Standard cell library design for sub-threshold operation

Citation for published version (APA):
Liu, B. (2014). *Standard cell library design for sub-threshold operation*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Electrical Engineering]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR782367

DOI:
10.6100/IR782367

Document status and date:
Published: 01/01/2014

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Standard Cell Library Design for Sub-threshold Operation

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. C.J. van Duijn, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op donderdag 11 december 2014 om 16:00 uur

door

Bo Liu

geboren te Harbin, China

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. A.C.P.M. Backx
1e promotor:	prof.dr. J. Pineda de Gyvez
2e promotor:	prof.dr. K.G.W. Goossens
leden:	prof.dr.ir. P.G.M. Baltus
	prof.dr. W. Dehaene (KU Leuven)
	dr. S. Hamdioui (TU Delft)
adviseur(s):	dr. H. Jiao
	dr. M. Ashouei (IMEC-NL)

This work is supported by Holst Centre/imec-nl and Eindhoven University of Technology

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-3737-2

ACKNOWLEDGEMENT

I wish to thank various people for their contributions to my PhD study and beyond. First thanks is reserved for dr. Yifan He, who enlightened me to explore in the world of digital circuit and who set a great example to dedicate fully in this field. I am also particularly grateful for the support and guidance given by Maurice Meijer, Cas Groot and Leo Sevat during my master graduation project at NXP. I would like to express my thankfulness to Bala Asirvatham, Alessio Filippi, Robert Rutten Marino Strik, Maarten Vertregt and Leo Warmerdam for giving me the chance to know more about NXP. The encouragement from them all was a significant factor leading to my decision to follow the PhD program as the next step.

These five years of PhD study at the Technical University of Eindhoven and Holst Centre was filled with excitement and appreciations. I would like to thank Jos Huiken for fitting me into the Ultra low power DSP group, Jun Zhou for the meaningful discussions that shed light on quite a few doubts I had in the beginning, as well as Hans Giesene, my office buddy, for all the idea exchanges. I also wish to acknowledge the help provided by all the people from Holst Centre for my PhD projects, Christian Bachmann, Benjamin Busze, Arjan Breeschoten, Tobias Gemmeke, Harmke de Groot, Mario Konijnenburg, Gert-Jan van Schaik, Jan Stuijt and Rik van de Wiel. I am particularly grateful for Maryam Ashouei, my daily supervisor for the PhD program, to offer her time and support so generously that contributed substantially to the progress of the study and my personal development.

I would like to use this opportunity to express my appreciation to the defense committee members, prof.dr.ir Ton Backx, prof.dr. Anton Tijhuis, prof.dr.ir. Peter Baltus, prof.dr. Wim Dehaene, prof.dr. Kees Goossens, dr. Said Hamdioui and dr. Hailong Jiao, for the review and feedback on the defense.

I am indebted to the faculty, staff, and graduate students of the Electronic system group for their help and support at all times. My special appreciation is to dr. Dongrui She for coding and scripting support, to almost dr. Hao Gao for the valuable discussions, to Bindi Wang for help offered in finalizing the thesis and to dr. Hamid Pourshaghaghi and Sebastian Londono for completing the first paper in my PhD time.

I would also like to acknowledge the support and encouragement provided by my dear family and friends during these years of PhD study, especially to those friends for constantly asking me the awkward question: "When will you graduate?" Certainly most people mentioned above and the person mentioned below became good friends of mine as a souvenir from the PhD study. As an important part of the souvenir, Yue Cui, without you I would not have a happy PhD life for the past five years.

I leave it to the last, special thanks to my friend Jose, as known as prof. dr. J. Pineda de Gyvez, without whom I would not be where I am today. I will always remember that the goal of a PhD is for GLORY as he says to me and illustrates to me.

Finally, I want to express my sincere appreciation to Edaboard, Google, Cadence manual database and Synopsys manual database.

Standard Cell Library Design for Sub-threshold Operation

Scaling the voltage to the sub-threshold region is a convincing technique to achieve low power in digital circuits. The problem is that process variability severely impacts the performance of circuits operating in the sub-threshold domain. Among other reasons, this mainly stems from the fact that sub-threshold current follows a widely spread Log-Normal distribution.

There are many interesting aspects to explore in the sub-threshold digital design. This thesis is dedicated to optimizing a standard cell library at 0.3V. The reason why we start with optimizing the standard cell library is that, standard cells (normally provided by the foundry) are the basic elements of digital circuits yet easily to be overlooked. However, the standard cell library is normally designed for super-threshold operations. The timing and power features of the standard cell library at sub-threshold are not optimized. What is worse, in sub-threshold region, the process variations severely impact the performance of the circuit. When designing digital circuits, the process variations need to be considered. Therefore, a sub-threshold standard cell library optimization methodology is needed to allow designers to optimize standard cells at different voltages in sub-threshold region at various technology nodes with the consideration of process variations.

Different from the previous works in sub-threshold region design, the main contributions of this work include: a variation aware sizing methodology to optimize standard cells for sub-threshold operations; based on the proposed methodology, custom standard cell libraries with over 100 standard logic cells have been developed at 90nm and 40nm technology nodes; test chips have been designed to provide silicon measurement results for methodology verification. Our optimization strategy relies on balancing the mean currents of the N and P network based on statistical formulations. The equivalent N and P networks are derived based on the best and worst case transition times. The slack available in the best-case timing arc is reduced by using smaller transistors on that path, while the timing of the worst-case timing arc is improved by using bigger transistors. The optimization is done such that the overall area remains constant with regard to the area before optimization. Two sizing styles are applied, one based on both transistor width and length tuning, and the other one based on width tuning only.

The optimization methodology has been evaluated on various sets of benchmark circuits and through silicon prototypes. Without loss of generality, when comparing results to the use of a super-threshold library characterized for low voltage operation in CMOS 40nm, we observed about 50% power delay product improvement on the benchmarks, and about 50% lower dynamic energy consumption in the silicon prototypes through the use of our sub-threshold library.

TO TELL A STORY,

START WITH A SIMPLE SENTENCE

AND WRITE IT REAL

CONTENTS

1 INTRODUCTION 1

1.1 WHY SUB-THRESHOLD 1
1.2 CHALLENGES IN SUB-THRESHOLD DESIGN 1
1.3 PROBLEM STATEMENT 4
1.4 CONTRIBUTIONS 6
1.5 THESIS OVERVIEW 6

2 TRANSISTOR LEVEL DESIGN TRADE-OFFS IN SUB-THRESHOLD ... 7

2.1 NMOS TRANSISTOR BEHAVIOR 10
2.2 PMOS TRANSISTOR BEHAVIOR 21
2.3 CONCLUSIONS OF CHAPTER 2 28

3 STANDARD CELL TRANSISTOR SIZING IN SUB-THRESHOLD 29

3.1 STACKED TRANSISTOR BEHAVIOR 29
3.2 PARALLEL TRANSISTOR BEHAVIOR 39
3.3 CONCLUSIONS OF CHAPTER 3 46

4 STANDARD CELL SIZING OPTIMIZATION IN SUB-THRESHOLD .. 47

4.1 SIZING METHODOLOGY 47
4.2 SIZING OPTIMIZATION FOR PUN AND PDN 50
4.3 CONCLUSIONS OF CHAPTER 4 75

5 STANDARD CELL LIBRARY IMPLEMENTATION AND COMPARISONS 77

5.1 LAYOUT CHOICES 77
5.2 LIBRARY CHARACTERIZATION 81
5.3 LIBRARY CELL COMPARISON 86
5.4 VERIFICATION OF STANDARD CELL OPTIMIZATION IN SILICON 94

6 SUB-THRESHOLD DESIGN USING THE OPTIMIZED LIBRARIES .. 107

6.1 LOGIC SYNTHESIS BENCHMARKING: EVALUATION OF LOW VOLTAGE LIBRARIES ON ACTUAL DESIGNS 107
6.2 SILICON RESULTS 112

7 CONCLUSIONS AND FUTURE WORK 121

7.1 CONCLUSIONS 121

7.2 FUTURE WORK 122

APPENDIX 125
BIBLIOGRAPHY 127
LIST OF PUBLICATION 135
ABOUT ME 137

LIST OF TABLES

Table 1 Current variation in series-connected transistors.....32
Table 2 Mean current of parallel connected transistors.....40
Table 3 Inverter sizing example at 90nm.....57
Table 4 Two-input NAND gate input pattern and current analysis.....60
Table 5 Setup time corner simulation summary.....74
Table 6 Comparison of inverter chains with two growing methods.....79
Table 7 Comparison of critical path delays of ISCAS circuit C6288 94
Table 8 Frequency statistics of 16-stage inverter chain [0.3V, 0.5V] 98
Table 9 Frequency statistics of 40 64 80-stage inverter chain at 0.35V.....98
Table 10 2MHz case study data summary 104
Table 11 Frequency of the NAND-NOR and XOR chains [0.3V, 1.1V] 105
Table 12 ISCAS benchmark circuit comparison 108
Table 13 ITC benchmark circuit synthesis results for maximum speed.....110
Table 14 ITC benchmark circuit synthesis results (equal area) 111
Table 15 ITC benchmark circuit synthesis results (same delay).....111

CHAPTER 1

INTRODUCTION

1.1 Why Sub-threshold

Due to the rapid growth of battery-operated portable applications such as personal digital assistants, cellular phones, medical applications, wireless receivers, and other portable communication devices, the demand for power sensitive design is expanding significantly [1]. Well-known methods of low power design (for example voltage scaling, switching activity reduction, architectural techniques of pipelining and parallelism, Computer-Aided Design (CAD) techniques of device sizing, interconnect, and logic optimization) may not be sufficient in many applications such as portable computing gadgets or medical electronics, where ultra low power consumption with medium frequency of operation (tens to hundreds of MHz) is the primary requirement. To cope with this insufficiency, the design of digital sub-threshold logic has been investigated [2-8]. In the sub-threshold region, the sub-threshold leakage current is used for computation. Operating in the sub-threshold region results in a quadratic reduction in dynamic power at the cost of performance degradation. Considering the frequency requirements of portable computing devices, design techniques based on sub-threshold logic have gained a wide interest in recent years.

1.2 Challenges in Sub-threshold Design

In [9], the advantages and challenges of design techniques making use of sub-threshold logic have been summarized based on simulation results in CMOS 90nm and 65nm technology nodes. The advantages are static power, dynamic power, and short circuit power savings. The low power consumption benefit of working in the sub-threshold region does not come for free, however. The cost is performance degradation and increased sensitivity to process variation when compared to super-threshold operations. These are two of the main problems that need to be solved before crossing the barrier of optimizing operations in the sub-threshold region.

The first challenge in sub-threshold design is the relatively low current flow that results in low frequency of operation. In [9], it is shown that sub-threshold logic is only relevant for systems and components operating at up to tens of MHz in a 65nm CMOS process. The frequency degradation of sub-threshold logic is only a challenge for high speed applications. In some applications, the use of both sub-threshold and super-threshold logic provides means to bypass these limitations. Furthermore, lowering the system supply voltage for different use-case modes, such as low performance mode and always-on mode, is another possibility for using sub-threshold logic.

The other main challenge in sub-threshold design, discussed in [1, 10-18], is the impact of process variations. The sub-threshold current is exponentially dependent on the transistor's threshold voltage. Accordingly, process variations that substantially affect the threshold voltages can cause a large variance in the behavior of sub-threshold circuits. In extreme cases, this can even cause circuits to malfunction. Special means, therefore, need to be taken to deal with this issue.

The choice of design sign-off points is also a major challenge in sub-threshold design. This challenge can be discussed from two aspects: temperature variation and cross corner variation. In the super-threshold region, a rise in temperature generally slows down circuits due to carrier mobility degradation. However, a rise in temperature causes a fall in threshold voltage, which exponentially increases the sub-threshold current. At a certain temperature, this increase overtakes the carrier mobility degradation. Sub-threshold circuits, therefore, become faster with increased temperature. On the other side of the temperature scale, cooling down a circuit not only increases carrier mobility but also minimizes sub-threshold leakage. At very low temperatures, the drain leakage becomes so minimal that the temperature-insensitive gate leakage current becomes the dominant leakage current. A further drop in temperature leads to negligible changes of leakage current. In sub-threshold designs, the scaling behavior among distinct signoff corners is different from super-threshold designs due to the impact of process and temperature variations. In the sub-threshold region, the behaviors of digital cells have problems in SF (slow NMOS fast PMOS) and FS (fast NMOS slow PMOS) corners due to the unbalanced PMOS and NMOS transistors. The difference between the high and low noise margin at those two corners is larger than the other three process corners (TT: typical NMOS typical PMOS, SS: slow NMOS slow PMOS, FF: fast

NMOS fast PMOS), which leads to a very low low-noise margin in the FS corner and a low high-noise margin in the SF corner.

Though challenges exist, sub-threshold design still commands wide interest to the digital design community. In [11], the energy efficiency of sub-threshold design is analyzed considering process variations, and a method is proposed to analyze the impact of the process variations on the optimal energy efficiency point. In [3], a FFT processor working at 180mV is proposed. Special circuit design considerations are employed to maintain the functionality of the standard cells for ultra-low voltage operation. Gate width sizing is used to improve the noise margin at different corners. Furthermore, the number of stacked devices is limited to reduce the series leakage dissipation. In [7], the authors present a design technique for (near) sub-threshold operation that achieves ultra low energy dissipation at throughputs of up to 100 MB/s (suitable for digital consumer electronic applications). A threshold voltage balancer structure is used to compensate the threshold mismatch between the PMOS and NMOS transistors. Finger-structured transistors are employed to improve current drivability. Furthermore, a sub-threshold standard cell selection procedure is described for higher reliability. In [19], another FFT processor core is presented with different design technologies. Additional pipeline latches are used to improve the performance and energy savings in the sub-threshold region. The logic depth is also deeply reduced for energy reduction and supply voltage minimization. At 0.27V, the FFT core operates at 30 MHz compared to typical ultra low voltage application working at tens of kilo hertz. In [20-24], the transmission gate structure is widely used in different sub/near-threshold circuit applications for variation resilient purposes. In [23], LVT devices with transmission gate structure are used to achieve MHz performance and sub-pJ energy consumption in sub-threshold region. In [21, 22, 24], pipelining and time borrowing design techniques are used for sub-threshold DSP design. In [20], a JPEG encoder with deeply pipelined architecture is fabricated in CMOS 40nm technology. All the logic gates are implemented based on a differential transmission gate structure to guarantee a variation resilient design with 5MHz performance at 210mV supply voltage.

There are other works that have contributed to the development of sub-threshold design, mainly related to sub-threshold transistor sizing and library characterization. In [25, 26], the authors calculate the optimum supply voltage to minimize energy consumption. They also claim that, theoretically, minimum sized cells are optimal for energy reduction. However, in this thesis, it is shown

that under speed constraints and when process variability is taken into account, this is not always the case. In[27], the authors explain the benefit of technology choices, power supply scaling, and body bias adaptability for circuits working in the sub-threshold region. It is implied that standard cell timing could be improved using the mentioned design techniques. The concept of sub-threshold logical effort for complex gate sizing is presented in [16]. Particularly interesting is a closed form current equation derived for stacked transistors in relation to other transistors in the same stack. In [28], the reverse short channel effect in the sub-threshold region is used for transistor sizing optimization, where the channel length is increased to have an optimal threshold voltage. This causes the transistors to have a higher current, to be less sensitive to random variations, and to have a smaller area. With a higher current and a lower gate capacitance, both the delay and power consumption are reduced. Furthermore, in [28], the channel lengths are increased to achieve the maximum currents for both NMOS and PMOS transistors. In [6], a standard cell library in 65nm CMOS technology is presented where, by upsizing the channel length of all transistors in each given cell, the energy per operation is reduced by about 15%. In this thesis, the standard cells are tuned individually, with various length and width selections, to have balanced transition currents. In [13], a searching algorithm is presented. The algorithm is based on multiple objectives through a free space search to optimize one cell. The approach is exhaustive, and the searching effort is huge for a complete library. In [29], a 45nm standard cell library optimized for 0.35V is proposed. The proposed PMOS-to-NMOS transistor ratio optimization is based on the optimal energy-delay product, not on balanced rise and fall times. In this thesis, the rise and fall times are balanced by taking into account the effect of process variations.

With knowledge of the existing techniques [2-4, 6, 7, 9, 11, 13, 14, 16, 18, 20-25, 27-37] and the requirements of an application area, this doctoral research is formulated in the next section.

1.3 Problem Statement

In the previous section, some application areas and the challenges of the sub-threshold logic design were discussed. Among many applications, wireless sensor networks (WSN) have created the ability to know the world in every dimension [38-41]. Currently, the concept of sensor network is extended to the network around the human body. The body area network helps people to understand themselves much better than ever [42-45]. Even further developed

small sensors can be swallowed or implanted into human body [46-49]. The above developments and other low power applications have inspired the digital circuit design society to search for solutions to those low power applications. Working in the sub-threshold region provides a very promising low power feature for applications with low/medium performance requirement such as the WSN and body area networks. However, there is a need to improve design methodologies in sub-threshold operations due to the performance sensitivity to process, voltage, and temperature variations. In this thesis, the problem of energy-efficient robust sub-threshold design is approached by developing standard cells that exhibit better robustness and performance over the current state of the art.

There are two interesting aspects to explore in the field of sub-threshold digital design: the lack of optimization of timing and power features, and the impact of process variations.

- This thesis is dedicated to optimizing a standard cell library at 0.3V. The reason why the standard cell library is optimized is that although standard cells (normally provided by the foundry) are the basic elements of digital circuits, they can easily be overlooked. In the current state of the art, the standard cell library is normally designed for super-threshold operations; thus the timing and power features of the standard cell library in sub-threshold operations are not optimized.
- What is worse, in the sub-threshold region, process variations severely impact the performance of the circuit, and the current state of the art libraries cannot properly address this challenge. Obviously, when designing digital circuits, process variations need to be considered. Therefore, a sub-threshold standard cell library optimization methodology is needed to allow designers to optimize standard cells at different voltages taking into account the impact of process variations.

In addition to the main problem statements, there is another interesting point to mention: the supply voltage value of 0.3V is chosen for two reasons. First, the transistors operate in the *deep* sub-threshold region. Secondly, working at 0.3V is convenient for the connection of wireless power sources, which is highly desirable for body area networks and WSN applications. In other words, working at 0.3V provides both power reduction and frequency benefits for body area network applications.

1.4 Contributions

Different from previous works for sub-threshold design, the main contributions of this work include the following:

- A statistical formulation of the sub-threshold current suitable to design process-aware low voltage standard cells [14, 36] (Chapter 2 and Chapter 3);
- A variation-aware gate sizing methodology to optimize standard cells for sub-threshold operation [14, 37] (Chapter 4);
- The development of custom standard cell libraries with over 100 standard logic cells in both 90nm and 40nm CMOS technology nodes [30, 37, 50] (Chapter 5); and
- Test chips to generate silicon measurement results to verify the proposed standard cell libraries in actual designs (Chapter 6).

1.5 Thesis Overview

The thesis is organized as follows. The transistor design trade off in sub-threshold region is presented in Chapter 2. The transistor sizing strategies in sub-threshold region is shown in Chapter 3. Chapter 2 and Chapter 3 provide the base for the proposed methodology. The standard cell optimization methods are explained in detail in Chapter 4. The standard cell library characterization and benchmarking based on simulation and silicon measurement results are shown in Chapter 5. The comparison among different libraries based on the benchmark circuit simulation and silicon results are presented in Chapter 6.

CHAPTER 2

TRANSISTOR LEVEL DESIGN TRADE-OFFS IN SUB-THRESHOLD

The sub-threshold region, also called the weak inversion region, was first defined as the "parabolic region" by Garett and Brattain in their Metal-Insulator-Semiconductor(MIS) diode study [51]. Their studies show that as soon as some voltage is applied between the source and the drain of a MOS transistor structure, the minority carriers move by diffusion, thereby producing a drain current: *a sub-threshold current*. The weak inversion region is defined as the sub-threshold region, where the supply voltage of the transistor is below the threshold voltage in modern CMOS technology nodes.

The sub-threshold current was neglected for many years since it was at the sub microampere level. In [52], the MOS transistor current was measured at very low levels, and the exponential dependence of the drain current to the gate voltage was made evident. This paper marked the onset of the interest in ultra low power digital design.

An illustration of an n-type MOSFET operating in the sub-threshold region is shown in Figure 2.1. The drawings of the n-type MOSFET in the linear region and saturation regions, respectively, are shown Figure 2.2.

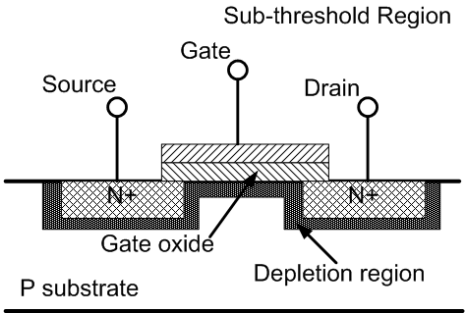


Figure 2.1 Cross section view of enhanced mode n-channel MOSFET operating in the sub-threshold region.

Compared to the linear region, and saturation region shown in Figure 2.2, when the gate to source voltage is smaller than the threshold voltage, there is no conduction channel between the drain and source terminals, and the area between the source and drain terminals is depleted. There is no inversion layer, as depicted in Figure 2.1. Considering the effect of thermal energy on the Fermi–Dirac distribution of electron energies, some of the more energetic electrons at the source terminal flow to the drain terminal through the depletion region. This electron flow results in the sub-threshold current [12, 53].

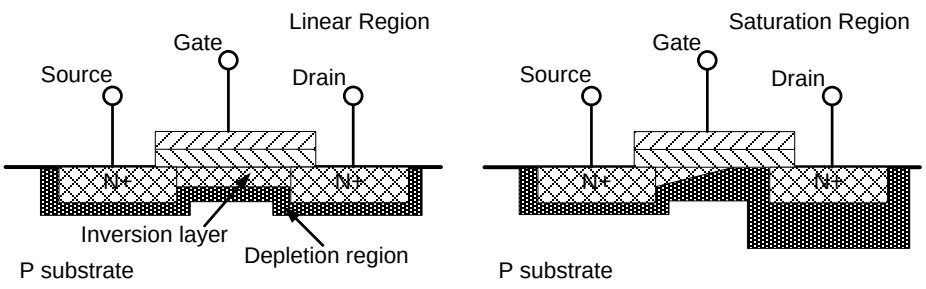


Figure 2.2 Cross section view of enhanced mode n-channel MOSFET at linear and saturation regions.

With the work of [35, 54-57], in [58], a compact model to describe the flowing drain current was developed, and it is written here as Equation (2.1). The model was further developed in detail in [59], and it is shown as Equation (2.2)

$$I_D = I_{D0}\, exp\left(\frac{V_{gs}}{nU_T}\right)\left(exp\frac{-V_s}{U_T} - exp\frac{-V_{ds}}{U_T}\right) \tag{2.1}$$

$$I_{Sub} = \mu_{eff} C_d \frac{W}{L}\left(\frac{kT}{q}\right)^2 e^{\frac{V_{gs}-V_{th}}{mkT/q}}\left(1 - e^{-\frac{V_{ds}}{kT/q}}\right) \tag{2.2}$$

where

I_{D0}	Drain current when zero biasing,
μ_{eff}	Effective mobility,
C_{ox}	Oxide capacitance,
C_d	Depletion capacitance,
W	Transistor width,
L	Transistor length,
m	$1 + C_d/C_{ox}$,

U_T	Thermal voltage (kT/q);
V_{gs}	Gate-source voltage,
V_{th}	Threshold voltage, and
V_{ds}	Drain-source voltage.

This current model (without the DIBL effect) has been shown in many research works for low voltage low power applications [7, 15, 29, 34, 60]. In this model, the exponential dependency of the sub-threshold current to the threshold voltage is clearly shown.

The current is very small in the sub-threshold region. Furthermore, process variations severely impact the performance of a transistor through threshold voltage variation. Due to the exponential relationship of the current to threshold voltage, the sub-threshold current follows a log-normal distribution. Three thousand Monte-Carlo simulations of a minimum sized NMOS transistor in a CMOS 40nm Low Power technology ($|V_{GS}| = |V_{DS}| = 0.3$V, T = 25°C) were used to obtain the statistical distributions of the threshold voltage and sub-threshold current shown in Figure 2.3. The simulation uses a BSIM4 (V4.5) model. The wide spread of the transistor current in the sub-threshold region makes it difficult to be modeled for the design of digital circuits in the sub-threshold regime. The modeling and the behavior of transistors in the sub-threshold region are addressed in the following sections.

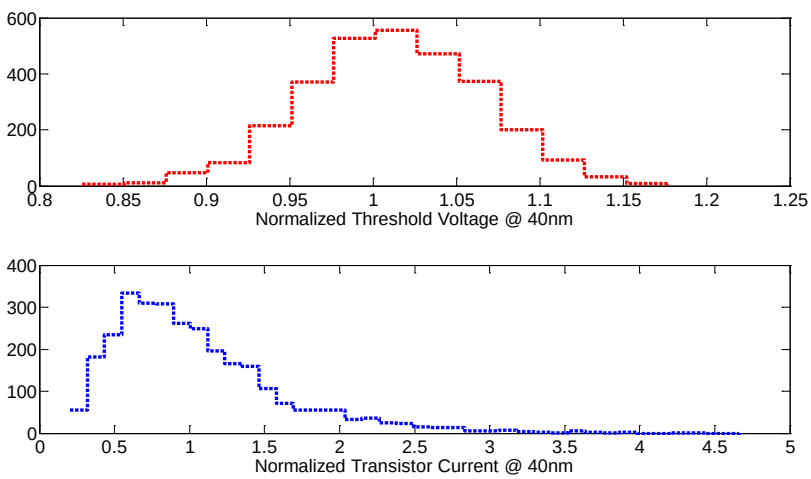


Figure 2.3 Threshold voltage and sub-threshold current distributions of a minimum sized NMOS transistor in a 40nm CMOS technology (W=0.15μm and L=40nm. $V_{GS} = V_{DS} = 0.3$V. T = 25°C).

In the sub-threshold region, the transistor behaves differently from that in the super-threshold region. In the previous section, the fact that the sub-threshold current follows a log-normal distribution has already been pointed out. To model the distribution, the mean $E[]$ and standard deviation $Std[]$ of the sub-threshold current are introduced.

Based on Equation (2.2) and [11, 61], the distribution of the sub-threshold current (shown in Figure 2.3) can be statistically modeled as,

$$E[I_{Sub}] = \mu_{eff} C_d \frac{W}{L} \left(\frac{kT}{q}\right)^2 e^{\frac{V_{gs}-E[V_{th}]}{nkT/q} + \frac{Std^2[V_{th}]}{2(nkT/q)^2}} \left(1 - e^{\frac{-V_{ds}}{kT/q}}\right)$$

$$Std[I_{Sub}] = \sqrt{\left(e^{\frac{Std^2[V_{th}]}{(nkT/q)^2}} - 1\right)}\left(E[I_{Sub}]\right) \tag{2.3}$$

where

$E[V_{th}]$	Mean value of the threshold voltage, and
$Std[V_{th}]$	Standard deviation of threshold voltage.

In [59], the channel length modulation, short-channel effect, and narrow-channel effect are identified to be related to the pinch-off voltage in field-effect transistors, which is the same concept of the threshold voltage in MOS transistors. In the next section, simulation results of transistors with different widths and lengths are shown to demonstrate how the sizes of transistors affect the sub-threshold current and, therefore, the distribution of the sub-threshold current.

2.1 NMOS Transistor Behavior

In this section, a series of NMOS transistors with different sizes in a CMOS 40nm technology are simulated. Monte Carlo simulations with 3000 iterations based on a commercial foundry model are used to show the influence of the sizes of transistors on the distributions of the sub-threshold current and threshold voltage.

2.1.1 Width tuning

The inverse narrow width effect causes a non-proportional relationship between the drain current and the transistor width: as the transistor width is increased from the minimum width, the drain current increases more slowly than expected due to the increased threshold voltage. The inverse narrow

width effect is shown in [10]. It is pointed out that the inverse narrow width effect increases the threshold voltage rapidly when the transistor width increases from the minimum value to 0.5μm in 90nm, 65nm, and 40nm technology nodes [10]. This effect becomes significant at low supply voltages, especially in the near/sub-threshold region, because the sub-threshold current is exponentially dependent on the threshold voltage. As a result, the transistor width has to be significantly increased to obtain a proportional increase of the drain current. Wider transistors, however, lead to larger area, higher power consumption, and longer delay.

Based on the work of [10], the statistical features of the narrow width effects are investigated. NMOS transistors with 40nm channel length in a CMOS 40nm technology were simulated (V_{GS} = V_{DS} = 0.3V, T = 25°C) and the results are shown in Figure 2.4. The distribution parameters, ($E[V_{th}]$ and $Std[V_{th}]$ of V_{th}), are derived from the Monte Carlo simulation. Furthermore, two values $E[V_{th}] + 3 * Std[V_{th}]$ and $E[V_{th}] - 3 * Std[V_{th}]$ are calculated to represent the upper bound and lower bound, respectively, of the V_{th} distribution.

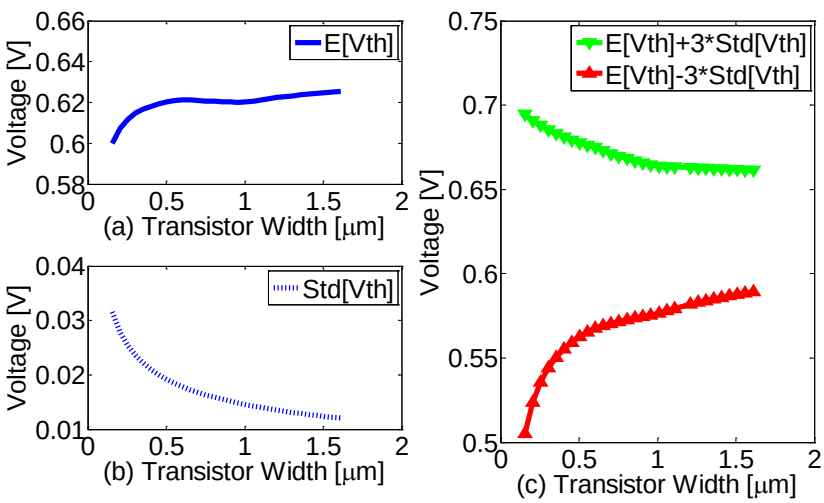


Figure 2.4 The distribution parameters of NMOS transistor threshold voltages vs. transistor width (L = 40nm. V_{GS} = V_{DS} = 0.3V. T = 25°C).

The minimum allowed transistor channel width in the simulated CMOS 40nm technology is 0.15μm. In Figure 2.4 (a), the inverse narrow width effect [10] is clearly observed when the transistor width varies from 0.15μm to

0.50μm. When the transistor width increases, $E[V_{th}]$ increases rapidly in the inverse narrow width effect range. The corresponding $Std[V_{th}]$ is shown in Figure 2.4 (b). When the transistor width increases, the $Std[V_{th}]$ decreases. In Figure 2.4 (c), the distributions of the upper bound and lower bound of the V_{th} are shown. When the transistor width is increased above 0.50μm, the inverse narrow width effect diminishes. As shown in Figure 2.4, when the width increases from 0.5μm to 1.5μm, the $E[V_{th}]$ is in the range of [0.620V, 0.625V] while the $Std[V_{th}]$ decreased from 0.020V to 0.012V. The increase of the $E[V_{th}]$ and the decrease of the $Std[V_{th}]$ lead to V_{th} spread reduction.

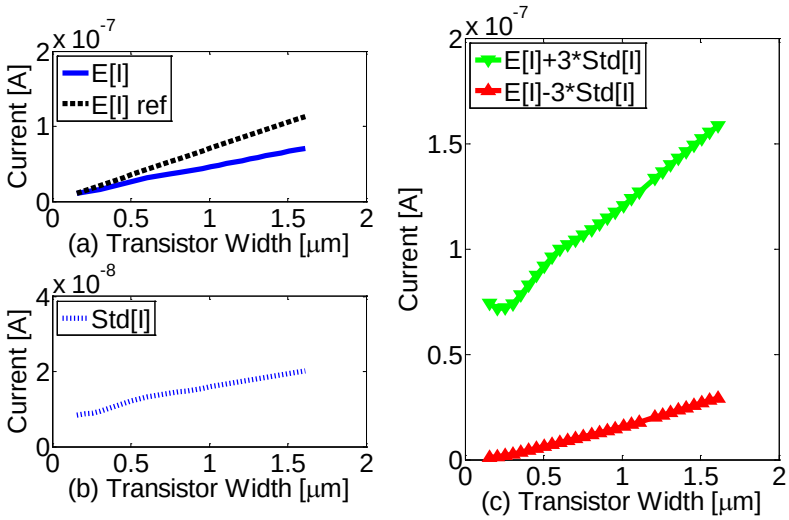


Figure 2.5 The distribution parameters of the drain currents of NMOS transistors vs. transistor width (L = 40nm. V_{GS} = V_{DS} = 0.3V. T = 25°C).

Transient analysis is used to measure the drain switching current through NMOS transistors, with 40nm channel length. The measured transistor is loaded with four copies of the same transistor at the output node. The results are shown in Figure 2.5. As shown in Figure 2.4, the distribution of the V_{th} becomes narrower when the width of the transistor is increased. From Equation (2.3), it is shown that when $E[V_{th}]$ increases and $Std[V_{th}]$ decreases (the distribution of the V_{th} is narrowed down) $E[I_{Sub}]$ decreases exponentially. However, due to the linear dependency of $E[I_{Sub}]$ on the channel width, when the channel width increases, $E[I_{Sub}]$ increases linearly. In Figure 2.5 (a) and Figure 2.5 (b), $E[I_{Sub}]$ and $Std[I_{Sub}]$ increase sub-linearly when the channel width increases. As stated in [10], if there is no inverse short channel effect, the increase of $E[I_{Sub}]$ is pro-

portional to the increase of the channel width as illustrated by the dashed line $E[I_{Sub}]\ ref$ in Figure 2.5 (a). The actual slope of $E[I_{Sub}]$ is smaller than the slope of the dashed $E[I_{Sub}]\ ref$ line due to the inverse short channel effect. In [10], it is shown that as the channel width increases, the threshold voltage increases as well, which leads to a slower increase on the drain current. It follows then that width tuning in the sub-threshold region is not as effective as in the super-threshold region to increase the drain current. The $\pm 3 * Std$ value of the drain current is shown in Figure 2.5 (c). Note that, in sub-threshold region, the current follows a lognormal distribution. In order to calculate the $E[I_{Sub}] \pm 3 * Std[I_{Sub}]$ value, the data from Monte Carlo simulation needs to be transformed to a normal distribution first. After the calculation, the natural exponential function is applied to the calculated value. The detailed steps are listed in the Appendix. The trend of $E[I_{Sub}] + 3 * Std[I_{Sub}]$, shown in green, indicates that the distribution of the upper bound of the drain current decreases initially and increases afterwards. As shown in Figure 2.5 (c), the upper bounds are equal when the channel width is 0.155μm and 0.355μm. The minimum upper bound is observed when W = 0.205μm. Another interesting point regarding Figure 2.5 (c) is that the spread of the sub-threshold current increases with increased channel width, which means wider channel width does not lead to smaller current variation.

From a "current variation" perspective, the transistor width should be minimized. From the $E[I_{Sub}]$ perspective, the linear dependency on the channel width makes increasing the channel width a preferable choice to increasing $E[I_{Sub}]$. However, note that the current increase (when the transistor width is increased) in the sub-threshold region is not as effective as in the super-threshold region.

2.1.2 Length Tuning

The length modulation effect is explained in [59]. In [28], it is shown that in the sub-threshold region, length tuning is a promising sizing method to find an optimal threshold voltage which causes the transistors to have a higher current. In this section, the length tuning effect is investigated (statistically) while the channel width of the transistor is kept as a fixed value.

