SUBTHRESHOLD CIRCUIT DESIGN AND OPTIMIZATION

Except where reference is made to the work of others, the work described in this thesis is my own or was done in collaboration with my advisor. This thesis does not include proprietary or classified information.

_____

Sungil Kim

Certificate of Approval

_____	_____
Vishwani D. Agrawal James J. Danaher Professor Department of Electrical and Computer Engineering	Melissa J. Baumann Director The Honors College

SUBTHRESHOLD CIRCUIT DESIGN AND OPTIMIZATION

Sungil Kim

A Thesis

Submitted to the

Honors College at Auburn University

In Partial Fulfillment of the

Requirement for

University Honors Scholar

Auburn, Alabama

May 8, 2016

SUBTHRESHOLD CIRCUIT DESIGN AND OPTIMIZATION

Sungil Kim

_____
Signature of Author

_____
Date of Graduation

VITA

Sungil Kim was born in Daegu, South Korea on February 11, 1995. He and his family immigrated to the United States in 2009. He graduated from the Loveless Academic Magnet Program (LAMP) High School in Montgomery, Alabama. As the first generation to attend college, he entered Auburn University, majoring in Electrical and Computer Engineering. His extracurricular activities include membership in Phi Kappa Phi, Tau Beta Pi, Eta Kappa Nu, Pi Gamma Tau, IEEE, and ACM. His research interest is integrated circuit (IC) and systems design with an emphasis on low power or energy-efficient circuits and analog/RF IC design. Other interests include VLSI architectures for signal processing, bio-sensing, and biomedical electronics. As an undergraduate, he actively participated in research, teaching, and mentoring. As an independent researcher, he surveyed subthreshold circuit design and optimization for his honors thesis and has worked on his undergraduate research fellowship that focuses on analysis and forecast of stock market prices using signal and statistical analysis and genetic algorithm. Other collaborative research activities include being a research assistant in the VLSI Design/Automation laboratory at the University of Michigan, the Auburn University MRI center, Nanotech group, and Hill laboratory. He expects to graduate Summa Cum Laude with a Bachelor of Electrical Engineering (with Computer Option) on May 8, 2016. In the fall of 2016, he will seek a Ph.D. in Electrical Engineering with the focus on IC design.

THESIS ABSTRACT

SUBTHRESHOLD CIRCUIT DESIGN AND OPTIMIZATION

Sungil Kim

Bachelor of Electrical Engineering, May 8, 2016

78 Typed Pages

Directed by Prof. Vishwani D. Agrawal

Modern electronics, whether medical imaging electronics, sensors, portable devices, or high-performance computers, are constrained by their power. It is well known that subthreshold circuit design where the supply voltage is less than the device threshold voltage can reduce the energy. That power reduction comes with significant performance drawback and susceptibility to process, voltage, and temperature (PVT) variations.

The energy saving and operation of the subthreshold circuit is demonstrated, and its advantages and limitations are discussed here. The analytical model to estimate circuit delay is also analyzed, particularly Alpha-Power Law. The estimated circuit delay by Alpha-Power Law is proven to be not effective in the subthreshold region because the subthreshold drain current exponentially depends on the gate-source voltage and subjects to PVT variations. To better estimate the circuit delay and understand the effect of variations in the subthreshold region, a variations-aware analytical model is proposed and verified through simulations. It is found that the circuit delay exponentially depends on

$\frac{\lambda(V_{DD2}-V_{DD1})}{2mV_T}$, where λ is the drain-induced barrier lowering (DIBL) coefficient, m is the subthreshold slope factor, and V_T is the thermal voltage. In a 45nm BSIM bulk CMOS technology from the PTM model [21], λ is 0.001 and m is 27. With an average error of 15% in the subthreshold region, the proposed analytical model is proved to be a more effective measure of subthreshold circuit delay. Moreover, the effect of variations is analyzed, and the smaller technology nodes are found to have greater errors.

To optimize both performance and power consumption, dual-threshold circuit design is explored, and a gate assignment algorithm is formulated using linear optimization (linear programming). The usage of both low threshold gates (fast and greater leakage power) and high threshold gates (slow and less leakage power) can improve performance while leakage power is reduced, and the circuit still operates properly.

Because wire capacitance does not scale with the supply voltage, and global wire delay is increasing with technology scaling, on-chip global interconnects causes significant performance degradation. Therefore, two techniques—repeater insertions and tapered interconnect driver—are discussed. In the subthreshold region where the driver delay dominates the overall interconnect delay, repeater insertions that superthreshold interconnects often use proved to be ineffective. An optimally sized, tapered driver can reduce up to 75% of the power-delay product. Also, the effect of interconnect length is discussed, and it is found that as interconnect length increases, a tapered driver is more effective.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Dr. Vishwani Agrawal, for guiding my research and advising and encouraging me. Throughout the process, my appreciation and interest in integrated circuit design have greatly increased. Also, I would like to thank Auburn University Honors College and the Department of Electrical and Computer Engineering for this undergraduate research opportunity. Finally, I would like to thank my family and friends for their encouragement, motivation, and inspiration.

Style manual or journal used Auburn University's Graduate School's *Guide to Preparation and Submission of Theses and Dissertations.* Bibliography follows *Institute of Electrical and Electronics Engineers (IEEE) Citation Reference* and is sorted in alphabetical order

Computer software used Microsoft Word

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

Moore's law describes the rapidly increasing trend in the number of transistors in an integrated circuit as specifically doubling every two years [9]. Concerns about power dissipation and its management are important in modern technology, especially for medical electronics and sensors that require ultralow power consumption [5, 17].

As one of the method to potentially solve the power consumption issue, this paper explores the subthreshold operation region. Although subthreshold operation design reduces power consumption according to past studies, its performance degradation and susceptibility to noise and variations of temperature have prevented its application. Thus, dual-threshold subthreshold circuit design and the optimization of global interconnects are discussed in this paper. Further, the analytical model for delay of the circuit (Alpha-Power law) is analyzed, and a more accurate analytical model is proposed for the subthreshold region.

This paper begins with a discussion of the types of power dissipation and methods to minimize power and the benefit of subthreshold operation CMOS design and its power advantage in terms of low leakage current and less power consumption. It further discusses the advantage of subthreshold circuit in energy saving. Then, two major problems of subthreshold CMOS design are reviewed, i.e., performance drawback and sensitivity to process/voltage/temperature (PVT) variation. It proposes a new design to

deter performance drawback using dual-threshold circuit design and on-chip global interconnect optimization. The types of variations are explained, and the Alpha-Power law is verified through simulations. The analytical model for a subthreshold circuit is analyzed, and a more accurate, variations-aware analytical model is proposed. The comparison between the subthreshold and near-threshold circuit design is also examined.

Chapter 2

Background of Power Dissipation and Estimation

This Chapter begins with a summary of the components of power dissipation, the effect of technology sizing and the variation of other parameters (such as threshold voltage), and the methods used to reduce power consumption. In Chapter 2.1, three types of power dissipation and the components within each type are analyzed. The effect of transistor sizing is discussed in Chapter 2.2. Chapter 2.3 examines several techniques for reducing power consumption.

2.1. Types of Power Dissipation

There are two types of power dissipation—dynamic and static power. Dynamic power weighs more than static power.

$$P_{total} = P_{dyn} + P_{stat} = P_{trans} + P_{sc} + P_{stat} \quad (2.1)$$

Dynamic power occurs during signal transitions (P_{trans}). Dynamic power is dissipated mainly by logic activity, glitches, and short-circuits (P_{sc}). Static power (P_{stat}), also known as leakage, occurs, however, regardless of whether the circuit is transitioning.

2.1.1. An Estimation of Transition Power

Transition power can be estimated using Equation 2.2. As electronics operate more frequently without any idle (sleep) mode, they will have more transitions, increasing the

activity factor α. In modern biomedical electronics (i.e. blood pressure monitor), the activity factor and operating frequency are low, and thus, the constraint on power consumption is more important than is the performance requirement.

$$P_{trans} = \frac{\alpha f_{ck} C V^2}{2} \text{ , } \alpha = \text{activity factor and } f_{ck} = \text{clock frequency} \quad (2.2)$$

2.1.2. An Estimation of Short-Circuit Power

Short-circuits power can be estimated using equation 2.3.

$$P_{sc} = \alpha f_{ck} E_{sc} \quad (2.3)$$

$$E_{scf} = \int_{t_B}^{t_E} V_{DD} i_{sc}(t) dt = \frac{(t_E - t_B) I_{scmaxf} V_{DD}}{2} = \frac{t_f (V_{DD} - |v_{Tp}| - V_{Tn}) I_{scmaxf}}{2}$$

$$E_{scr} = \frac{t_r (V_{DD} - |v_{Tp}| - V_{Tn}) I_{scmaxr}}{2} \quad (2.4)$$

Short circuit current increases with the size of the transistors and decreases with load capacitance. It is largest when load capacitance equals zero [19]. Short circuit power can be often negligible because most of the power dissipation depends heavily on dynamic and static power. Also, it should be noted that short circuit power remains the smallest fraction of total power regardless of whether the circuit operates in superthreshold or subthreshold operation region.

Short circuit power increases with rise and fall times of input and decreases for larger output load capacitance because a large capacitor takes most of the current. Short circuit power is approximately 5~10% of dynamic power.

2.1.3. An Estimation of Static Power

$$P_{static} = I_{static} V_{DD} \quad (2.5)$$

$$I_{static} = I_{sub} + I_D + I_G + I_{PT} + I_{GIDL} \quad (2.6)$$

As shown in equation 2.6, static current can be estimated by adding five components: the subthreshold conduction, reverse bias PN junction conduction, gate tunneling, drain source punchthough due to short channel and high drain-source voltage, and gate induced drain leakage due to tunneling at the gate-drain overlap. The two major components are I_{sub} and I_D while I_G tends to become a greater fraction of total power with scaling. In the subthreshold region, I_{sub} increases, and leakage dominates.

2.2. Effects of Transistor Sizing

In today's submicrometer circuits, the sizing of the transistor significantly affects power consumption of the circuit. Because the parasitic capacitances and the velocity saturation effect are larger in submicrometer circuits, the power reduction achievable from the transistor sizing ranges from 50% to 30% [23]. When this technique is coupled with technology scaling, even larger savings is achievable. However, the critical delay of the original circuit should remain the same to ensure there is no timing error in the circuit.

2.3. Power-Reducing Mechanisms

Using the above equations, there are various power-reducing mechanisms for reducing supply voltage to device scaling. Because supply voltage can reduce the dynamic power quadratically, it has been a popular mechanism to use to reduce the total dissipated power and energy. However, reducing supply voltage to the point of the subthreshold region increases the subthreshold current and leakage with an expense of performance. Thus, it

suits some electronic devices, such as medical sensors, which do not need higher frequency or stay on and off. Further, it suits mobile devices (cell phones) since the reduction of supply voltage ensures portability and increases battery life.

Another mechanism includes architectural modification, such as transistor sizing. As more and more transistors are equipped in devices due to more advanced packaging and sizing technology, technologies have been sized down tremendously, now heading toward less than 10 ns. Reduction of size has also been proven to be power-efficient as discussed in Chapter 2.2.

Chapter 3

Subthreshold Circuit Operation

Subthreshold, or weak inversion, circuit design uses a supply voltage that is less than the threshold voltage of the transistors. Unlike the traditional circuit design that uses a supply voltage greater than a threshold voltage, a subthreshold circuit design saves energy at the expense of performance. In this Chapter, the origin, motivation, advantages, and limitations of the subthreshold circuit design are discussed.

3.1. Origin of Subthreshold CMOS Design

While the weak inversion region had been ignored for years, the need to limit the power consumption of the electronic watch drew the attention of the digital design community [27]. While modern electronics still heavily operate in a superthreshold region, or a strongly inverted region, due to their performance specifications, some devices that require low to ultra-low power consumption or have a low activity factor are ideal devices to operate in the subthreshold or near-threshold region.

Watchmaker's Electronic Center first started its design in bipolar technology and then switched to CMOS technology. After characterizing MOS transistors at a very low drain current level, they demonstrated that the drain current exponentially depends on the gate voltage (V_{gs}) as shown in Figure 3.1 below.

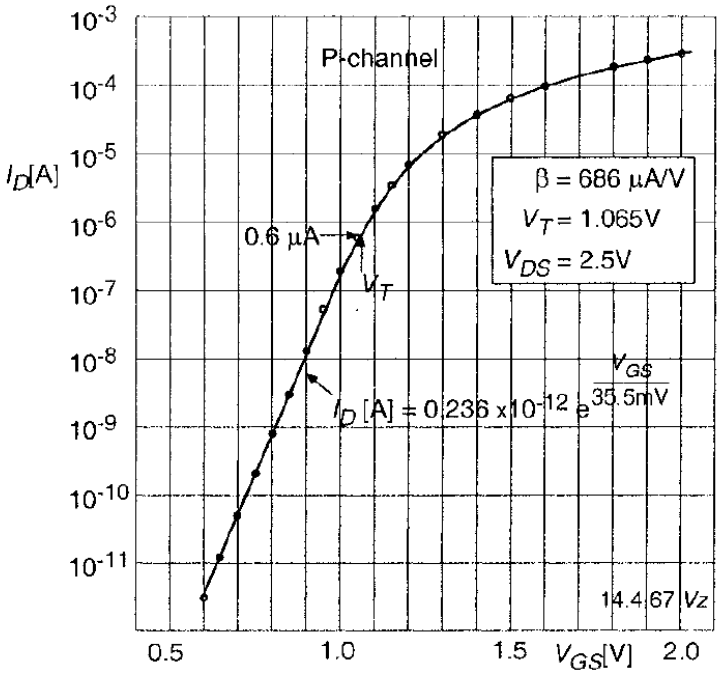

Figure 3.1: Early measurement of the $l_D(V_{GS})$ characteristics of a P-channel metal-gate MOS transistor (cleaned-up plot from E. Vittoz' notebook, CEH, 1967) [28].

Then, in 1972, Barron demonstrated the exponential dependence on the surface potential. Swanson and Meindl explained this effect by relating it to the gate voltage and a capacitive divider. Their paper in 1976 at the second European Solid-State Circuits Conference showed that the compact model characterizeed the drain current as Equation 3.1.

$$I_D = I_{D0}\frac{W}{L}e^{\left(\frac{kV_g}{U_T}\right)}\left(e^{\frac{V_S}{U_T}} - e^{-\frac{V_D}{U_T}}\right) \tag{3.1}$$

While the utilization of a subthreshold was heavily limited to analog circuits in the 1980s, greater importance being placed on the management of power allowed the exploration of the subthreshold circuit operation. Thus, in the 21st century, many devices

that do not have a high activity level employ this technique to reduce both dynamic power and overall power consumption.

3.2. Motivation

Subthreshold is the region below the threshold voltage that can operate with a current less than the superthreshold current, thus achieving low energy with a performance penalty in terms of frequency. When using a supply voltage (V_{DD}) less than a threshold voltage, transistors have a much smaller ratio of $\frac{I_{on}}{I_{off}}$. Thus, the subthreshold circuit design reduces dynamic power and saves active energy quadratically because power is dependent on supply voltage squared [9]. However, subthreshold current tends to become higher when the circuit operates in the subthreshold region, suggesting that leakage of current and delay will vary significantly due to slight changes in V_{DD}. Thus, a clear understanding of the susceptibility to variations can make subthreshold circuit design more reliable and applicable to various devices. Furthermore, algorithms to deter a performance penalty while operating in subthreshold or near-threshold region are essential for broad applications of subhtreshold circuit design.

3.3. Advantages

The power saving of the subthreshold circuit design has been tested through the simple CMOS inverter. The CMOS inverter was designed in Mentor Graphics Design Architecture after importing the Verilog netlist synthesized on LeonardoSpectrum using TSMC 0.18 μm technology with optimization for delay. Then, the operation was tested

using HSPICE by generating the transfer function and observing the output in EZWAVE. As shown in Figure 3.3, the input was inverted to output with some time delay.

