<table>
<thead>
<tr>
<th>Title</th>
<th>Performance Modeling and On-Chip Memory Structures for Minimum Energy Operation in Voltage-Scaled LSI Circuits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Shiomi, Jun</td>
</tr>
<tr>
<td>Citation</td>
<td>Kyoto University (京都大学)</td>
</tr>
<tr>
<td>Issue Date</td>
<td>2017-11-24</td>
</tr>
<tr>
<td>URL</td>
<td><a href="https://doi.org/10.14989/doctor.k20778">https://doi.org/10.14989/doctor.k20778</a></td>
</tr>
<tr>
<td>Right</td>
<td>表示できない文字が含まれています。URLにてご確認下さい。</td>
</tr>
<tr>
<td>Type</td>
<td>Thesis or Dissertation</td>
</tr>
<tr>
<td>Textversion</td>
<td>ETD</td>
</tr>
</tbody>
</table>

Kyoto University
Performance Modeling and On-Chip Memory Structures for Minimum Energy Operation in Voltage-Scaled LSI Circuits

Jun Shiomi
Acknowledgments

This thesis has been supported by a lot of people. I would like to express my appreciation to all the people who have encouraged me to pursue my Ph.D. degree.

Firstly, I would like to express my deepest gratitude to my adviser, Professor Hidetoshi Onodera in Kyoto University for his support over the years. I cannot thank him enough for his tremendous support and help. I deeply thank him for giving me precious opportunity and excellent environment to study as a Ph.D student. He is the best professor I have ever met and my role model as an engineer.

Next, I would like to express my sincere gratitude to Associate Professor Tohru Ishihara in Kyoto University for his tremendous support. He has read all my papers and gave me many valuable comments to keep me on track. He always discussed with me and guided me whenever I need his help. The completion of this thesis would have been difficult without his continuous support.

I deeply thank Associate Professor Akira Tsuchiya in University of Shiga Prefecture. He gave me hundreds of valuable suggestion on my research when he was in Kyoto University. His vast knowledge on analog circuits and electromagnetism always surprised me. His excellent comments broadened my horizons.

I am grateful to Professor Takashi Sato and Professor Sadao Kurohashi in Kyoto University for their efforts to read through my thesis and their helpful comments to make the thesis better. Their sophisticated comments considerably improved my thesis.

I express my gratitude to all the Onodera Laboratory members. I thank Assistant Professor Shinichi Nishizawa and Assistant Professor A.K.M Mahfuzul Islam for their precious supports. Their vast knowledge on performance variability of CMOS circuits and the design methodology on low-power digital circuits helped me come up with new ideas in my research. Their kind support in designing test chips was indispensable to finalize my thesis. Special thanks go to Mr. Tatsuya Kamakari. His analytical stability model for estimating yields of latch cells has played an important role in designing latch cells for memory circuits in Chapter 4 of this thesis. I also deeply thank Mr. Masamichi Fujiwara. Discussion and chatting with Mr. Tatsuya Kamakari and Mr. Masamichi Fujiwara were
essential for my research activity and my academic life. I am also grateful to Mr. Norihiro Kamae for his generous support. His wealth of knowledge about LSI circuit design greatly helped the test chip design. I am also grateful to Mr. Kondo Masahiro who designed a number of standard-cell libraries in various process technologies. Without his libraries, designing the processor in Chapters 5 would have been impossible. I also thank Ms. Kim SinNyong, Assistant Professor Jun Furuta, Assistant Professor Takashi Matsumoto, Mr. Xiu Qi, Mr. Taro Amagai, Mr. Shohei Nishimura, Mr. Toshihiro Takeshita, Mr. Shinichi Nakanishi and the other students for their kind support. I would like to deeply thank laboratory Secretary Seiko Jinno for her generous support through my student life.

The test chips are fabricated by the support of VLSI Design and Education Center (VDEC), the University of Tokyo. I would like to deeply thank them for the tremendous support. I would like to express my appreciation to Japan Society for the Promotion and Science (JSPS) for their financial support as a Fellowship for Young Scientists (DC1).

