Effect of MOSFET Threshold Voltage Variation on High-Performance Circuits

by

Siva G. Narendra

Bachelor of Engineering in Electronics and Communication Engineering
Government College of Technology, Coimbatore, India, June 1992.

Master of Science in Computer Engineering
Syracuse University, Syracuse, NY, June 1994.

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Electrical Engineering and Computer Science
at the
Massachusetts Institute of Technology

January 2002

© 2002 Siva G. Narendra. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part.

Signature of Author ____
Department of Electrical Engineering and Computer Science
January 31, 2002

Certified by____
Anantha Chandrakasan, Ph.D.
Associate Professor of Electrical Engineering
Thesis Supervisor

Certified by____
Dimitri Antoniadis, Ph.D.
Professor of Electrical Engineering
Thesis Supervisor

Accepted by ____
Arthur Smith, Ph.D.
Professor of Electrical Engineering
Graduate Officer

Effect of MOSFET Threshold Voltage Variation on High-Performance Circuits

by

Siva G. Narendra

Submitted to the Department of Electrical Engineering and Computer Science
on January 31, 2002 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

The driving force for the semiconductor industry growth has been the elegant scaling nature of CMOS technology. In future CMOS technology generations, supply and threshold voltages will have to continually scale to sustain performance increase, limit energy consumption, control power dissipation, and maintain reliability. These continual scaling requirements on supply and threshold voltages pose several technology and circuit design challenges. One such challenge is the expected increase in threshold voltage variation due to worsening short channel effect. This thesis will address three specific circuit design challenges arising from increased threshold voltage variation and present prospective solutions. First, with supply voltage scaling, control of *die-to-die threshold voltage variation* becomes critical for maintaining high yield. An analytical model will be developed for existing circuit technique that adaptively biases the body terminal of MOSFET devices to control this threshold voltage variation. Based on this model, recommendations on how to effectively use the technique in future technologies will be presented. Second, with threshold voltage scaling, sub-threshold leakage power is expected to be a significant portion of total power in future CMOS systems. Therefore, it becomes imperative to accurately predict and minimize leakage power of such systems, especially with increasing *within-die threshold voltage variation*. A model that predicts system leakage based on first principles will be presented and a circuit technique to reduce system leakage without reducing system performance will be discussed. Finally, due to different processing steps and short channel effects, threshold voltage of devices of same or different polarities in the same neighborhood may not be matched. This will introduce mismatch in the device drive currents that will not be acceptable in some high performance circuits. In the last part of the thesis, voltage and current biasing schemes that minimize the impact of *neighborhood threshold voltage mismatch* will be introduced.

Thesis Supervisor: Anantha Chandrakasan
Title: Associate Professor of Electrical Engineering

Thesis Supervisor: Dimitri Antoniadis
Title: Professor of Electrical Engineering

Thesis Supervisor: Vivek De
Title: Principal Engineer, Intel Corporation

Thesis Reader: Charles Sodini
Title: Professor of Electrical Engineering

To Appa…for proving that learning never ceases,
and to Amma…for teaching the art of learning.

Acknowledgements

Throughout the course of implementing this work, I had the privilege of interacting with some of the best in the field of electrical engineering, for which I am very grateful. Foremost, I am indebted to my thesis advisors – Prof. Anantha Chandrakasan, for his vision and motivation in guiding me to explore the bridge between devices and circuits; Prof. Dimitri Antoniadis, for being a patient and inspiring teacher; and Dr. Vivek De of Intel Labs, for being an invaluable technical mentor. I am also extremely grateful for their trust in the choices I made to complete this thesis.

I want to thank Prof. Charles Sodini for his time, encouragement, and technical guidance. I would like to express my gratitude to several EECS faculty members, especially, Prof. Duane Boning, Prof. Don Troxel, Prof. Clifton Fonstad, Prof. Jesus delAlamo, Prof. Judy Hoyt, and Prof. Raphael Reif for their valuable support. My stay at MIT was stimulating and entertaining, thanks to the friendship of Dr. Kush Gulati, Dr. James Kao, Dr. Andy Wei, Dr. Isabel Yang, Dr. Mark Armstrong, Dr. Anthony Lochtefeld, Dr. Keith Jackson, Jeremy Milikow, and Prasanth Duvvur. I also would like to acknowledge Marilyn Pierce at the EECS graduate office and Margaret Flaherty for their help in getting this thesis in order.

I am appreciative of the support from all of my colleagues at Intel Labs. Specifically, Brad Bloechel, Jim Tschanz, Matt Haycock, and Greg Dermer for their invaluable support in the lab; Shekhar Borkar, Richard Hofsheier, and Justin Rattner for their technical leadership and for funding this research; Nitin Borkar and team for the coffee breaks, guidance in design, and for silicon real-estate. I am also grateful to Dr. Soumyanath Krishnamurthy for energizing me to complete the research and to Dr. Ali Keshavarzi, Dr. Yibin Ye, and Dr. Dinesh Somasekhar for numerous technical discussions. Many thanks to Greg Ruhl, Dan Klowden, and Zachary Keer for their interest and participation.

All of this was made possible by the love and encouragement of my family. I am deeply indebted to my parents, Dr. M. R. P. Gurusami and Varamani, for teaching me the value in accumulating the wealth of knowledge. I am also very fortunate to have the collective guidance of five older siblings, Nallini, Madhavi, Ezhil, Aravanan, and Senthil. Each sibling and their family have an irreplaceable influence in my life, for which I am very grateful. I want to acknowledge my parents-in-law, Ashok and Meena, and sister-in-law Aditi for their support. Thanks to Inji for her playful company during the process of writing this thesis. Finally, I am exceedingly thankful for the immense love and friendship from my soul mate Monika.

Contents

Chapter 1 Introduction **21**

1.1 Thesis organization 24

Chapter 2 Background **25**

2.1 Technology scaling and threshold voltage variation 25

2.2 Threshold voltage variation categories 29

Chapter 3 Die-to-die and Within-die Threshold Voltage Variations **33**

3.1 Adaptive body bias 33

3.1.1 Adaptive body bias and short channel effect (SCE) 34

3.1.2 Scaling of required body bias and SCE increase 33

3.1.3 Impact on within-die threshold voltage variation 41

3.1.4 Summary 43

3.2 Bi-directional adaptive body bias 43

3.3 Body bias circuit impedance requirement 50

Chapter 4 Within-die Threshold Voltage Variation and Leakage Power **57**

4.1 Estimation of chip leakage current 57

4.1.1 Present leakage current estimation techniques 57

4.1.2 Leakage current estimation including within-die variation 58

4.1.3 Measurement results 61

4.2 Leakage reduction 62

4.2.1 Model for stack effect factor 64

4.2.2 Leakage reduction using forced-stacks 69

4.2.3 Stack effect vs. channel length increase 71

4.2.4 Case study and summary 73

Chapter 5 Neighborhood Threshold Voltage Variation 75

5.1 Voltage biasing 76

5.1.1 Application of voltage bias to low-voltage sense-amplifiers 77

5.2 Current biasing 81

5.2.1 Basic iso-current biasing and two-phase clock generation 81

5.2.2 Process insensitive current biasing 84

5.2.2.1 Process insensitive constant current generation 85

Chapter 6 Conclusion 91

6.1 Contributions 91

6.2 Suggestions for future work 93

Bibliography 95

List of Figures

Figure 1-1: Timeline on technology scaling and new microprocessor architecture introduction. ...22

Figure 1-2: Basic form of Moore's law. 22

Figure 1-3: Relative die sizes of the last nine microprocessor generations. 23

Figure 2-1: Barrier height lowering due to channel length reduction and drain voltage increase in an nMOS 26

Figure 2-2: Barrier lowering (BL) resulting in threshold voltage roll-off with channel length reduction. Drain induced barrier lowering (DIBL) reduces threshold voltage for short channel devices and increases threshold voltage roll-off. For short channel devices channel length variation (ΔL) translates to threshold voltage variation (ΔV_T) 26

Figure 2-3: Dependence of threshold voltage variation on channel length and drain voltage; n is the number of MOS device samples measured. 27

Figure 2-4: Relationship between threshold voltage (Vt) and sub-threshold leakage current (Ioff).27

Figure 2-5: Trend in sub-threshold leakage and switching power with technology scaling. 28

Figure 2-6: Threshold voltage variation categories covered in the thesis. 30

Figure 3-1: Die-to-die threshold voltage distributions (a) Conventional approach without adaptive body bias (b) Adaptive body bias approach. 33

Figure 3-2: Reduction in *Vt* modulation with reverse body bias with reduction in *Vt*. 35

Figure 3-3: Increase in Vt-roll-off with *Vt* reduction and reverse body bias increase....35

Figure 3-4: Increase in DIBL due to increase in reverse body bias..36

Figure 3-5: (a) Adaptive body bias reduces the die-to-die *Vt* variation. (b) Within-die *Vt* variation increases for die samples that require body bias to match their mean *Vt* to the target *Vt*. *Vt-target* is the target saturation threshold voltage for a given technology. *Vt-low* and *Vt-nom* are the minimum and mean threshold voltages of the die-to-die distribution. 38

Figure 3-6: Trend in mean saturation threshold voltage of different die samples before adaptive body bias under (a) 30% *Vt* scaling and (b) 20% *Vt* scaling scenarios...............................40

Figure 3-7: Matching of mean saturation threshold voltages of different die samples with adaptive body bias under (a) 30% *Vt* scaling and (b) 20% *Vt* scaling scenarios...............................40

Figure 3-8: Increase in within-die threshold voltage variation due to increase in short channel effect with adaptive body bias under (a) 30% *Vt* scaling and (b) 20% *Vt* scaling. We assume that the dominant reason for within-die *Vt* variation is critical dimension variation. The results shown here assume within-die variation in *Lg* of 5%.......................................42

Figure 3-9: Die-to-die threshold voltage distributions (a) Conventional approach without adaptive body bias (b) traditional adaptive body bias approach – die sample that requires maximum reverse body bias is 2ΔVt2 away from Vt-target (c) bi-directional adaptive body bias approach – die sample that requires maximum reverse body bias is ΔVt1 away from Vt-target. Note: ΔVt2 > ΔVt1 since SCE of devices with lower *Vt* will be more. 44

Figure 3-10: Chip micrograph of a sub-site..45

Figure 3-11: Circuit block diagram of each sub-site. 46

Figure 3-12: Demonstration of frequency adapting to meet target and list of possible on-chip bias modes...46

Figure 3-13: Die-to-die variation in frequency and leakage for no body bias (NBB), 0.2 V static forward body bias (FBB), and adaptive body bias applied to compensate die-to-die variation (ABB). 47

Figure 3-14: Frequency vs. number of critical paths that determine the frequency. 47

Figure 3-15: Comparison of variations in within-die device current and frequency. 48

Figure 3-16: Die-to-die variation in frequency and leakage for adaptive body bias applied to (i) compensate die-to-die variation (ABB) and (ii) compensate within-die variation (WID-ABB). 49

Figure 3-17: Histogram of bias voltages within a die sample and effect of bias resolution on frequency distribution. 50

Figure 3-18: Communications router chip architecture with PMOS body bias. 51

Figure 3-19: Measurement of body and Vcc current. 51

Figure 3-20: Overview of body bias generation and distribution. 52

Figure 3-21: Buffer impedance requirements and body bias noise comparisons with NBB. 53

Figure 3-22: LBG buffer implementations and comparisons. 54

Figure 3-23: Frequency vs. Vcc of FBB and NBB chips. 55

Figure 3-24: Leakage reduction from active to standby mode in FBB chips. 56

Figure 3-25: Micrograph of communications router chip with PMOS body bias and of chip characteristics. 56

Figure 4-1: Comparison of calculated leakage versus measured leakage for (a) existing leakage current estimation techniques and (b) leakage current estimation technique introduced in this thesis. 61

Figure 4-2: Ratio of measured to calculated leakage current ratio distribution for $I_{leak\text{-}u}$, $I_{leak\text{-}l}$, and $I_{leak\text{-}w}$ techniques (Sample size: 960). 61

Figure 4-3: Leakage current difference between a single *off* device and a stack of two *off* devices. As illustrated by the energy band diagram, the barrier height is modulated to be higher for the two-stack due to smaller drain-to-source voltage resulting in reduced leakage. 62

Figure 4-4: Trade-off between standby leakage and performance by forcing a two-stack under iso-input load. An NMOS two-stack will reduce leakage when input stays at logic "0" 63

Figure 4-5: Load line analysis showing the leakage reduction in a two-stack. 65

Figure 4-6: Measurement results showing the relationship between stack effect factor *X* for a two-stack to the universal exponent *U*. Lines indicate the relationship as per the analytical model and symbols are from measurement results. White symbols are for nominal channel devices and gray symbols are for devices smaller than the nominal channel length. Triangle, circle, and square symbols are for V_{dd} of 1.5, 1.2, and 1.1 V respectively. Zero body bias is when the body-to-source diode of the device closet to the power supply is zero biased and reverse body bias is when the diode is reverse biased by 0.5 V. 66

Figure 4-7: Measurement results indicate a slower rate of increase in leakage of two-stack compared to that of a single device. This should translate to reduction in the variation of effective threshold voltage. 67

Figure 4-8: Nominal channel length device measurement results showing stack effect factor across two technology generations. The increase in stack effect factor is attributed to worsening of short channel effect, λ_d, which is predicted by the analytical model. The higher stack effect factor for the low-V_t device in 0.13-µm technology generation is attributed to the same reason. Lines are from analytical model and symbols are from measurement. 67

Figure 4-9: Nominal channel length device measurement results indicating the scaling of stack effect factor from 0.18µm to 0.13µm low-V_t under different V_{dd} scaling conditions. The

low-V_t device will dominate leakage in 0.13μm technology, so the comparison is made with the low-V_t device. 68

Figure 4-10: Prediction in the scaling of stack effect factor for two V_{dd} scaling scenarios in nominal channel length devices. V_{dd} for 0.18μm is assumed to be 1.8 V............................68

Figure 4-11: Stack forcing and dual-V_t can reduce leakage of gates in paths that are faster than required...69

Figure 4-12: Simulation result showing the nominal channel length delay versus mean leakage trade-off that can be achieved by stack forcing technique under iso-input load conditions. Iso-input load is achieved by making the gate area after stack forcing identical to before stack forcing. Several such conditions are possible, which enhances delay-leakage trade-off possible by stack forcing. The two-stack condition with the least delay is for $w_u=w_l=½w$. This trade-off can be used with or without high-V_t transistors. 70

Figure 4-13: A sample path where natural stack is used to reduce standby leakage by applying a predetermined vector during standby. No delay penalty is incurred with this technique.....70

Figure 4-14: Using stack-forcing technique the number of logic gates in stack mode can be increased. This will enable further leakage reduction in standby mode. Increase in delay under normal mode of operation will be incurred. 71

Figure 4-15: If a gate can have its input as either “0” or “1” and still force stack effect then that gate will have reduced active leakage. The more the number of inputs that can be either “0” or “1” the higher the probability that stack effect will reduce active leakage. 71

Figure 4-16: Comparing device leakage reduction due to channel length increase with two-stack leakage. The channel length is given by η x 0.18 μm. Stack leakage is a two stack of devices with η=1 and $wu=wl=½w$. Leakage numbers are obtained from simulation under iso-input load. 72

Figure 4-17: Energy-delay trade-off of inverter under different configurations with fan-out of 1 and iso-input load. The simulation-based comparison clearly shows that the two-stack configuration's delay is less than increasing channel length, especially when compared to iso-standby leakage (η=3) configuration. 72

Figure 4-18: Summary of delay-leakage trade-off comparison between two-stack and channel length. 73

Figure 5-1: Die-micrograph of mismatch structures testchip. 76

Figure 5-2: Linear threshold voltage mismatch for 500 mV forward body bias, zero body bias and 500 mV reverse body bias. 77

Figure 5-3: Traditional sense-amplifier. 78

Figure 5-4: Body voltage for the traditional sense-amplifier. 78

Figure 5-5: Dependence of saturation threshold voltage mismatch on body bias. 79

Figure 5-6: New no body bias sense-amplifier. 79

Figure 5-7: Total delay verses input differential for iso-output differential at 1.5 V, 1 mV/pS ramp rate, and 110 Celsius, for the traditional and the new sense-amplifiers. 79

Figure 5-8: Total delay (sense-amplifier delay plus ramp development delay) improvement due to input offset reduction in the new sense-amplifier at 1.5 V, 1 mV/pS ramp rate and 110 Celsius. 80

Figure 5-9: Basic iso-current biasing scheme 82

Figure 5-10: Standard two-phase clock generator design 82

Figure 5-11: Iso-bias current based non-overlapping two-phase clock generator. 83

Figure 5-12: Performance comparison of two-phase generators. 84

Figure 5-13: Process insensitive current biasing scheme. 85

Figure 5-14: Measured process variation in a long-channel, wide-width, process-uncompensated, device current (I_u). Measurements were carried out across wafer on identical devices with 0.9 V gate drive. Both raw data and statistical information are presented above................87

Figure 5-15: Normalized process variation in I_{ref} for different device size ratios when a=2 and b=5. Measurement confirms process variation in I_{ref} minimizes at $z1/z2$ ratio predicted by the theoretical model. 88

Figure 5-16: Measured variation in I_{ref} for a=2, b=5, and $z1/z2$=1/8. Device current and V_t measurements were carried out across wafer on two devices with appropriate gate drives and device sizes given by the theoretical model..88

Figure 5-17: Circuit schematics showing generation of V_t and I_{ref}. Since generated V_t will not be accurate, device size ratio $z1/z2$ was optimized with a=2, b=5 and V_{dd}=0.9 V to minimize I_{ref}'s process variation. 89

Figure 5-18: Circuit simulation results with a=2, b=5, $z1/z2$=1/6, V_{dd}=0.9 V, showing variation in I_u and I_{ref}. With respect to typical process corner I_u varied by +22% and -16% while variation in I_{ref} was –5% and –5%. Total variation in normalized I_u across all process corners is 0.38 while it is 0.05 for normalized I_{ref}. 89

List of Tables

Table 3-1: Technology parameters under two scaling scenarios. 37

Table 3-2: With adaptive body bias short channel effect of devices increase, indicated by DIBL (λ_d in mV/V) increase and body effect reduction factor (λ_b) decrease. This SCE increase is worse for *Vt-low* devices, compared to *Vt-nom* devices, as they require larger body bias to match *Vt-target*. The required bias values (in V) are indicated within parentheses. 41

Table 5-1: Total delay improvement under different supply voltage and ramp rate conditions for input differential of 150 mV for the traditional sense-amplifier and 118 mV for the new zero body bias sense-amplifier at 110 Celsius. Larger improvement is correlated to faster sense-amplifier resulting in input offset and ramp development delay reductions more critical. 80

Table 5-2: Sub-set of parameters that satisfy equations (6)-(7) to minimize process impact on $I_{ref} = I_1 - I_2$. 87

Table 5-3: Low voltage operation enabled by redesigning Vt generation circuit............................ 90

Chapter 1
Introduction

MOS transistor based integrated circuits have transformed the world we live in. It is estimated that there are more than 15 billion silicon semiconductor chips currently in use with an additional 500,000 sold each day [1]. The ever shrinking size of the MOS transistors that result in faster, smaller, and cheaper systems have enabled ubiquitous use of these chips. Among these semiconductor chips, a prevalent component is the high-performance general-purpose microprocessor. Figure 1-1 illustrates the timeline on technology scaling and new high-performance microprocessor architecture introductions in the past three decades [2]. This trend holds in general for other segments of the semiconductor industry as predicted by Moore's law [3]. In 1965, Gordon Moore showed that for any MOS transistor technology there exists a minimum cost that maximizes the number of components per integrated circuits. He also showed as transistor dimensions are shrunk (or scaled) from one technology generation to the next, the minimal cost point allows significant increase of the number of components per integrated circuit as shown in Figure 1-2.

Historically, technology scaling resulted in scaling of vertical and lateral dimensions by 0.7X each generation resulting in delay of the logic gates to be scaled by 0.7X and the integration density of logic gates to be increased by 2X. From the timeline shown in Figure 1-1 it is clear that there were two distinct eras in technology scaling – constant voltage scaling and constant electric field scaling.

Constant voltage scaling era (*First two decades*): Technology scaling and new architectural introduction in this era happened every 3.6 years. Technology scaling should scale delay by 0.7X translating to 1.4X higher frequency. However, frequency scaled by 1.7X with the additional

increase primarily brought about by increase in the number of logic transistors. As it can be seen from Figure 1-1 the number of logic transistors increased by 3.3X in each of the new introductions. Technology scaling itself would have provided only 2X – the additional increase was enabled by increase in die area of about 1.5X every generation [4].

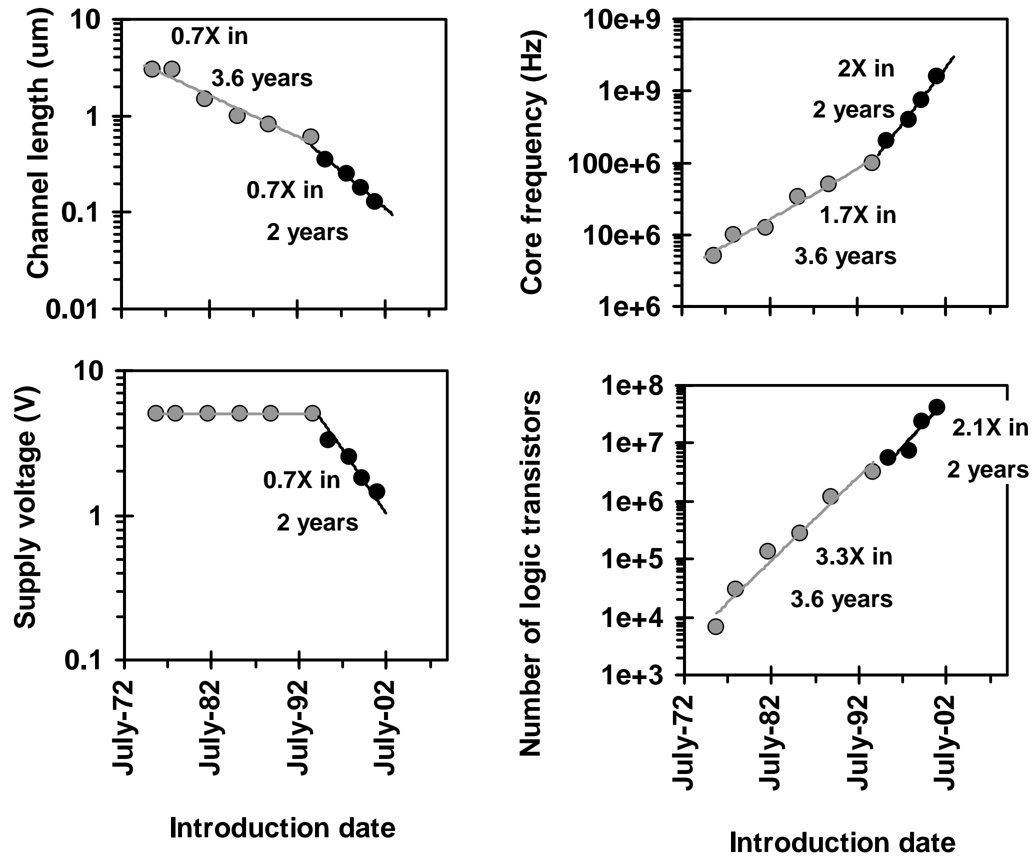


Figure 1-1: Timeline on technology scaling and new microprocessor architecture introduction.