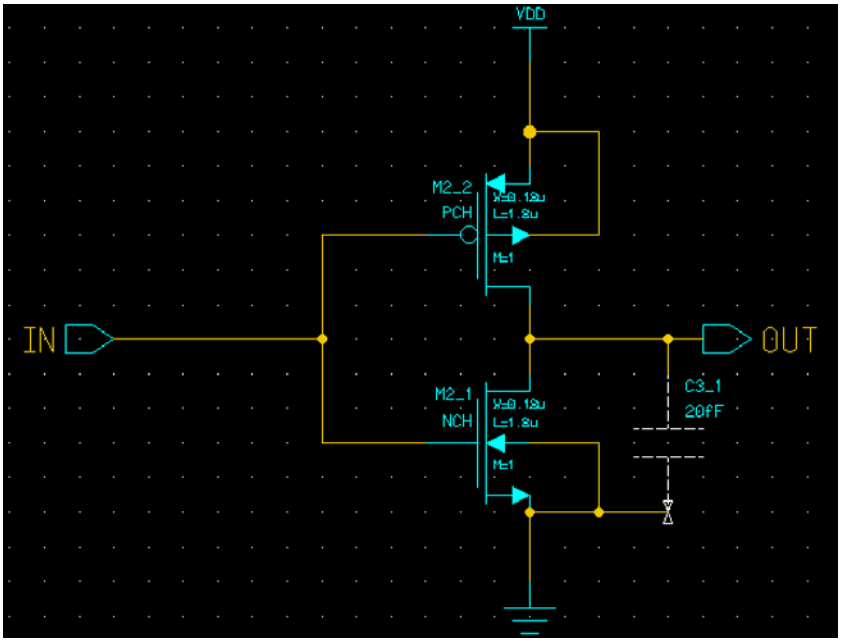

Figure 3.2: CMOS inverter

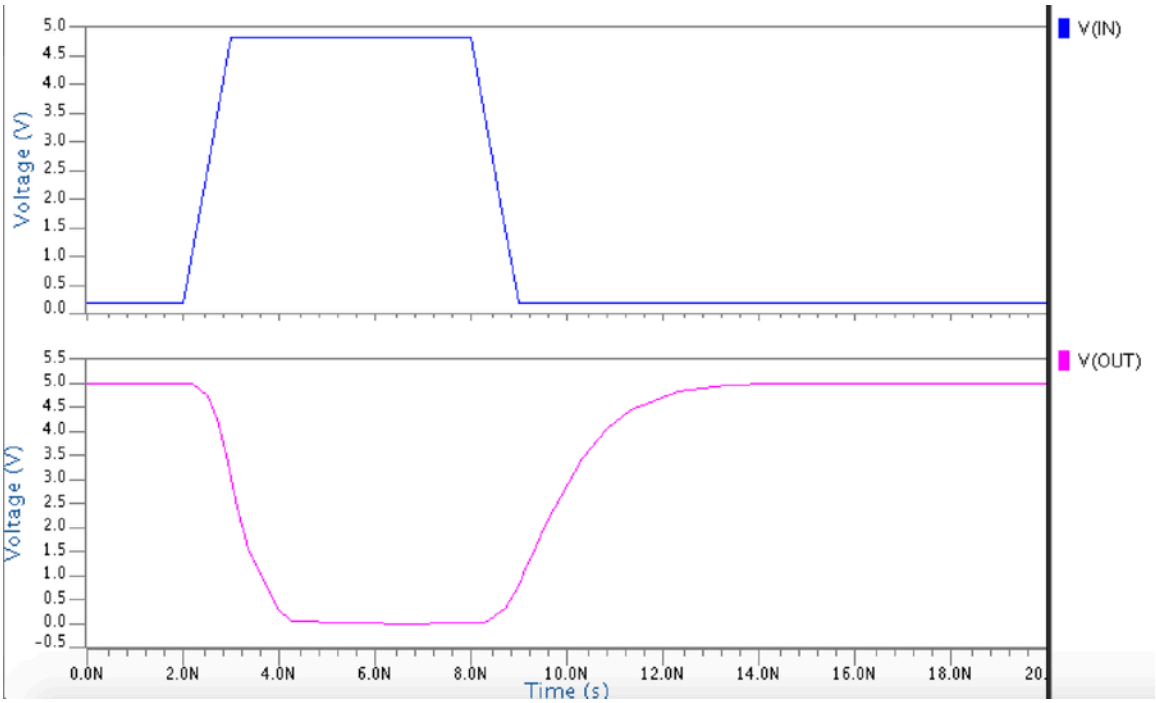

Figure 3.3: CMOS inverter characteristics

The DC analysis is shown in Table 3.1. As the supply voltage decreases, the power consumption decreases significantly while the delay increases. Since the supply voltage quadratically influences the dynamic (transitioning) power as shown in Figure 3.4, the subthreshold circuit design reduces the overall power by reducing the dynamic power quadratically. Further, because the dynamic power composes the most of the total power consumption, reducing the dynamic power is effective. Additionally, the CMOS circuit operates correctly under a variation of timing, supply voltage, temperature, and process, while it is, however, more susceptible to the variations.

$$P_{trans} = \frac{\alpha f_{ck} C V^2}{2}$$, α = activity factor and f_{ck} = clock frequency

Table 3.1. Power for various voltages of the CMOS inverter

Supply Voltage (V)	Total Power (pW)
0.1	0.01
0.2	0.08
0.3	0.18
0.4	0.32
0.5	0.50
0.6	0.72
0.7	0.98
0.8	1.28
0.9	1.60
1.0	2.00
1.5	4.50
2.0	8.00
2.5	12.00
3.0	18.00
3.5	24.00
4.0	32.00
4.5	40.00
5.0	50.00

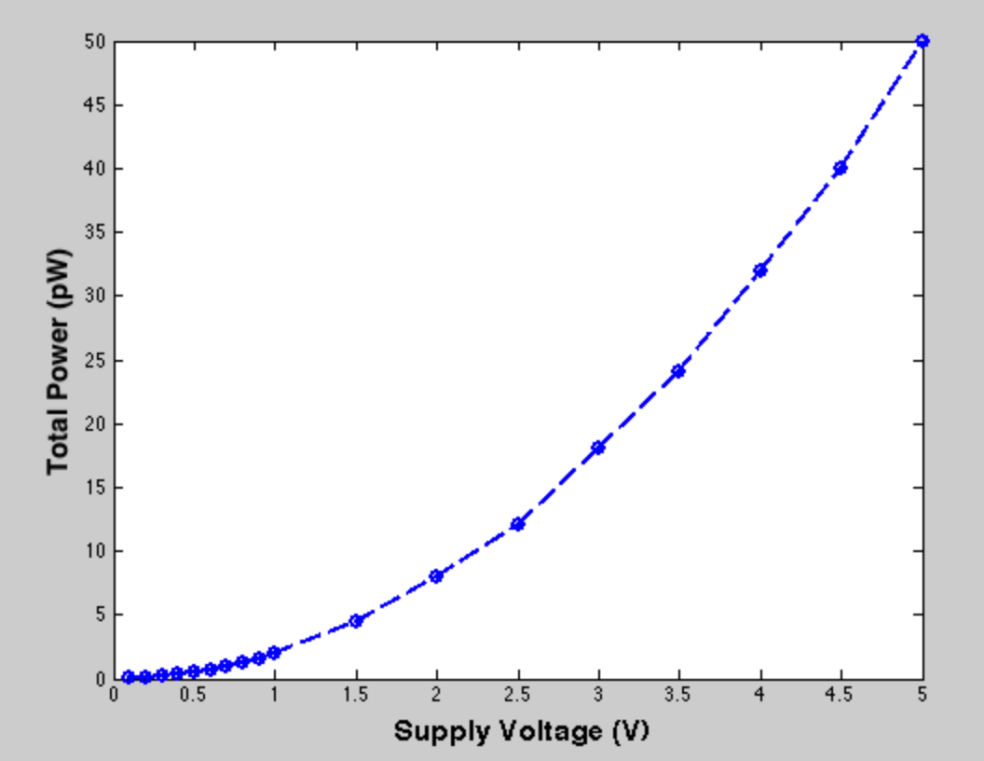

Figure 3.4: Quadratic growth of total power with an increase of V_{DD}

A 45nm high performance PTM model [21] is used to analyze the energy efficiency at the subthreshold region. At a supply voltage of 0.45V, the optimal power-delay product is achieved. For this model, the threshold voltage is 0.47V. The power saving at this region compared to 1V is 95%, and the energy saving is 65%.

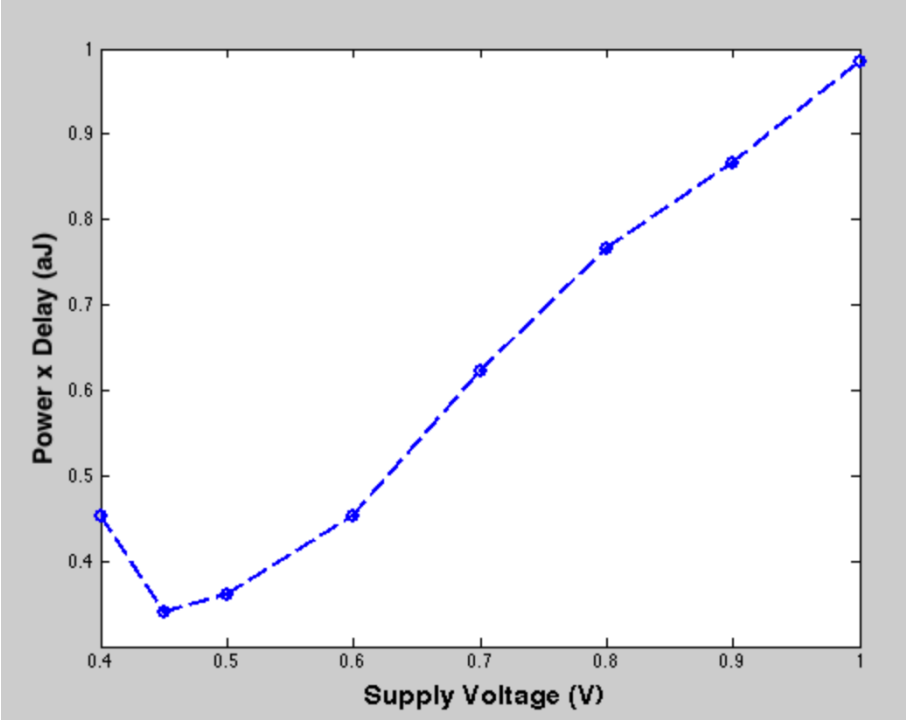

Figure 3.5: The power-delay product for the CMOS inverter

The benchmark circuit is also used to demonstrate the optimal operation region defined as the minimum power-delay product. In this case, a 16 x 16 bit multiplier (ISCA85 c6288) with 32nm LP PTM model [21] was used. This technology has a threshold voltage of 0.63V. Obtained from a HSPICE simulation, test vectors including critical path delay were applied. The power and delay is shown in Table 3.2. Shown in Figure 3.4, when compared to the 1V supply voltage operation (above threshold), the operation at the subthreshold voltage (specifically at 0.55V) achieves a 99% reduction in total power. The simulation verifies that the circuits behave correctly at the subthreshold region. Also, an 8-bit and 16-bit ripper carry adder has a minimum energy operation point, where the circuit consumes the least energy per cycle, below the threshold voltage [13, 14]. Thus, when energy consumption is in main concern, operating at subthreshold supply voltage is ideal.

Table 3.2. Power and delay for various voltages of the multiplier

Supply Voltage (V)	Total Power (μW)	Delay (ns)	Power-Delay Product
0.20	0.05	4700	235
0.40	0.15	120	18.0
0.45	0.60	41	24.6
0.50	1.40	12	16.8
0.55	1.80	4	7.20
0.60	6.00	1.60	9.60
0.80	110	0.18	19.8
1.00	640	0.07	44.8

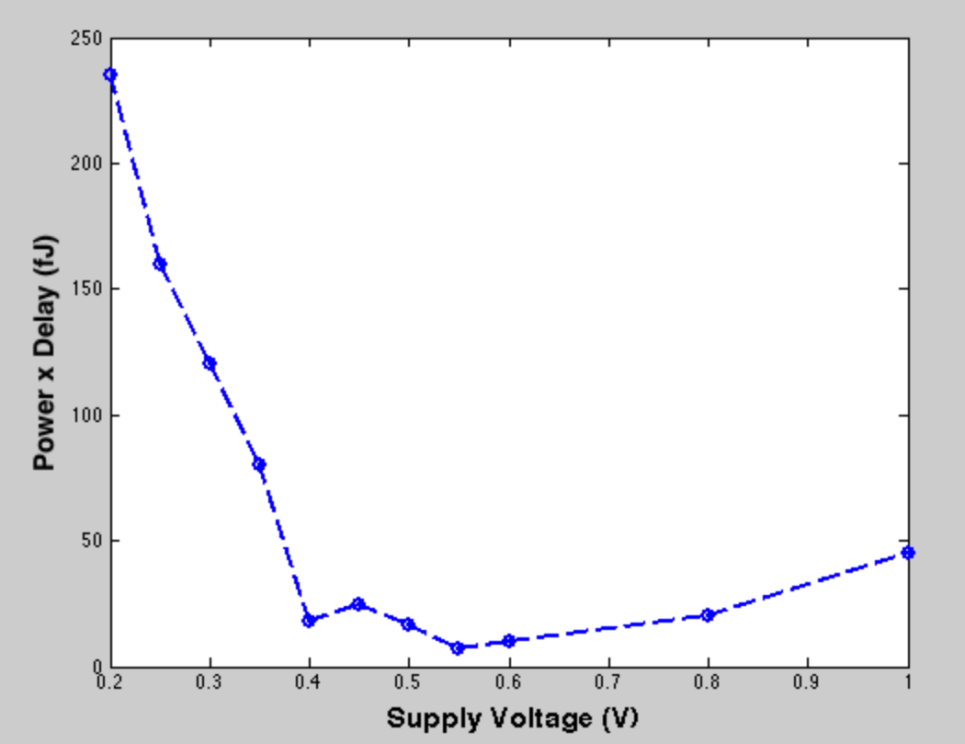

Figure 3.6: The power-delay product for a multiplier

While the subthreshold operation region does not have any area overhead, its effect on variations needs to be analyzed. After NMOS was characterized as shown in Figure 3.5, the DIBL (drain-induced barrier lowering) effect on the MOS transistors was analyzed. As the technology node decreases, greater leakage current (thus leakage power) occurs

due to a shorter channel length between the source and the drain. In the subthreshold region, the delay of the circuit increases exponentially, and leakage current decreases. Figure 3.6 shows that as V_{ds} decreases from 0.9V to 0.2V, the drain current also decreases. That also means that as V_{ds} increases, the threshold voltage decreases. As shown in the DIBL effect, the mechanism used to reduce leakage power becomes more and more important in modern electronics.

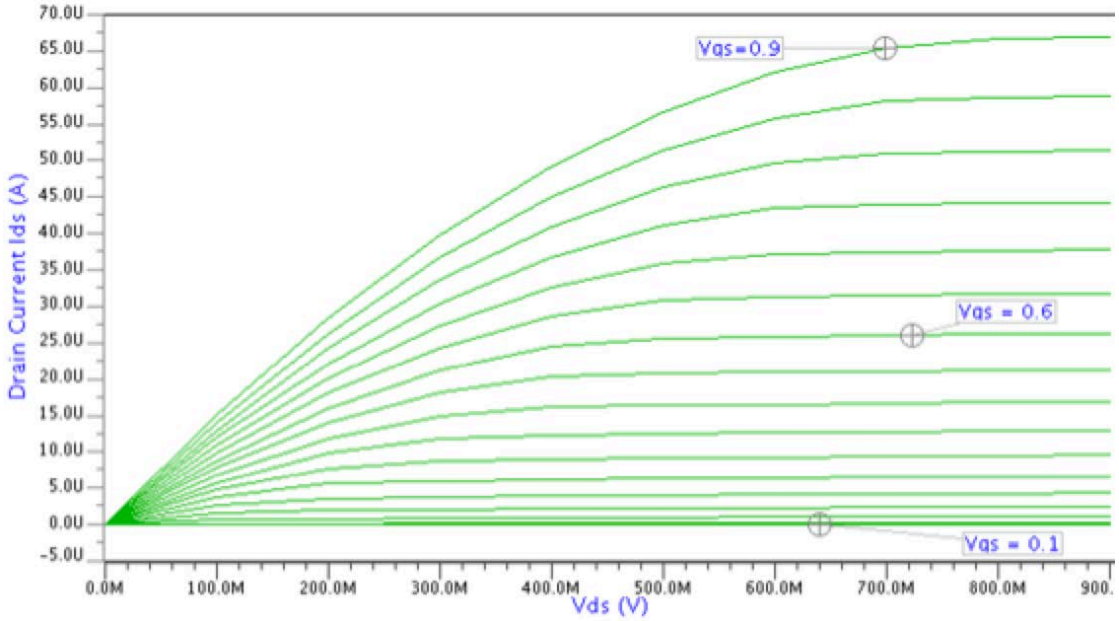

Figure 3.7: NMOS I-V characteristics

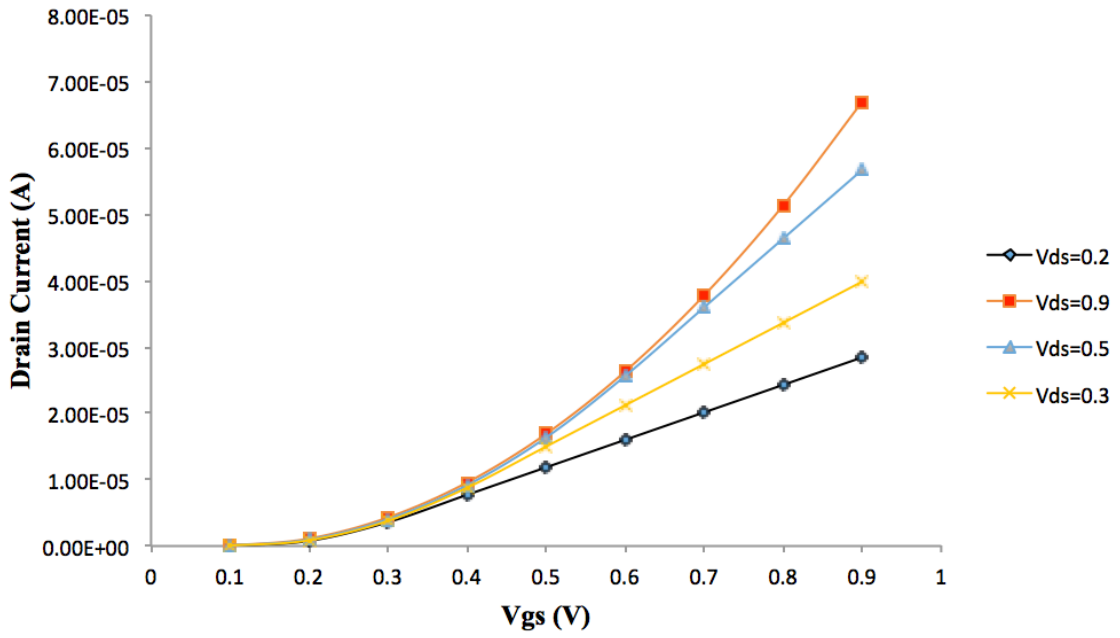

Figure 3.8: The DIBL effect

3.4. Limitations and Summary

Although the subthreshold circuit design reduces the dynamic and static power, it does have several drawbacks and limitations. The major limitation of a subthreshold circuit design is the rapid increase of subthreshold current and leakage and performance degradation. Further, global interconnect delays in the subthrehsold region dominate the overall circuit delay, resulting in performance degradation. Therefore, optimizing the subthrehsold interconnect is important. Also, as the dynamic power decreases quadratically, the ratio of static power to the total power dissipation also increases. Thus, the focus of further modification rests in a reduction of static power (leakage) and a mechanism to recover the performance drawback and susceptibility of PVT variations.