In Figure 2.6, the super-threshold length tuning effect on V_{th} is shown with transistors in the 40nm technology node with a fixed channel width (0.3μm). $V_{GS} = V_{DS} = 1.1V$. In [28], the short channel effect (SCE) and reverse short chan-

nel effect (RSCE) are investigated. SCE (or V_{th} roll-off) is an undesirable phenomenon in short channel devices where V_{th} decreases as the channel length is reduced, as shown by the black arrows in Figure 2.6. The SCE is exacerbated with an increased drain induced barrier lowering (DIBL) effect. Traditionally, non–uniform HALO doping is used to reduce the DIBL effect. As a byproduct of HALO doping, a short channel device shows RSCE, where the V_{th} decreases as the transistor channel length is increased [62, 63]. The RSCE is not a major concern in conventional super-threshold designs since SCE is dominant in the minimum channel length region. However, in the sub-threshold region, RSCE becomes dominant due to the reduced DIBL effect. This leads to the monotonous decrease of $E[V_{th}]$ when the transistor channel length is increased, as shown in Figure 2.7 (a). $E[V_{th}]$ decreases rapidly when the channel length is increased from a minimum channel length of 0.04μm to 0.1μm. From 0.1μm to 0.2μm, the decreasing slope of $E[V_{th}]$ becomes smaller. The range of [0.04μm, 0.1μm] is the interesting length tuning range for V_{th} tuning. The spread between the upper and lower bounds follows the same trend as $E[V_{th}]$ in the sub-threshold region, as shown in Figure 2.7 (c). Furthermore, as shown in Figure 2.7 (b), as the channel length is increased, the area of the transistor increases, thereby reducing the variation of V_{th} [64].

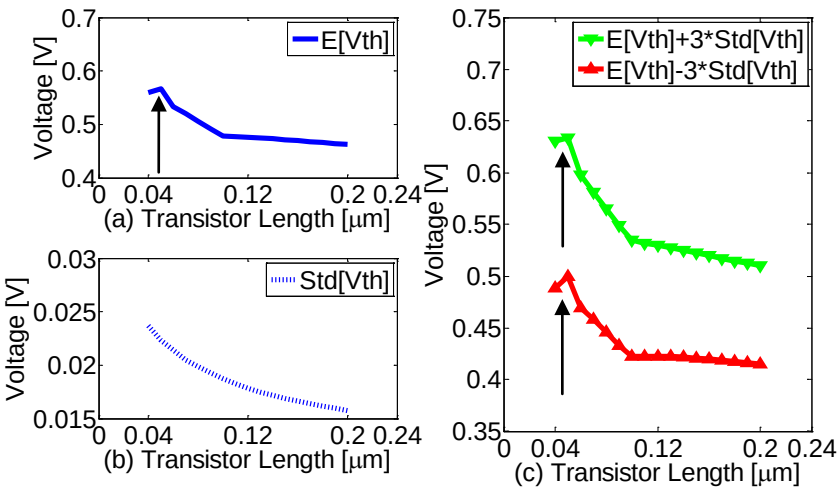


Figure 2.6 The distribution parameters of an NMOS transistor threshold voltage vs. transistor length (W = 0.3μm. $V_{GS} = V_{DS} = 1.1$V. T = 25°C).

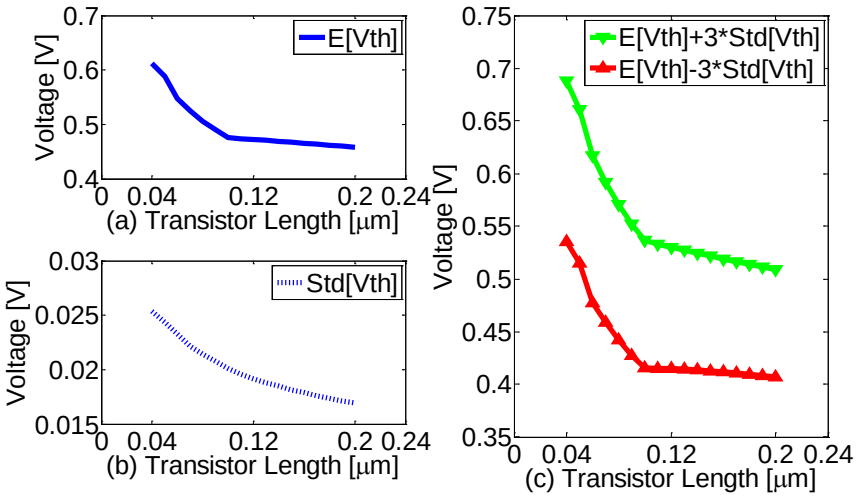


Figure 2.7 The distribution parameters of NMOS transistor threshold voltages vs. transistor length (W = 0.3μm. $V_{GS} = V_{DS} = 0.3V$. T = 25°C).

As the variation of threshold voltage is more significant in the sub-threshold region as compared to in the super-threshold region, the drain current trend changes accordingly. In the super-threshold region, the drain current is modeled with the alpha power model also shown in [65].

$$I_{D-superthreshold} = \mu_{eff} C_d \frac{W}{L} T_{fit} (V_{gs} - V_{th})^{\alpha} \tag{2.4}$$

where

T_{fit}	Technology fitting parameter, and
α	Velocity saturation index.

As it is implied from Equation (2.4), when the channel length is increased, the super-threshold drain current decreases. The mean values of the drain currents of NMOS transistors in the super-threshold region (1.1V supply voltage) are shown in Figure 2.8 (a). As indicated by the black arrow, the optimum length to achieve maximum $E[I_{D-superthreshold}]$ is at 0.05μm. In Figure 2.8 (b) and Figure 2.8 (c), the V_{th} roll-off effect in the sub-threshold regime is more pronounced compared to that in the super-threshold regime. Furthermore, as shown in Equation (2.4), when V_{th} is reduced, the drain current increases. However, when the transistor channel length is increased, the drain current

decreases more rapidly than the drain current increase caused by reducing V_{th} in the super-threshold region.

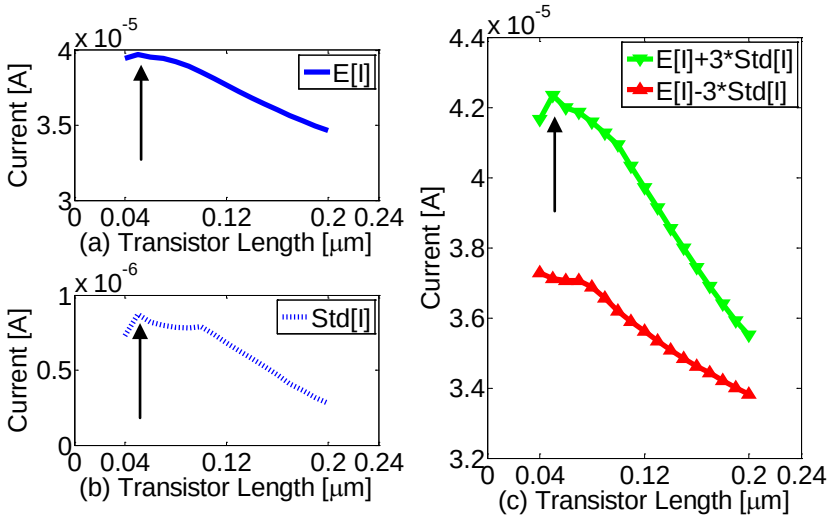


Figure 2.8 The distribution parameters of the drain currents of NMOS transistors vs. transistor length (W = 0.3μm. V_{GS} = V_{DS} = 1.1V. T = 25°C).

In the sub-threshold region, due to the exponential dependency of the drain current on V_{th}, the drain current trend at 0.3V behaves differently, as shown in Figure 2.9. $E[I_{Sub}]$ is shown in Figure 2.9 (a). Unlike the trend shown in [28], when the transistor channel length increases from 0.04μm to 0.05μm, $E[I_{Sub}]$ decreases. This is because $Std[V_{th}]$ of the transistor with minimum channel length (0.04μm) is higher than when L = 0.05μm. From 0.05μm to 0.1μm, the RSCE causes $E[I_{Sub}]$ to increase. After reaching the 0.1μm point, the increasing trend starts to bend. When the channel length is further increased, the $E[I_{Sub}]$ increases slowly. From 0.04μm to 0.1μm, $Std[I_{Sub}]$ follows the same trend as the $E[I_{Sub}]$. When the channel width is wider than 0.1μm, the $Std[I_{Sub}]$ decreases as the channel width is increased. Therefore, the spread of I_{Sub} is narrower, as shown in Figure 2.9 (c).

In the super-threshold region, the channel length is often kept at the minimum value when sizing the transistors for speed and power perspectives. Due to RSCE, the minimum channel length is not the optimal length for speed and variation tolerance in the sub-threshold region. In the sub-threshold region, when the channel length of the NMOS transistors is increased, the sub-

threshold current increases, the current spread decreases, and the transistors are more resilient to process variations.

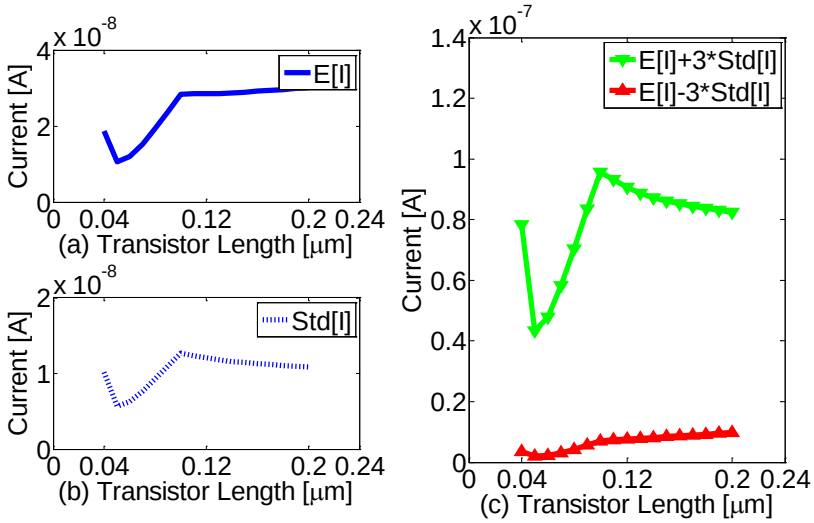


Figure 2.9 The distribution parameters of the drain currents of NMOS transistors vs. transistor length (W = 0.3μm. V_{GS} = V_{DS} = 0.3V. T = 25°C).

2.1.3 Finger Effect

In the previous sections, the width and length tuning effects on NMOS transistors are shown. In commercial CMOS integrated circuit designs, large transistors are often implemented as finger structures. There are two important parameters in figure structures: the finger size and the minimum distance between two fingers. The minimum finger sizes are determined based on the layout design rules. The minimum distance is the distance used to separate each finger to minimize the area overhead. In sub-threshold designs, finger structured transistors have a different behavior from the corresponding structures in the super-threshold region, due to the fact that transistors with larger area are more immune to process variations [64]. In this section, the effect of finger structures on NMOS transistors in the sub-threshold region is studied.

In Figure 2.10, the distribution parameters of the drain current of the fingered transistors with minimum channel length (V_{GS} = V_{DS} = 0.3V, T = 25°C) are shown. The red and green curves represent the parameters of transistors implemented with four fingers (NF4) and two fingers (NF2), respectively. The

blue curves represent the parameters of the comparison reference (NF1) where the finger structure is not used. Note that, in Figure 2.10, the X-axis is the total width of the different transistor structures. Due to the minimum channel width (0.15μm) requirements, the minimum total width of NF2 transistors is 0.30μm (2x0.15μm), and the minimum total width of NF4 transistors is 0.60μm (4x0.15μm).

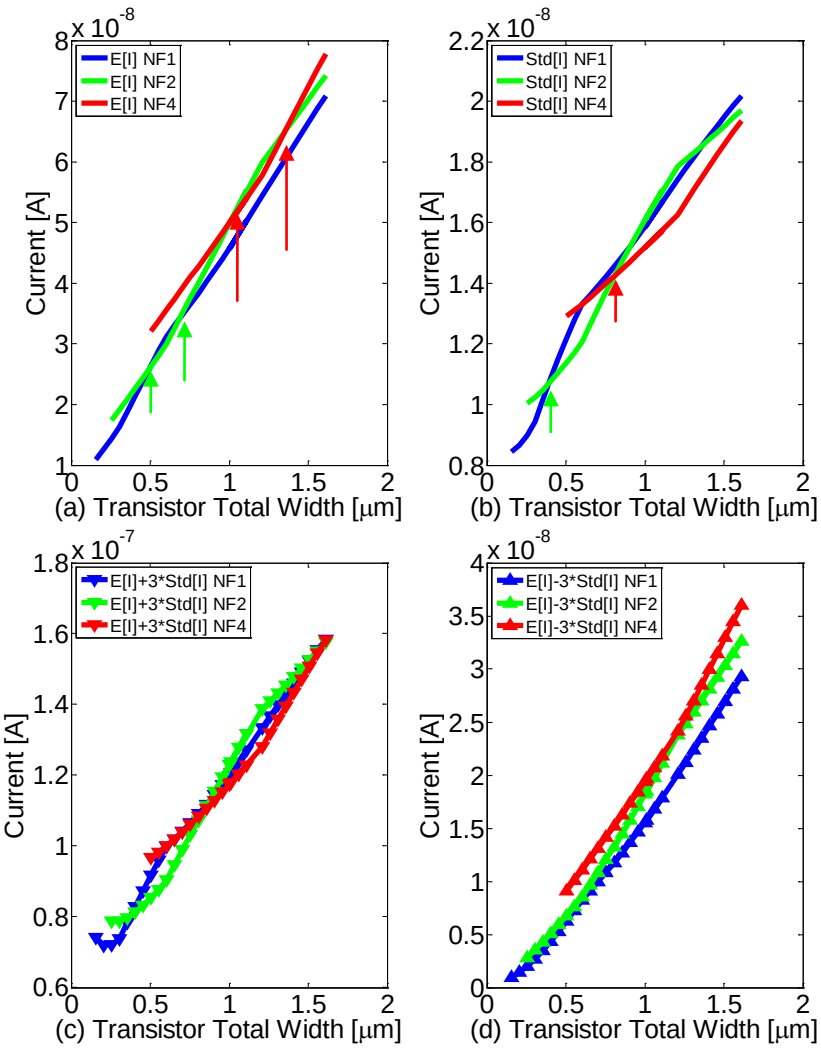


Figure 2.10 The distribution parameters of the drain currents of fingered NMOS transistors vs. transistor width (L = 40nm. $V_{GS} = V_{DS} = 0.3V$. T = 25°C).

In Figure 2.10 (a), the $E[I_{Sub}]$ of transistors with different finger structures is compared. Increasing the number of fingers has a positive influence on increasing the mean current of the transistors. The NF4 NMOS transistor has a larger mean current compared to the NF1 NMOS transistor. As indicated by the green arrows, there is an overlap region where the number of fingers is increased from one to two which does not increase the drain current. Similarly, the overlap region between NF2 and NF4 is shown by red arrows. In Figure 2.10 (b), the $Std[I_{Sub}]$ of transistors with different finger structures is shown. Although transistors NF1, NF2, and NF4 have the same total width, the $Std[I_{Sub}]$ of all three transistors shows different trends. The green and red arrows indicate where the NF2 and NF4 become a better choice for variation resilience, respectively, compared to NF1. Similarly, the spread parameters of transistors with different finger structures are plotted in Figure 2.10 (c) and Figure 2.10 (d). Similar to the trends shown in the $E[I_{Sub}]$ and $Std[I_{Sub}]$ figures, the spread parameters of different transistor structures are twisted at different regions.

The optimal choice for number of fingers, therefore, varies for different targets. The improvement of drain currents that are produced by fingered transistors with minimum channel length is very limited. $E[I_{Sub}]$ of transistors NF1, NF2, and NF4 is shown in Figure 2.11. When the channel length increases, the gap between the multi-fingered transistors and non-fingered transistors becomes larger, which means that $E[I_{Sub}]$ increases when the number of fingers is increased. Compared with the trends in Figure 2.11 (a), when the channel length is increased from 0.06μm to 0.08μm, $E[I_{Sub}]$ decreases. When the channel length is increased further to 0.1μm, $E[I_{Sub}]$ increases. This trend also confirms the trend shown in Figure 2.9. Another interesting point is to compare the starting point of all three trends in Figure 2.11 (b), Figure 2.11 (c), and Figure 2.11 (d), where the minimum total transistor width is allowed for the maximum number of fingers. It is clear that, if the minimum total width for multi-fingered transistor is allowed, using multiple fingers improves the current drivability at various lengths.

2.1.4 Conclusions of Section 2.1

In this section, the effects of width tuning, length tuning, and finger structure on the electrical characteristics of NMOS transistors operating in the sub-threshold region are studied. More specifically, the impact of inverse narrow width effect and reverse short channel effect on the NMOS transistor behavior were investigated from the current drivability and spread perspectives.

The inverse narrow width effect reduces the threshold voltage when reducing the transistor channel width in the sub-threshold region. This leads to the observation that when the channel width is increased, the drain current increase in the sub-threshold region is less than the current increase in the super-threshold region. At both the sub-threshold and super-threshold regions, transistor current has a linear dependency on channel width, which means that when the channel width of the NMOS transistor is increased, the current drivability increases and the current variation decreases.

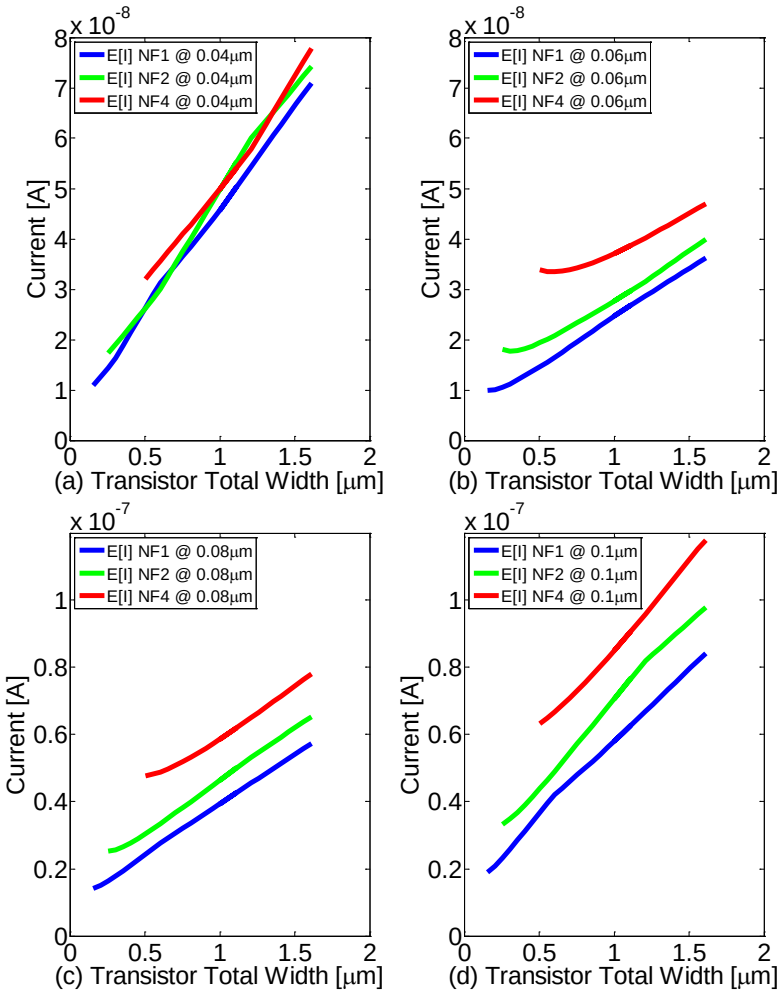


Figure 2.11 The mean currents of fingered NMOS transistors vs. transistor total width (L = 40nm. VGS = VDS = 0.3V. T = 25°C).

Length tuning has an opposite effect on current tuning in the sub-threshold region when compared to the super-threshold region due to RSCE. In super-threshold, when the channel length is increased, the transistor's drain current decreases. Alternatively, in sub-threshold, when the channel length is increased, the NMOS transistors become faster and more resilient to process variations, which makes length tuning an effective approach for cell sizing in the sub-threshold regime.

When increasing the drive strength of the NMOS transistors, increasing the number of fingers is more efficient than increasing the channel width in sub-threshold region. Multi-fingered NMOS transistors also have smaller current spread compared to non-fingered transistors, which makes finger structure with minimum finger width an attractive approach in improving the current drivability and reducing the variation in the sub-threshold region.

2.2 PMOS Transistor Behavior

Similar to the previous section, the width tuning, length tuning, and finger effects on PMOS transistors in sub-threshold region are examined here.

2.2.1 Width tuning

The inverse narrow width effect also impacts the $|V_{th}|$ and I_{Sub} of the PMOS transistors. In this section, Monte Carlo simulation results of PMOS transistors are shown to examine width tuning effects on PMOS transistors in the sub-threshold region ($|V_{GS}| = |V_{DS}| = 0.3$V. T = 25°C).

The distribution parameters of the $|V_{th}|$ of PMOS transistors are shown in Figure 2.12. As shown in Figure 2.12 (a) and Figure 2.12 (b), when the channel width is increased from the minimum value, the $E[|V_{th}|]$ increases and the $Std[|V_{th}|]$ decreases as the total area increases. In Figure 2.12 (c), the spread of $|V_{th}|$ is shown. When the transistor channel width is increased, the spread of the $|V_{th}|$ is narrowed down. Increasing the channel width has a negative impact on improving the strength of the transistor (due to the increase of $|V_{th}|$). However, the spread of the $|V_{th}|$ of wider transistors is narrower, which means that increasing the channel width renders a weak yet variation-resilient PMOS transistor. The increase of $E[|V_{th}|]$ impacts the current-width relationship in sub-threshold as stated in [10] and Figure 2.13.

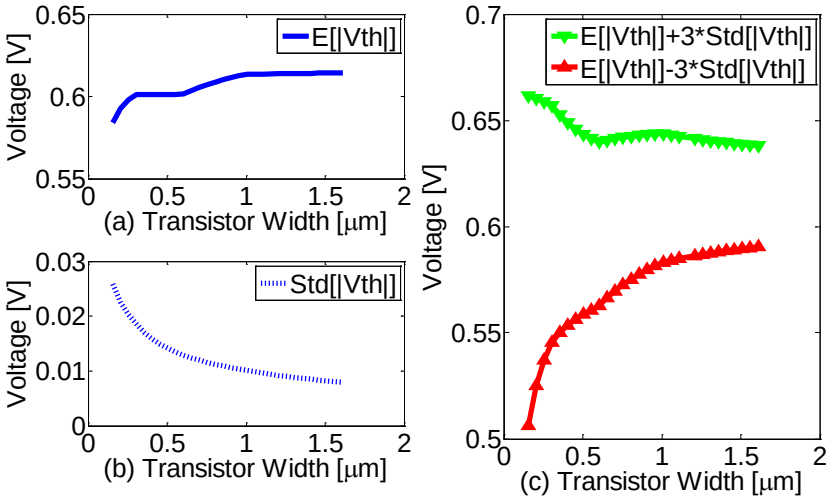


Figure 2.12 The distribution parameters of the $|V_{th}|$ of PMOS transistors vs. transistor width (L = 40nm. $|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

When the channel width is increased, the increase of V_{th} leads to a negative effect in increasing the I_{Sub}. Similar to Figure 2.5, in Figure 2.13 (a), the mean value of the sub-threshold current $E[I_{Sub}]$ (shown by the solid line) and $E[I_{Sub}]\,ref$ (shown by the dashed line, which represents the predicted proportional increase trend of the drain current by ignoring the inverse narrow width effect) indicate that the actual increase of the current is not linearly proportional to the increase of the channel width. Width tuning is not effective for PMOS transistors in the sub-threshold region due to the inverse narrow width effect [10]. In Figure 2.13 (b) and Figure 2.13 (c), the standard deviation and the spread parameters of I_{Sub} are shown. When the channel width is increased, the standard deviation and the drain current spread increase.

To increase the current drivability in the sub-threshold region, increasing the width of the PMOS transistor is not as effective as in the super-threshold region. Considering the variation and spread increase caused by transistor width increase, the traditional width-tuning only method is not an optimal choice in the sub-threshold region for current optimization. Therefore, other parameters of the transistor need to be investigated. In the next section, the transistor channel length tuning effects are studied.

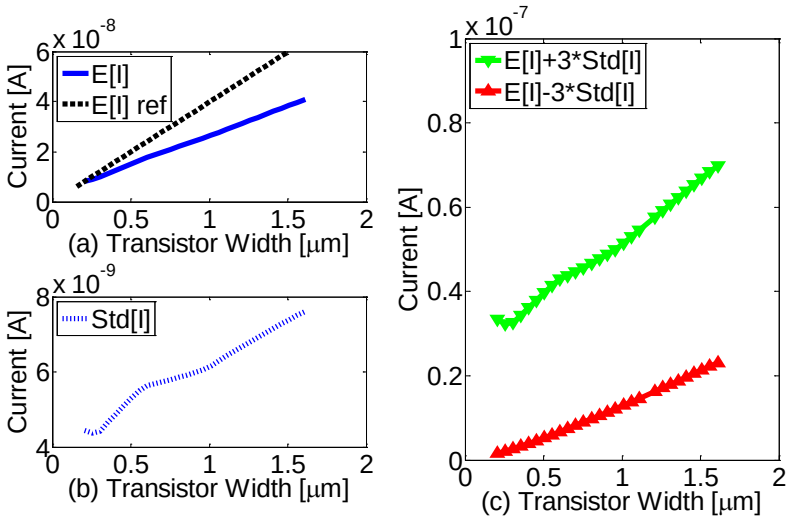


Figure 2.13 The drain current distribution parameters of the PMOS transistors vs. transistor width (L = 40nm. $|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

2.2.2 Length tuning

In the super-threshold region, the channel length of the transistors is commonly kept at the minimum value at different technology nodes to achieve high current drivability. However in the sub-threshold, the impact of channel length on threshold voltage effect plays an important role due to the exponential relationship between the sub-threshold current and the threshold voltage as indicated in Equation (2.2). In order to understand how the transistor channel length tuning affects the PMOS transistors, Monte Carlo simulations (at $V_{GS} = V_{DS} = |0.3V|$, T = 25°C) with global and local variation enabled are used to show the distribution parameters of V_{th} and I_{Sub} in the sub-threshold region. When tuning the channel length, the channel width is kept constant in the simulations.

In Figure 2.14, the distribution parameters of the $|V_{th}|$ of the PMOS transistors are shown. Unlike increasing the channel width, when the channel length is increased, all the distribution parameters shown in Figure 2.14 decrease. Namely, increasing the channel length results in smaller $|V_{th}|$, smaller $|V_{th}|$ variation, and narrower $|V_{th}|$ spread.

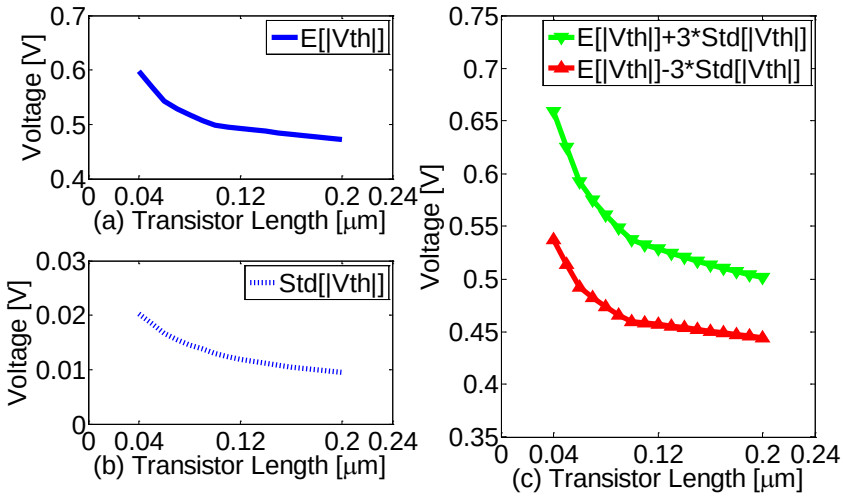


Figure 2.14 The distribution parameters of the $|V_{th}|$ of PMOS transistors vs. transistor length (W = 0.3μm. $|V_{GS}| = |V_{DS}| = 0.3$V. T = 25°C).

Increasing the channel length decreases the $|V_{th}|$ as shown in Figure 2.14. Smaller $|V_{th}|$ leads to higher transistor current. As is implicated from Equation (2.2), however, the transistor current is inversely proportional to the channel length, which means increasing the channel length decreases the transistor current. These two opposite effects on the drain current because of increasing the transistor channel length are shown in Figure 2.15.

At the minimum channel length, the PMOS transistor has the highest $E[I_{Sub}]$ and $Std[I_{Sub}]$ values and widest spread in the comparison range. When the channel length is increased from the minimum length, the effect of the decreasing threshold voltage is dominant, which leads to the increase of the transistor current until 0.1μm as shown in Figure 2.15 (a). The inverse proportional relationship starts to dominate and the transistor current starts to decrease. A similar trend of the $Std[I_{Sub}]$ is shown in Figure 2.15 (b). However, the difference (as indicated by Equation (2.3)) is that the $Std[V_{th}]$ decreasing effect is more pronounced. Together with the trends of $E[I_{Sub}]$ and $Std[I_{Sub}]$, the spread parameters are also shown in Figure 2.15 (c). When the channel length is increased, the spread of the drain current of the PMOS transistor is narrowed down.

In the sub-threshold region, increasing the channel length is more efficient in reducing the current variation as compared to increasing the channel width alone. Furthermore, when the channel length is increased, the current drop can be compensated by the decrease of the threshold voltage caused by the reverse short channel effect.

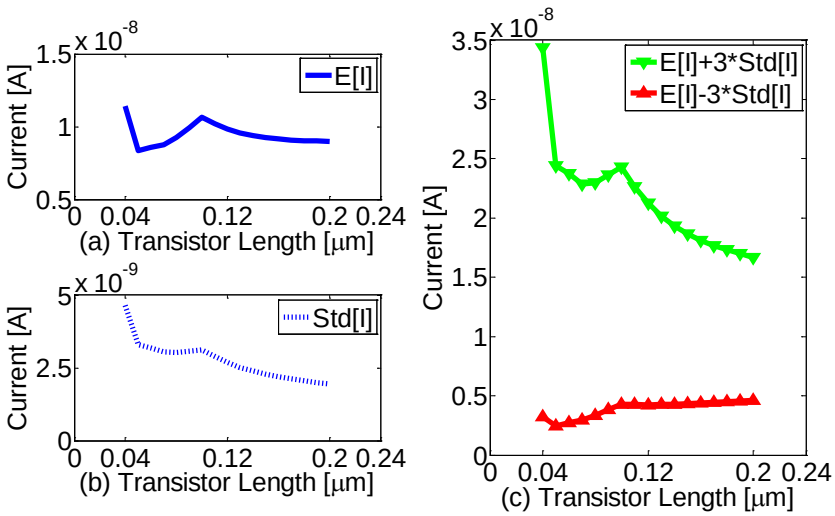


Figure 2.15 The drain current distribution parameters of the PMOS transistors vs. transistor length (W = 0.3μm. $|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

2.2.3 Finger effect

This section shows how a finger structure impacts the behavior of the PMOS transistors in the sub-threshold region.

Compared to the finger effect on the NMOS transistor, the gap between the multi-finger PMOS and non-finger PMOS transistors is wider in $E[I_{Sub}]$, $Std[I_{Sub}]$, and the total spread as shown in Figure 2.16.

The finger effect is more pronounced for PMOS transistors for increasing the current drivability than NMOS transistors as shown in Figure 2.16 (a). It is clearly shown that at the same total transistor width, the more fingers are used, the more current the fingered transistors can provide, and the more variation the transistors can tolerate as shown in Figure 2.16 (b) and Figure 2.16 (c). However, as is shown in the length tuning effect section, when the channel length is increased, the current variation decreases and the spread narrows.

This is also true for multi-finger PMOS transistors. The $Std[I_{Sub}]$ of the fingered PMOS transistor with different channel lengths and total widths are shown in Figure 2.17. The $Std[I_{Sub}]$ trends shown by the dashed lines are when the transistor channel length is kept at minimum (40nm in Figure 2.17). $Std[I_{Sub}]$ trends shown by the solid lines, are when the transistor channel length is increased (ranges from 0.06μm to 0.1μm). Comparing the solid lines to the dashed line in Figure 2.17 (b-d), it is noted that the $Std[I_{Sub}]$ of non-fingered transistors and multi-fingered transistors decrease when the channel length is increased. Furthermore, when the total width is increased, the gap between the $Std[I_{Sub}]$ of transistors with the minimum channel length and the $Std[I_{Sub}]$ of transistors with long channel increases.

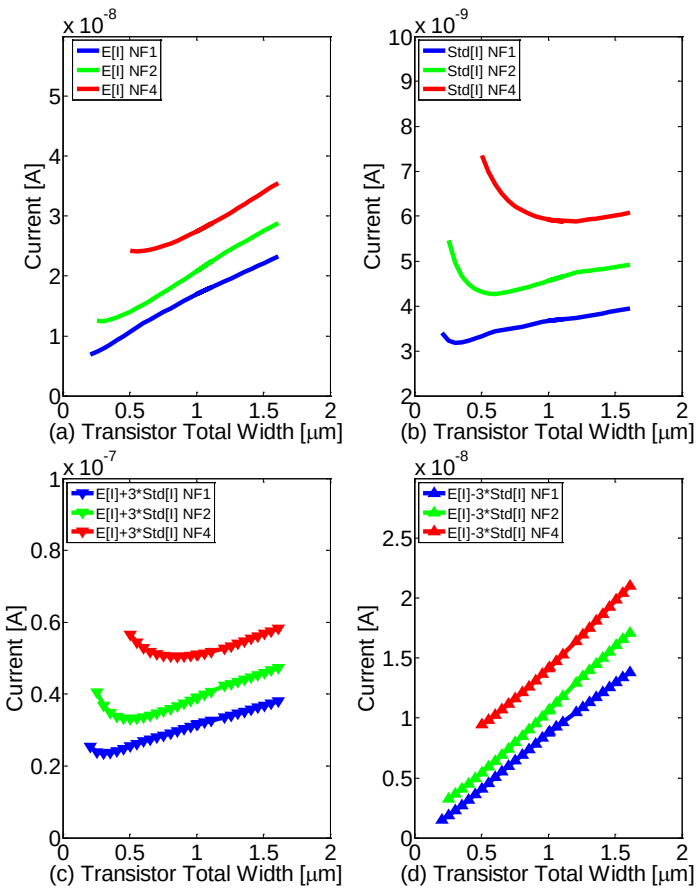


Figure 2.16 The drain current distribution parameters of the fingered PMOS transistor vs. transistor total width (L = 40nm. $|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

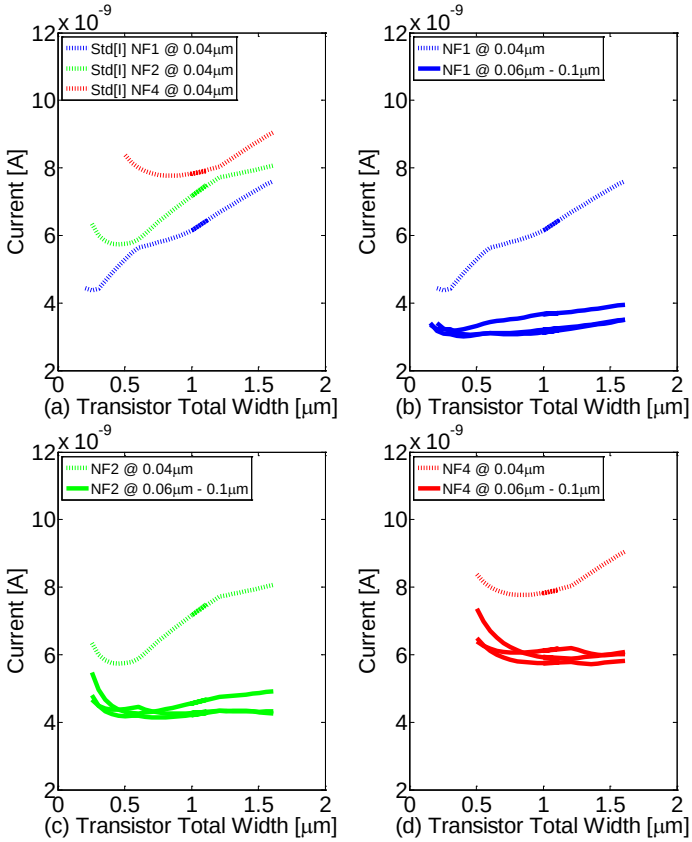


Figure 2.17 The standard deviation of the fingered PMOS transistors vs. transistor total width (L = 40nm and 0.06μm to 0.1μm. $|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

2.2.4 Conclusions of Section 2.2

In this section, the width tuning, length tuning, and finger effects are studied for PMOS transistors. Width tuning certainly can impact the current drivability of the PMOS transistors. However, due to the inverse short channel effect, the current increase is not proportional to the width increase as expected from the experience in the super-threshold region. Width tuning in the sub-threshold region is not as efficient as in the super-threshold region. Another drawback is that when the channel width is increased, the variation of the PMOS transistors increases.