Finally, I deeply thank my parents, Tomonori Shiomi and Hisayo Shiomi for their support of my long student life.

塩見 準
Jun Shiomi
in Kyoto
October 2017
Abstract

Performance Modeling and On-Chip Memory Structures for Minimum Energy Operation in Voltage-Scaled LSI Circuits
by
Jun Shiomi
Doctor of Philosophy in Informatics
Kyoto University
Professor Hidetoshi Onodera, Chair

With the rapid development of the information and communication technology, energy-efficient Large Scale Integration (LSI) circuits are highly required. Scaling the supply voltage is a promising approach to reduce the energy consumption of LSI circuits. The goal of this thesis is thus to minimize the energy consumption of LSI circuits using the voltage scaling technique. However, designing voltage-scaled LSI circuits is associated with a considerable design effort. A major concern is the performance variation of the voltage-scaled circuits, which makes the circuit design strategies different depending on the operating voltage. The first goals of this thesis are thus performance modeling of voltage-scaled circuits and developing circuit design strategies for energy-efficient voltage-scaled circuits.

This thesis discusses the architectural-level circuit design strategies for both the nominal voltage operation and the low voltage operation. Based on the simple transregional performance modeling of LSI circuits, this thesis shows that random variations in transistors have different impacts on the circuit delay depending on the operating voltage. Although techniques estimating the performance variation of LSI circuits are widely studied in the nominal voltage circuits, the counterparts in the aggressively voltage scaled region are not fully studied. This thesis firstly develops the basic statistical operations including the SUM and MAX operations in the aggressively voltage scaled region, which enables
to estimate the delay variation of LSI circuits in a closed form. After that, we prove several theorems that help consider architectural design strategies for high performance and energy efficient LSI circuits. Based on the theorems, this thesis derives different architectural-level circuit design strategies depending on the operating voltage.

On-chip memory is one of the most vulnerable components in aggressively voltage-scaled LSI circuits. Therefore, designing on-chip memories for the aggressively scaled voltage operation is highly required to improve the energy-efficiency of LSI circuits. Recently, several papers have proposed Standard-Cell Memory (SCM) structures as an alternative to a traditional on-chip SRAM for the low voltage operation. Since the SCM consists of digital standard-cells only, it can thus reduce the custom design effort to the level of fully automated cell based design with keeping their stability in aggressively voltage-scaled region. However, SCM occupies a large part of the chip area compared with full-custom SRAM macros, which directly leads to the loss of the computational power of LSI circuits. To solve this problem, this thesis focuses on improving area-efficiency using minimum height standard-cells. Unlike conventional SCMs, the proposed SCM has standard-cells with a minimum possible cell height allowed by the logic design rule of the target technology. This thesis also presents energy efficient readout and write schemes for reducing dynamic energy consumption. Post layout simulation using a 65-nm technology shows that the proposed SCM achieves the area of 6.82 $\mu$m$^2$ per bit ($682F^2$ per bit), which is 20% smaller than that of the state of the art SCMs. The results also show that the energy consumption of the proposed SCM is 21% and 31% smaller than that of the state of the art SCMs and low voltage SRAMs, respectively.