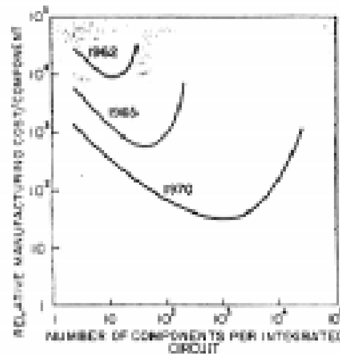


Figure 1-2: Basic form of Moore's law.

Constant electric field scaling era (*Past decade*): Technology scaling and new architectural introduction in this era happened every 2 years along with voltage scaling of 0.7X. As always technology scaling should scale delay by 0.7X translating to 1.4X higher frequency, but frequency increased by 2X in each new introduction. The additional increase in frequency was primarily brought by decrease in logic depth through architectural and circuit design advancements. The number of logic transistors grew only by about 2.1X every generation, which could be achieved without significant increase in die area. Since switching power is proportional to Area x ε/distance x Vdd x Vdd x F, it increased by (1 x 1/0.7 x 0.7 x 0.7 x 2 =) 1.4X every generation. Although the die size growth is not required for logic transistor integration, it is important to note that the total die area did continue to grow at the rate of 1.5X per generation [4] due to increase amount of integrated memory. Relative die areas for the last nine microprocessor generations are shown in Figure 1-3.

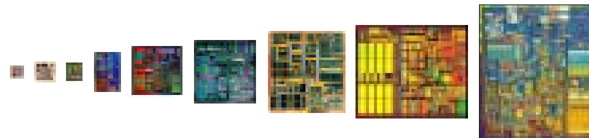


Figure 1-3: Relative die sizes of the last nine microprocessor generations.

In the past decade, technology and new architecture product cycles reduced from 3.6 years to 2 years. From an operational perspective, this requires concurrent engineering in product design, process design, and manufacturing supply lines [5]. The past decade also required supply voltage scaling imposed by oxide reliability and the need to slow down the switching power growth rate. From the process design stand point supply voltage scaling requires threshold voltage scaling [6, 7] so that the technology scaling can continue to provide 1.4X frequency increase. To prolong the tremendous growth the industry has experienced in the past three decades threshold voltage scaling and concurrent engineering has to continue. These requirements pose several challenges in the coming years including increase in process variation, worsening interconnect RC delay, and increase in sub-threshold, gate, and tunneling leakage components [7, 8]. This thesis will focus on one of the challenges – the increasing importance of threshold voltage variation and how it impacts digital CMOS circuits used in microprocessors and other high-performance integrated circuits.

1.1 Thesis organization

In the subsequent chapters the effects of MOSFET threshold voltage variation on the leakage power, delay, and operation of high-performance digital CMOS circuits, and potential circuit solutions that alleviate these effects will be presented in the following order:

- Chapter 2 provides a brief background on the reasons for the increasing importance of threshold voltage variation, existing solutions, and a detailed overview of the research concepts investigated in this thesis.
- Chapter 3 focuses on different aspects of die-to-die threshold voltage variation and its impact on delay and power of the integrated circuit. Ineffectiveness of prior published circuit solution to minimize die-to-die threshold voltage variation as technology scales and the detrimental interaction this solution introduces between die-to-die and within-die threshold voltage variations are identified. An improved circuit solution that is void of these defects is described.
- Chapter 4 introduces (i) the importance of taking into account the influence within-die threshold voltage variation will have on system's leakage power especially as technology scales and (ii) a circuit technique to reduce system leakage power.
- Chapter 5 describes circuit techniques to reduce the impact threshold voltage mismatch between MOS devices in the same neighborhood.
- Summary of this work is described in Chapter 6. Suggestions for future work are also discussed in Chapter 6.

Chapter 2
Background

Conventionally, CMOS technology has been scaled to provide 30% smaller gate delay with 30% smaller dimensions, resulting in CMOS systems operating at about 40% higher frequency in half the area with reduced energy consumption. Scaled CMOS systems, such as new generation microprocessors, achieve at least an additional 60% frequency increase with augmented architecture and circuit techniques. This complexity increase results in higher energy consumption, peak power dissipation and power delivery requirements [4].

To limit the energy and power increase in future CMOS technology generations supply voltage will have to continually scale. The amount of energy reduction depends on the magnitude of supply voltage scaling [9]. Along with supply voltage scaling, MOSFET device threshold voltage will have to scale to sustain the traditional 30% gate delay reduction. This supply and threshold voltage scaling requirements pose several technology and circuit design challenges [4, 8, 10]. One such challenge is the expected increase in threshold voltage variation due to worsening short channel effects. This is explained in the following section.

2.1 Technology scaling and threshold voltage variation

With technology scaling, the MOSFET's channel length is reduced. As the channel length approaches the source-body and drain-body depletion widths, the charge in the channel due to these parasitic diodes become comparable to the depletion charge due to the MOSFET gate-body voltage [11], rendering the gate and body terminals to be less effective. As the band diagram illustrates in Figure 2-1, the finite depletion width of the parasitic diodes do not influence the energy barrier height to be overcome for inversion formation in a long channel device. However, as the channel length becomes shorter both channel length and drain voltage reduce this barrier height. This two-

dimensional effect makes the barrier height to be modulated by channel length variation resulting in threshold voltage variation as shown in Figure 2-2. The amount of barrier height lowering, threshold voltage variation, and gate and body terminal's channel control loss will directly depend on the charge contribution percentage of the parasitic diodes to the total channel charge. Figure 2-3 shows measurements of 3σ threshold voltage variations for three device lengths in a 0.18-μm technology confirming this behavior. It is essential to mention that in sub-micron technologies variation in several physical and process parameters lead to variation in the electrical behavior of the MOS device. The discussions in this thesis will address variation in the electrical behavior manifested as threshold voltage variation because of parameter variation. In addition, the threshold voltage variations addressed here are due to short channel effect in scaled MOS devices and not on threshold voltage variation due to random dopant fluctuation effect. Random dopant fluctuation effect is expected to be one of the significant sources of threshold voltage variation in devices of small area [12].

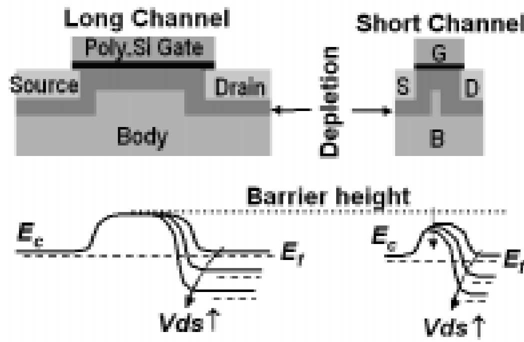


Figure 2-1: Barrier height lowering due to channel length reduction and drain voltage increase in an nMOS.

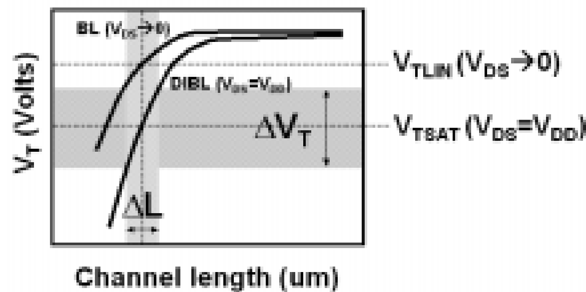


Figure 2-2: Barrier lowering (BL) resulting in threshold voltage roll-off with channel length reduction. Drain induced barrier lowering (DIBL) reduces threshold voltage for short channel devices and increases threshold voltage roll-off. For short channel devices channel length variation (ΔL) translates to threshold voltage variation (ΔV_T)

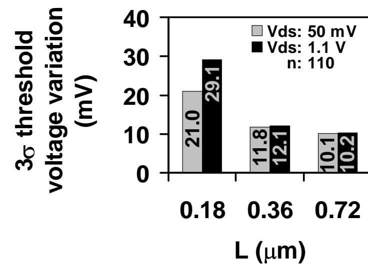


Figure 2-3: Dependence of threshold voltage variation on channel length and drain voltage; n is the number of MOS device samples measured.

It was mentioned in Chapter 1 that in order to maintain the performance increase trend with technology scaling threshold voltage would have to be scaled along with supply voltage. However, reduction in threshold voltage increases the sub-threshold leakage current significantly. Relationship between threshold voltage and sub-threshold leakage is illustrated in Figure 2-4. Typically, reduction in threshold voltage of about 85 mV, as shown in Figure 2-4, will increase the sub-threshold leakage current by 10X.

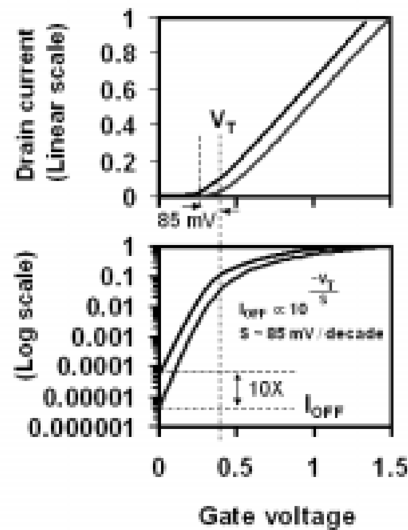


Figure 2-4: Relationship between threshold voltage (Vt) and sub-threshold leakage current (Ioff).

As indicated Chapter 1 switching power increases by 1.4X per generation. With scaling of threshold voltage sub-threshold leakage power will increase at a very rapid rate due to its strong dependence on the threshold voltage. Figure 2-5 illustrates the comparison between the increase in the switching power and sub-threshold leakage power with technology scaling. As it is evident from the figure sub-threshold leakage power will be comparable to the switching power in the immediate future. This ‘inefficient’ leakage power manifests itself as active leakage that influences the total power budget during operation and as standby leakage that influences the battery life of hand held systems. It therefore becomes important to not only reduce sub-threshold leakage power but also accurately estimate it.

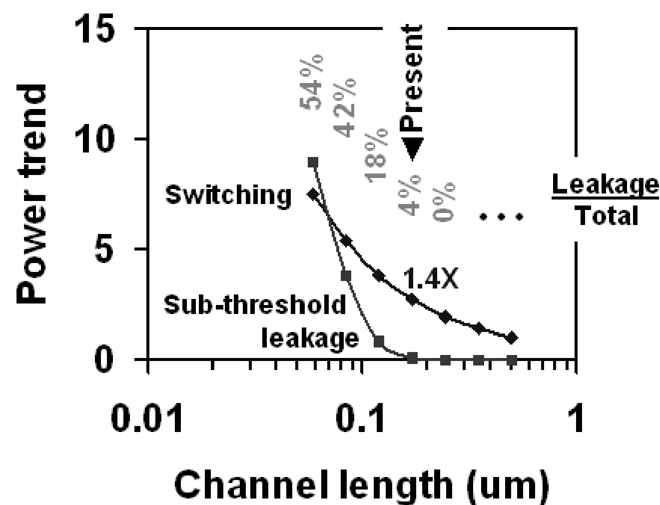


Figure 2-5: Trend in sub-threshold leakage and switching power with technology scaling.

With supply and threshold voltage scaling, control of threshold voltage variation becomes essential for achieving high yields and limiting worst-case leakage [13]. Maintaining good device aspect ratio, by scaling gate oxide thickness is important for controlling threshold voltage tolerances [7]. With the silicon dioxide gate dielectric thickness approaching scaling limits due rapid increase in gate tunneling leakage current [14, 15] researchers have been exploring several alternatives, including the use of high permittivity gate dielectric, metal gate, novel device structures and circuit based techniques [16, 17, 18, 19, 20, 21]. The use of high permittivity gate dielectric will result in thicker and easier to fabricate dielectric for iso-gate oxide capacitance with potential for significant reduction in gate leakage. Identification of a proper high permittivity dielectric material that has good interface states with silicon along with limited gate leakage is in progress [16]. However, it has also been shown that use of high permittivity gate dielectric has limited return [17]. Use of metal gate prevents poly-depletion resulting in a thinner effective gate dielectric. However, identification of dual metal gates to replace the n+ and p+ doped polysilicon is essential to maintain threshold voltage scaling. In addition, novel device structures such as self-aligned double gate planar MOSFETs provide better device aspect ratio [18]. Other than material and device based solutions, circuit design solutions such as threshold canceling logic [19] and adaptive body bias [20, 21] enable supply and threshold voltage scaling. Threshold canceling logic mimics threshold voltage scaling by defining the MOS off state with $|V_{gs}| > 0$, instead of $|V_{gs}| = 0$. Although threshold canceling logic enables threshold voltage scaling, it requires larger area due to increase in logic complexity and number of power grids.

2.2 Threshold voltage variation categories

The three threshold voltage variation categories illustrated in Figure 2-6, which impact high-performance circuit design, will be covered in the next three chapters. In Chapter 3 of this thesis an analytical model will be developed, to show that traditional adaptive reverse body bias circuit solution to reduce die-to-die threshold voltage variation is not scalable for future generations and this technique results in increased within-die threshold voltage variation [22]. Use of bi-directional adaptive forward and reverse body bias to limit threshold voltage variation is more promising [23]. Forward body bias can be used not only to reduce threshold voltage [24, 25], but also to reduce die-to-die and within-die threshold voltage variations as will be shown in Chapter 3. Bias circuit impedance requirements for on-chip body bias are also discussed in Chapter 3.

It is important to note that threshold voltage variation not only affects supply voltage scaling but also the accuracy of leakage power estimation. Accurate leakage power estimation is very critical for future CMOS systems since the leakage power is expected to be a significant portion of the total power due to threshold voltage scaling [4]. In Chapter 4, leakage power estimation that takes into account within-die threshold voltage variation will be presented. In a leakage dominant CMOS system, it also becomes inevitable to identify techniques to reduce this variation and leakage power. In Chapter 4 the use of stacked devices to reduce system leakage power without reducing system performance will be shown. An analytical model to predict the scaling nature of this stack effect and verification of the model through statistical device measurements will be presented. Measurements also show reduction in threshold voltage variation for stacked devices compared to non-stack devices. Comparison of stack effect to the use of high threshold voltage or longer channel length devices for leakage reduction will also be discussed [26].

Chapter 5 of this thesis will deal with the variation in the threshold voltage of matched devices that are in the same neighborhood. The devices that are in close proximity can be either of the same polarity or of different polarity. Matched devices of the same polarity are used as sense-amplifier input devices for low voltage swing sensing among other applications [27]. Any mismatch in threshold voltage of this input device pair will appear as input offset resulting in degraded performance. A simple voltage-biasing scheme that reduces the mismatch between matched transistor pair of same polarity will be discussed.

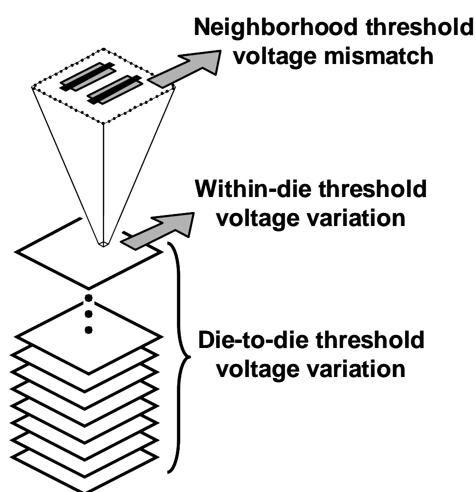


Figure 2-6: Threshold voltage variation categories covered in the thesis.

In addition, for some digital CMOS circuits a known PMOS to NMOS drive current ratio is required either to achieve a well-defined switching threshold or to achieve equal rising and falling delays. Since the processing steps such as threshold voltage implants for the PMOS and NMOS devices are not correlated there could be significant variation between the required and achieved threshold voltages for the two device types. The short channel effects further worsen this variation. The net variation will change the drive current ratio of PMOS to NMOS devices and can affect the operation of high performance circuits that depend on a pre-determined skew between the two device types. Ability to adjust the charging and discharging currents by sensing the skew difference can alleviate this problem. In Chapter 5 current biasing schemes that maintain the relationship between the charging and discharging currents, independent of the process skew is explained. The first current scheme that is the simplest, guarantees constant ratio between charging and discharging currents no matter the change in the relative skews of the PMOS and NMOS devices. Although this scheme maintains the relationship between charging and discharging delays, it doesn't provide constant delay as the threshold voltages vary. A true process insensitive current generation theory and circuit will be described in Chapter 5 [28]. This can then be used as bias current for the charging PMOS and the discharging NMOS networks enabling a threshold voltage variation and skew variation insensitive circuit. Example circuits that benefit from these biasing schemes will be presented. Apart from the digital circuits, a true process insensitive current can be used for numerous biasing applications in analog circuits.

Chapter 3
Die-to-die and Within-die Threshold Voltage Variations

3.1 Adaptive body bias

Supply voltage (*Vdd*) and threshold voltage (*Vt*) scaling is the most effective approach to keep active power dissipation under control while maintaining performance improvement [9]. One of the limits to *Vdd* scaling is the expected increase in *Vt* variation [8, 13]. Increase in die-to-die *Vt* variation will result in slow dies that do not meet the frequency target and fast dies that exceed the allowed power limits due to excessive leakage. The resulting reduction in yield will lead to increases in manufacturing cost and time to market, neither of which is acceptable especially with the technology life cycle shrinking from 3.6 to 2 years (Figure 1-1). Adaptive body bias schemes have been proposed in the past to reduce this expected increase in die-to-die *Vt* variation [20, 21].

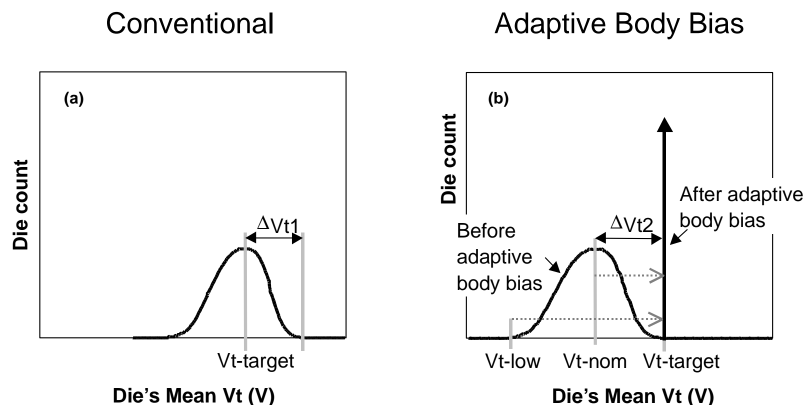


Figure 3-1: Die-to-die threshold voltage distributions (a) Conventional approach without adaptive body bias (b) Adaptive body bias approach.

Figure 3-1(a) illustrates that in a conventional approach without adaptive body bias the mean *Vt* of

all the die samples do not match the target *Vt*. By using adaptive body bias, a sharper distribution in die-to-die *Vt* variation can be achieved, as shown in Figure 3-1(b). Adaptive body bias first requires modification of the process so that mean *Vt* of all the dies are lower than the target *Vt*, as depicted in Figure 3-1(b). This lowering of *Vt* for a given technology is accomplished by reducing the channel doping which increases the depletion width of the MOSFET parasitic junction diodes. It was shown in Section 2.1 that this would result in increased *Vt* variation due to worsened short channel effect (SCE)! Therefore, $\Delta Vt2 > \Delta Vt1$ in Figure 3-1. After this process modification, depending on the mean *Vt* of a die sample an adaptive amount of reverse body bias is applied to the entire die so that its mean *Vt* will be increased to match the target *Vt*, as illustrated in Figure 3-1(b).

Reverse body bias increases the depletion width of the MOSFET parasitic junction diodes [29]. It was shown in Section 2.1 that this would result in increased *Vt* variation due to worsened short channel effect (SCE)! The research objectives in Section 3.1 are (1) to study the effectiveness of adaptive body bias in controlling die-to-die *Vt* variation as technology is scaled and (2) to determine impact of adaptive body bias on within-die *Vt* variation. It will shown that as MOSFET technology is scaled, the body bias required for compensating die-to-die *Vt* variation increases, which in turn further increases SCE, and, because of this increase in SCE, within-die *Vt* variation becomes worse. It will also be shown that the die that requires larger body bias to match its mean *Vt* to the target *Vt* will end up with a higher within-die *Vt* variation. The resulting increase in within-die *Vt* variation due to adaptive body bias can impact clock skew, worst-case gate delay, worst-case device leakage current, total chip leakage power, and analog circuit performance. More importantly, increase in within-die *Vt* can also reduce the frequency of operation in high performance designs that have increasingly lesser logic stages between flip-flops [32, 34]. This will be elaborated in the second of this chapter. In the rest of this section, the effectiveness of adaptive body bias and within-die *Vt* variation due to adaptive body bias will be analytically quantified for three technology generations. To reiterate the point from Section 2.1, the focus of *Vt* variation in this thesis is due to worsening SCE with technology scaling and channel length variation.

3.1.1 Adaptive body bias and short channel effect (SCE)

For adaptive body bias the *Vt* of the process technology has to be re-targeted to be lower as shown in Figure 3-1. In a given technology this is achieved by lower channel doping that will result in lower body effect to begin with. Since adaptive body bias depends on body effect to modulate *Vt*

with reverse body bias, lowering *Vt* will render adaptive body bias less effective. The body effect is further reduced in short channel devices because lower *Vt* with reduced channel doping will increase diode depletion charge and SCE. Figure 3-2 illustrates the reduction in body effect due to *Vt* lowering in a 0.25 μm technology. For an MOS device with *Vt* of 0.4 V, reverse body bias of 0.6 V increased the *Vt* by 25%. *Vt* modulation for the same amount of reverse body bias reduces to less than 8% for an MOS device with Vt of 0.25 V.

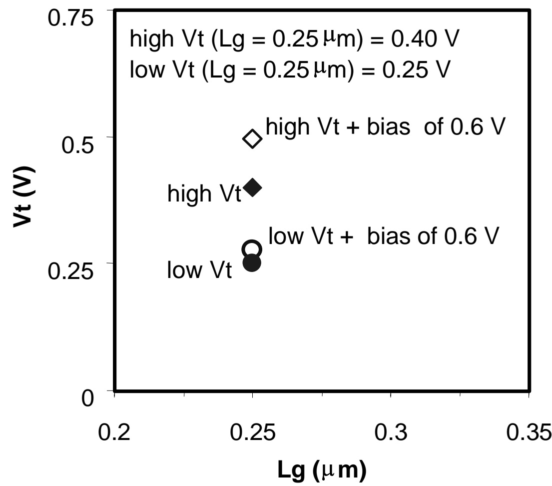


Figure 3-2: Reduction in *Vt* modulation with reverse bodv bias with reduction in *Vt*.

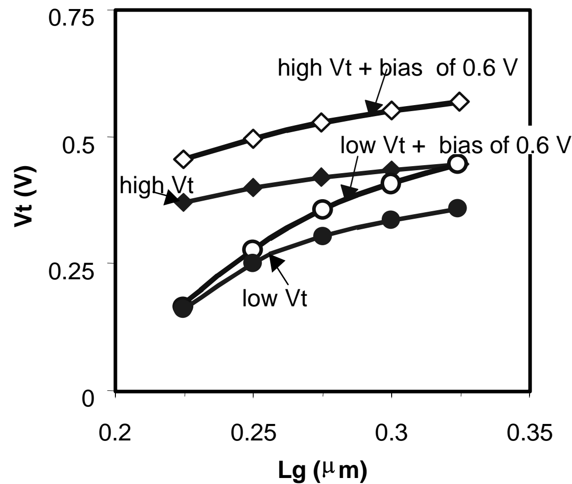


Figure 3-3: Increase in Vt-roll-off with *Vt* reduction and reverse body bias increase.