Length tuning and finger structures, on the other hand, are more efficient in increasing current drivability and reducing the variation of the PMOS transistors in the sub-threshold region compared to operating in the super-threshold region.

2.3 Conclusions of Chapter 2

In section 2.1, single NMOS and PMOS transistor behaviors in the sub-threshold region are studied. Compared to the super-threshold region, in the sub-threshold region, the NMOS and PMOS transistors behave differently. In the super-threshold region, width tuning is used to increase the current drivability of the transistors. However, in the sub-threshold region, due to the inverse narrow width effect, width tuning is not as efficient as in the super-threshold region when increasing the current is the tuning target. Alternatively, length tuning is very efficient in increasing the current and reducing the variation in the sub-threshold region. In the finger structure section, it is shown that, for a fixed total width, the more fingers used, the better total current the structure can achieve. All the three parameters, channel width, channel length, and number of fingers, allow the designers to explore the transistor sizing space in sub-threshold region for stronger current and less variation.

CHAPTER 3

STANDARD CELL TRANSISTOR SIZING IN SUB-THRESHOLD

Standard cells encapsulate combinations of series and parallel connected transistors. The way these transistors are sized has a direct impact on current drivability and sensitivity to process variability. This chapter presents the design strategies to properly size transistors under the previously mentioned connectivity types. The proposed approach considers the Log-Normal behavior of the sub-threshold current, therefore includes the impact of process variability in the transistor sizing calculations. This chapter presents first the study on series connected (stacked) transistors, followed by the study on parallel connected transistors. The analysis includes current drivability, robustness, and signal slew results.

3.1 Stacked Transistor Behavior

In [16, 32], the sub-threshold current equation (Equation 2.2) was extended to stacked transistors. The current of the upper and lower transistors can be formulated as in Equation (3.1).

$$I_L = TW_L e^{\frac{(V_{dd}-V_X)-(V_{th}+(m-1)V_X+\lambda(V_{dd}-V_X))}{mkT/q}}\left(1-e^{\frac{-(V_{dd}-V_X)}{kT/q}}\right)$$
$$I_U = TW_i e^{\frac{(V_{dd})-(V_{th}+\lambda(V_X))}{mkT/q}}\left(1-e^{\frac{-(V_X)}{kT/q}}\right) \tag{3.1}$$
$$T = \mu_{eff} C_d \frac{1}{L}\left(\frac{kT}{q}\right)^2$$

where

W_i	Width of the "upper" transistor (PMOS: closer to VDD; NMOS: closer to ground);
W_L	Width of the "lower" transistor (PMOS: away from VDD; NMOS: away from ground);

λ DIBL effect coefficient; and
V_X The voltage of the Node X shown in Figure 3.1

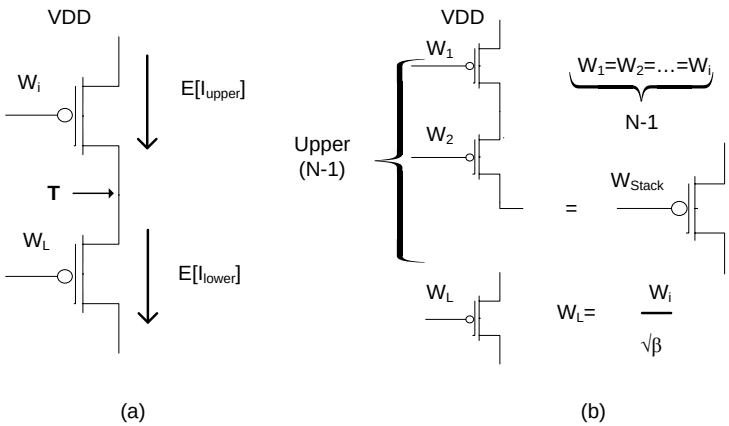


Figure 3.1 Stacked PMOS transistors (a) Stack 2 (b) Stack N

By equalizing the transistor current of the upper and lower transistors in Equation (3.1), the voltage at node X, shown in Figure 3.1 (a), can be obtained using Equation (3.2).

$$V_X = \frac{kT}{q}\ln\left(1 + \frac{\beta W_L}{W_i}\right)$$
$$\beta = e^{\frac{-\lambda V_{dd}}{mkT/q}} \tag{3.2}$$

The current flow in the serial connected transistors can be determined by Equation (3.3).

$$I_U = I_L = T\frac{\beta W_i W_L}{\beta W_L + W_i} e^{\frac{(V_{dd})-(V_{th})}{mkT/q}} \tag{3.3}$$

By solving the differential equation $\frac{\partial I_U}{\partial W_i} = 0$ or $\frac{\partial I_L}{\partial W_L} = 0$, the optimal width ratio between the two serial connected transistors for maximizing the stack current can be determined in Equation (3.4).

$$W_i = \frac{W_i + W_L}{1 + \sqrt{\beta}} \tag{3.4}$$

$$W_L = \frac{W_i + W_L}{1 + \sqrt{\beta}} \sqrt{\beta}$$

For any two transistors connected in a series to achieve the maximum current, the transistor farthest away from the rails should be sized $\sqrt{\beta}$ larger compared to the other transistor. Note that the sizing ratio β is a technology and voltage dependent parameter.

Similarly, the calculation is extended to the current distribution parameters of the stack transistor structure. By equalizing the mean current of the upper and lower transistors as shown in Figure 3.1 (a),

$$E[I_{Sub}]_i = E[I_{Sub}]_L \tag{3.5}$$

The mean current of the 2-transistor stack can be seen in Equation (3.6). The equivalent transistor width is also shown. Note that the channel length parameter of the transistors is excluded from the equation for simplification purposes as it is assumed that both transistors have the same channel length.

$$\begin{gathered} E[I_{stack2}] = TK_E W_{Stack2} e^{\frac{V_{dd} - E[V_{th}]}{mkT/q} + \frac{Std^2[V_{th}]}{2(mkT/q)^2}} \\ W_{Stack2} = \frac{\beta W_L}{\left(1 + \beta W_L \frac{1}{W_i}\right)} \\ \beta = e^{\frac{-\lambda V_{dd}}{mkT/q}} \end{gathered} \tag{3.6}$$

where

K_E Technology fitting parameter of the mean current, and
W_{Stack2} Equivalent width for two series connected transistors.

Furthermore, the mean current equation is extended to N transistors in a stack as shown in Figure 3.1 (b). The equivalent width of N transistors connected in series is shown in Equation (3.7)

$$W_{StackN} = \frac{\beta W_L}{\left(1 + \beta W_L \left(\sum_1^{N-1} \frac{1}{W_i}\right)\right)} \tag{3.7}$$

Note that, when sizing the stacked transistors, the sizing ratio within the upper PMOS (lower NMOS) transistors does not have a significant impact on the stack current. For layout purposes, the size of the upper PMOS (lower NMOS) transistors can be kept equal. If maximizing the sub-threshold current drive is the sizing optimization target, without increasing the total size of the

stacked transistors, the bottom PMOS (the top NMOS) transistor in Figure 3.1 (b) must be sized $\sqrt{\beta}$ larger compared to the upper PMOS (lower NMOS) transistors.

The variance of the stack is determined by the variance of each transistor in the stack. Since each transistor has the same impact on the total variance, the stack variance is the sum of the variances of each transistor divided by the square of the number of transistors in the stack[61].

$$Std^2[I_{stackN}] = \frac{1}{N^2}\left(Std^2[I_L] + \sum_{i=1}^{N-1} Std^2[I_i]\right)$$
$$Std^2[I_{stackN}] = \left(\frac{\beta(\sum_1^{N-1} W_i) + W_L}{K_{Std} N^2 W_{stackN}}\right)(e^{\frac{Std^2[V_{th}]}{(mkT/q)^2}} - 1)E^2[I_{stackN}] \tag{3.8}$$

where

K_{Std} Technology fitting parameter of the standard deviation of the current.

From Equations (3.6), (3.7), and (3.8), the current variation of N transistors in a stack can be derived, as shown in Equation (3.9).

$$\frac{Std[I_{stack}]}{E[I_{Stack}]} \propto \sqrt{\frac{\left(\beta\sqrt{\beta}(N-1)+1\right)\left(1+\sqrt{\beta}(N-1)\right)}{K_{Std} N^2 \beta}} \tag{3.9}$$

The current distribution parameter equations are very important for sub-threshold designs. Based on these equations, the number of transistors that can be connected in one stack can be calculated under certain constraints such as mean current constraints, standard deviation constraints, and area constraints. Note that, the equations are technology and supply voltage dependent. As the technology node and supply voltage scales, the technology fitting parameter and supply voltage related parameter need to be adjusted accordingly.

Table 1 Current variation in series-connected transistors

Number of transistors in series	Simulation results		Normalized $\frac{Std[I_{stack}]}{E[I_{stack}]}$	Calculation from Equation (2.13)
	$E[I_{Stack}]$	$Std[I_{stack}]$		
2x1.50μm	1.45E-07	59.98%	1	1
3x1.00μm	7.29E-08	67.35%	1.123	1.107
4x0.75μm	3.59E-08	70.42%	1.174	1.161
5x0.60μm	2.43E-08	73.55%	1.226	1.192

Three-thousand Monte-Carlo simulations were used to calculate the distribution parameters of 2, 3, 4, and 5 PMOS transistors in a stack working in sub-threshold ($|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C). All transistor channel lengths are made equal. The total width (sum of individual transistor widths) for each simulation case is set to 3μm. In Table 1, it is shown that Equation (3.9) predicts correctly the trend of the variation. The mismatch between the calculation and the simulation values is because, in the sizing equations, V_{th} is treated as a given technology dependent parameter (source bulk modulation is not taken into account).

With the understanding of the stacked transistor current distribution in the next two sections, the behavior of the stacked NMOS and PMOS transistors in the sub-threshold region is discussed.

3.1.1 NMOS Transistor Stack

Understanding the behavior the stacked NMOS transistors in the sub-threshold region can help designers define the design constraints and characterization of the standard cell library. In section 3.1.1, the focus of the study is on the current, and the output slope distribution parameters of the stacked NMOS transistors.

3.1.1.1 Current Distribution Parameters

In this section the stack current distribution for 2, 3, and 4 transistors in one stack is compared in a CMOS 40nm technology. Within the stack, the width and length of each individual transistor is set to be the same. To compare the stack effects on the mean current, a number of cases with various transistor channel widths and lengths are simulated in the sub-threshold region ($V_{GS} = V_{DS} = 0.3V$, T = 25°C). Monte Carlo simulations with 3000 iterations were performed to calculate the mean value of the transistor drain current. The results are shown in Figure 3.2.

The X-axis is set to be the number of transistors within the stack. The Y-axis is set to be the mean stack current. In Figure 3.2 (a-d), the channel length values (represented by L) are 0.04μm, 0.06μm, 0.08μm, and 0.1μm, respectively. The width of individual transistors in the stack is represented by W. One can see from lines 1 to 5 in Figure 3.2 (a-d), that 1) when the channel width is increased, the mean current increases, 2) when the number of the transistors is increased, the mean current decreases. To achieve the same mean current of an equivalent wide transistor, the sizes of the corresponding stacked transistors

need to be increased. For example, in Figure 3.2 (a), the black dashed line shows that to achieve the same mean current of an NMOS transistor with W = 0.405μm and L = 0.04μm (as shown by line 5 at X=1), the NMOS transistor stack needs to be sized up 2.5x for a two NMOS transistors stack (as shown by line 3 at X=2) and 4.0x for a three NMOS transistors stack (as shown by line 1 at X=3). In Figure 3.2 (b-d), a similar ratio can be found.

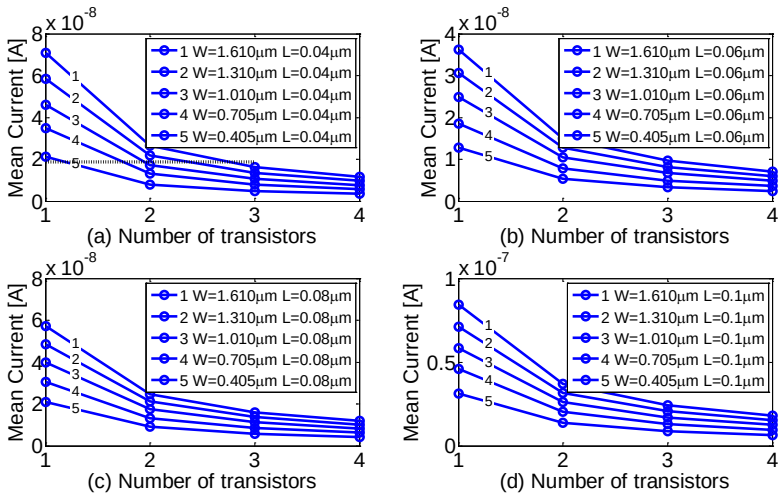


Figure 3.2 Mean stack current vs. Number of transistor in an NMOS transistor stack with various channel widths (0.405μm-1.610μm) and lengths: (a) 0.04μm, (b) 0.06μm, (c) 0.08μm, and (d) 0.1μm (V_{GS} = V_{DS} = 0.3V. T = 25°C).

From the mean current perspective, the more transistors in one stack, the less the mean current is. From the variation perspective, when the size of the transistor is kept the same and the number of stacked transistors is increased, the variation decreases as shown in Figure 3.3. The trends of different channel lengths of 0.04μm, 0.06μm, 0.08μm, and 0.1μm are shown in Figure 3.3 (a), Figure 3.3 (b), Figure 3.3 (c), and Figure 3.3 (d), respectively. The dashed black lines show the standard deviation of the current for Stack1 (1 transistor in the stack) NMOS transistor. The blue, green, and red solid lines show the trend of a stack with 2, 3, and 4 transistors, respectively. As indicated by the dashed line and the solid lines, when number of transistors in a stack is increased, the standard deviation of the stack current decreases.

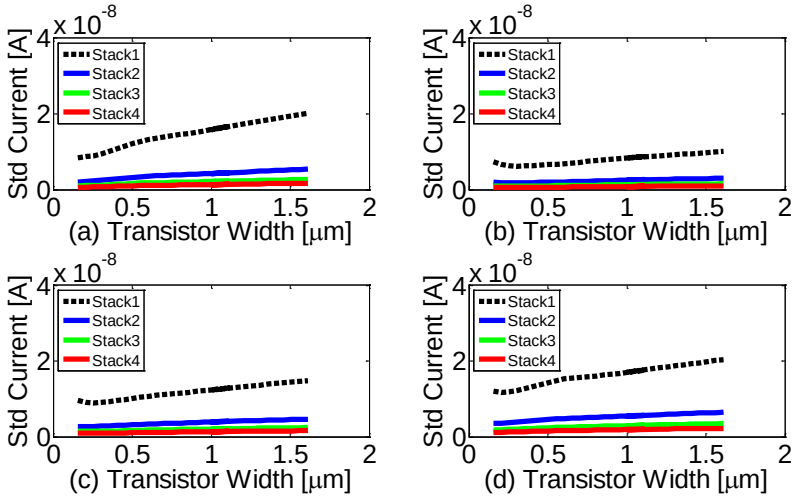


Figure 3.3 The standard deviation of the stack current vs. Number of transistor in an NMOS transistor stack with various channel lengths: (a) 0.04μm, (b) 0.06μm, (c) 0.08μm, and (d) 0.1μm ($V_{GS} = V_{DS} = 0.3V$. T = 25°C).

3.1.1.2 Output Slew Distribution Parameters

In this section, another interesting parameter is introduced: the output slew. It measures the transition time (from 10% to 90% of the supply voltage) at the output node of a capacitive loaded stack. In the sub-threshold, the variation of the input/output slew can cause serious timing violations and extra energy dissipation. Therefore, it is important to understand the relationship between the slew parameter and the transistor sizing parameters, especially for the clock element cell design and clock tree design [66].

In Figure 3.4, the slew distribution parameters of the Stack2, Stack3, and Stack4 NMOS transistors are shown. In the simulation setup, the channel length of all transistors in the stack is set to be 0.04μm while the channel width of each transistor in the stack varies from 0.15μm to 1.61μm. For the measured transistor, a fan-out 4 structure is used as the output node. The input of the measured transistor is connected to a three-stage inverter chain. The influence of the input slew setting on the output slew is negligible. From the comparison of the stack mean current in Section 3.1.1.1, it is shown that, as the number of the transistors in a stack is increased, the mean stack current decreases as shown in Figure 3.2, which leads to a mean output slew increase as shown in Figure 3.4 (a). Limiting the number of transistors in a stack has a positive im-

pact on the transition response at the output node of the stack, especially when the width of the transistors in the stack is small. Furthermore the standard deviation of the output slew is smaller for fewer transistors in a stack.

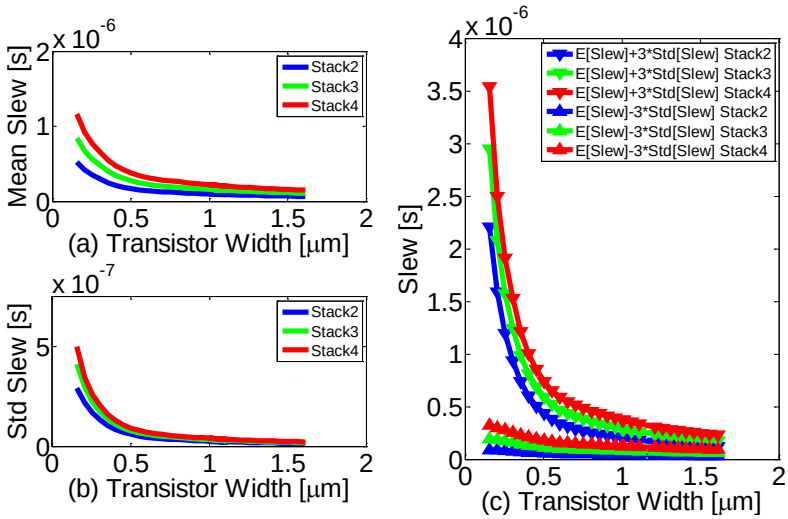


Figure 3.4 The output slew distribution parameters of an NMOS transistor stack vs. transistor width (L=40nm. $V_{GS} = V_{DS} = 0.3V$. T = 25°C).

3.1.2 PMOS Transistor Stack

Similar to the analysis of the NMOS transistor stack of Section 3.1.1, in this section, the distribution parameters of the stack current and the output slew of the PMOS transistor stack are shown for various channel widths and lengths.

3.1.2.1 Current distribution parameters

In this section, the stack current distribution of different channel widths and channel lengths of two, three, and four PMOS transistors in one stack is compared. The test setup is the same as the setup used in Section 3.1.1. The mean current of the stack PMOS transistors is shown in Figure 3.5.

Here we also see that when the channel width is increased, the mean current increases, and also when the number of the transistors is increased, the mean current decreases. To achieve the same mean current as that of an equivalent wide transistor, the size of the PMOS transistors in a stack needs to be increased as shown by the black dashed line in Figure 3.5 (a). The sizing up ratios to meet the current of the Stack1 PMOS transistor for the channel width

of the PMOS transistor in Stack2 and Stack3 are very similar to the ratios for NMOS transistors.

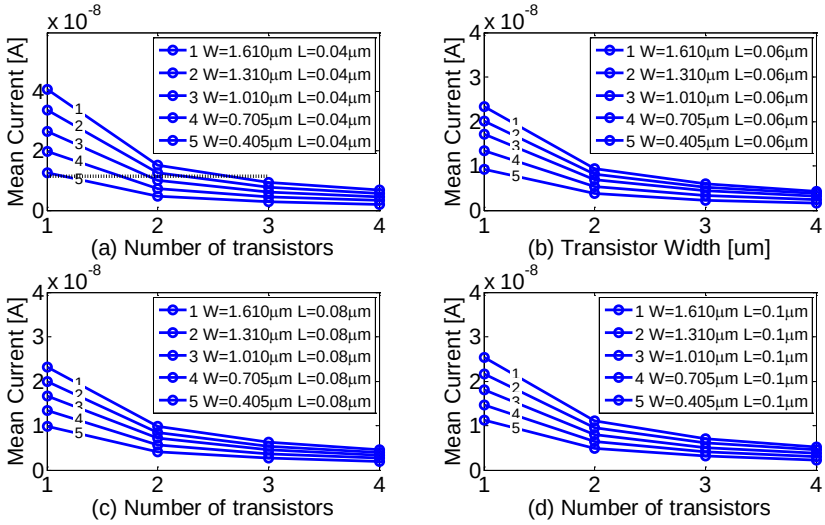


Figure 3.5 Mean stack current vs. Number of transistor in a PMOS transistor stack with various channel lengths: (a) 0.04μm, (b) 0.06μm, (c) 0.08μm, and (d) 0.1μm ($|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

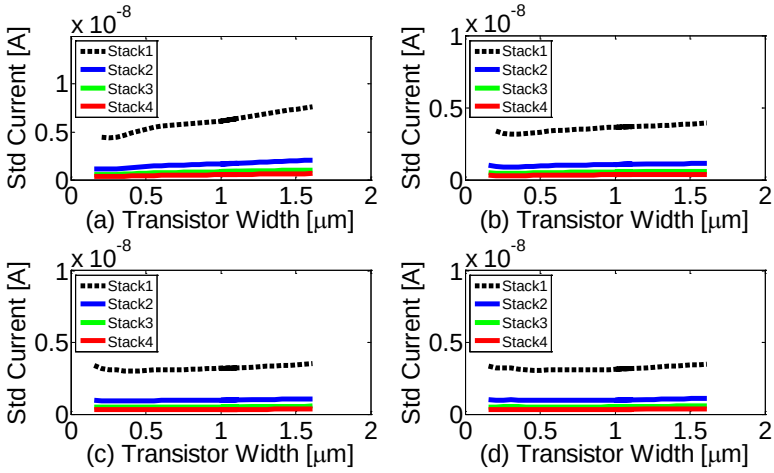


Figure 3.6 The standard deviation of the stack current vs. Number of transistor in a PMOS transistor stack with various channel widths and lengths (a) 0.04μm, (b) 0.06μm, (c) 0.08μm, and (d) 0.1μm ($|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

The standard deviation trends of stacked PMOS transistors with different channel lengths of 0.04μm, 0.06μm, 0.08μm, and 0.1μm are shown in Figure 3.6 (a), Figure 3.6 (b), Figure 3.6 (c), and Figure 3.6 (d), respectively. The X-axis shows the width tuning range. When the channel width is increased, the standard deviation of the current increases as well. Comparing a different number of transistors in a stack, it is noted that when the number of transistors in a PMOS transistor stack is increased, the standard deviation of the current flowing in the stack decreases. Comparing Figure 3.6 (a) to Figure 3.6 (d), we can see when the transistor channel length is increased, the standard deviation of the stack current decreases.

3.1.2.2 Output slew distribution parameters

The output slew distribution parameters of the PMOS transistor stack is shown in Figure 3.7. All the PMOS transistors have the same channel length 0.04μm. Similar to the output slew distribution of the NMOS transistor stack, when the number of transistors in a stack is increased, the mean slew increases, while the slew distribution shifts to the slower side as shown in Figure 3.7 (c).

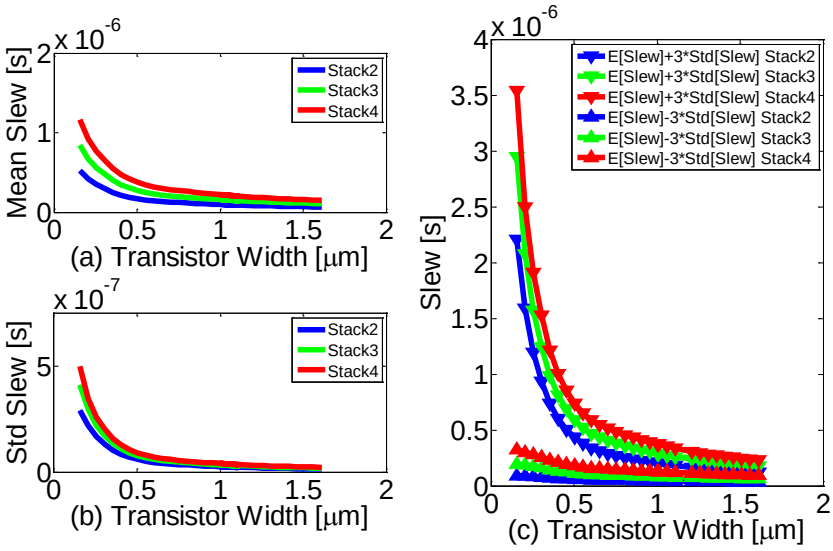


Figure 3.7 The output slew distribution parameters of a PMOS transistor stack vs. transistor width (L=40nm. $|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

3.1.3 Conclusions of Section 3.1

The behavior of stacked NMOS and PMOS transistors in the sub-threshold region was analyzed in this Section. Stacked NMOS or PMOS structures are often used in a multiple input structure in most of the standard logic cells. In general, when the number of the transistors in a stack is increased, the standard deviation of the current decreases, the mean current decreases, and the output slew increases.

3.2 Parallel Transistor Behavior

For parallel sizing, the mean and variance of the total current of N transistors connected in parallel is calculated by the sum of the Log-Normal distributions. The sum of Log-Normal distributions with the same variance can be approximated by one equivalent Log-Normal distribution[67]. However, the sum does not represent the actual current since the summation assumes uncorrelated parallel transistors. Hence, a correlation factor ρ_p for V_{th} needs to be introduced. This correlation factor is not needed in series-connected transistors because, in that case, the source-bulk modulation overshadows the correlation. The mean and variance of the current of N identical parallel connected transistors are,

$$\begin{aligned} E[I_{para}] &= NK_{Ep}\frac{W_{one}}{L}\, e^{\frac{V_{gs}-E[V_{th}]}{nU}+\frac{Std^2[V_{th}]}{2(nU)^2}+\frac{N^2}{\rho_p}} \\ K_{Ep} &= \mu C e^{1.8} U^2 (1 - e^{\frac{-V_{ds}}{U}}) \\ \rho_p &\propto Std^2[V_{th}] \\ Std^2[I_{para}] &= (e^{\frac{Std^2[V_{th}]}{(nU)^2}+\frac{2N}{\rho_p}} - 1)\,(E[I_{para}])^2/N^2 \end{aligned} \tag{3.10}$$

where

W_{one} the width of one single transistor.

The equivalent width (W_{Para}) for parallel connected transistors can therefore be calculated from Equation (3.11),

$$\begin{aligned} W_{Para} &= \gamma(N) W_{one} \\ \gamma(N) &= N e^{\frac{N^2}{\rho_p}} \end{aligned} \tag{3.11}$$

Hence, the width of one single transistor, which has the same mean current as the one of N transistors in parallel, is $\gamma(N)$ times the width of the transistors in parallel.

Three thousand Monte-Carlo simulations were used for one to six NMOS transistors in parallel, with a total width of 1.2μm to investigate the current distribution. The simulation and calculation results are listed in Table 2. It is worth observing these results in more detail. Namely, the joint correlated Log-Normal distribution indicates that the mean current is bigger than that of the uncorrelated sum of individual transistor currents [7].

Table 2 Mean current of parallel connected transistors

Number of parallel transistors	Simulated E[I] (A)	Normalized E[I]	Calculation from (3.10)
1x1.20μm	3.54E-07	1.00	1.00
2x0.60μm	3.71E-07	1.05	1.15
3x0.40μm	4.13E-07	1.17	1.33
4x0.30μm	5.09E-07	1.44	1.54
5x0.24μm	6.29E-07	1.78	1.77
6x0.20μm	7.64E-07	2.16	2.04

3.2.1 Parallel NMOS Transistors

In this section the distribution parameters of the current and slew of the parallel connected NMOS transistors are shown for a supply voltage of 0.3V. The results of multiple channel width and length choices presented in Figure 3.8 for NMOS transistors with channel lengths of 0.04μm, 0.06μm, 0.08μm, and 0.1μm are shown in Figure 3.8 (a), Figure 3.8 (b), Figure 3.8 (c), and Figure 3.8 (d), respectively.

Unlike the comparison of the stack transistor section, the X-axis shows the width tuning range of the total width of all the transistors connected in parallel. The reason for this is that, while for the stack comparison the main concern is how many transistors can be connected within a stack, the main concern for parallel-connected transistors is to determine whether it is beneficial to spilt one big transistor into multiple transistors. The total transistor width is a sizing constraint to make a fair comparison between a single transistor, shown in black dashed lines, noted as Para1, and the multiple parallel transistors (Para2, Para3, and Para4), shown in blue, green, and red, respectively.

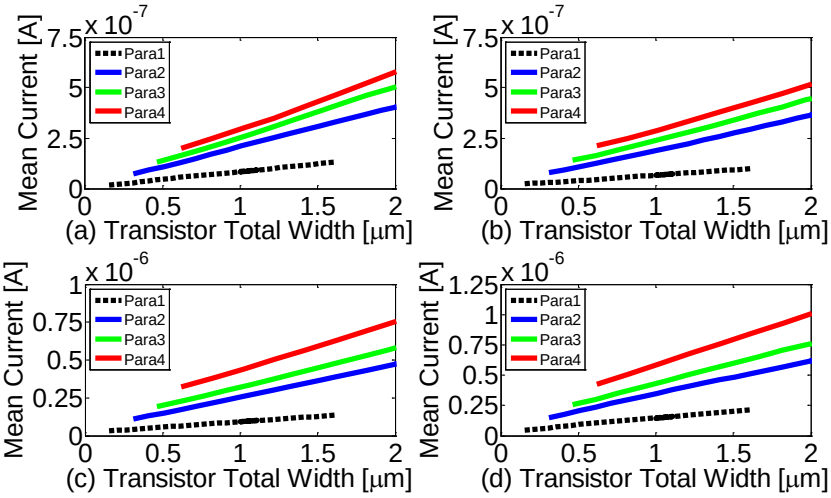


Figure 3.8 Mean current of parallel NMOS transistors vs. total width at different channel lengths: (a) 0.04μm, (b) 0.06μm, (c) 0.08μm, and (d) 0.1μm ($V_{GS} = V_{DS} = 0.3$V. T = 25°C).

In general, the more transistors connected in parallel, the bigger the mean current. For different lengths, the $\gamma(N)$ ratio of the parallel connection equivalent is different.

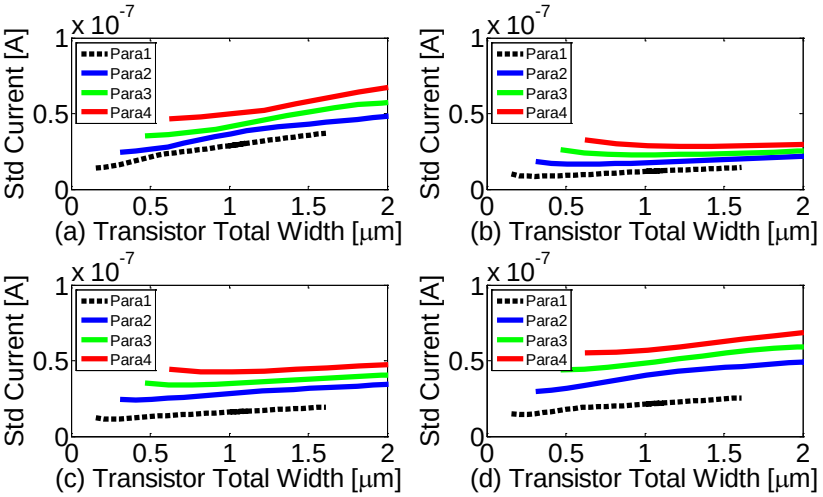


Figure 3.9 The standard deviation of the current of parallel NMOS transistors vs. total width at different channel lengths: (a) 0.04μm, (b) 0.06μm, (c) 0.08μm, and (d) 0.1μm ($|V_{GS}| = |V_{DS}| = 0.3$V. T = 25°C).

When the number of identical transistors connected in parallel is increased, the mean current flowing through the whole set of transistors increases. However, the standard deviation of the current also increases as shown in Figure 3.9. Looking at the four trends in Figure 3.9 (a), Figure 3.9 (b), Figure 3.9 (c), and Figure 3.9 (d), the more transistors connected in parallel, the larger the current standard deviation of the current is. However, when the channel length is 0.06μm (Figure 3.9 (b)) or 0.08μm (Figure 3.9 (c)), the standard deviation of the current of the parallel connected transistors is in the same range as for the single transistor with minimum channel length (shown in the dashed line in Figure 3.9 (a)).

When the channel length of the transistor is increased together with the number of the transistors connected in parallel, the mean current increases. However the standard deviation of the current is maintained in a similar range as a single transistor with minimum channel length.

The comparison is extended to the output slew of the two, three, and four transistors connected in parallel. The mean, standard deviation, and the upper and lower boundaries of the output slew distribution are shown in Figure 3.10. The distribution is calculated from 3000 Monte Carlo simulations.

The differences among all three trends of Para2, Para3, and Para4 are not significant when the total width is above 1μm. However, when looking at the starting points of each trend of the mean, the standard deviation, and the upper and lower boundaries of the output slew, the benefit of the parallel connection is shown clearly: the mean slew is reduced and the distribution of the slew is narrowed. With a total channel width of 0.465μm, the mean slew of Para3 is 20% lower, while the improvement on the standard deviation is 24% compared to the mean slew of Para2. With a total channel width of 0.62μm, the mean slew of Para4 is 15% and 25% lower compared to the mean slew of Para3 and Para2, respectively. Furthermore, the standard deviation of the slew is 17% and 32% lower compared to Para3 and Para2, respectively.

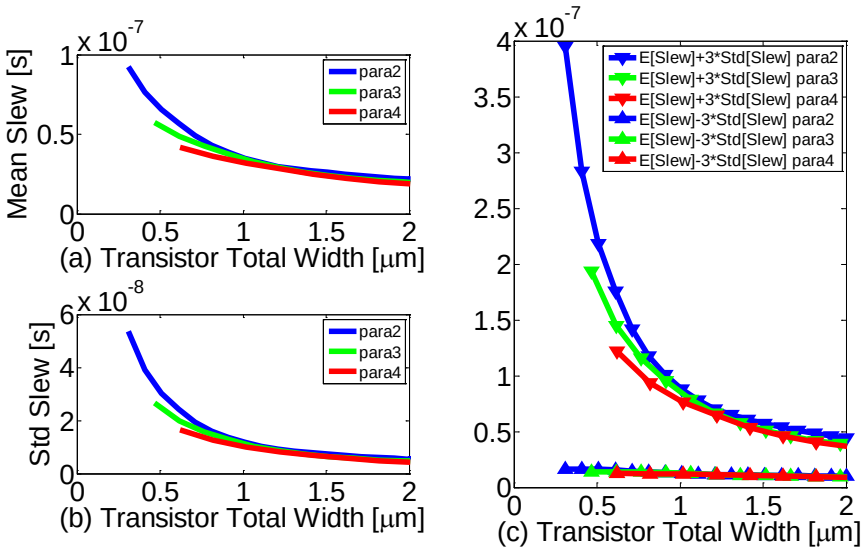


Figure 3.10 The output slew distribution parameters of parallel NMOS transistors vs. total width (L= 40nm. $V_{GS} = V_{DS} = 0.3V$. T = 25°C).

3.2.2 Parallel PMOS Transistors

Similar distribution parameter comparison is carried out for the PMOS transistors in this section. The mean current comparison is shown in Figure 3.11. The corresponding standard deviation comparison of the current is shown in Figure 3.12. The output slew distribution parameter comparison is shown in Figure 3.13.

Similar to the comparison between the parallel connected NMOS transistors and the one single NMOS transistor, the parallel PMOS transistors show higher mean current and larger standard deviation at the same channel length and total width. When the channel length is increased, the standard deviation of the current of both the parallel connected PMOS transistors and the one single PMOS transistor with the same total width decreases, as shown in Figure 3.12. With the transistor channel width and length tuning together, at 0.3V, the parallel connected PMOS transistors achieves higher mean current and smaller variation compared to the one single PMOS transistor with minimum channel length.