Finally, a technique dynamically tuning the supply voltage ($V_{DD}$) and the threshold voltage ($V_{th}$) over a wide operating performance range is discussed as a post-silicon tuning method. Scaling $V_{DD}$ and $V_{th}$ dynamically has a strong impact on energy efficiency of LSI circuits. Therefore, techniques for optimizing $V_{DD}$ and $V_{th}$ simultaneously under dynamic workloads are required but not fully studied. In this thesis, we refer to the optimum pair of $V_{DD}$ and $V_{th}$, which minimizes the energy consumption of a circuit under a specific performance constraint, as a minimum energy point (MEP). Based on the simple transregional models of a CMOS circuit, this thesis derives a simple necessary and sufficient condition for the MEP operation. The simple condition helps find the MEP of CMOS circuits. Measurement results using SCMs implemented into an embedded processor which is fabricated using a 65-nm process technology also validate the condition derived in this thesis. The measurement results show that the MEP operation using simultaneous tuning of $V_{DD}$ and $V_{th}$ achieves 44% less energy consumption at the best case than the conventional DVFS technique.
Voltage scaling is a key technique to reduce the energy consumption of LSI circuits. The techniques proposed in this thesis help to design energy-efficient LSI circuits operating over a wide supply range. The proposed area-efficient SCM structures improve the voltage scalability of LSI circuits. The LSI circuits integrating the SCMs can operate over a wide supply range down to the sub-threshold region where \( V_{DD} \) is below \( V_{th} \), which further improves the energy-efficiency by voltage scaling. The simultaneous optimization of \( V_{DD} \) and \( V_{th} \) proposed in this thesis enables LSI circuits to operate with the minimum possible energy consumption without violating the application deadline. Since the energy consumption is a limiting factor in LSI circuits, the proposed energy-efficient voltage-scaled LSI circuits support the continuous development of the information and communication technology.
# Contents

Acknowledgments i

Abstract iii

1 Introduction 1
   1.1 Background .................................................. 1
   1.2 Motivation: Voltage Scaling ................................. 3
   1.3 Challenges toward Voltage Scaling .......................... 4
       1.3.1 Performance Variability in the Low Voltage Operation 4
       1.3.2 Variability-Aware On-Chip Memory ....................... 5
       1.3.3 Optimum Supply Voltage and Threshold Voltage Selection 5
   1.4 Research Goal and Thesis Contribution ..................... 6
   1.5 Thesis Organization ......................................... 8

2 Literature Review 9
   2.1 Performance Models for LSI Circuits ....................... 9
       2.1.1 Delay and Energy Models ............................... 9
       2.1.2 Performance Variability in CMOS Circuits ............... 10
   2.2 Techniques toward Energy-Efficient Circuit Design ........ 13
       2.2.1 Variability-Aware Architectural Design ................ 13
       2.2.2 On-Chip Memory for Low Voltage Operation ............. 15
       2.2.3 Transistor Sizing ..................................... 16
   2.3 Post-Silicon Tuning for Energy-Efficiency Improvement .... 17
       2.3.1 Dynamic Voltage and Frequency Scaling ................. 18
       2.3.2 Adaptive Body Biasing ................................. 18
       2.3.3 Minimum Energy Point Tracking using Combined Supply and Threshold Voltage Scaling ............... 19
   2.4 Summary ..................................................... 19
## 3 Statistical Timing Modeling for Voltage-Scaled Circuit Design 21

3.1 Introduction ................................. 21

3.2 Preliminaries ................................. 22
  3.2.1 Properties of Lognormal Distribution .......... 22
  3.2.2 Delay Distribution for the Super-Threshold Voltage Operation .. 22
  3.2.3 Delay Distribution for the Near-Threshold Voltage Operation ... 23
  3.2.4 Delay Distribution for the Sub-Threshold Voltage Operation ... 24

3.3 Lognormal Timing Model ....................... 24
  3.3.1 SUM Operation for Lognormal Distributions ......... 25
  3.3.2 Averaging Effect of Paths .................... 27
  3.3.3 Delay Distribution of Parallel Paths ............... 28
  3.3.4 Impact of Gate Sizing on Delay Variation .......... 31

3.4 Validation with a 28-nm Process Technology Model .......... 32
  3.4.1 Delay Distribution of a Buffer Chain ............. 32
  3.4.2 Validation of Averaging Effect .................. 33
  3.4.3 Validation of Fmax Degradation Speed .............. 34
  3.4.4 Validation of Gate Sizing Effect ................ 37
  3.4.5 Validation of Properties in Other Logic Cells .... 39

3.5 Summary of Properties .......................... 42

3.6 Application Example .......................... 44
  3.6.1 Pipelining ................................ 44
  3.6.2 Standard-Cell Memory Readout Circuit ............ 45