Chapter 4

Background of Variations

4.1. Types of Variations

There are mainly three types of variations predominant in a subthrehsold circuit, which are also known as PVT: Process, voltage, and temperature corners. Because of the exponential relationship between subthreshold current and threshold voltage variations, subthreshold circuit designs are highly sensitive to variations [12]. These variations cause variations in timing as well as in power for system on chip (SoC) designs [1]. Therefore, the assumption made during the design stage may not be valid under such fluctuations; thus additional efforts should be made to address the PVT variations. Previous research indicates a statistical timing analysis as well as the effect of PVT variability [10, 15]. Work by Zhai et al. was able to that random dopant fluctuations (RDF) and suggested design strategies to maintain variability levels of less than 30% while demonstrating a 24% energy reduction through pipelining [31].

4.2. Comparison to Near-threshold Design

When supply voltage scales down to the near-threshold voltages, both energy and performance decrease by a factor of 10 [5]. Further, from the near-threshold region to the subthreshold region, energy decreases by a factor of two, while time delay rises 50 to 100

times [5]. Thus, a comparison between the subthreshold and near-threshold circuit design is crucial when performance drawback is being considered.

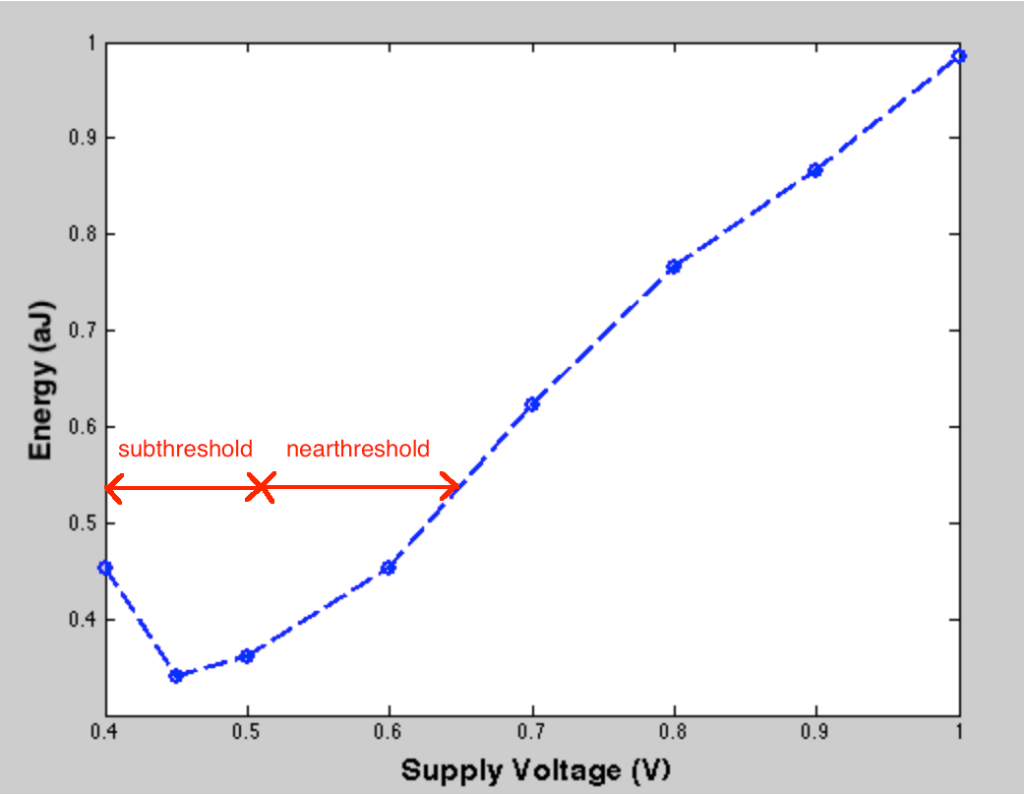

Figure 4.1: Subthreshold and nearthreshold region

4.3. Summary

While the subthreshold design is susceptible to variations, careful circuit design that considers and addresses such variations can ensure functionality of the circuit. In the subthreshold region, random dopant fluctuations dominate. Also, the proper usage of pipelining and system-level and architecture-level optimization will achieve variability-aware low-power design.

Chapter 5

Analytical Circuit Delay Model: The Alpha-Power Model and the Proposed Model

The Alpha-Power law developed by Takayasu Sakurai presents a simple, yet useful, characterization of MOSFET, which considers the velocity saturation effects of carriers and applies them to recent submicrometer technology. The model can be verified in various environments: such as via the variations of supply voltage (V_{DD}) and the size of the CMOS inverter chain. Our analysis concluded that The Alpha-Power law predicts propagation delay within 5% of the measured delay; however, it is not applicable to a near-threshold operation region or a large circuit where the delay is greater. Thus, a new analytical model for a subthreshold and near-threshold region is needed for accurate simulations and analytical treatment of circuit behavior.

5.1. Introduction

Simple, yet practical, models for MOSFETs are useful for analytical expressions of circuit behavior and simulations. The historical Shockley model has been used extensively; however, its application is not effective to recent short-channel submicrometer MOSFETs because that model does not consider velocity saturation effects of a carrier [24]. Thus, as a new approach to overcome these limitations, the Alpha-Power law MOSFET model developed by Takayasu Sakurai defines the voltage-current characteristics of MOSFETs and derives simple analytical expressions for the

drain current, short-circuit power, logic threshold voltage, and propagation delay.

CMOS inverter delay was first characterized by Burns [2], but more research on the circuit behavior is needed to consider the submicrometer region, parasitic delay, voltage dependence, and gate-source capacitance.

5.2. Comparison to Shockley Model

5.2.1. Shockley Model

The Shockley model expresses drain current as follows:

$$I_D = \begin{cases} 0 & , V_{GS} \le V_{th} \text{ (cutoff region)} \\ K\{(V_{GS} - V_{th})V_{DS} - 0.5V_{DS}^2\} & , V_{DS} < V_{DSAT} \text{ (linear region)} \\ \frac{K}{2}(V_{GS} - V_{th})^2 & , (V_{DS} \ge V_{DSAT} \text{ (saturation region)} \end{cases}$$

where K is a drivability factor, and V_{DSAT} is drain saturation voltage.

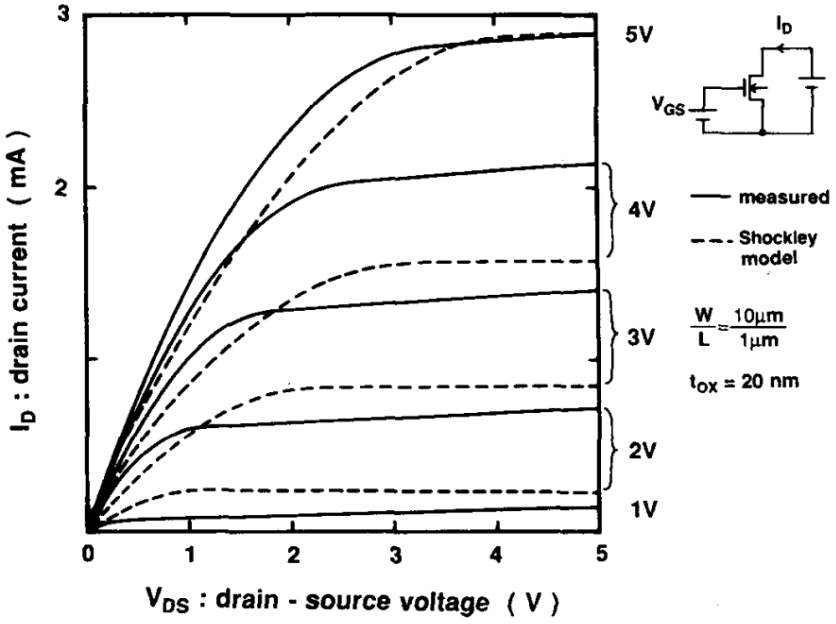

Figure 5.1. Measured V_{DS}-I_D characteristics and the Shockley model [24]

The Shockley Model does not accurately predict drain saturation voltage as it fails to consider the velocity saturation effects found in short-channel MOSFETs.

5.2.2. Alpha-Power Model

The Alpha-Power law model expresses drain current as follows:

$$I_D = \begin{cases} 0 & , V_{GS} \le V_{th} \text{ (cutoff region)} \\ \left(\frac{I'_{D0}}{V'_{D0}}\right) V_{DS} & , V_{DS} < V_{DSAT} \text{ (linear region)} \\ I'_{D0} & , (V_{DS} \ge V_{DSAT} \text{ (saturation region)} \end{cases}$$

where

$$I'_{D0} = I_{D0}\left(\frac{V_{GS} - V_{th}}{V_{DD} - V_{th}}\right)^{\alpha}$$

$$V'_{D0} = V_{D0}\left(\frac{V_{GS} - V_{th}}{V_{DD} - V_{th}}\right)^{\alpha/2}$$

From the above expressions, the α-power law depends on V_{th} (threshold voltage), α (velocity saturation index), V_{D0} (drain saturation voltage), and I_{D0} (drain current).

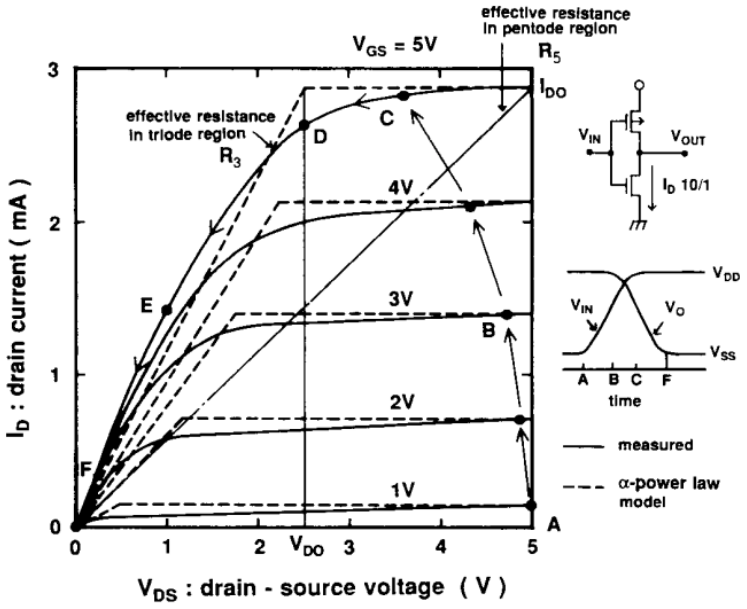

Figure 5.2. α–power law MOS model [24]

The Alpha-Power law does not characterize a subthreshold (below threshold) or near-threshold region as shown in Figure 5.2.

5.3. Summary of Alpha-Power Law

5.3.1. Drain Current

The Alpha-Power law model characterizes the drain current as the following equation. K is the device constant.

$$I_{DS} = K(V_{DD} - V_{th})^{\alpha} \tag{5.1}$$

5.3.2. Delay

From the Alpha-Power law MOS model, the delay is inversely dependent on $(V_{DD} - V_{th})^{\alpha}$ [25] To express equation 3.2 as equality, device constant K, different from the K in equation 3.1, can be used as a multiplier.

$$Delay \propto \frac{V_{DD}}{(V_{DD} - V_{th})^{\alpha}} \tag{5.2}$$

5.3.3. Frequency

The expression for frequency can be obtained from delay.

$$Frequency \propto \frac{(V_{DD} - V_{th})^{\alpha}}{V_{DD}} \tag{5.3}$$

5.4. Device Parameters

To verify the Alpha-Power law model, several parameters (K, V_{th}, and α) were extracted using HSPICE and MATLAB. First, I-V Characteristics for 45nm BSIM bulk CMOS library from an ASU PTM model [21] were obtained through HSPICE. Then,

necessary data was extracted: V_{GS} 0.5V to 1.0V for NMOS and V_{GS} 0V to 0.5V for PMOS (V_{DS} was set to 0.5V). The data for these values were most consistent.

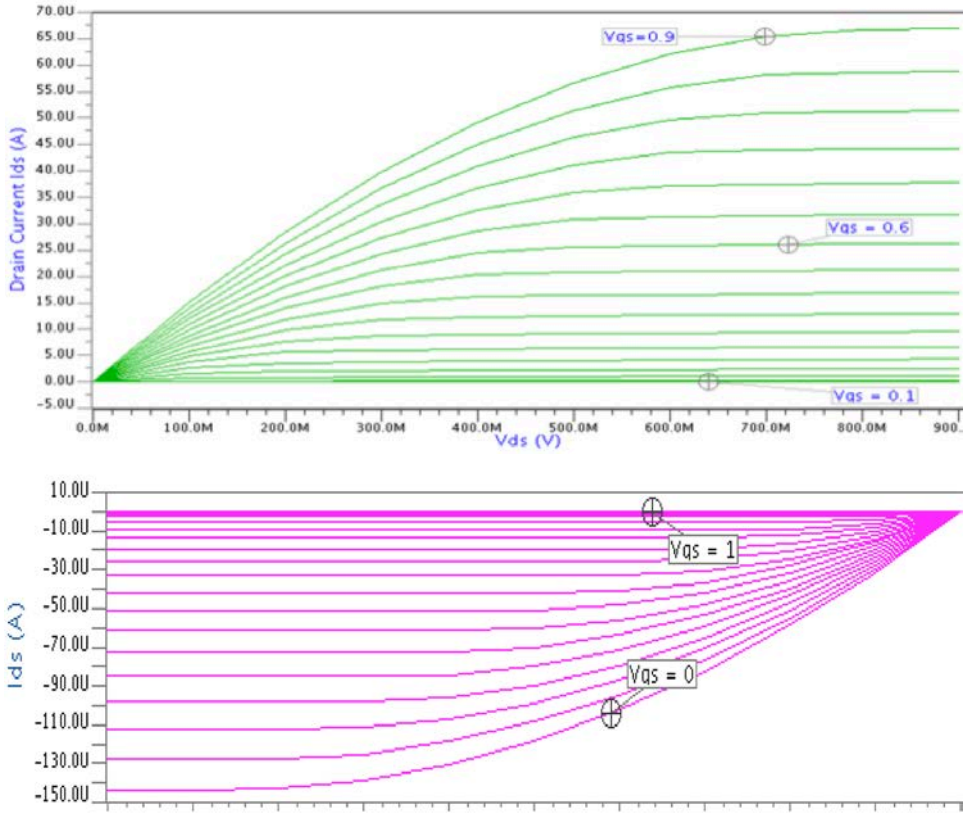

Figure 5.3. I-V Characteristics for NMOS (top) and PMOS (bottom)

Using MATLAB, the best fits for 11 samples were found for a NMOS and a PMOS (Figure 5.3). According to the Alpha-Power law, the drain current can be expressed using Equation 3.1.

For a NMOS device, a threshold voltage (V_{th}) of 0.25 and alpha of 1.3 best fit the samples. Additionally, when the threshold increased, the drain current was underestimated. When alpha decreased, the slope of the curve increased.

For a PMOS device, a threshold voltage (V_{th}) of 0.80 and an alpha of 1.5 best fit the samples. Also, when the threshold increased, the slope of the curve also increased. When alpha increased to 1.5, the curve moved closer to the best-fit scenario.