Furthermore, since *Vt* reduction degrades short channel effect, Vt-roll-off with channel length

reduction should be more for the lower-Vt device. In addition, reverse body bias will further increase the Vt-roll-off as shown in Figure 3-3.

It is known that increase in reverse body bias worsens MOSFET's short channel effect. Figure 3-4 shows sub-threshold characteristics of a 0.25 μm NMOS device. Using Drain Induced Barrier Lowering (DIBL) which is ΔVt observed for a given ΔVds, as another figure of merit to indicate short channel effect, we see that increasing reverse body bias (*Vsb*) from 0 V to 2 V increases ΔVt and hence DIBL, by 88%.

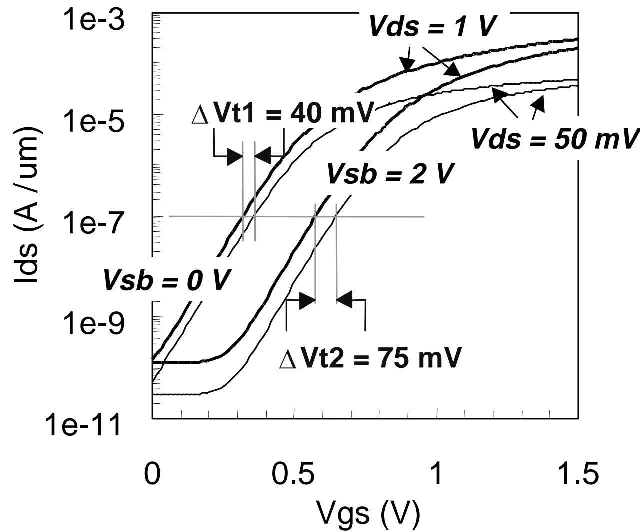


Figure 3-4: Increase in DIBL due to increase in reverse body bias.

3.1.2 Scaling of required body bias and SCE increase

Increase in Vt-roll-off due to adaptive body bias will lead in increase in within-die *Vt* variation. To quantify the impact of adaptive body bias on within-die *Vt* variation, we first determine the bias required to reduce die-to-die *Vt* variation, for two scaling scenarios, starting from a 0.25 μm technology as shown in Table 1. Once we determine bias required to reduce die-to-die *Vt* variation, we then determine, the SCE increase indicated by Drain Induced Barrier Lowering (DIBL) and the resulting increase in within-die *Vt* variation.

In Table 3-1, *Lg*, *Tox*, *Xj*, *Vdd*, and *Vt-linear* are gate length, oxide thickness, junction depth, supply voltage, and linear threshold voltage respectively. In both scaling scenarios *Lg*, *Tox*, *Xj*, and

Vdd scale by 30%. While *Vt-linear* scales by an aggressive 30% in the first scenario, it scales by a less aggressive 20% in the second. Equation (1) gives threshold voltage for a short channel NMOS by including body effect reduction factor, λ_b, from [30] and DIBL, λ_d [31]. Using (1) with $Vsb = 0$ and $Vds \rightarrow 0$, we can determine the channel doping *N*, for a given *Vt-linear*. The calculated values of *N* for the target devices are also listed in Table 3-1.

30% Vt scaling

Lg (um)	Elec. Tox(A)	Xj (um)	Vdd (V)	Vt-linear Reqd (V)	N (cm-3)
0.25	50	0.050	2.5	400e-3	5.99E+17
0.18	35	0.035	1.8	280e-3	7.37E+17
0.13	25	0.025	1.2	196e-3	9.26E+17

20% Vt scaling

Lg (um)	Elec. Tox(A)	Xj (um)	Vdd (V)	Vt-linear Reqd (V)	N (cm-3)
0.25	50	0.050	2.5	400e-3	5.99E+17
0.18	35	0.035	1.8	320e-3	8.52E+17
0.13	25	0.025	1.2	256e-3	1.21E+18

Table 3-1: Technology parameters under two scaling scenarios.

With channel doping known, we can determine DIBL, λ_d, using equation (2), which has been verified for accuracy down to Lg = 0.1 μm [31]. It is important to note that equation (2) is empirical and therefore its form cannot be explained using physical reasoning. With λ_d and *Vt-linear* known, we can now estimate *Vt-target*, the saturation threshold voltage for the target device.

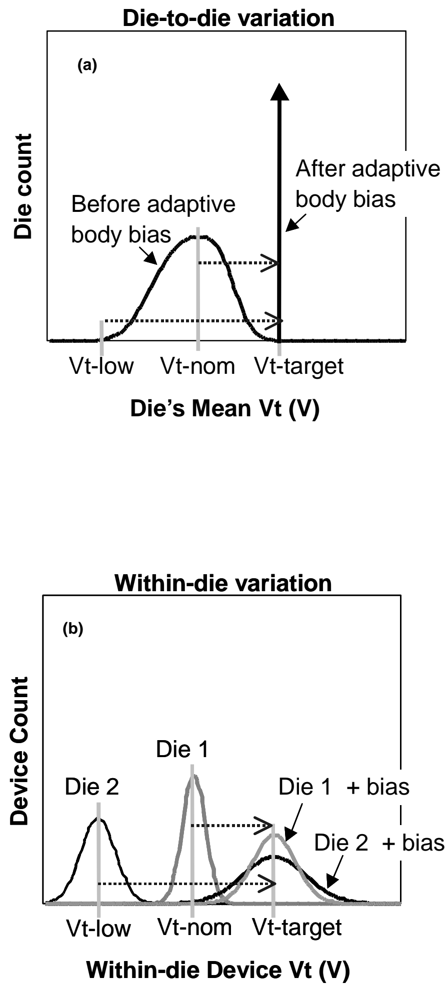


Figure 3-5: (a) Adaptive body bias reduces the die-to-die *Vt* variation. (b) Within-die *Vt* variation increases for die samples that require body bias to match their mean *Vt* to the target *Vt*. *Vt-target* is the target saturation threshold voltage for a given technology. *Vt-low* and *Vt-nom* are the minimum and mean threshold voltages of the die-to-die distribution.

$$\left.\begin{array}{l} V_t = V_{fb} + \left|2\phi_p\right| + \dfrac{\lambda_b}{C_{ox}}\sqrt{2qN\varepsilon_s(|2\phi_p| + V_{sb})} - \lambda_d V_{ds}; \quad \lambda_d \equiv DIBL \\ \lambda_b = 1 - \left(\sqrt{1 + \dfrac{2W}{X_j}} - 1\right)\dfrac{X_j}{L}; \quad W = \sqrt{\dfrac{2\varepsilon_s}{qN}(|2\phi_p| + V_{sb})}; \quad L = L_g - 2X_j \end{array}\right\}(1)$$

$$\left.\begin{array}{l} \lambda_d = \left[\dfrac{L}{2.2\mu m^{-2}(T_{ox} + 0.012\mu m)\,(W_{sd} + 0.15\mu m)\,(X_j + 2.9\mu m)}\right]^{-2.7} \\ W_{sd} = (W_s + W_d); W_s = \sqrt{\dfrac{2\varepsilon_s}{qN}(\phi_{bi} + V_{sb})}\,; W_d = \sqrt{\dfrac{2\varepsilon_s}{qN}(\phi_{bi} + V_{sb} + V_{ds})} \end{array}\right\}(2)$$

$$\left.\begin{array}{l} \dfrac{\Delta V_t}{\Delta L} = \dfrac{\partial V_t}{\partial \lambda_d}\,\dfrac{d\lambda_d}{dL} + \dfrac{\partial V_t}{\partial \lambda_b}\,\dfrac{d\lambda_b}{dL} \quad from\ (1) \\ \therefore \Delta V_t = \left[2.7V_{dd}\,\lambda_d + \dfrac{1}{C_{ox}}\sqrt{2qN\varepsilon_s(\left|2\phi_p\right| + V_{sb})}\,(1 - \lambda_b)\right]\dfrac{\Delta L}{L} \\ assume\, V_{ds} = V_{dd} \end{array}\right\}(3)$$

Let us now define *Vt-nom* and *Vt-low* to be the mean saturation threshold voltages of two different die samples as shown in Figure 3-5(b). *Vt-nom* is also the mean saturation threshold voltage of the die-to-die distribution as shown in Figure 3-5(a), and is due to 2.5% reduction in *Lg*, *Tox*, and *N*, and 2.5% increase in *Xj*, from the target device. Similarly, *Vt-min* is the minimum saturation threshold voltage of the die-to-die distribution, and is due to 5% change in *Lg*, *Tox*, *N*, and *Xj* from the target device. The values of *Vt-target*, *Vt-nom*, and *Vt-min,* before adaptive body bias are illustrated in Figure 3-6. Using equation (1) we can determine the body bias required to increase the saturation threshold voltage of the *Vt-nom* and *Vt-min* devices to *Vt-target*. The resulting saturation threshold voltages after adaptive body bias are depicted in Figure 3-7. The required bias values to match the saturation threshold voltages under the two scaling scenarios are given in Table 3-2 within parenthesis.

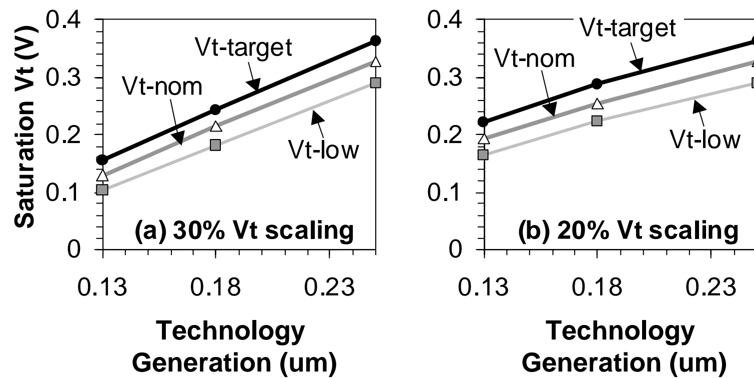


Figure 3-6: Trend in mean saturation threshold voltage of different die samples before adaptive body bias under (a) 30% *Vt* scaling and (b) 20% *Vt* scaling scenarios.

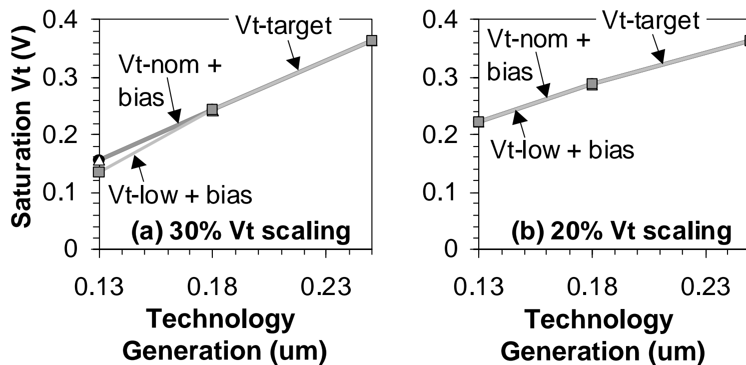


Figure 3-7: Matching of mean saturation threshold voltages of different die samples with adaptive body bias under (a) 30% *Vt* scaling and (b) 20% *Vt* scaling scenarios.

Comparing Figure 3-6 and Figure 3-7, it is clear that adaptive body bias will reduce die-to-die *Vt* variation. It is also clear from Table 3-2 that the bias required to match die-to-die *Vt* variation increases with scaling. Note from Figure 3-7 (a) that under 30% *Vt* scaling, adaptive body bias was unable to increase *Vt-low* (103 mV) to *Vt-target* (156 mV) for 0.13 μm technology due to body effect reduction with bias [30]. For body bias above 1.34 V the saturation threshold voltage of this device saturates at 134 mV.

DIBL increase and body effect factor reduction for the different devices with and without body bias can be estimated using equation (2), and the values are listed in Table 3-2. As expected, SCE (DIBL increase and body effect reduction) becomes worse with scaling and degrades further with body bias. In addition, the increase in SCE due to adaptive body bias escalates with technology scaling, since the amount of bias required for reducing die-to-die *Vt* variation increases.

30% Vt scaling

Lg (um)	λb, λd for Vt-target	λb, λd for Vt-nom	λb, λd for Vtnom with (bias)	λb, λd for Vt-low	λb, λd for Vtlow with (bias)
0.25	0.78, 15	0.76, 17	0.74, 18 (0.24)	0.74, 20	0.68, 25 (0.66)
0.18	0.74, 21	0.72, 24	0.69, 27 (0.31)	0.70, 29	0.59, 40 (1.13)
0.13	0.70, 32	0.68, 38	0.62, 44 (0.49)	*0.65, 44*	*0.52, 64 (1.34)*

20% Vt scaling

Lg (um)	λb, λd for Vt-target	λb, λd for Vt-nom	λb, λd for Vtnom with (bias)	λb, λd for Vt-low	λb, λd for Vtlow with (bias)
0.25	0.78, 15	0.76, 17	0.74, 18 (0.24)	0.74, 20	0.68, 25 (0.66)
0.18	0.75, 19	0.73, 23	0.71, 25 (0.28)	0.71, 27	0.63, 34 (0.84)
0.13	0.73, 28	0.71, 33	0.67, 36 (0.34)	0.68, 39	0.57, 53 (1.26)

Table 3-2: With adaptive body bias short channel effect of devices increase, indicated by DIBL (λ_d in mV/V) increase and body effect reduction factor (λ_b) decrease. This SCE increase is worse for *Vt-low* devices, compared to *Vt-nom* devices, as they require larger body bias to match *Vt-target*. The required bias values (in V) are indicated within parentheses.

3.1.3 Impact on within-die threshold voltage variation

If for a given technology within-die *Vt* variation is primarily due to variation in critical dimension, equation (3) shows that within-die *Vt* variation of a device depends on its DIBL (λ_d) and body effect reduction factor (λ_d). Hence, the increase in DIBL and decrease in body effect with

adaptive bias will be translated to increase in within-die *Vt* variation. In other words, the within-die *Vt* variation of a die sample whose mean saturation threshold voltage was made to align with *Vt-target* using body bias, will be worse than that of the die sample whose mean saturation voltage was *Vt-target* to begin with.

For example, for the 0.25 μm technology with 5% (12.5 nm) variation in within-die *Lg*, the die sample whose mean saturation threshold voltage was *Vt-target* to begin with, is estimated to have a within-die *Vt* variation of 8.2 mV. On the other hand, after adaptive body bias, the within-die *Vt* variation for the die sample with *Vt-low* (*Vt-nom*) as the mean threshold voltage is estimated to be 15.7 mV (11 mV). So, the saturation threshold voltage ranges for the *Vt-target*, *Vt-nom*, and *Vt-low* die samples will be 363 mV ± 8.2 mV, 363 mV ± 11 mV, and 363 mV ± 15.7 mV respectively.

If we assume that within-die variation in critical dimension is 5% of target *Lg* then the percentage variation in *Vt* can be calculated using equation (3) and is illustrated in Figure 3-8. Clearly, with scaling within-die *Vt* variation due to adaptive body bias increases and is more pronounced for aggressive *Vt* scaling. This increase in within-die *Vt* variation can impact clock skew, worst-case gate delay, worst-case device leakage current, total chip leakage power, and analog circuit performance.

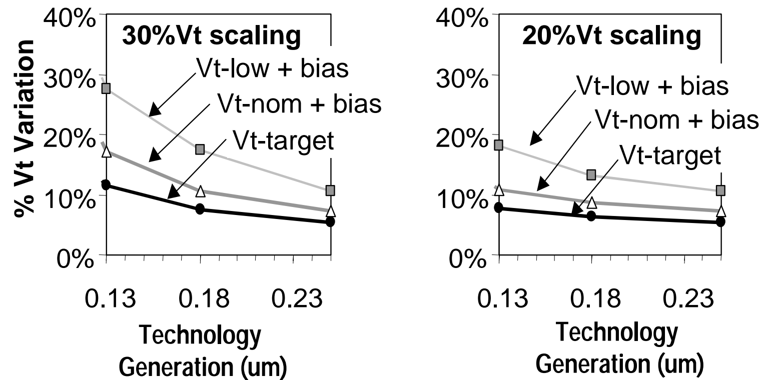


Figure 3-8: Increase in within-die threshold voltage variation due to increase in short channel effect with adaptive body bias under (a) 30% *Vt* scaling and (b) 20% *Vt* scaling. We assume that the dominant reason for within-die *Vt* variation is critical dimension variation. The results shown here assume within-die variation in *Lg* of 5%.

3.1.4 Summary

We showed that although adaptive body bias reduces die-to-die *Vt* variation it increases within-die *Vt* variation, due to increase in short channel effect. Moreover, we quantified this increase under two *Vt* scaling scenarios. The analysis showed that the increase in within-die *Vt* variation due to adaptive bias worsens with scaling and is more pronounced for aggressive *Vt* scaling. Consequently, to make effective use of the traditional adaptive body bias scheme one should consider (a) the maximum acceptable within-die *Vt* variation increase that can be tolerated for a given design and (b) the use of multiple adaptive bias generators within-die on a triple well process. Even if these techniques are employed to minimize impact of adaptive body bias on with-die *Vt* variation, adaptive body bias is still destined to become less effective with scaling due to increased SCE and weakening body effect. In addition, circuits that cannot tolerate increase in short channel effect due to reverse body bias should be isolated not to receive body bias. This will require triple-well process if adaptive body bias needs to applied for both PMOS and NMOS devices.

In the next section, a scheme called bi-directional adaptive body bias is introduced. This scheme does not require process modification for *Vt* re-targeting, minimizes die-to-die *Vt* variation without impacting *Vt* within-die variation, and more importantly, its effectiveness scales better with technology compared to the traditional adaptive body bias. The bi-directional adaptive body bias scheme discussed in the next section is designed to minimize the variation in microprocessor operating frequency due to within-die and die-to-die *Vt* variations. The testchip was designed in collaboration with James Kao (MIT Ph.D. 2001). My contributions were to (i) study the impact that within-die variation plays on the microprocessor frequency distribution and (ii) determine the proper bias circuit impedance required to ensure minimal impact of noise on the stability of the bias value. The details of the testchip and measurement results are discussed in Section 3.2 and the bias circuit impedance requirement and measurement results are discussed in Section 3.3.

3.2 Bi-directional adaptive body bias

Both die-to-die and within-die *Vt* variations, which are becoming worse with technology scaling, impact clock frequency and leakage power distributions of microprocessors in volume manufacturing [32]. In particular, they limit the percentage of processors that satisfy both minimum frequency requirement and maximum active switching and leakage power constraints. Their

impacts are more pronounced at the low supply voltages used in processors for mobile systems where the active power budget is limited by constraints imposed by heat removal, power delivery and battery life considerations.

In bi-directional adaptive body bias the mean *Vt* of all die samples are matched to the target *Vt* by applying both forward and reverse body bias. Forward body bias is applied to die samples that are slower than the target and reverse body bias is applied to die samples that are faster than the target, as shown in Figure 3-9. It is important to note that while forward bias reduces Vt it also increases the junction current. Hence, there is a maximum forward bias beyond which the junction current increase will inhibit proper operation of CMOS circuits. It has been determined that at a temperature of 110ºC the maximum amount of forward bias that can be applied is 450 mV. This increases to 750 mV at an operating temperature of 30ºC [33]. Since both *Vt* reduction and increase are possible, process re-targeting to reduce *Vt* is not required. By avoiding process re-targeting increase in within-die *Vt* variation due increase in SCE for lower *Vt* transistors is prevented. In addition, the die samples that forward body bias since it reduces the diode depletion improves SCE and hence reduces within-die *Vt* variation and maximum reverse body bias required under bi-directional adaptive body bias clearly would be smaller. So, this scheme will always scale better than the traditional adaptive body bias. This technique was first reported in [23] as a follow-up to [21] and [22]. In rest of this section, improvements over [23] will be presented.

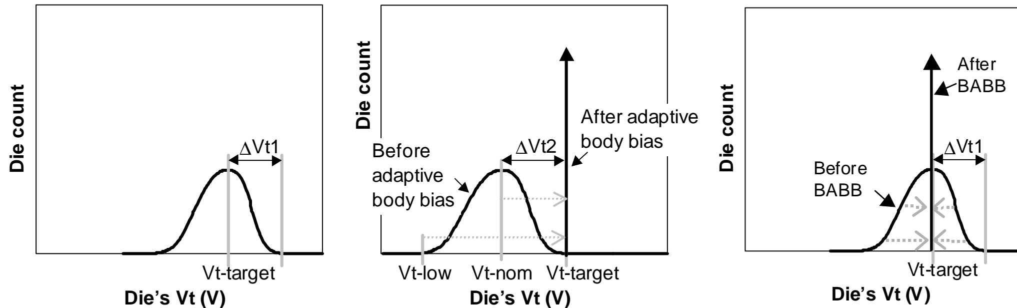


Figure 3-9: Die-to-die threshold voltage distributions (a) Conventional approach without adaptive body bias (b) traditional adaptive body bias approach – die sample that requires maximum reverse body bias is 2ΔVt2 away from Vt-target (c) bi-directional adaptive body bias approach – die sample that requires maximum reverse body bias is ΔVt1 away from Vt-target. Note: ΔVt2 > ΔVt1 since SCE of devices with lower *Vt* will be more.

A testchip (Figure 3-10) has been implemented in a 150 nm CMOS technology to evaluate effectiveness of the bi-directional adaptive body bias technique for minimizing impacts of both die-to-die and within-die *Vt* variations on processor frequency and active leakage power [34]. The testchip contains 21 "sub-sites" distributed over a 4.5 x 6.7 mm^2 area in two orthogonal orientations. Each sub-site has (i) a circuit block (CUT) containing key circuit elements of a microprocessor critical path, (ii) a replica of the critical path whose delay is compared against an externally applied target clock frequency (ϕ) by a phase detector, (iii) a counter which updates a 5-bit digital code based on the phase detector output, and (iv) a "resistor-ladder D/A converter + op-amp driver" which, based on the digital code, provides one of 32 different body bias values to PMOS transistors in both the CUT and the critical path delay element. The circuit block diagram of each sub-site is shown in Figure 3-11. N-well resistors are used for the D/A converter implementation. For a specific externally applied NMOS body bias, this on-chip circuitry automatically generates the PMOS body bias that minimizes leakage power of the CUT while meeting a target clock frequency, as demonstrated by measurements in. Different ranges of unidirectional – forward (FBB) or reverse (RBB) – or bi-directional body bias values (Figure 3-12) can be selected by using appropriate values of V_{REF} and V_{CCA}, and by setting a counter control bit. Adaptive body biasing can also be accomplished by using the phase detector output (PD) to continually adjust off-chip bias generators through software control, instead of using the on-chip circuitry, until the frequency target is met.

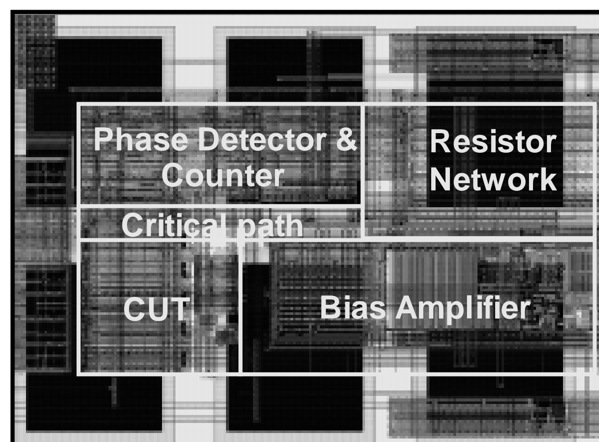


Figure 3-10: Chip micrograph of a sub-site.

