be directed to only optimize for robustness and use only non-INWE optimized cells. The same cannot be said for the minimum sizing strategy that only contains cells of maximum INWE optimization.

7.3.3 Time Dependent Dielectric Breakdown (TDDB)

Section 2.3.3 described a breakdown mechanism associated with the generation of oxide defect sites as a result of direct quantum tunneling through the gate oxide. These form stochastically in the oxide and therefore breakdown of the oxide occurs at a critical defect density. The model presented indicted that the time to breakdown therefore depended on the transverse field and the area of the device. Both of these are manipulated to take advantage of the INWE in the subthreshold regime for the minimum and full diffusion sizing strategies. The previous subsection discussed the impact on the transverse field, which also holds true for this failure mechanism. As the spacers are introduced to induce the INWE, the transverse field increases as a result of the fringing field, increasing the defect site generation and increasing TDDB degradation.

Conversely, this same operation reduces the area of the gate. As the gate area is reduced, the statistical probability of a percolation path forming is reduced; therefore the probability of oxide breakdown is also reduced, increasing the expected lifetime of the device.

The overall effect on TDDB is therefore unclear and is likely technology node dependent. Should the impact of the transverse field dominate, the regular sizing strategy would be superior. Should area reduction dominate, the minimum sizing strategy would be superior.

Should the full diffusion sizing strategy be chosen, the synthesis tool could be instructed to favor either option and therefore provides an excellent compromise.

The studies presented in Section 2.3.3 only showed meaningful degradation down to 1.2V. Therefore operating solely in the subthreshold regime for the chosen technology node, TDDB is of little concern. For bulk planar technology nodes 45nm or less, gate leakage becomes orders of magnitude more important; therefore this degradation mechanism may be of more concern should a higher level of scaling be required for a design. However, few energy aware designs for IOT application venture this deep into submicron bulk planar technologies due to the rapid increase in leakage current below 65nm.

7.3.4 Line Depletion Electromigration

Section 2.3.4.2.1 outlined how insufficient conductor allocation can result in the drift of copper atoms in the interconnect, leading to a failure mechanism known as line depletion electromigration. For self-heating (Joule) systems, an increase in operating temperature of

only 5% was shown to result in a 30% reduction in expected lifetime. Whilst strict operation in the subthreshold regime eliminates joule heating, operation in voltage scaling schemes such as ultra wide dynamic voltage scaling may be of concern.
As diffusion is reduced and replaced for STI spacers, the requirement to connect all diffusion areas to the correct nets becomes more complex and the via density becomes greater. In order to use the full diffusion area available in the 12 track library in the given technology, the power rails had to be scaled to around a third of their initial size. This is not an issue in the subthreshold regime therefore no sizing strategy could be deemed superior. In voltage scaling schemes, special consideration must be added during synthesis to ensure that the number of consecutive cells sharing the same rail does not exceed the level whereby joule heating becomes a factor. Whilst this does mean added complexity during the synthesis and placement stages of circuit design, the effects of line depletion electromigration can be mitigated. Therefore following the correct procedures removes this degradation mechanism as an area of concern.

7.3.5 Via Depletion Electromigration

Section 2.3.4.2.2 described how copper vias in deep submicron bulk planar technologies are lined to prevent cooper migration and ultimately via failure. Two preventative measures were discussed; interconnect overhang and via redundancy.
Interconnect overhang is a minimum extension of copper interconnect beyond the via on the upper interconnect layer. This overhang reduces the fill angle during the liner deposition and creates a superior via liner, reducing processing errors and time to failure. The DRM for the chosen technology node recommends an interconnect overhang of at least 40nm on two opposing sides of a via. All cells in the Full Diffusion library abide by this layout rule, sometimes with dedicated T structures to maintain beneficial inter route spacing. However, so do the cells from remaining two sizing strategies. Therefore, no strategy can be deemed superior.
Via redundancy is the process of connecting diffusion areas to nets through multiple vias. If a single via fails, this leaves functioning vias that are still able to provide the net connection. For large diffusion areas (as in the regular sizing strategy), multiple vias are possible and highly recommended. For the reasons discussed in Section 2.3.4.2.2, vias are always recommended at the minimum aspect ratio (90 x 90 nm in the chosen technology). As well as lowering variation through etchant rate similarity, this also allows the technology node to specify the resistance associated with each via (8 Ohms in the chosen

technology). The use of multiple vias is therefore analogous to multiple resistances in parallel, with multiple vias reducing the RC parasitics of net connections. If a via in a multiple via connection fails, the gate would suffer an increase in RC parasitics but the underlying transistors would not suffer any alteration.

For smaller diffusion areas (as in the minimum sizing strategy), the via-to-via spacing rules in the technology prohibit multiple vias in the transistor-to-transistor dimension. However, each device is essentially split into multiple devices. Therefore, in the event of a via failure, the result would be the removal of one of the transistors in the multiple transistor device. This impacts the underlying strengths of the pull up / pull down networks, which would degrade the output swing of the overall gate, leading to an increase in probability of logical failure. This is more serious than a simple change in the RC parasitics of the gate.

The Full Diffusion sizing strategy falls in between these two options. The lower fingered gates (1F/2F) allow multiple vias, the higher fingered gates (3F/4F) do not. It could therefore be argued on an iso-area via depletion robustness basis, the superior sizing strategy is the regular sizing strategy, followed by the full diffusion sizing strategy with some via redundancy and minimum sizing strategy with no via redundancy. On a non-iso-area basis, no sizing strategy holds the advantage as the widths of the cells could simply be increased to allow via redundancy on all cells in the non transistor-to-transistor dimension. Moreover, as technology nodes mature, the iterative refinement eventually reaches the point where via failure becomes insignificant.

7.3.6 Negative Bias Temperature Instability (NBTI)

Section 2.3.5 described the degradation mechanism by which the device threshold voltage and carrier mobility are negatively affected as tied off dangling bonds in the interface structure are broken in response to large negative biasing and high temperatures. This is therefore known as negative bias temperature instability. The models presented showed that the mechanism is dependent on how far the freed hydrogen diffuses away from the interface, preventing the re-annealing of the bonds. This in turn is dependent on the width of the device. For larger width devices, the hydrogen diffuses at angles which keep the free hydrogen closer to the interface, allowing the hydrogen re-annealing process to extend further into the lifetime of the device. For small width devices, the hydrogen is forced to diffuse away from the interface at a narrower angle closer to orthogonal. This results in the hydrogen diffusing further away from the interface, removing the hydrogen

from the possibility of being re-annealed and therefore increasing the effect of NBTI over the lifetime of the device.

In terms of sizing strategy superiority, the regular sizing strategy always uses the maximum width device and therefore suffers least from NBTI. The Minimum width sizing strategy always uses the minimum width device and therefore suffers most. The Full Diffusion sizing strategy falls in between these two, depending on the composition of the cells chosen. Studies presented in Section 2.3.5 showed that the onset of NBTI at 300K is around 3V, and therefore this effect is virtually non-existent in targeted subthreshold circuits. However, if voltage scaling is used and the circuit is operated for extended periods of time in elevated temperatures, NBTI may become a design factor.

7.3.7 Robustness Overview

In terms of robustness, the sizing strategy favored to reduce the degradation mechanisms outlined in this subsection is the regular sizing strategy. Maximum width devices have the lowest INWE influences and therefore the transverse field is not increased which leads to an increase in the probability of many of the degradation mechanisms occurring. The maximum width also affords adequate area for via redundancy and widens the diffusion angle during NBTI. The only negative implication of using the largest area device is that it increases the probability of a percolation path and therefore the resultant TDDB breakdown.

The sizing strategy least favored is the minimum sizing strategy, for the exact opposite reasoning outlined above. The Full Diffusion sizing strategy falls in between these two strategies. It contains both the minimum sizing strategy cell (4F) and regular sizing strategy cell (1F). Should robustness be the critical design feature, the Full Diffusion sizing strategy could be chosen and the synthesis stage optimized for robustness. Moreover, should an acceptable level of robustness be specified, the Full Diffusion sizing strategy could be used such that the initial cell choice is made from high robustness cells to meet the robustness constraint and then the synthesis iterated towards higher performance from the fingered cells until the robustness constraint was violated. This would produce a faster / more energy efficient design than the regular sizing strategy library could for the robustness criterion.

7.4 Variation

7.4.1 Mechanical Stress

Section 2.4.1 described the affect of the mechanical stress induced into the existing silicon after the STI processing has been performed. The stress is shown to decrease electron mobility and increase hole mobility. The magnitude of the deviation from nominal is dependent on the distance from the isolation wall (often termed the Length of Oxide Definition, LOD), falling into insignificance after 2µm. For dense digital logic, this distance is simply too large to incorporate into the cells, which can be as small as 600nm wide. Therefore, the increase in variability is accepted as a necessary evil for highly packed digital logic design.