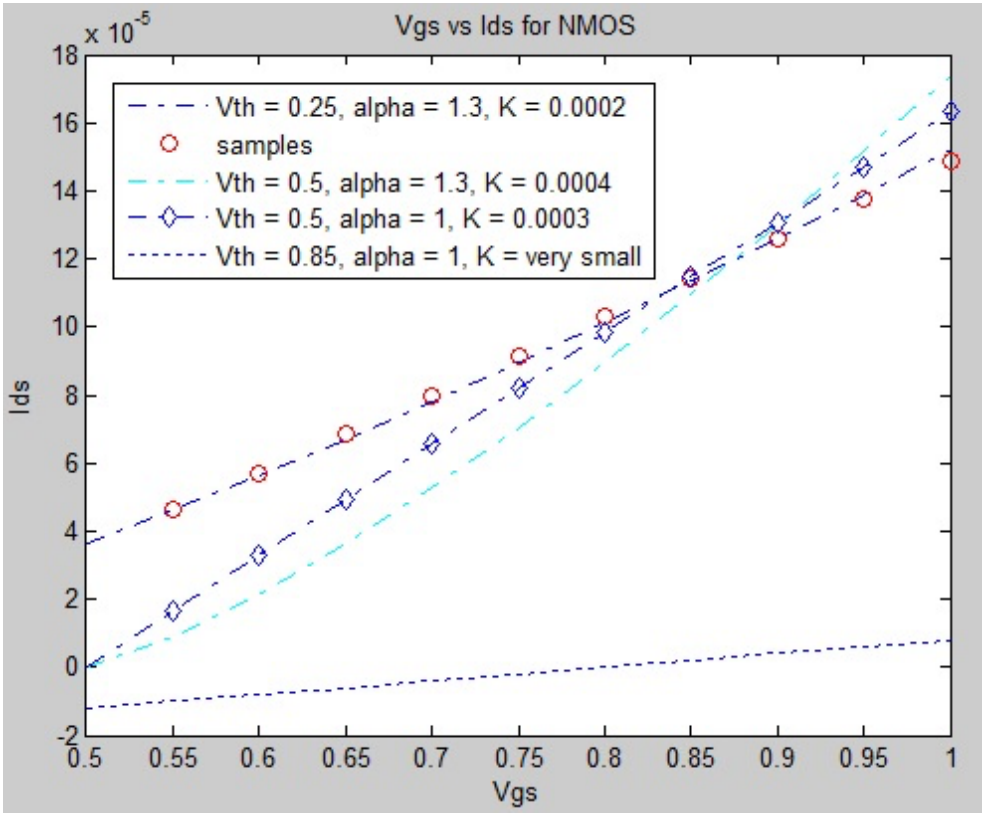

Figure 5.4. Curve fit for an NMOS

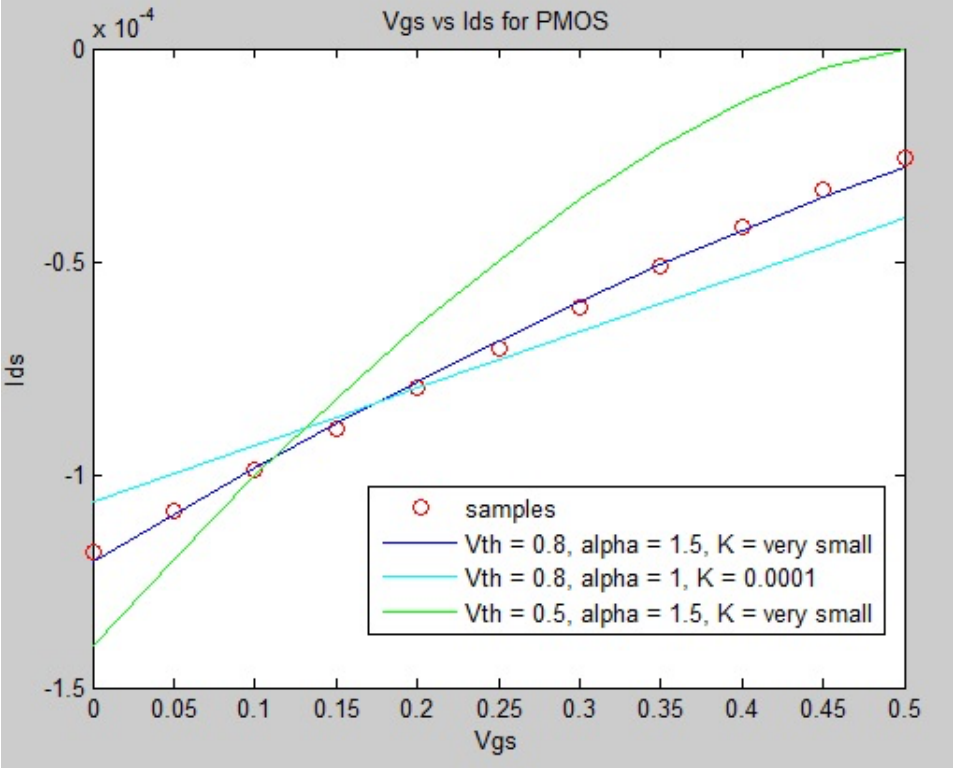

Figure 5.5. Curve fit for an PMOS

5.5. Simulation Results

Using a 45nm BSIM bulk CMOS library from the PTM model [21], a transient analysis was done using HSPICE. An input pulse wave was applied to check its timing and functionality. An inverter chain with a size of 5 was created by repeating the previously designed inverter with a capacitor of 10 pF using HSPICE. A 10 pF capacitor was used since average propagation delay (t_{PAVG}), low-to-high delay (t_{PHL}), and high-to-low delay (t_{PLH}) were measured as shown in Table 5.1. Figure 5.7 shows the average propagation delay.

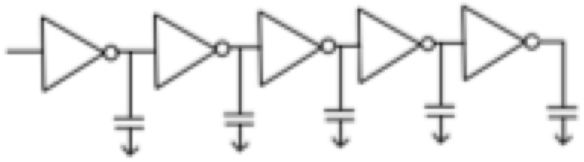

Figure 5.6. The size 5 inverter chain

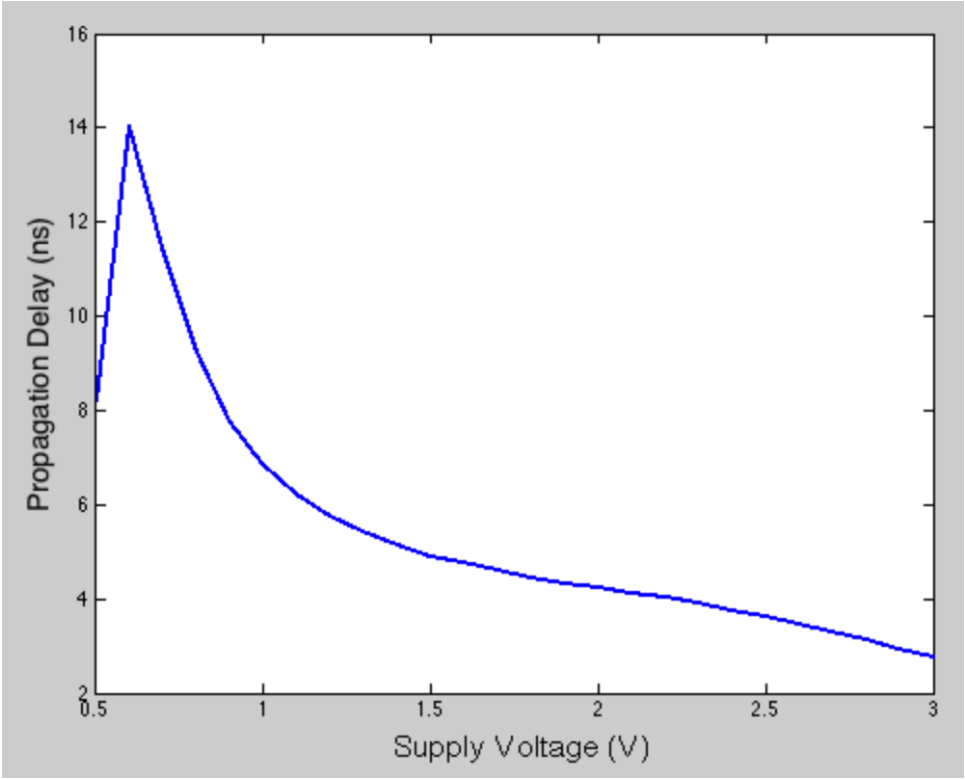

Figure 5.7. Delay for various V_{DD}

Table 5.1. Delay in the CMOS inverter chain for various V_{DD}

V_{DD} (V)	t_{PHL} (ns)	t_{PLH} (ns)	t_{PAVG} (ns)
3.0	0.95	4.60	2.77
2.9	1.02	4.86	2.94
2.8	1.09	5.13	3.11
2.7	1.16	5.40	3.28
2.6	1.25	5.66	3.45
2.5	1.32	5.90	3.61
2.4	1.39	6.14	3.76
2.3	1.46	6.34	3.90
2.2	1.52	6.51	4.02
2.1	1.58	6.66	4.12
2.0	1.63	6.82	4.23
1.9	1.67	6.99	4.33
1.8	1.72	7.17	4.45
1.7	1.76	7.40	4.58
1.6	1.80	7.71	4.76
1.5	1.84	7.97	4.90
1.4	1.88	8.36	5.12
1.3	1.95	8.85	5.40
1.2	2.03	9.47	5.75
1.1	2.14	10.29	6.21
1.0	2.29	11.42	6.86
0.9	2.53	13.02	7.77
0.8	2.87	15.55	9.21
0.7	3.44	19.43	11.43
0.6	4.77	23.26	14.01
0.5	8.83	7.56	8.20

When NMOS is discharging the output capacitors, the effect of PMOS can be ignored [7]. The predicted delay of t_{PHL} and the measured delay of the inverter chain are shown in Figure 5.8. A case wherein PMOS is charging the output capacitors is shown in Figure 5.9. For the delay when the circuit is operating at a near-threshold or subthreshold region, the Alpha-Power law does not accurately characterize the delay. The V_{th} for NMOS was 0.25, and the V_{th} for PMOS was 0.8. For PMOS where V_{DD} equals V_{th}, the delay was predicted to be extremely large because V_{DD}-V_{th} approaches zero. Thus, the estimation for the near-threshold region is not accurate for PMOS as shown in Figure 5.9.

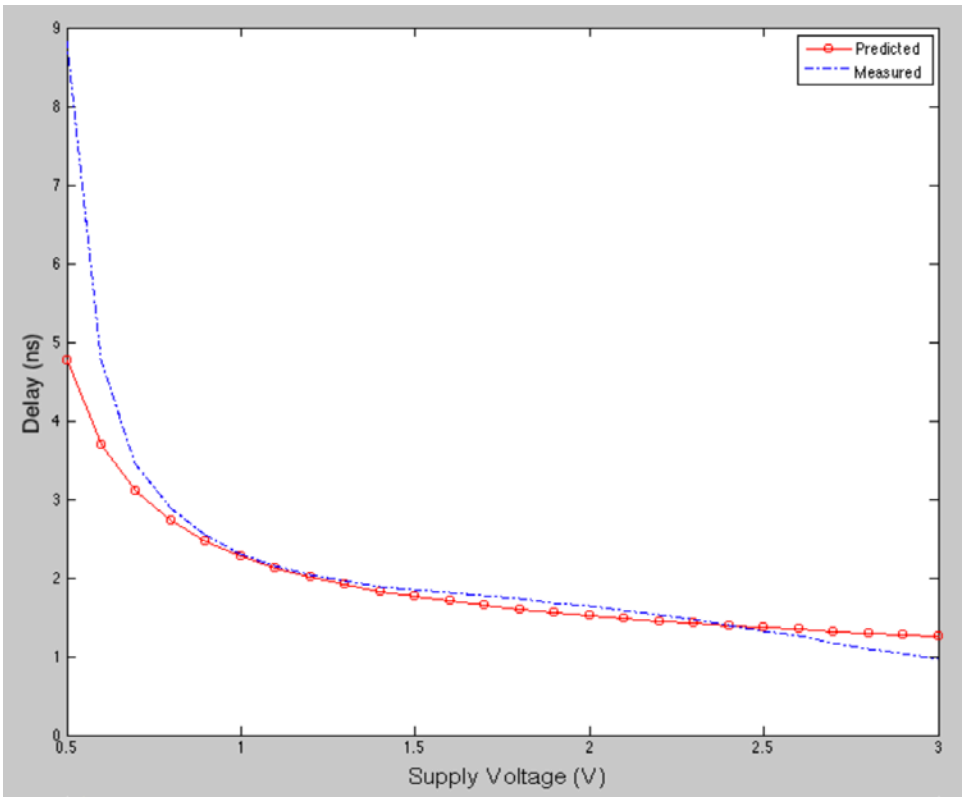

Figure 5.8. T_{PHL} at various V_{DD}

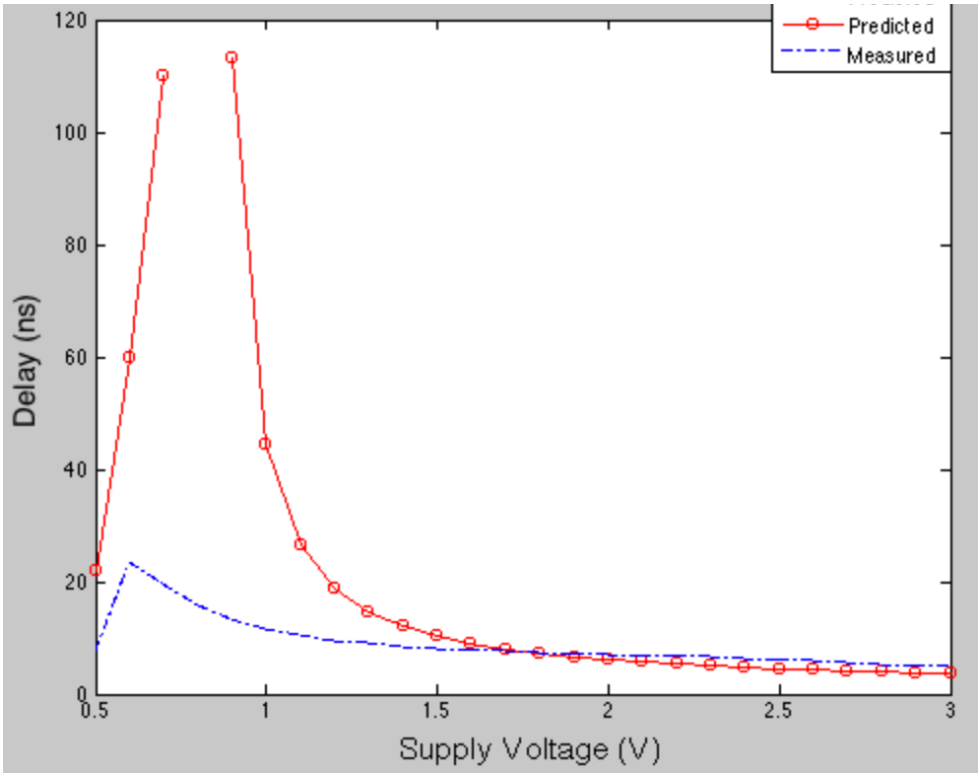

Figure 5.9. T_{PLH} at various V_{DD}

The absolute error between the predicted and the measured delay is shown in Figure 5.10. For the near-threshold or a V_{DD} approaching 3V, the percentage error was 46.12% and 33.10%, respectively. In between those extreme points, the percent error was less than 10%. Thus, the Alpha-Power law closely characterizes the propagation delay for the regions other than those near-threshold or for a large supply voltage.

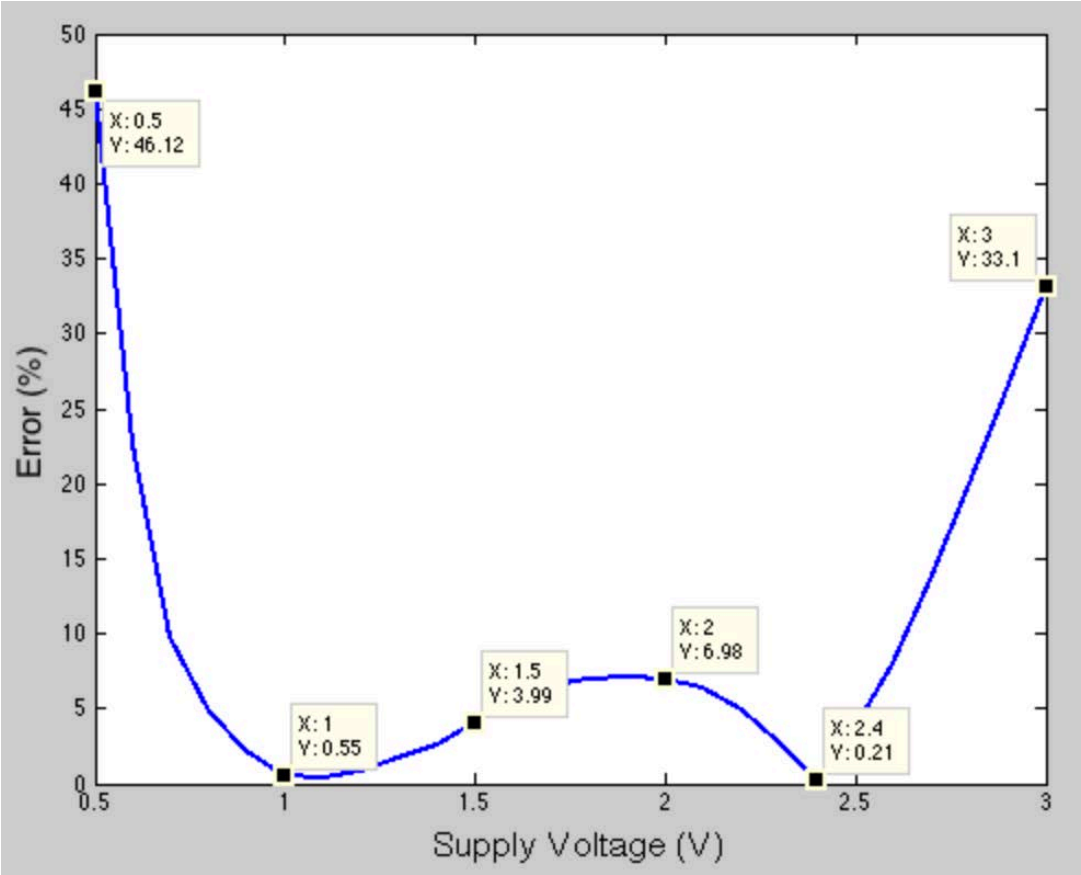

Figure 5.10. Percentage error between measured and estimated delays

To compare the results to a different size of an inverter chain, a total of five different sizes of inverter chain were simulated via HSPICE. A supply voltage of 2.4V, 2.5V, and 2.6V were chosen because the minimal percentage error was found at these operation regions.