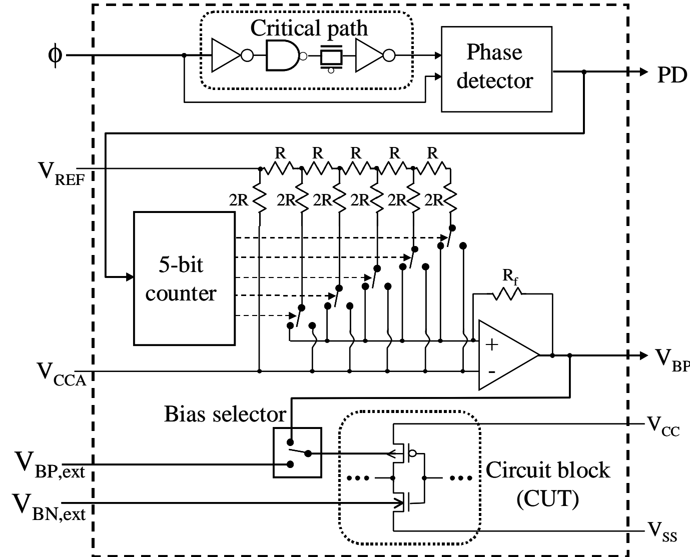
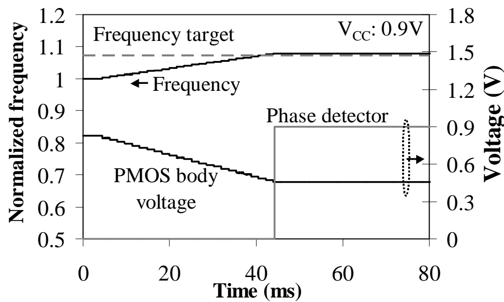


Figure 3-11: Circuit block diagram of each sub-site.

Clock frequency, switching power and active leakage power of the 21 CUT's per die are measured independently at 0.9V V_{CC} and 110C, for 62 dies on a wafer. Die clock frequency is the minimum of the CUT frequencies, and active leakage power is sum of the CUT leakages. When no body bias (NBB) is used, 50% of the dies meet both the minimum frequency requirement and the maximum active leakage constraint set by a total power density limit of 20 W/cm^2 (Figure 3-13). Using 0.2V forward body bias (FBB) allows all of the dies to meet the minimum frequency requirement, but most of them fail to satisfy the leakage constraint. As a result, only 20% of the dies are acceptable even though variations are reduced slightly by FBB due to improved short-channel effects [23].



Bias Mode	Condition	Range
NBB → FBB	$V_{CCA} = V_{CC}$ $V_{REF} > V_{CCA}$	FBB: 0 → V_{REF}-V_{CCA}
NBB → RBB	$V_{CCA} = V_{CC}$ $V_{REF} < V_{CCA}$	RBB: 0 → V_{CCA}-V_{REF}
FBB → RBB	$V_{CCA} < V_{CC}$ $V_{REF} < V_{CCA}$	FBB: V_{CC}-V_{CCA} → RBB: $2V_{CCA}$-V_{REF}-V_{CC}

Figure 3-12: Demonstration of frequency adapting to meet target and list of possible on-chip bias modes.

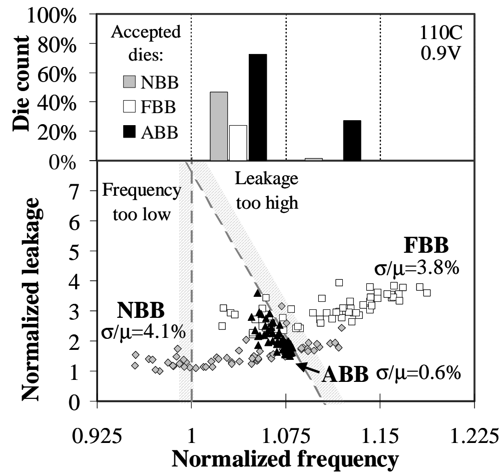


Figure 3-13: Die-to-die variation in frequency and leakage for no body bias (NBB), 0.2 V static forward body bias (FBB), and adaptive body bias applied to compensate die-to-die variation (ABB).

Bi-directional ABB is used for both NMOS and PMOS devices to increase the percentage of dies that meet both frequency requirement and leakage constraint. For each die, we use a single combination of NMOS and PMOS body bias values that maximize clock frequency without violating the active leakage power limit. As a result, die-to-die frequency variations (σ/μ) reduce by an order of magnitude, and 100% of the dies become acceptable (Figure 3-13). In addition, 30% of the dies are now in the highest frequency bin allowed by the power density limit when leakage is negligible.

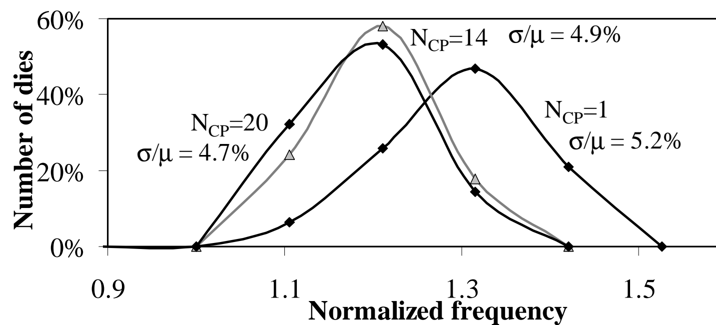


Figure 3-14: Frequency vs. number of critical paths that determine the frequency.

In a simpler ABB scheme, within-die variations can be neglected [23] and the required body bias for a die can be determined from measurements on a single CUT per die. However, testchip measurements in Figure 3-14 show that as the number of critical paths (N_{CP}) on a die increases, WID delay variations among critical paths cause both μ and σ of the die frequency distribution to become smaller. This is consistent with statistical simulation results [32] indicating that the impact of WID parameter variations on die frequency distribution is significant. As N_{CP} exceeds 14, there is no change in the frequency distribution with N_{CP}. Therefore, using measurements of 21 critical paths on the testchip to determine die frequency is sufficiently accurate for obtaining frequency distributions of microprocessors, which contain 100's of critical paths. Previous measurements [23] on 49-stage ring oscillators showed that σ of the WID frequency distribution is 4X smaller than σ of the device saturation current (I_{ON}) distribution. However, measurements on the testchip containing 16-stage critical paths (Figure 3-15) show that σ's of WID critical path delay distributions and NMOS/PMOS I_{ON} distributions are comparable. Since typical microprocessor critical paths contain 10-15 stages, and this number is reducing by 25% per generation [4], impact of within-die variations on frequency is becoming more pronounced. This is further evidenced by the fact that the number of acceptable dies reduces from 100% to 50% in the simpler ABB scheme which neglects within-die variations, although die count in the highest frequency bin increases from 0% to 11% when compared with NBB.

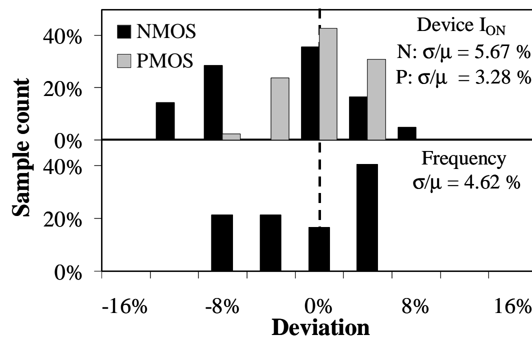


Figure 3-15: Comparison of variations in within-die device current and frequency.

The ABB scheme, which compensates primarily for die-to-die parameter variations by using a single NMOS/PMOS bias combination per die, can be further improved to compensate for WID variations as well. In this WID-ABB scheme, different NMOS/PMOS body bias combinations are

used for different circuit blocks on the die. A triple-well process is needed for NMOS implementation. For each CUT, the NMOS body bias is varied over a wide range using an off-chip bias generator. For each NMOS bias, the on-chip circuitry determines the PMOS bias that minimizes leakage power of the CUT while meeting a particular target frequency. The optimal NMOS/PMOS bias for the CUT at a specific clock frequency is then selected from these different bias combinations as the one that minimizes CUT leakage. This produces a distribution of optimal NMOS/PMOS body bias combinations for the CUT's on a die at a specific clock frequency. If the die leakage power exceeds the limit at that frequency, the target frequency is reduced and the process is repeated until we find the maximum frequency where the leakage constraint is also met.

WID-ABB reduces σ of the die frequency distribution by 50%, compared to ABB (Figure 3-16). In addition, virtually 100% of the dies are accepted in the highest possible frequency bin, compared to 30% for ABB. Distribution of optimal NMOS/PMOS body bias combinations (Fig. 6) for a sample die in the WID-ABB scheme reveals that while RBB is needed for both PMOS and NMOS devices, FBB is used mainly for the PMOS devices. In addition, body bias values in the range of 0.5V RBB to 0.5V FBB are adequate. Finally, measurements (Figure 3-17) show that ABB and WID-ABB schemes need at least 300mV and 100mV body bias resolutions, respectively, to be effective. The 32mV bias resolution provided by the on-chip circuitry in the testchip is, therefore, sufficient for both ABB and WID-ABB.

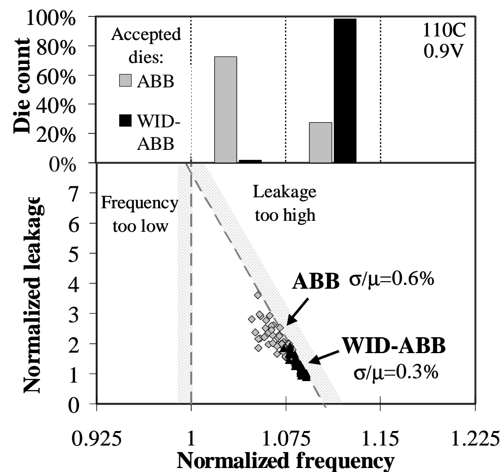
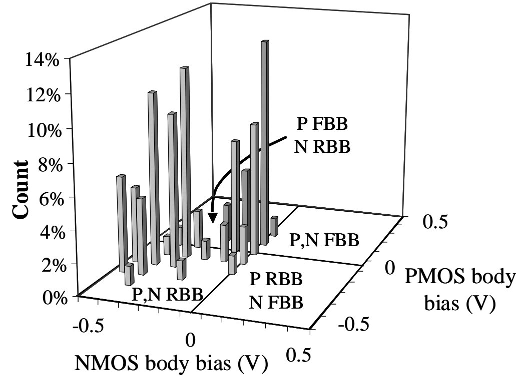


Figure 3-16: Die-to-die variation in frequency and leakage for adaptive body bias applied to (i) compensate die-to-die variation (ABB) and (ii) compensate within-die variation (WID-ABB).



Bias resolution	Die-to-die ABB		Within-die ABB	
	dies, F > 1	σ/μ	dies, F > 1.075	σ/μ
0.5	79 %	2.87 %	2 %	1.89 %
0.3	100 %	1.47 %	66 %	0.50 %
0.1	100 %	0.58 %	97 %	0.25 %

Figure 3-17: Histogram of bias voltages within a die sample and effect of bias resolution on frequency distribution.

3.3 Body bias circuit impedance requirement

Since adaptive body bias circuit technique require on-chip biasing, it is important to determine impedance requirement for the on-chip bias voltage generator circuit. In this section a method to determine proper bias circuit impedance and sample bias circuits are described. To verify the design of the bias circuit a 6.6 million transistors communications router chip [3536, 37], with on-chip circuitry to provide forward body bias (FBB) [38] during active operation and zero body bias (ZBB) during standby mode, has been implemented in a 150nm CMOS technology (Figure 3-18). FBB is applied during active mode and it is withdrawn during standby mode to reduce leakage power. Power and performance of the chip are compared with the original design that has no body bias (NBB). The FBB and NBB router chips reside adjacent to each other on the same reticle to allow accurate comparisons by measurement. If the on-chip bias circuit has proper impedance then (i) FBB chip in FBB mode should increase the frequency of operation at a given supply voltage (Vcc) (ii) FBB chip in ZBB mode should have lower standby leakage and (iii) FBB chip with ZBB should have the same frequency of operation as that of the NBB chip on the same reticle.

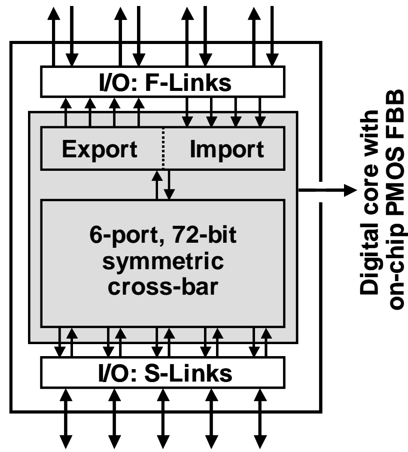


Figure 3-18: Communications router chip architecture with PMOS body bias.

In the FBB testchip, body bias is used for the PMOS devices in the digital core of the chip. Total biased PMOS transistor width is 2.2 meters. Body bias generator circuits and bias distribution across the chip have been optimized to minimize area overhead, and to provide a constant 450mV FBB with sufficient robustness against various noises, as well as variations in process, Vcc and temperature (PVT).

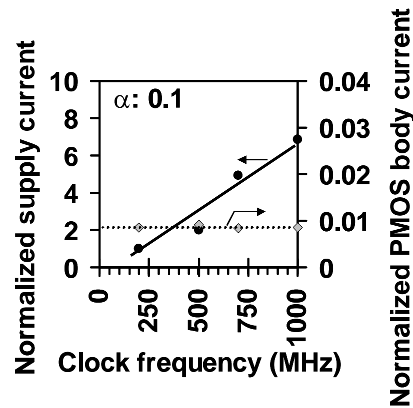


Figure 3-19: Measurement of body and Vcc current

Testchip measurements (Figure 3-19) show that current in the body grid is at least two orders of magnitude smaller than the Vcc current across a range of operating frequencies. Therefore,

overhead of body bias routing is minimal compared to the Vcc grid. Distributed bias generator architecture has been implemented to minimize variation of the body-to-source voltage (Vbs) due to *global* coupling and Vcc noises (Figure 3-20). A central bias generator (CBG) uses a scaled bandgap circuit [39] to generate a PVT-insensitive 450mV voltage with reference to Vcca. This reference voltage is routed to 24 local bias generators (LBG), distributed around the digital core of the chip. Global routing of this 450mV differential reference voltage uses Vcca tracks on both sides for proper shielding and adequate common-mode noise rejection. Each LBG has a reference translation circuit that converts the Vcca-450mV reference voltage to a voltage 450mV below the local Vcc. This voltage is driven by a buffer stage and routed locally to the PMOS devices in the core to provide 450mV FBB during active operation. Local body bias routing tracks are placed adjacent to the local Vcc tracks to improve common-mode noise rejection, and thus reduce noise-induced variations in the target 450mV Vbs in the biased PMOS devices. The voltage buffer and the local decoupling capacitor at the buffer output have been designed to minimize Vbs variations induced by *local* coupling and Vcc noises, with a small area and power overhead. Full-chip area overhead of the biasing circuitry is 2% and power overhead is 1%.

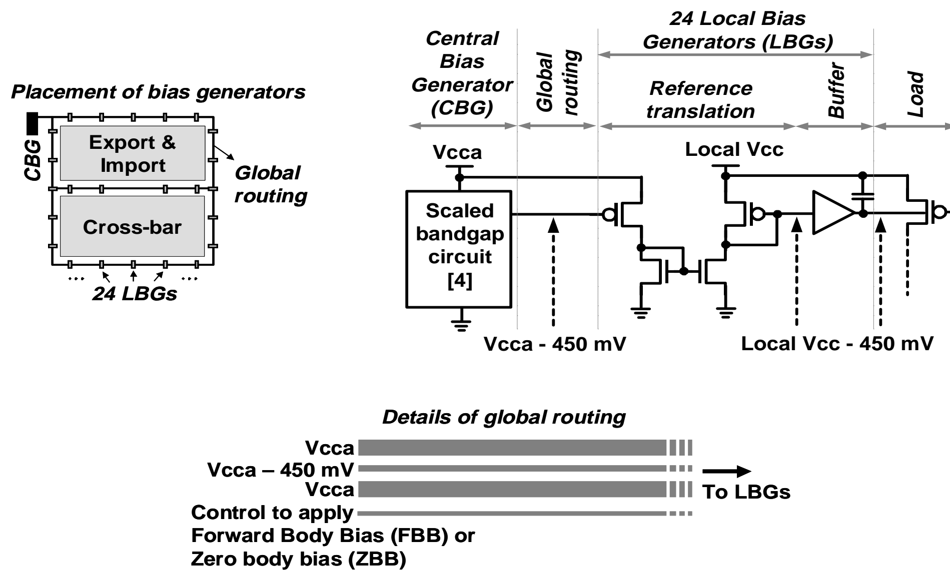
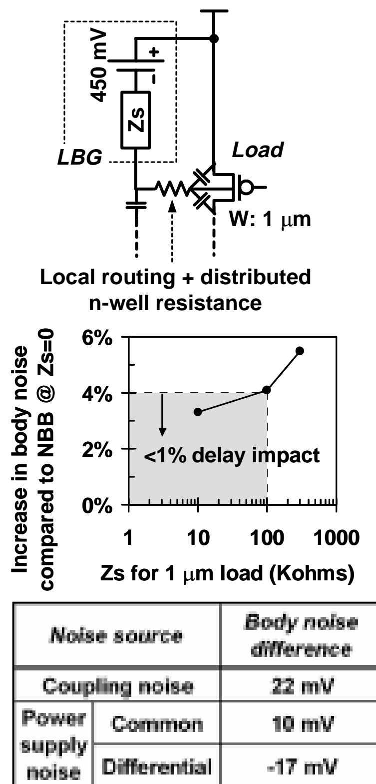


Figure 3-20: Overview of body bias generation and distribution.

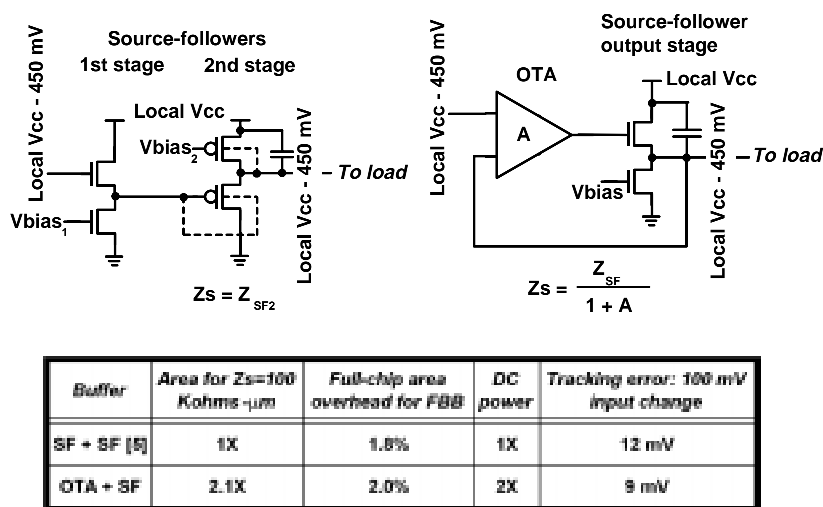
Three different sources of noise can induce variations in the target Vbs value. First, coupling to the body node from logic circuit output transitions can change Vbs of a victim transistor during switching. This noise is transmitted to the victim through the bias grid and the n-well. Circuit simulations in a 150nm technology, with a two-dimensional distributed RC model for the n-well, show that the width of this noise pulse is several hundred pico-seconds for 1.5-2KΩ/sq n-well sheet resistance. Therefore, this noise impacts switching delay of the victim circuit. However, since different circuits switch in opposite directions at the same time in a large logic design, a small fraction (<10%) of the total transistor width accounts for the simultaneous unidirectional switching that couples noise to the body. Second, Vcc noise common to both the Vbs generator and a biased logic circuit can causes Vbs to vary (common Vcc noise). Finally, any difference in Vcc values of the Vbs generator and the biased logic circuit induces variation in Vbs (differential Vcc noise).



Noise source		Body noise difference
Coupling noise		22 mV
Power supply noise	Common	10 mV
	Differential	-17 mV

Figure 3-21: Buffer impedance requirements and body bias noise comparisons with NBB.

The voltage buffer and the decoupling capacitor in the LBG have been designed to provide worst-case output impedance (Zs) of 100KΩ per μm of *effective* (simultaneous unidirectional switching) biased PMOS width. Simulations (Figure 3-21) for this design show that the total Vbs variation induced by all three noises increases by 4% from the NBB design, where the body is tied locally to Vcc. The resulting impact on circuit delay is 1%. Vbs variations due to coupling and 10% common Vcc noise increase by 10-20mV, whereas that due to 10% differential Vcc noise reduces by 17mV. Since n-well sheet resistance is relatively high in logic technologies, and since the maximum distance allowed between n-well taps are several tens of microns, significant deviations are observed (by simulations) in the zero bias value for NBB designs.



Buffer	Area for Zs=100 Kohms -μm	Full-chip area overhead for FBB	DC power	Tracking error: 100 mV input change
SF + SF [5]	1X	1.8%	1X	12 mV
OTA + SF	2.1X	2.0%	2X	9 mV

Figure 3-22: LBG buffer implementations and comparisons.

Figure 3-22 shows implementations of two different LBG buffers – the “SF+SF” on a 5GHz 32-bit integer execution core chip [40] in a 130nm dual-Vt CMOS technology, and the “OTA+SF” on this chip. In the SF+SF implementation, the overall impedance is determined by the output impedance Z_{SF2} (~$1/g_m$) of the second stage, while the first stage is designed to meet bandwidth requirements. FBB, already available in the LBG, is used for the PMOS devices in the output SF stage to improve g_m by 30%, thus reducing Zs for the same area. The OTA+SF implementation uses a high-gain OTA and an output SF stage. The overall impedance is determined by the output impedance (Z_{SF}) of the SF stage and the voltage gain (A) of the OTA. The design is optimized to

obtain impedance less than the target value up to 10GHz frequency, while minimizing area. In this optimization, the gain and corresponding bandwidth of the OTA are traded-off against the amount of decoupling capacitance needed at the buffer output. Comparisons of the two implementations (Figure 3-22) show that full-chip area overheads are about the same for both. OTA+SF consumes double the area and power, while providing better accuracy in input voltage tracking. The OTA+SF design was used in the communications router chip as well as the adaptive body bias chip described in Section 3.2.

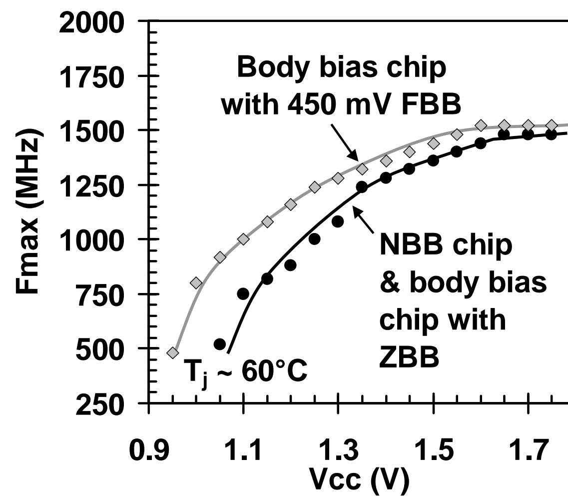


Figure 3-23: Frequency vs. Vcc of FBB and NBB chips.