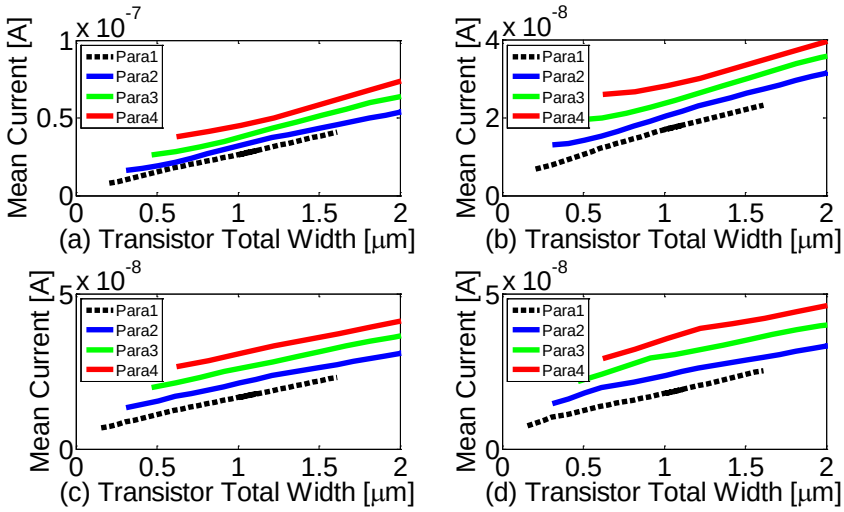


Figure 3.11 Mean current of parallel PMOS transistors vs. total width at different channel lengths: (a) 0.04μm, (b) 0.06μm, (c) 0.08μm, and (d) 0.1μm ($|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

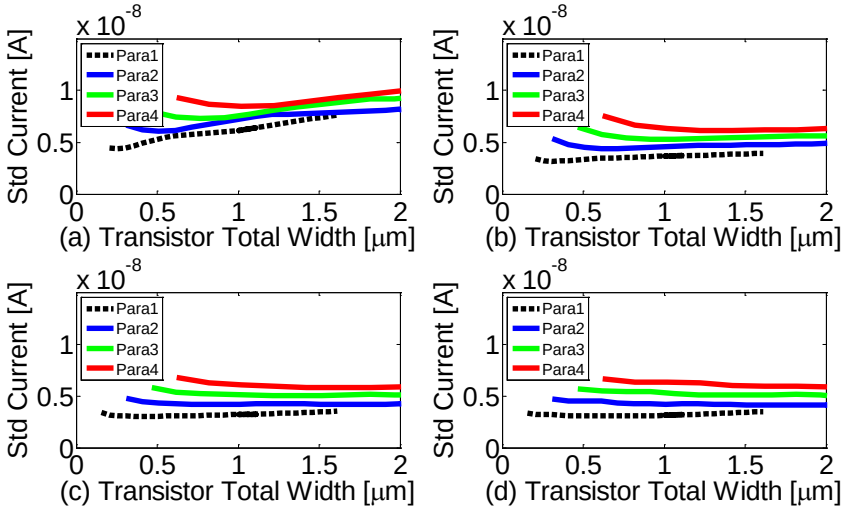


Figure 3.12 The standard deviation of the current of parallel PMOS transistors vs. total width at different channel lengths: (a) 0.04μm, (b) 0.06μm, (c) 0.08μm, and (d) 0.1μm ($|V_{GS}| = |V_{DS}| = 0.3V$. T = 25°C).

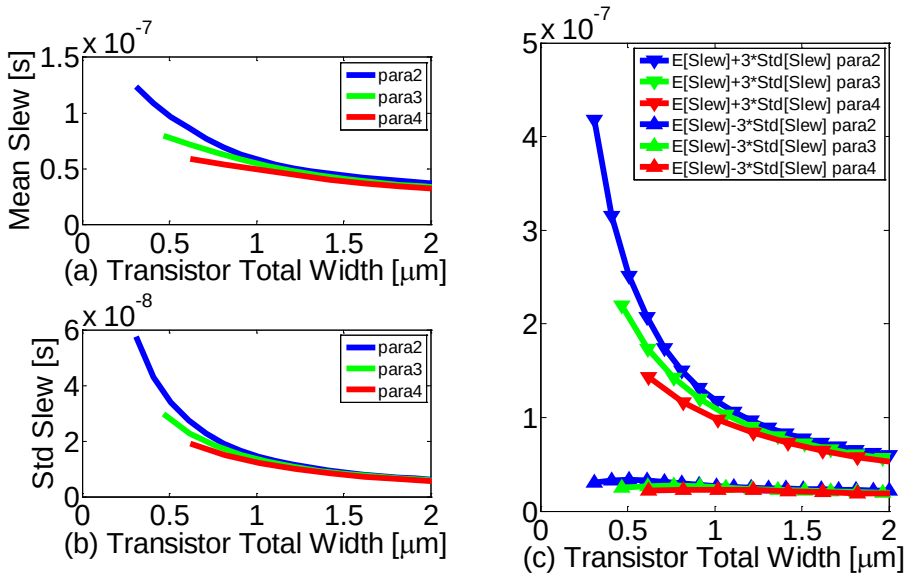


Figure 3.13 The output slew distribution parameters of parallel PMOS transistors vs. total width (L= 40nm. $|V_{GS}|=|V_{DS}|=0.3V$. T = 25°C).

The output slew distribution parameter of the parallel PMOS transistors is shown in Figure 3.13. When the total width is above 1μm, the difference between the trends is very small. However when looking at the starting points of each trend of the mean, standard deviation, and the upper and lower boundaries of the output slew, the benefit of parallel connection is shown clearly. Namely, the mean slew is reduced and the distribution of the slew is narrowed down. With a total width of 0.465μm, the mean slew of Para3 is 23% lower compared to the mean slew of Para2. Furthermore, the standard deviation is 23% smaller. With a total width of 0.62μm, the mean slew of the Para4 is 19% and 32% lower compared to the mean slew of Para3 and Para2, respectively. Furthermore, the standard deviation of the slew is 17% and 31% smaller compared to Para3 and Para2, respectively.

3.2.3 Conclusions of Section 3.2

The behavior of the parallel connected transistors has been studied in this section. Both NMOS and PMOS transistors have higher mean current and lower output slew when the number of transistor connected in parallel is increased. When the channel length is increased, the current variation of the parallel connected transistors decreases. In general, increasing the number of parallel con-

nected transistors is beneficial for increasing the current drivability of both NMOS and PMOS transistors. The drawback of the variation increase can be compensated by increasing the channel length of the transistors.

3.3 Conclusions of Chapter 3

In this chapter, the transistor behavior in the sub-threshold region is shown analytically and further proven by Monte-Carlo simulation results at 40nm CMOS technology node. Based on the distribution parameter model, the mean and variation of the transistor currents can be calculated. Inverse narrow width effect, reverse short channel effect, and finger effect are studied for one single transistor, serial connected transistors, and parallel connected transistors. The difference of the transistor behaviors in the sub-threshold region compared to the super-threshold region is discussed in detail. The three effects together create a transistor sizing space to optimize standard cells in the sub-threshold region for different purposes such as current drivability improvement, variation reduction, and area reduction.

CHAPTER 4

STANDARD CELL SIZING OPTIMIZATION IN SUB-THRESHOLD

4.1 *Sizing Methodology*

It is well known that the quality of designs produced by the logic synthesis tools depends on the library that is used [68]. Thus, two main aspects are included in standard cell library development: standard cell sizing and proper cell function selection. This section focuses on the first aspect, standard cell sizing.

4.1.1 Conventional Sizing

Conventional standard cell library sizing focuses on three aspects: low power, high speed, and minimal area as stated in [69]. In general, there exists three distinct optimization strategies to calculate the size of the NMOS and PMOS transistors. *i)* when reducing power consumption is the primary concern, the preference is to minimize the transistor sizes within noise margin requirements; *ii)* for high speed, the standard cell height and the width of the transistors are increased during the sizing optimization; and *iii)* the standard cell height is kept as short as in the sizing optimization for minimum area. The delay of the standard cells with minimum cell height is optimized based on the topology. During all three optimizations, the rise/fall time, the fan-in/fan-out, and the noise margin are set as constraints. The transistor width, length, and layout height are adjusted accordingly.

In [70], transistor sizing is geared to find the available size range for which the error, due to delay mismatch between different sizes, is minimized. The selection includes equal-spacing or exponential-spacing strategies. In [71], the optimum ratio between the NMOS and PMOS network is derived based on speed and noise immunity using a delay chain. The optimum cell height is determined based on the number of routing channels over one cell. In general,

the sizing methodology presented in [71] is targeted to achieve maximum cell speed/ layout area efficiency.

4.1.2 Proposed Sub-Threshold Sizing: Balancing

Compared to other sizing methods [6, 13, 16, 28, 29], the proposed sub-threshold sizing method treats the threshold voltage variation as one of the statistical parameters in the current/delay equation and optimizes cells to have balanced current/delay distributions. The sizing method is derived from the observation that the transistor's current distribution in the sub-threshold region follows a Log-Normal spreading, whereas conventional super-threshold and state-of the-art sub-threshold sizing treats the transistor's current as a Normal distribution, which is only true for the super-threshold region. Considering the above mentioned fact, and the observation that process variability can be mapped onto threshold voltage variability to within a first order approximation, a methodology for robust standard cell design has been developed.

The focus of the proposed sub-threshold sizing methodology is to balance the drive strength of the Pull-Down-Network (PDN) and the Pull-Up-Network (PUN) while considering the impact of the threshold voltage variation in the sub-threshold region.

4.1.2.1 Why Balancing

Balancing the drive strengths of the PDN and PUN is a commonly used sizing method in the super-threshold region. Balancing is also very important in the sub-threshold. Previous sub-threshold sizing methods in [6, 13, 16, 25-29] take advantage of different transistor sizing effects, such as short channel effect (SCE), different sizing optimization setups, or simulation methods to build a set of standard cells.

In the sub-threshold, cell sizing based on individual transistors without considering the impact of process variations cannot compensate the big gap between the drive strength distributions of the PDN and PUN due to active (sub-threshold) current. Matching the drive strength distribution of the PDN and PUN can provide equal pull-up and pull-down delay distributions and improve the worst case transition delay. By properly balancing the PDN and PUN distributions, the cell delay spread can be narrowed, which makes the cell more robust.

4.1.2.2 Before and After Balancing

The threshold voltage variation severely impacts the current distribution in the sub-threshold region. In Figure 4.1, the distributions of V_{th} and transistor current in the sub-threshold regime of two single SVT NMOS transistors (W=0.3μm, L=0.1μm from 90nm CMOS technology and W=0.15μm, L=0.04μm from 40nm CMOS technology) working with a supply voltage of 0.3V at 25°C are shown. The results are obtained from 3000 Monte Carlo simulations and are normalized to the mean values at corresponding technology nodes. The results from the 90nm CMOS technology are shown in blue and the results from the 40nm CMOS technology are shown in red color. From these simulations, it is evident that the V_{th} distribution obeys a Normal distribution and that the current obeys a Log-Normal distribution. The normalized distribution of V_{th} in the 40nm technology node is very similar to the one of the 90nm technology node. The normalized log-normal distributed sub-threshold current has a larger span compared to the one of the 90nm CMOS technology node as the 3sigma points show.

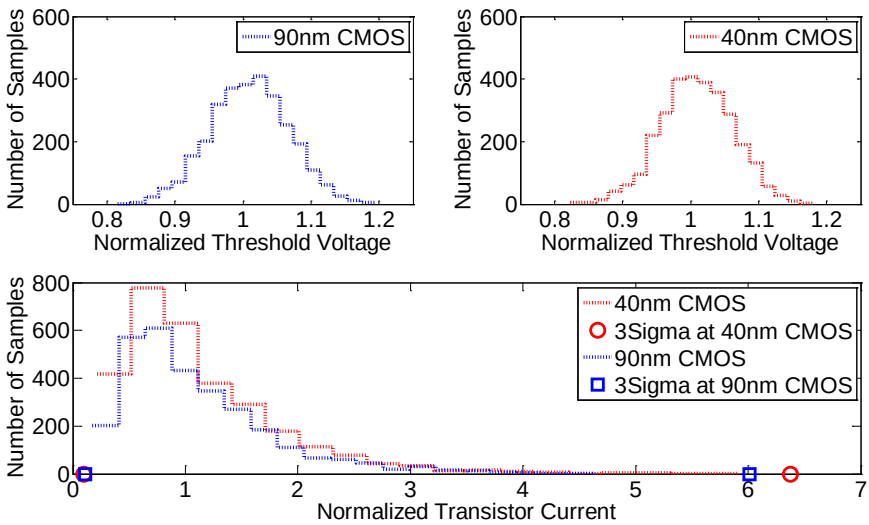


Figure 4.1 The normalized $V_{t\bullet}$ and transistor current distributions (W=0.3μm and L=0.1μm. T=25°C. 90nm CMOS technology and W=0.15μm and L=0.04μm. T=25°C. 40nm CMOS technology).

The widely spread current in the sub-threshold region is clearly shown in Figure 4.1. For one set of sizing parameters, due to the spread, the normalized current can be any value in the range of 0.1(-3sigma) to 6.4(+3sigma) as shown

in red circle, or even in a larger range as the technology node scales down. Therefore, it is very important to consider the impact of process variations and the proper sizing parameters for the balancing scheme.

Without loss of generality and as a way of illustration, consider an inverter sized under the conventional super-threshold approach. The current distributions before and after balancing are shown in Figure 4.2 (a) and Figure 4.2 (b), respectively. The green lines represent the normalized current distribution of the PMOS transistor, while the blue lines represent the normalized current distribution of the NMOS transistor. Note that all the values are normalized to the mean current of the NMOS transistor before balancing is applied. Before applying the sizing optimization, the NMOS and PMOS transistors follow the same sizing parameters as the conventional super-threshold sizing. The big current difference between NMOS and PMOS networks leads to unbalanced rise and fall transitions. After the proposed balanced sizing is applied, the gap of the current distributions of the NMOS and PMOS transistors can be reduced even without area penalty as shown in Figure 4.2 (b). The idea of the balancing is to slow down the fast transition path and speed up the slow transition path within the PDN and PUN so that the current distributions are properly matched and the worst case transition is improved.

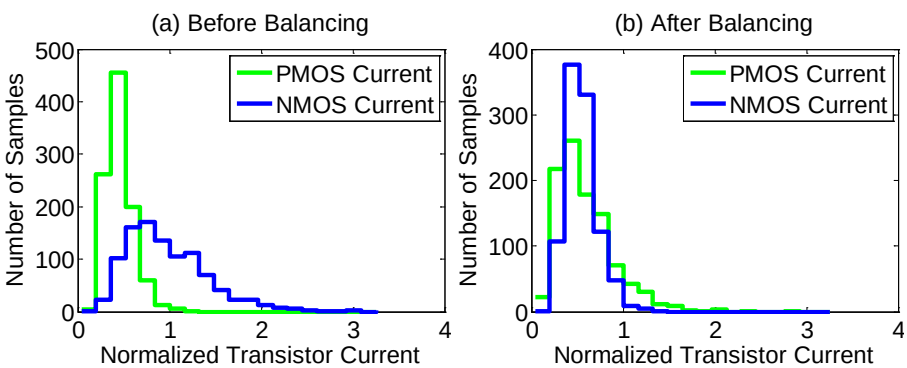


Figure 4.2 Comparison of the transistor current distributions (a) before and (b) after balancing.

4.2 *Sizing optimization for PUN and PDN*

It is clearly shown in the previous section that the balancing method can bring the current distributions of the PDN and PUN together. In the following sections, the balancing-based sizing optimization process is explained in detail.

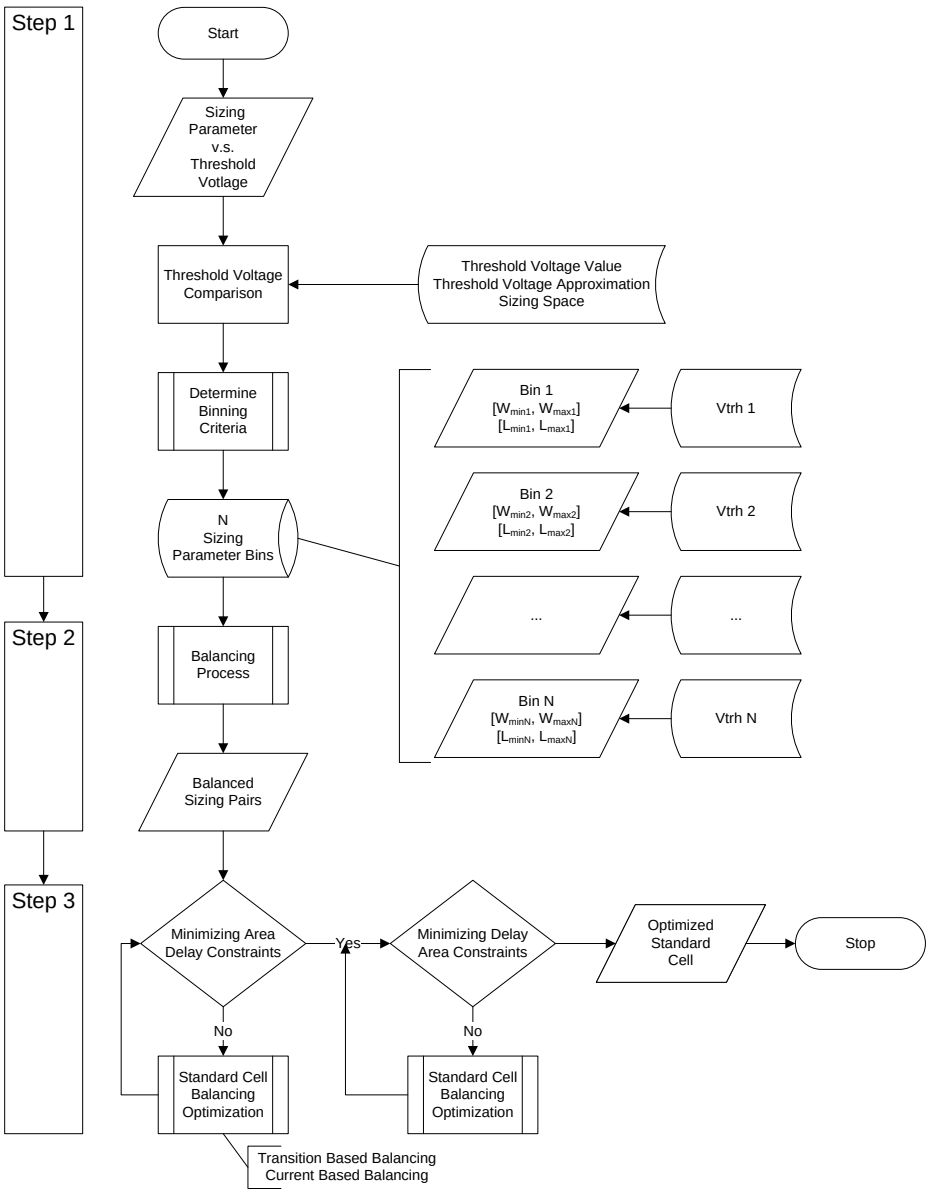
The transistor threshold voltage and sub-threshold current behaviors were illustrated in Chapter 2. The study shows that the width, length, and finger effects play important roles in the transistor's behavior for certain sizing ranges. Thus, during the sizing optimization, the first step is to do size binning to find out which effect is more pronounced and in what geometry range. The size binning is done based on the evaluation of the threshold voltage distribution. In each binned range, the threshold voltage is assumed to be constant. The channel width, channel length, and the number of fingers are set as tunable in that range.

4.2.1 Balancing the current distribution of PUN and PDN

Finding the threshold voltage as a function of transistor geometry is the first step of the sizing optimization process. For any two consecutive transistor geometries, there is one unique threshold voltage value. For example, in Flowchart 1, the threshold voltage value is V_{thN} in the range of [W_{minN}, W_{maxN}] and [L_{minN}, L_{maxN}]. The contour plots of the threshold voltage distribution parameters as a function of transistor geometry can be generated starting with a series of Monte Carlo simulations. In Figure 4.3, the threshold voltage as a function of the geometry, (width [0.155μm, 1.61μm] and length [0.04μm, 0.20μm]), is shown in the contour plots, where the x-axis represents the channel width in μm units and the y-axis represents the channel length in μm units. These contour plots are used to build the binning strategy shown in Algorithm 2. After merging in each sizing bin, the width and length are set to be fixed values determined by the average of the maximum and minimum in the bin.

Different reference geometry pairs can be chosen for different optimization purposes. When the optimization target is to increase the mean current, the binning is based on the threshold voltage mean value, and the standard deviation values are used as the second reference. When the optimization target is for the worst-case optimization, the lower boundary of the threshold voltage spread parameter is used as the main reference in the binning algorithm as shown in Algorithm 2. Another possible reference pair is the variation and mean value of the threshold voltage when the target is to reduce the spread of the current distribution.

Note that the binning is based on individual transistors. The stacking and finger effects are calculated and taken into account through the algorithms shown in the following sections.



Flowchart 1. Standard cell sizing optimization process.

Algorithm 2 Binning algorithm.

Algorithm 2: $E[V_{th}]$ and $Std[V_{th}]$are indicated as two references: one is the main reference (V) and the other one is (T). For n bins, the sets of available transistor width w_i, length l_j ($i, j \in [1, n]$,) have the same step size of δ (e.g.$0.05\mu m$ in CMOS 40nm).

Possible reference pairs:

For worst-case optimization, use $E[V_{th}] - 3 \times Std[V_{th}]$ as V, and $Std[V_{th}]$ as T

For variation (σ/μ) optimization, use $Std[V_{th}]/E[V_{th}]$ as V, and $E[V_{th}]$ as T

BEGIN Binning

INITIALISATION

SET $\forall\ i, j \in [1, n]$

$V_{min}(i) \leftarrow \min\ (V\{\ w_i, l_j\})$

$V_{max}(i) \leftarrow \max\ (V\{\ w_i,\ l_j\})$

$T_{min}(i) \leftarrow \min\ (T\{\ w_i,\ l_j\})$

$T_{max}(i) \leftarrow \max\ (T\{\ w_i,\ l_j\})$

SET Bin index $i\ = 1$

End INITALISATION

WHILE $i < n$

IF $(1 - \epsilon < \frac{V_{max}(i+1)}{V_{min}(i)} < 1 + \epsilon)$ **&** $(1 - \epsilon < \frac{T_{max}(i+1)}{T_{min}(i)} < 1 + \epsilon)$

Merge the ith **bin and the** $(i+1)^{th}$**bin**

ENDIF

$i = i + 1$

ENDWHILE

END Binning

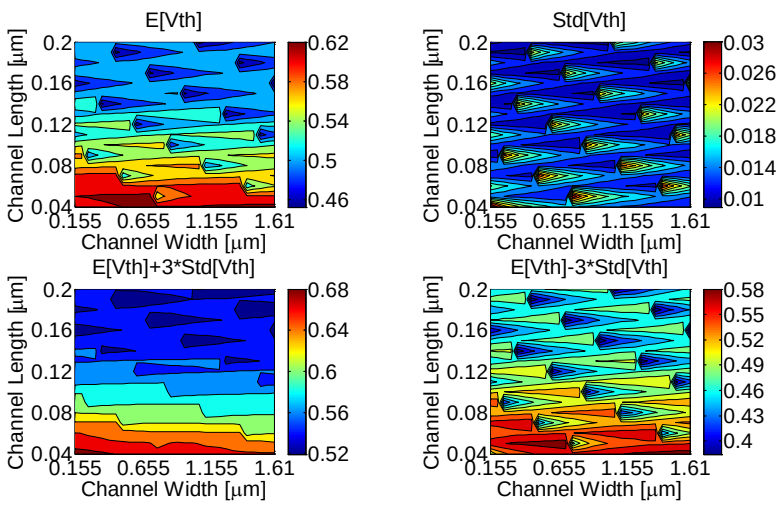


Figure 4.3 The distribution parameters of the threshold voltage (W=[0.155, 1.61]μm, L=[0.04, 0.2]μm. 40nm CMOS technology).

After fixing the sizing parameters in the binning step, the bins are used as an input to the balancing process. In the balancing process, the NMOS and PMOS sizing pairs can be found based on the balancing between the distribution parameters of the PUN and PDN current/delay. Equation (4.1) shows the condition in which to balance the current distributions of the PUN and PDN's.

$$E[I_{PDN}] = E[I_{PUN}] \tag{4.1}$$

Let us assume that the transistors in the PUN and PDN can be represented by an equivalent transistor, respectively. Introducing Equation (2.3) into (4.1), the balancing equation can be written as

$$\frac{W_{PDN}L_{PUN}}{W_{PUN}L_{PDN}} = \alpha e^{\frac{E[V_{thPDN}]-E[|V_{thPUN}|]}{nU}} e^{\frac{Std^2[|V_{thPUN}|]-Std^2[V_{thPDN}]}{2(nU)^2}} \tag{4.2}$$

The area and current optimizations are described in Algorithm 3. The area of the standard cells can be optimized based on the balancing equation (4.2) and Algorithm 3.

Algorithm 3. Area and current optimization algorithm.

Algorithm 3: *Area* and *current* are two distinct optimization targets. The goal is to determine the equivalent transistor sizes of PUN and PDN. In the pool (G) of all possible sizing pairs of PUN and PDN, P_i presents the ith pair of geometries. GB indicates the pool of balanced pairs of PUN and PDN where PB_j is its jth pair.

BEGIN Optimization for Area
INITIALISATION
  SET
  A_{opt} is the target area for optimization,
  $A_{opt} \leftarrow \infty$
  SET searching index: $i = 1$
  SET determine index: $j = 1$
End INITALISATION
  While ($P_i \in G$)
    IF $P_i = \left(\frac{W_{PDN}L_{PUN}}{W_{PUN}L_{PDN}}\right)_i = \alpha e^{\frac{E[V_{thPDN}]-E[|V_{thPUN}|]}{nU}} e^{\frac{Std^2[|V_{thPUN}|]-Std^2[V_{thPDN}]}{2(nU)^2}}$
      $GB \leftarrow P_i$
    ENDIF
    $i++$
  END While

  While ($PB_j \in GB$)

IF $(Area(\ PB_j\ ) < A_{opt}\ )$
$A_{opt} \leftarrow Area(\ PB_j\ )$
$j++$
END While

END Optimization for Area

BEGIN Optimization for Current
INITIALISATION
SET
$I_{opt} \leftarrow 0$
I_{opt} is the current for optimization under certain working condition ,
SET searching index: $i = 1$
SET determine index: $j = 1$
End INITALISATION

While $(P_i \in G)$
IF $P_i = \left(\frac{W_{PDN}L_{PUN}}{W_{PUN}L_{PDN}}\right)_i = \alpha e^{\frac{E[V_{thPDN}]-E[|V_{thPUN}|]}{nU}} e^{\frac{Std^2[|V_{thPUN}|]-Std^2[V_{thPDN}]}{2(nU)^2}}$
$GB \leftarrow P_i$
ENDIF
$i++$
END While

While $(PB_j \in GB)$
IF $(I(\ PB_j\ ) > I_{opt}\ )$
$I_{opt} \leftarrow I(\ PB_j\ )$
$j++$
END While

END Optimization for Current

The area of the standard cells can be optimized using the current-constrained area optimization. Standard cell area is often related to the loading capacitance for the previous stage. When the area is decreased, the loading capacitance is decreased. With a smaller loading capacitance at the output node, the charging time is decreased, and the transition speed of the cell is therefore increased. The room for optimization is proportional to the differences of the current and physical area of the PDN and PUN and the total area of the standard cells. The larger the standard cell and the more difference the

PDN and PUN have in terms of area and current, the larger the optimization room is.

Using the area-constrained current optimization, the standard cells can be optimized without area penalty and the worst-case current is increased due to the reverse short channel effect and the balancing. The current optimization process compares currents of PDN and PUN with the sizing pairs which fulfill the balancing criteria. The sizing pair which can generate the maximum current of the PDN and PUN is the output of the optimization.

In the above two optimization settings, the standard cell area is taken into account as a constraint or as an optimization target. The drawback is that, when the starting standard cell area is close to the minimum cell area, for example, an inverter cell with zero drive strength (the area of inverter with zero drive strength is very close to the minimum size allowed by the technology), the sizing room is limited when the area is not allowed to increase, therefore, the improvement is limited. The optimization improvement of the smaller standard cells is less compared to higher drive strength cells or cells with complex topologies.

There are two different methods to find the balanced equivalent PDN and PUN pair. One is a current-based balancing method, which balances the PDN and PUN in general. The other one is a transition-based balancing method, which balances the worst and best transition paths of the PDN and PUN. The two different balancing methods are explained in the next section.

4.2.2 Current-Based Balancing Method

The sizing method balances the currents of the PDN and PUN in general. Any two PDNs and PUNs are treated like an equivalent inverter with one NMOS and one PMOS transistor. The input vector and the cell topology are not taken into account for all the standard logic cells. Two examples are used to show the optimization process.

The simulation set-up used to compare the current and timing before and after optimization is illustrated in Figure 4.4. The target cell marked with lines is the cell used to measure the current and delay. Four copies of the target cell are used as the output loading capacitance at each stage. Three copies are used at the input node of the target cell to regulate the input slope. Monte-Carlo simulation with 1000 iterations is used to generate the distribution parameters.

The simulation is done at 25°C with a commercial foundry model including the global and local variations.

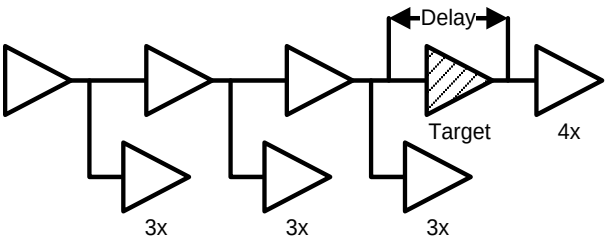


Figure 4.4 Simulation set-up used for sizing optimization.

Consider an inverter as an example of the current sizing method. An inverter with twice the size of the minimum area inverter is chosen. Two optimizations are used. One is based on area optimization and the other on current. The comparison is carried out at two technology nodes, 90nm and 40nm. As way of comparison, in Table 3, the new sized inverter is compared against a regular inverter sized for super-threshold operation but characterized to work in sub-threshold. The optimization is constrained to the area and current capability of the conventional inverter.

Table 3 Inverter sizing example at 90nm

	Area Constraints		Current Constraints	
	Sub-threshold Sizing	Super-threshold Sizing	Sub-threshold Sizing	Super-threshold Sizing
Size (W/L) [μm]	N: 0.3/0.2 P: 0.31/0.36	N: 0.6/0.1 P: 0.8/0.1	N: 0.4/0.2 P: 0.41/0.26	N: 1.2/0.1 P: 1.6/0.1
Area [μm^2]	0.14	0.14	0.19	0.28
E[I] [nA]	63.89	38.43	69.30	69.68
Std[I]/E[I]	31.27%	41.57%	32.12%	41.68%
100% yield [ns]	27.4	64.2	27.2	44.6

It can be easily seen that when constraining the area, the optimized inverter has 1.67x higher mean current when compared to the super-threshold sized cell. When the current is constrained, the area of the optimized cell is 1.5x

smaller, and the speed at 100% the yield point is increased by about 1.6x. The optimized inverter has 10% smaller variation than the conventional library cell.

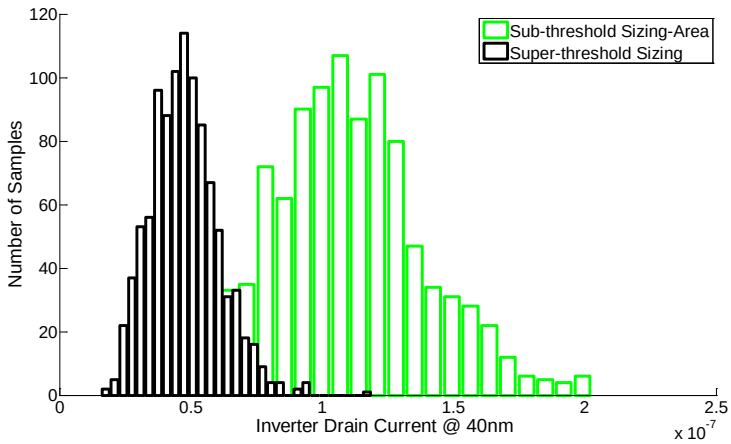


Figure 4.5 Comparison of the current distributions using an inverter sizing example (40nm CMOS technology).

In the 40nm technology node, a similar optimization process is repeated for the inverters. The area constraint optimization results are compared with the super-threshold sizing results in Figure 4.5. The green color represents the optimized inverter and the black color represents the super-threshold sizing inverter. The mean current improvement is 2.3x, while the variation and the area of the cell is the same before and after the optimization.

The comparison of the transition delay (average of the high-to-low delay and low-to-high delay) among the super-threshold sized inverter, the optimized inverter constrained by area, and the optimized inverter constrained by current is shown in Figure 4.6. The mean delay improvement of the inverter with the same area with regard to the super-threshold one, is 2.2x. The delay improvement of the area constrained optimization is because the current at the output node is improved. Hence, with the same loading capacitance (four copies of the measured inverter), the delay is improved. The improvement of the inverter with the constrained current is 1.6x. This improvement also has 30% area savings with regard tothe conventional inverter. As the gate capacitance is proportional to the area, with the FO4 measurement set-up shown in Figure 4.4, the loading capacitance of the current constrained inverter is decreased. The delay improvement comes from the decrease of the loading capacitance.

The inverter sizing examples compare the optimization results at two different technology nodes. At both 90nm and 40nm nodes, the proposed current sizing method improves the inverters in terms of speed, current, or area based on different constraints and optimization target settings.

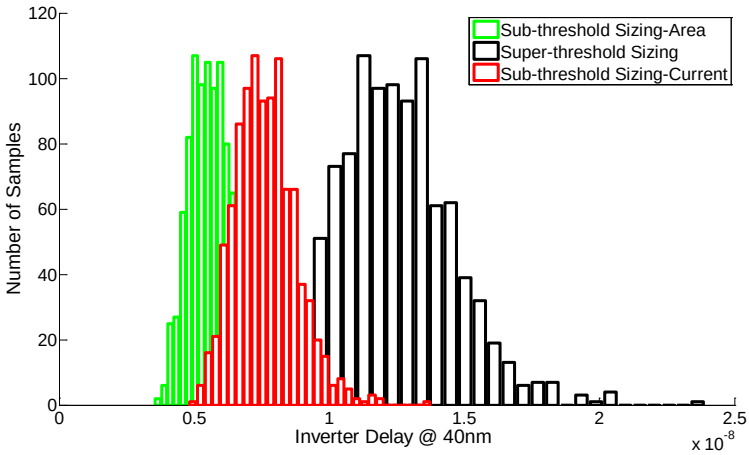


Figure 4.6 Comparison of the delay distributions comparison using an inverter sizing example (40nm CMOS technology).

4.2.2.1 NAND Sizing Example

In the inverter example, it is easy to determine the PDN and the PUN for the balancing optimization process. In the NAND gate topology, for a simple two-input NAND gate, the PMOS transistors are parallel connected and the NMOS transistors are in a stack topology. The equivalent sizes of the PDN and PUN need the stack and parallel sizing equations (2.10) to (2.15) of Chapter 2.

Finding the equivalent transistors of the PDN and PUN is simple. The equivalent size of the PDN and PUN for a two input NAND gate is shown in Figure 4.7. The sizing parameters β and ϱ are based on the technology nodes and the supply voltage. In the current balancing method, the on/off state of different transistors within the PDN or PUN is not considered. The equivalent transistors of the PDN and PUN are determined based on the fact that all the transistors within the PDN and PUN are on. The size of the equivalent transistors of the PDN and PUN is shown in Figure 4.7. The equivalent size of the PUN is calculated by two PMOS transistors of the same size and connected in parallel based on Equation (2.15). The equivalent size of the PDN is calculated by two serial connected NMOS transistors based on Equation (2.11).

With the equivalent sizes of the PUN and PDN, the NAND gate is translated into an inverter-like structure. The NAND gate can be optimized following the same sizing method stated in the previous section.

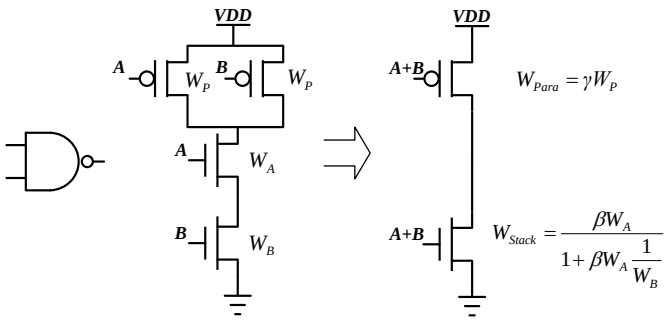


Figure 4.7 The current sizing example using a two-input NAND.