Table 5.2. Absolute error for various inverter chain size and V_{DD}

Size	VDD = 2.4		VDD = 2.5		VDD = 2.6	
	Delay (ns)	Error (%)	Delay (ns)	Error (%)	Delay (ns)	Error (%)
5	1.39	0.21	1.32	3.69	1.25	7.92
10	1.38	3.53	1.31	0.2	1.23	4.01
100	1.24	3.99	1.17	0.07	1.10	4.41
1000	1.03	6.94	0.94	0.06	0.85	8.00
10,000	0.97	23.67	0.72	0.13	0.49	44.68

The error becomes larger as the size of the inverter chain increases as shown in Figure 5.11. It was also observed that the predicted delay is not accurate for the circuit when the delay is too small or too large. When an output capacitor of 100 fF was used, the delay was too small, such that the dependence on delay according to the Alpha-Power law could not be clearly observed.

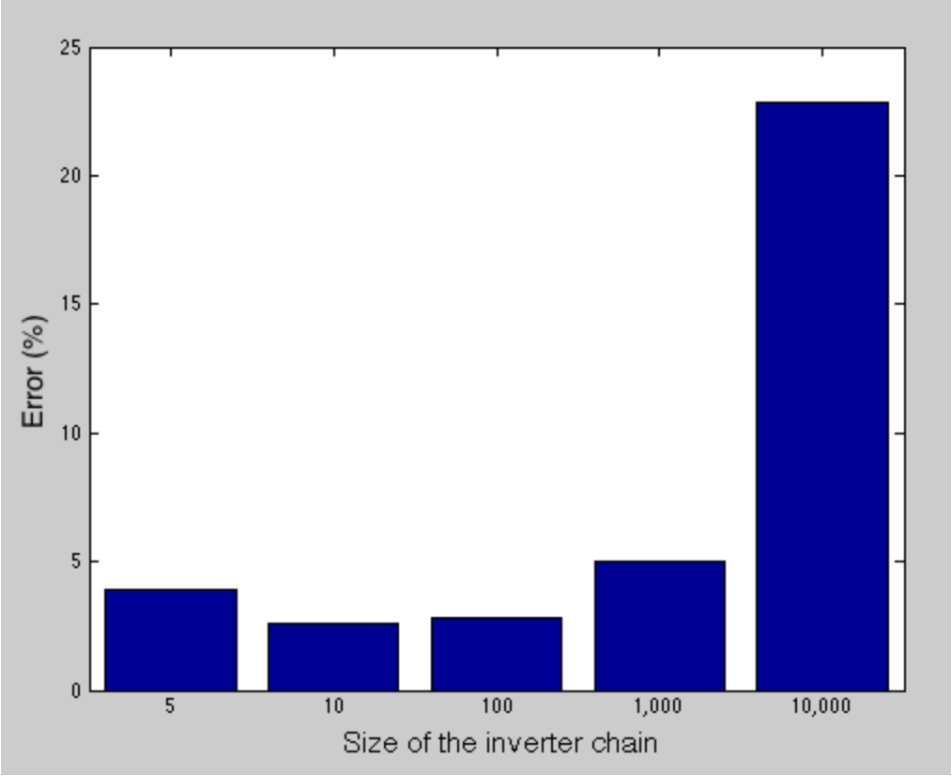

Figure 5.11. Percentage error for various sizes of inverter chain

5.6. Proposed Analytical Model

For the subthreshold region, a new relationship was proposed to estimate subthreshold circuit delay. This relationship can be used to observe the future trend of the subthreshold circuit, and it has been simulated and verified. The Alpha-Power law is based on the fact that the drain current of short-channel MOSFETs is proportional to $(V_{GS}\text{-}V_{TH})^{\alpha}$, where the α (carrier velocity saturation index) is around 1.3 for modern technology [25]. However, the delay in the subthreshold region increases at a much faster pace than what the Alpha-Power law predicts. The past research [6] presents the variation-aware analytical model that can be used for a subthreshold circuit region in addition to a superthreshold region. The inverter delay for a 0 to 1 transition can be estimated using the following equation:

$$\tau_{step} = ln2 * \frac{2mV_T}{I_o \lambda^2 V_{DD}} * \left[\frac{\frac{\lambda V_{DD}}{2} + mV_T}{e^{\frac{\lambda V_{DD}}{2mV_T}}} - \frac{\lambda V_{DD} + mV_T}{e^{\frac{\lambda V_{DD}}{mV_T}}} \right] * (C_L + C_{int}) \qquad (5.4)$$

where $I_o = \mu C_{OX} * \frac{W}{L}(m-1)V_T^2 e^{V_{GS}-V_{th}}/mV_T$. μ is the carrier mobility, C_{OX} is the oxide capacitance for the unit area, V_T is the thermal voltage, m is the subthrehsold slope factor, λ is the DIBL coefficient, C_L is the external load and C_{int} is its internal subthreshold capacitance.

Using the parameters of a 45nm BSIM bulk CMOS library from the PTM model [21], where V_T is 26mV at room temperature, m is 27, λ is 0.001, the dependence of delay on a different supply voltage can be written as Equation 5.5.

$$Delay \propto \frac{\tau_{step,V_{DD1}}}{\tau_{step2,V_{DD2}}} = \frac{V_{DD2}}{V_{DD1}} * \frac{\left[\frac{\frac{\lambda V_{DD1}}{2} + mV_T}{e^{\frac{\lambda V_{DD1}}{2mV_T}}} - \frac{\lambda V_{DD1} + mV_T}{e^{\frac{\lambda V_{DD1}}{mV_T}}}\right]}{\left[\frac{\frac{\lambda V_{DD2}}{2} + mV_T}{e^{\frac{\lambda V_{DD2}}{2mV_T}}} - \frac{\lambda V_{DD2} + mV_T}{e^{\frac{\lambda V_{DD}2}{mV_T}}}\right]} \quad (5.5)$$

Because $e^{\frac{\lambda V_{DD}}{2mV_T}} \approx e^{\frac{\lambda V_{DD}}{mV_T}}$, the above equation can be further reduced to the following:

$$Delay \propto \frac{V_{DD2}}{V_{DD1}} * \frac{\left[\frac{-\frac{\lambda V_{DD1}}{2}}{e^{\frac{\lambda V_{DD1}}{2mV_T}}}\right]}{\left[\frac{-\frac{\lambda V_{DD2}}{2}}{e^{\frac{\lambda V_{DD2}}{2mV_T}}}\right]}$$

$$= \frac{e^{\frac{\lambda V_{DD2}}{2mV_T}}}{e^{\frac{\lambda V_{DD1}}{2mV_T}}} = e^{\frac{\lambda V_{DD2}}{2mV_T} - \frac{\lambda V_{DD1}}{2mV_T}} = e^{\frac{\lambda (V_{DD2} - V_{DD1})}{2mV_T}} \quad (5.6)$$

It is observed that delay exponentially depends on $\frac{\lambda(V_{DD2} - V_{DD1})}{2mV_T}$. HSPICE simulation was run to verify the proposed analytical model. As shown in Figures 5.12 and 5.13, the proposed model is more accurate in the subthreshold region while the Alpha-Power law estimates circuit delay more accurately in that superthreshold region. The average percentage error of a high to low switch in the subthreshold region is 15%, and the maximum error of 32% is found at the subthreshold region. The average percentage error for low to high switch in the subthreshold region is 15%, while the maximum error is 19% in the subthrehold region and 77% in the superthreshold region

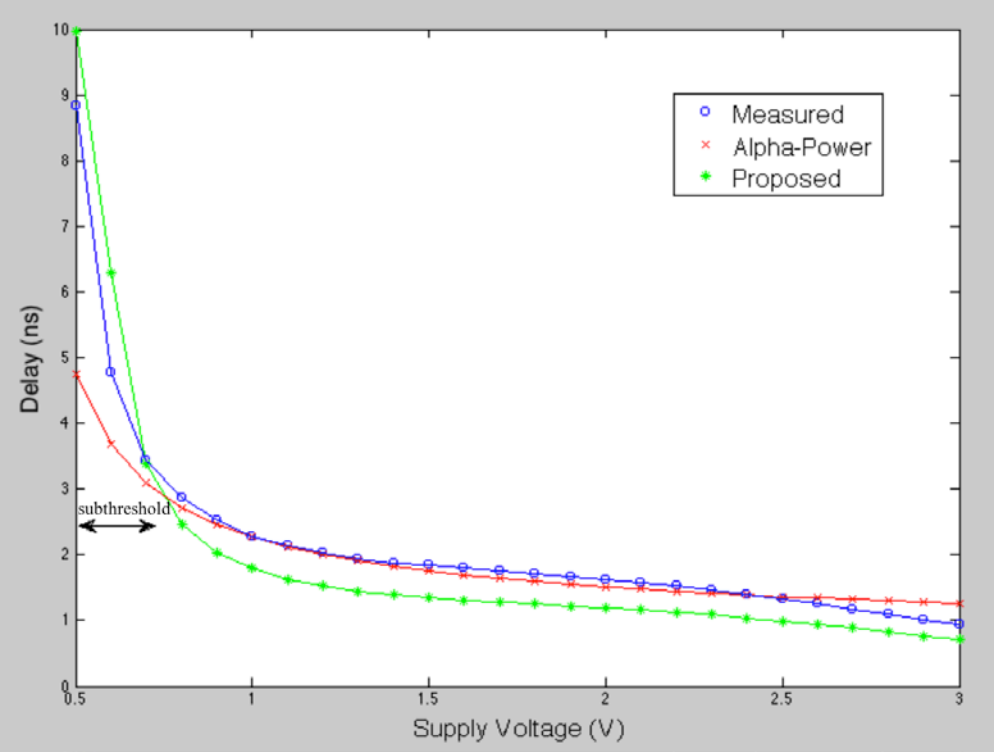

Figure 5.12. T_{PHL} at various V_{DD}

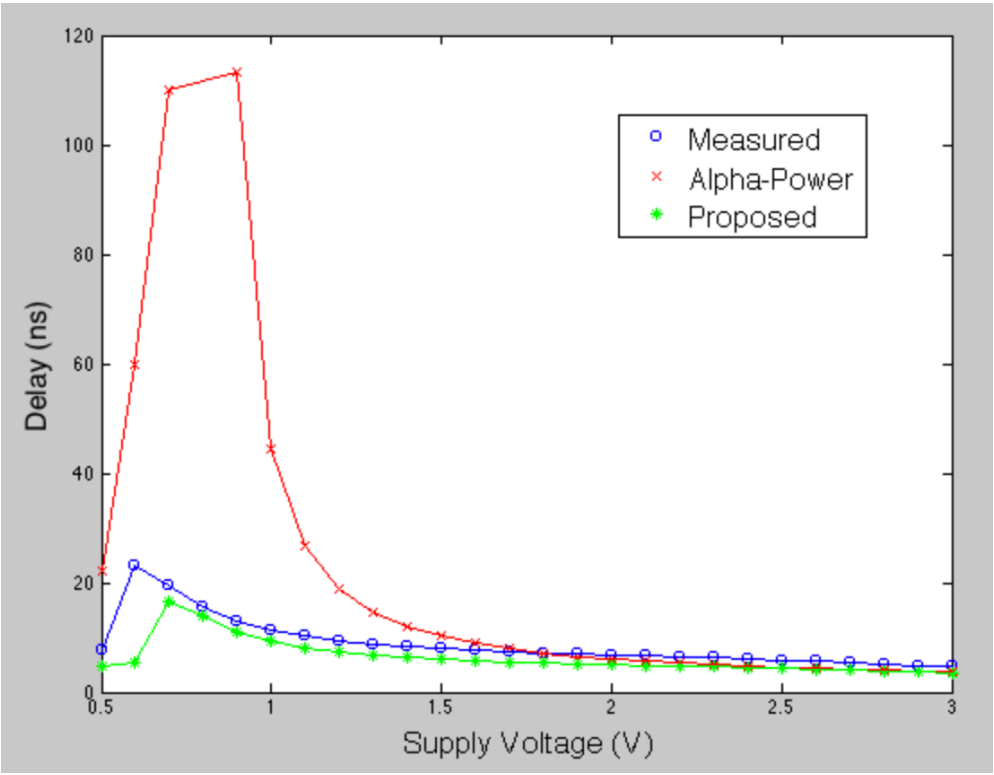

Figure 5.13. T_{PLH} at various V_{DD}

5.7. Effect of Variations

PVT variations affect the delay of the subthreshold circuit. Using the models developed by Lin, et al [16], PVT variations were analyzed and simulated using the 32nm PTM model [21]. A study [16] proposed and verified the following analytical models that account for PVT variations for the BSIM4 level-54 model.

5.7.1. Process Variation

In the subthreshold region, threshold voltage variation is the dominant factor and has an exponential relationship to circuit delay variation [16]. Therefore, the following equation shows the impact of threshold voltage variation on the delay.

$$\Delta T_{delay} = \frac{T_{delay2}}{T_{delay1}} * 100\% = exp\left(\frac{V_{t2} - V_{t1}}{nV_{th}}\right) * 100\% \qquad \ldots \qquad (5.5)$$

Because previous research focused on 130nm technology, further simulation was carried out for the 32nm LP PTM model [21], whose threshold voltage is 0.63. A supply voltage of 0.5V was used. The HSPICE simulation result is shown in Figure 5.14.

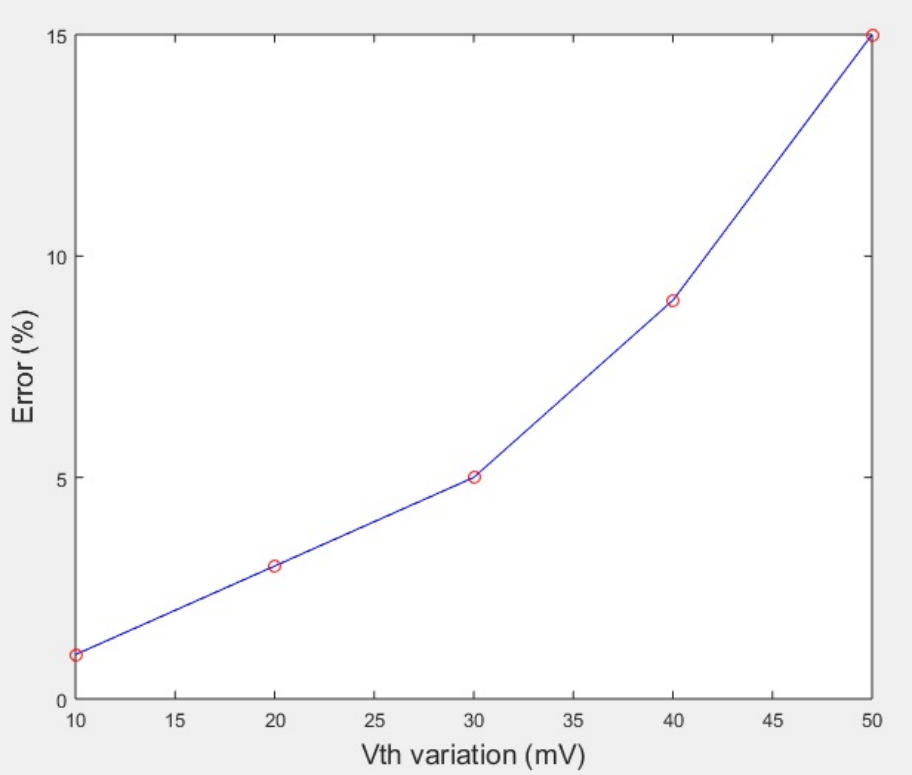

Figure 5.14. Percentage error for V_{TH} variation

5.7.2. Supply Voltage Variation

Voltage variation affects the supply voltage of the circuit, and delay as well as power consumption can vary depending on the supply voltage. The following equation shows the exponential relationship of supply voltage variation on the delay.

$$\Delta T_{delay} = \frac{T_{delay2}}{T_{delay1}} * 100\% = \frac{V_{DD1}}{V_{DD2}} exp\left(\frac{V_{DD1} - V_{DD2}}{nV_{th}}\right) * 100\% \quad \ldots \quad (5.6)$$

A simulation result shows a maximum 26% error of the analytical model on a 50mV variation of supply voltage.