Maximum frequency (Fmax) of the NBB and FBB router chips are compared from 0.9V to 1.8V Vcc at 60C (Figure 3-23). Fmax values are measured by sending data in through the F-link input port and verifying data at the F-link output port after the data has passed through the import, crossbar and export units. The FBB chip with forward body bias achieves 1GHz operation at 1.1V, compared to 1.25V required for the NBB chip and FBB chip with ZBB. As a result, switching power is 23% smaller at 1GHz. The frequency of the FBB chip is 33% higher than the NBB chip at 1.1V. The frequency improvement is more pronounced as Vcc is further reduced. Also, there is no observable performance impact of potentially larger Vcc noise in the FBB design due to the absence of n-well to substrate junction capacitance for local Vcc decoupling. Chip leakage currents are measured for 74 dies on a wafer with FBB and ZBB. Leakage current during active mode is set

by FBB, which is withdrawn in standby mode to reduce leakage. Histogram (Figure 16.4.6) of active-to-standby leakage ratio – I(FBB)/I(ZBB) – shows 2X to 8X leakage reduction, with an average reduction of 3.5X. Clearly, this leakage reduction capability will not available in the NBB chip, if the performance is improved by lowering *Vt* in the process technology. The die micrograph and chip characteristics are shown in Figure 3-25.

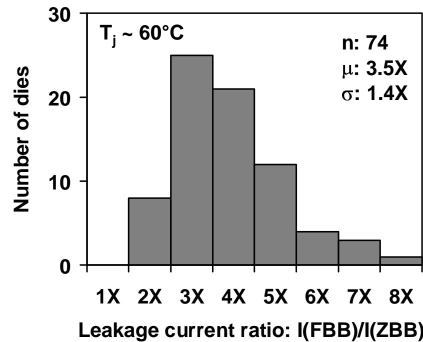
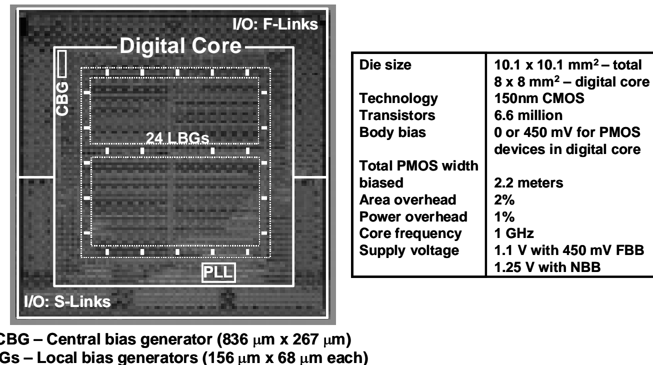


Figure 3-24: Leakage reduction from active to standby mode in FBB chips.



Die size	10.1 x 10.1 mm² – total 8 x 8 mm² – digital core
Technology	150nm CMOS
Transistors	6.6 million
Body bias	0 or 450 mV for PMOS devices in digital core
Total PMOS width biased	2.2 meters
Area overhead	2%
Power overhead	1%
Core frequency	1 GHz
Supply voltage	1.1 V with 450 mV FBB 1.25 V with NBB

Figure 3-25: Micrograph of communications router chip with PMOS body bias and of chip characteristics.

It was shown that (i) the FBB chip in FBB mode operates at 23% higher frequency at 1.1 V, (ii) the FBB chip in ZBB mode on an average has 3.5X lower standby leakage and (iii) the FBB chip with ZBB has virtually the same frequency of operation as that of the NBB chip at any given Vcc. This proves that the method of targeting the bias circuit impedance to be 100KΩ per μm of *effective* (simultaneous unidirectional switching) biased MOS width is a successful rule-of-thumb.

Chapter 4
Within-die Threshold Voltage Variation and Leakage Power

4.1 Estimation of chip leakage current

It has been established that to limit the energy and power increase in future CMOS technology generations, the supply voltage (V_{dd}) will have to continually scale. The amount of energy reduction depends on the magnitude of V_{dd} scaling. Along with V_{dd} scaling, the threshold voltage (V_t) of MOS devices will have to scale to sustain the traditional 30% gate delay reduction. These V_{dd} and V_t scaling requirements pose several technology and circuit design challenges. One such challenge is the rapid increase in sub-threshold leakage power due to V_t scaling. Should the present scaling trend continue it is expected that the sub-threshold leakage power will become as much as 50% of the total power in the 0.09 μm generation as shown in Figure 2-5. Under this scenario, it is important to be able to predict sub-threshold leakage power more accurately. Present leakage current estimation techniques do not take into account the variation in within-die threshold voltage. It will be shown that this assumption leads to significant inaccuracies. A mathematical model for chip leakage current that considers within-die threshold voltage variation will be derived. Microprocessor measurements that verify the improvement in leakage estimation with the new model are also presented. In rest of the chapter, the term leakage refers to sub-threshold leakage.

4.1.1 Present leakage current estimation techniques

Due to the wide variation expected threshold voltage of MOS devices from die-to-die and within-die during the life time of a process, present leakage current estimation techniques provide lower and upper bounds on the leakage current. The upper and lower bounds are at least an order of

magnitude apart and leakage power of most chips lies between the two bounds as shown in [41]. In older technology generations, basing system design on the two leakage current bounds was acceptable since leakage power was a negligible component of the total power. In most systems, the worst case bound is assumed for the design. In future technology generations where as much as half of the system power during active mode can be due to leakage, depending the worse case bound will lead to extremely pessimistic and expensive design solutions. One cannot base the system design on the lower bound since it will lead to overly optimistic and unreliable design solutions. Therefore, it will be crucial to estimate leakage current as accurately as possible. The upper and lower bound estimate equations and measurements are provided in the next part of this section. The lower bound leakage current estimation of a chip is given as follows,

$$I_{leak\text{-}l} = \frac{w_p}{m_p} I_p^o + \frac{w_n}{m_n} I_n^o$$

where, w_p and w_n are the total PMOS and NMOS device widths in the chip; m_p and m_n are factors that determine percentage of PMOS and NMOS device widths that are in off state; I^o_p and I^o_n are the expected mean leakage currents per unit width of PMOS and NMOS devices in a particular chip. The mean leakage current is obtained for devices with mean threshold voltage or channel length. The upper bound leakage current estimation of a chip is related to the device leakage as follows,

$$I_{leak\text{-}u} = \frac{w_p}{m_p} I^{3\sigma}_{off-p} + \frac{w_n}{m_n} I^{3\sigma}_{off-n}$$

where, $I^{3\sigma}_{off\text{-}p}$ and $I^{3\sigma}_{off\text{-}n}$ are the worst-case leakage current per unit width of PMOS and NMOS devices. The worst-case leakage current is obtained for devices with threshold voltage or channel length 3σ lower than the mean leakage currents per unit width of PMOS and NMOS devices in a particular chip.

4.1.2 Leakage current estimation including within-die variation

To include the impact of within-die threshold voltage or channel length variation it is necessary to consider the entire range of leakage currents, not just the mean leakage or the worst-case leakage. Let us assume that the within-die threshold voltage or channel length variation follows a normal distribution with respect to transistor width, with μ being the mean and σ being the sigma of the distribution. Let I^o be the leakage of the device with the mean threshold voltage or channel

length. Then by performing the weighted sum of devices of different leakage, we can estimate the total leakage of the chip. This is achieved by integrating the threshold voltage or channel length distribution multiplied by the leakage, as shown below.

$$I_{leak} = \frac{I^o w}{m} \frac{1}{\sigma\sqrt{2\pi}} \int_{x\,min}^{x\,max} e^{\frac{-(x-\mu)^2}{2\sigma^2}} e^{\frac{(\mu-x)}{a}} dx$$

In the above equation, the first exponent estimates the fraction of the total width for the device leakage estimated by the second exponent. If the distribution considered within-die is threshold voltage variation then x in the above equation represents threshold voltage and a will be equal to $n\phi_t$. If the distribution considered is channel length then x in the above equation will represent channel length and a will be equal to λ. λ can be estimated for a technology by measuring the relationship between channel length and device leakage. In the rest of this section, we will assume that the distribution of interest is the channel length, since this parameter is used to characterize a technology. The derivation of the chip leakage is then given as follows,

$$I_{leak} = \frac{I^o w}{m} \frac{1}{\sigma\sqrt{2\pi}} \int_{l\,min}^{l\,max} e^{\frac{-(l-\mu)^2}{2\sigma^2}} e^{\frac{(\mu-l)}{\lambda}} dl$$

$$= \frac{I^o w}{m} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\sigma^2}{2\lambda^2}} \int_{l\,min}^{l\,max} e^{\frac{-(l-\mu)^2}{2\sigma^2}} e^{\frac{(\mu-l)}{\lambda}} e^{\frac{-\sigma^2}{2\lambda^2}} dl$$

$$= \frac{I^o w}{m} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\sigma^2}{2\lambda^2}} \int_{l\,min}^{l\,max} e^{-\left[\frac{l-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}\right]^2} dl$$

Let,

$$t = \left[\frac{l-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}\right] \Rightarrow dl = \sqrt{2}\sigma dt$$

$$\therefore I_{\mathrm{leak}} = \frac{I^o w}{m} \frac{1}{\sqrt{\pi}} e^{\frac{\sigma^2}{2\lambda^2}} \int_{\frac{l\,min-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}}^{\frac{l\,max-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}} e^{-[t]^2} dt$$

The integral can be rewritten as,

$$I_{\text{leak}} = \frac{I^o w}{2m} \, e^{\frac{\sigma^2}{2\lambda^2}} \left[\frac{2}{\sqrt{\pi}} \int_o^{\frac{l\,max-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}} e^{-t^2} \, dt \; - \; \frac{2}{\sqrt{\pi}} \int_o^{\frac{l\,min-\mu}{\sqrt{2}\sigma}+\frac{\sigma}{\sqrt{2}\lambda}} e^{-t^2} \, dt \right]$$

$$= \frac{I^o w}{2m} \, e^{\frac{\sigma^2}{2\lambda^2}} \left[erf\left(\frac{l\,max-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda} \right) - erf\left(\frac{l\,min-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda} \right) \right] \quad \because erf(z) = \frac{2}{\sqrt{\pi}} \int_o^z e^{-t^2} \, dt$$

$$= \frac{I^o w}{2m} \, e^{\frac{\sigma^2}{2\lambda^2}} \left[erf\left(\frac{l\,max-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda} \right) + erf\left(\frac{\mu - l\,min}{\sqrt{2}\sigma} - \frac{\sigma}{\sqrt{2}\lambda} \right) \right] \quad \because erf(-z) = -erf(z)$$

Since,

$$erf(z) \rightarrow 1 \quad \text{if} \quad z > 1 \quad \text{and} \quad \frac{l\,max-\mu}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}\lambda} \;,\; \frac{\mu - l\,min}{\sqrt{2}\sigma} - \frac{\sigma}{\sqrt{2}\lambda} >> 1$$

$$\Rightarrow I_{leak} = \frac{I^o w}{m} \, e^{\frac{\sigma^2}{2\lambda^2}}$$

Using the above result we can now estimate the leakage of a chip that has both PMOS and NMOS devices including within-die variation as follows,

$$I_{leak-w} = \frac{I_p^o w_p}{m_p} \, e^{\frac{\sigma_p{}^2}{2\lambda_p{}^2}} \; + \; \frac{I_n^o w_n}{m_n} \, e^{\frac{\sigma_n{}^2}{2\lambda_n{}^2}}$$

where, w_p and w_n are the total PMOS and NMOS device widths in the chip; m_p and m_n are factors that determine percentage of PMOS and NMOS device widths that are in off state; $I^o{}_p$ and $I^o{}_n$ are the expected mean leakage currents per unit width of PMOS and NMOS devices in a particular chip; σ_p and σ_n are the standard deviation of channel length variation within a particular chip; λ_p and λ_n are constants that relate channel length of PMOS and NMOS devices to their corresponding sub-threshold leakages. It is also worth pointing out that from the formula for I_{leak}, if I_{leak} can be measured for a chip, a macroscopic standard deviation (σ) representing parameter variation in that chip can be determined as,

$$\sigma = \lambda \sqrt{2\,ln\left(\frac{m}{w} \frac{I_{leak}}{I^o} \right)}$$

4.1.3 Measurement results

Leakage power measurements on several samples of a 0.18-μm 32-bit microprocessor were carried out. The current and effective channel length measurements on test devices that accompany each microprocessor were measured to determine I^o_p, I^o_n, λ_p, and λ_n. σ_p and σ_n were assumed as a constant percentage of the measured channel length in the test device of each sample. Using these individual device measurements, with w_p and w_n obtained from the design the leakage power was calculated using the $I_{leak\text{-}l}$, $I_{leak\text{-}u}$, and $I_{leak\text{-}w}$ formulae. In addition, we assumed that on an average half of the devices will be in off state, that is, $m_p = m_n = 2$. The three calculated leakages are then compared with the measured leakage.

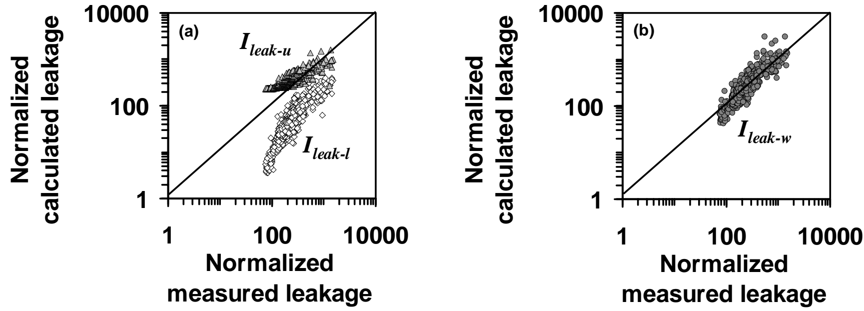


Figure 4-1: Comparison of calculated leakage versus measured leakage for (a) existing leakage current estimation techniques and (b) leakage current estimation technique introduced in this thesis.

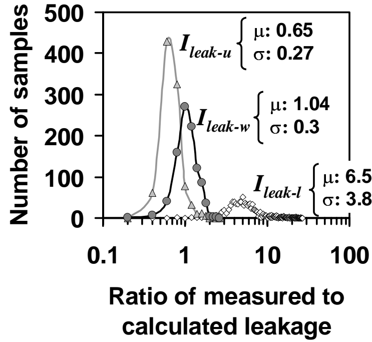


Figure 4-2: Ratio of measured to calculated leakage current ratio distribution for $I_{leak\text{-}u}$, $I_{leak\text{-}l}$, and $I_{leak\text{-}w}$ techniques (Sample size: 960).

Figure 4-1(a) clearly illustrates that the upper bound technique overestimates the leakage current of the chips while the lower bound techniques underestimates the leakage current. However, the estimation technique introduced in this thesis that includes within-die variation

matches the measurement better, as illustrated in Figure 4-1(b). Data shown in Figure 4-1 is summarized in Figure 4-2. As the figure indicates the leakage power for most of the samples are underestimated by 6.5X if the lower bound technique is used and overestimated by 1.5X if the upper bound technique is used. The measured-to-calculated leakage ratio for majority of the device samples is 1.04 for the new technique described in this thesis. The calculated leakage is within ±20% of the measured leakage for more than 50% of the samples, if the new $I_{leak\text{-}w}$ technique is used. Only 11% and 0.2% of the samples fall into this range for the $I_{leak\text{-}u}$ and $I_{leak\text{-}l}$ techniques respectively. $I_{leak\text{-}w}$ technique can be used to predict chip level leakage with better accuracy once device level leakage, parameter variation, and total transistor widths are known.

4.2 Leakage reduction

To reiterate, should the present scaling trend continue it is expected that the sub-threshold leakage power will become as much as 50% of the total power in the 0.09 μm generation [4]. Under this scenario, it is not only important to be able to predict sub-threshold leakage power more accurately as discussed in the previous section, it becomes crucial to identify techniques to reduce this leakage power component. It has been shown previously that the stacking of two *off* devices has significantly reduced sub-threshold leakage compared to a single *off* device [42, 43, 44]. This concept of stack effect is illustrated in Figure 4-3.

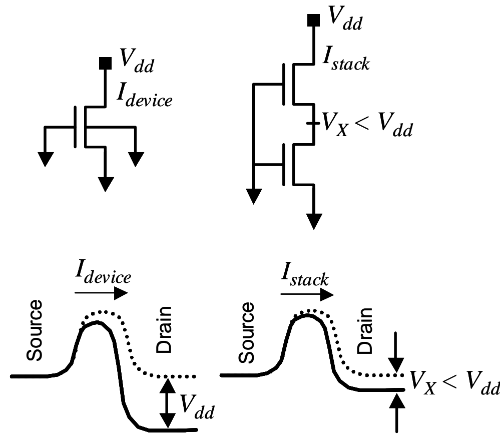


Figure 4-3: Leakage current difference between a single *off* device and a stack of two *off* devices. As illustrated by the energy band diagram, the barrier height is modulated to be higher for the two-stack due to smaller drain-to-source voltage resulting in reduced leakage.

In this section, a model is derived that predicts the stack effect factor, which is defined as the ratio of the leakage current in one *off* device to the leakage current in a stack of two *off* devices. Model derivation based on device fundamentals and verification of the model through statistical device measurements from 0.18 μm and 0.13 μm technology generations are presented in Section 4.2.1. The scaling nature of the stack effect leakage reduction factor is also discussed in the next section.

One solution to the problem of ever-increasing leakage is to force a non-stack device to a stack of two devices without affecting the input load, as shown in Figure 4-4. By ensuring iso-input load, the previous gate's delay and the switching power will remain unchanged. Logic gates after stack forcing will reduce leakage power, but incur a delay penalty, similar to replacing a low- V_t device with a high-V_t device in a dual-V_t design [45]. In a dual-V_t design, the low-V_t devices are used in performance critical paths and the high-V_t devices in the rest [46]. Usually a significant fraction of the devices can be high-V_t or forced-stack since a large number of the paths are non-critical. This will reduce the overall leakage power of the chip without impacting operating clock frequency. In Section 4.2.2 we discuss the stack forcing method to reduce leakage in paths that are not performance critical. This stack forcing technique either can be used in conjunction with dual-V_t or can be used to reduce the leakage in a single-V_t design. Differences between achieving leakage reduction through forced-stacks and channel length increase are discussed in Section 4.2.3. Case study and summary are presented in Section 4.2.4.

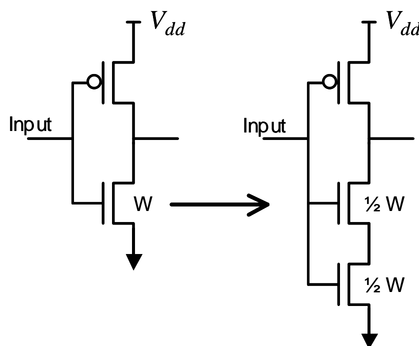


Figure 4-4: Trade-off between standby leakage and performance by forcing a two-stack under iso-input load. An NMOS two-stack will reduce leakage when input stays at logic "0"

4.2.1 Model for stack effect factor

Let I_1 be the leakage of a single device of unit width in *off* state with its $V_{gs} = V_{bs} = 0$ V and $V_{ds} = V_{dd}$. If the gate-drive, body bias, and drain-to-source voltages reduce by ΔV_g, ΔV_b, and ΔV_d respectively from the above-mentioned conditions, the leakage will reduce to,

$$I'_1 = I_1 \quad 10^{-\frac{1}{S}\left[\Delta V_g + \lambda_d \Delta V_d + k_\gamma \Delta V_b\right]}$$

where S is the sub-threshold swing, λ_d is the drain-induced barrier lowering (DIBL) factor, and k_γ is the body effect coefficient. The above equation assumes that the resulting $V_{ds} > 3kT/q$ [47]. For a two-device stack shown in Figure 4-5, a steady state condition will be reached when the intermediate node voltage V_{int} approaches V_x such that the leakage currents in the upper and lower devices are equal. Under this condition, the leakage currents in the upper and lower devices can be expressed as,

$$I_{stack\text{-}u} = w_u I_1 \;\; 10^{\frac{-(1+\lambda_d + k_\gamma)V_x}{S}}$$

$$I_{stack\text{-}l} = w_l I_1 \quad 10^{\frac{-\lambda_d (V_{dd} - V_x)}{S}}$$

and the intermediate node voltage will be,

$$V_x = \frac{\lambda_d V_{dd} + S \log \frac{w_u}{w_l}}{1 + k_\gamma + 2\lambda_d}$$

For short channel devices the body terminal's control on the channel is negligible compared to gate and drain terminals, implying $k_\gamma << 1 + 2\lambda_d$. Hence, the steady state value, V_x, of the intermediate node voltage can be approximated as,

$$V_x \approx \frac{\lambda_d V_{dd} + S \log \frac{w_u}{w_l}}{1 + 2\lambda_d}$$

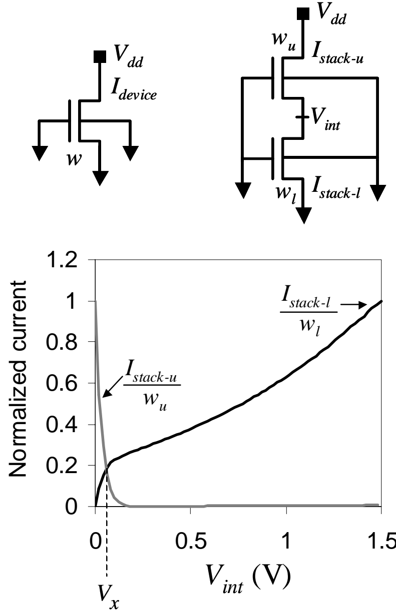


Figure 4-5: Load line analysis showing the leakage reduction in a two-stack.

Substituting V_x in either $I_{stack\text{-}u}$ or $I_{stack\text{-}l}$ will yield the leakage current in a two-stack given by,

$$I_{stack} = w_u^{\alpha} w_l^{1-\alpha} I_1 \ 10^{\frac{-\lambda_d V_{dd}}{S}(1-\alpha)}$$

$$where \quad \alpha \approx \frac{\lambda_d}{1+2\lambda_d}$$

The leakage reduction achievable in a two-stack comprising of devices with widths w_u and w_l compared to a single device of width w is given by,

$$X = \frac{I_{device}}{I_{stack}} = \frac{w}{w_u^{\alpha} w_l^{1-\alpha}} 10^{\frac{\lambda_d V_{dd}}{S}(1-\alpha)}$$

$$= 10^{\frac{\lambda_d V_{dd}}{S}(1-\alpha)} \quad when \quad w_u = w_l = w$$

The stack effect factor, when $w_u = w_l = w$, can be rewritten as,

$$X = 10^{\frac{\lambda_d V_{dd}}{S}\left(\frac{1+\lambda_d}{1+2\lambda_d}\right)} = 10^U$$

where U is the universal two-stack exponent which depends only on the process parameters, λ_d and S, and the design parameter, V_{dd}. Once these parameters are known, the reduction in leakage due to

a two-stack can be determined from the above model. It is essential to point out that the model assumes the intermediate node voltage to be greater than $3kT/q$.

To confirm the model's accuracy we performed device measurements on test structures fabricated in 0.18 μm and 0.13 μm process technologies. Results discussed in the rest of the section are from NMOS device measurements, but similar results hold true for PMOS devices as well.

Figure 4-6 shows NMOS device measurements under different temperature, V_{dd}, body bias, and channel length conditions for 0.18-μm technology generations, which prove the accuracy of the theoretical model. It is important to note that the model discussed above doesn't include the impact of diode junction leakages that originate at the intermediate stack node. In Figure 4-6, the model's accuracy deviates the most under reverse body bias for nominal channel length devices, where the ratio of diode junction leakage to sub-threshold leakage current increases.