4.2.3 Transition Based Balancing Method

In the NAND gate sizing shown in section 4.2.2.1, the way to translate the NAND gate into an inverter-like structure is based on an assumption that all the transistors within the PDN are on and the transistors within PUN are off or vice versa. For multiple input logic signals, different input patterns lead to different transition paths from supply/ground to the output node. The assumption used in the current-sizing method cannot guarantee to find the worst-case and best-case transition paths. An input pattern analysis is included in the transition-sizing method to find the worst-case and best-case transition paths.

4.2.3.1 Input Pattern Analysis

For a two-input NAND gate as shown in Figure 4.7, there are six input patterns as shown in Table 4.

Table 4 Two-input NAND gate input pattern and current analysis

Input pattern	Output	Current at output node
00→11	1→0	$I_{onStack} - 2I_{offP}$ [a]
01→11	1→0	$I_{onUp} - I_{offP}$ [b]
10→11	1→0	$I_{onLow} - I_{offP}$ [b]
11→00	0→1	$2I_{onP} - I_{offStack}$
11→01	0→1	$I_{onP} - I_{offUp}$
11→10	0→1	$I_{onP} - I_{offLow}$

The red text shows that the worst pull-down delay occurs when the input vector changes from 00 to 11, while the worst pull-up delay occurs when the input vector changes from 11 to 10. Thus, for a 2-input NAND gate, the balancing is applied between one PMOS transistor and the stack NMOS transistor as the worst-case balancing criteria. In this method, the best-case paths are weakened to compensate the worst-case paths, such that the worst-case paths are optimized without area overhead.

The sizing criterion can be derived with the help of the stack and parallel sizing models. By balancing one active PMOS transistor with the NMOS stack, the sizing ratio is described in Equation (4.5).

$$\frac{W_{stack}}{W_p} = \alpha e^{\frac{E[V_{thn}]-E[|V_{thp}|]}{nU}} e^{\frac{Std^2[|V_{thp}|]-Std^2[V_{thn}]}{2(nU)^2}} \tag{4.5}$$

With the sizing ratio above, the two-input NAND gate can be further optimized.

Monte-Carlo simulations were done for a super threshold library NAND cell (characterized at low voltage) and the optimized NAND cell. Three stages at the input and a FO4 structure are used to measure the transition delay. The three stages at the input node are used to rule out the influence of a steep input slope. The results are shown in Figure 4.8 for 90nm and Figure 4.9 for 40nm. The cumulative distribution function (CDF) curves are presented for both the best and the worst case delays. The optimization procedure reduces the biggest current to compensate the smallest current through the PDN and PUN. Therefore, the best case delay is slightly increased, while the worst case delay is reduced. The advantage of using the statistical distribution of the current for sizing purposes is that one can optimize the worst case current without any area penalty.

In Figure 4.8 and Figure 4.9, the solid and dashed lines of each color serves as the envelop curves of all the transition delays for the two-input NAND gate. As shown in both figures, the solid blue lines are on the right side of the red solid lines, which means that the best-case transition delay is increased after optimization. Further, the dashed blue lines are on the left side of the dashed red lines, which means the worst-case transition delay is improved by the optimization. After the transition-based sizing optimization, the best- and worst-case transition envelopes gets smaller, the difference between the best-case and worst-case transition delays is decreased.

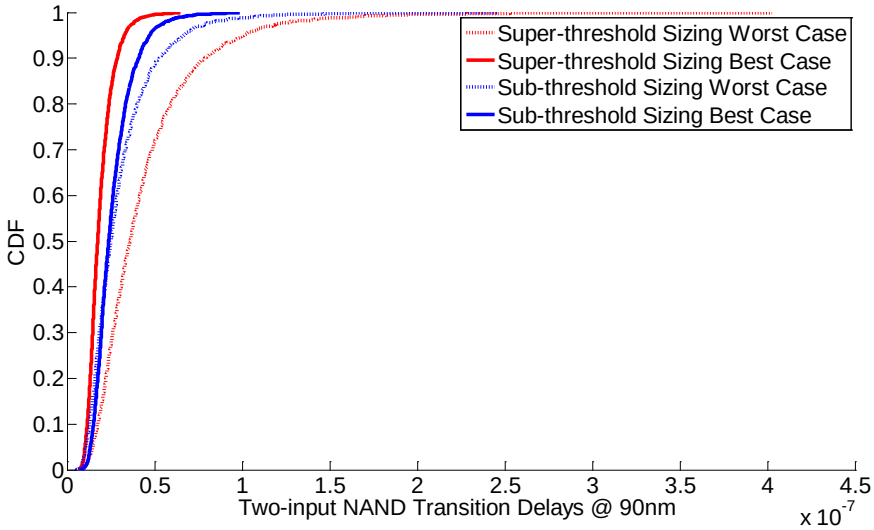


Figure 4.8 The best- and worst- case transition delays of a NAND gate (CMOS 90nm).

In the 90nm technology node, the mean delay of the worst-case transition decreases from 43.87ns to 29.95ns, the standard deviation decreases from 31.08ns to 19.03ns, and the variation (standard deviation/mean) decreases from 70.85% to 63.54%.

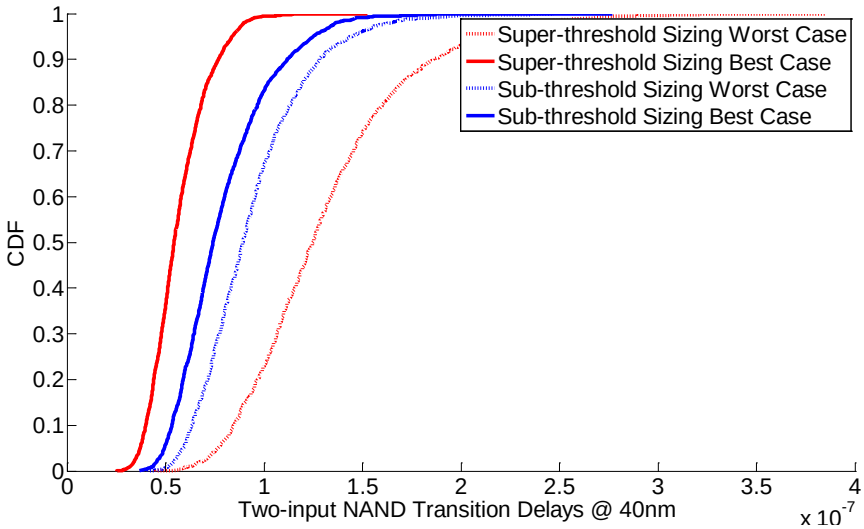


Figure 4.9 The best- and worst- case transition delays of a NAND gate (CMOS 40nm).

In the 40nm technology node, the mean delay of the worst-case transition is improved by 28%, and the standard deviation is improved by 35% at the cost of increased best-case transition delay.

Comparing the CDFs of the 40nm NAND gate to the 90nm NAND gate, we note that the difference of the best-case and worst-case in 40nm NAND gate is larger at different yield points (except at 100% yield point). The bigger the difference is, the more optimization room the cell has. From 90nm to 40nm, for a NAND gate, the optimization method can generate cells with better worst-case performance at low voltages compared to the conventional super-threshold sizing method. A similar improvement can be expected in smaller technology nodes.

4.2.3.2 NOR Gate Example

A two-input NOR gate is used as another example of the transition sizing method (see Figure 4.10). The relationship of the NMOS and PMOS transistors within the PDN and PUN is opposite to the NAND gate. Based on the transition sizing method, the balancing happens between the PMOS transistor stack and one single NMOS transistor.

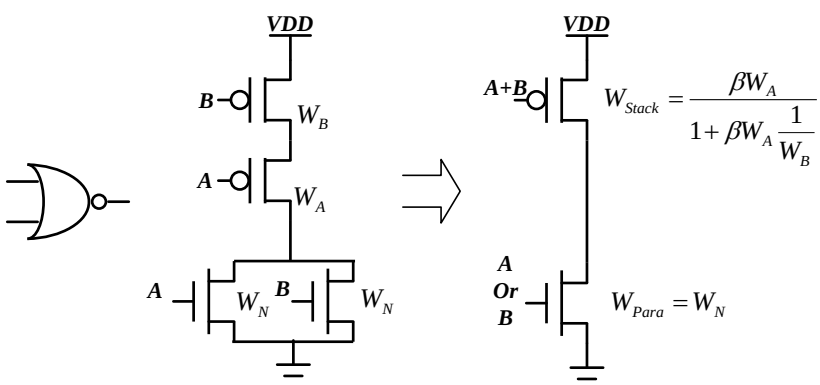


Figure 4.10 The transition sizing example using a two-input NOR gate.

The results of the NOR gate optimization are shown in Figure 4.11. After the optimization, at 0.3V, the target voltage, the mean delay of the worst case transition is improved by 36%, and the standard deviation is improved by 35%.

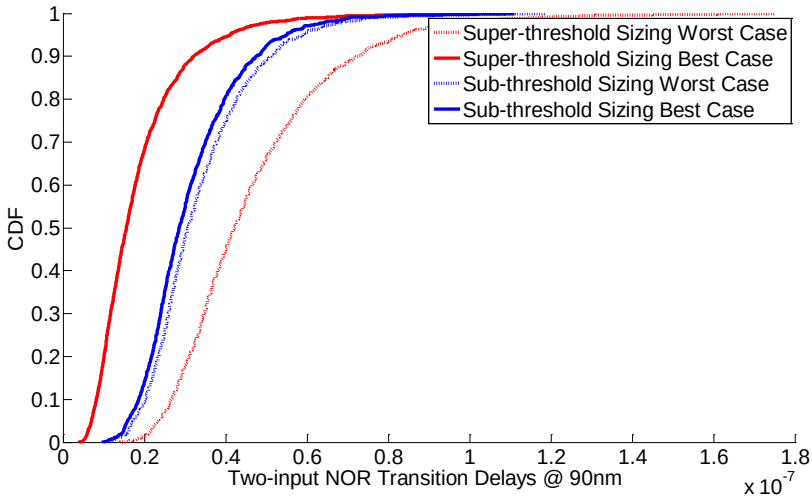


Figure 4.11 The best- and worst- case transition delays of a NOR gate (90nm CMOS).

4.2.3.3 Complex Gate Translation

The transition sizing method is explained in the NAND and NOR gate examples. In the standard cell library, the topology of the gates often involves combinations of parallel and serial connections. In this section, a complex gate is chosen to explain the translation theory: how to find the equivalent transistors of the PDN and PUN with combinations of parallel and serial connections. The schematic of the gate is shown in Figure 4.12.

Before finding the equivalent transistors for the PDN and PUN, it is important to determine the sizing ratio of different transistors within the PDN or PUN. The principle to find the sizing ratio within the PDN or PUN is to maximize the current drive of the PDN or PUN.

In the sizing process within the PUN and PDN, the starting point can be set at any transistor within the network. The proposed method is two-step iteration. First is to find if there is any direct parallel connection, such as the NMOS transistor B and C in Figure 4.12, for parallel connected transistors, applying the sizing rule R_p as shown in Algorithm 4. The second step is to find the path which has the most transistors on it. For example, in PUN, shown in Figure 4.12, from VDD to the output node, there are two paths: one has PMOS transistor A and D on it, the other one has B, C, and D on it. The sizing rule $\boldsymbol{R_s}$ is applied to PMOS transistors B, C, and D. Note that, the sizing rules are only applied to two transistors.

To find the sizing ratio within any complex cell, a translation algorithm that contains two rules to reduce the parallel and serial connection is used. The algorithm is shown in Algorithm 4.

Algorithm 4 Complex cell translation algorithm.

Algorithm 4: Parallel connected transistors qualify for the sizing rule (***R**P*) of $W_{Para} = \gamma(N)W_{one}$ and $(N) = Ne^{\frac{N^2}{\rho_p}}$.

And serial connected transistors qualify the sizing rule (***R**s*) of $W_{StackN} = \frac{\beta W_L}{\left(1+\beta W_L\left(\sum_1^{N-1}\frac{1}{W_i}\right)\right)}$. There are *n* transistors in parallel and *m* transistors in series.

BEGIN Complex Sizing
    Do
        If (n transistors in parallel**)**
            Follow ***R**P* for sizing
        Else
            If (m transistors in stack**)**
                Follow ***R**s* for sizing
            End if
        End if
    Until no complex gate connection
END Complex sizing

Let us consider now the example of Figure 4.12. The initial size of B and C are set to be the same and equal to W_N. Then, the size of the equivalent NMOS transistor of B // C is γW_N as calculated through Equation 2.15. With transistor A in series connection, the size of A become $\gamma W_N/\sqrt{\beta}$ through Equation 2.10. The equivalent transistor size of A, B // C is defined by Equation 2.10 as $\gamma W_N/\beta(\sqrt{\beta}+1)$. The size of transistor D is equal to the size of the equivalent parallel-connected transistors. A similar procedure can be followed to size the transistors of the PUN.

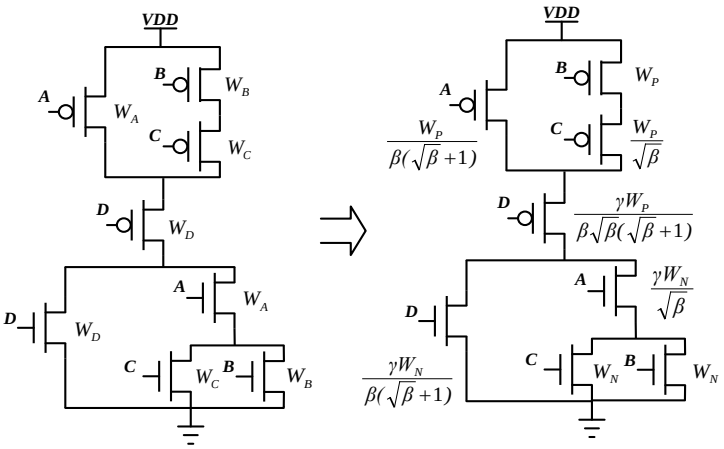


Figure 4.12 The schematic and sizing ratio of an example complex gate.

After determining the sizing ratio within the PDN and PUN, the logic cell shown in Figure 4.12 can easily be translated to an inverter-like structure based on different sizing methods, like the current sizing or transition sizing methods shown in the previous sections.

In Figure 4.13, two sets of balancing pairs are shown. In *case I*, the best PUN transition path and the worst PDN transition path are shown, and in *case II*, the worst PUN transition path and best PUN transition path are shown.

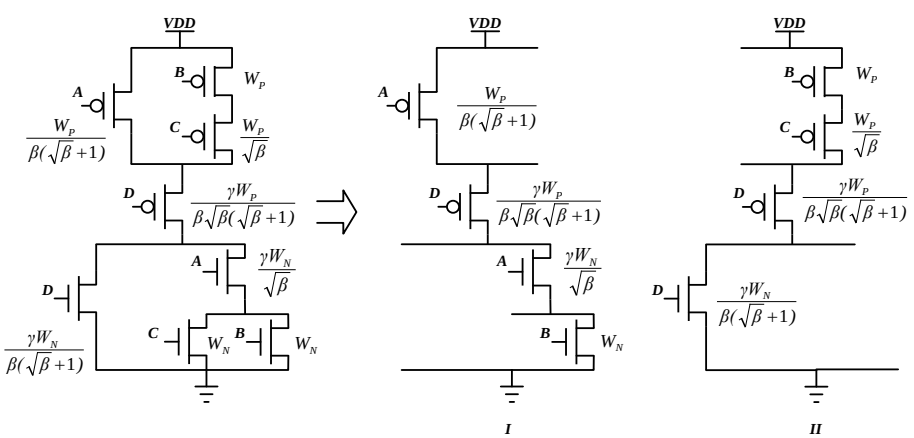


Figure 4.13 The equivalent pairs of the example complex gate.

In Figure 4.14, the two sizing methods, current balancing and transition balancing, are compared using the complex gate of Figure 4.13. One thousand Monte Carlo trials are used to generate the delay distribution. The worst-case transition CDF curves of the complex gate and the optimized complex gate are shown in Figure 4.14. The red dotted line represents the super-threshold sized complex gates, the green dashed line shows the worst-case transition delay sized by the current balancing method, and the blue solid line shows the result from the transition balancing method. The mean delay of the worst-case transition is reduced to 79% and 67% by the current balancing and transition balancing methods, respectively. The maximum delay of the worst case transition is reduced to 81% and 65%, accordingly. For the delay spread, the savings are 81% and 65%, respectively. Comparing the current balancing method and the transition balancing method, the transition balancing method has 16% less mean delay, 21% less maximum worst-case delay, and 20% less variation.

4.2.3.4 Flip-Flop Sizing

The approach to size Flip-Flops is different than the approach to size combinational logic cells due to the feedback loop for data retention. Therefore, it is important to analyze the operation of a Flip-Flop, and to understand the relationship between the different timing parameters, setup and hold time, and the sizing of various transistors in the Flip-Flop.

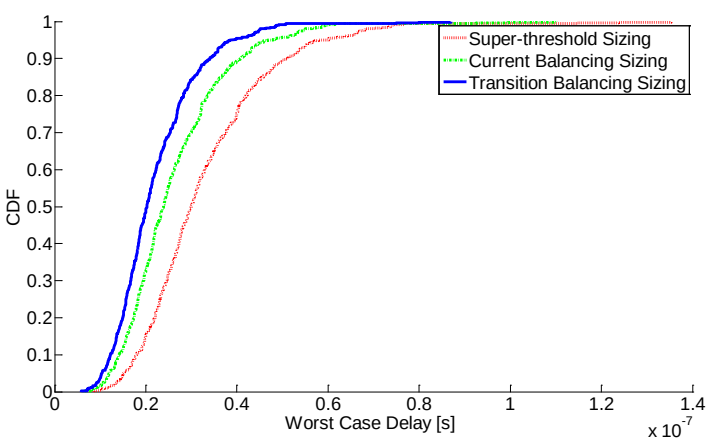


Figure 4.14. Comparison of the worst-case delay distributions of the complex cell (CMOS 40nm).

4.2.3.4.1 Flip-Flop Operation

In Figure 4.15, a basic D Flip-Flop schematic is shown for presentation purposes. In this D Flip-Flop, a transmission gate is connected to the clock trigger signal, and two back-to-back inverters are used to latch or retain the logic value.

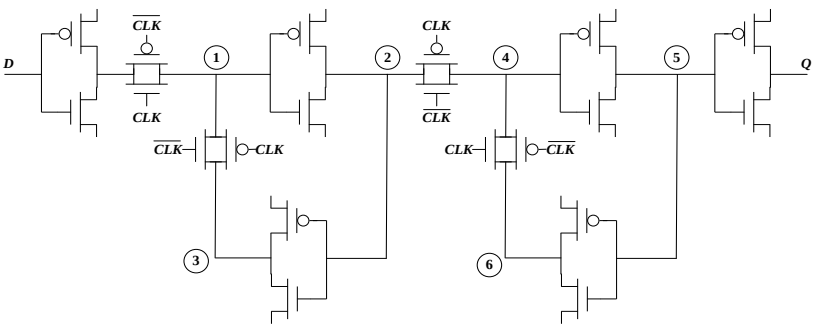


Figure 4.15 The schematic example of a D Flip-Flop.

The operation of the Flip-Flop is demonstrated in Figure 4.16. The dark line shows the actual transmission path at different operation stages. As shown in Figure 4.16 (a), when the clock signal (noted as CLK) is low, the logic signal at input node D is transmitted to node 3 through nodes 1 and 2. When CLK is high, as shown in Figure 4.16 (b), the path between input node D and node 1 is closed, and the logic value stored at node 3 is transmitted via nodes 1, 2, 4, and 5 to the output node Q and node 6.

When the CLK is low again, as shown in Figure 4.16 (c), the latching circuit within nodes 4, 5, and 6 is enabled. The value at the output node Q will not change. The value at the input node D is stored in the latching circuit within node 1, 2, and 3. When the next CLK is pulled up, the value will be transmitted to the output node. In summary, if D changes, the change would reflect only at node 3 when CLK is low and it would appear at the output only when the CLK is high.

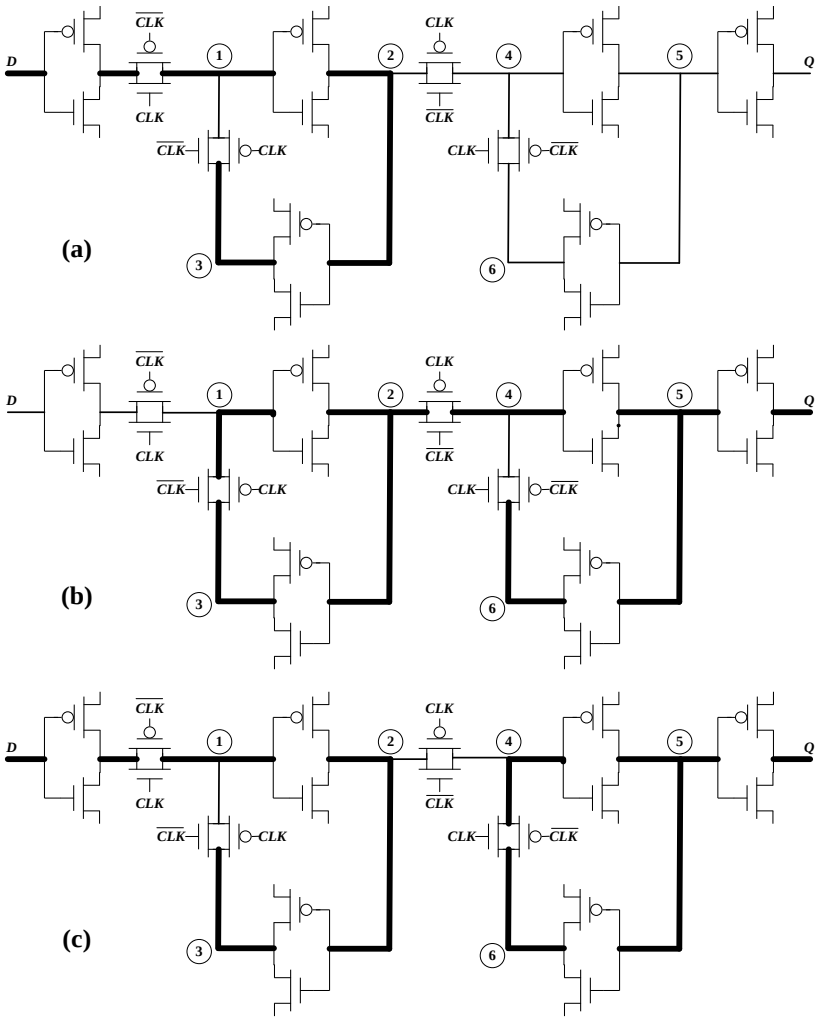


Figure 4.16 The operation stages of a Flip-Flop (a) setup time related path, (b) when clock is high, and (c) when clock is low.

After understanding the distinct conducting paths of the different operations of the Flip-Flop, the relationship between the timing parameters and the transistors in the Flip-Flop can be concluded, and further analysis on the key timing parameters, setup and hold time, are shown in the next sections.

4.2.3.4.2 Sub-threshold sizing for Setup Time

Setup time is defined as the minimum time that the data must be stable at the input node of the flip flop before the clock edge arrives. In the schematic view, it can be seen that, when CLK is low, the setup time is actually the time that it takes data at node D to propagate to node 3 as emphasized by the darkened line shown in Figure 4.16 (a). When CLK is high, the path between D and node 1 is off, and the path between node 3 and node 1 is on, so that the latching is enabled and the data is thus properly stored.

The proposed sub-threshold setup time sizing involves the gates on the conducting path D-1-2-3. Each gate serves as the output load of the previous gate. In sub-threshold, because the current drive is very small compared to the current in nominal voltage, the large leakage current of the PMOS transistor in path 2-3 fails the flip-flop operation as shown in [26, 34]. Therefore, it is important to balance the inverter on path 1-2 and the inverter on path 2-3 to maintain a correct operation with the consideration of the loading capacitance of each path. To include the loading capacitance, the balancing criterion Equation (4.1) is extended,

$$E[Delay]_{path1} = E[Delay]_{path2} \tag{4.6}$$

where

$$E[Delay] = \frac{(kT/q)F\left(\frac{V_{gs}}{kT/q}\right)C_{Load}e^{-V_{gs}+E[V_{th}]/nkT/q+Std^2[V_{th}]/2(nkT/q)^2}}{\mu C\frac{W}{L}}$$
$$F\left(\frac{V_{gs}}{kT/q}\right) = log\left(\frac{1-e^{V_{gs}/(kT/q)}}{1-e^{V_{gs}/2(kT/q)}}\right) \tag{4.7}$$

The loading capacitance of the inverter on path 1-2 consists of two inverters and one transmission gate, while the loading capacitance of the inverter on path 2-3 consists of one inverter and one transmission gate. After the extended balancing is applied, the difference between the two different loading capacitances makes the inverter on path 1-2 stronger than the inverter on path 2-3. In [26, 34], a similar Flip-Flop sizing strategy can be found. The difference is that the proposed method balances the delay instead of the current drive. Note that the rise and fall transitions of the transmission gates are balanced before applying the extended balancing.

4.2.3.4.3 Sub-threshold sizing for hold time

Hold time is defined as the minimum amount of time after the clock edge arrives in which data must be stable. Hold time margin is also an important parameter for Flip-Flops. The hold time involves the time to switch on the transmission gate (T_{TG}), and the time it takes data at the input node to reach the transmission gate (T_D), as shown in Figure 4.17. Note that CLK and CLK BAR often have a finite delay. Normally, an inverter is used to generate the CLK BAR signal, which sums up with the conducting time of the transmission gate is the T_{TG}. The gates used to generate the CLK BAR also needs proper sizing when sizing for hold time.

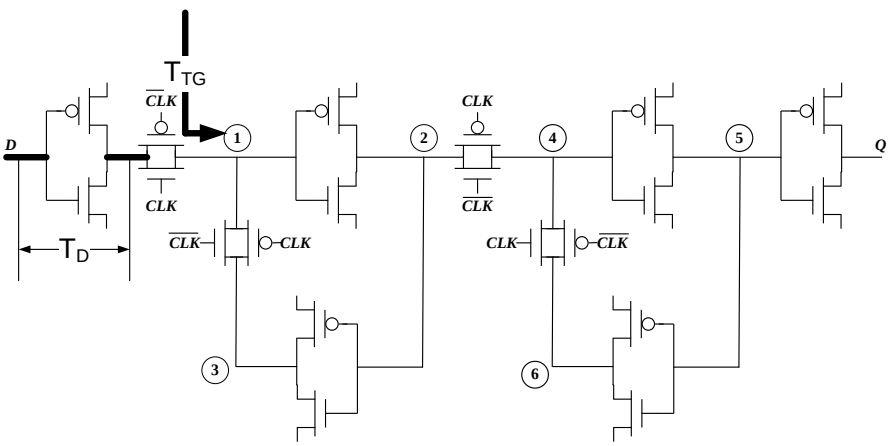


Figure 4.17 The hold time related path of a Flip-Flop.

Unlike the setup time which always has a positive delay, the hold time can be positive, zero or negative because of different combinations of T_D and T_{TG}. When T_{TG} is greater than T_D, the hold time is positive; when they are equal, the hold time is zero. Normally, T_D is related to the function of the Flip-Flops, such as, set-reset-enable, scan-enable. Based on different functions of the Flip-Flops, the hold time margin can be adjusted accordingly.

In the proposed sub-threshold library, the flip-flop is tuned to have negative hold time by making T_D bigger than T_{TG}. In the shown example, the transmission gate between D and node 1 is the fastest transmission gate among all the other transmission gates. Negative hold time helps to avoid hold time issues caused by the slow clock input when scaling the voltage from sub-threshold region to super-threshold region.

4.2.3.4.4 Comparison of sizing results in 40nm

The Flip-Flop shown in the previous section is used to compare the difference between the conventional super-threshold sizing method and the proposed sub-threshold sizing method.

First, the setup time and CLK-Q delays are compared based on corner simulation results at 0.3V with FO4 loading at the data output node. In the SS corner, the maximum setup time of both super-threshold and sub-threshold sized Flip-Flops is shown in Figure 4.18. The X-axis shows the setup time and the Y-axis shows the CLK-Q delay at of corresponding setup times.

In Figure 4.18, it can be seen that, when the setup time is above 0.50µs as shown by arrow A, changing the setup time has no influence on the CLK-Q delay of the super-threshold sized Flip-Flop. The CLK-Q delay is stable around 0.63µs. For the sub-threshold sized Flip-Flop, when the setup time is above 0.20µs as shown by arrow B, the CLK-Q delay is stable around 0.56µs. Comparing A and B, one can see that with the proposed sizing method, at 0.3V, the setup time can be improved from 0.50µs to 0.20µs, while the CLK-Q delay is improved by 12% in the SS corner. When the setup time is below 0.20µs, the super-threshold sized Flip-Flop can no longer function properly; however, the sub-threshold sized Flip-Flop can still give correct output data at the cost of a small increase of CLK-Q delay. The sub-threshold Flip-Flop can tolerate 0.06µs as minimum functional setup time.

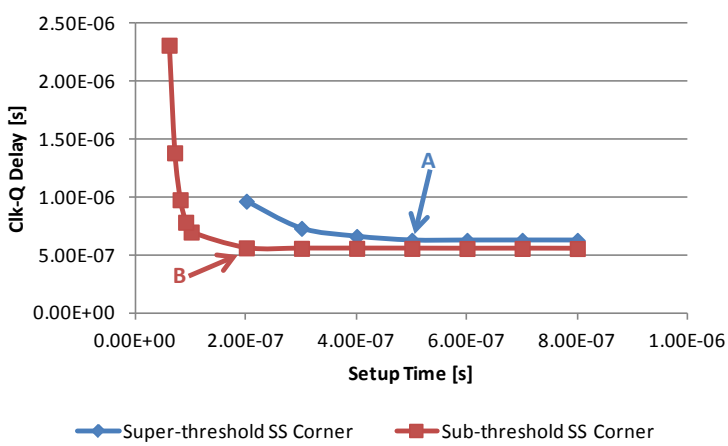


Figure 4.18 The setup time of the super-threshold and sub-threshold sized Flip-Flop comparison at SS corner.

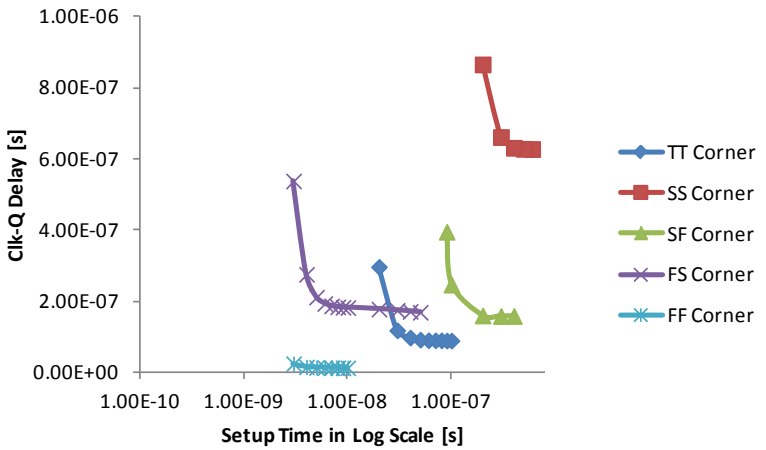


Figure 4.19 The setup time of the super-threshold sized Flip-Flop at different process corners.

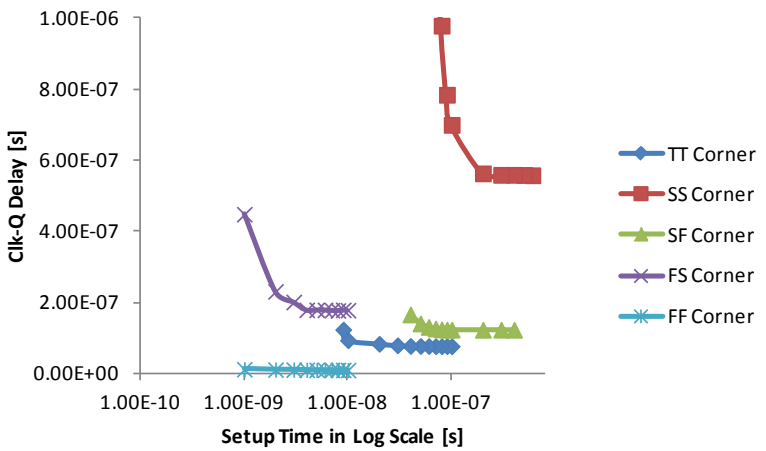


Figure 4.20 The setup time of the sub-threshold sized Flip-Flop at different process corners.

The setup time simulation results of all five corners for both the super-threshold and sub-threshold sized Flip-Flops are shown in Figure 4.19 and Figure 4.20, respectively. Note that in Figure 4.20 the Y-axis is limited so as to have the same range as the Y-axis used in Figure 4.19 in order to gain a better visualized comparison.

In Figure 4.19 and Figure 4.20, one can see that, at 0.3V the SS corner has the worst CLK-Q delay and setup time performance for both super-threshold and sub-threshold sized Flip-Flops. Comparing the two figures, it is clearly shown that the sub-threshold sized Flip-Flop has the faster and more stable CLK-Q delay as well as a better setup time tolerance. To summarize the two figures, Table 5 is used to list the minimum allowed setup time with stable CLK-Q delay, and minimum function setup time value at all five corners.

Table 5 Setup time corner simulation summary

	Minimum Setup Time [μs] (with Stable CLK-Q Delay)		Minimum Functional Setup Time [μs]	
Corner	Super-threshold	Sub-threshold	Super-threshold	Sub-threshold
FF	0.01 (0.13)	0.005 (0.10)	0.003	0.001
FS	0.01 (0.18)	0.006 (0.18)	0.003	0.001
TT	0.06 (0.09)	0.040 (0.07)	0.020	0.009
SF	0.30 (0.16)	0.080 (0.12)	0.090	0.040
SS	0.50 (0.63)	0.200 (0.56)	0.200	0.060

Among all five corners, the maximum improvement on minimum setup time (with stable CLK-Q delay) is approximately 3.75x at SF corner, and for minimum functional setup time improvement, the improvement number is 3.33x at SS corner. The stable CLK-Q delay is compared by 3000 Monte Carlo Simulation at 0.3V with a 1μs setup time. The CLK-Q delay distribution comparison between the super-threshold sized Flip-Flop and the sub-threshold sized Flip-Flop is shown in Figure 4.21.

Based on the proposed sizing method, at 0.3V, the mean CLK-Q delay can be reduced by 20% and the standard deviation can be reduced by 16%.

The hold time comparison is not shown, because based on different purposes, the hold time can be tuned from positive value to negative value to guarantee proper functionality.

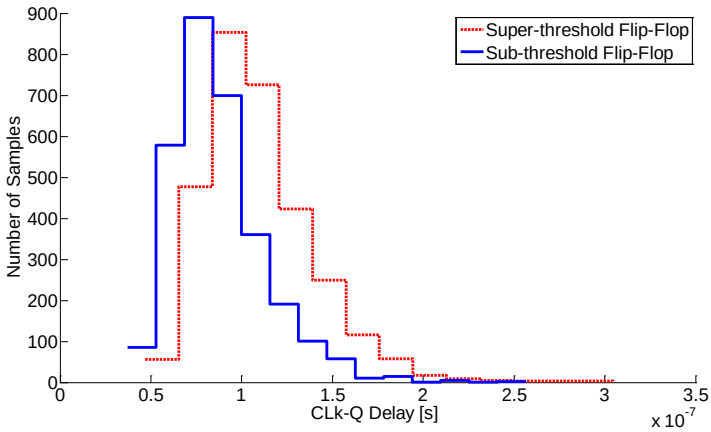


Figure 4.21 The distributions of the CLK-Q delay.

4.3 *Conclusions of Chapter 4*

In this section, the sizing method and the optimization process are explained in detail. Two sizing methods are proposed based on the transistor behavior in the sub-threshold region: one is to balance the standard cells without considering the input patterns (namely, the current sizing method); the other it to balance the standard cells with the consideration of the input patterns. Both methods render cells with faster transitions, better yields, and more robust behaviors compared to the conventional super-threshold sizing and state-of-the-art sizing. During sizing, the complex cell topology translation algorithm is used to find the equivalent transistor of different transition paths within the combinational cells. In addition to the combinational cells, the Flip-Flops are analyzed. The relationship between the transistor sizing parameters and the Flip-Flop timing parameters is studied.

CHAPTER 5

STANDARD CELL LIBRARY IMPLEMENTATION AND COMPARISONS

A variety of strategies to integrate a standard cell library from individually optimized cells operating in the sub-threshold regime is presented in this chapter. The standard cell library development conforms to the standards of the Synopsys liberty timing format. The optimized libraries are then compared against the corresponding super-threshold libraries in different technology nodes, different corners, and different voltages.