3.7 Summary ...................................... 48

## 4 Area- and Energy-Efficient Standard-Cell Memory using Minimum Height Standard-Cells 51

4.1 Introduction .................................. 51

4.2 Minimum Height Standard-Cell ................... 52
  4.2.1 Concept .................................. 52
  4.2.2 Physical Design ............................ 53
  4.2.3 Simplified Cell Library ...................... 55
  4.2.4 Simplified Latch for Energy- and Area-Efficiency .... 55

4.3 Energy-Efficient Memory Architecture ............. 56
  4.3.1 Write Scheme ................................ 57
  4.3.2 Readout Scheme ............................ 58
4.4.1 Stability Analysis for Latch Cells and Standard-Cell Design .... 59
4.4.2 Comparison with Prior-Art SCMs ........................................ 63
4.5 Summary .............................................................................. 68

5 Minimum Energy Point Operation using Supply and Threshold Voltage Scaling ................................................................. 71
5.1 Introduction .......................................................................... 71
5.2 Modeling for Minimum Energy Point Operation ....................... 73
5.2.1 Necessary Condition for Minimum Energy Point Operation .... 74
5.2.2 Minimum Energy Curve in Sub-threshold Region ................. 75
5.2.3 Minimum Energy Curve in Near-threshold Region ............... 77
5.2.4 Minimum Energy Curve in Super-threshold Region ............... 78
5.2.5 Impact of DIBL on Minimum Energy Curve ....................... 80
5.2.6 Impacts of Temperature and Activity Factor ....................... 81
5.2.7 Summary of Properties .................................................... 82
5.3 Silicon Measurements ............................................................ 83
5.3.1 Test Chip Architecture ..................................................... 83
5.3.2 Minimum Energy Curves of Standard-Cell Memories .......... 86
5.3.3 Impacts of Activity Factor ................................................ 92
5.4 Voltage Scaling Strategy for Minimum Energy Point Operation .... 93
5.5 Summary .............................................................................. 94

6 Conclusion .............................................................................. 95
6.1 Summary of This Thesis ....................................................... 95
6.2 Energy Reduction by Standard-Cell Memories with the Minimum Energy Point Operation ..................................................... 97
6.3 Future Work ........................................................................... 98

Publication List .................................................................. 111
List of Figures