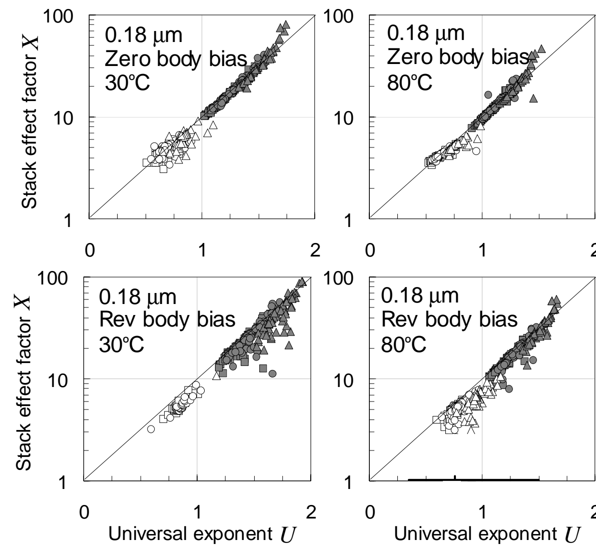


Figure 4-6: Measurement results showing the relationship between stack effect factor X for a two-stack to the universal exponent U. Lines indicate the relationship as per the analytical model and symbols are from measurement results. White symbols are for nominal channel devices and gray symbols are for devices smaller than the nominal channel length. Triangle, circle, and square symbols are for V_{dd} of 1.5, 1.2, and 1.1 V respectively. Zero body bias is when the body-to-source diode of the device closet to the power supply is zero biased and reverse body bias is when the diode is reverse biased by 0.5 V.

It is known that the stack effect factor strongly depends on λ_d as suggested by the model. In addition, a decrease in the channel length (L) will increase λ_d in a given technology [48]. So, any increase in the leakage of a single device due to decrease in L will not increase leakage of a two-stack at the same rate. This is illustrated in Figure 4-7 where increase in two-stack leakage is at a slower rate than that of a single device. Therefore, variation in L will result in smaller effective threshold voltage variation for a two-stack compared to a single device. Figure 4-8 illustrates the average stack effect factor for the nominal channel devices in both 0.18 μm and 0.13 μm technology generations obtained from both the measurements and the model. The increase in stack effect factor at a given V_{dd} with technology scaling is attributed to increase in λ_d, which is predicted by the analytical model. The higher stack effect factor for the low-V_t device in 0.13μm technology generation is due to the same effect.

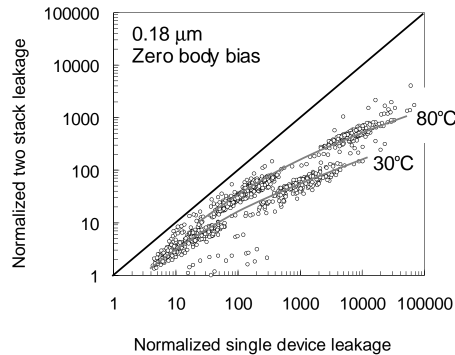


Figure 4-7: Measurement results indicate a slower rate of increase in leakage of two-stack compared to that of a single device. This should translate to reduction in the variation of effective threshold voltage.

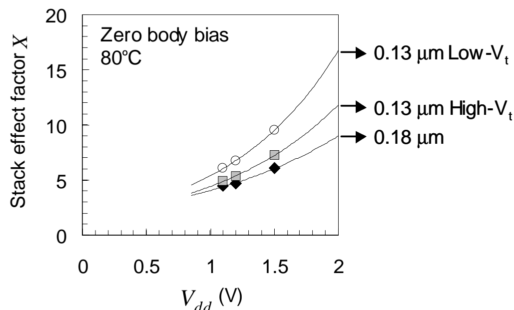


Figure 4-8: Nominal channel length device measurement results showing stack effect factor across two technology generations. The increase in stack effect factor is attributed to worsening of short channel effect, λ_d, which is predicted by the analytical model. The higher stack effect factor for the low-V_t device in 0.13-μm technology generation is attributed to the same reason. Lines are from analytical model and symbols are from measurement.

In 0.13-μm generation, the low-V_t device will dominate chip leakage. Figure 4-9 shows the scaling of stack effect from a 0.18-μm device to a 0.13-μm low-V_t device based on device measurements under different V_{dd} scaling scenarios. Since λ_d is expected to increase due to worsening device aspect ratio and since V_{dd} scaling will slow down due to related challenges [49], stack effect leakage reduction factor is expected to increase with technology scaling. The predicted scaling of stack effect factor from 0.18 μm to 0.06 μm is depicted in Figure 4-10.

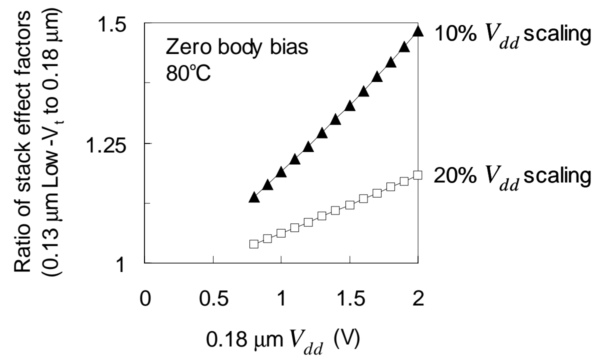


Figure 4-9: Nominal channel length device measurement results indicating the scaling of stack effect factor from 0.18μm to 0.13μm low-V_t under different V_{dd} scaling conditions. The low-V_t device will dominate leakage in 0.13μm technology, so the comparison is made with the low-V_t device.

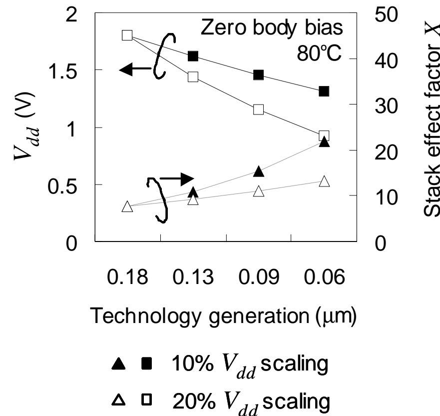


Figure 4-10: Prediction in the scaling of stack effect factor for two V_{dd} scaling scenarios in nominal channel length devices. V_{dd} for 0.18μm is assumed to be 1.8 V.

This scaling nature of stack effect factor makes it a powerful technique for leakage reduction in future technologies. In the next section, we describe a circuit technique for taking advantage of stack effect to reduce leakage at a functional block level.

4.2.2 Leakage reduction using forced-stacks

As shown earlier, stacking of two devices that are *off* has significantly reduced leakage compared to a single *off* device. However due to the iso-input load requirement and due to stacking of devices, the drive current of a forced-stack gate will be lower resulting in increased delay. So, stack forcing can be used only for paths that are non-critical, just like using high-V_t devices in a dual-V_t design [45, 46]. Forced-stack gates will have slower output edge rate similar to gates with high-V_t devices. Figure 4-11 illustrates the use of techniques that provide delay-leakage trade-off. As demonstrated in the figure, paths that are faster than required can be slowed down which will result in leakage savings. Such trade-offs are valid only if the resulting path still meets the target delay. Figure 10 shows the delay-leakage trade-off due to n-stack forcing of an inverter with fan-out of 1 under iso-input load conditions in a 0.13 µm technology [50].

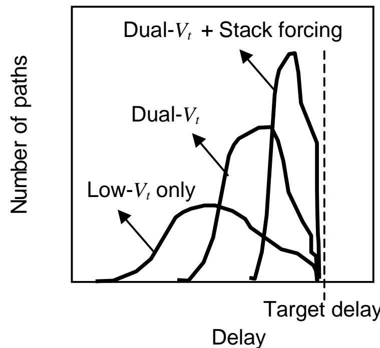


Figure 4-11: Stack forcing and dual-V_t can reduce leakage of gates in paths that are faster than required.

By properly employing forced-stack, one can reduce standby and active leakage of non-critical paths even if a dual-V_t process is not available. This method can also be used in conjunction with dual-V_t. Stack forcing provides wider coverage in the delay-leakage trade-off space as illustrated in Figure 4-12.

Functional blocks have naturally stacked gates such as NAND, NOR, or other complex gates. By maximizing the number of natural stacks in *off* state during standby by setting proper input vectors, the standby leakage of functional block can be reduced. Since it is not possible to force all

natural stacks in the functional block to be in *off* state the overall leakage reduction at a block level will be far less than the stack effect leakage reduction possible at a single logic gate level [42]. With stack-forcing the potential for leakage reduction will be higher. Figure 4-13 and Figure 4-14 illustrates such an example.

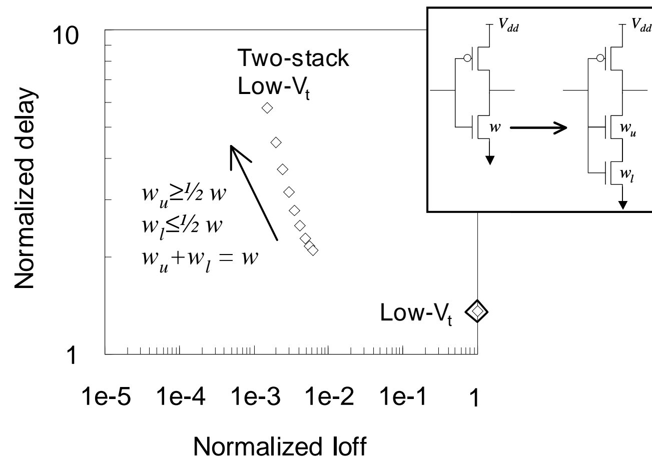


Figure 4-12: Simulation result showing the nominal channel length delay versus mean leakage trade-off that can be achieved by stack forcing technique under iso-input load conditions. Iso-input load is achieved by making the gate area after stack forcing identical to before stack forcing. Several such conditions are possible, which enhances delay-leakage trade-off possible by stack forcing. The two-stack condition with the least delay is for w_u=w_l=½w. This trade-off can be used with or without high-V_t transistors.

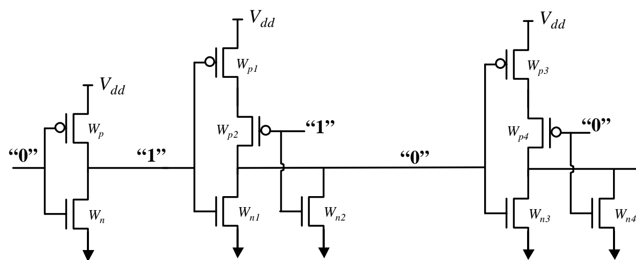


Figure 4-13: A sample path where natural stack is used to reduce standby leakage by applying a predetermined vector during standby. No delay penalty is incurred with this technique.

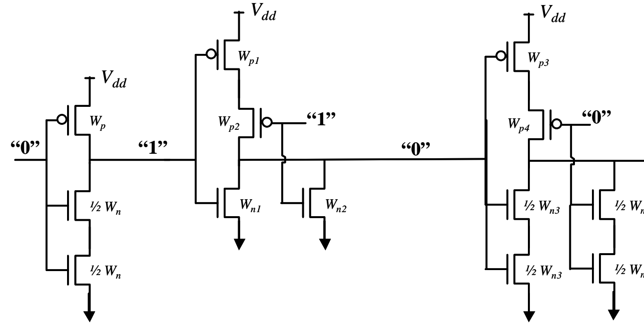


Figure 4-14: Using stack-forcing technique the number of logic gates in stack mode can be increased. This will enable further leakage reduction in standby mode. Increase in delay under normal mode of operation will be incurred.

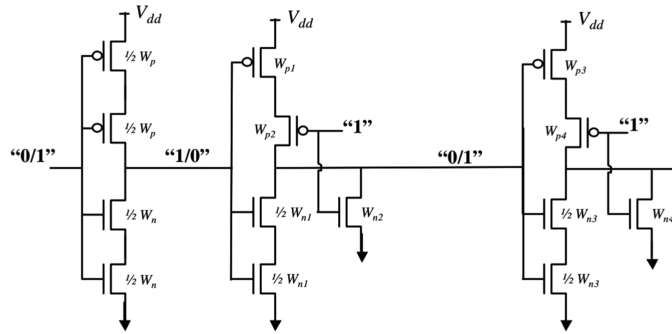


Figure 4-15: If a gate can have its input as either "0" or "1" and still force stack effect then that gate will have reduced active leakage. The more the number of inputs that can be either "0" or "1" the higher the probability that stack effect will reduce active leakage.

Forcing a stack in both n- and p-networks of a gate will guarantee leakage reduction due to stacking, independent of the input logic level. Such an example is shown in Figure 4-15. To reiterate, stack forcing can be applied to paths only if increase in delay due to stacking does not violate timing requirements. Gates that can force stack effect independent of its input vectors will automatically go into leakage reduction mode when the intermediate node of the stack reaches the steady state voltage. This will boost standby and active leakage reduction since a specific input vector need not be applied.

4.2.3 Stack effect vs. channel length increase

It is possible to facilitate delay-leakage trade-off by increasing the channel length of devices [51] that are in non-critical paths. To maintain iso-input load the channel width will have to be reduced along with increase in the channel length. Figure 4-16 shows the mean leakage reduction

achievable by increasing the channel length. In Figure 4-16 the channel length of interest is given by η x 0.18 μm and stack leakage is for a stack of two devices with η of 1 and w_u=w_l=½w. As it is clear from Figure 4-16, the channel length has to be increased 3 times as that of the nominal channel length to match the mean leakage of a two-stack of 0.18μm devices. The reason for such a large increase is attributed to the reverse short channel effect that is present due to halo doping [49] where V_t reduces with increase in channel length. It is important to note that stacking two devices of nominal channel length is different from doubling the channel length due to the two dimensional nature of barrier-lowering and drain induced barrier lowering effects described in Section 2.1.

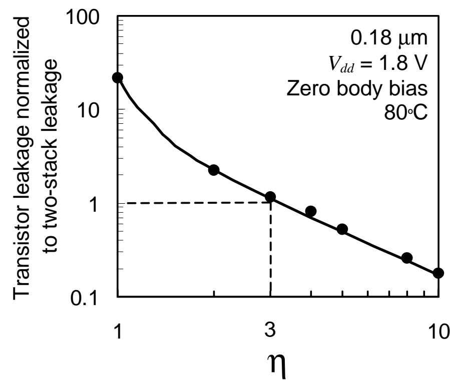


Figure 4-16: Comparing device leakage reduction due to channel length increase with two-stack leakage. The channel length is given by η x 0.18 μm. Stack leakage is a two stack of devices with η=1 and w_u=w_l=½w. Leakage numbers are obtained from simulation under iso-input load.

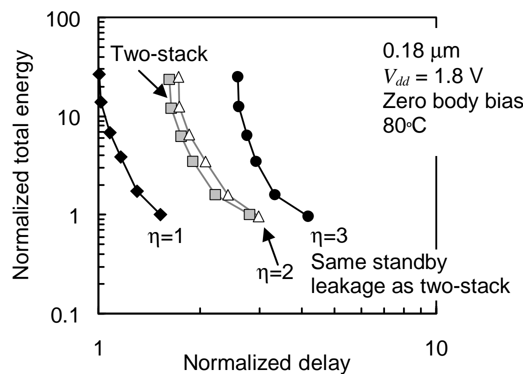


Figure 4-17: Energy-delay trade-off of inverter under different configurations with fan-out of 1 and iso-input load. The simulation-based comparison clearly shows that the two-stack configuration's delay is less than increasing channel length, especially when compared to iso-standby leakage (η=3) configuration.

Figure 4-17 shows the energy-delay trade-off of an inverter under different configurations with fan-out of 1 and iso-input load. The simulation-based comparison clearly shows that the two-stack configuration's delay is less than delay due to increasing channel length, especially when compared to iso-standby leakage ($\eta \approx 3$) configuration. As summarized in Figure 4-18, η of 2 has about the same delay as that of the two-stack with η of 1 but with a 2.3X higher mean leakage. On the other hand, η of 3 provides about the same mean leakage as the two-stack but with 60% higher delay.

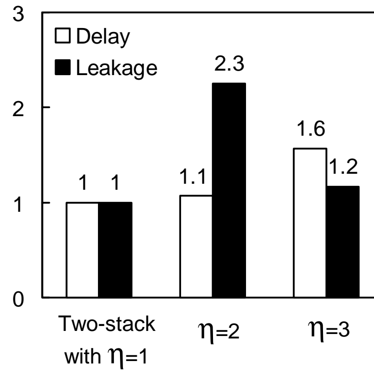


Figure 4-18: Summary of delay-leakage trade-off comparison between two-stack and channel length.

4.2.4 Case study and summary

Two-stack assignment of low-*Vt* transistors was applied to a 32-bit microprocessor's instruction decode block in 0.13 µm technology. Stack assignment was done so that the all-low-*Vt* maximum frequency of 1 GHz is preserved at 1.4 V. Switching power of 45.9 mW at 1.4 V was also preserved since iso-input load was maintained during stack assignment. All low-*Vt* leakage power was 39.1 mW. Iso-frequency stack assignment allowed conversion about 70% of transistor width to two-stack, resulting in leakage power reduction of 3X. If high-*Vt* assignment is used then about 95% of the transistor width became high-*Vt*, resulting in 4.3X leakage reduction.

A model based on device fundamentals that predicted the scaling nature of stack effect based leakage reduction was presented. Device measurements verified the model's accuracy across different temperatures, channel lengths, body bias values, supply voltages, and process technologies. Modes for using stack forcing to reduce standby and active leakage components were discussed and the advantage of stack forcing over channel length increase for delay-leakage trade-

off was demonstrated. Case study showed the potential for leakage reduction at a block level without reducing the maximum frequency of operation.

Chapter 5
Neighborhood Threshold Voltage Variation

So far, this thesis has addressed effects of macroscopic level threshold voltage variation on high-performance circuits. This chapter will deal with the variation in the threshold voltage of matched devices that are within a few microns apart in the same neighborhood. The devices of interest that are in close proximity can be either of the same or different polarity. Matched devices of the same polarity are used as sense-amplifier input devices for low voltage swing sensing among other applications [27]. Any mismatch in threshold voltage of this input device pair will appear as input offset resulting in degraded performance. A simple voltage-biasing scheme that reduces the mismatch between matched transistor pair of same polarity will be discussed.

In addition, for some digital CMOS circuits a known PMOS to NMOS drive current ratio is required either to achieve a well-defined switching threshold or to achieve equal rising and falling delays. Since the processing steps such as threshold voltage implants for the PMOS and NMOS devices are not correlated there could be significant variation between the required and achieved threshold voltages for the two device types. The short channel effects further worsen this variation. The net variation will change the drive current ratio of PMOS to NMOS devices and can affect the operation of high performance circuits that depend on a pre-determined skew between the two device types. Ability to adjust the charging and discharging currents by sensing the skew difference can alleviate this problem. Current biasing schemes that maintain the relationship between the charging and discharging currents, independent of the process skew is explained. The first current scheme that is the simplest, guarantees constant ratio between charging and discharging currents no matter the change in the relative skews of the PMOS and NMOS devices. Although this scheme

maintains the relationship between charging and discharging delays, it doesn't provide constant delay as the threshold voltages vary. A true process insensitive current generation theory and circuit will also be described [28]. This can then be used as bias current for the charging PMOS and the discharging NMOS networks enabling a threshold voltage variation and skew variation insensitive circuit.

5.1 Voltage biasing

For short channel devices as the channel length approaches the source-body and drain-body depletion widths, the charge in the channel due to these parasitic diodes become comparable to the depletion charge due to the MOSFET gate-body voltage, rendering the gate and body terminals to be less effective. To reiterate from Section 2.1, the finite depletion width of the parasitic diodes do not influence the energy barrier height to be overcome for inversion formation in a long channel device. However, as the channel length becomes shorter both channel length and drain voltage reduce this barrier height. This two-dimensional effect makes the barrier height to be modulated by channel length variation resulting in threshold voltage variation. Even for matched MOS device pairs there will be non-zero channel length mismatch between the two devices, resulting in threshold voltage mismatch.

The amount of barrier height lowering, threshold voltage mismatch, and gate and body terminal's channel control loss will directly depend on the charge contribution percentage of the parasitic diodes to the total channel charge. This contribution can be reduced by applying forward bias to these parasitic diodes [41]. As this voltage bias is applied, the threshold voltages of the MOS devices reduce and the mismatch in threshold voltage for a given mismatch in physical parameter reduces. A 0.18-μm testchip with different mismatch pairs were fabricated (Figure 5-1). Measurement results of linear threshold voltage mismatch (V*t-lin*) on wide-width and short-channel device pair are shown in Figure 5-2. It is clear from that 500 mV forward bias reduces threshold voltage mismatch by (*i*) 34% compared to zero bias with V*t-lin* reduction of 34% and (*ii*) 37% compared to 500 mV reverse bias with V*t-lin* reduction of 45%.

Figure 5-1: Die-micrograph of mismatch structures testchip.

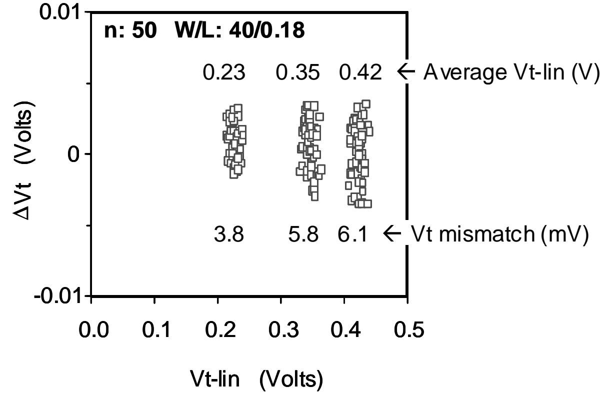


Figure 5-2: Linear threshold voltage mismatch for 500 mV forward body bias, zero body bias and 500 mV reverse body bias.

Measurements also confirm that mismatch sensitivity to body bias is maintained as V_{ds} is increased for the device to operate in the saturation region. For wide width and short channel devices, the drain current can be approximated as follows,

$$I_{\mathrm{dsat}} \propto (V_{gs} - V_{tsat})$$

$$\Rightarrow I_{\mathrm{dsat1}} \propto (V_{gs} - V_{tsat1}) \qquad I_{\mathrm{dsat1}} \propto (V_{gs} - V_{tsat2})$$

$$\therefore I_{mismatch} \propto V_{tsat1} - V_{tsat2} \quad \& \quad I_{avg} \propto (V_{gs} - V_{tsat}^{avg})$$

The percentage mismatch in the drain currents can then be written as,

$$\frac{I_{mismatch}}{I_{avg}} = \frac{V_{tsat1} - V_{tsat2}}{V_{gs} - V_{tsat}^{avg}}$$

Therefore, as the body bias is changed from reverse mode to forward mode both threshold voltage mismatch and threshold voltage reductions contribute positively in reducing the percentage mismatch in drain currents. It is evident that the impact of threshold voltage reduction on drive current mismatch will be more pronounced for smaller value of gate voltage.

5.1.1 Application of voltage bias to low-voltage sense-amplifiers

Sense-amplifiers are used to amplify a low-voltage swing differential signal to full-swing digital signal. They are widely used as receivers at the end of long interconnection such as memory to enhance performance. The minimum input differential that can be sensed by a sense-amplifier depends of several factors including input offset due to threshold voltage mismatch of input device pair. A traditional sense-amplifier is shown in Figure 5-3. Here the body of the input pair that is

used to sense the voltage is connected to the power supply. As the sense-amplifier's strobe is enabled to sense the input differential the body bias voltage of the matched input devices starts at reverse bias as shown in Figure 5-4. This will result in increased input offset, as the sense-amplifier is about to amplify the input differential.