As the method used to induce the INWE in the minimum and Full Diffusion sizing strategies is STI placement, the levels of mechanical stress induced into the silicon is increased. The fingered devices are therefore of a higher variability, with mechanical stress one of many contributing factors.
Mechanical stress is incorporated into the BSIM 4.5 models, therefore the variability analysis in Chapter 4 includes this form of variability. Moreover, as the effect degrades NMOS devices and improves PMOS devices, the inherent P-to-N strength ratio is brought closer to parity.

In terms of sizing strategies, the superiority is strictly technology dependent. The variation analysis conducted in Chapter 4 showed that the least variable gate in the RVT device permutation was the 1 finger device. This support the mechanical stress argument as this device has the lowest amount of STI induced mechanical stress. However, the same analysis showed that the 4 fingered gate displayed the lowest variability in the LVT devices. In this case, the mechanical stress, along with the other variability contributors, was outweighed by the averaging effect of multiple identical devices.
Both the regular sizing strategy and minimum sizing strategy contain only one device width (maximum or minimum). The Full Diffusion sizing strategy contains both extremes plus a scale in between; therefore it could always be used to generate the lowest variability design during synthesis optimization.

7.4.2 Litho-friendly Design (LFD)

7.4.2.1 Polysilicon Pitch

Section 2.4.2.1 described the variation induced into highly scaled technologies where the minimum features size is less than the wavelength of the lithographic process. The optical proximity effects in a 65nm process were calculated to have an interaction distance of 775nm, far greater than the polysilicon pitch in cells with several polysilicon lines. As this distance is governed by the manner in which devices are stacked in series, not parallel, there is nothing in the minimum or Full Diffusion sizing strategies that exacerbates, alleviates or simply has any influence on this metric whatsoever. No superior sizing strategy therefore emerges for variation induced by polysilicon pitch.

7.4.2.2 Device Orientation

Section 2.4.2.2 described variability induced between devices forcing carriers through differing orientations in the underlying monocrystalline silicon lattice structure. The benefits gained from INWE in the minimum and Full Diffusion sizing strategies are achieved without having to orientate devices in the vertical or any orientation other than horizontal. As all strategies obey this orientation, no superior strategy emerges based on this metric

7.4.2.3 Polysilicon Neighborhood

Section 2.4.2.3 described the variation imparted by differing polysilicon etching rates depending on the polysilicon structures in the vicinity of the polysilicon gate. As synthesized logic placement is automated, the best form of defense is to position the polysilicon structures towards the center of the gate. The inducement of the INWE by intentional introduction of STI spacers does not force a deviation away from this design strategy. Therefore the minimum and Full Diffusion sizing strategies suffer this form of variation no more than the regular sizing strategy does.

7.4.2.4 Polysilicon Rounding

Section 2.4.2.4 described the variation imparted by polysilicon rounding encroaching into critical gate areas by the abutment of polysilicon routing close to the gate structures. To prevent this from of variation, the DRM for the technology node chosen recommends a minimum abutment distance of 50nm from any gate structures. For complex cells that required polysilicon routing such as the flip-flop, the narrowing of the power rails

permitted this minimum recommended distance to be adhered to. Therefore the STI techniques required to induce the INWE do not come at the cost of additional variation imparted by polysilicon rounding.

7.4.2.5 Contact Density

Section 2.4.2.5 described the variation induced by mechanical stress imparted from vias, with a difference of up to 5% reported between sparse and dense via arrangements. The STI spacers required for the minimum and Full Diffusion sizing strategies can be placed without interfering in gate performance and without increasing silicon area. For cells with large diffusion areas, the lower fingered (1F/2F) Full Diffusion cells group vias in pairs, this is because the design rules for two vias and three or more vias are different, with three or more vias requiring a greater equidistant spacing, summing to a greater overall area requirement. For the 1F Full Diffusion cell, a choice was therefore presented between 4 paired but closely spaced vias and 3 equidistant vias. For the redundancy arguments presented in Section 7.3.5, the former option was chosen. The STI spacers required for the minimum and Full Diffusion sizing strategies ensure that via spacing larger than the minimum recommended pitch is adhered to. Therefore, no additional variation is imparted by using this technique, even though the overall number of vias required increases as the number of STI spacers increases.

7.4.2.6 Polysilicon Counter-Doping

Section 2.4.2.6 described a variation mechanism by which alternate device types (P/N) experience a variation of up to 10% when placed in close proximity and sharing a single polysilicon gate structure. The spacer techniques used to create the minimum or Full Diffusion sizing strategies add no further restrictions on the spacing between alternate active diffusion areas. Therefore there is no additional variation from polysilicon counter-doping for these sizing strategies.

7.4.2.7 Line Edge Roughness

Section 2.4.2.7 described a stochastic source of variation derived from the non-ideal dimensions of the final polysilicon structure. Whilst stochastic, the variation is proportional to the overall perimeter of the device.

Correct RSCE sizing for the subthreshold regime showed that depending on the process corner and device variant, the optimal length is frequently greater than minimum. The

largest optimal length for the chosen technology node was shown to be 230nm. Increasing the length increases the device perimeter and therefore lowers variation from device edge roughness.

When introducing the STI spacers to induce the INWE, the total device perimeter for the pull up and pull down network increases. However, devices have to be considered as individual transistors. The variation from line edge roughness therefore forms part of the parametric variation of a single device. The overall effect of variation must then be considered as outlined in Section 4.5. Line edge roughness is included in the BSIM 4.5 model and therefore the variation analysis of Chapter 4 accounted for this form of variation. Again the LVT gates showed minimum variation at 4F and the RVT at 1F. The sizing strategy advantage is therefore technology node dependent.

7.4.3 Random Dopant Fluctuation (RDF)

Section 2.4.3 showed that RDF can account for up to 70% of the total device variation in deep submicron bulk planar technology nodes. The variation analysis in Chapter 4 showed that this this is not only technology dependent but also depends on the choice of VT device in the same technology node.

For the technology node chosen, the 4F design was least variable in LVT and the 1F device least variable in RVT. Should the designer intend to select only a single VT synthesis flow, this would make the minimum sizing strategy and regular sizing strategies the most appealing respectively. However, the Full Diffusion sizing strategy contains both of these cells at both VTs. Therefore the Full Diffusion sizing strategy always contains the least variable cell and therefore can be used to synthesize the lowest variable design via synthesis optimization. Having the device characteristic range, it could also provide a faster design than the regular sizing strategy and more energy efficient design than the minimum sizing strategy given a variability constraint. It could therefore be argued that it is the superior choice for this type of variation.

7.4.4 Well Proximity Effect (WPE)

Section 2.4.4 described a variation mechanism in which ions are scattered into the channel by the edge of the photo resist, lowering carrier mobility and increasing the threshold voltage. The distance beyond which this effect becomes negligible was shown to be 1µm. The distance whereby the majority of scattered ions reach the channel center was shown to be as high as 400nm.

The inclusion of additional STI spacers to induce the INWE increases the occurrence of ion scattering and the distances affect all devices in all sizing strategies, growing in significance from the regular sizing strategy to the minimum sizing strategy.
Again, the BSIM 4.5 compact model accounts for the WPE. The variation analysis in Chapter 4 therefore includes this source of variation. The same argument for the earlier effects apply, which is to say that the LVT gate shown to have the lowest variability was the 4F device, which theoretically contained the highest variation from the WPE. The reduction in variation from the averaging effect of multiple devices exceeded the additional variation induced from the WPE. The superior sizing strategy is therefore technology node dependent.

7.4.5 Variation Overview

In terms of overall variation, the analysis from Chapter 4 and the points outlined in the previous subsections highlight the fact that the superior sizing strategy is not only technology node dependent, but also VT dependent within the same technology node.
The 65nm LVT devices showed that the best method to reduce variation was to increase in the number of fingers, as the averaging effect proved greater than increasing area. This would make the minimum sizing strategy the superior choice.
The 65nm RVT devices showed that the best method to reduce variation was to increase the area of a single device, as the inverse quadrature proportionality proved greater than the averaging effect. This would make the regular sizing strategy the superior choice.
The Full Diffusion sizing strategy included the variability optimal cells at both VTs. Therefore, its choice could always produce the lowest variability design via synthesis optimization. It could therefore be considered the superior sizing strategy of all three listed.

7.5 Performance

7.5.1 Reverse Short Channel Effect (RSCE)

Section 2.5.1 outlined the physics of the RSCE and speculated on the effects this would have on device characteristics in the subthreshold regime.
Theoretically it was speculated that the RSCE would increase the depletion depth in the channel and therefore decrease the gate capacitance. The capacitance analysis in Chapter 4 showed minor deviation in the geometric relationship of the device capacitance, but

nothing of any design value. Therefore the desired reduction in capacitance as a result of RSCE was not seen in the chosen technology node.