2.1 Atomic operations in SSTA. (a) SUM operation. (b) MAX operation. . . . 12
2.2 The definition of $k\sigma$ worst case delay. ............................... 12
2.3 (a) Normal architecture. (b) Parallelism. (c) Pipelining. .................. 14
3.1 Definition of averaging effect ratio. ........................................ 28
3.2 Buffer chain example where all buffers have the same fan-out. ......... 32
3.3 Buffer chain simulation result ($V_{DD} = 0.4$ V). $\mu = -21$, $\sigma = 0.21$ and $r = 0.32$. ......................................................... 33
3.4 Averaging effect ratio for buffer chains. .................................. 34
3.5 Logic depth vs. the $4\sigma$ worst case delay. ................................. 34
3.6 Parallelism of 8-stage-buffer chains. ...................................... 35
3.7 The number of critical paths $N_{cp}$ vs. the $3\sigma$ worst case delay. The logic depth is 8. ....................................................... 36
3.8 The number of critical paths $N_{cp}$ vs. the $3\sigma$ worst case delay. The logic depth is 1. ......................................................... 36
3.9 Delay distributions for different gate sizes. ................................ 37
3.10 Buffer size X vs. $4\sigma$ worst case delay. ................................. 38
3.11 The $4\sigma$ worst case delay for different gate sizes. ....................... 38
3.12 Test circuit structure for NAND2 and NOR2. ............................. 39
3.13 Averaging effect ratio for NAND2/NOR2 chains. ......................... 39
3.14 Logic depth vs. $4\sigma$ worst case delay for NAND2/NOR2 chains. .... 40
3.15 The number of critical paths $N_{cp}$ vs. the $3\sigma$ worst case delay for NAND2/NOR2 chains. .................................................. 41
3.16 Gate size X vs. $4\sigma$ worst case delay for NAND2/NOR2 chains. ... 41
3.17 The $4\sigma$ worst case delay of NAND2/NOR2 chains with different gate sizes. ................................................................. 42
3.18 $p$-parallel $n$-stage buffer chains where all buffers in chains have the same gate size X. ......................................................... 43
3.19 Performance gain by pipelining. .............................................. 45
3.20 Memory readout structure. (a) SRAM (b) SCM. .......................... 47
3.21 CDF versus readout delay. .................................................. 48
4.1 The concept of minimum height standard-cells. ......................... 52
4.2 An inverter cell with minimum cell height. .............................. 53
4.3 Simplified latch schematic and clock-shared 4-bit latch. ............... 56
4.4 Proposed SCM structure. ..................................................... 56
4.5 Write clocking scheme of the proposed SCM. ........................... 57
4.6 Readout scheme of the proposed SCM. ................................... 59
4.7 (a) Schematic of cross-coupled inverters. (b) Butterfly curve of cross-coupled inverters. ......................................................... 60
4.8 Verification of the analytical stability model of latch cells (4.2). ..... 62
4.9 Yields of latch cells for various gate widths. ............................ 62
4.10 The layout of minimum height standard-cells. .......................... 64
4.11 Layouts of the proposed 16 kb SCM (512x32). ........................ 64
4.12 Area-comparison between the proposed SCMs, prior-art SCMs and SRAMs. The area of the SCMs in [1] is multiplied by \((100 \text{ nm}/50 \text{ nm})^2 = 4\). ..... 65
4.13 Estimated maximum operating frequency with a scaled \(V_{DD}\). ........ 66
4.14 Estimated write energy consumption per bit with a scaled \(V_{DD}\). 69
4.15 Estimated read energy consumption per bit with a scaled \(V_{DD}\). 69
4.16 Estimated sleep energy consumption per bit with a scaled \(V_{DD}\). 69
4.17 Leakage power per bit with a scaled \(V_{DD}\). .............................. 70
5.1 Test circuit: 50-stage fanout-4 inverter chain. ......................... 73
5.2 Energy and performance contours for a 50-stage inverter chain. Solid line: energy contour. Dashed line: performance contour. Bold line: minimum energy curve. ......................................................... 74
5.3 Minimum energy points in sub-threshold region. ....................... 76
5.4 Minimum energy curve of a circuit designed with a 28-nm process technology. ................................................................. 77
5.5 Minimum energy points in super-threshold region. .................... 80
5.6 Minimum energy curves for different temperature and activity. ....... 81
5.7 Chip photograph of the fabricated RISC processor in a 65-nm process technology. ................................................................. 84
5.8 The SCM structure. ............................................................. 85
5.9 Minimum energy curve of the SCM. Solid line: energy contour [nJ/cycle]. Dashed line: Fmax contour. Bold line: minimum energy curve. ......... 87
5.10 Energy consumption for various operating frequencies. $V_{BB}$ is fixed at
−0.4 V for the conventional DVFS technique. .......................... 88
5.11 Energy consumption for various operating frequencies. $V_{BB}$ is fixed at
−0.97 V for the conventional DVFS technique. ......................... 89
5.12 $E_d/E_s$ ratio on MEPs. ....................................................... 89
5.13 $E_d/E_s$ ratio on 391 kHz Fmax contour. ............................. 90
5.14 $E_d/E_s$ ratio on 8 MHz Fmax contour. ............................... 91
5.15 $E_d/E_s$ ratio on 28.57 MHz Fmax contour. ......................... 91
5.16 Definition of the parameter $\alpha_M$. ................................. 92
5.17 Minimum energy curve of the SCM for $\alpha_M = 0.1$. Solid line: energy
contour [nJ/cycle]. Dashed line: Fmax contour. Bold line: minimum
energy curves. ................................................................. 93
List of Tables

3.1 Summary of properties. C: Corollary. L: Lemma. T: Theorem. \( p \): degree of parallelism. \( N \): logic depth. \( W \): gate width. \( L \): gate length. STV: Super-Threshold Voltage. ......................................................... 42