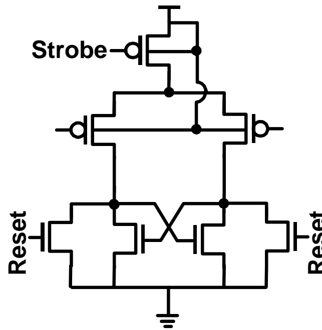


Figure 5-3: Traditional sense-amplifier.

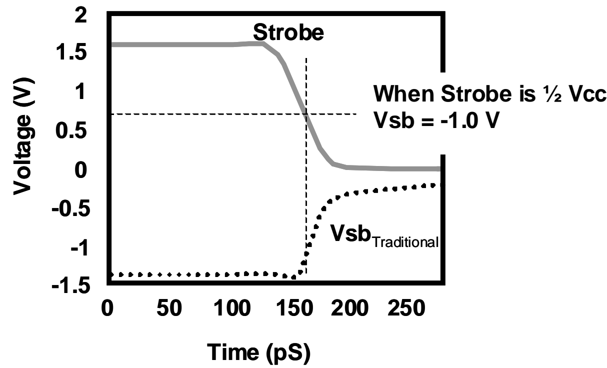


Figure 5-4: Body voltage for the traditional sense-amplifier.

Statistical measurement of saturation threshold voltage mismatch sensitivity to body bias for typical sense-amplifier input pair device width of 2 μm in a 0.18 μm technology at V_{ds} of 1.5 V is illustrated in Figure 5-5. For the traditional sense-amplifier shown in Figure 5-3, the mismatch is measured to be 65 mV compared to 33 mV under zero body bias condition – a difference of 32 mV. This increase in mismatch will require increased minimum input differential for the traditional sense-amplifier to function compared to if the body bias can be maintained as zero.

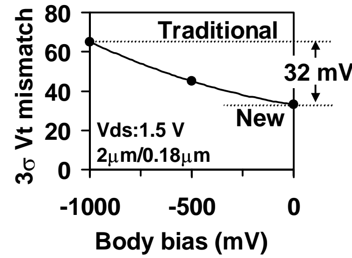


Figure 5-5: Dependence of saturation threshold voltage mismatch on body bias.

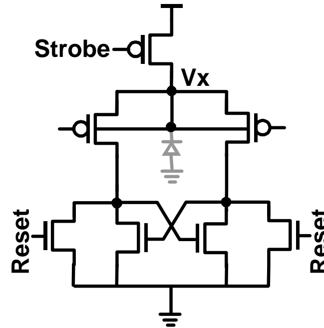


Figure 5-6: New no body bias sense-amplifier.

A modified no body bias sense-amplifier (Figure 5-6) where the input device pair's shared body is shorted to the shared source terminal achieves input differential sensing under zero body bias. This will result in the reduction of the minimum input differential required for the sense-amplifier by 32 mV and hence will translate to reduced total delay. Total delay includes the sense-amplifier delay plus the input differential development delay. Delay simulations for the two sense-amplifiers for ramp-rate of 1 mV/pS, at 1.5 V and 110 C are shown in Figure 5-7.

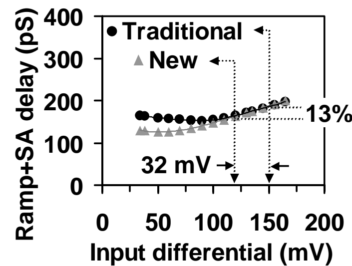


Figure 5-7: Total delay verses input differential for iso-output differential at 1.5 V, 1 mV/pS ramp rate, and 110 Celsius, for the traditional and the new sense-amplifiers.

Since the source of the input device pair is shorted to the body, this will result in a slight increase in the capacitance at the intermediate node. Simulations show that delay of the modified

no body bias sense-amplifier is 12-18% smaller (Figure 5-8), in spite of additional capacitive loading of the source node by the n-well to substrate junction. The input differential requirement for the traditional sense-amplifier was set at 150 mV to include other sources of noise such as strobe timing and power supply variations. The input differential requirement can be reduced to 118 mV for the no body bias sense-amplifier – resulting in 13% reduction in total delay, majority of which comes from reduction in the input differential development delay. Performance improvement summary is listed in Table 3-1 for different supply voltages and ramp rates.

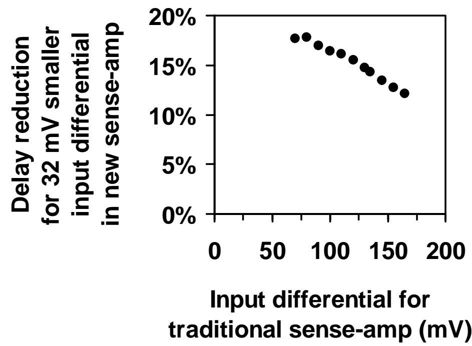


Figure 5-8: Total delay (sense-amplifier delay plus ramp development delay) improvement due to input offset reduction in the new sense-amplifier at 1.5 V, 1 mV/pS ramp rate and 110 Celsius.

Vcc (V)	Delay improvement	
	Ramp rate: 1 mV/pS	Ramp rate: 2.5 mV/pS
1.1	10%	1%
1.3	12%	4%
1.5	13%	7%
1.8	15%	10%

Table 5-1: Total delay improvement under different supply voltage and ramp rate conditions for input differential of 150 mV for the traditional sense-amplifier and 118 mV for the new zero body bias sense-amplifier at 110 Celsius. Larger improvement is correlated to faster sense-amplifier resulting in input offset and ramp development delay reductions more critical.

The benefit of this voltage-biasing technique was illustrated for a PMOS sense-amplifier in an n-well process. If this technique has to be applied for both PMOS and NMOS sense-amplifiers, either a triple-well process or PD-SOI with substrate contact process will be required. Operational amplifiers can also benefit from this voltage-biasing scheme.

5.2 Current biasing

In the previous section, voltage-biasing technique was introduced to reduce the impact of electrical mismatch between devices of the same polarity (NMOS vs. NMOS or PMOS vs. PMOS). In some digital circuits, it is necessary to minimize electrical mismatch between devices of opposite polarity (NMOS vs. PMOS). Digital CMOS circuits that require a known PMOS to NMOS drive current ratio is either to achieve a well-defined switching threshold or to achieve equal rising and falling delays. Since the processing steps such as threshold voltage implants for the PMOS and NMOS devices are not correlated there could be significant variation between the required and achieved threshold voltages for the two device types. The short channel effects further worsen this variation. The net variation will change the drive current ratio of PMOS to NMOS devices and can affect the operation of high performance circuits that depend on a pre-determined skew between the two device types.

Ability to adjust the charging and discharging currents by sensing the skew difference can alleviate this problem. Current biasing schemes that maintain the relationship between the charging and discharging currents, independent of the process skew is explained in the second half of this chapter. The first current scheme that is the simplest, guarantees constant ratio between charging and discharging currents no matter the change in the relative skews of the PMOS and NMOS devices. Application of this current biasing scheme to generate true non-overlapping two-phase clock is presented. Although this scheme maintains the relationship between charging and discharging delays, it doesn't provide constant delay as the threshold voltages vary. A true process insensitive current generation theory and circuit will be described [28]. This can then be used as bias current for the charging PMOS and the discharging NMOS networks enabling a threshold voltage variation and skew variation insensitive circuit. This second current biasing scheme can be used to maintain switching threshold of CMOS gates independent of the PMOS to NMOS process skew among other biasing applications.

5.2.1 Basic iso-current biasing and two-phase clock generation

The iso-current biasing scheme, illustrated in Figure 5-9, guarantees same charging and discharging currents no matter the change in the relative skews between the PMOS and NMOS devices. Therefore, it enables identical charging and discharging delays. The computation block comprising of the PMOS and NMOS networks in Figure 5-9 can be a simple inverter as a repeater,

or a single phase to two-phase clock generation logic, or any other CMOS gate that requires a preset switching threshold.

Figure 5-9: Basic iso-current biasing scheme

True and complement clock generation circuits are used today in a variety of CMOS based designs. Each clock generation circuit takes a single clock input, sends a clock to one output, and sends its complement to the other output. True non-overlapping two-phase clock circuits must adhere to the following requirements: (1) phase one's falling edge should align with phase two's rising edge and (2) phase two's rising edge should align with phase one's falling edge. This implies that the two outputs are exactly 180° out of phase with each other and that their edges cross at ½Vcc.

Figure 5-10: Standard two-phase clock generator design

Standard two-phase clock generator designs typically use an inverter for one output and an optimized delay element for the other output to accomplish these design requirements (Figure 5-10). The delay of the inverter and the transmission gate delays are matched at one design corner,

assuming the capacitive loads at the two outputs are matched. Unfortunately, a variety of issues, such as temperature, voltage, and process fluctuations affect the operation of the optimized design. With clock period decreasing at a faster rate than transistor delay, the overlapping duration between the true and complement edges of these clock circuits under non-ideal conditions becomes increasingly critical. The non-idealities arise from the fact that PMOS charging current can vary differently from NMOS discharging current with process variation.

The iso-current biasing scheme discussed earlier is used to design a circuit that will maintain PMOS and NMOS currents to be equal under all process corners. Since the circuit self-adjusts to temperature, supply voltage, and process fluctuations, a reduction in the overlap time duration of the two phases can achieved compared to the present state of the art.

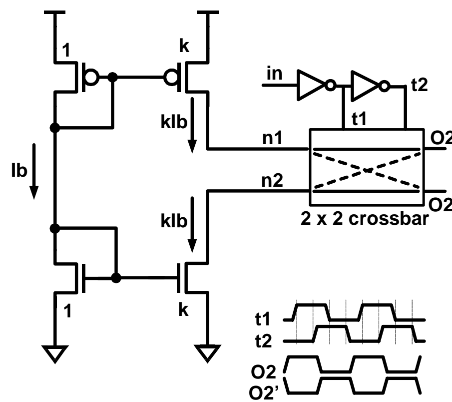


Figure 5-11: Iso-bias current based non-overlapping two-phase clock generator.

The iso-bias current based non-overlapping two-phase clock (Figure 5-11) has three main components. The first component is a current mirror that forces a constant current ratio through nodes n1 and n2 for majority of the charging and discharging time when the charging and discharging devices are in the saturation region. The second segment is the 2 x 2 transmission gate based crossbar switch, which has two steady state positions. In the first position the PMOS charges output O2 while the NMOS discharges output O2', in the second position while the PMOS charges O2' the NMOS discharges O2. The third component is two inverters that drive the crossbar switch. Although these two inverters are susceptible to the problems mentioned in the previous paragraph, the delay between the two signals shouldn't impact the final output, since the delay difference

affect the both output phases almost equally. The current mirror governs these charging and discharging rates, which help to ensure equal rising and falling slopes on the outputs. When both selection inputs are high, all the transmission gates in the switch are turned on. The PMOS charges the output while the NMOS discharges the path at the same time – essentially a short to ground with the transistors acting as resistors in series.

Simulation data was gathered for 63 different operating conditions that included seven different process skews, three different temperatures (30C, 80C, and 110C), and three different voltages (1.1V, 1.3V, and 1.5V). As shown in the Figure 5-12, the iso-bias current based two-phase clock design improves mean and standard deviation clock skew by a factor of 3X over the current best known method. This benefit comes at a cost of about 100% increase in transistor width and power due to additional complexity for iso-edge rate.

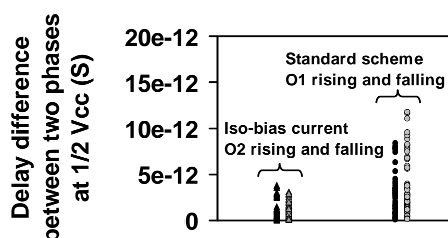


Figure 5-12: Performance comparison of two-phase generators.

5.2.2 Process insensitive current biasing

The first current scheme that is the simplest, guarantees constant ratio between charging and discharging currents no matter the change in the relative skews of the PMOS and NMOS devices. Application of this current biasing scheme to generate true non-overlapping two-phase clock is presented. Although this scheme maintains the relationship between charging and discharging delays, it doesn't provide constant delay as the threshold voltages vary. A true process insensitive current generation theory and circuit will be described. This current is then used as biasing current (Figure 5-13) to provide process insensitive charging and discharging delays. Another application of the scheme will be set a predetermined ratio between the charging and discharging currents to obtain a process skew insensitive switching threshold for CMOS logic gates.

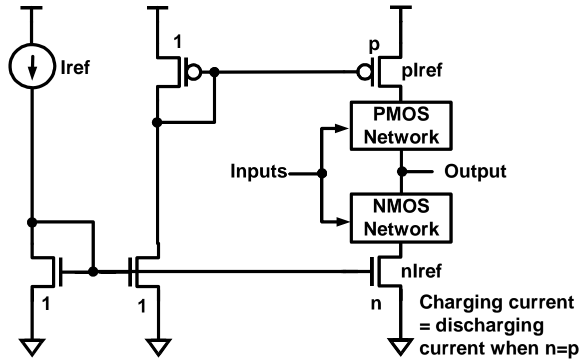


Figure 5-13: Process insensitive current biasing scheme.

5.2.2.1 Process insensitive constant current generation

Prior published works on current references fall into one of the following three categories: (*i*) translation of bandgap voltage to reference current, (*ii*) translation of MOSFET based 'reference' voltage to reference current, or (*iii*) direct generation of reference current using MOSFET transistors.

The first category requires generation of a bandgap voltage and an off-chip resistor [52]. Off-chip resistors not only increase system cost, but also limit the usage of such current reference circuits. In addition, even the best-known bandgap voltage generation circuit [53] will not work well with CMOS technology supply voltage expected to reach sub-0.7 V [54]. The voltage-scaling problem is solved in the second category of circuits by replacing the bandgap voltage with a MOSFET based 'reference' voltage [55]. However, MOSFET based 'reference' voltage cannot be truly process independent, since it generates threshold voltage of a device at 0 K, as shown in equation (1). In addition, these circuits still require an off-chip resistor. Although resistor-less voltage to current translation has been proposed [56], the remaining problems associated with voltage to current translation are still unsolved.

$$\begin{aligned} V_{ref}^{MOS} &= V_t + k_c V_{PTAT} \\ &\approx (V_{TO} - k_a T) + k_c V_{PTAT} \\ &= V_{TO} \qquad \cdots \quad (1) \end{aligned}$$

A third category of circuits in which MOSFET devices are used directly to generate reference current, solve problems associated with scaling, cost, and usage flexibility. The circuit discussed in [57] is one such example, where a positive temperature coefficient current is added to a negative temperature coefficient current. This circuit while providing a temperature insensitive current, however, does not provide process insensitivity. As the effective gate oxide thickness approaches sub-35 Å [58], it will be necessary to design MOSFET current reference circuits that are insensitive to gate oxide thickness variation. In this section, a CMOS current reference concept that addresses scaling, cost, usage flexibility, and process variations that exist in deep sub-micron MOSFET devices.

$$I_1 \approx \beta z_1 (V_{gs1} - V_t)^2 \qquad I_2 \approx \beta z_2 (V_{gs2} - V_t)^2 \quad \ldots \quad (2)$$

$$\beta = \mu C_{ox}; \quad V_t \approx \frac{\sqrt{qN_a \varepsilon_{si} \phi}}{C_{ox}}; \quad z_1 = \frac{W_1}{2L}; z_2 = \frac{W_2}{2L} \quad \ldots \quad (3)$$

$$\frac{dI_1}{dP} \approx z_1 (V_{gs1} - V_t)^2 \frac{d\beta}{dP} - 2\beta z_1 (V_{gs1} - V_t) \frac{d\beta}{dP} \frac{dV_t}{d\beta} \quad \ldots \quad (4)$$

$$\frac{dI_2}{dP} \approx z_2 (V_{gs2} - V_t)^2 \frac{d\beta}{dP} - 2\beta z_2 (V_{gs2} - V_t) \frac{d\beta}{dP} \frac{dV_t}{d\beta} \quad \ldots \quad (5)$$

Let $I_{ref} = I_1 - I_2$, $V_{gs2} = aV_t$, and $V_{gs1} = bV_t$ $\quad \ldots \quad (6)$

By equating $z_1 (V_{gs1} - V_t)^2 = 2\beta z_2 (V_{gs2} - V_t) \frac{dV_t}{d\beta}$

and $z_2 (V_{gs2} - V_t)^2 = 2\beta z_1 (V_{gs1} - V_t) \frac{dV_t}{d\beta}$ we get

$$\frac{z_1}{z_2} = \frac{(a^2 - 1)}{(b^2 - 1)} \quad \ldots \quad (7)$$

This will ensure $\frac{dI_{ref}}{dP} = \frac{dI_1}{dP} - \frac{dI_2}{dP} \approx 0$

and $I_{ref} = I_1 - I_2$ to be non-zero as long as $a \neq b$

The idea behind process compensated current, I_{ref}, is to take saturation current of two MOSFET devices, I_1 and I_2, and use the natural variation in these two currents to cancel out variations in the difference of the two currents, *i.e.* $I_{ref} = I_1 - I_2$. We use long-channel wide-width MOSFET devices

to avoid process variation related to small lateral dimensions. Equations (2) and (3) show the saturation currents I_1 and I_2. We assume that the devices generating I_1 and I_2 are laid out to have proper matching.

Process parameters that are expected to impact the magnitudes of currents I_1 and I_2 are β and V_t. Equations (4) and (5) show the change in two currents, I_1 and I_2, with respect to process for MOSFET devices operating in the saturation region, assuming mobility is not a strong function of channel doping. To achieve a non-zero process compensated current, I_{ref}, circuit parameters are set such that $d\beta/dP$ term of one current is canceled with dV_t/dP term of the other current, so that $dI_{ref}/dP = dI_1/dP - dI_2/dP$, will be zero, but $I_{ref} = I_1 - I_2$ will be non-zero. The necessary and sufficient conditions to achieve process compensation for I_{ref} are given by equations (6) and (7). Table 5-2 shows a possible set of values for the circuit parameters a, b, and z_1/z_2 that satisfy equations (6)-(7).

a	*b*	*z1 / z2*
2	5	1 / 8
2.33	4	8 / 27
2.6	3.5	1 / 2
3.5	2.6	2 / 1
4	2.33	27 / 8
5	2	8 / 1

Table 5-2: Sub-set of parameters that satisfy equations (6)-(7) to minimize process impact on $I_{ref} = I_1 - I_2$.

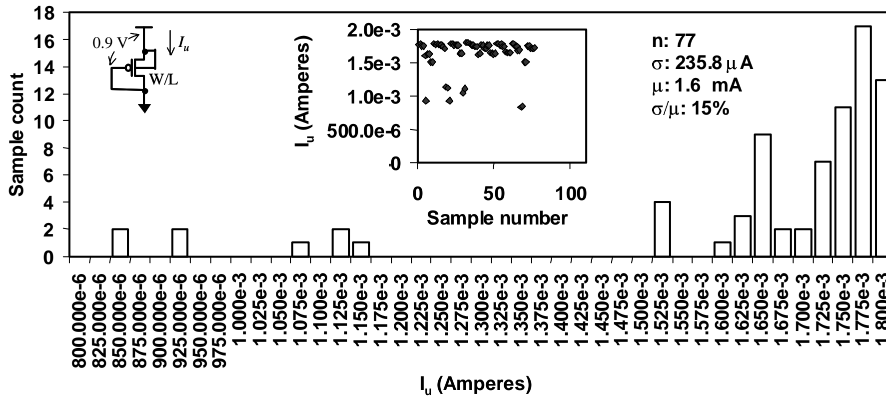


Figure 5-14: Measured process variation in a long-channel, wide-width, process-uncompensated, device current (I_u). Measurements were carried out across wafer on identical devices with 0.9 V gate drive. Both raw data and statistical information are presented above.

Long channel device measurements were carried out at 30ºC across a single wafer in 0.18μm technology generation [58]. As a control experiment we measured variation in a process-uncompensated diode-connected device current, I_u with 0.9 V gate drive as shown in Figure 5-14. Results show that normalized variation (σ/μ) in such uncompensated device current to be 15%.

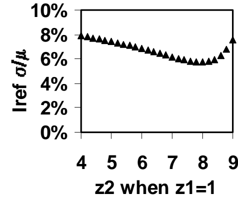


Figure 5-15: Normalized process variation in I_{ref} for different device size ratios when a=2 and b=5. Measurement confirms process variation in I_{ref} minimizes at $z1/z2$ ratio predicted by the theoretical model.

For I_{ref} generation the gate voltages for I_1 and I_2 as required by (6) were calculated by first measuring device V_t. Figure 5-15 illustrates σ/μ in I_{ref}, with $a = 2$ and $b = 5$, for various values of z_1/z_2. Measurements clearly indicate that variation in I_{ref} is reduced compared to the uncompensated device current. Also, the best process compensation occurs when $z_1/z_2 = 1/8$, as predicted by the theoretical model. Figure 5-16 shows statistical distribution in I_{ref} for $a = 2$, $b = 5$, and $z_1/z_2 = 1/8$. The normalized variation, σ/μ, for I_{ref} across a single wafer was 5.7%, which translates to 2.6X reduction in variation compared to the uncompensated device current.

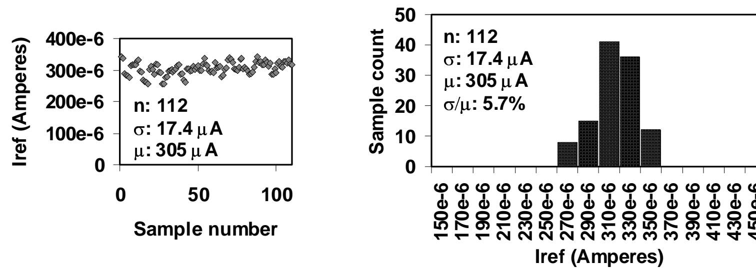


Figure 5-16: Measured variation in I_{ref} for a=2, b=5, and $z1/z2$=1/8. Device current and V_t measurements were carried out across wafer on two devices with appropriate gate drives and device sizes given by the theoretical model.

We also performed circuit simulations to predict the improvement in process compensation in I_{ref} across a process lifetime. The circuit used to generate process compensated I_{ref} for a = 2 and b = 5 is depicted in Figure 5-17. Block A of the circuit in Figure 5-17 illustrates method for generation of aV_t and bV_t voltages. Since it is not possible to accurately generate V_t, the current reference device size ratio z_1/z_2 was optimized for minimizing variation in I_{ref}. Figure 5-18 shows simulation result for the process compensated I_{ref} at V_{dd} of 0.9 V and 30ºC. Results indicate that process variation in uncompensated device current I_u to be 0.48 while the variation in I_{ref} for z_1/z_2=1/6 remained within 0.05, an improvement of 7.6X. It is important to note that for the circuit to operate at sub-1V we need enough head room for the V_t generation circuit, especially with b = 5. With long channel V_t of ~100 mV we were able to successfully generate I_{ref} at V_{dd} of 0.9 V. Further reduction in supply voltage is possible by designing the subtraction for b=3 and a=2 as shown in Table 5-3. Supply voltage dependence of the reference current is also summarized in Table 5-3.