It was also speculated that the lowering of dopant density in the center of the channel via RSCE optimal sizing would lower the device threshold voltage and therefore increase drain current for a given V_{DS}. This was shown to be accurate in the subthreshold current geometry sweeps presented in Chapter 3.

Even without lowering the capacitance, the additional current and therefore decrease in propagation delay make RSCE optimization a worthwhile endeavor. The length optimizations provided in Chapter 5 show that the RSCE optimal length is dependent on the device width, device VT and even the technology node. The sizing strategy must therefore be determined before the RSCE optimal length can be calculated.

As all three sizing strategies benefit from the RSCE, no sizing strategy can really be deemed superior to the other. The drain current geometric sweeps of Chapter 3 show that the RSCE is much more important to minimum width devices as this is where the RSCE optimal length deviates most from the minimum length SCE optimal. This was corroborated in the study by [98] who attempted a minimum width sizing strategy without RSCE optimization and received silicon results with a 10% performance degradation instead of the 50% improvement seen in simulation.

7.5.2 Inverse Narrow Width Effect

Section 2.5.5 outlined the physics of the INWE and speculated on the effects this would have on device characteristics in the subthreshold regime.

It was again theoretically speculated that the INWE would increase the depletion depth of the channel and therefore decrease the gate capacitance. The geometric capacitance analysis presented in Chapter 3 also disproved this to be the case. However, the reduction in gate area as a result of the increasing number of STI spacers was shown to lower the gate capacitances in the gate characterization analysis shown in Chapter 5. A reduction in gate capacitance directly equates to a decrease in dynamic energy consumption, which means for energy-optimized circuits, the highest level of INWE possible should be applied to circuits dominated by dynamic energy consumption.

It was also speculated that the increase in transverse electric field thanks to the addition of the fringing field would induce a larger current. This was shown to be true for both the active and leakage currents. For some device configurations, the reduction in propagation delay and decrease in gate capacitance resulted in a gate with a performance improvement

greater than the cost in increased leakage current. Most gates however showed leakage current increased at a greater proportion.
The Full Diffusion sizing strategy presented a novel way to negate the deleterious increase in leakage current by providing a full range of usable gates under high variation constraints to safely re-introduce multi-Vt synthesis into the subthreshold regime. This benefit is only available to complex designs that have a large range in inherent path delays.
In terms of the INWE, the superior sizing strategy is therefore design dependent. For simple designs with path delays of a similar size, the additional leakage of the Full Diffusion sizing strategy would negate the benefits in its use and the superior sizing strategy of choice would be the minimum sizing strategy. For larger complex designs with a large range of path delays, the speed improvement and availability to perform leakage recover from multiple VT devices leads the Full Diffusion sizing strategy to be the superior choice.

7.5.3 Stack Forcing

Section 2.5.9 outlined the physics of the stack forcing. It was postulated that forcing a duplicate device in series with an existing device would create a condition whereby the V_{GS} on one device would become negative. That device would be pushed into a state of supercutoff and would therefore reduce leakage current for the network.
Chapter 3 showed that the INWE is equally inducible in stacked devices and Chapter 5 showed that all basic combinational logic gates can be stack forced. However, whilst the simulation data from Chapter 3 showed that both the active and leakage current were reduced, no evidence of the supercutoff condition was observed in the chosen technology node. Moreover, the Ion/Ioff ratio was actually degraded. The reduction in leakage current would create a reduction in leakage energy, however, the additional devices essentially double the amount of gate capacitance and therefore dynamic energy. Chapter 5 showed that some of the redundant nets can be eliminated in stack forcing but the gates remain inferior to non-stacked gates.
The stacked gates did find use in extending the range of the Full Diffusion sizing strategy. These provide vital stepping-stones between the LVT and RVT devices and provide a mechanism by which mutli-Vt synthesis can be safely re-introduced back into the subthreshold regime for any design. For the aforementioned reasons however, they should be used sparingly in energy critical designs.

In terms of how stack forcing impacts the three sizing strategies, stack forcing was shown to impact them equally, but provided pragmatic use only in the Full Diffusion sizing strategy to extend the effective characteristic range.

7.5.4 Performance Overview

In terms of overall performance, the superior sizing strategy is actually design dependent. Both the RSCE and stack forcing can be used to provide additional benefits to all three sizing strategies. The regular sizing strategy was shown to provide little performance benefit in the subthreshold regime. For this reason, it should realistically be precluded from voltage scaled design. The minimum sizing strategy is the sizing strategy of choice in the field. It provides the greatest performance increase at the cost of leakage energy increase. For the low activity factor designs typically used in the subthreshold regime, this could be a problem. However, the Full Diffusion sizing strategy does not solve this problem for simple designs with similar path delays. There is therefore no benefit in using the Full Diffusion sizing strategy over the minimum sizing strategy for these types of simple circuits.

However, most circuits in the subthreshold regime are complex circuits with large path delays. The Full Diffusion sizing strategy helps alleviate the leakage increase associated with the use of the INWE by offering a characteristic range of cells that spread across the two device VTs available. This allows fast LVT cells to be placed in a small amount of slow paths to increase performance, but less leaky RVT cells to be used in the vast majority of paths to reduce leakage. The result is a circuit with a superior energy profile.

7.6 Hypothesis Two

The second hypothesis presented in the thesis posed:

Can the Full Diffusion sizing strategy be used, along with any necessary stacked interstitial libraries, to create a more energy efficient design by the increase in granularity of multi-vt synthesis in the subthreshold regime?

The AES core testing and results provided in Chapter 6 were aimed at determining whether the Full Diffusion sizing strategy could produce a more energy efficient complex design, using an identical Verilog hardware description, over a state-of-the-art proven subthreshold library. Additional metrics of the design and test results shall be explored to determine whether a superior design was indeed created. These are area, activity factor, minimum energy point (MEP), performance (frequency), variation and minimum operating voltage (MOV).

7.7 Design Metrics

7.7.1 Area

One of the key metrics for IOT viability is cost, as throwaway silicon devices have to be manufactured at a price point low enough that their disposability does not influence the intended application. This is particularly true for the envisaged application of tracking fast moving consumable items. By using cheaper mature bulk planar technology nodes, the primary cost factor becomes the silicon area of the design.

The results showed that the silicon areas of the designs on a strict utilized area only basis were 45742 μm^2 for the ARM LE library, 49812 μm^2 for the Full Diffusion RVT library and 49315 μm^2 for the Full Diffusion Multi-Vt library. These represent increases of 8.9% and 7.8% respectively. Interestingly, the cell counts for the designs were 5626 for the ARM LE library, 7318 for the Full Diffusion RVT library and 7244 for the Full Diffusion Multi-Vt library. These represent increases of 30% and 28.6% respectively. This lack of direct correlation supports the fact that the cell count increase is as a result of the synthesis tool's ability to perform logical decomposition using the richer variety of cells in the Full Diffusion libraries. Therefore, the switching of combinations of the smaller cells in the Full Diffusion libraries for a single larger cell in the ARM LE constituted little overall change in the silicon area.

The actual source of silicon area increase was the RSCE length optimization that created wider cells quantized to the unit cell. Whilst the RSCE optimization for the LVT devices was larger than the RVT cells (150nm to 100nm respectively), only 4% of the cells in the multi-Vt design were LVT cells and the overall cell count was fewer. It can therefore be argued that the Full Diffusion sizing strategy alone has little impact on silicon area.

7.7.2 Activity Factor

The activity factor of a design is a direct measure of the active toggling during operation and therefore directly influences the dynamic energy consumption of the design. The activity factors for the AES core designs were 63.38% for the ARM LE library, 60.97% for the Full Diffusion RVT library and 47.39% for the Full Diffusion Multi-Vt library, representing reductions of 2.41% and 15.99% respectively.

This reduction correlates with an increase in the range of cell performance offered in the libraries. As the range increases, the number of viable circuit topologies available to the synthesis tool for a given time constraint increases. This additional choice allows the synthesis tool to select circuits that perform identical functionality with a finer granularity in the logic. This means that fewer gates toggle during operation, shifting the energy-per-cycle contributions away from the dynamic energy consumption and towards leakage, which can be mitigated or eliminated using power gating techniques. This effect is an excellent candidate for further research in subthreshold energy optimization.

As the purpose of voltage scaling is to reduce dynamic energy consumption, this additional reduction serves as an advantage to the Full Diffusion sizing strategy. It also serves to push the MEP higher up the supply voltage scale, which is beneficial for the reasons outlined in the next subsection.