4.1 5.5-track minimum height standard cell library in the target 65-nm FD-SOI process technology. ......................................................... 63

4.2 Comparison between Prior-Art SCMs and SRAMs. ........................ 67

5.1 5.5-track minimum height standard cell library in the target 65-nm FD-SOI process technology. ......................................................... 85
Chapter 1

Introduction

The goal of this thesis is to minimize the energy consumption of Large Scale Integration (LSI) circuits using voltage scaling techniques. Downscaling the supply voltage of LSI circuits is the most effective way to reduce their energy consumption. However, LSI performance variability is severely degraded in the low supply voltage, which causes the LSI malfunction and thus limits the voltage scalability of the LSI circuits. Therefore, variability-aware circuit design techniques are highly required. This chapter discusses the background and motivation of this research. Challenges toward the voltage scaling techniques and contribution of this thesis are presented.

1.1 Background

Large Scale Integration (LSI) circuit is one of the most essential technologies in our information society. With the aggressive scaling of Metal-Oxide-Semiconductor Field-Effect Transistors (MOSFETs), the number of MOSFETs which can be integrated to a single chip has been continuously increasing, which enables a single chip to perform various complex functions today. This has accelerated the rapid growth of new technologies in our information society. LSI circuits are embedded to almost all commercial products for various purposes from high-performance computing to battery powered low-power devices. Thus, they play important roles in our information society.

Over several decades, the development in LSI circuits was mainly aiming at enhancing computational power. However, in a modern information society, power and energy consumptions have also become key metrics in LSI circuits. For example, designing low-power and energy-efficient LSI circuits is essential in the following reasons.

- **High-Performance Computing (HPC).** In 2010s, Artificial Intelligence (AI) including deep learning is a hot topic and one of the most rapidly growing areas in
HPC platforms. In HPC platforms, power dissipation rather than a computational power is a limiting factor since cooling systems can no longer handle ever increasing heats dissipated from the platforms. Energy-efficient platforms are also highly required from the viewpoint of low-carbon societies.

- **Mobile Computing.** Laptops, smartphones, tablet computers and wearable devices are essential mobile computing devices in our daily lives. One of major concerns in the mobile computing platforms is a battery lifetime, which pushes designers to make LSI circuits with a low energy consumption. Recently, mobile computing devices have also required SoCs to handle complex multimedia data. Therefore, achieving both high computational power and high energy-efficiency is extremely crucial for these SoCs.

- **Sensor-based Computing.** In Wireless Sensor Networks (WSN), which is a core technology in Internet of Things (IoT), billions of wireless sensor devices are interconnected each other, which enables an autonomous exchange of information. While a computational power is not a key metric, sensor devices in WSN typically require ultra low power circuits operating only with a limited battery capacity or an ambient energy source.

Traditionally, the transistor scaling is one of the most effective techniques to reduce the energy consumption, the power dissipation and a design cost. The horizontal and vertical feature sizes of MOSFETs, along with the supply voltage and the threshold voltage, are reduced by the factor of $k$ per generation, where $k$ is typically 0.7. In 1965, Gordon Moore observed that the number of MOSFETs integrated on a single chip increases by a factor of two every two years [2]. Based on the simple performance models, Robert Dennard pointed out that the transistor scaling gives designers better speed, less power dissipation, less energy consumption and smaller area than those in a scenario without the scaling [3], which has encouraged the continuous transistor scaling. In 2017, the minimum feature size of commercial microprocessors has reached 14 nm while it was 10 $\mu$m in 1974.