Process variation is a very important issue in the sub-threshold region. It is also an important factor in filtering the minimum number of cells that integrate the library. In Chapter 2, the variability study for stacked transistors unveiled the maximum number of series-connected transistors within a given process variation tolerance requirement. As listed in Table 1, if a tolerance of 30% to 40% is acceptable, the maximum number of transistors connected in a series is three. Thus, all cells in the library will have three or less inputs. The second filtering criterion is the drive strength. In low power applications, speed is not the primary concern. Thus, high drive-strength cells are not widely selected by the synthesis tool. Moreover, the proposed sizing method increases the drive current without increasing the area of the cells, which means that after optimization, the same drive current can be achieved by cells with smaller area. The third filtering criterion is the frequency of occurrence of the cells that are used in various types of applications. Based on a set of circuit synthesis reports and the first two filtering criteria, a set of 166 standard cells, including 109 basic logic cells (such as inverter, NAND gate, and NOR gate), 35 buffers, 10 flip-flops (different functions), 8 adders, and 4 MUX cells, is chosen to build the sub-threshold standard cell library in this work.

5.1 Layout Choices

Standard cell layout choices are studied in this section. In low power applications the physical libraries are optimized for low power and small area since

performance is not the main consideration. The height of the cells is made preferably six or seven routing tracks high to make the low drive strength cells more area-efficient for better power savings. If higher drive strength is needed, more fingers or parallel-connected transistors within the standard logic cells are used. The cells grow in width to increase their drive strength. But is it always beneficial to use low cell height (six or seven tracks) in sub-threshold region? In this section, a comparison between the width-growth and height-growth is carried out. It is shown that it is not always beneficial to keep the number of tracks as low as possible and to grow only the width of the cells in the low voltage domain.

Without loss of generality, the example of an inverter with single drive strength, shown in Figure 5.1 (b), is used to compare height and width cell layout expansion strategies. One way to enhance the drive strength from one to two is to make the inverter wider, namely Width Growing, as shown in Figure 5.1 (a). The source (drain) is shared by two parallel-connected transistors to decrease the total area of the inverter. The other way of growing is to grow in vertical direction, namely Height Growing, as shown in Figure 5.1 (c). The drain and source of two transistors are connected via Metal 1.

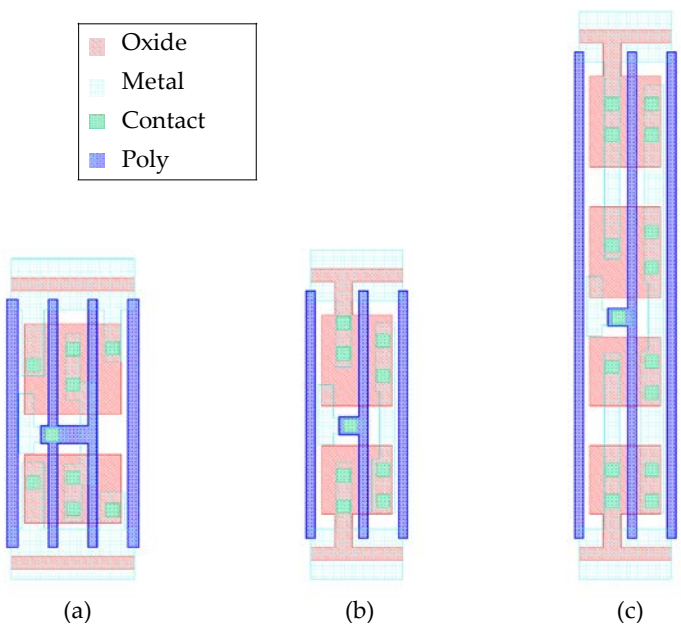


Figure 5.1 The layout view of inverters.

To compare different layouts of inverters with drive strength 2, three thousand Monte Carlo simulations are used to show the difference in terms of the delay and variation. The results of the two layout styles of a 15-stage inverter chain are shown in Figure 5.2.

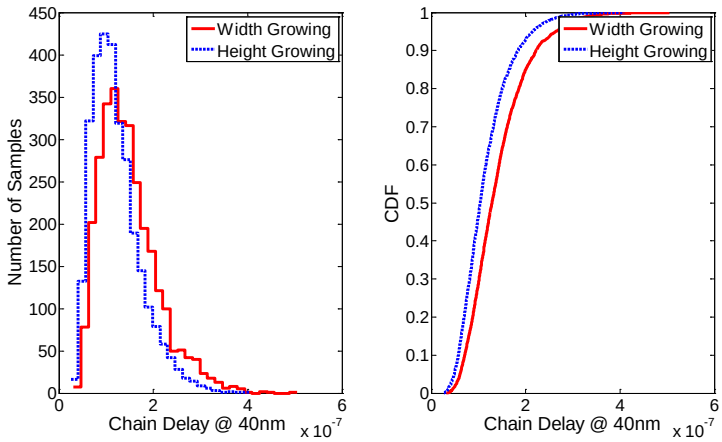


Figure 5.2 The delay distributions of two layout styles at 0.3V.

One can see that in Figure 5.2 the delay distribution of the inverter chain with Height Growing is left shifted. Based on the CDF curve, the Height Growing inveter displays a better yield. The data is summarized in Table 6.

Table 6 Comparison of inverter chains with two growing methods

	Width Growing	Height Growing	Improvement
Mean of delay [μs]	0.1536	0.1240	19%
Standard deviation of delay [μs]	0.0813	0.0683	16%
Mean+3*standard deviation of delay [μs]	0.5938	0.4955	17%
Mean of leakage power [fW]	0.0176	0.0181	-2.5%

Compared to the Width Growing method, the Height Growing method yields the 15-stage inverter chain with 19% lower mean of delay and 16% smaller standard deviation. Note that in order to calculate the (mean + 3 * standard deviation) point for the lognormal distribution, a natural logarithm is

performed to transfer the distribution to a normal distribution. Then an exponential function is applied to the (mean+3*standard deviation) of the normal distribution to calculate the correct value. Comparing the leakage power of the two methods, it can be seen that the Height Growing method has 2.5% higher leakage power at 0.3V.

The improvement of the Height Growing method is due to reduced internal capacitance. After the capacitance extraction for the two different inverters was performed, the capacitance between each node is shown in Figure 5.3.

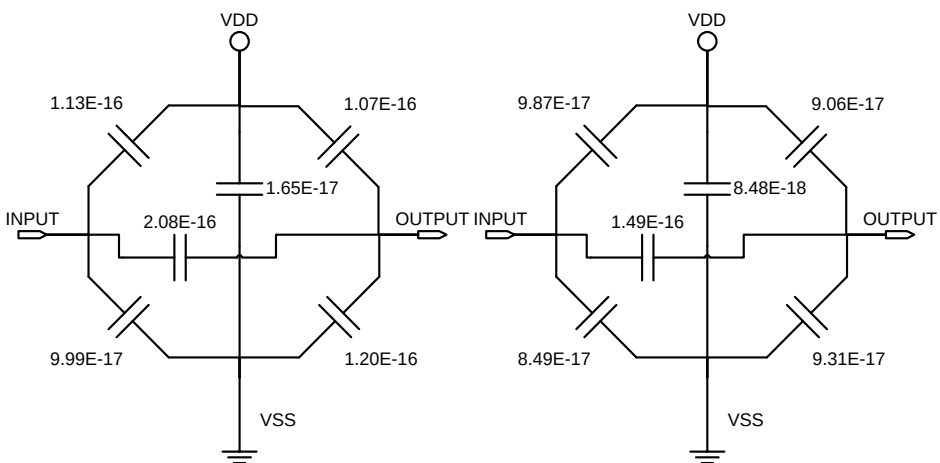


Figure 5.3 The capacitance models of Width Growing and Height Growing.

In general, the Height Growing method shown on the right side of the figure has lower capacitance between different nodes than the Width Growing method shown on the left side as shown in Figure 5.3. The delay improvement mainly comes from the lower capacitance between the input and output nodes.

Growing in either direction leads to different benefits. As stated previously, it is easier to make the area of the cells with multiple drive strengths more compact with width growth because of the shared drain/source region. Growing in height has delay and variation benefits because of the reduced internal capacitance. When increasing the drive strength of the cells with the Height Growing method, there is no shared drain/source region with adjacent transistors. The area efficiency of the Height Growing method is therefore lower compared to the Width Growing method. Furthermore, since the cells grow in

the vertical direction, routing for power connections is as easy as the cells following the Width Growing method. However, since the height growing cells are higher than the width growing cells, the number of routing tracks is increased.

When choosing between Width Growing and Height Growing, another parameter that needs to be considered is the number of routing tracks in the given technology. Normally, the number of tracks is proportional to the cell height. The comparison of Width Growing and Height Growing helps to determine the optimum number of tracks (cell height) in low voltage domain. It is clearly shown that the Height Growing approach brings benefits in delay and variation improvement. When drawing the layout of the standard cells in the low voltage domain, the height of the cell should be as high as possible for delay and variation purposes. In terms of number of tracks, the more tracks the better. From the parallel transistor analysis in Chapter 2, a finger structure is also very efficient in delay and variation improvement. Together with the Height Growing method, using small transistors with multiple fingers within the highest allowed cell height is a promising choice for low power standard cell sizing.

5.2 Library Characterization

After the sizing parameters are chosen based on the optimization and the layout choice, the physical standard cell library can be developed. The next step is to characterize the timing and power of the standard cells at different conditions in the "liberty" format [72].

In the liberty format of standard cell libraries, there are three main categories of parameters: timing slew, loading capacitance, and flip-flop related parameters. In this section, these three different sets of parameters are explained together with strategy on how to determine those parameters for different cells.

5.2.1 Timing and Power Model

In this work, a Non-Linear-Delay-Model (NLDM) [72] is used to specify the standard cell library. The NLDM model for delay/power is presented in a two-dimensional matrix table, with the two independent variables being the input transition time and the output loading capacitance. The entries in the table denote the timing delay/power. The model shown in Figure 5.4 is an example of such a table for the delay of the output pin of an inverter cell.

```
  pin (Z) {
    direction : output;
    max_transition : Value;
    max_capacitance : Value;
    timing () {
      related_pin : "IN";
      timing_sense : negative_unate;
      timing_type : combinational;
      cell_rise (delay_template_7x7) {
        index_1 ("S1, S2, S3, S4, S5, S6, S7");
        index_2 ("C1, C2, C3, C4, C5, C6, C7");
        values ( \
                 "T11,T12,T13,T14,T15,T16,T17", \
                 " T21,T22,T23,T24,T25,T26,T27", \
                 "T31...T37", \
                 "T41...T47", \
                 "T51...T57", \
                 "T61...T67", \
                 "T71,T72,T73,T74,T75,T76,T77" \
        );
 }
```

Figure 5.4 An example of NLDM model

In the example shown in Figure 5.4, the two dimensional matrix table of the timing delay of the output pin Z is described. This portion of the cell description contains the rising delay models for the timing arc from pin IN to pin Z, as well as the max_transition and max_capacitance values at pin Z. There are separate models for the rising and falling delays (for the output pin), which are labeled as cell_rise and cell_fall, respectively. The type of indices and the order of table lookup indices are described in the lookup table template delay_template_7x7. The timing_sense has negative_unate and positive_unate. Negative_unate means cell output logic is the inverted version of input logic. In inverter Z is negative unate with respect to IN. Positive_unate means cell output logic is the same as that of the input.

This look-up table template specifies that the first variable in the table is the input slew (index_1), and the second variable is the output loading capacitance (index_2). There are seven entries for each variable, thereby resulting in a 7-by-7 table. In most cases, the entries for the table are also formatted like a table. The first index (index_1) can then be treated as a row index. The second index (index_2) becomes the column index. For example, when the input transition is S_2 and the output loading capacitance is C_3, the delay value is T_{23}.

The delay template serves as a look-up table in the NLDM model. During the actual synthesis, if the input transition time and the output capacitance match a table entry, the table lookup is trivial since the timing value corresponds directly to the value in the table. However, often enough, there is no match to any of the entries available in the table. In this case, a two-dimensional interpolation is utilized to provide the resulting timing value. The two nearest table indices in each dimension are chosen for the table interpolation. For example, for a given set with input transition time denoted as S_0 ($S_1 < S_0 < S_2$) and output load C_0 ($C_2 < C_0 < C_3$), four values in the table are sen: T_{12}, T_{22}, T_{13}, and T_{23}. The T_{00} value for S_0 and C_0 is calculated by interpolation and shown as

$$T_{00} = \frac{X_1 T_{12} + X_2 T_{13} + X_3 T_{22} + X_4 T_{23}}{X_5}$$

where

$$X_1 = (S_2 - S_0)(C_3 - C_0)$$

$$X_2 = (S_2 - S_0)(C_0 - C_2)$$

$$X_3 = (S_0 - S_1)(C_3 - C_0)$$

$$X_4 = (S_0 - S_1)(C_0 - C_2)$$

$$X_5 = (S_2 - S_1)(C_3 - C_2)$$

The table index is critical for correct interpolation. Therefore, it is very important to find the correct index range and steps for the delay and load capacitance in order to get an accurate delay/power model for backend flow.

5.2.2 Defining the Slew

In standard cell libraries, the slew rate is measured as the time it takes for a signal to transition between two specific voltage levels. The slower the slew the larger the transition time, and vice versa.

The reason to define the *correct* slew range is that the transition time for a standard cell is different at different voltages. If the input slew is defined at nominal voltage, the transition time of the cell using this slew at low voltage is inaccurate because the slew is orders of magnitude faster than the actual input slew. Based on the NLDM model, the accuracy of the timing value after inter-

polation is not the actual response, either. An accurate delay model can only be derived with the correctly aligned input slew values. In fact, the input slew value in sub-threshold can vary up to three orders of magnitude for different cells depending upon PVT conditions. In the proposed library characterization process, the slew is defined based on a simple cell and variable loading capacitance simulation setup. The process is shown in Figure 5.5.

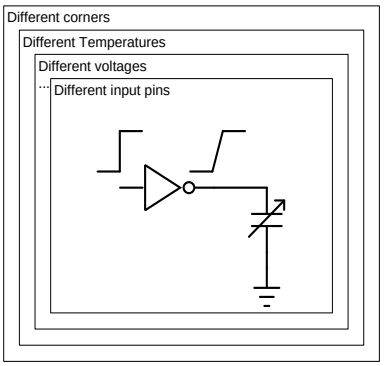


Figure 5.5 The simulation setup for defining the slew.

For each specific loading capacitance used in the simulation, a set of input slews is used to stimulate the cell. The input slew value is chosen only when the output slew matches the input slew at each specific loading capacitance. The simulations are repeated at different corners, temperatures, and voltages for different input pins to collect the slew for libraries at different working conditions. In this way, the input slew can be matched correctly with different conditions and can provide the right interpolation space for backend simulation.

5.2.3 Defining the Load

The loading capacitance is defined based on the extracted input capacitance of the standard cell. With the extraction, the internal capacitance of all the cells in the standard cell library is updated. The range of the output capacitance is defined based on the minimum and maximum capacitance of the standard cell library. The minimum capacitance is defined by input capacitance of the minimum sized standard cell. The maximum capacitance is defined by four times the input capacitance of the largest standard cell.

5.2.4 Sequential Logic

In sequential logic timing model, the delay and power values have a similar format as the combinational logics. With certain input slews and loading capacitances, a delay/power value can be found or interpolated based on the values in the NLDM matrix. The differences are the timing check models to ensure a correct logic function for sequential circuit elements. In this section, the relative timing checks that need to be carefully considered at low voltage range are discussed.

5.2.4.1 Synchronous Check

The setup and hold time checks verify that the data input is unambiguous at the active edge of the clock and that the proper data is in fact latched at the active edge. These timing checks validate whether the data input is stable around the active clock edge.

The template for the timing check is similar to the delay template. The main difference between them is the index. In the timing check template, index_1 is the input transition times of the constrained pin (data transition for setup and hold time), which is the same as the delay template. The index_2 however is also a transition index which shows the transition time of a related pin (clock pin for setup and hold time check).

The hold timing check index may show negative values. It normally happens when the path from the pin of the flip-flop to the internal latch point for the data is longer than the corresponding path for the clock. Thus, a negative hold check implies that the data pin of the flip-flop can change ahead of the clock pin and still meet the hold time check.

The setup timing check index can also be negative. However, both the setup and hold time cannot be negative at the same time. The sum of the setup and hold time defines the pulse width during which the data pin needs to be steady for proper logic operations.

5.2.4.2 Pulse Width Check

In addition to the synchronous and asynchronous timing checks, there is a check which ensures that the pulse width at an input pin of a cell meets the minimum requirement. For example, if the width of a pulse at the clock pin is smaller than the specified minimum, the clock may not latch the data properly. The pulse width checks can be specified for relevant synchronous and asyn-

chronous pins as well. The minimum and maximum pulse widths are related to the input slew range of the clock pin and data pin. Therefore, at different voltages and different conditions, the minimum and maximum pulse widths need to be determined separately.

5.3 Library Cell Comparison

Based on the proposed standard cell sizing methodology, the layout strategy, and the library characterization settings previously discussed, new standard cell libraries are composed, optimized, and characterized for low voltage operation in two different technology nodes (40nm and 90nm CMOS technologies). Commercial libraries, designed to operate in the super-threshold region in both technology nodes, are characterized now in the low voltage region and used as reference libraries.

In this section, the comparison between the proposed libraries and the reference libraries is shown at different supply voltages and process corners in terms of power and performance effectiveness. This comparison is done both in simulation using primitive gates and also verified by measurement results in the 40nm CMOS technology. More specifically, the libraries compared in this section include:

- Commercial low power libraries characterized for low voltages in 40nm LP CMOS technology and 90nm LP CMOS technology;
- Optimized libraries:
    - Width-tuned library optimized and characterized for low voltages in 90nm LP CMOS technology;
    - Width and length-tuned library optimized and characterized for low voltages in 40nm LP CMOS technology and 90nm LP CMOS technology.

The differences between the two optimized libraries are the sizing methods and optimization targets. In Chapter 2, the width and length tuning effects are studied. Recall that both the reverse short channel effect and the inverse narrow width effect help improve the speed of the transistors. The reverse short channel effect helps to decrease the variation of the transistors in sub-threshold region. Tuning the width and length has an opposite effect on the speed and

variation of the transistors in the sub-threshold region as compared to the super-threshold region.

The width-tuned only library uses the transition-based method to improve the worst-case delay of the standard cells from the sub-threshold to super-threshold region. The width and length tuned library is a full option package at low voltage. Together with the transition-based sizing method, both width and length are tuned to make use of the reverse short channel and narrow width effects in the sub-threshold region. Since at different voltages the width and length tunings have different effects, the width and length tuned library is a voltage sensitive library.

The first part of the comparison is at the cell level. The timing, power, and area parameters are compared cell by cell at low voltage to show the improvement of the proposed sizing method. The comparison based on different corners is shown in Section 5.3.1, while the comparison across different voltages is shown in Section 5.3.2.

In each standard cell library, the delay and power values are measured using circuit-level simulations at different input slews and loading capacitances. The maximum and minimum values of the slews and loading capacitance indexes differ by more than two orders of magnitude. It is not convenient to carry out a straightforward comparison in a large range. Instead, the average delay and power values are presented across all combinations of the slews and loading capacitances. From here on, the average values are referred to by the pin-delay and pin-power parameters for timing and power parameters, respectively.

5.3.1 Comparison across Different Corners

The comparison among different sets of libraries is performed in three corners: TT, FF, and SS. The comparison covers timing, power, and variation in both 90nm and 40nm technology nodes.

5.3.1.1 Timing and Power Comparison at TT Corner

Firstly, the comparison between the reference library and the optimized library is done at 0.3V at the TT corner. In the comparisons, the max cell delay is the maximum value of the pin-delay among different transitions of each cell. It actually shows the worst average transition of each cell. The corresponding pin-delay and pin-power are used to compare the power-delay product (PDP)

of each cell. The max cell delay and the max cell PDP are compared at each technology node.

In Figure 5.6, the width and length-tuned library is compared with the commercial library at TT corner with supply voltage at 0.3V. The comparison is carried out at 90nm and 40nm technology nodes. The values shown in the figures are normalized to the minimum delay or minimum PDP value of the super-threshold library at corresponding technology node.

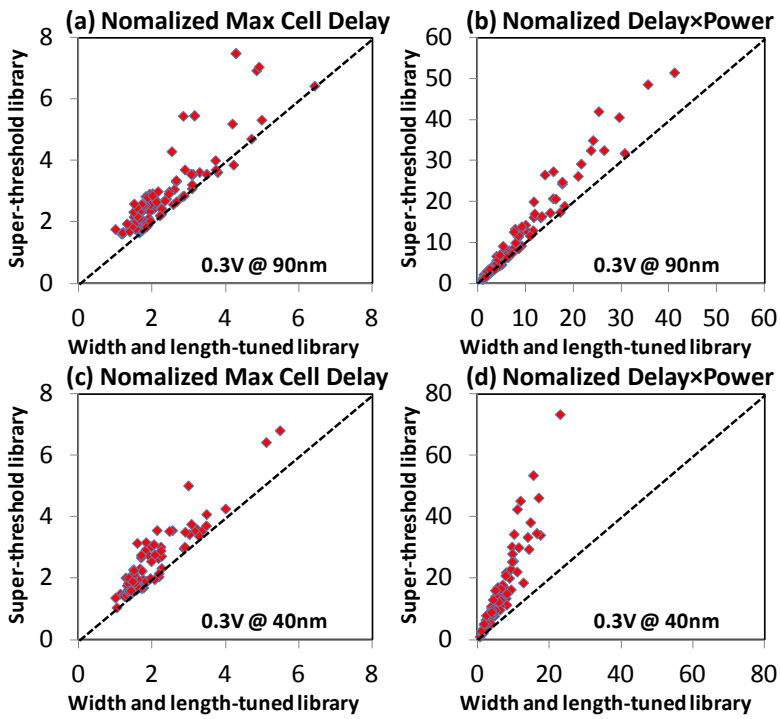


Figure 5.6 The normalized max cell delay and PDP comparison at TT corner with 90 nm and 40 nm CMOS technologies. (a) Delay comparison at 90nm; (b) PDP comparison at 90nm; (c) delay comparison at 40nm; (d) PDP comparison at 40nm.

As shown in Figure 4.5, most of the points lie above the reference 45 degree dashed line, which means that the cells from the width- and length-tuned library have better timing properties. Those cells that lie on the reference line are the minimum sized cells which cannot be further optimized by using the proposed sizing method. After optimization, the complex cells and cells with larg-

er drive strength have better performance improvement compared to the rest of the logic cells.

On average, the 90nm cells with width and length tuning have 38% better timing for worst-case transitions without introducing extra area cost. On average, the cells from the width and length tuning library achieve 31% better PDP at worst transition. In the 40nm technology node, the width- and length- tuned library cells have 49% average timing improvement for worst-case transitions and 55% better average PDP compared to the super-threshold library reference at 40nm.

5.3.1.2 Variation Comparison

Three thousand Monte Carlo simulations with both global and local variations were done for each cell in the libraries to compare their timing variation sensitivity in the sub-threshold region. The results of the delay variations in the 90nm and 40nm technology nodes are shown in Figure 5.7 and Figure 5.8, respectively. The red circles represent the width- and length-tuned library cells. The blue squares represent the cells from the commercial super-threshold library. The X-axis shows the normalized mean delay. The values of both libraries are normalized to the minimum mean delay of the cells in the super-threshold library for each technology node. The y-axis shows the delay variation value calculated by the standard deviation over the mean. The marker size of each data point is a crude indication of cell area. As known, bigger cells have smaller variation [73]. However, in Figure 5.7 and Figure 5.8, this is not always true. Most of the cells lie in the standard deviation/mean range from 50% to 70%. There is no clear indication that increasing the area will lead to variation savings in the sub-threshold region. Width- and length-tuned cells have in fact better speed and variation compared to the cells with minimum channel length and the same area.

In Figure 5.7, the width- and length-tuned cells are mainly distributed in the lower left corner, which means that the timing and the robustness of those cells are better than the cells from the super-threshold library. On average, the width- and length-tuned cells have 22% smaller variation and 1.69x timing improvement.

Similar to Figure 5.7, in Figure 5.8, the width- and length-tuned cells on average have 11% smaller variation and 2.17× timing improvement in the 40nm technology node. Among all the compared cells, the width and length tuning have maximally 45% variation savings for a two-input NOR gate NR2D1 and

4.12x maximum performance improvement for the NR2XD8 without any area penalty.

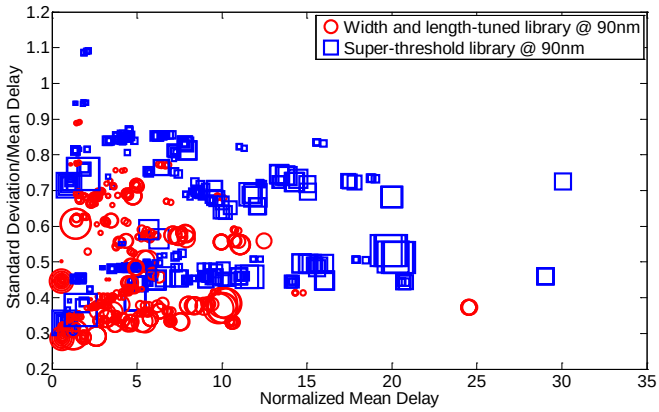


Figure 5.7 The variation of the delay and area comparison (V_{GS}=V_{DS}=0.3 V. 90nm CMOS technology).

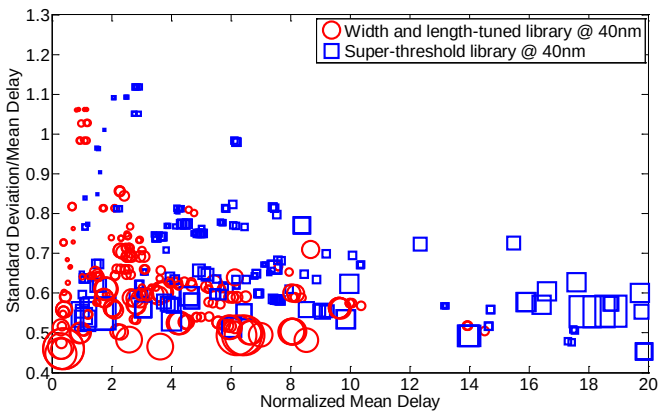


Figure 5.8 The variation of the delay and area comparison (V_{GS}=V_{DS}=0.3 V. 40nm CMOS technology).

5.3.1.3 Comparison at FF/SS Corners

The comparison of the timing and power-delay product of the standard cells in the FF and SS corners is shown in Figure 4.9 and Figure 4.10, respectively. The performance of the standard cells in the SS and FF corners determines the lower and upper bounds of the delay and power distributions. It should be

noted that the values shown in Figure 5.9 and Figure 5.10 are normalized to the same value in Figure 5.6 to compare values across different corners.

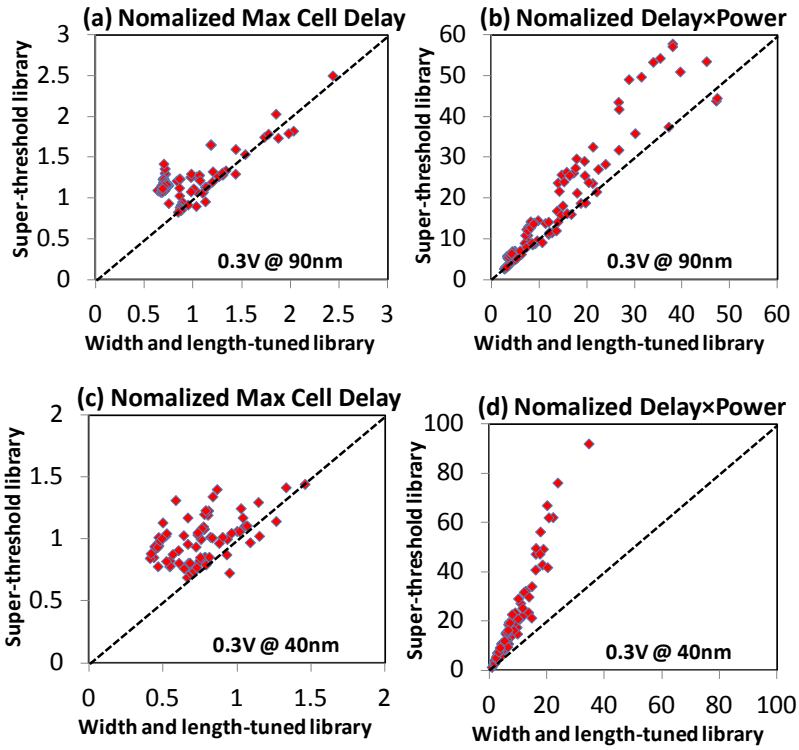


Figure 5.9 The normalized max cell delay and PDP comparison at FF corner with 90 nm and 40 nm CMOS technologies. (a) Delay comparison at 90nm; (b) PDP comparison at 90nm; (c) delay comparison at 40nm; (d) PDP comparison at 40nm.

Similar to Figure 5.6, most of the points in Figure 5.9 lie above the 45-degree reference line, which means that at the FF corner the width- and length-tuned library is faster compared at 0.3V. In the 90nm technology node, the cells from the width- and length-tuned library have on average 22% better timing and 21% better PDP. The maximum timing and PDP improvement is 94% and 51%, respectively. In the 40nm technology node, the average timing improvement is 24% while the PDP improvement is 53%. The maximum timing and PDP improvements among all the cells in the library are 56% and 71%, respectively.

In the 40nm technology node, the sizing effects in the sub-threshold region are stronger than in the 90nm technology node, thereby leading to more significant timing and power improvement.

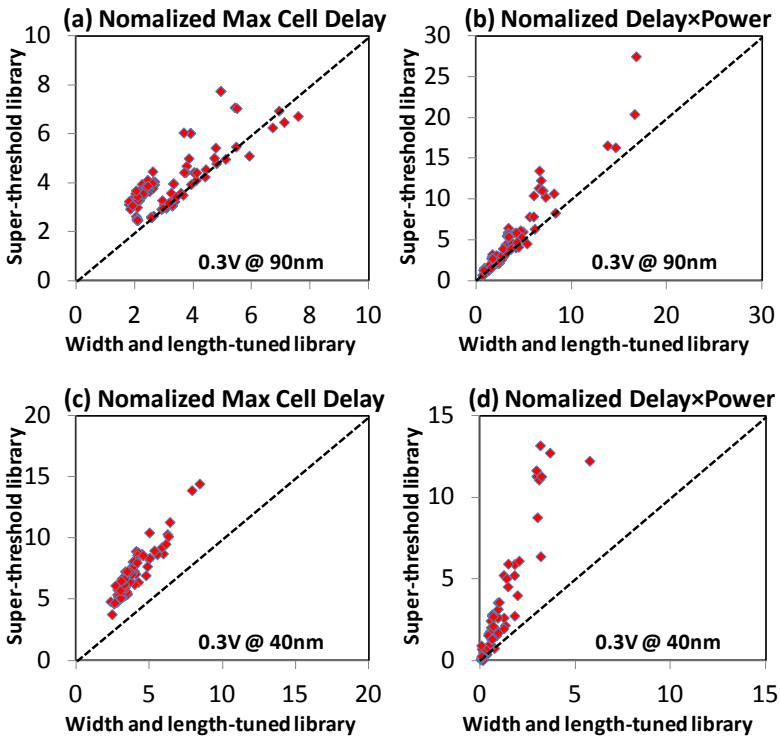


Figure 5.10 The normalized max cell delay and PDP comparison at SS corner with 90 nm and 40 nm CMOS technologies. (a) Delay comparison at 90nm; (b) PDP comparison at 90nm; (c) delay comparison at 40nm; (d) PDP comparison at 40nm.

The results for the SS corner, presented in Figure 5.10, show similar improvements to the ones shown in Figure 4.8, where most of the points are above the 45 degree reference line. In the 90nm technology node, the average timing and PDP improvements are 21% and 23%, respectively. The maximum improvements of the timing and PDP values are 44% and 50%, respectively. In the 40nm technology node, the average timing and PDP improvements are 45% and 54%, respectively.

Average timing and PDP improvements of the 90nm library in the TT corner, are 10% more than in the SS and FF corners on average. The optimization

target in 90nm was to improve the speed in the TT corner and to reduce the variation. Therefore, the delay improvement in FF and SS corners is not as high as for the TT corner. The power efficiency improvement for both FF and SS corners is similar to the improvement in the TT corner. The 10% lower delay improvement therefore results in 10% lower PDP improvement. The PDP improvements among all three compared corners are more than 50%.

5.3.2 Comparison across Voltage Ranges

The voltage scalability of distinct 90nm libraries is presented in Figure 5.11. Timing improvement is calculated by comparing the timing value of each transition of each cell with the corresponding cell in the super-threshold library. The average value of all improvements is calculated at each compared voltage. The width- and length-tuned library shows around 49% better timing at 0.3V. When the supply voltage increases to 0.65 V, the improvement drops to 0. Above 0.65 V, the width- and length-tuned library works slower than the super-threshold library. The width-tuned only library shows 10% to 11% better average timing from 0.3 V to 1.2 V when compared to the super-threshold library.

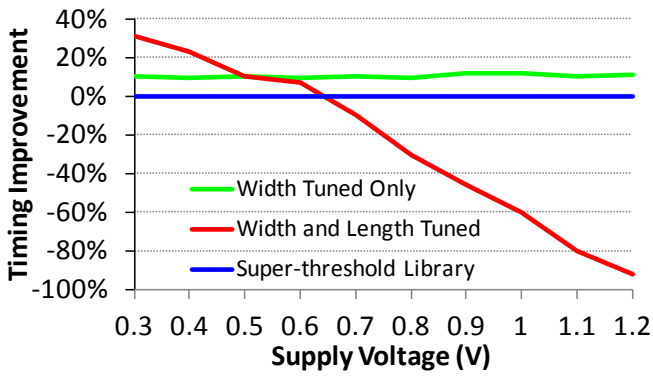


Figure 5.11 The average cell timing improvement (V_{GS}=V_{DS}=[0.3V, 1.1V]. 90 nm CMOS technology).

The variation of the three libraries is also compared using Monte Carlo simulations on extracted critical paths of the circuit C6288 of the ISCAS benchmark. The same circuit is synthesized from 0.3V to 1.2V. Note that the critical paths of the three different libraries are different. Moreover, at different voltages, the critical paths from the same library are different as well. The

mean and the lower boundary of the distribution are summarized in Table 7. The comparisons are carried out at three different voltage nodes: 0.3V, 0.6V, and 1.2V.

Besides the mean delay, an extra column is included to show the $(\mu + 3\sigma)$ of the delay. Observing that since the libraries have different mean delays, the conventional figure of merit, i.e. standard deviation over mean, does not represent the real variation of the delay. Thus, the $\mu + 3\sigma$ delay is used to benchmark the lower bound of the delay distribution. As can be concluded from Table 7, at 0.3V, the proposed width- and length-tuned library achieves a 28% faster circuit with 25% smaller variation. At 0.6V, the width- and length-tuned library has a similar behavior to the super-threshold library with respect to the delay distribution. At 1.2V, the width- and length-tuned library is 15% slower than the super-threshold library in terms of mean delay and 17% slower in $\mu + 3\sigma$ delay.

Table 7 Comparison of critical path delays of ISCAS circuit C6288

Sizing	Width and Length		Super-threshold		Width only	
	μ	$\mu + 3\sigma$	μ	$\mu + 3\sigma$	μ	$\mu + 3\sigma$
0.3V	0.23E-6	0.67E-6	0.32E-6	0.91E-6	0.29E-6	0.82E-6
0.6V	0.34E-8	0.55E-8	0.37E-8	0.57E-8	0.34E-8	0.53E-8
1.2V	0.67E-9	0.76E-9	0.57E-9	0.63E-9	0.51E-9	0.55E-9

In general, at different voltages, the proposed width-tuned library shows 10% improvement of mean delay on average and 10% improvement of the $\mu + 3\sigma$ delay compared to the super-threshold library.

5.4 Verification of Standard Cell Optimization in Silicon

The chip studied in this section was developed at Holst Centre/imec-nl, taped-out in December 2012. It became available for test in June 2013. It contains several ring oscillator structures of various lengths and different cell types from two libraries (the width- and length-tuned and width-tuned only libraries). The die and layout views are shown in Figure 5.12.