7.7.3 Minimum Energy Point (MEP)

The minimum energy point signifies where on the supply voltage scale the minimum energy-per-cycle for a design resides. This is the point where the contributions of the leakage energy and dynamic energy are equal. Therefore a reduction or increase in one of these underlying contributions shifts the MEP along the voltage supply scale.

The MEPs for nominal operation (20 °C) were 250mV for the ARM LE library, 270mV for the Full Diffusion RVT library and 300 mV for the Full Diffusion Multi-Vt library. Figure 72 in Section 5.10 showed that this shift is the result of both an increase in leakage energy and reduction in dynamic energy for the Full Diffusion libraries. The temperature

analysis in Section 6.4 showed that the Full Diffusion MEP's were consistently higher up the supply voltage scale across the full temperature range tested of 0 °C to 85 °C.

The location of the MEP is important for two reasons. The first is that variation is voltage dependent. Increasing supply voltage deceases variation, therefore operation at a higher supply voltage aides in counteracting the negative effects of variation. The second reason is that voltage-scaling schemes like Ultra Wide Dynamic Voltage Scaling are often at the mercy of the DC-to-DC voltage conversion efficiency. It is more efficient to convert DC voltages that are closer together than it is to convert DC voltages that are further apart. For this reason, operating at an MEP higher in supply voltage reduces losses in DC-to-DC conversion.
From the discussion above, the Full Diffusion sizing strategy can be considered superior to the comparative ARM LE library.

7.7.4 Frequency (Performance)

As the primary concern of subthreshold operation is energy minimization, the frequency of operation is a primary concern. In terms of energy-per-cycle, the frequency of operation determines how long non-switching devices remain leaking whilst switching devices perform computation. Moreover, in terms of absolute energy consumption in duty-cycled systems, the faster a design can complete the required task, the faster the design can be power gated, reducing the active period in the duty cycle.
At nominal operation (20 °C), the maximum functional frequencies of the libraries were 17kHz for ARM LE, 34kHz for Full Diffusion RVT and 147kHz for Full Diffusion Multi-Vt. These were taken at their respective MEP to provide a fair comparison. This represents a performance improvement of 2X and 8.65X respectively, indicating a substantial advantage for the Full Diffusion libraries.
Due to temperature inversion when operating in the subthreshold regime, devices fail at low temperatures. At low temperature operation (0 °C), the maximum functional operating frequencies of the libraries were 4kHz for ARM LE, 7kHz for Full Diffusion RVT and 41kHz for Full Diffusion Multi-Vt. This represents performance improvements of 1.75X and 10.25X respectively. The ring oscillator results from Chapter 4 indicated that the INWE fillip in performance comparatively increased at lower temperatures. This is clearly shown in the Multi-Vt library but not in the RVT library. Interestingly the Multi-Vt library design consisted of 96% RVT gates. It is therefore likely that the

additional stability provided by the INWE did exist in the AES core experiment, but that the quantization of 1kHz on the clock generator skewed the result for the RVT only comparison. The results from the high temperature testing revealed frequency improvements of 1.48X and 3.6X. The conclusion from this is that the Full Diffusion sizing strategy always produces a faster design, but that this performance enhancement is degraded as temperature increases.

7.7.5 Energy

The purpose of subthreshold operation is to reduce energy consumption and therefore extend battery life. The energy-per-cycle of a design is therefore of critical importance. At nominal operation (20 °C), the energy-per-cycle for the libraries were 4.12pJ/cycle for ARM LE, 3.84pJ/cycle for Full Diffusion RVT and 3.11pJ/cycle for Full Diffusion Multi-Vt. These were taken MEP-to-MEP for fair comparison. This represents an energy improvement of 7% and 24% respectively.

At low temperature operation (0 °C), the energy-per-cycle for the libraries were 3.57pJ/cycle for ARM LE, 3.21pJ/cycle for Full Diffusion RVT and 2.85pJ/cycle for Full Diffusion Multi-Vt. This represents an energy improvement of 10% and 20% respectively. This shows that even providing the additional performance stability at lower temperatures, the comparative energy improvement remains fairly constant. This is interesting as the ring oscillator results from Chapter 4 showed that at the device level, the leakage current increases mirroring the frequency improvement. Whilst at the device level this appears as a disadvantage, at the circuit level it serves only the purpose of pushing the MEP higher up the supply voltage scale in tandem with the reduction in dynamic energy. The energy-per-cycle improvement is therefore maintained.

At high temperature operation (85 °C), the comparative improvements were 6.4% for Full Diffusion RVT and 27.3% for Full Diffusion Multi-Vt. This shows that the Full Diffusion sizing strategy provides energy-per-cycle improvements consistent across the full temperature range tested.

7.7.6 Variation

Variation is a key metric of interest as it is exacerbated by operation in the subthreshold regime and is directly related to yield. An experiment to test design variation at nominal operation was therefore included in the AES testing.

The coefficient of variation (σ/μ) for maximum frequency was 53% higher for the Full Diffusion RVT library comparative to the ARM LE library. The variation analysis in Chapter 4 showed that an increase of up to 28% could be attributed directly to the underlying devices. The remaining variation is likely due to the disproportional increase the standard deviation as a result of increased logical decomposition in comparison to the increase in mean frequency.
The coefficient of variation for maximum frequency was 32% higher for the Full Diffusion Multi-Vt library comparative to the ARM LE library. Again the reasoning behind this increase is likely the functional decomposition. Interestingly, Chapter 4 showed that the LVT device variants had an inherently lower variability. As the LVT devices were used to improve the critical paths, it is likely that this accounts for the lower maximum frequency variability displayed comparative to the RVT only Full Diffusion library.

The worst case deviation from the mean was also calculated for each library at their MEP. The ARM LE library has a worse case deviation of 7.7%, the Full Diffusion RVT library 5.9% and the Full Diffusion Multi-Vt library 9.1%. It is interesting that the RVT only library deviates less than the ARM LE library as the coefficient of variation shows the opposite effect. This suggests there may be a deviation in the Gaussian distribution and although the probability of having a mean frequency implementation is diminished, the worst-case outliers are superior for the Full Diffusion RVT library. Alternatively, this may just be an erroneous result symptomatic of the restricted sample size of 10. The proportional increase in worst case deviation for the Full Diffusion Multi-Vt library is likely due to global tracking differences in the LVT and RVT device processing steps.

The coefficient of variation for the leakage current was 29% higher for the Full Diffusion RVT library but 21% lower for the Full Diffusion Multi-Vt library. The increase in variation for the RVT library is extremely close to the 28% that could be attributed directly to the underlying devices. The decrease for the Multi-Vt library is also likely as a direct consequence of the addition of the underlying devices. The AES results of Chapter 6 show a marked increase in leakage current (increase in μ) yet 96% of the gates retain the standard deviation of the RVT devices. This disproportionately alters the coefficient of variation.

The coefficient of variation for the energy-per-cycle was 115% higher for the Full Diffusion RVT library and 54% higher for the Full Diffusion LVT library. These results are likely an amalgamation of performance and leakage variations previously addressed. As energy is the metric of power over time, the large frequency improvements provided by the INWE exacerbate the variation in energy-per-cycle. Even if the standard deviations of the underlying devices remained the same, the frequency improvement results in a lower mean and therefore increases the coefficient of variation. There is no method to mitigate this and therefore it must be accepted as the cost of lowering the average energy-per-cycle. As the energy-per-cycle is a second order metric, this increase in variability does not induce functional or temporal errors into the design.

7.7.7 Minimum Operating Voltage

The minimum operating voltage is the supply voltage point at which functional logic errors occur and the design fails. Much emphasis is placed on this metric in the field and a claim to its direct impact on yield is often made.

The AES core results in Chapter 6 showed that the minimum operating voltages for the Full Diffusion libraries were consistently higher than for the ARM LE library. However, these findings must be taken in context. Given that the minimum energy points also shift an equal amount higher in supply voltage, no degradation in yield can be claimed by the use of the Full Diffusion sizing strategy.

7.7.8 Hypothesis Discussion

In terms of the design metrics outlined in this subsection, it is safe to conclude that the Full Diffusion sizing strategy can be used to create a superior design by allowing the safe and successful reintroduction of multi-Vt synthesis into the subthreshold regime.

The decision was taken to allocate one of the three allotted core slots to an RVT only implementation of the Full Diffusion sizing strategy to determine whether the range created for a single VT library could still provide a marked improvement over an existing commercial subthreshold library. The implementation proved nominally twice as fast and 7% more energy efficient for almost no additional silicon area. Therefore, if only a single VT is available in a bulk planar technology library, the Full Diffusion sizing strategy is still worth pursuing.