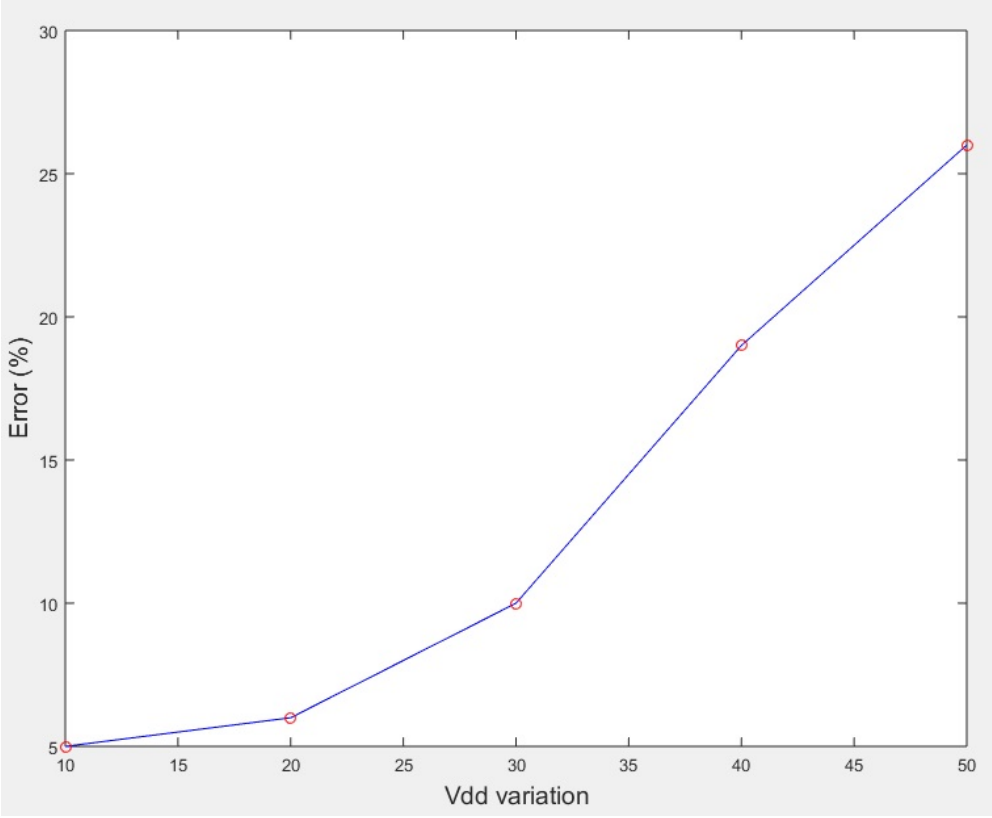

Figure 5.15. Percentage error for V_{DD} variation

5.7.3. Temperature Variation

Temperature variation affects delay exponentially. Parameter UTE (average value of -1.85) is mobility temperature constant, and K_{T1} (average value of -0.25) is temperature coefficient for threshold voltage.

$$\Delta T_{delay} = \frac{T_{delay2}}{T_{delay1}} * 100\%$$
$$= \left(\frac{T_1}{T_2}\right)^{2+UTE} exp\left((V_t - K_{T1} - V_{DD}) * \frac{(T_1 - T_2)q}{nkT_1T_2}\right) * 100\% \quad \ldots \quad (5.7)$$

Temperature was varied from 5°C to 25°C. The simulation result is shown in Figure 5.16, and the maximum error was 3%.

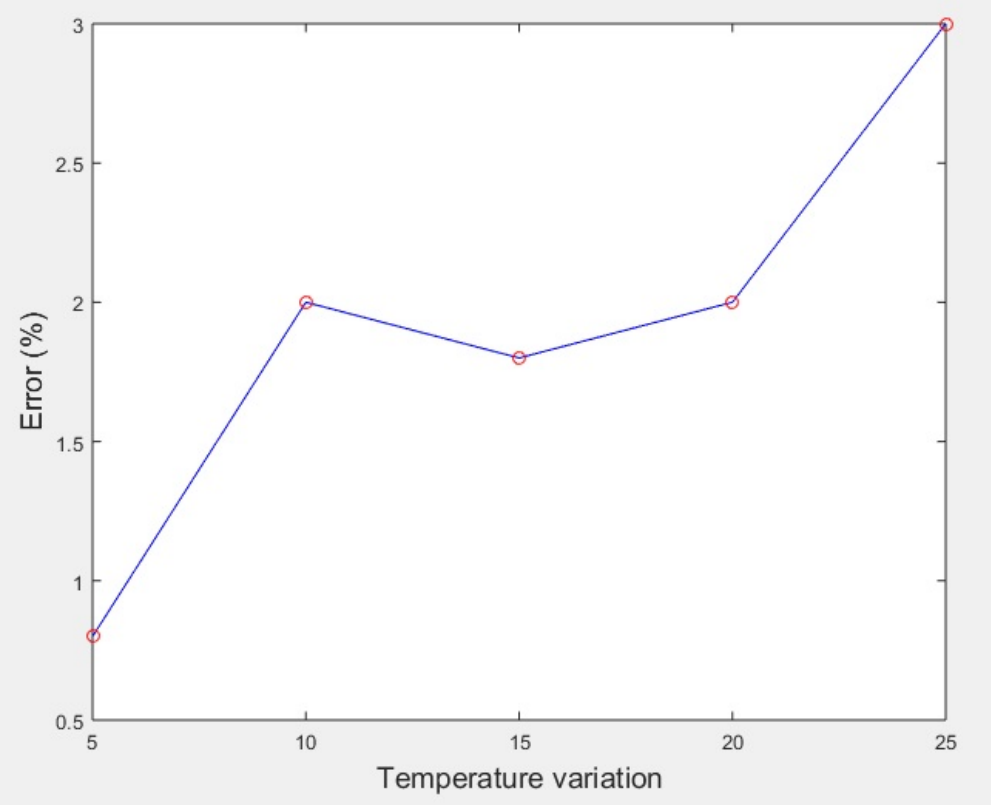

Figure 5.16. Percentage error for temperature variation

5.8. Summary

The Alpha-Power law provides a simple and yet practical analytical model for MOSFET. As the size of the inverter chain increases, the error of the predicted delay als o increases. Average error is less than 5% for an above-threshold operation region unless the delay is significantly large or significantly small. When the size of the inverter exceeds 1,000, the error becomes larger and approaches 25%. When the circuit is operating at a near-threshold or subthreshold region, the error can be as large as 50%.

As The Alpha-Power law does not accurately characterize the circuit behavior of a subthreshold or near-threshold operation region, an alternative variation-aware analytical model is proposed. The proposed model has an average of 15% error in the subthreshold

region while The Alpha-Power law has a 21% error in high to low switch and a 404% error in low to high switch. The maximum error improved in the proposed model.

Furthermore, PVT variations were analyzed using analytical models and HSPICE simulation. It was found that error is greater in smaller technology nodes. Along with the analytical model and the effect of PVT variations, the subthreshold delay trend can be better analyzed, and its design can be improved so as to consider its variations.

Chapter 6

Optimization of Subthreshold Circuit using Linear Programming

Using Linear Optimization (LP) to optimize the subthreshold circuit was proven to be effective, especially when various design approaches are possible. In the subthreshold circuit, there are multiple parameters: Supply voltage, projected leakage, estimated delay, performance degradation, and sensitivity. In this study, since the focus is on improving the performance while taking advantage of low power consumption, LP was utilized to demonstrate how much power can be increased when a certain amount of performance degradation is recovered. In this section, a brief introduction of linear programming along with a discussion of the methodology to improve the subthreshold circuit performance.

6.1. Linear Programming

Linear Programming (or linear optimization) is set by the objective function defined as either a maximization or minimization problem and a set of constraints defined by linear equalities and inequalities. After defining the objective function, the linear programming solver utilizes several versions of a simplex algorithm to solve the optimization problem and produce one of three possible cases: (1) optimal, (2) unbounded, and (3) infeasible. Optimal solution(s) can be obtained from the optimal case, while such is impossible in an unbounded or infeasible case due to either a lack of

necessary constraint or an unfeasible objective function. How the simplex algorithm works is briefly described in the following section.

6.2. Motivation

Whether circuit designers aim for ultra-low power or energy-efficient design, the optimization of the circuit is essential. Often, optimization for area, power, or delay is the main interest. One example would be minimization of power, given a certain delay while the circuits do not violate a critical delay. Through optimization, circuits can achieve robustness, ultra-low power, energy-efficiency, or a smaller area.

6.3. Optimization Methodology

The optimization considerations are mainly computational complexity and accuracy. To achieve near complete accuracy, computational complexity must increase, and it takes more time and effort. Thus, a heuristic approach is often used for large circuits. A comparison of the complexity and accuracy between a greedy heuristic and linear programming can indicate which method is best to use. Also, values for parameters do not have to be exact for initial optimization, as they only serve to give one a general idea of the trend. These values can be fixed later for better accuracy. After the circuits are optimized, SPICE simulation can verify the functionality of the circuit. When the simulation does not match the expected result, one or more constraints may be missing, or the input parameters may not be accurate enough.

6.4. Summary

As circuits contain various characteristics that designers want to optimize, their interest often conflicts. Thus, an optimization model with the necessary constraints was developed to find the optimal solution. When computational complexity is the main interest, a greedy heuristic can be used instead after analyzing the trade-off in terms of optimization accuracy. Optimized circuits need to function properly; therefore, software simulations are important to make sure of their precise functionality.

Chapter 7

The Dual-Threshold Voltage CMOS Design

To reduce the leakage of a subthreshold circuit, the dual-threshold voltage design technique is discussed in this chapter. While the low threshold gate is fast and produces more leakage power, the high threshold gate is slow and produces less leakage power. By assigning two types of gates optimally through linear programming, a reduction of leakage power is achieved. Moreover, further architectural modification and due considerations of different technology size are discussed to achieve ultra-low power consumption. This algorithm takes into account the variation of threshold and supply voltage, while the previous research assumed that the threshold and supply voltage would remain relatively stable, which is often not the case in the subthreshold circuit.

7.1. History

While a dual threshold circuit design has been prevalent for a strong inversion region, its exploration has been largely limited. Yao [29, 30] presents a gate-slack- based, dual-threshold subthreshold circuit design and reduction of energy consumption per cycle. A experimental study of a 32-bit ripple carry adder verifies the reduction of energy consumption per cycle at 29% compared to the single threshold design [29, 30].

7.2. Motivation

As dynamic power depends on the capacitance and the supply voltage, reducing the leakage (static power) has been a main focus in the subthreshold circuit. Especially for modern submicrometer technology, the leakage is consistently increasing while the dynamic power is manageable in the subthreshold circuit by lowering the supply voltage significantly below the threshold voltage.

7.3. Algorithm

First, the given circuit is analyzed using HSPICE simulation to extract the total power consumption and critical path delay. Then, a library is characterized for delay and power consumption. For example, the PTM model [21] contains both low power (LP) and high performance (HP) models with a different threshold. High performance gates consume more power while producing less delay, and low power gates consume less power while having more delay. Gates in the critical path with an assigned low threshold are assigned to maximize the speed and ensure functionality of the circuit, and gates in a non-critical path are analyzed. If a certain delay condition is met, those gates can be changed. Thus, the following integer programming model was utilized.

A simplified version of the dual-threshold circuit design developed by Yao [29, 30] was employed to build the integer programming to reduce the computational complexity while still maintaining reasonable accuracy and energy saving.

T_i is defined as the signal arrival time at gate i output, and T_j is defined as signal arrival time at j^{th} input of gate i. $f_o(i)$ is defined as the number of fanouts from gate i. X_i is equal to zero when a low threshold (high performance) gate is assigned and equal to one

when a high threshold (low power) gate is assigned. P_{LT} and P_{HT} correspond to the power consumption for a low threshold and high threshold gate, respectively, and D_{LT} and D_{HT} correspond to the delay for a low threshold and high threshold gate, respectively.

$$Minimize\ E = \sum_{all\ gates} P_{LT}(1 - X_i) * f_o(i) + P_{HT}X_i * f_o(i)$$

$$Subject\ to\ T_i \geq T_j + D_{LT}(1 - X_i) * f_o(i) + D_{HT} * X_i * f_o(i)\ for\ all\ j$$

After the gate assignment according to the output of the integer programming, the circuit is simulated, and energy consumption is measured. Because all low threshold gates are used for the critical path delay, the performance remains the same, and the optimization mainly reduces the leakage power by replacing the low threshold gates with high threshold gates but without violating the critical delay.

7.4. Simulation Results

Using the 32nm PTM model [21], the simulation for the 32-bit adder was obtained. There were two models used: Low power (LP) and high performance (HP). While the low power consumes less power, its delay is greater than the high performance gates. Figure 9.3 shows the major four cases: (1) assignment of all low V_{th}, (2) assignment of all high V_{th}, (3) random assignment, and (4) optimal assignment. The optimal assignment requires no change in the critical path delay, as the increase in that delay may trigger the malfunction of the circuit while the power is minimized. In this simulation, the variation of threshold and supply voltage was not taken into account. The integer programming was solved using MATLAB.

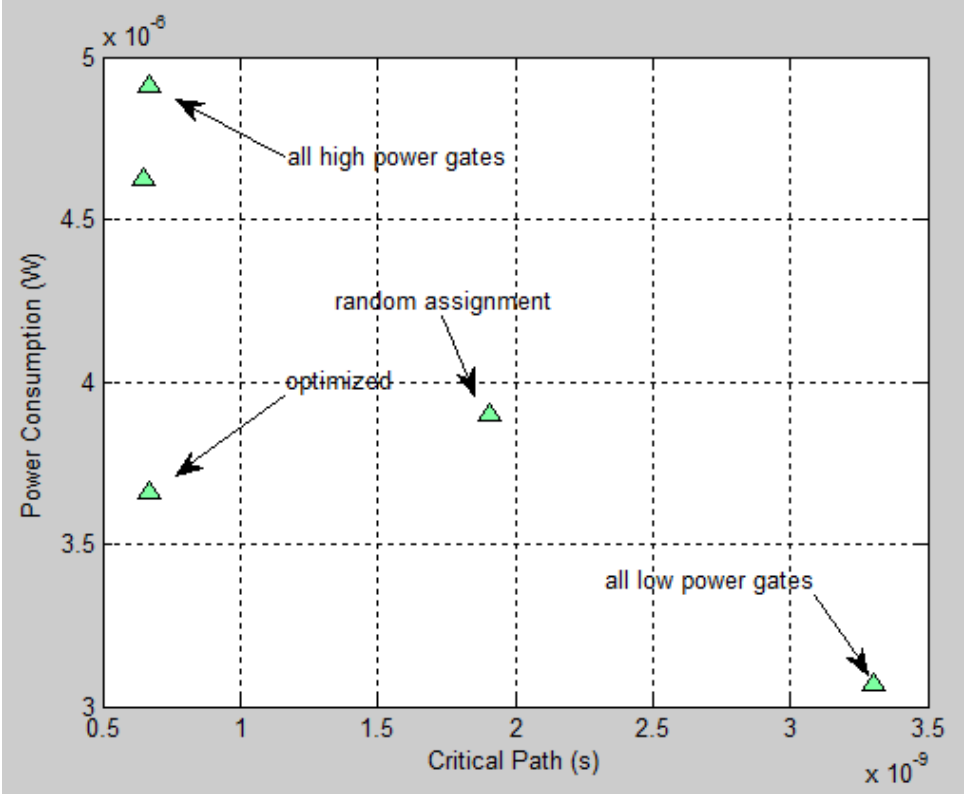

Figure 7.1. Optimal assignment for 32nm 32-bit adder

The optimized circuit demonstrated a 25% reduction in power consumption with no change in critical path delay. While a more thorough algorithm [29, 30] could achieve 29% reduction in energy consumption per cycle, the algorithm used here is more simplified and thus takes less computational time.

7.5. Effect of Technology Scaling

Figure 7.2 demonstrates the consistently increasing leakage as the technology node decreases. For instance, from 90nm to 16nm, the leakage increases 40 times. As shown in Figure 7.3, the static component of the total power is, on average, 25% greater than for the dynamic power. Also, considering that the supply voltage was 2.4V for this result, the

subthreshold circuit has even greater portions of static power. Thus, the trend for sizing down technology shows there is a need to reduce static power while still managing dynamic power using a reduced supply of voltage.

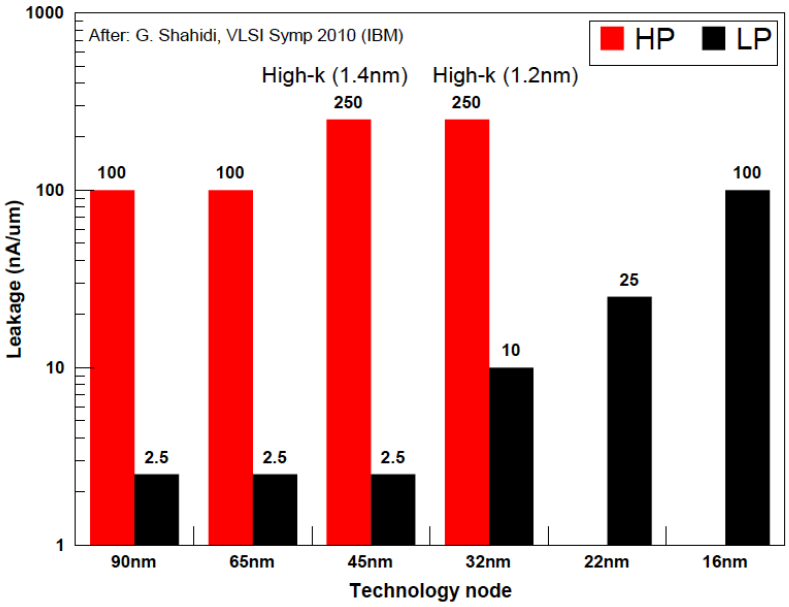

Figure 7.2. Leakage for various technologies [26]

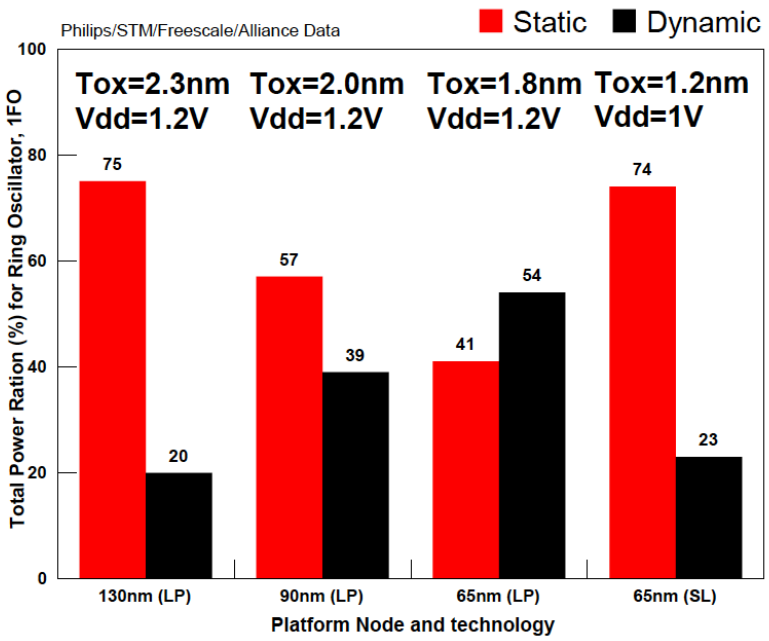

Figure 7.3. Total power for various technologies [26]

7.6. Summary

As the technology node decreases, the need for a circuit design that reduces the leakage power reduction increases. Because focusing entirely on power consumption leads to severe performance degradation and thus limits the application of the circuit, a dual-threshold circuit design that maintains the same critical path delay and performance while reducing power consumption about 25% is beneficial. The simplified integer programming model was demonstrated, simulated, and verified in this Chapter.