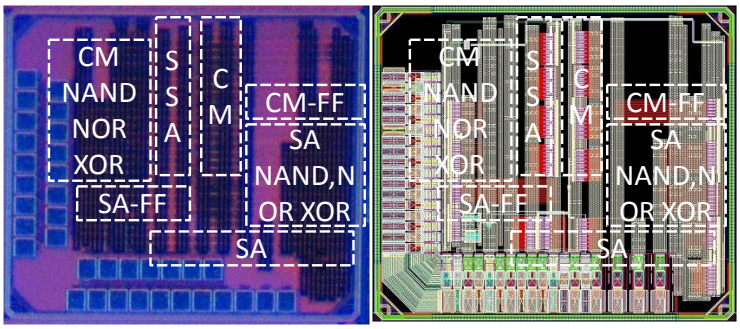


Figure 5.12 The die and layout views of the test chip.

This test chip has four modules. Within the module, ring oscillators are used to compare the timing of cells from the two libraries. Level shifters are used between each ring oscillator and the I/O pad. In the inverter module, inverters from different libraries have separate power supplies to be able to measure power consumption of the two libraries.

Inverter module contains:

- 16-, 40-, 64-, and 80-stage inverter chains, based on the inverter of size D4 from a commercial 40nm library. See CM in Figure 5.12.
- 16-, 40-, 64-, and 80-stage inverter chains, based on an inverter from the sub-threshold library with an equivalent area as the commercial D4 inverter. See SA in Figure 5.12.
- 16-, 40-, 64-, and 80-stage inverter chains, based on an inverter from the sub-threshold library with less area than the commercial D4 inverter. See SSA in Figure 5.12.

NAND NOR module (shown in Figure 5.12 together with the XOR Module)

- 72-stage NAND-NOR chains, based on the gates of size D4 from the commercial 40nm library. See CM NAND NOR in Figure 5.12.
- 72-stage NAND-NOR chains, based on the gates of size D4 from sub-threshold library. See SA NAND NOR in Figure 5.12.

XOR module (shown in Figure 5.12 together with the NAND and NOR Module)

- 200-stage XOR chains, based on the gates of size D2 from the commercial 40nm library. See CM XOR in Figure 5.12.
- 200-stage XOR-chain, based on the gates of size D2 from the sub-threshold library. See SA XOR in Figure 5.12.

Flip-Flop (FF) module

- 200-stage FF chains, based on the flip-flop of size D1 from the commercial 40nm library. See CM-FF in Figure 5.12.
- 200-stage FF chains, based on the sub-threshold library flip-flop with equivalent area as the commercial flip-flop. See SA-FF in Figure 5.12.

In each module, for each chain length and cell type, there are 16 repetitions of the chain.

5.4.1 Inverter Chain

There are three inverters compared in this section, labeled as CM, SA, and SSA. The Inverter denoted as CM is the Inverter with size D4 from the commercial 40nm library. Based on the proposed sizing methodology, the inverter sizing is optimized with the same area constraint in the sub-threshold region, which provides the inverter labeled as SA. The SSA inverter is an inverter optimized for sub-threshold operation and has weaker drive strength than the SA inverter. In terms of area, SSA is smaller than an inverter with size D4 from the commercial library.

5.4.1.1 Frequency Comparison in Sub-threshold Region

The frequency comparison of the 16-stage inverter chain at 0.3V is shown in Figure 5.13. The blue color represents the commercial library inverters, namely CM-INV. The red color represents the inverters from the sub-threshold library with the same area as the CM-INV, namely SA-INV. The green color represents the sub-threshold library inverter with smaller drive strength, namely SSA-INV. The SA and SSA inverter chains are faster than the CM inverter chain. On average, the SA and SSA inverters are 1.47x and 1.25x faster, respectively, compared to the CM inverter. Here, a parameter $\frac{\mu+\sigma}{\mu-\sigma}$ is introduced to represent the spread of the frequency (note that, the μ and σ values are calculated from 64 samples).

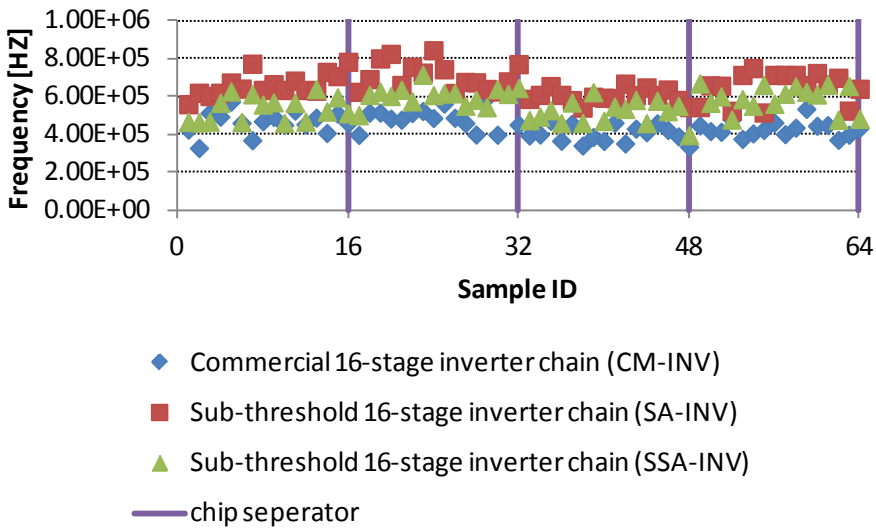


Figure 5.13 The frequency comparison using a 16-stage inverter chain (V_{GS}=V_{DS}=0.3V).

As listed in Table 8, in sub-threshold, the SA and SSA inverters from the sub-threshold library have 1.5x and 1.3x faster frequency, respectively, compared to the commercial library. Because the mean value is changed, comparing the standard deviation alone does not represent the frequency spread. As has been mentioned, $\frac{\mu+\sigma}{\mu-\sigma}$ is used to compare the spread. SA and SSA inverters have narrower spread compared to the CM inverter. The average improvement with SA and SSA inverters is 5% with regard to the commercial library. The frequency comparison of the inverter chains with a different number of stages is shown in Table 9.

The reason to compare the different number of stages is to see if the variation is suppressed or not by adding more cells in a chain. As listed in Table 9, when the number of stages is increased from 40 to 64 or 80, the variation of the measured samples does not change significantly. The variation is around 10% for all three different cells. When comparing the unit delay of the 40-stage chain to the 64/80-stage chain, the unit delay mean value decreases, which means that when the number of stages is increased from 40 to 64/80, the frequency per stage increases. The speed-up comes from the variation cancellation of longer chains [74].

Table 8 Frequency statistics of 16-stage inverter chain [0.3V, 0.5V]

		Commercial	Sub-threshold (SA)	Sub-threshold (SSA)
0.3V	μ	4.48E+05	6.57E+05	5.62E+05
	σ	6.35E+04	7.35E+04	6.97E+04
	$\frac{\mu+\sigma}{\mu-\sigma}$	1.33	1.25	1.28
0.4V	μ	3.92E+06	5.67E+06	5.05E+06
	σ	5.09E+05	6.45E+05	5.86E+05
	$\frac{\mu+\sigma}{\mu-\sigma}$	1.30	1.26	1.26
0.5V	μ	2.50E+07	3.26E+07	3.02E+07
	σ	2.41E+06	2.69E+06	2.59E+06
	$\frac{\mu+\sigma}{\mu-\sigma}$	1.21	1.18	1.19

Table 9 Frequency statistics of 40 64 80-stage inverter chain at 0.35V

		40-stage inverter	64-stage inverter	80-stage inverter
CM	μ	6.65E+05	4.36E+05	3.43E+05
	σ	6.99E+04	4.34E+04	3.41E+04
	$\frac{\sigma}{\mu}$	10.5%	9.9%	9.9%
	$\frac{\mu+\sigma}{\mu-\sigma}$	1.23	1.22	1.22
SA	μ	9.90E+05	6.69E+05	5.56E+05
	σ	1.04E+05	6.97E+04	4.93E+04
	$\frac{\sigma}{\mu}$	10.5%	10.4%	8.8%
	$\frac{\mu+\sigma}{\mu-\sigma}$	1.23	1.23	1.19
SSA	μ	8.97E+05	5.90E+05	4.83E+05
	σ	9.98E+04	5.79E+04	4.80E+04
	$\frac{\sigma}{\mu}$	11.1%	9.8%	9.9%
	$\frac{\mu+\sigma}{\mu-\sigma}$	1.25	1.22	1.22

5.4.1.2 Power Comparison

The normalized leakage power is shown in Figure 5.14. Note that all the power figures are normalized to the average leakage power of the CM inverter chains at 0.3V. In the sub-threshold, from 0.3V to 0.5V, the SA consumes 38% to 51% less leakage power with regard to CM. Similarly, SSA has 57% to 60% lower leakage power compared to CM. From 0.6V to 1.1V, the SA and SSA show on average 53% and 60% leakage power savings, respectively, compared to CM.

The leakage savings in the sub-threshold region is because of the balancing between the N and P network. The faster network is slowed down to speed up the slower network, which means that the larger current in the faster network is reduced to increase the lower current in the slower network. In this way, the circuit can be faster in the worst case and less leaky in idle mode [14, 30]. During the sizing optimization, both channel length and width are tuned: the length of the transistor is increased to speed up the transistor in the sub-threshold region and the channel width is reduced to maintain the same area. In the super-threshold region, increasing the channel length decreases the transistor current, which leads to leakage power savings as shown in the lower part of Figure 5.14.

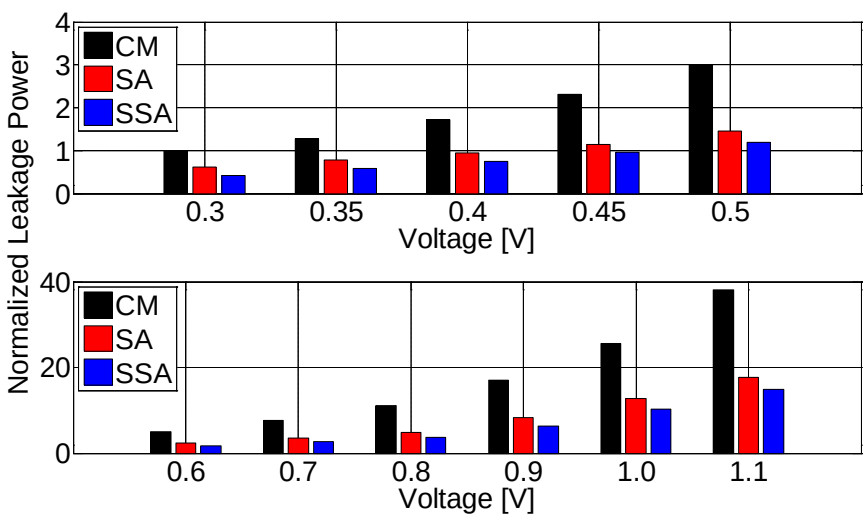


Figure 5.14 The leakage power comparison using inverters.

The leakage power of a single 16-stage chain, shown in Figure 5.14, is divided into two ranges to have a better comparison scale: sub-threshold [0.3V, 0.5V] and from near threshold to nominal voltage [0.6V, 1.1V]. Notice that the sub-threshold libraries have around 55% lower leakage power compared to the super-threshold library. The leakage savings at nominal voltage is 53%.

5.4.1.3 Comparison with Body Biasing

Body biasing [43, 75] and voltage scaling [4] are post-silicon techniques commonly used in digital circuit performance tuning. In the sub-threshold, the performance of the cells is very sensitive to supply voltage and body biasing voltage change, which makes post silicon tuning techniques very efficient for performance compensation or for performance boosting. In this section, the chains are compared under body biasing and voltage scaling.

Both forward and reverse body biasing voltages are applied to a 16-stage inverter chain at the 0.3V supply voltage to compare the body biasing effect on all three inverters (see Figure 5.15). The body biasing voltage range is from -0.2V to 0.4V. Similar to the super-threshold as shown in [43, 75], when applying forward body biasing, the logic circuits can speed up. Reverse body biasing has the opposite effect. Different color lines show the mean values of the delays of different inverter chains.

When the forward body bias voltage varies from -0.2V to 0.3V, the slope of the body biasing tuning curve is increased at each voltage. When forward body biased is applied at 0.4V, the slope of the curve decreases compared to the one at smaller biasing voltages. Thus the efficiency of body biasing at 0.4V decreases. A similar trend can be found in the sub-threshold region for the 40nm technology node. The SA and SSA inverter chains are faster than the CM inverter chains at each body biasing voltage. However, compared to the performance at zero body biasing voltage, CM inverter chains on average speed up by 5x, while for the SA and SSA inverter chains, the speed-up is 4.5x at 0.3V. CM inverter chains are, therefore, more sensitive to forward body biasing at 0.3V than the SA and SSA inverter chains.

The influence of the body biasing voltages can also be recognized as threshold voltage process variation. CM inverter chains are, therefore, more sensitive to process variation, which verifies the balancing theory and corresponds to the Monte-Carlo simulation results.

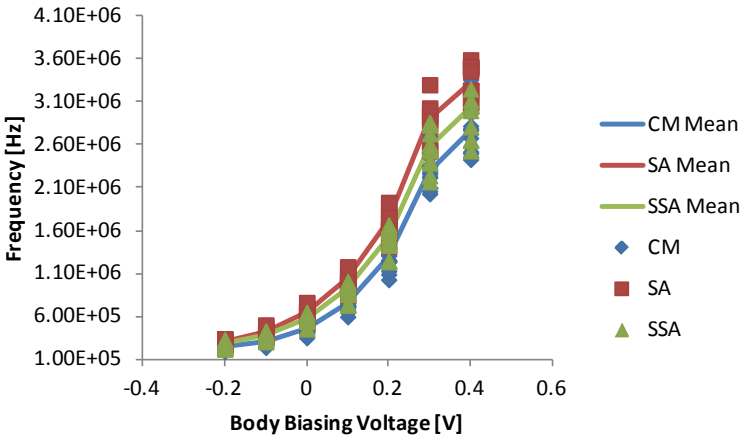


Figure 5.15 The frequency comparison using the inverter chain with body biasing (V_{GS}=V_{DS}=|0.3V|).

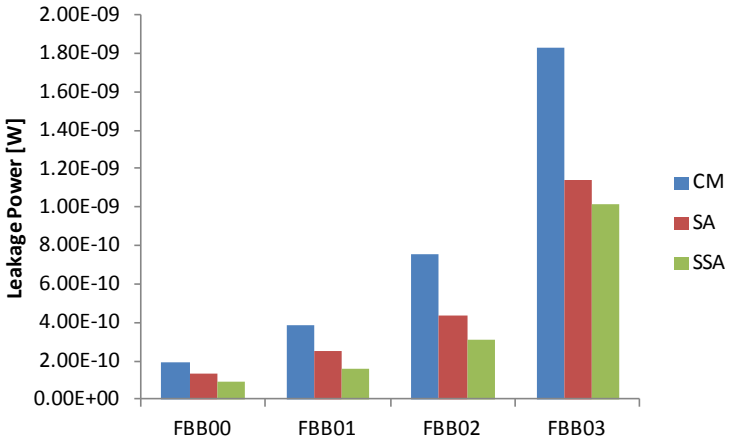


Figure 5.16 The leakage power comparison using the inverter chain with body biasing (V_{GS}=V_{DS}=|0.3V|).

The leakage power consumption at different biasing voltages is shown in Figure 5.16. When the supply voltage is 0.3V, the same dynamic energy is maintained for different body biasing voltages. The increased energy consumption is mainly provided by the leakage energy. The trend of the leakage power of the 16-stage inverter chain with FBB from 0 to 0.3V can also be observed in

Figure 5.16. At each body biasing voltage, SA and SSA cells consume lower leakage power compared to the CM inverters.

To have a more detailed comparison, a frequency target of 2MHz is set to compare the dynamic power and the leakage power of all three different inverter chains under supply voltage scaling and body biasing scaling.

The frequency and the leakage power of the two library cells working at 0.3V with forward body biasing of up to 0.3V are shown in Figure 4.16. Note that when forward body biasing the transistors, the dynamic energy is not affected. Hence, the dynamic energy trend is excluded in the body biasing comparison. The dashed lines represent the frequency, and the solid lines represent the leakage power of each kind of cell. When the forward body biasing voltage is increased, all three cells speed up and consume higher leakage power. When taking a closer look at the solid lines, one can see that the slope of the black solid line is larger than the slope of the red and blue solid lines. This means that the leakage power of the CM cells is more sensitive to forward body biasing than for cells from the sub-threshold library. When the target frequency is set to be 2MHz, as the green line indicates, the SA and SSA cells need smaller body biasing voltages and are less leaky compared to CM cell. Data are further evaluated in Table 10.

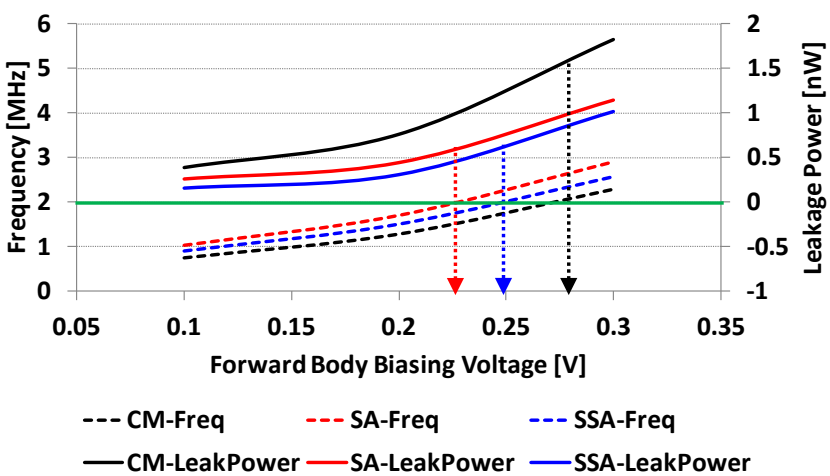


Figure 5.17 The frequency and leakage power comparison using 2MHz case study with body biasing.

In Figure 5.18, the frequency and dynamic energy trends are shown for two library cells with voltage scaling from 0.3V to 0.4V. Note that when scaling the supply voltage, the dynamic energy is dominant. The leakage trend is not shown in Figure 4.17. At the 2MHz target frequency (indicated by the green line), SA and SSA need lower supply voltages than CM and consume lower dynamic energy. Further results combining supply and body biasing tuning results are shown in Table 10.

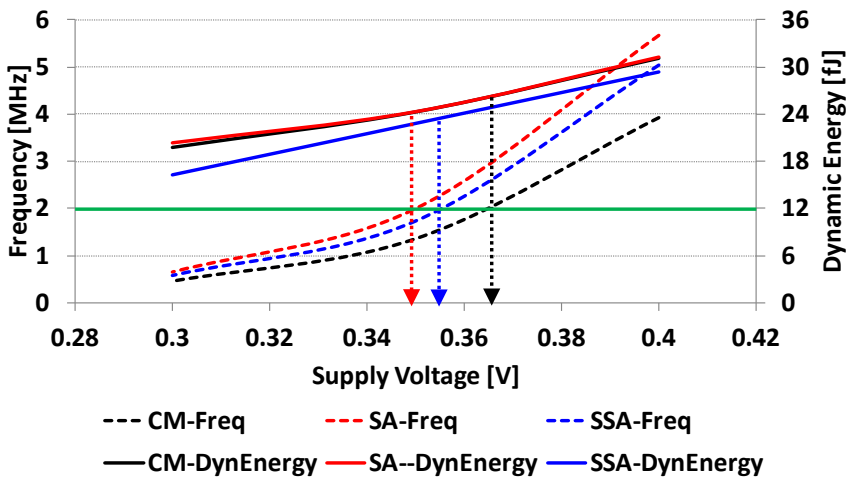


Figure 5.18 The frequency and leakage power comparison using 2MHz case study with voltage scaling.

Among all six options working at 2MHz, the SSA cells working at 0.3V with 0.24V forward biasing consume the lowest dynamic energy. SSA saves 40% of the dynamic energy with regard to CM working at 0.365V with zero biasing, and 20% at 0.3V with 0.27V body biasing. When leakage power saving is the target, SSA working at 0.355V provides the highest leakage savings. Compared to the CM options 1 and 4, the leakage reduction factors are 92% and 51%, respectively. Assuming that the forward body biasing voltage can be turned off during idle time, the leakage power is reduced, as shown in-between the parentheses. The leakage savings of the SSA and SA sub-threshold libraries is 50% and 67%, respectively.

Another interesting conclusion is that, when minimizing the supply or the body biasing voltage, the SA cells offer the most voltage-to-frequency efficien-

cy: SA needs 0.35V with zero body biasing, or 0.3V with 0.22V biasing voltage to achieve 2MHz frequency.

Table 10 2MHz case study data summary

		Supply [V]	Body Biasing [V]	Dynamic Energy[fJ]	Leakage Power [nW]
1	CM	0.3	0.27	20.3	1.5(0.195@zero BB)
2	SSA	0.3	0.24	16.3	0.5(0.087@zero BB)
3	SA	0.3	0.22	19.8	0.5(0.120@zero BB)
4	CM	0.365	0	27	0.27
5	SSA	0.355	0	23	0.13
6	SA	0.350	0	24	0.15

In general, compared to the commercial library cells, the sub-threshold library cells have lower dynamic energy, lower leakage power, and need lower supply voltage to work at the same frequency when post-silicon tuning techniques are used in sub-threshold region.

5.4.2 NAND-NOR and XOR Chain

The frequency of different chains as a function of supply voltage is listed in Table 11. The relative frequency ratio is shown in Figure 5.19.

The data in Figure 5.19 are calculated from the average frequency of SA chains over the average frequency of CM chains at each voltage. As can be concluded from Figure 5.11, when comparing against the commercial library cells, the NAND-NOR gates (shown in red), and the XOR gates (shown in green), from the proposed sub-threshold library show higher frequency at sub/near-threshold. At 0.3V, the NAND-NOR chain shows 2× higher frequency while the XOR chain shows only 1.3× higher frequency. The difference between the improvement factors is because the size of the NAND and NOR gates used in comparison are twice more than the minimum size, thereby allowing a larger sizing optimization search space. The size of the XOR gate in the comparison is close to minimum size, limiting the optimization search space. From 0.7V to 1.1V, the frequency of the proposed sub-threshold library is worse than the commercial library [14, 30].

Table 11 Frequency of the NAND-NOR and XOR chains [0.3V, 1.1V]

	72-stage NX-chain								
Voltage [V]	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1
Commercial [Hz]	92.5K	0.76M	4.62M	17.7M	46.3M	88.9M	143M	198M	250M
Sub-Vt [Hz]	183K	1.17M	6.35M	22.6M	45.0M	77.9M	112M	148M	179M
	200-stage XOR-chain								
Voltage [V]	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1
Commercial [Hz]	10.3K	94.2K	0.62M	2.41M	5.98M	11.1M	17.1M	24.5M	29.8M
Sub-Vt [Hz]	13.4K	123K	0.78M	2.65M	5.84M	9.94M	14.5M	19.1M	23.5M

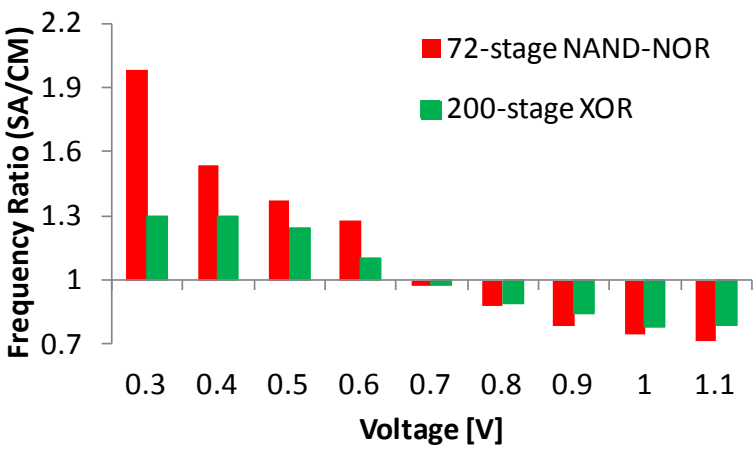


Figure 5.19 The frequency ratio comparison using NAND-NOR and XOR chains (V_{GS}=V_{DS}=[0.3V,1.1V]).

5.4.3 Conclusions of Section 5.4

Silicon measurements were presented in this section. The inverter chain/ring oscillator structure is used to compare the frequency, leakage, and the influence of body biasing in sub-threshold region. The body biasing comparison indicates that, with the same performance target in the sub-threshold region, the cells from the sub-threshold library have lower leakage power and dynamic energy consumption with regard to the ones of the commercial library. The NAND-NOR and XOR chain comparison results prove that the width and length-tuned library cells have better performance in the sub-threshold region yet lower speed in the super-threshold region.

CHAPTER 6

SUB-THRESHOLD DESIGN USING THE OPTIMIZED LIBRARIES

The behavior of the standard cells of different libraries was compared with both simulation and measurement results at different supply voltages and different process corners in the previous chapter. In this chapter, the standard cell library containing 166 cells is evaluated from a design perspective. The libraries are used to synthesize circuits from the ISCAS and ITC benchmarks to assess gate count, speed and energy. The results are compared against a corresponding super-threshold library that was characterized to operate in the sub-threshold regime. A test-chip was taped out in 40nm CMOS technology. The test-chip consists of several multiply accumulate modules implemented with the libraries described in this thesis and a commercial library characterized in the sub-threshold regime. Silicon measurements confirm the advantages in terms of power and speed of using the new libraries.

6.1 Logic Synthesis Benchmarking: Evaluation of Low Voltage Libraries on Actual Designs

6.1.1 ISCAS Benchmarks

In Table 12, the proposed library is compared with the super threshold library in 90nm CMOS technology node. One can see that for all shown circuits in the SS corner, the proposed library has up to 57% and 69% improvement in timing and energy, respectively. In the TT corner, the timing and energy improvements are 55% and 74%, respectively. In general, the synthesis with the proposed libraries yields a lower gate count number. It is also noticed that with the proposed library, the logic synthesis tool tends to use more complex cells. This is because the transition balancing theory is based on the topology of the standard cells. Complex cells allow bigger optimization room. Therefore, the optimized complex cells have more balanced and larger currents when com-

pared to the commercial library cells. This also proves that the balancing theory is effective for delay and power savings without area penalty.

Table 12 ISCAS benchmark circuit comparison

		Commercial Library				This Thesis				Improvement Comparison %				
		Timing	Power (uw)		Gate	Timing	Power (uw)		Gate	Our Work		[16]	[28]	
		(ns)	Leakage	Dynamic	Count	(ns)	Leakage	Dynamic	Count	Timing	Energy	Timing	Timing	Energy
C6288	SS	199.32	9.86	137.31	1592	113.45	11.28	69.24	1356	43.1	68.8	12.6	10.1	29.8
	TT	34.86	16.48	274.23		20.69	20.76	150.24		40.6	65.1			
C3450	SS	138.75	8.79	112.28	1423	63.31	15.32	65.12	1320	54.4	69.7	4.4	Not available	
	TT	30.78	15.74	217.54		14.65	27.48	131.28		52.4	67.6			
C1355	SS	57.21	2.24	34.25	286	36.32	3.03	17.31	217	36.5	64.6	13.5	10.4	41.2
	TT	13.44	4.12	68.50		7.14	5.19	30.06		46.9	74.2			
74283	SS	220.91	0.85	7.76	44	122.35	0.97	4.42	27	44.6	65.3	5.3	10.4	22.7
	TT	41.19	1.40	13.89		22.32	1.87	7.72		45.8	66.0			
74L85	SS	272.74	1.71	14.27	60	140.28	2.84	8.92	43	48.6	62.1	6.6	9.1	37.8
	TT	55.48	3.12	27.54		30.01	3.65	15.28		45.9	66.6			
74182	SS	167.15	0.77	6.29	22	72.25	1.27	4.58	18	56.8	64.2	33.1	7.8	12.4
	TT	31.43	1.39	12.41		14.32	1.95	8.21		54.4	66.5			

In Table 12, a comparison with regard to the work of [16] and [28] is also shown. Due to the lack of information in [16, 28], the synthesis experiments cannot be scaled to the same technology node to compare the balancing approach directly with their proposed sizing method. Only the benchmark improvement data (with regard to their own reference library in their own technology) is compared. The difference between the approach in [16] and the proposed sub-threshold sizing method is that V_{th} variation is considered in the proposed method. Furthermore, transition-based sizing allows us to compensate for the worst-case delay by increasing the best-case delay. This allows us to match the rise and fall delays. In [28], the channel length is increased to decrease V_{th}, leading to a higher current through the transistors. Because the channel length is increased to minimize V_{th}, the width is decreased to minimize the gate capacitance in order to be able to optimize the delay. As mentioned in Chapter 2, Chapter 3, and Chapter 4, the proposed sizing optimization goal is to balance the rise and fall transitions, not to maximize the current. Observe

that the proposed library has on average 3.3x and 4.8x better timing improvement than the work in [16, 28], respectively, and has 1.6x better energy improvement when compared to [28].

6.1.2 ITC Benchmarks

ITC benchmark circuits [76] were synthesized for minimum delay to compare the effect of the 40nm CMOS libraries at 0.3 V. Before the results of the entire benchmark suite are presented, a speed-area study on one benchmark circuit is carried out first: namely, the synthesis results of the B14 circuit from ITC benchmark circuit [76] are analyzed in detail. The critical paths of the B14 circuit were then extracted. One thousand Monte Carlo simulations were performed to generate the delay distributions of each critical path to compare the variability of different libraries. The results are shown in Figure 6.1. As can be seen from Figure 6.1, the critical path delay follows a log-normal distribution. Without any sizing optimization, the critical path has a wide distribution with a long tail as the blue line shows. The delay distribution of the critical path of the proposed library cells is left-shifted and narrowed down. The mean delay decreases from 4.25μs to 2.59μs, and the variation is reduced from 44% to 30%.

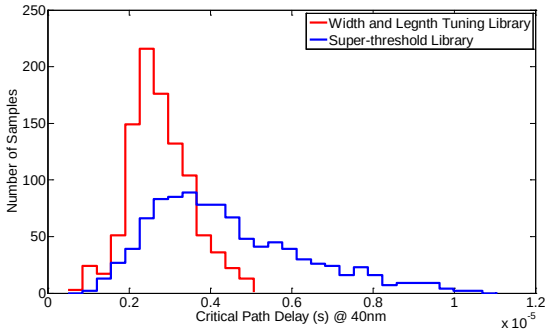


Figure 6.1 Comparison of the critical path delay distribution of the B14 circuit in 40nm CMOS technology.

The trends of the synthesized delay versus the area of the ITC B14 circuit at 0.3 V are shown In Figure 6.2. The three black arrows show different constraints. With the width and length-tuned library, the circuit can work at a faster speed. Arrow C indicates that, when delay is a constraint, the circuit synthesized by the width and length-tuned library requires a 14% smaller area as compared to the circuit synthesized with the super-threshold library. When area is the constraint (Arrow B), the circuit synthesized by the width- and

length-tuned library is 1.8× faster. Without any constraints, the circuit can be sped up 2.1× with 1.08× area compared to the circuit synthesized by the super-threshold library, as indicated by arrow A.

The delay, area, and power information of the B14 circuit with different constraints are shown in Table 13, Table 14, and Table 15. In Table 13, the speed of the circuit synthesized by the proposed library is pushed to the highest possible value (like Arrow A in Figure 6.2). The delay improvement and power savings when the area is constrained is listed in Table 14 (as Arrow B in Figure 6.2). The area and power savings when the same target delay is applied are shown in Table 15 (as the Arrow C in Figure 6.2).

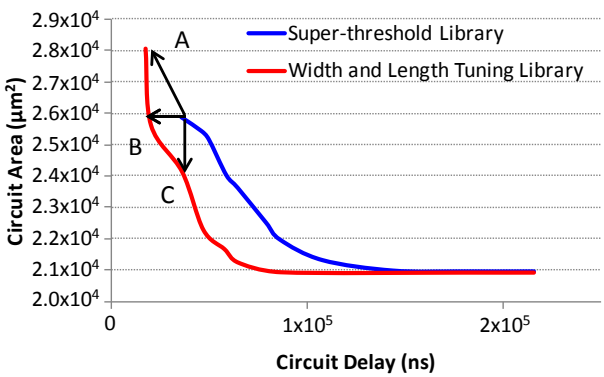


Figure 6.2 The delay and area comparison using synthesized B14 circuit (V_{GS}=V_{DS}=0.3V. 40nm CMOS technology).

Table 13 ITC benchmark circuit synthesis results for maximum speed

	Delay (ns)		%	Area (μm²)		%	Total Power (nW)		%
	Super-threshold library	Width and length tuned		Super-threshold library	Width and length tuned		Super-threshold library	Width and length tuned	
B01	850	480	43.5	320	334	−4.4	0.502	0.308	38.6
B02	780	450	42.3	213	227	−6.6	0.237	0.161	32.1
B03	880	510	42.0	582	660	−13.4	0.229	0.164	28.4
B04	1170	630	46.2	2120	2525	−19.1	1.267	0.865	31.7
B05	1820	1030	43.4	3118	3664	−17.5	1.336	0.920	31.1
B14	3600	1720	52.2	25866	28056	−8.5	3.795	2.780	26.7

The results indicate that, in the 40nm CMOS technology node, the circuits synthesized by the proposed width- and length-tuned library have better timing, smaller area, and lower power consumption when compared to the super-threshold library at 0.3 V. For the delay driven comparison shown in Table 13, one can observe a maximum timing improvement of 52% and power savings of 39%. If the same area constraint is applied, the maximum timing improvement is 44% and the power savings is 41%. When the delay target is set the same for both libraries, the width- and length-tuned library achieves up to 24% area savings and 53% power savings.

Table 14 ITC benchmark circuit synthesis results (equal area)

	Delay (ns)		%	Area (μm²)		Total Power (nW)		%
	Super-threshold library	Width and length tuned		Super-threshold library	Width and length tuned	Super-threshold library	Width and length tuned	
B01	850	500	41.2	320	315	0.502	0.298	40.6
B02	780	490	37.2	213	204	0.237	0.148	37.6
B03	880	750	14.8	582	555	0.229	0.138	39.7
B04	1170	810	30.8	2120	2077	1.267	0.765	39.6
B05	1820	1200	34.1	3118	3114	1.336	0.826	38.2
B14	3600	2000	44.4	25866	25614	3.795	2.466	35.0

Table 15 ITC benchmark circuit synthesis results (same delay)

	Delay (ns)		Area (μm²)		%	Total Power (nW)		%
	Super-threshold library	Width and length tuned	Super-threshold library	Width and length tuned		Super-threshold library	Width and length tuned	
B01	850	850	320	243	24.1	0.502	0.238	52.6
B02	780	780	213	177	16.9	0.237	0.144	39.2
B03	880	880	582	536	7.9	0.229	0.139	39.3
B04	1170	1170	2120	1671	21.2	1.267	0.877	30.8
B05	1820	1820	3118	2726	12.6	1.336	0.723	45.9
B14	3600	3600	25866	24121	6.7	3.795	2.852	24.8

6.2 *Silicon Results*

In addition to the simulation results of the benchmark circuits, a test chip was designed to evaluate the library sizing techniques through silicon measurements. The test chip was fabricated in the 40nm CMOS technology. On this test chip, a hardware accelerator is used to enable an efficient ECG algorithm mapping. The accelerator is designed to be connected with an ARM Cortex M3 processer via the high performance on chip bus (AHB bus). An ARM Cortex M3 processor is a 32-bit processor, and the ECG algorithms typically require data types which are a multiplier of 12-bit. Therefore, the proposed accelerator uses native ARM data types of 16-bit and 32-bit signed values. Though designed as a hardware accelerator for an ARM processor, in this test chip, the accelerator is not connected to an ARM processor. It uses random data generated by a linear feedback shift register (LFSR) as input to compare the frequency and power consumption of different libraries at multiple voltages.

For benchmarking purposes, in the test chip, the hardware accelerator is implemented three times, synthesized using different libraries with different frequency and power targets. A high-level architecture of the test chip is shown in Figure 6.3.

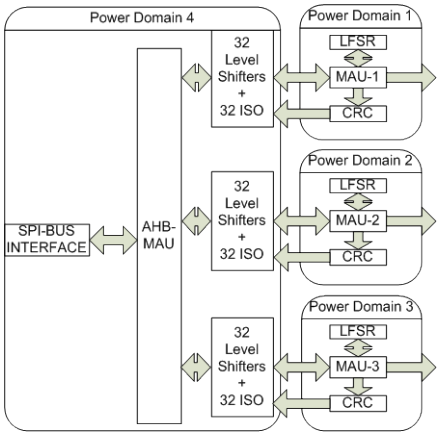


Figure 6.3 The architecture of the test chip.