The Full Diffusion sizing strategy provides an impressive improvement when the Multi-Vt library is used. Nominally it showed an 8.65X frequency improvement and a 24%

reduction in energy consumption. It consistently outperformed the baseline library across the temperature range tested in frequency and energy-per-cycle again without any meaningful increase in silicon area.

The cost of this was primarily an increase in variability. However, the increase in functional metric variation for the Multi-Vt library was never more than 32% and no core came close to failure when operated at the corresponding minimum energy point.

7.8 Prior Art Comparison

Table 18 shows a comparison of the Full Diffusion sizing strategy to other sizing strategies outlined in Chapter 2. Direct comparison is not possible due to the differences in process node and benchmark circuit, but the claimed improvements may be compared. The RSCE sizing strategy proposed in [67] synthesized ISCAS benchmark circuits in a 120nm bulk planar technology process. Simulation at a nominal temperature of 27 °C showed a maximum delay improvement of 10.38% and a maximum power improvement of 34.38%. These results were not corroborated in silicon. The study justifies a reduction in energy by a local minimum in gate capacitance at a length 3X the minimum device length (360 nm) assuming the device width is optimized for iso-current. There are many caveats to this approach. The first is that the methodology precludes the use of the INWE, which shows a greater benefit than RSCE (demonstrated as up to 2.4X in this work as opposed to the 10.38% claimed for RSCE). Therefore if only one optimization is to be made than governs the other, the optimization of choice should be the device width, not length. The simulation work presented earlier also shows that the optimal length and potential gain in RSCE is highly dependent on process corner, drifting from a high as 230 nm in the SS corner for LVT devices, right to minimum length for the same device in the FF corner. Conversely, the optimal device width across all process corners was always determined to be where INWE was highest (i.e. minimum width). Therefore the global process corner does not cause optimization of this dimension to drift. Finally the work presented in Chapters 5 and 6 showed that even using VCD power analysis, the tool under estimated the power and therefore energy consumption by 26% for non-INWE optimized sizing strategies. Therefore it is difficult to accept studies based only on simulated data as accurate without corroboration in silicon.

The minimum width sizing strategy proposed in [98] synthesized an 8-bit 8-tap FIR filter in a 180nm bulk planar technology node. The study followed the established commercial synthesis choice of signing off the design at the SS corner to guarantee functionality at a fixed frequency. The design was then committed to silicon and tested. Measured at the energy optimal supply voltage for the design of 250 mV, the silicon exhibited an increase in delay of 50%. Conversely during testing, the SS corner simulation showed a 10% improvement. There is no evidence of RSCE-aware geometric sizing on the length. Had the author failed to perform any optimization, the length sizing was likely minimum length which is less than optimal and highly variable in subthreshold. Had the author optimized device length for the SS corner correctly along with the synthesis signoff corner, it is likely that the decade old process node returned TT silicon and the device lengths were massively oversized. Even if the author had used TT sized devices and signed off at the SS corner, the synthesized design would sill have been limited by the slower expectation of the circuit. The analysis presented in Chapter 4 indicated that performance improvements of up to 1.8X/2.4X for LVT/RVT devices are to be expected using this type of minimum width sizing strategy. The results of the study are therefore limited by the author's lack in understanding of subthreshold semiconductor physics. Using the analysis from Chapter 4 to make an educated guess at what the outcome of this study should have been, it is likely that this sizing strategy exhibits the aforementioned performance increase but the increase in leakage generates a larger energy-per-cycle than the Full Diffusion sizing strategy would. The minimum width sizing strategy and Full Diffusion sizing strategy share the same fastest cell, therefore no speed advantage is displayed between one or the other. It could therefore be argued that the lower energy-per-cycle and larger cell choice favors the Full Diffusion sizing strategy.

The constant yield sizing strategy proposed in [80] synthesized and simulated Kogge-Stone adder circuits. All data for the designs in the work are presented in relative sizing and therefore the process node used is indeterminate. The sizing strategy argues that in order to meet a constant failure rate of 0.13%, the device widths must be upsized by at least 63% and as much as 4.43X when operating at a voltage of 240 mV. Monte Carlo simulation of 1000 iterations showed that this reduced the variation of first order characteristics such as leakage current by around 10% but increased variation in energy-per-cycle by as much as 53% comparative to the single minimum sized device (most variable). The Full Diffusion Multi-Vt AES core displayed a 21% decrease in leakage

Device Sizing Strategy	Benchmark Circuits	Process Node	Fabricated	Claimed Improvement
RSCE-aware Sizing [67]	ISCAS Benchmark	120nm	No	10.38% Delay improvement 34.38% Power improvement
Minimum Width Sizing [98]	8-bit 8-tap FIR Filter	180nm	Yes	10% SS Delay improvement 50% TT Delay degradation No RSCE optimization
Constant-Yield Sizing [80]	Kogge-Stone Adders	N/A	No	0.13% Failure rate achieved 9.94% Variation reduction in leakage current 53.48% Variation increase in energy-per-cycle
INWE-Aware Sizing [75]	Base-band Processor	40nm	No	26.47% Leakage power reduction 15.43% Dynamic power reduction 7.46% Improvement in area
Full Diffusion Sizing	128-bit AES with LBIST	65nm	Yes	2X/8.65X frequency improvement for RVT/Multi-Vt 7%/24% energy-per-cycle improvement for RVT/Multi-Vt 1.8% reduction in max frequency deviation at MEP for RVT

Table 18: Prior art comparison

current variation and a 54% increase in energy-per-cycle variation comparative to the largest width device (least variable). This information, along with all device sizing at least 4X of the minimum device width in the 65nm technology node suggests that the Full Diffusion sizing strategy would also meet the fixed failure outlined in the study. Therefore, for the criterion of fixed failure rate, it is likely that the Full Diffusion sizing strategy offers improvement over the constant yield sizing strategy. For a simple circuit with close path timings, the caveat to this would be an increase in leakage. For large complex circuits with a large variety in path timing, this potential loss is claimed back as was demonstrated by the AES core synthesized from the Multi-Vt library.

The INWE-Aware sizing strategy proposed in [75] synthesized a base-band processor in a 40nm bulk planar technology. The study uses minimum width quantized cells mixed with a superthreshold-sized library to compare against a superthreshold-sized library at 300 mV. The mixed library does not produce the range required to successfully perform multi-vt synthesis. Therefore, all cells are RVT only. The mixed library shows a 26.47% reduction in leakage power and a 15.43% reduction in dynamic power. However the delay improvement was only 20%. As the Multi-Vt Full Diffusion AES core proved 8.65X faster at nominal temperature, the power reduction displayed within the study would not equate to the same energy per cycle reduction displayed in the Multi-Vt core. Whilst this sizing strategy comes closest to matching the Full Diffusion sizing strategy, it is still far surpassed in energy and performance for complex circuits with large path timing variation.

7.9 Summary of Novel Contribution and Critique

The work conducted towards this thesis has provided several novel contributions and encountered several limitations. These are now discussed.

7.9.1 Novel Contributions

7.9.1.1 Impact of the Inverse Narrow Width Effect over Multiple Widths

The focus of contributions in the field thus far has been on minimum width devices. Therefore beyond a simple width sweep to show where the INWE is most prevalent, no data has been present on using the INWE beyond the minimum width. This thesis is the first to present a methodological analysis of the INWE on devices of varying widths. Moreover, this study is the first to corroborate this in measured silicon results. The results

showed that the INWE affects all devices subjected to fringing fields from thick field oxide gate overlap encountered wherever a shallow trench isolation spacer is placed. This affect can then be used to create a range of devices, rather than simply optimizing for highest performance at minimum width.

7.9.1.2 Evaluation of the Inverse Narrow Width Effect over a Full Supply Voltage Range

The focus in the field has been to use the Inverse Narrow Width Effect solely in the subthreshold regime. The silicon results from the ring oscillators showed that the INWE can be used to provide a performance improvement for all combinational gates tested right up to 800 mV. Moreover, the maximum performance penalty for this at nominal voltage was only 5%. This was a surprising result and showed that the INWE was an excellent candidate for use in voltage scaling techniques such as Ultra Wide Dynamic Voltage Scaling.

7.9.1.3 Evaluation of the Inverse Narrow Width Effect over a Wide Temperature Range

No systematic, silicon corroborated testing for the INWE over a large temperature range has been presented in the field before this thesis. The measured results showed that the contribution of the INWE increases as temperature decreases. This props up the circuit performance under temperature inversion in subthreshold and therefore provides an additional form of circuit stability.

7.9.1.4 Proposal and Testing of the Full Diffusion Sizing Strategy

Finally, this work explored all avenues of cell design within the design space of a standard cell 12 track library operating in the subthreshold regime. A new sizing strategy was proposed that used the INWE to create a range of cell characteristics. Cells were designed using this strategy and characterized. A complex circuit was synthesized using a commercial digital synthesis design flow and committed to silicon. This was then measured against a design synthesized from an identical hardware description and state-of-the-art subthreshold standard cell library. The design proved to consistently provide a higher performance at a lower energy-per-cycle.