Chapter 8

Optimization of Subthreshold Global Interconnects

In this Chapter, global interconnect optimization is surveyed, and several techniques used to optimize global interconnects are considered. As wire capacitance does not scale with the supply voltage, an on-chip global interconnect suffers significant performance degradation. For example, a conventionally used repeater insertion technique for the superthreshold circuit applied to the subthreshold global interconnects proved to be ineffective. Thus, a tapered driver is used to optimize the circuit delay and energy (power-delay product). In Chapter 8.1, past research on global interconnect optimization for both superthreshold and subthreshold is surveyed. Motivation for the proposed technique is discussed in Chapter 8.2. Chapter 8.3 and Chapter 8.4 discuss method and results, respectively, and Chapter 8.4 is devoted to a summary of results.

8.1. History

As demand for low power consumption is increasing, more challenges regarding circuit performance have been presented. As the technology node is decreasing, the portion of global wire delay heavily overweighs logic delay as shown in Figure 8.1. In 32nm, global wire delay is responsible for 99% of total delay without repeater and 91% with repeater [8, 11]. Especially, performance degradation by global interconnects significantly reduces overall circuit performance and more significantly for subthreshold

circuit. Furthermore, the effect of PVT should be considered for optimizing any subthreshold global interconnect.

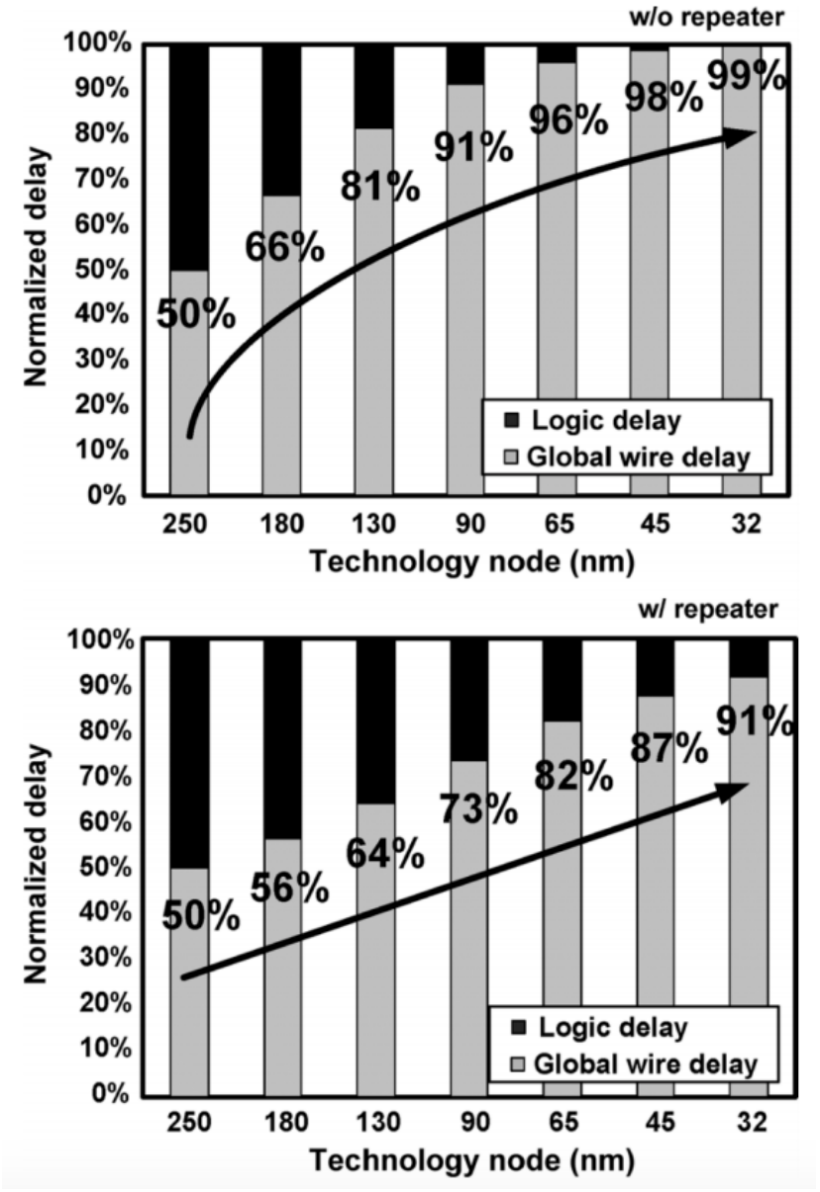

Figure 8.1. Scaling trend of logic delay and interconnect delay [8, 11]

For electronics that require a balance between energy and performance, optimization of the CMOS interconnect is crucial to obtain the required throughput for the system. Introduced by Deodhar, the throughput per bit-energy proves to offer the optimal

interconnect design between throughput and energy per bit [3]. The optimal number of repeater insertions in the global interconnect achieves maximum throughput per bit-energy. The case study of various supply voltage from 0.8V to 2.5V reveals that 1V is the optimal supply voltage without a loss in throughput, and also that the circuit cannot achieve the required throughput at 0.8V.

Furthermore, a study by Nalarnalpu and Burleson [18] presents a complex model for the delay and optimal repeater spacing and sizing. It utilizes the Alpha-Power law to model the propagation delay. The estimated performance achieves a maximum error of only 5% in 0.13 pm CMOS technology [18]

However, the study by Deodhar and Davis [3] does not include any discussion on subthreshold global interconnect optimization, and the Alpha-Power law used by Nalarnalpu and Burleson [18] has great discrepancy in circuits operating in the subthreshold region as discussed in Chapter 5. As shown in Figure 8.2, as the supply voltage decreases from 1V to 0.2V, while logic delay decreases, global interconnect delay remains the same. This suggests that a global interconnect technique for the subthreshold region should be different from one for a superthreshold region.

As subthreshold gains more attention, more studies have been conducted on the interconnect technique for subthreshold circuits, including those on circuit level [11] and device level optimization [22]. A study by Kil and Kim [11] presented the capacitive boosting technique, internally boosting the gate voltage of the driver to shows a 2.6x faster switching speed and 2.4x less delay sensitivity under temperature variations. A study by Rahaman and Chowdhury [22] utilized the negative capacitance effect in

composite ferroelectric semiconductor MOSFET to present a theoretical model in the device level.

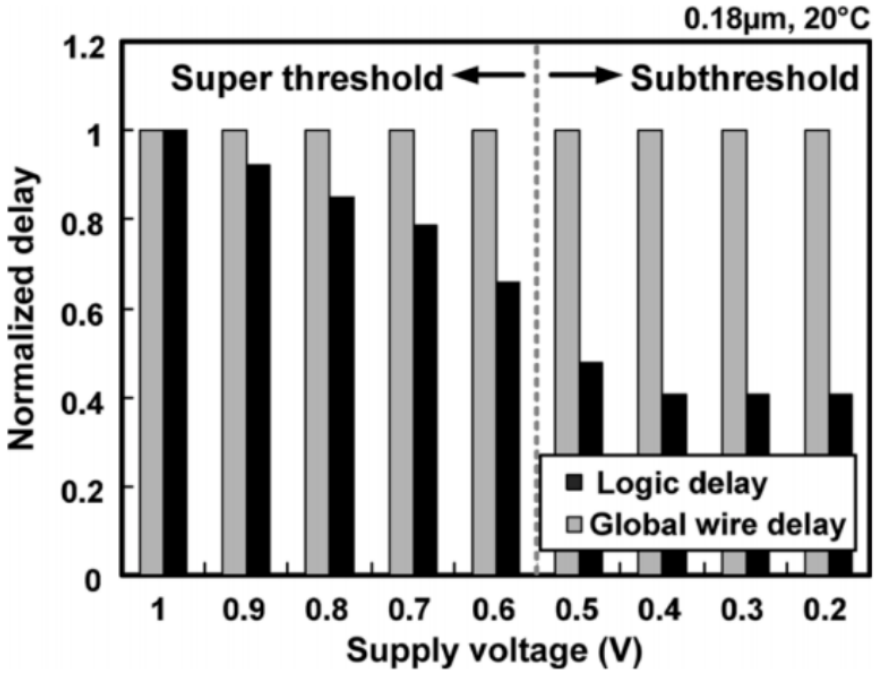

Figure 8.2. Global interconnect delay in a subthreshold region [11]

8.2. The Repeater Insertion Technique

8.2.1. Motivation

As authors in [3] explore the benefit of repeater insertion to find the minimum dynamic power loss, the study of its application to a subthreshold operation region has not been explored thoroughly. It is suggested that a subthreshold circuit can utilize the same repeater insertion technique to achieve maximum throughput per bit-energy and compensate for performance drawback. Without the change in wire area, repeater insertion to the global interconnect may be able to increase the throughput of a subthrehsold circuit and compensate for the performance drawback caused by the supply

voltage being below threshold. Thus, the efficiency of repeater insertion in the subthreshold region is discussed here.

8.2.2. Method and Simulation Results

Considerations for optimizing a global interconnect includes threshold voltage, the number of repeater insertions, and the technology being used. As static energy consumption is becoming more and more important for modern technology as seen in Figure 8.3, the focus of the power reducing mechanism for the global interconnect lies in reducing leakage current and static power consumption. Therefore, a high threshold model is used for simulation, and the static power consumption is measured during the PTM model [21] on a HSPICE simulation using a subthreshold operation region. For simplicity, a simple CMOS inverter was chosen for the repeater. Also, to characterize wire and its parasitic effects accurately, a RLC network was used. The additional cost of each repeater insertion was a delay and power consumption of an inverter. Various numbers of repeaters were used, and the simulation result is shown in Figure 8.4.

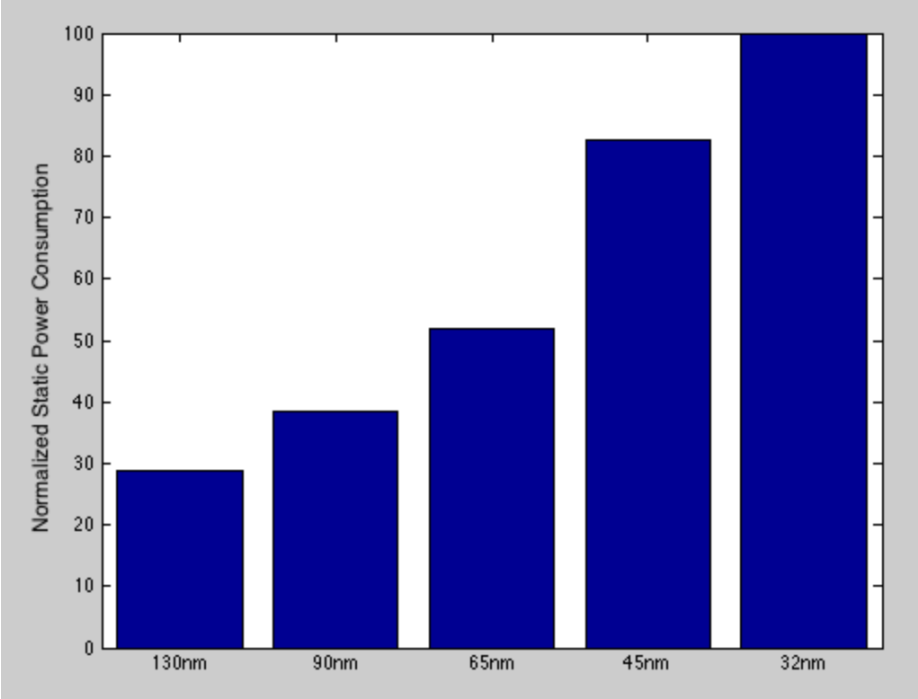

Figure 8.3. Static power consumption for various size of technology nodes

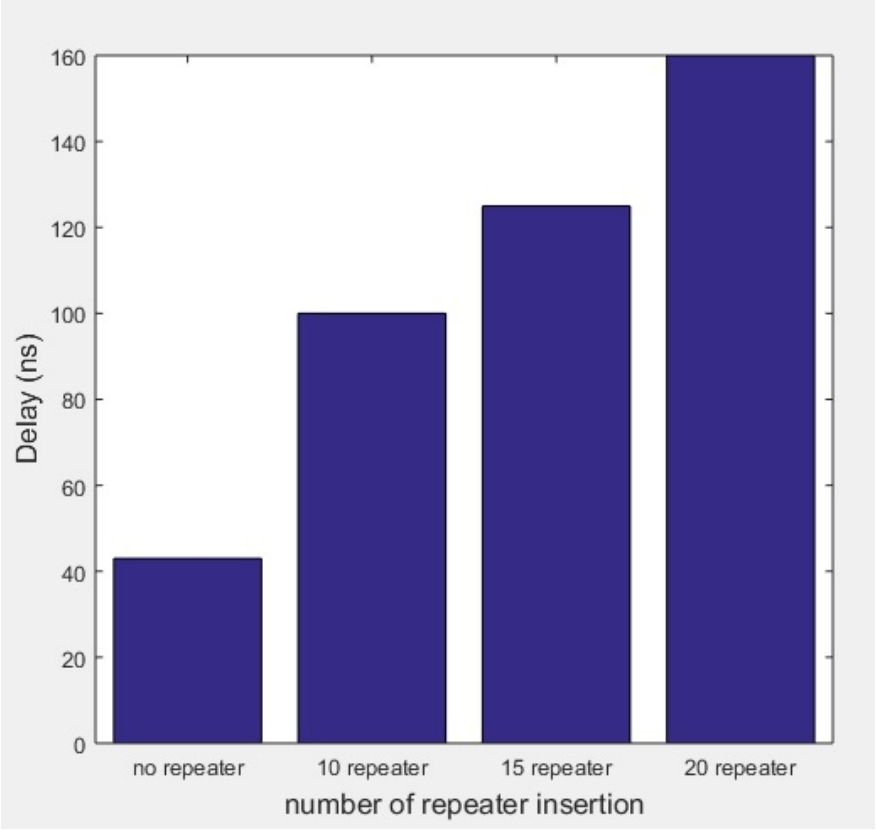

Figure 8.4. Effect of repeater insertions for a subthreshold global interconnect

As the repeater is inserted into a subthreshold global interconnect, delay increases more than 2x. When 20 repeaters are inserted, a 4x increase in delay is observed, and this drastic increase in delay likely makes the circuit not able to meet the performance requirement. When a repeater is inserted in a superthreshold global interconnect, the throughput is increased as illustrated in Figure 8.5. The same PTM model is used in the simulation, and a CMOS inverter is used for the repeater.

As driver size increases, driver delay dominates the interconnect delay [20] as shown in Figure 8.6. In subthreshold operation, the driver delay dominates the interconnect delay; therefore, a larger driver is needed to reduce the total delay, while area penalty exists [20]. Thus, optimizing a driver, such as a tapered driver, is more useful for a subthreshold global interconnect than is the repeater insertion technique.

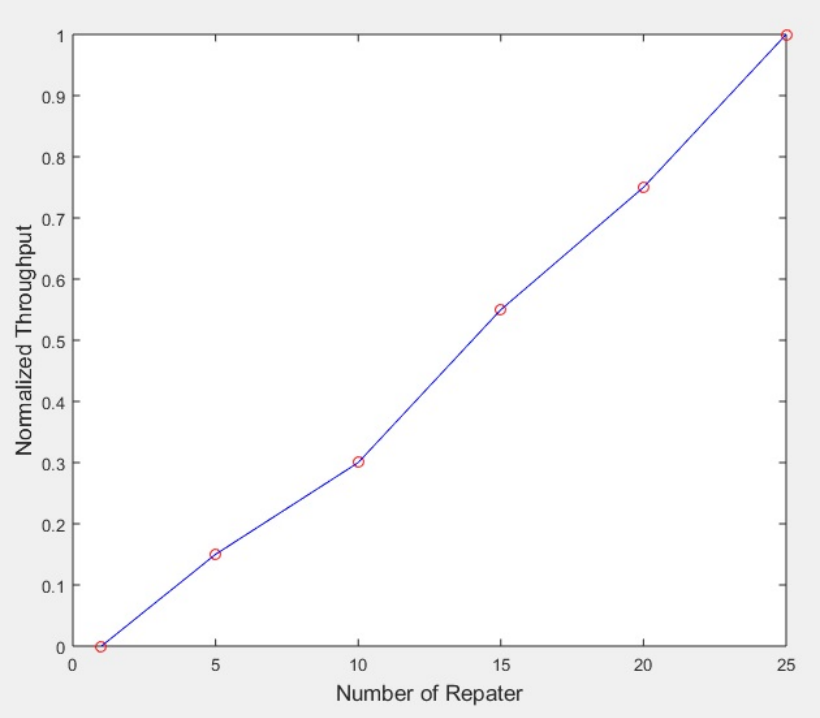

Figure 8.5. Effect of repeater insertion for a superthreshold global interconnect

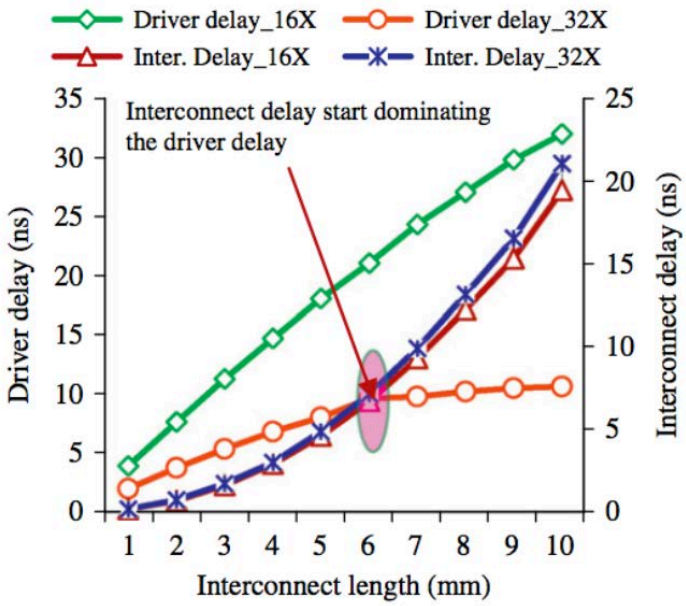

Figure 8.6. Driver and the interconnect delay [20]

8.3. Proposed Method

8.3.1. Motivation

As discussed in Chapter 8.2, the repeater insertion technique is not effective for a subthreshold global interconnect because the driver, not interconnects, dominates the circuit delay in a subthreshold region. Furthermore, this technique is susceptible to PVT variations. Therefore, a subthreshold global interconnect optimization technique that can reduce PVT variations and alleviate performance degradation issues is needed.