The following outlines the blocks of the accelerator module in Figure 6.3. MAU stands for multiply accumulate unit. ISO is short for the isolation cell, which is needed to enable or disable the transition across two power domains. LFSR is the linear feedback shift register used to generate the input data for the

MAU. CRC is the cyclic redundancy check, which is used to determine whether the output is correct or not. SPI is the standard serial-parallel-interface. The AHB-MAU block is designed to connect to an ARM processer. In this design, the AHB-MAU block is, however, modified to connect the SPI interface with the outputs of three MAUs.

Multiple power domains are used to measure the power consumption separately. Power domains 1, 2, and 3 are the MAU domains synthesized at low voltage, in which different libraries and different synthesis settings are used. Power domain 1 contains cells from the proposed sub-threshold library with maximum frequency as synthesis target (as point A shown in Figure 6.2). Power domain 2 uses cells of the super-threshold library synthesized for maximum frequency. Power domain 3 has the same frequency target as power domain 2 but the cells are of the sub-threshold library. This power domain intends to show the low power and small area potential of the sub-threshold library (as point C, shown in Figure 6.2). Note that, during synthesis, setting up a different clock frequency for each module requires different clock domains. In this design, there is no connection between MAUs. Therefore a multiple clock domain setup is not required. The clock frequency setting is explained later in Section 6.2.1. Power domain 4 is an intermediate power domain operating at SPI clock speed, in which the voltage is around the threshold voltage. Because the whole IO of the chip is working at nominal voltage 1.1V, power domain 4 serves as a transition stage between the low voltage domains and the rest of the chip working at 1.1V.

The layout view and die photo are shown in Figure 6.4.

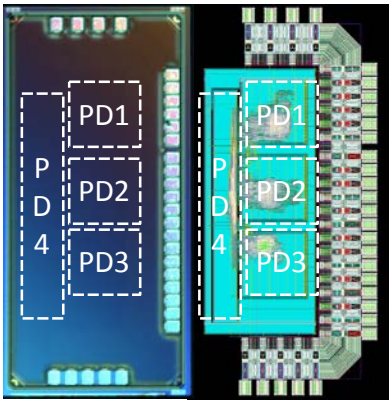


Figure 6.4 The die and layout view of the test chip.

6.2.1 Design Strategies for the Hardware Accelerators

Since multiple power domains are implemented in this design, the frequency target settings for different power domains at different modes need to be considered. When synthesizing the complete chip with all the modules included, the frequency of each module cannot be set separately without enabling multiple clock domain settings. The easiest way to set frequency targets of different MAUs is to set the maximum path delay value of the critical paths of each MAU. To identify the critical paths and maximum path delay value of each MAU, a separate synthesis process is needed. Each MAU is synthesized together with the SPI module in power domain 4 to ensure the correct data transfer between two modules in the sub-threshold region. As mentioned in the previous section, power domain 4 serves as an intermediate section between the IO and the MAUs. The working voltage of the SPI interface is around the technology's threshold voltage and with a dedicated SPI clock. After the maximum path delay and the critical paths are determined for each MAU, the command *set_max_delay* of Cadence RTL Compiler is used to set the frequency target of each MAU during top level synthesis with multiple power domains and power modes stated in the common power format (CPF) file. Note that the single MAU and SPI synthesis is just used to find the maximum delay and the input and output pins of the critical paths. In this design, the focus is the performance in the sub-threshold region. Therefore, the frequency at nominal voltage is irrelevant in having a negligible influence on the frequency in the sub-threshold region.

With the clock frequency and maximum path delay values properly aligned at different voltages for the different power domains, the next settings that need to be determined are power and area related parameters. MAU1 is aggressively designed by setting the optimization target with minimum area and zero leakage power to collect the area and leakage power numbers. MAU1 is designed to represent Arrow A in Figure 6.2. Therefore the zero area and zero leakage synthesis targets are also set. MAU2 (synthesized with the super-threshold library characterized at low voltages) is the reference design. The area of MAU2 is used to tune the area of MAU3 (synthesized by the proposed sub-threshold library) together with the timing constraints in order to compare the leakage power of the two MAUs when the area is similar, as shown by Arrow B in Figure 6.2.

The libraries at different corners and ±10% voltages near the synthesized voltage need to be included in the low power CPF flow settings, based on the recommended operating point settings of the foundry model. These libraries are used for the Multi Mode Multi Corner check during the backend phase of the design. During the library characterization, the setup and hold check matrix needs to be verified at different voltages. This is because, in the sub-threshold region, the timing values of the standard cells are very sensitive to supply voltage change. It is not unusual to see that the frequency difference can be up to 10× from 0.3V to 0.4V.

Another important step in the backend flow is the clock tree synthesis (CTS). In the sub-threshold region, the slew and delay of the cells are much smaller compared to the super-threshold region. For that reason, the CTS process tends to have a perfect clock tree, which means that the clock waveforms at the output node of the clock tree are adjusted very steeply. The skew is also minimized. The perfect clock tree consumes considerable power because big clock buffers are inserted along the clock tree. However, in the sub-threshold region, a sufficiently good clock tree can tolerate the slower transition delay as well as flip-flops with negative hold time requirements. Therefore, a more relaxed clock skew setting is used in the CTS. The maximum allowed clock buffer size is limited to drive strength 4, which is large enough to generate proper clock output in the clock tree simulation.

6.2.2 Measurement Results and Evaluation of Design Characteristics

In this section, the measurement results of the test chip are shown to compare the advantages and disadvantages of different libraries at different voltages.

The three modules are connected to an SPI structure. The writing and reading waveforms of the SPI are shown in Figure 6.5. The writing process is shown on the left side, while the waveform when the SPI is reading from the modules is shown on the right side.

Figure 6.5 The writing and reading waveforms of the SPI.

The currents of three modules are shown in Figure 6.6. The current is measured by connecting a 10MΩ resistor at the supply node. The actual supply voltage is scaled to maintain the same supply voltage that would otherwise be directly applied to the three modules. Both idle current and active current are shown. Note that the three current trends are measured individually and restored to one screen for comparison purposes. The leakage current comparison is shown in Figure 6.8.

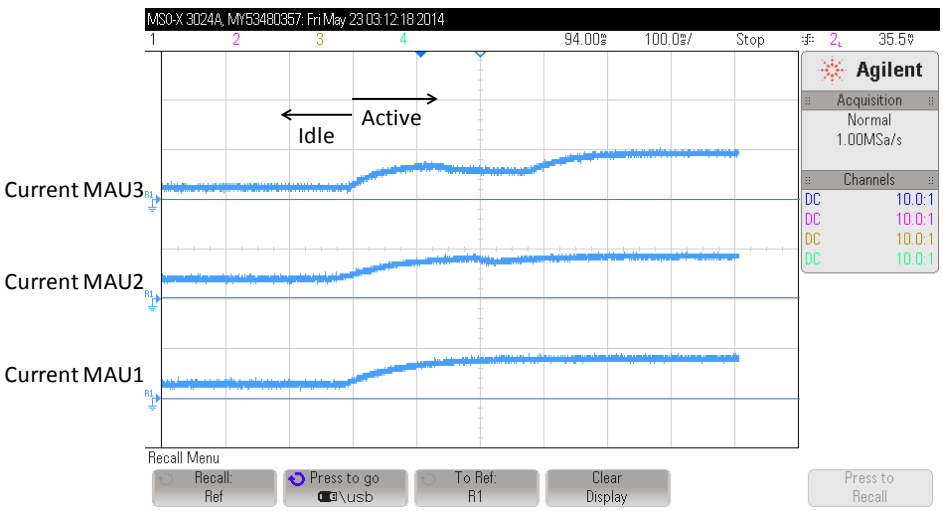


Figure 6.6 The current trends of MAU1, MAU2, and MAU3.

The comparisons of the frequency and dynamic energy are carried out among the three MAUs at different voltages (see Figure 6.7). MAU1, shown in green, is the synthesized module which has the fastest speed in the sub-

threshold region from all three modules. MAU2, shown in blue, and MAU3, shown in red, have the same synthesis timing target. The only difference between MAU2 and MAU3 is that MAU2 is synthesized by the characterized super-threshold library cells, whereas MAU3 is synthesized by the proposed sub-threshold library as MAU1. Note that, the comparison of frequency and dynamic energy presented in Figure 6.7 shows a cross voltage comparison. All three modules are measured through the same voltage range. The horizontal axis shows the dynamic energy while the vertical axis shows the frequency. The voltage level scales from 0.4V to 1.1V as can be seen from the bottom left to the top right direction in Figure 6.7.

When looking at the vertical axis, and comparing the frequency of three modules, MAU2 has the highest frequency at nominal voltage as shown at the top right side of Figure 6.7. MAU1 has the highest frequency at the lowest comparison voltage as shown at the bottom left side of the Figure 6.7. A similar trend was found in the inverter chain measurements in Chapter 5. MAU1 is 1.5x faster compared with MAU2 at the lowest voltage. Compared with MAU3, MAU1 is 2.0x faster. Comparing the frequency of the three modules alone at the same supply voltage, the advantage of using the proposed library is not as great as in the chain comparison. This is because frequency is not the only target during synthesis. The energy consumption is also considered. MAU1 is not only faster in the sub-threshold region, but also shows improvements in terms of dynamic energy savings. From the sub-threshold to super-threshold region, the green line always stays on the left side of the blue line, which means MAU1 is more dynamic energy efficient compared to MAU2. On average, at the same frequency, MAU1 can achieve more than 50% dynamic energy savings. When comparing MAU2 with MAU3 with the same synthesis target, the two modules are seen to have similar speed and dynamic energy consumption in the sub-threshold region.

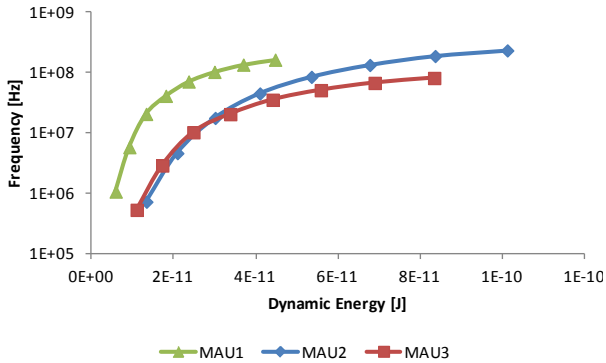


Figure 6.7 The dynamic energy consumption and frequency comparison of MAU1, MAU2, and MAU3.

The reason to include MAU3 in the test chip is to show the leakage savings advantage of the sub-threshold library. The leakage comparison from 0.4V to 1.1V is shown in Figure 6.8. In the compared voltage range, MAU1 and MAU3 modules are less leaky compared to MAU2. On average, MAU1 and MAU3 consume 33% and 40% lower leakage power, respectively, compared to MAU2.

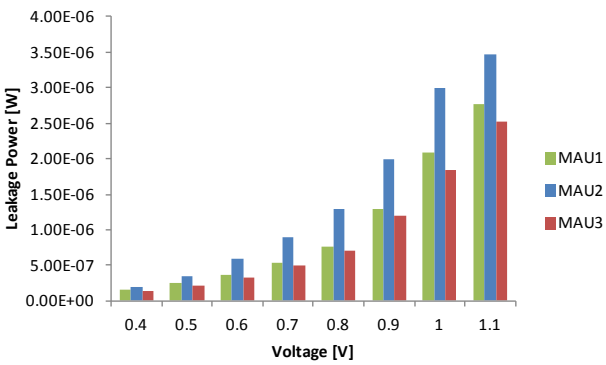


Figure 6.8 The leakage power consumption comparison of MAU1, MAU2, and MAU3.

6.2.3 Conclusions of Section 6.2

In this section, the libraries are compared based on silicon measurement results of a low power hardware accelerator. Three copies of the hardware accelerator are used to compare the libraries from both the frequency and power perspectives. From the frequency and dynamic energy comparison, the accelerator generated by the proposed sub-threshold library (MAU1) saves 50%

dynamic energy in the sub-threshold region when running at the same clock frequency of the super-threshold library based module (MAU2). From the frequency comparison between the MAU1 and MAU2, at the minimum working voltage, the proposed sub-threshold library cells can achieve 2x better frequency compared to the commercial library cells. When comparing the leakage power of the two libraries, from sub-threshold to super-threshold, on average, it is found that the proposed sub-threshold library cells consume 33% to 40% lower leakage power compared to the super-threshold library cells.

Based on different synthesis targets, the sub-threshold library generated by the proposed sub-threshold sizing method provides a faster and more power efficient solution in the sub-threshold region compared to the super-threshold library.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In Section 7.1, the most important conclusions which are made in the previous chapters are recapped. The future research directions that we are interested in are discussed in Section 7.2.

7.1 Conclusions

Sub-threshold operation is an essential technique for low power applications. Working in the sub-threshold region not only provides power reduction benefits, but also leads to many well known problems, such as process variations, logic failure, and slow frequency. Therefore, the study and optimization of the sub-threshold operation is very important for the robustness of low power applications. In this work, the focus is on the sub-threshold digital circuit design, more specifically, the standard cell sizing optimization in the sub-threshold region.

The research starts from the behavior analysis of the transistors working in the sub-threshold region in Chapter 2 and Chapter 3. In the behavior analysis section, the sizing effects are studied for single NMOS and PMOS transistors, stacked transistors, parallel connected transistors, and transistors within other different transistor topologies. The differences between the super-threshold and sub-threshold operations are pointed out. Based on the behavior analysis and the transistor sizing effects, several possible sizing strategies are proposed for sub-threshold operation optimization.

With the understandings of how transistors with different geometry behave in the sub-threshold region, the research continued in Chapter 4. In Chapter 4, a balancing based voltage and technology independent sizing methodology is proposed. Based on the balancing methodology, two sizing methods, the width- and length-tuning method and the width-tuning-only method, are presented.

In Chapter 5, the standard cell libraries are developed in 40nm and 90nm technology nodes. The standard cell layout style and the library characterization process in the sub-threshold region are studied to guarantee an optimized and accurate library model at the target voltages.

In Chapter 6, the low power design backend flow is discussed, based on the behavior of the standard cells in the sub-threshold region. With the carefully tuned backend parameters, a test circuit is successfully developed in 40nm technology node to compare the proposed library with the commercial library at different voltages.

Overall, to benchmark the contribution of this thesis, comparisons between the proposed sub-threshold sizing methodology and libraries and the super-threshold sizing methodology and commercial libraries are carried out. At the cell level comparison, at different corners, up to 4x performance improvement and 40% variation savings can be achieved in the sub-threshold region. On average, the proposed standard cells are 40% faster compared to the commercial standard cells. When comparing the leakage power consumption, it is found that the proposed library is 50% less leaky than the commercial library cells from the sub-threshold region to the super-threshold region. When the comparison is carried out in the circuit level, similar improvement can be found in the simulation comparisons of the ISCAS and ITC benchmark circuits. In the test circuit measurement comparison, the module made by the proposed library outlines the module made by the commercial library in terms of dynamic energy (50% lower), frequency (2x higher), and leakage power (40% lower) in the sub-threshold region.

7.2 Future Work

Though a sizing methodology is proposed and other low voltage related issues have been studied, there are still some remaining tasks and interesting topics that can be explored further:

- In this thesis, the libraries are optimized for 0.3V only. Even though the optimized frequency is improved to 2x compared to the commercial library, it is still in the KHz range. The same optimization methodology can be applied at higher voltages to broaden the application area.

- Due to the layout customization effort, only 166 cells are chosen to be optimized in this thesis, a more complete set of standard cells is needed to fully evaluate the proposed sizing methodology.
- A sub-threshold cell pruning strategy is needed. For some cells, the variation is around 100% or even larger. For some cells, when the loading capacitance is over 10x of the gate capacitance, the output slew is 100x greater than when the loading capacitance is around 2x of the gate capacitance. Such cells need to be redesigned or removed from the sub-threshold library.

There are interesting topics to explore, not only in the development of the standard cell library, but also in the usage of the standard cell library in the low power backend flow.

The first topic of interest is the clock tree synthesis (CTS). In the sub-threshold region, the speed of the NMOS and PMOS transistors is very slow. Therefore, the tolerance of the input slew is larger than in the super-threshold region. The sub-threshold CTS timing constrains need to be adjusted according to the supply voltage setting. From circuit design perspectives, the design of new clock buffers and flip-flops for sub-threshold operation is also an interesting topic.

Another aspect that is worth investigating is how to design a system that has robust operation from the super-threshold region to the sub-threshold region. Such a system can benefit from the high-speed computing from the super-threshold mode and the low-energy consumption from the sub-threshold mode. An alternative solution is near-threshold computing. By optimizing the circuit at near-threshold voltage, the circuit can have medium frequency and still maintain lower power consumption than in the super-threshold region. Another possible solution is voltage-island techniques: dividing the whole system into several voltage islands. Like the chip shown in Chapter 5, a hardware accelerator is working in the sub-threshold region and is used as a matrix multiplexer unit of a main core working at nominal voltage. The tricky part in the circuit design of this solution is the communication between the low-voltage island and the high-voltage island. A specified flip-flop and level shifter design is needed.

Finally, in the circuit optimization process, individual customization is applied for each different cell for the best results. The optimization effort is huge

for a complete standard cell library. Therefore, either a more generic optimization strategy or an automatic optimization and verification flow is needed. What is meant by a generic optimization strategy is to cluster a set of the cells and apply one simple and effective optimization for one cell cluster. And what this simple and effective optimization means is to find a sweet point between the optimization effort and optimization result. Additionally, an automatic optimization and verification process includes the netlist optimization, layout generation, and simulation verification.

APPENDIX

Log-normal distribution parameter calculation

Suppose X is the a set of data which obeys a log-normal distribution

$Log(X)$ obeys a normal distribution

$E[Log(X)]$ and $Std[Log(X)]$ are the mean and standard deviation values of the normal distribution, respectively.

Then the mean and standard deviation values of X are calculated as

$$E[X] = e^{E[Log(X)]+\frac{1}{2}Std^2[Log(X)]}$$

$$Std[X] = \sqrt{e^{Std^2[Log(X)]} - 1}E[X]$$

And the upper and lower bound parameters (E±3Std) used in Chapter 2 are calculated as

$$E \pm 3Std[X] = e^{E[Log(X)]\pm 3Std[Log(X)]}$$

BIBLIOGRAPHY

[1] R. Vaddi, S. Dasgupta, and R. P. Agarwal, "Device and Circuit Co-Design Robustness Studies in the Subthreshold Logic for Ultralow-Power Applications for 32 nm CMOS," *Electron Devices, IEEE Transactions on,* vol. 57, pp. 654-664, 2010.

[2] B. Zhai, L. Nazhandali, J. Olson, A. Reeves, M. Minuth, R. Helfand*, et al.*, "A 2.60pJ/Inst Subthreshold Sensor Processor for Optimal Energy Efficiency," in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, 2006, pp. 154-155.

[3] A. Wang and A. Chandrakasan, "A 180mV FFT processor using subthreshold circuit techniques," in *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*, 2004, pp. 292-529 Vol.1.

[4] M. Srivastav, M. B. Henry, and L. Nazhandali, "Design of low-power, scalable-throughput systems at near/sub threshold voltage," in *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, 2012, pp. 609-616.

[5] D. Bol, D. Kamel, D. Flandre, and J.-D. Legat, "Nanometer MOSFET effects on the minimum-energy point of 45nm subthreshold logic," presented at the Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design, San Fancisco, CA, USA, 2009.

[6] D. Bol, J. De Vos, C. Hocquet, F. Botman, F. Durvaux, S. Boyd*, et al.*, "SleepWalker: A 25-MHz 0.4-V Sub-mm^2 7- uW/MHzMicrocontroller in 65-nm LP/GP CMOS for Low-Carbon Wireless Sensor Nodes," *Solid-State Circuits, IEEE Journal of,* vol. 48, pp. 20-32, 2013.

[7] Y. Pu, J. Pineda de Gyvez, H. Corporaal, and H. Yajun, "An Ultra-Low-Energy Multi-Standard JPEG Co-Processor in 65 nm CMOS With Sub/Near Threshold Supply Voltage," *Solid-State Circuits, IEEE Journal of,* vol. 45, pp. 668-680, 2010.

[8] M. Ashouei, J. Hulzink, M. Konijnenburg, Z. Jun, F. Duarte, A. Breeschoten*, et al.*, "A voltage-scalable biomedical signal processor running ECG using 13pJ/cycle at 1MHz and 0.4V," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, 2011, pp. 332-334.

[9] S. Fisher, A. Teman, D. Vaysman, A. Gertsman, O. Yadid-Pecht, and A. Fish, "Digital subthreshold logic design - motivation and challenges," in *Electrical and Electronics Engineers in Israel, 2008. IEEEI 2008. IEEE 25th Convention of*, 2008, pp. 702-706.

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

[11] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Low Power Electronics and Design, 2005. ISLPED '05. Proceedings of the 2005 International Symposium on*, 2005, pp. 20-25.

[12] P. R. van der Meer, A. van Staveren, and A. H. M. van Roermund, *Low-Power Deep Sub-Micron CMOS Logic: Sub-threshold Current Reduction*: Kluwer Academic, 2004.

[13] M. Blesken, Lu, x, S. tkemeier, and U. Ruckert, "Multiobjective optimization for transistor sizing sub-threshold CMOS logic standard cells," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 2010, pp. 1480-1483.

[14] B. Liu, M. Ashouei, J. Huisken, and J. Pineda de Gyvez, "Standard cell sizing for subthreshold operation," in *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, 2012, pp. 962-967.

[15] A. Wang, B. H. Calhoun, and A. P. Chandrakasan, *Sub-Threshold Design for Ultra Low-Power Systems*: Springer, 2006.

[16] J. Keane, E. Hanyong, K. Tae-Hyoung, S. Sapatnekar, and C. Kim, "Subthreshold logical effort: a systematic framework for optimal subthreshold device sizing," in *Design Automation Conference, 2006 43rd ACM/IEEE*, 2006, pp. 425-428.

[17] T. Gemmeke and M. Ashouei, "Variability aware cell library optimization for reliable sub-threshold operation," in *ESSCIRC (ESSCIRC), 2012 Proceedings of the*, 2012, pp. 42-45.

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

[19] M. Seok, D. Jeon, C. Chakrabarti, D. Blaauw, and D. Sylvester, "A 0.27V 30MHz 17.7nJ/transform 1024-pt complex FFT core with super-

pipelining," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, 2011, pp. 342-344.

[20] N. Reynders and W. Dehaene, "27.3 A 210mV 5MHz variation-resilient near-threshold JPEG encoder in 40nm CMOS," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, 2014, pp. 456-457.

[21] N. Reynders and W. Dehaene, "A 190mV supply, 10MHz, 90nm CMOS, pipelined sub-threshold adder using variation-resilient circuit techniques," in *Solid State Circuits Conference (A-SSCC), 2011 IEEE Asian*, 2011, pp. 113-116.

[22] P. Weckx, N. Reynders, I. de Moffarts, and W. Dehaene, "Design of a 150 mV Supply, 2 MIPS, 90nm CMOS, Ultra-Low-Power Microprocessor," in *Integrated Circuit and System Design. Power and Timing Modeling, Optimization and Simulation*. vol. 7606, J. Ayala, D. Shang, and A. Yakovlev, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 175-184.

[23] N. Reynders and W. Dehaene, "Variation-Resilient Building Blocks for Ultra-Low-Energy Sub-Threshold Design," *Circuits and Systems II: Express Briefs, IEEE Transactions on,* vol. 59, pp. 898-902, 2012.

[24] N. Reynders and W. Dehaene, "Variation-resilient sub-threshold circuit solutions for ultra-low-power Digital Signal Processors with 10MHz clock frequency," in *ESSCIRC (ESSCIRC), 2012 Proceedings of the*, 2012, pp. 474-477.

[25] J. Kwong, Y. Ramadass, N. Verma, M. Koesler, K. Huber, H. Moormann, *et al.*, "A 65nm Sub-Vt Microcontroller with Integrated SRAM and Switched-Capacitor DC-DC Converter," in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, 2008, pp. 318-616.

[26] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *Solid-State Circuits, IEEE Journal of,* vol. 40, pp. 1778-1786, 2005.

[27] D. Bol, D. Flandre, and J.-D. Legat, "Technology flavor selection and adaptive techniques for timing-constrained 45nm subthreshold circuits," presented at the Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design, San Fancisco, CA, USA, 2009.

[28] T.-H. Kim, E. Hanyong, J. Keane, and C. Kim, "Utilizing Reverse Short Channel Effect for Optimal Subthreshold Circuit Design," in *Low Power*

Electronics and Design, 2006. ISLPED'06. Proceedings of the 2006 International Symposium on, 2006, pp. 127-130.

[29] F. Abouzeid, S. Clerc, F. Firmin, M. Renaudin, and G. Sicard, "A 45nm CMOS 0.35v-optimized standard cell library for ultra-low power applications," presented at the Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design, San Fancisco, CA, USA, 2009.

[30] B. Liu, J. Pineda de Gyvez, and M. Ashouei, "Library tuning for subthreshold operation," in *Subthreshold Microelectronics Conference (SubVT), 2012 IEEE*, 2012, pp. 1-3.

[31] D. Bol, D. Flandre, and J.-D. Legat, "Nanometer MOSFET Effects on the Minimum-Energy Point of Sub-45nm Subthreshold Logic---Mitigation at Technology and Circuit Levels," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 16, pp. 1-26, 2010.

[32] H. Al-Hertani, D. Al-Khalili, and C. Rozon, "A new subthreshold leakage model for NMOS transistor stacks," in *Circuits and Systems, 2007. NEWCAS 2007. IEEE Northeast Workshop on*, 2007, pp. 972-975.

[33] D. Bol, J. De Vos, C. Hocquet, F. Botman, F. Durvaux, S. Boyd, *et al.*, "SleepWalker: A 25-MHz 0.4-V Sub-mm 7- uW/MHz Microcontroller in 65-nm LP/GP CMOS for Low-Carbon Wireless Sensor Nodes," *Solid-State Circuits, IEEE Journal of,* vol. 48, pp. 20-32, 2013.

[34] A. Wang, B. H. Calhoun, and A. P. Chandrakasan, *Sub-threshold design for ultra low-power systems*. New York: Springer, 2006.

[35] R. R. Troutman, "Subthreshold slope for insulated gate field-effect transistors," *Electron Devices, IEEE Transactions on,* vol. 22, pp. 1049-1051, 1975.

[36] B. Liu, H. R. Pourshaghaghi, S. M. Londono, and J. P. de Gyvez, "Process Variation Reduction for CMOS Logic Operating at Sub-threshold Supply Voltage," in *Digital System Design (DSD), 2011 14th Euromicro Conference on*, 2011, pp. 135-139.

[37] B. Liu, J. de Gyvez, and M. Ashouei, "Sub-Threshold Standard Cell Sizing Methodology and Library Comparison," *Journal of Low Power Electronics and Applications,* vol. 3, pp. 233-249, 2013.

[38] N. Bulusu and S. Jha, *Wireless sensor networks*: Artech House Boston, 2005.

[39] F. L. Lewis, "Wireless sensor networks," *Smart environments: technologies, protocols, and applications,* pp. 11-46, 2004.

[40] C. S. Raghavendra, K. M. Sivalingam, and T. Znati, *Wireless sensor networks*: Springer, 2004.

[41] F.-Y. Ren, H.-N. Huang, and C. Lin, "Wireless sensor networks," *Journal of software,* vol. 14, pp. 1282-1291, 2003.

[42] L. Seulki, Y. Long, R. Taehwan, H. Sunjoo, and Y. Hoi-Jun, "A 75 uW Real-Time Scalable Body Area Network Controller and a 25 uW ExG Sensor IC for Compact Sleep Monitoring Applications," *Solid-State Circuits, IEEE Journal of,* vol. 47, pp. 323-334, 2012.

[43] M. Meijer and J. P. de Gyvez, "Body bias driven design synthesis for optimum performance per area," in *Quality Electronic Design (ISQED), 2010 11th International Symposium on*, 2010, pp. 472-477.

[44] F. Di Franco, C. Tachtatzis, B. Graham, M. Bykowski, D. C. Tracey, N. F. Timmons*, et al.*, "The effect of body shape and gender on wireless Body Area Network on-body channels," in *Antennas and Propagation (MECAP), 2010 IEEE Middle East Conference on*, 2010, pp. 1-3.

[45] Y. Hoi-Jun and B. Joonsung, "Low energy wireless body area network systems," in *Wireless Symposium (IWS), 2013 IEEE International*, 2013, pp. 1-2.

[46] A. Arami, A. Barre, R. Berthelin, and K. Aminian, "Estimation of prosthetic knee angles via data fusion of implantable and wearable sensors," in *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, 2013, pp. 1-6.

[47] A. Bolz, V. Lang, B. Merkely, and M. Schaldach, "First results of an implantable sensor for blood flow measurement," in *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, 1997, pp. 2341-2343 vol.5.

[48] A. B. Islam, M. R. Haider, A. Atla, S. K. Islam, R. Croce, S. Vaddiraju*, et al.*, "A potentiostat circuit for multiple implantable electrochemical sensors," in *Electrical and Computer Engineering (ICECE), 2010 International Conference on*, 2010, pp. 314-317.

[49] R. D. Black, "Recent Advances in Translational Work on Implantable Sensors," *Sensors Journal, IEEE,* vol. 11, pp. 3171-3182, 2011.

[50] B. Liu, M. Ashouei, T. Gemmeke, and J. P. de Gyvez, "Sub-threshold custom standard cell library validation," in *Quality Electronic Design (ISQED), 2014 15th International Symposium on*, 2014, pp. 257-262.

[51] C. G. B. Garrett and W. H. Brattain, "Physical Theory of Semiconductor Surfaces," *Physical Review,* vol. 99, pp. 376-387, 07/15/ 1955.

[52] F. Wanlass and C. Sah, "Nanowatt logic using field-effect metal-oxide semiconductor triodes," in *Solid-State Circuits Conference. Digest of Technical Papers. 1963 IEEE International*, 1963, pp. 32-33.

[53] P. R. Gray, *Analysis and design of analog integrated circuits*. New York: Wiley, 2001.

[54] E. Vittoz and J. Fellrath, "New Analog CMOS IC'S Based on Weak Inversion Operation," in *Solid State Circuits Conference, 1976. ESSCIRC 76. 2nd European*, 1976, pp. 12-13.

[55] R. W. J. Barker, "Small-signal subthreshold model for i.g.f.e.t.s," *Electronics Letters,* vol. 12, pp. 260-262, 1976.

[56] R. R. Troutman and S. N. Chakravarti, "Subthreshold characteristics of insulated-gate field-effect transistors," *Circuit Theory, IEEE Transactions on,* vol. 20, pp. 659-665, 1973.

[57] T. Masuhara, J. Etoh, and M. Nagata, "A precise MOSFET model for low-voltage circuits," *Electron Devices, IEEE Transactions on,* vol. 21, pp. 363-371, 1974.

[58] E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operations," *Solid-State Circuits, IEEE Journal of,* vol. 12, pp. 224-231, 1977.

[59] C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integr. Circuits Signal Process.,* vol. 8, pp. 83-114, 1995.

[60] C. H. Kim, H. Soeleman, and K. Roy, "Ultra-low-power DLMS adaptive filter for hearing aid applications," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 11, pp. 1058-1067, 2003.

[61] E. L. Crow and K. Shimizu, *Lognormal Distributions: Theory and Applications*: M. Dekker, 1988.

[62] C. Subramanian, J. Hayden, W. Taylor, M. Orlowski, and T. McNelly, "Reverse short channel effect and channel length dependence of boron penetration in PMOSFETs," in *Electron Devices Meeting, 1995. IEDM '95., International*, 1995, pp. 423-426.

[63] N. C. C. Lu and J. M. Sung, "Reverse short-channel effects on threshold voltage in submicrometer salicide devices," *Electron Device Letters, IEEE,* vol. 10, pp. 446-448, 1989.

[64] J. M. Rabaey, *Low Power Design Essentials*: Springer-Verlag US, 2009.

[65] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *Solid-State Circuits, IEEE Journal of,* vol. 25, pp. 584-594, 1990.

[66] J. R. Tolbert, Z. Xin, L. Sung-Kyu, and S. Mukhopadhyay, "Analysis and Design of Energy and Slew Aware Subthreshold Clock Systems," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 30, pp. 1349-1358, 2011.

[67] L. Fenton, "The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems," *Communications Systems, IRE Transactions on,* vol. 8, pp. 57-67, 1960.

[68] K. Scott and K. Keutzer, "Improving cell libraries for synthesis," in *Custom Integrated Circuits Conference, 1994., Proceedings of the IEEE 1994*, 1994, pp. 128-131.

[69] C. Fisher, R. Blankenship, J. Jensen, T. Rossman, and K. Svilich, "Optimization of standard cell libraries for low power, high speed, or minimal area designs," in *Custom Integrated Circuits Conference, 1996., Proceedings of the IEEE 1996*, 1996, pp. 493-496.

[70] F. Beeftink, P. Kudva, D. Kung, and L. Stok, "Gate-size selection for standard cell libraries," in *Computer-Aided Design, 1998. ICCAD 98. Digest of Technical Papers. 1998 IEEE/ACM International Conference on*, 1998, pp. 545-550.

[71] T. Mozdzen, "Design methodology for a 1.0u cell-based library efficiently optimized for speed and area," in *ASIC Seminar and Exhibit, 1990. Proceedings., Third Annual IEEE*, 1990, pp. P12/3.1-P12/3.5.

[72] J. Bhasker and R. Chadha, *Static Timing Analysis for Nanometer Designs: A Practical Approach*: Springer, 2009.

[73] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *Solid-State Circuits, IEEE Journal of,* vol. 24, pp. 1433-1439, 1989.

[74] D. Blaauw, K. Chopra, A. Srivastava, and L. Scheffer, "Statistical Timing Analysis: From Basic Principles to State of the Art," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 27, pp. 589-607, 2008.

[75] M. Meijer, B. Liu, R. Van Veen, and J. P. de Gyvez, "Post-silicon tuning capabilities of 45nm low-power CMOS digital circuits," in *VLSI Circuits, 2009 Symposium on*, 2009, pp. 110-111.

[76] F. Corno, M. S. Reorda, and G. Squillero, "RT-level ITC'99 benchmarks and first ATPG results," *Design & Test of Computers, IEEE,* vol. 17, pp. 44-53, 2000.

LIST OF PUBLICATION

B. Liu, H. R. Pourshaghaghi, S. M. Londono, and J. P. de Gyvez, "Process Variation Reduction for CMOS Logic Operating at Sub-threshold Supply Voltage," in *Digital System Design (DSD), 2011 14th Euromicro Conference on*, 2011, pp. 135-139.

B. Liu, M. Ashouei, J. Huisken, and J. Pineda de Gyvez, "Standard cell sizing for subthreshold operation," in *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE,* 2012, pp. 962-967.

B. Liu, J. Pineda de Gyvez, and M. Ashouei, "Library tuning for subthreshold operation," in *Subthreshold Microelectronics Conference (SubVT), 2012 IEEE*, 2012, pp. 1-3.

B. Liu, J. de Gyvez, and M. Ashouei, "Sub-Threshold Standard Cell Sizing Methodology and Library Comparison," *Journal of Low Power Electronics and Applications*, vol. 3, pp. 233-249, 2013.

B. Liu, M. Ashouei, T. Gemmeke, and J. P. de Gyvez, "Sub-threshold custom standard cell library validation," in *Quality Electronic Design (ISQED), 2014 15th International Symposium on*, 2014, pp. 257-262.

B. Liu, H. Jiao, T. Gemmeke, and J. Pineda de Gyvez, "Ultra-Low Power Circuit Design Based on An Ultra-Low Voltage Standard Cell Library" in *International Symposium on Circuits and Systems (ISCAS), 2015. Submitted*

T. Gemmeke, M. Ashouei, B. Liu, M. Meixner, T. G. Noll, and H. de Groot, "Cell libraries for robust low-voltage operation in nanometer technologies," *Solid-State Electronics,* vol. 84, pp. 132-141, 2013.

About Me

Bo Liu was born on 18-06-1984 in Harbin China.

After finishing Bachelor of Science in 2007 at Zhejiang University in Hangzhou, China, he studied Electrical Engineering at Eindhoven University of Technology in Eindhoven, the Netherlands. In 2009 he graduated within the Electronic system group on Run-Time Performance boosting with Body Biasing Islands in CMOS Digital Circuits. From Nov 2009 he started a PhD project at Eindhoven University of Technology at Eindhoven, the Netherlands of which the results are presented in this dissertation. Since 2014 he is employed at Holst Centre/Imec-nl.