7.9.2 Critique

7.9.2.1 Minimum Sizing Strategy

For completeness, it would have been beneficial to also synthesize and measure an AES core designed from the minimum sizing strategy. This would have required little additional work as the Full Diffusion libraries already contain the 4F cell. The only reason this was not performed was due to area restrictions on the test chip.
The simulation and analysis work covered within the thesis suggests that the minimum sizing strategy would also generate the performance improvements seen in the Full Diffusion silicon measurements. However, without the range of cells to perform leakage recovery, the energy-per-cycle at the minimum energy point would be higher and therefore inferior.

7.9.2.2 Full Voltage AES Core Testing

It would have been interesting to see how the frequency and energy-per-cycle measured above 500mV and all the way up to the nominal 1.2V. This was a limitation in a part of the design (clock generator) that was not designed by the author but was provided as part of the test chip. As such this was beyond the control of the author and could be left for future work.

7.10 Chapter Summary

This chapter reflected on the design considerations of Chapter 2 to determine whether a superior subthreshold library could be designed by taking advantage of the underlying subthreshold physics. The three sizing strategies introduced in Chapter 4 were compared on robustness, variability and performance and the appropriate sizing strategy matched to a particular design goal.
The Full Diffusion sizing strategy was the then compared to the ARM LE library by drawing on the AES core results presented in Chapter 6. The work concludes that the Full Diffusion library is able to synthesize superior circuits but only if those circuits are complex enough to exhibit large variation in path delays. Fortunately, this accounts for most IOT designs.
A comparison to prior art was then presented, with the Full Diffusion sizing strategy showing favorable results. Finally the novel contributions of the work were outlined and the work undertaken scrutinized for improvement and extension to further work.

Chapter 8: Conclusion

The work conducted and presented in this thesis showed that the underlying device physics in the subthreshold regime can be leveraged to produce a superior subthreshold standard cell library. By conducting geometric sweeps of active current, leakage current, Ion/Ioff ratio and gate capacitance, optimal sizing strategies based on the Reverse Short Channel Effect and Inverse Narrow Width Effect were derived for the chosen technology node (65nm Bulk Planar). The entire design space for a subthreshold standard cell library was explored, and three distinct sizing strategies defined. These sizing strategies were then laid out in the chosen technology, parasitically extracted, simulated and compared to determine their strengths and weaknesses and it was demonstrated how these should be matched to the circuit to be designed. A novel sizing strategy, named the Full Diffusion sizing strategy, was created by geometrically introducing shallow trench isolation spacers to induce the Inverse Narrow Width Effect. The aim of this was to create a set of cells that were usable under strict variation conditions with varying performance and leakage characteristics.

Ring oscillators were created from the newly designed cells, simulated and committed to silicon. Measured results showed that performance improvement extended up to 800 mV for all basic combinational cells in the technology node studied. This lends well to the implementation of ultra wide dynamic voltage scaling. Temperature analysis also showed that the performance fillip from the inverse narrow width effect increased as temperature decreased, providing an additional form of stability to temperature inversion in the subthreshold regime. Stack forcing was explored to determine the viability of supercutoff leakage enhancement in the chosen technology node. Post layout simulation showed no supercutoff enhancement and degradation in the Ion/Ioff ratio and dynamic energy. Stack forcing was shown as a viable way to augment the characteristic cell range provided by the Full Diffusion sizing strategy, creating a platform to successfully reintroduce multi-Vt synthesis into the subthreshold regime.

The Full Diffusion sizing strategy was used to synthesize 128-bit AES cores to demonstrate how the characteristic range provides a methodology to construct a superior complex subthreshold circuit. By introducing fast fingered cells into critical paths and slow, less leaky cells into fast paths during leakage recovery, frequency improvements of up to 10.15X and energy-per-cycle improvements of up to 27% were measured comparative to a state-of-the-art subthreshold standard cell library.

Throughout the synthesis methodology, several areas of interest were highlighted as requiring deeper investigation beyond the scope of this contribution. The two most prominent are investigation into streamlining synthesis methodologies to determine maximum operating frequency and detailed analysis on the impact of the Full Diffusion sizing strategy on Activity Factor. These are left as opportunities for future work.

References

[1] R. H. Dennard, "Design of Micron Switching Devices," in *IEDM*, Washington D.C., 1972, pp. 168-170.

[2] Semiconductor Ind. Assoc., "The International Technology Roadmap for Semiconductor," San Jose, 2007.

[3] R. Gonzalez, "Supply and threshold voltage scaling for low power CMOS," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, pp. 1210-1216, August 1997.

[4] M. Kumar, "Effects of Scaling on MOS Device Performance," *IOSR Journal of VLSI and SIgnal Processing*, vol. 5, no. 1, pp. 25-28, January 2015.

[5] T. Ghani, "Scaling Challenges and Device Design Requirements for High Performance Sub-50 nm Gate Length Planar CMOS Transistors," in *Symposium on VLSI Technology*, 2000, pp. 174-175.

[6] R. Singh, "Analysis of the Effect of Temperature Variations on Sub-threshold Leakage Current in P3 and P4 SRAM Cells at Deep Sub-micron CMOS Technology," *International Journal of Computer Applications*, vol. 35, no. 5, pp. 8-13, December 2011.

[7] M. Alioto, "Ultra-Low Power VLSI Circuit Design Demystified and Explained: A Tutorial," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 1, pp. 3-29, January 2012.

[8] W. Shockley, "The Theory of p-n Junction in Semiconductors and p-n Junction Transistors," *Bell. Sys. Tech*, vol. 435, p. 28, 1949.

[9] P. A. Stolk, "Modeling statistical dopant fluctuations in MOS transistors," *IEEE Transactions on Electron Devices*, vol. 45, no. 9, pp. 1960-1971, September 1998.

[10] Chenming Hu, "Lucky-Electron Model of Channel Hot Electron Emission," *1979 International Electron Devices Meeting*, pp. 22-25, 1979.

[11] B. Ricco, "Low Voltage Hot-Electron Effects in Short Channel MOSFETs," *1984 International Electron Devices Meeting*, pp. 92-95, 1984.

[12] E. Sangiorgi, "Hot Electrons and Holes in MOSFET's Biased Below the Si-SIO2 Interfacial Barrier," *IEEE Electron Device Letters*, vol. 6, no. 10, pp. 513-515, Oct 1985.

[13] Y. Nakagome, "New Observation of Hot-Carrier Injection Phenomena," *Japanese Journal of Applied Physics*, vol. 22, no. 22-1, pp. 99-102, 1982.

[14] F. Driussi, "Observation of a New Hole Gate Current Component in p(+)-poly Gate p-channel MOSFET's," in *30th European Solid-State Device Research Conference*, 2000, pp. 136-139.

[15] F. Driussi, "On the Electrical Monitor for Device Degradation in the CHISEL Stress Regime," *IEEE Electron Device Let.*, vol. 24, pp. 357-359, May 2003.

[16] S. Mahapatra, "A comprehensive study of hot-carrier induced interface and oxide trap distributions in MOSFETs using a novel charge pumping technique," *IEEE Transactions on Electron Devices*, vol. 47, no. 1, pp. 171-177, January 2000.

[17] H. P. Tuinhout, "Impact of parametric mismatch and fluctuations on performance and yield of deep-submicron CMOS technologies," in *ESSDERC*, Firenze, 2002, pp. 95-101.

[18] L. L. Lewyn, "Analog Circuit Design in Nanoscale CMOS Technologies," *Proceedings of the IEEE*, vol. 97, no. 10, pp. 1687-1714, October 2009.

[19] C. H. Choi, "Impact of gate direct tunneling current on circuit performance: a simulation study," *IEEE Transactions on Electron Devices*, vol. 48, no. 12, pp. 2823-2829, December 2001.

[20] I. C. Chen, "Electron trap generation by recombination of electrons and holes in Si02 ," *J. Appl. Phys*, vol. 61, no. 9, p. 4544, 1987.

[21] J. H. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits," *IEEE Transactions on Device and Materials Reliability*, vol. 1, no. 1, pp. 43-45, March 2001.

[22] E. Cartier, "Passivation and Depassivation of Silicon Dangling Bonds at the Si/SiO2 Interface by Atomic Hydrogen," *Appl. Phys. Lett.*, vol. 63, pp. 1510-1512, 1993.

[23] L. do Thanh, "Elimination and Generation of Si-SiO2 Interface Traps by Low Temperature Hydrogen Annealing," *J. Electrochem Soc.*, vol. 135, pp. 1797-1801, 1988.