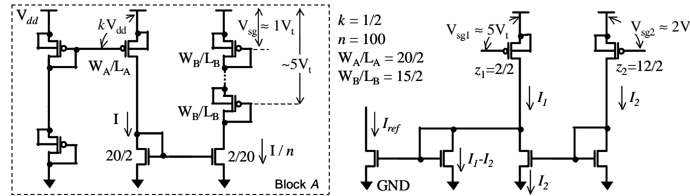


Figure 5-17: Circuit schematics showing generation of V_t and I_{ref}. Since generated V_t will not be accurate, device size ratio $z1/z2$ was optimized with a=2, b=5 and V_{dd}=0.9 V to minimize I_{ref}'s process variation.

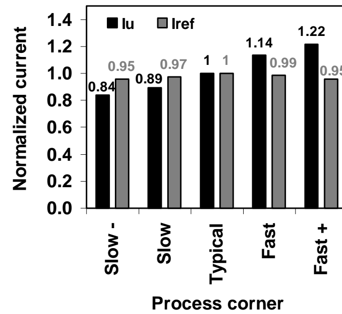


Figure 5-18: Circuit simulation results with a=2, b=5, $z1/z2$=1/6, V_{dd}=0.9 V, showing variation in I_u and I_{ref}. With respect to typical process corner I_u varied by +22% and -16% while variation in I_{ref} was –5% and –5%. Total variation in normalized I_u across all process corners is 0.38 while it is 0.05 for normalized I_{ref}.

b, a	V_{dd}^{min} (V)	Temp (°C)	I_{ref} variation	V_{dd} sensitivity
5, 2	0.9	30	5.0%	0.3% per 100 mV
3, 2	0.6	30	5.2%	0.4% per 100 mV

Table 5-3: Low voltage operation enabled by redesigning Vt generation circuit

A sub-1 V CMOS process compensated current reference generation that reduces process sensitivity was demonstrated. Device measurements and circuit simulations show 2.6X to 7.6X reduction in process variation of the generated current compared to device current without process compensation. This process compensated current can be used to bias the CMOS logic shown in Figure 5-13.

Chapter 6
Conclusion

A 50 GHz microprocessor with 2 billion logic transistors (additionally an order of magnitude memory transistors) using 22 nm drawn channel length (7 nm of effective channel length) devices operating at 250 mV supply voltage by first half of the next decade – this is the expected roadmap should the scaling trends continue. Can we achieve this – maybe, maybe not! To be able to even dream about such a processing system, it is important to be able to do predictive design. The old and easy way of designing for worst-case will not be adequate. It is important to accept that process variation is a reality and that one has to design circuits, with variation in mind.

One of the important device parameters that impact the design of circuits is its threshold voltage. Since the variation in the threshold voltage is expected to increase with scaling, it is imperative to understand the nature of its impact, models to predict the magnitude of impact, and techniques to reduce its impact.

6.1 Contributions

This thesis focused on the impact short channel induced threshold voltage variation will have on high-performance circuits. Three separate threshold voltage variation categories were considered in depth. In Chapter 3 of this thesis an analytical model was developed, to show that traditional adaptive reverse body bias circuit solution to reduce die-to-die threshold voltage variation is not scalable for future generations and the fact that this technique results in increased within-die threshold voltage variation. Use of bi-directional adaptive forward and reverse body bias to limit threshold voltage variation was shown to be a better alternative through 150 nm testchip

measurements. Design methodology to choose proper bias circuit impedance for on-chip body bias generators were described in Chapter 3 as well.

It was shown that threshold voltage variation not only affects supply voltage scaling but also the accuracy of leakage power estimation. Accurate leakage power estimation is very critical for future CMOS systems since the leakage power is expected to be a significant portion of the total power due to threshold voltage scaling. In Chapter 4, leakage power estimation that takes into account within-die threshold voltage variation was presented. Measurement results from 960 0.18-μm 32-bit microprocessor samples verified the model's accuracy. In a leakage dominant CMOS system, it also becomes inevitable to identify techniques to reduce this variation and leakage power. In Chapter 4 the use of stacked devices to reduce system leakage power without reducing system performance was shown. Analytical model to predict the scaling nature of this stack effect and verification of the model through statistical device measurements was presented. Measurements also show reduction in threshold voltage variation for stacked devices compared to non-stack devices. Comparison of stack effect to the use of high threshold voltage or longer channel length devices for leakage reduction was discussed.

Chapter 5 of this thesis dealt with the variation in the threshold voltage of matched devices that are in the same neighborhood. The devices in that are in close proximity can be either of the same polarity or of different polarity. Matched devices of the same polarity are used as sense-amplifier input devices for low voltage swing sensing among other applications. Any mismatch in threshold voltage of this input device pair will appear as input offset resulting in degraded performance. A simple voltage-biasing scheme that reduces the mismatch between matched transistor pair of same polarity was discussed. In addition, for some digital CMOS circuits a known PMOS to NMOS drive current ratio is required either to achieve a well-defined switching threshold or to achieve equal rising and falling delays. Since the processing steps such as threshold voltage implants for the PMOS and NMOS devices are not correlated there could be significant variation between the required and achieved threshold voltages for the two device types. The short channel effects further worsen this variation. The net variation will change the drive current ratio of PMOS to NMOS devices and can affect the operation of high performance circuits that depend on a pre-determined skew between the two device types. Ability to adjust the charging and discharging currents by sensing the skew difference can alleviate this problem. In Chapter 5 current biasing schemes that

maintain the relationship between the charging and discharging currents, independent of the process skew was explained. The first current scheme that is the simplest, guarantees constant ratio between charging and discharging currents no matter the change in the relative skews of the PMOS and NMOS devices. Although this scheme maintains the relationship between charging and discharging delays, it doesn't provide constant delay as the threshold voltages vary. A true process insensitive current generation theory and circuit was described in Chapter 5. This can then be used as bias current for the charging PMOS and the discharging NMOS networks enabling a threshold voltage variation and skew variation insensitive circuit. Example circuits that benefit from these biasing schemes was presented.

6.2 Suggestions for future work

This thesis touched upon a few techniques that can be used to reduce the impact of threshold voltage variation on the behavior of CMOS circuits. With increasing variation due to worse short channel effects it will become inevitable to consider variations explicitly and rigorously in all areas of design. This will require the development of new circuit solutions that are tolerant to process variation and methodologies that combine computational efficiency of simple-minded worst-case methods, with the precision of statistical design methods. New circuit techniques that can be explored includes:

(1) Combined adaptive modulation of threshold voltage and supply voltage to minimize impact of process variation.

(2) Collaborative architectural and circuit effort to better estimate the impact of process variation on leakage, such as more accurate estimation of m in Section 4.1.

(3) Technique that accomplish both leakage reduction and process sensitivity reduction such as combining MTCMOS sleep transistor technique for leakage reduction [59] and the current biasing scheme introduced in Chapter 5.

One important aspect that was not covered in this thesis is the increasing importance of supply voltage variation. The variation in supply voltage is due to iR and L di/dt drops in the power grid with non-zero parasitic resistance (R) and non-zero loop inductance (L). Ideally, one would like to maintain the historical 10% variation is supply voltage. This is becoming harder due to increase in the current level and the rate of change of current due to faster switching as technology is scaled. In addition, the parasitic resistance and inductance have not been reducing at the same rate as the

increase properties of current flow [60]. This problem is compounded by the fact that the supply voltage is expected to scale with technology. Traditionally, passive on-die and off-die decoupling capacitors were used to filter power supply noise. Delivering 500 W at 250 mV supply voltage is a very challenging problem due to high power supply current and low power supply voltage. Recently, researchers have shown use of active on-die voltage regulation to be more efficient in controlling supply voltage variation [61, 62]. To further improve the power distribution efficiency, an on-chip voltage down regulation scheme should be explored. In this scheme the power distribution is done at higher voltages and converted locally to lower logic voltage level, thereby reducing not only the AC and DC current levels but also the percentage voltage drop in the package and global distribution grids. Such solutions will benefit from integrated magnetic inductors [63].

The challenge of power dissipation goes hand-in-hand with that of power delivery. Subsequent to the computation, power delivered to CMOS VLSI circuits gets dissipated as heat. Increase in power delivered with scaling results in increased power dissipation and higher power density [4]. In order to maintain junction temperature constant with increased power dissipation it maybe necessary to use more exotic cooling and enhanced heat spreading solutions such as carbon nano-tubes [64] and electrokinetic microchannel cooling [65]. Since sub-threshold leakage power will become more dominant with scaling, total power dissipated will have strong junction temperature dependence. Therefore, instead of keeping junction temperature constant with scaling, it might be beneficial to decrease the temperature. This temperature scaling will not only reduce leakage power but also improve drive current, interconnect resistance and reliability [66, 67]. Further optimization of device and circuits for low temperature operation can provide additional scaling benefits [19, 66, 68]. Main challenges to achieve temperature scaling for future CMOS generations include (i) understanding the relationship between junction temperature, total power, reliability, and speed and (ii) invention of low-cost integrated cooling solutions.

As a final note – to be able to successfully implement the techniques discussed in this thesis and other techniques that exist or that will become in vogue, it will be necessary to develop new computer aided design tools enabling designers to efficiently manage the risks of variation. Especially predictive models and computer aided design tools that bridge the gap between process variation and its impact on circuits blocks and the resulting impact on architectural parameters can be quite valuable.

Bibliography

[1] R. Smolan and J. Erwitt, *One Digital Day – How the Microchip is Changing Our World*, Random House, 1998.

[2] http://www.intel.com/research/silicon/mooreslaw.htm

[3] G.E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, April 19, 1965.

[4] V. De and S. Borkar, "Technology and Design Challenges for Low Power & High Performance," *Intl. Symp. Low Power Electronics and Design*, pp. 163-168, Aug. 1999.

[5] K.G. Kempf, "Improving Throughput across the Factory Life-Cycle," *Intel Technology Journal*, Q4, 1998.

[6] S. Thompson, P. Packan, and M. Bohr, "MOS Scaling: Transistor Challenges for the 21st Century," *Intel Technology Journal*, Q3, 1998.

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

[8] D. Antoniadis and J.E. Chung, "Physics and Technology of Ultra Short Channel MOSFET Devices," *Intl. Electron devices Meeting*, pp. 21-24, 1991.

[9] A. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-Power CMOS Digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473-484, Apr. 1992.

[10] Z. Chen, J. Shott, J. Burr, and J. D. Plummer, “CMOS Technology Scaling for Low Voltage Low Power Applications,” *IEEE Symp. Low Power Elec*., pp. 56-57, 1994.

[11] H.C. Poon, L.D. Yau, R.L. Johnston, D. Beecham, “DC Model for Short-Channel IGFET's,” *Intl. Electron Devices Meeting*, pp. 156-159, Dec. 1973.

[12] A. Asenov, G. Slavcheva, A.R. Brown, J.H. Davies, and S. Saini, “Increase in the Random Dopant Induced Threshold Fluctuations and Lowering in Sub-100 nm MOSFETs due to Quantum Effects: A 3-D Density-Gradient Simulation Study,” *IEEE Transactions on Electron Devices*, vol. 48, no. 4, pp. 722-729, April 2001.

[13] S. W. Sun and P. G. Y. Tsui, “Limitation of Supply Voltage Scaling by MOSFET Threshold-Voltage variation,” *Custom Integrated Circuits Conf*., pp. 267-270, 1994.

[14] D.A. Muller, T. Sorsch, S. Moccio, F.H. Baumann, K. Evans-Lutterodt, and G. Timp, “The Electronic Structure at the Atomic Scale of Ultrathin Gate Oxides,” *Nature*, vol. 399, pp. 758-761, June 1999.

[15] M. Schulz, “The End of the Road for Silicon,” *Nature*, vol. 399, pp. 729-730, June 1999.

[16] C. H. Lee, S. J. Lee, T. S. Jeon, W. P. Bai, Y. Sensaki, D. Roberts, and D. L. Kwong, “Ultra Thin ZrO(2) and Zr(27)Si(10)O(63) Gate Dielectrics Directly Prepared on Si-Substrate by Rapid Thermal Processing,” *SRC Techcon*, pp. 46, Sep. 2000.

[17] N. R. Mohapatra, M. P. Desai, S. Narendra, and V. R. Rao, “The Impact of High-K Gate Dielectrics on Sub 100 nm CMOS Circuit Performance,” *IEEE Transactions on Electron Devices*, To be published, 2002.

[18] J. Lee, G. Tarachi, A. Wei, T. A. Langdo, E. A. Fitzgerald, D. Antoniadis, “Super self-aligned double-gate (SSDG) MOSFETs utilizing oxidation rate difference and selective epitaxy,” *Intl. Electron Devices Meeting*, pp. 71-74, 1999.

[19] I. Kohno, T. Sano, N. Katoh, and K. Yano, “Threshold Canceling Logic (TCL): A Post-CMOS Logic Family Scalable Down to 0.02 µm,” *Intl. Solid-State Circuits Conf*., pp. 218-219, 2000.

[20] T. Kuroda, T. Fujita, S, Mita, T. Nagamatsu, S. Yoshioka, K. Suzuki, F. Sano, M. Norishima, M. Murota, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, “A 0.9-V, 150-MHz, 10-mW, 4-mm^2, 2-D Discrete Cosine Transform core Processor with Variable Threshold-Voltage (VT) Scheme,” *IEEE J. Solid-State Circuits*, vol. 31, pp. 1770-1779, Nov. 1996.

[21] M. Miyazaki, H. Mizuno, and K. Ishibashi, “A Delay Distribution Squeezing Scheme with Speed-Adaptive Threshold-Voltage CMOS (SA-Vt CMOS) for Low Voltage LSIs,” *Intl. Symp. Low Power Electronics and Design*, pp. 48-53, Aug. 1998.

[22] S. Narendra, D. Antoniadis, and V. De, “Impact of Using Adaptive Body Bias to Compensate Die-to-die Vt variation on Within-die Vt variation,” *Intl. Symp. Low Power Electronics and Design*, pp. 229-232, Aug. 1999.

[23] M. Miyazaki, G. Ono, T. Hattori, K. Shiozawa, K. Uchiyama, and K. Ishibashi, “A 1000-MIPS/W Microprocessor using Speed Adaptive Threshold-Voltage CMOS with Forward Bias,” *Intl. Solid-State Circuits Conf.*, pp. 420-421, 2000.

[24] V. De, “Forward Biased MOS Circuits,” *United States Patent*, Patent number: 6,166,584, Filed: June 1997, Issued: Dec. 2000.

[25] C. Wann, J. Harrington, R. Mih, S. Biesemans, K. Han, R. Dennard, O. Prigge, C. Lin, R. Mahnkopf, and, B. Chen, “CMOS with Active Well Bias for Low-Power and RF/Analog Applications,” *Symp. on VLSI Technology*, pp. 158-159, 2000.

[26] S. Narendra, S. Borkar, V. De, D. Antoniadis, and A. Chandrakasan, “Scaling of Stack Effect and its Application for Leakage Reduction,” *Intl. Symp. Low Power Electronics and Design*, pp. 195-200, Aug. 2001.

[27] R. Kraus, “Analysis and reduction of sense-amplifier offset,” *IEEE J. Solid-State Circuits*, vol. 24, no. 4, pp. 1028-1033, Aug. 1989.

[28] S. Narendra, D. Klowden, and V. De, “Sub-1 V Process Compensated MOS Current Generation without Voltage Reference,” *Symp. on VLSI Circuits*, pp. 143-144, 2001.

[29] Y.P. Tsividis, *Operation and Modeling of The MOS Transistor*, McGraw Hill, New York, 1987.

[30] H.C. Poon et al., *Intl. Electron Devices Meeting*, pp. 156-159, 1973.

[31] K.K. Ng, S.A. Eshraghi, and T.D. Stanik, “An improved generalized guide for MOSFET scaling,” *IEEE Transactions on Electron Devices*, vol. 40, pp. 1895-1897, Oct. 1993.

[32] K. Bowman, S. Duvall, and J. Meindl, “Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution”, *Intl. Solid-State Circuits Conf*., pp. 278-279, 2001.

[33] A. Keshavarzi, S. Narendra, B. Bloechel, S. Borkar, and V. De, “Forward Body Bias for Microprocessors in 130nm Technology Generation and Beyond,” *Submitted for review, 2002 Symposium on VLSI circuits*.

[34] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, “Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage,” *Intl. Solid-State Circuits Conf*., Paper 25.7, 2002.

[35] S. Narendra et.al., “16.4 1.1V 1GHz Communications Router with On-Chip Body Bias in 150nm CMOS,” *Intl. Solid-State Circuits Conf*., Paper 16.4, 2002.

[36] M. Haycock et.al., “3.2 GHz 6.4Gb/s per Wire Signaling in 0.18mm CMOS,” *Intl. Solid-State Circuits Conf*., pp. 62-63, 2001.

[37] R. Nair et.al., “A 28.5 GB/s CMOS Non-Blocking Router for Terabits/s Connectivity between Multiple Processors and Peripheral I/O Nodes,” *Intl. Solid-State Circuits Conf*., pp. 224-225, 2001.

[38] Y. Oowaki et.al., “A Sub-0.1μm Circuit Design with Substrate-over-Biasing,” *Intl. Solid-State Circuits Conf*., pp. 88-89, 1998.

[39] H. Banba et.al., “A CMOS Band-gap Reference Circuit with Sub-1V operation,” *Symp. on VLSI Circuits*, pp. 228-229, 1998.

[40] S. Vangal et.al., “5GHz 32-bit Integer Execution Core in 130nm Dual-Vt CMOS,” *Intl. Solid-State Circuits Conf*., Paper 25.2, 2002.

[41] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, V. De, "Effectiveness of reverse body bias for leakage control, in scaled dual Vt CMOS ICs," *Intl. Symp. Low Power Electronics and Design*, pp. 207-212, Aug. 2001.

[42] Y. Ye, S. Borkar, and V. De, "A Technique for Standby Leakage Reduction in High-Performance Circuits," *Symp. of VLSI Circuits*, pp. 40-41, 1998.

[43] J.P. Halter and F. Najm, "A gate-level leakage power reduction method for ultra-low-power CMOS circuits," *Custom Integrated Circuits Conf.*, pp. 475-478, 1997.

[44] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks," *Intl. Symp. Low Power Electronics and Design*, pp. 239-244, 1998.

[45] L. Su, R. Schulz, J. Adkisson, K. Beyer, G. Biery, W. Cote, E. Crabbe, D. Edelstein, J. Ellis-Monaghan, E. Eld, D. Foster, R. Gehres, R. Goldblatt, N. Greco, C. Guenther, J. Heidenreich, J. Herman, D. Kiesling, L. Lin, S-H. Lo, McKenn, "A high-performance sub-0.25μm CMOS technology with multiple thresholds and copper interconnects," *Intl. Symp. on VLSI Technology, Systems, and Applications*, pp. 18-19, 1998.

[46] D. T. Blaauw, A. Dharchoudhury, R. Panda, S. Sirichotiyakul, C. Oh, and T. Edwards "Emerging power management tools for processor design," *Intl. Symp. Low Power Electronics and Design*, pp. 143-148, 1998.

[47] A. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits*, IEEE Press, pp. 46-47, 2000.

[48] Z. Liu, C. Hu, J. Huang, T. Chan, M. Jeng, P. Ko, and Y. Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFET's," *IEEE Transactions on Electron Devices*, vol. 40, no. 1, pp. 86-95, January 1993.

[49] Y. Taur, "CMOS Scaling beyond 0.1μm: how far can it go?" *Intl. Symp. on VLSI Technology, Systems, and Applications*, pp. 6-9, 1999.

[50] S. Tyagi, M. Alavi, R. Bigwood, T. Bramblett, J. Bradenburg, W. Chen, B. Crew, M. Hussein, P. Jacob, C. Kenyon, C. Lo, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen, L. Rumaner, R.

Schweinfurth, S. Sivakumar, M. Stettler, S. Thompson, B. Tufts, J. Xu, S. Yang, and M. Bohr, "A 130 nm Generation Logic Technology Featuring 70 nm Transistors, Dual Vt Transistors and 6 layers of Cu Interconnects," *Intl. Elec. Devices Meeting*, pp. 567-570, December 2000.

[51] D. Dobberpuhl, "The Design of a High Performance Low Power Microprocessor," *Intl. Symp. Low Power Electronics and Design*, pp. 11-16, 1996.

[52] E. Vittoz, "The Design of High-Performance Analog Circuits on Digital CMOS Chips," *IEEE J. Solid-State Circuits*, pp. 657-665, June 1985.

[53] H. Banba, H. Shiga, A. Umezawa, T. Miyaba, T. Tanzawa, S. Atsumi, and K. Sakui, "A CMOS band-gap reference circuit with sub 1 V operation," *Symp. on VLSI Circuits*, pp. 228-229, 1998.

[54] http://public.itrs.net/files/1999_SIA_Roadmap/ORTC.pdf.

[55] E. Vittoz et al., "CMOS analog integrated circuits based on weak inversion operations," *IEEE J. Solid-State Circuits*, pp. 224-231, June 1977.

[56] H.J. Oguey and D. Aebischer, "CMOS current reference without resistance," *IEEE J. Solid-State Circuits*, pp. 1132-1135, July 1997.

[57] C.H. Lee and H.J. Park, "All-CMOS temperature independent current reference," *Electronics Letter*, pp. 1280-1281, July 1996.

[58] S. Yang, S. Ahmed, B. Arcot, R. Arghavani, P. Bai, S. Chambers, P. Charvat, R. Cotner, R. Gasser, T. Ghani, M. Hussein, C. Jan, C. Kardas, J. Maiz, P. McGregor, B. McIntyre, P. Nguyen, P. Packan, I. Post, S. Sivakumar, J. Steigerwald, "A high performance 180 nm generation logic technology," *Intl. Elec. Devices Meeting*, pp. 197-200, Dec. 1998.

[59] S. Mutoh et al, "1-V power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS," *IEEE J. Solid-State Circuits*, pp. 847-854, Aug. 1995.

[60] P. Larrson, "Power Supply Noise in Future ICs: A Crystal Ball Reading," *Custom Integrated Circuits Conference*, pp. 467-474, 1999.

[61] M. Ang, R. Salem, and A. Taylor, “An On-chip Voltage Regulator using Switched Decoupling Capacitors”, *Intl. Solid-State Circuits Conf.*, pp. 438-439, 2000.

[62] D. Takashima, Y. Oowaki, S. Watanabe, and K. Ohuchi, “Noise Suppression Scheme for Gigabit-Scale and Gigabyte/s Data-Rate LSI’s”, *IEEE J. Solid-State Circuits*, vol. 33, pp. 260-267, Feb. 1998.

[63] D. Gardner et al., “High frequency (GHz) and low resistance integrated inductors using magnetic materials,” *Interconnect Technology Conference*, pp. 101-103, 2001.

[64] J. Hone, B. Batlogg, Z. Benes, A. T. Johnson, and J. E. Fischer, “Quantized Phonon Spectrum of Single-Wall Carbon Nanotubes”, *Science*, vol. 289, pp. 1730-1733, Sep. 2000.

[65] Prof. Ken Goodson’s research at Stanford. http://www.stanford.edu/group/microheat/hex.html

[66] I. Aller, K. Bernstein, U. Ghoshal, H. Schettler, S. Schuster, Y. Taur, and O. Terreiter, “CMOS Circuit Technology for Sub-Ambient Temperature Operation”, *Intl. Solid-State Circuits Conf.*, pp. 214-215, 2000.

[67] U. Ghoshal and R. Schmidt, “Refrigeration technologies for Sub-Ambient Temperature Operation of Computing Systems”, *Intl. Solid-State Circuits Conf.*, pp. 216-217, 2000.

[68] K. Jackson, *Optimal MOSFET Design for Low Temperature Operation*, MIT EECS Doctoral Thesis, 2001.