8.3.2. Tapered Driver

A tapered driver has been suggested for reduced receiver delay and thus, an overall total path delay [20]. A tapered driver is to use to upsize the interconnect driver as shown in Figure 8.8. For interconnects, RLC networks have been used, and their values were determined using parameters by *International Technology Roadmap for Semiconductors*

(ITRS) [8]. Unlike the traditional interconnect circuit shown in Figure 8.7, a larger interconnect driver was utilized.

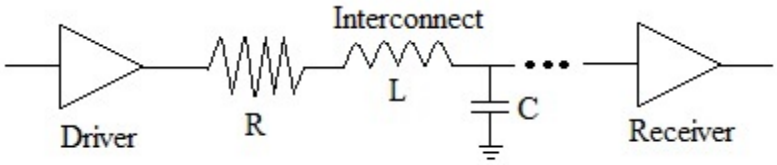

Figure 8.7. An interconnect driver

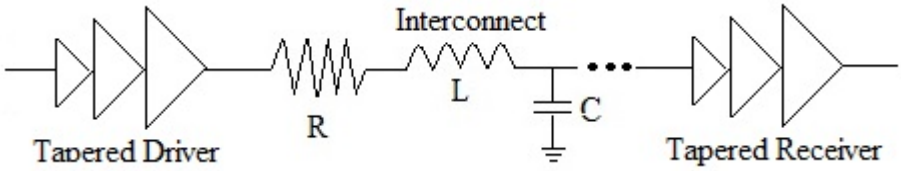

Figure 8.8. A tapered interconnect driver

Because the driver delay dominates the total path delay, using a tapered driver significantly reduces the total path delay with but a small increase in power consumption. However, since the total path delay decreases significantly, total energy— the power-delay product—is saved. The tapered driver (twice the larger size) was used for the same circuit that was simulated in Figure 8.4. without repeater. The simulation result is shown in Figure 8.9.

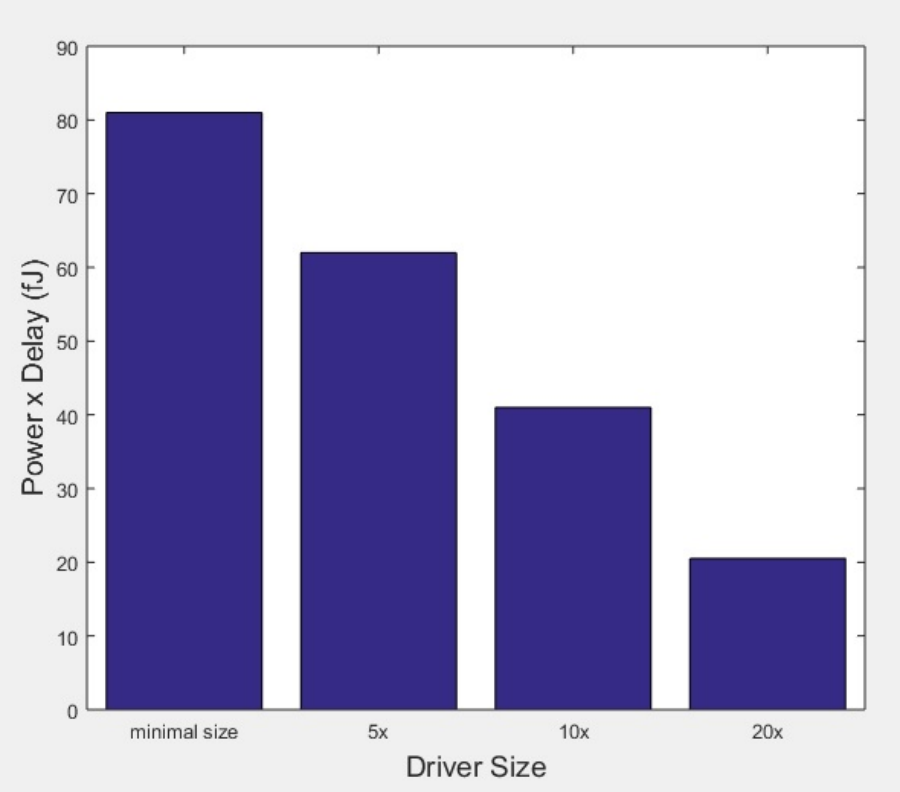

Figure 8.9. Power-delay product

The optimal energy (power-delay product) is achieved at 20x of minimal driver size. Because of a low swing under the subthreshold region, power dissipation increases only minimally [20]. Thus, while total delay is reduced significantly, energy savings is achieved with the tapered driver. With a 20x tapered driver, 75% of the energy saving is achieved. This result varies depending on the interconnect, interconnect length, type of driver, driver size, and the technology being used. Therefore, various lengths of interconnect were simulated, and that result is shown in Figure 8.10.

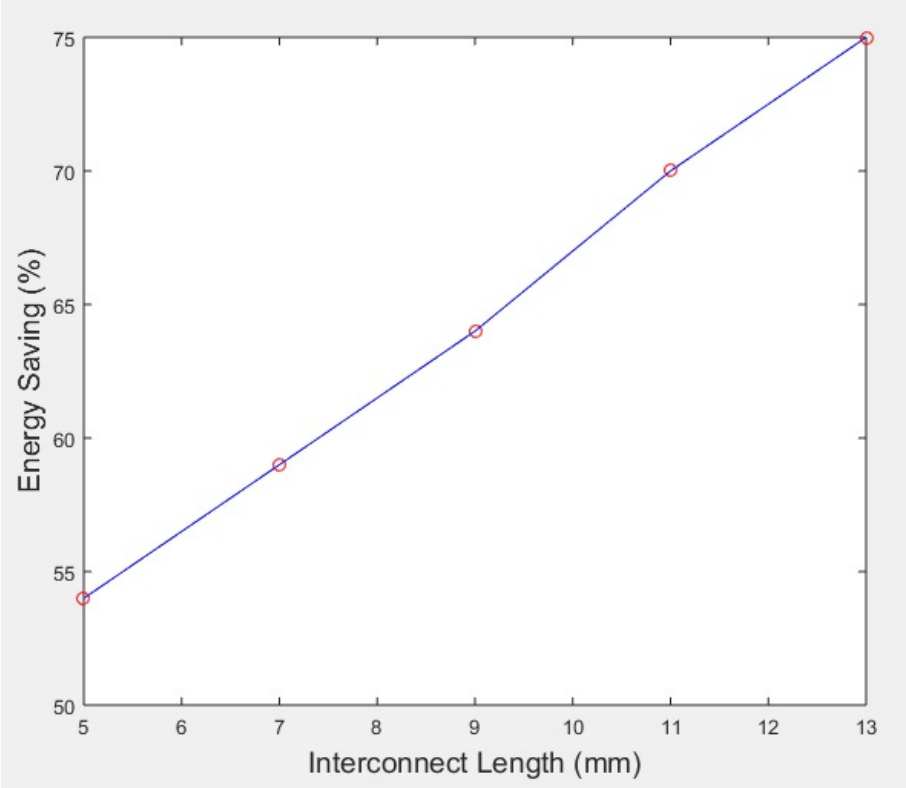

Figure 8.10. Energy saving for various interconnect length

As interconnect length increases, energy saving is also increasing from 60% to 80%. Because both driver delay and interconnect delay are larger in the longer interconnect, and driver delay dominates under the subthrehosld region, energy savings for the tapered driver in the longer interconnect is greater.

8.4. Summary

A commonly used repeater insertion technique for superthreshold was simulated for subthreshold interconnects. As the driver delay dominates, repeater insertion was shown to be ineffective for subthreshold interconnect optimization. Therefore, the tapered driver for global interconnect optimization was analyzed and simulated. That result shows that

total path delay is significantly reduced while power consumption is minimally increased due to much less power in the subthreshold circuit, thus reducing total energy and improving performance. With an optimally sized driver, a 75% reduction in power-delay product was achieved. In addition, the effect of interconnect length was analyzed. It was observed that as interconnect length increases in the subthreshold region, a tapered driver becomes more effective. As global wire delay is increasing more and more due to technology scaling, greater focus on subthreshold driver optimization is needed to deter further performance degradation in the subthreshold region.

Chapter 9

Conclusion

9.1. Conclusion

As the number of transistors is increasing due to technology scaling and improved lithography technology, the focus of study over the last decade has been the concern over power management.

In Chapter 2, the background regarding power dissipation and estimation was explored and the effects of transistor sizing discussed. Furthermore, several power-reducing mechanisms were suggested, and subthrehsold operation indicated as the focus of study in this thesis.

In Chapter 3, the discussion indicated that a circuit operates successfully in a subthreshold region. Its origin, advantages, and limitations are discussed. The two major limitations of subthreshold circuit design are (1) susceptibility to PVT variations and (2) performance degradation.

In Chapter 4, types of variations were discussed along with a concise comparison of them to near-threshold circuit design. In Chapter 5, the Alpha-Power model, analytical model was analyzed to estimate circuit delay. The simulation result revealed that the Alpha-Power law does not accurately characterize subthreshold operation. Thus, a new analytical model that considers variations was derived and verified through simulation.

Further, the effect of technology scaling on PVT variations was analyzed and verified through simulations.

In Chapter 6, the background of linear programming is explained, and in Chapter 7, linear optimization model and an algorithm were formed to assign either high-threshold or low-threshold gates for a dual-threshold voltage CMOS design. By using two different threshold voltages, the circuit consumes less power without circuit failure, thus ensuring that the critical path delay is not violated.

In Chapter 8, the global interconnect delay in superthreshold and subthrehsold was discussed, and it was found that driver delay dominates in the subthreshold region. Therefore, the repeater insertion technique often used in the superthrehsold is not ideal. A tapered driver can be used to optimize energy.

Because both performance degradation and variations are inevitable in both subthreshold and near-threhsold circuits, a variation-aware analytical model was developed, dual-threshold voltage design and its optimization model was then discussed, and a tapered driver optimization technique was used to increase throughput and proven to be more effective than repeater insertions.

9.2. Future Research

Further efforts and research on the combination effects of several techniques discussed in this paper will guide both circuit design and optimization technique that can deter performance degradation and variations susceptibility. For instance, dual-V_{DD} assignment [14] can be combined with dual-threshold voltage assignment to maximize energy saving.

Additionally, further investigation on the effect of technology scaling is needed, as more challenges will appear in the smaller and smaller technology nodes of the future.

Bibliography

[1] S. Borkar, T. Karnik, and V. De, “Design and reliability challenges in nanometer technologies,” in *DAC,* July 2004, pp. 75.

[2] J. R. Burns, “Switching response of complementary-symmetry MOS transistor logic circuits,” *RCA Rev*., vol. 25, pp. 627-661, Dec. 1964.

[3] V. Deodhar and J. Davis, “Voltage scaling and repeater insertion for high-throughput low-power interconnects,” in *Proceedings of the 2003 International Symposium on Circuits and Systems*, May 2003, pp. 349-352.

[4] V. Deodhar and J. Davis, “Optimal voltage scaling, repeater insertion, and wire sizing for wave-pipelined global interconnects,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, no. 4, pp. 1023–1030, May 2008.

[5] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, “Near-threshold computing: reclaiming Moore’s law through energy efficient integrated circuits,” *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, February 2010.

[6] F. Frustaci and P. Corsonello, “Analytical delay model considering variability effects in subthreshold domain,” *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 3, pp. 168-172, March 2012.

[7] N. Hedenstierna and K. O. Jeppson, “CMOS circuit speed and buffer optimization,” *IEEE Trans. Computer-Aided Design*, vol. CAD-6, no. 2. pp. 270-280, Mar. 1987.

[8] ITRS, “International Technology Roadmap for Semiconductors”

[9] R. C. Jaeger and T. N. Blalock. "Chapter 7. Complementary MOS (CMOS) Logic Design" *Microelectronic Circuit Design*, 4th ed. New York: McGraw-Hill, 2011, ch. 7, pp. 367-405.

[10] V. Khandelwal and A. Srivastava, “A general framework for accurate statistical timing analysis considering correlations,” in *DAC*, June 2005, pp. 89-94.

[11] J. Kil, J. Gu, and C. Kim, “A high-speed variation-tolerant interconnect technique for sub-threshold circuits using capacitive boosting,” *IEEE Trans. VLSI Systems*, vol. 16, no. 4, pp. 456-465, April 2008.

[12] C. H. Kim, H. Soeleman, K. Roy, “Ultra-low-power DLMS adaptive filter for hearing aid applications,” in *IEEE TVLSI*, December 2003, pp. 1058-1067.

[13] K. Kim, “Ultra low power CMOS design,” PhD thesis, Auburn University, Dept. of ECE, Auburn, Alabama, May 2011.

[14] K. Kim and V. D. Agrawal, “Ultra low energy CMOS logic using below-threshold dual-voltage supply,” Journal of Low Power Electronics, vol. 7, no. 4, pp. 460-470, Dec. 2011.

[15] J. Le, X. Li, and L. Pileggi. “STAC: statistical timing analysis with correlation”, in *DAC,* July 2004, pp. 343-348.

[16] T. Lin, K. Chong, B. Gwee, J. Chang, and Z. Qiu, “Analytical delay variation modeling for evaluating sub-threshold synchronous/asynchronous designs,” in *IEEE NEWCAS*, June 2010, pp. 69-72.
[17] D. Markovic, C. C. Wang, L. P. Alarcon, T Liu, and J. M. Rabaey. "Ultralow-power design in near-threshold region," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 237-252, February 2010.
[18] A. Nalarnalpu and W. Burleson, “Repeater insertion in deep sub-micron CMOS: ramp-based analytical model and placement sensitivity analysis,” in *IEEE ISCAS*, May 2000, pp. 766-769.
[19] M. A. Ortega and J. Figueras, “Short circuit power modeling in submicron CMOS,” in *PATMOS*, Aug. 1996, pp. 147-166.
[20] S. D. Pable and M. Hasan, "Ultra-low power signaling challenges for subthreshold global interconnects,” *Integration, the VLSI journal*, vol. 45, no. 2, pp. 186-196, March 2012.
[21] PTM: Predictive technology model. In: Nanoscale Integration and Modeling (NIMO) Group. Arizona State University, Arizona.
[22] M. S. Rahaman and M. H. Chowdhury, “Interconnect technique for sub-threshold circuits using negative capacitance effect,” in *IEEE MWSCAS*, August 2009, pp. 1122-1125.
[23] R. Rogenmoser and H. Kaeslin. “The impact of transistor sizing on power efficiency in submicron CMOS circuits.” *IEEE Journal of Solid-State Circuits*, vol. 32, no. 7, pp. 1142-1145, July 1997.
[24] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas." *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584-594, April 1990.
[25] T. Sakurai, “Alpha-Power Law MOS Model,” *IEEE Solid-State Circuits Society Newsletter*, vol. 9, no. 4, pp. 4-5, October 2004.
[26] E. Shauly, “CMOS leakage and power reduction in transistors and circuits: process and layout considerations,” *Journal of Low Power Electronics and Applications*, vol. 2, no. 1, pp. 1-29. January 2012.
[27] A. Wang, B. H. Calhoun, and A. P. Chandrakasan, “Sub-threshold Design for ultra low-power systems.” Springer. 2006.
[28] A. Wang, B. H. Calhoun, and A. P. Chandrakasan. "Origins of weak inversion (or sub-threshold) circuit design." *Sub-threshold Design for Ultra Low-power Systems.* New York: Springer, 2006.
[29] J. Yao and V. D. Agrawal, “Dual-threshold design of sub-threshold circuits,” in *Proceedings of IEEE SOI-3D-Subthreshold Microelectronics Technology Conference*, October 2013, pp. 77–78.
[30] J. Yao, “Dual-Threshold Voltage Design of Sub-Threshold Circuits,” PhD thesis, Auburn University, Dept. of ECE, Auburn, Alabama, May 2014.
[31] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, “Analysis and mitigation of variability in subthreshold design”, in *IEEE ISLPED*, August 2005, pp. 20-25.