[24] D. J. DiMaria, "Mechanism for Stress-Induced Leakage Currents in Thin Silicon Dioxide Films," *J. Appl. Phys.*, vol. 78, pp. 3883-3894, 1995.

[25] I. C. Chen, "Substrate Hole Current and Oxide Breakdown," *Appl. Phys. Lett.*, vol. 49, pp. 669-671, 1986.

[26] J. W. MacPherson, "Underlying Physics of the Thermochemical E Model in Describing Low-Field Time-Dependent Breakdown in SiO2 Thin Films," *J. Appl. Phys.*, vol. 84, pp. 1513-1523, 1998.

[27] W. W. Abadeer, "Reliability monitoring and screening issues with ultrathin gate dielectric devices," *IEEE Transactions on Device and Materials Reliability*, vol. 1, no. 1, pp. 60-68, March 2001.

[28] E. Harrari, "Dielectric Breakdown in Electrically Stressed Thin FIlms of Thermal SiO2," *J. Appl. Phys.*, vol. 49, pp. 2478-2489, April 1978.

[29] F. Chen, "A Comprehensive Study of Low-K SicOH TDDB Phenomena and its Reliability Model Development," *IEEE IRPS Proceedings*, pp. 46-53, 2006.

[30] J. C. A. Huang, "Some practical concerns on isothermal electromigration tests," *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 4, pp. 387-394, November 2001.

[31] C. Hau-Riege, "The Effect of Low-K ILD on the electromigration Reliability of Cu Interconnects with Differet Line Lengths," in *Proc. IEEE 41st Annu. Int. Rel. Phys. Symp.*, 2003, pp. 173-177.

[32] F. Chen, "Technology Reliability Qualification of a 65nm CMOS Cu/Low-K BEOL Interconnect," *2006 13th International Symposium on the Physical and Failure Analysis of Integrated Circuits*, pp. 97-105, 2006.

[33] M. A. Alam, "A critical examination of the mechanics of dynamic NBTI for PMOSFETs," in *IEEE International Electron Devices Meeting 2003*, Washington D.C., 2003, pp. 14.4.1-14.4.4.

[34] S. Mahapatra, "On the generation and recovery of interface traps in MOSFETs subjected to NBTI, FN, and HCI stress," *IEEE Transactions on Electron Devices*, vol. 53, no. 7, pp. 1583-1592, July 2006.

[35] H. Kufluoglu, "Theory of interface-trap-induced NBTI degradation for reduced cross section MOSFETs," *IEEE Transactions on Electron Devices*, vol. 53, no. 5, pp. 1120-1130, May 2006.

[36] S. Mahapatra, "A New Observation of Enhanced Bias Temperature Instability in Thin Gate Oxide p-MOSFETs," *IEDM Tech. Dig.*, p. 337, 2003.

[37] C. Y. Lu, "Reverse Short-Channel Effects on Threshold Voltage in Submicrometer Salicide Devices," *IEEE Electron Device Letters*, vol. 10, no. 10, pp. 446-448, October 1989.

[38] A. Sharma, "Multifinger MOSFETs' Optimization Considering Stress and INWE in Static CMOS Circuits," *IEEE Transactions on Electron Devices*, vol. 63, no. 6, pp. 2517-2523, June 2016.

[39] M. Miyamoto, "Impact of reducing STI-induced stress on layout dependence of MOSFET characteristics," *IEEE Transactions on Electron Devices*, vol. 51, no. 3, pp. 440-443, March 2004.

[40] H. Lee, "An anomalous device degradation of SOI narrow width devices caused by STI edge influence," *IEEE Transactions on Electron Devices*, vol. 49, no. 4, pp. 605-612, April 2002.

[41] C. Y. Chan, "Impact of STI Effect on Flicker Noise in 0.13-μm RF nMOSFETs," *IEEE Transactions on Electron Devices*, vol. 54, no. 12, pp. 3383-3392, December 2007.

[42] Y. M. Sheu, "Impact of STI mechanical stress in highly scaled MOSFETs," in *2003 International Symposium on VLSI Technology, Systems and Applications. Proceedings of Technical Papers*, 2003, pp. 269-272.

[43] G. Scott, "NMOS Drive Current Reduction Caused by Transistor Layout and Trench Isolation Induced Stress," in *IEDM Tech. Dig.*, 1999, pp. 827-830.

[44] M. Choi, "Impact on circuit performance of deterministic within-die variation in nanoscale semiconductor manufacturing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 7, pp. 1350-1367, July 2006.

[45] A. K. Wong, "Microlithography: trends, challenges, solutions, and their impact on design," *IEEE Micro*, vol. 23, no. 2, pp. 12-21, March 2003.

[46] A. K. Wong, *Resolution Enhancement Techniques in Optical Lithography*. Bellingham WA: SPIE Press, 2001.

[47] S. Saxena, "Variation in Transistor Performance and Leakage in Nanometer-Scale Technologies," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 131-144, January 2008.

[48] S. Eneman, "Layout impact on the performance of a locally strained PMOSFET," in *Digest of Technical Papers. 2005 Symposium on VLSI Technology*, Kyoto, 2005, pp. 22-23.

[49] J. Czochralski, "Ein neues Verfahren zur Messung der Kristallisationsgeschwindigkeit der Metalle," *Z. Phys. Chem.*, vol. 92, pp. 219-221, 1918.

[50] L. Xinfu, "A study of inverse narrow width effect of 65nm low power CMOS technology," in *2008 9th International Conference on Solid-State and Integrated-Circuit Technology*, Beijing, 2008, pp. 1138-1141.

[51] S. R. Nassif, "Within-Chip Variability Analysis," in *IEDM Tech. Dig.*, 1998, pp. 283-286.

[52] A. Tsiamis, "Electrical Test Structures for the Characterization of Optical Proximity Correction," *IEEE Transactions on Semiconductor Manufacturing*, vol. 25, no. 2, pp. 162-169, May 2012.

[53] T. Linton, "Determination of the line edge roughness specification for 34 nm devices," in *Digest. International Electron Devices Meeting*, San Francsico, 2002, pp. 303-306.

[54] J. A. Croon, "Line edge roughness: characterization, modeling and impact on device behavior," in *Digest. International Electron Devices Meeting*, San Fransisco, 2002, pp. 307-310.

[55] T. Mizuno, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *IEEE Transactions on Electron Devices*, vol. 41, no. 11, pp. 2216-2221, November 1994.

[56] Pelgrom M, "Matching properties of MOS transistors," in *ESSCIRC*, Manchester, November 1998, pp. 327-330.

[57] S. Wolf, "Silicon Processing for the VLSI Era - Vol. 2 - Process Integration," p. 389.

[58] T. B. Hook, "Lateral ion implant straggle and mask proximity effect," *IEEE Transactions on Electron Devices*, vol. 50, no. 9, pp. 1946-1951, September 2003.

[59] P. G. Drennan, "Implications of Proximity Effects for Analog Design," in *IEEE Custom Integrated Circuits Conference 2006*, San Jose, pp. 169-176.

[60] Y. M. Sheu et al, "Modeling the Well-Edge Proximity Effect in Highly Scaled MOSFETs," *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2792-2798, November 200.

[61] R. R. Troutman, "VLSI limitations from drain-induced barrier lowering," *IEEE Transactions on Electron Devices*, vol. 26, no. 4, pp. 461-469, April 1979.

[62] S. Narendra, "Full-chip sub-threshold leakage power prediction model for sub-0.18 µm CMOS," *Proceedings of the International Symposium on Low Power Electronics and Design*, vol. 39, no. 2, pp. 19-23, Febrauary 2002.

[63] W. R. Bandy, "A simple approach for accurately modeling the threshold voltage of short-channel mosts," *IEEE Journal Solid State Electronics*, vol. 20, no. 8, pp. 675-680, August 1977.

[64] Y. Bin, "Short-channel effect improved by lateral channel-engineering in deep-submicronmeter MOSFET's," *IEEE Transactions on Electron Devices*, vol. 44, no. 4, pp. 627-634, April 1997.

[65] B. C. Paul, "Device Optimization for Digital Subthreshold Logic Operation," *IEEE Transactions on Electron Devices*, vol. 52, no. 2, pp. 237-247, February 2005.

[66] A. P. Chandrakasan, "Technologies for Ultradynamic Voltage Scaling," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 191-214, February 2010.

[67] T. H. Kim, "Utilizing Reverse Short-Channel Effect for Optimal Subthreshold Circuit Design," *IEEE Transactions on Very Large Scale Integration (VLSI)*

Systems, vol. 15, no. 7, pp. 821-829, July 2007.