As we enter a deca-nanometer era, semiconductor fabrication technologies have faced physical limits. The gate oxide thickness, which is the most critical vertical dimension, has reached an atomic scale around a 130-nm process technology and has eventually stopped due to intolerable tunneling leakage current through the gate oxide [4]. By 2005, the scaling of the threshold voltage and the supply voltage has stopped due to an unacceptable amount of the sub-threshold leakage current [5]. As the horizontal feature size approaches an atomic scale, the short-channel effect including Drain-Induced Barrier Lowering (DIBL) and velocity saturation has not been a negligible effect, which degrades
1.2 Motivation: Voltage Scaling

Power and energy dissipation as well as a computational power are key metrics in modern LSI circuits. Supply voltage scaling is one of the most effective approaches to reduce power and energy dissipation. In typical circuits, the dominant source of energy consumption is a dynamic energy consumption which is consumed when Complementary MOS (CMOS) circuits charge or discharge load capacitors. Since the dynamic energy consumption is quadratically proportional to the supply voltage, the low voltage operation is a good choice to reduce the power and energy dissipation at the cost of degrading the computational power. Early studies related to the low-voltage operations are the sub-threshold operation where the supply voltage is scaled down below the threshold voltage. In 1972, Meindl et al. pointed out the merits by the sub-threshold operation by means of...
finding the minimum supply voltage for ideal CMOS inverters [12, 13]. In 1999, Soeleman et al. proposed sub-threshold digital circuits as a promising approach for ultra-low power circuits [14]. However, as the supply voltage is scaled down below the threshold voltage, the circuit performance is exponentially degraded. Typical operating frequency of sub-threshold processors such as Ref. [15] is in the order of kilohertz. Therefore, their applications will be limited to a specific application domain which does not need high performance but requires extremely low power consumptions. For HPC computing and mobile computing applications, the computational-power of the sub-threshold logic circuits is no longer satisfiable.

As a solution to this issue, a concept of near-threshold computing has emerged [16, 17]. It scales down the supply voltage near the threshold voltage, which brings quadratic dynamic energy savings with keeping the performance degradation of the circuit to a minimum. Ref. [17], for example, proposed the voltage-scalable processor across the wide voltage range from 280 mV to 1.2 V in a 32-nm process technology. It achieved 4.7 times energy-efficiency in the near-threshold voltage operation compared with a nominal voltage operation at the cost of performance degradation to one-ninth. The degraded performance is still acceptable compared with that of sub-threshold circuits. Thus, near-threshold circuits can be applied to several applications such as mobile computing where both a computational power and an energy-efficiency are required.

1.3 Challenges toward Voltage Scaling

1.3.1 Performance Variability in the Low Voltage Operation

The aggressively scaled voltage operation poses us several severe challenges. In the sub-/near-threshold voltage operation, the threshold voltage fluctuation has an exponential impact on the operating speed of LSI circuits [18]. Therefore, performance variability is a crucial problem in the aggressively scaled voltage region. For example, Ref. [16] pointed out that the relative delay variations in CMOS logic gates in the near-threshold voltage are expanded up to five times compared with those in the nominal voltage. Although the variation impacts on the circuit performance have been extensively studied over the past several decades, the counterparts in the sub-/near-threshold voltage operation are not fully studied. If we design the low voltage LSI circuits using the conventional worst-case design, too pessimistic margins are required, which leads to the overhead of the energy consumption and the operating speed. The expanded performance variation implies that the optimum architecture for LSI circuits are different depending on the operating voltage.
Since accurate modeling of the circuit performance leads to LSI circuits running with better performance and energy-efficiency, modeling of the performance variability and developing design strategies for low-voltage circuits are highly required.

The increased performance variation also brings functional failures in the voltage-scaled circuits. The exponential variation may cause timing violations between pipeline registers. It may also cause a retention failure of flip-flops. 6T SRAM macros may no longer successfully operate in readout, write or hold operation. Since the functional failure affects all LSI circuits from embedded processors to HPC processors, developing variability-aware sub-/near-threshold circuit design methodology is a must.