[68] J. S. Wang, "RSCE-aware ultra-low-voltage 40-nm CMOS circuits," in *2011 International SoC Design Conference*, Jeju, 2011, pp. 131-134.

[69] F. M. Klaassen, "Narrow-width Effects in Submicron MOS ICs," in *ESSDERC '89: 19th European Solid State Device Research Conference*, Berlin, 1989, pp. 105-108.

[70] P. A. H. Hart, "Device down scaling and expected circuit performance," *IEEE Journal of Solid-State Circuits*, vol. 12, no. 2, pp. 343-357, April 1979.

[71] L. Akers, "The Inverse-Narrow-Width Effect," *IEEE Electron Device Letters*, vol. 7, no. 7, pp. 419-421, July 1986.

[72] N. Shigyo, "A Review of Narrow-Channel Effects for STI MOSFET's: A Difference Between Surface- and Buried-Channel Cases," *Solid State Electronics*, vol. 43, May 1999.

[73] K. K. L. Hsueh, "Inverse-narrow-width effects and small-geometry MOSFET threshold voltage model," *IEEE Transactions on Electron Devices*, vol. 35, no. 3, pp. 325-338, March 1988.

[74] C. Pacha et al, "Impact of STI-induced stress, inverse narrow width effect, and statistical VTH variations on leakage currents in 120 nm CMOS," in *Proceedings of the 30th European Solid-State Circuits Conference*, 2004, pp. 397-400.

[75] J. Zhou, "A 40 nm Dual-Width Standard Cell Library for Near/Sub-Threshold Operation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 11, pp. 2569-2577, November 2012.

[76] S. Borkhar, "Circuit techniques for subthreshold leakage avoidance, control and tolerance," *IEDM Technical Digest. IEEE International Electron Devices Meeting*, pp. 241-242, 2004.

[77] J. S. Wang, "RSCE-aware ultra-low-voltage 40-nm CMOS circuits," in *2011 International SoC Design Conference*, Jeju, 2011, pp. 131-134.

[78] R. Liao, "Digital circuit design for robust ultra-low-power cell library using optimum fingers," in *2012 IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Boise, ID, 2012, pp. 446-449.

[79] M. Pons, "Ultra low-power standard cell design using planar bulk CMOS in subthreshold operation," in *2013 23rd International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, Karlsruhe, 2013, pp. 9-15.

[80] J. Kwong, "Variation-Driven Device Sizing for Minimum Energy Sub-threshold Circuits," in *ISLPED'06 Proceedings of the 2006 International Symposium on Low Power Electronics and Design*, Tegernsee, 2006, pp. 8-13.

[81] J. Keane, "Subthreshold logical effort: a systematic framework for optimal subthreshold device sizing," in *2006 43rd ACM/IEEE Design Automation Conference*, San Francisco, 2006, pp. 425-428.

[82] N. Weste D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective.*, 2010.

[83] T. Niiyama, "Increasing minimum operating voltage (VDDmin) with number of CMOS logic gates and experimental verification with up to 1Mega-stage ring

oscillators," in *2008 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, Bangalore, 2008, pp. 117-122.

[84] W. H. Kao, "Parasitic extraction: current state of the art and future trends," in *The 2001 IEEE International Symposium on Circuits and Systems*, vol. 5, Sydney, 2001, pp. 487-490.

[85] A. Sharma, S. Miryala and A. Bulusu C. I. Kumar, "A novel energy-efficient self-correcting methodology employing INWE," in *2016 13th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, Lisbon, 2016, pp. 1-4.

[86] M. Muker and M. Shams, "Designing digital subthreshold CMOS circuits using parallel transistor stacks," *Electronics Letters*, vol. 47, no. 6, pp. 372-374, March 2011.

[87] S. Jayapal, B. Busze, L. Huang and J. Stuyt J. Zhou, "A 40 nm inverse-narrow-width-effect-aware sub-threshold standard cell library," in *2011 48th ACM/EDAC/IEEE Design Automation Conference (DAC)*, New York, 2011, pp. 441-446.

[88] J. Myers, "8.1 An 80nW retention 11.7pJ/cycle active subthreshold ARM Cortex-M0+ subsystem in 65nm CMOS for WSN applications," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, San Francisco, 2015, pp. 1-3.

[89] S. Hanson et al, "Exploring Variability and Performance in a Sub-200-mV Processor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 881-891, April 2008.

[90] S. K Springer, "Modeling of Variation in Submicrometer CMOS ULSI Technologies," *IEEE Transaction on Electron Devices*, vol. 53, no. 9, pp. 2168-2178, September 2006.

[91] M. Quarantelli, "Characterization and modeling of MOSFET mismatch of a deep submicron technology," in *International Conference on Microelectronic Test Structures*, 2003, pp. 238-243.

[92] S. Saxena, "Test structures and analysis techniques for estimation of the impact of layout on MOSFET performance and variability," in *Proceedings of the 2004 International Conference on Microelectronic Test Structures*, 2004, pp. 263-266.

[93] T. Park, "Overview of methods for characterization of pattern dependencies in copper CMP," in *Proc. Chemical Mechanical Polish ULSI Multilevel Interconnection Conf*, Santa Clara, 2000, pp. 196–205.

[94] Y. Chen, "Area fill synthesis for uniform layout density," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 10, pp. 1132-1147, October 2002.

[95] M. Orshansky, "A general probabilistic framework for worst case timing analysis," in *Proceedings 2002 Design Automation Conference*, 2002, pp. 556-561.

[96] J. Kwong, "A 65nm Sub-Vt Microcontroller with Integrated SRAM and Switched-Capacitor DC-DC Converter," in *2008 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, San Francisco, 2008, pp. 318-616.

[97] K. A. Bowman, "Impact of Extrinsic and Intrinsic Parameter Fluctuations on CMOS Circuit Performance," *IEEE JSSC*, vol. 35, no. 8, pp. 1186-1193, August 2000.

[98] B. H. Calhoun, "Characterizing and modeling minimum energy operation for subthreshold circuits," in *Proceedings of the 2004 International Symposium on Low Power Electronics and Design*, Newport Beach, 2004, pp. 90-95.

[99] Ferdinand Braun, "Uber die Stromleitung durch Schwefelmetalle," *Ann. Phys. Chem.*, vol. 556, p. 153, 1874.

[100] W.H.Brattain J.Bardeen, "The Transistor, a Semiconductor Triode," *Phys. Rev.*, vol. 230, p. 71, 1948.

[101] J. S. Kilby, "Invention of the Integrated Circuit," *IEEE Trans Electron Devices*, vol. ED-23, p. 648, 1976.

[102] R. N. Noyce, "Semiconductor Device-and-Lead Structure," 2,981,877, 1959.

[103] D. Kahng M. M. Atalla, "Silicon-Silicon Dioxide Surface Device," in *IRE Device Research*, Pittsburgh, 1960, pp. 583-596.

[104] G. E. Moore, "Cramming More Components onto Integrated Circuits," *IEEE Solid-State Circuits Society Newsletter*, vol. 38, no. 8, p. 114, April 1965.

Appendix A: Historical Overview of Semiconductor Devices

Semiconductor Physics – The Origin

The predisposition of the 21st century engineer is to suppose that the advent of the study of modern semiconductor physics had its origins in the mid 20th century. In fact, the first notable contribution was by Ferdinand Braun in 1874, who systematically observed the dependence of resistance on the polarity and magnitude of the potential drop across the heterogeneous junctions of metals and metal sulfides [99]. This misconception is understandable, as any previous work in the field was subsequently dwarfed by the contributions of Bardeen and Brattain with their unveiling of the world's first (Point-Contact) Bipolar Junction Transistor [100] and William Shockley's consequent analytical evaluation of the P-N Junction [8], an endeavor which saw the trio receive the Nobel prize in physics in 1956. By 1960, patented designs of the first monolithic integrated circuits were beginning to emerge [101] [102] and the subsequent discovery of the Metal Oxide Semiconductor Field Effect Transistor by Kahng and Atalla [103] propelled integrated circuit design in the phenomenon we know today.

VLSI Design and the Era of Gordon Moore

In order to drive the nascent integrated circuit industry towards a high complexity, low cost and low power consumption design base, the industry committed to aggressively scaling device dimensions at a rate approximate to the square root of two every two years as outlined in Gordon Moore's prophetic 1965 paper [104].

The non-sustainability of power density in the founding technologies of Bipolar Junction Transistors and nMOS eventually led to the supremacy of the Complementary Metal Oxide Semiconductor (CMOS) technology during the 1990's, which has been the driving technology ever since. Monolithic integrated circuits evolved with the gradual introduction of processing technologies from designs consisting of a handful of components, to single dice consisting of billions of devices known as Very Large Scale Integrated (VLSI) circuits.