1.3.2 Variability-Aware On-Chip Memory

As mentioned in the previous subsection, designing low-voltage circuits without functional failures is a severe challenge. Among the components in LSI circuits, 6T SRAM macros, which are typical on-chip memories in LSI circuits, are the most vulnerable components to the threshold voltage variation. According to Ref. [19], random $V_{th}$ variation becomes large when transistor’s channel area is small. Therefore, SRAM whose bit cell is usually designed with the smallest size allowed by its design rule has been traditionally one of the most variability sensitive components. It is also pointed out that the peripherals of SRAM arrays are more sensitive to the $V_{th}$ variation than that of conventional digital circuits. Performance variations in the voltage scaled region may cause the wrong trigger timing of sense amplifiers and precharge circuitries. Mismatches in the sense amplifiers also may degrade the readability. Therefore, the conventional 6T SRAM macros are no longer satisfiable for voltage-scaled circuits.

1.3.3 Optimum Supply Voltage and Threshold Voltage Selection

The selection of the threshold voltage has also a large impact on energy-efficiency of voltage-scaled circuits. Refs. [20, 21] pointed out that optimizing the threshold voltage depending on the circuit architecture can reduce the energy consumption of LSI circuits. They experimentally show that optimizing the supply and threshold voltages for every functional block (e.g., logic circuits and memory macros) can effectively reduce the energy consumption without degrading the circuit operating speed. Therefore, the selection of the optimum threshold voltage is also important from the viewpoint of energy-efficiency.

What makes the selection complicated is that the optimum pair of the supply and threshold voltages depends on not only the circuit architecture but also a number of pa-
rameters including (1) an activity factor, (2) input vectors, (3) chip temperature, and (4) process variations [22]. Since these parameters dynamically change after the chip fabrication, post-silicon tuning is a must for energy-efficient LSI circuits. Although there are several techniques dynamically optimizing the threshold voltage (and the supply voltage) for improving energy-efficiency [22, 23], the techniques over a wide operating performance range are not fully studied. More specifically, there exists a pair of the supply voltage and the threshold voltage that minimizes the energy consumption of LSI circuits under a specific clock period. To effectively improve the energy-efficiency of LSI circuits, the optimum pair should be found in not only the nominal voltage operation but also the low voltage operation. However, the methods to find the optimum pair over the wide operating performance range are still the open problems.

1.4 Research Goal and Thesis Contribution

The goal of this thesis is to minimize the energy consumption of LSI circuits using voltage scaling techniques. Supply voltage scaling is a promising approach for developing energy-efficient LSI circuits. However, the supply voltage scaling poses us challenging issues such as performance variations and functional failures. The following techniques are required for achieving energy-efficient voltage-scaled LSI circuits:

1. Variability-aware performance models,

2. Circuit design strategies overcoming the variability issues, and

3. Optimum voltage selection to minimize the energy consumption.

Although the first two items have been fully studied for the nominal voltage operation, the counterparts in the low voltage operations are still the open problems. The methods to find the optimum pair of the supply voltage and the threshold voltage are also the open problems. To solve the problems, this thesis presents performance modeling, circuit design strategies and voltage scaling techniques for energy-efficient LSI circuits. The contribution of this thesis is summarized below.

- **Delay Variation Model for Voltage-Scaled Circuits**
  This thesis shows statistical models for random delay variability in low-voltage LSI circuits. In the aggressively scaled process technologies, the random performance fluctuations of MOSFETs are the critical concerns in designing LSI circuits. Since the fluctuations degrade the operating speed of LSI circuits, estimating the delay
variation of LSI circuits are highly required. In the last fifteen years, techniques for estimating delay variations for the nominal voltage operation have been widely studied. They assume that the delay distributions of the circuits follow the Gaussian distribution. This makes it simpler to perform the SUM and the MAX operations for estimating the performance variation of the targeting circuit. Recent literature \[18\] revealed that the delay distributions in the sub-/near-threshold voltage operation do not follow the Gaussian distribution, but follow lognormal distribution. Based on this fact, this paper derives delay variability models for voltage-scaled circuits. After that, this thesis proves several theorems that help consider design strategies for improving performance of voltage-scaled circuits in terms of an architectural-level circuit design including transistor sizing, pipelining and parallelization of LSI circuits. Based on the theorems, this thesis shows that lognormal delay distributions in voltage-scaled circuits have different impacts on architectural-level circuit design compared with those of the nominal voltage circuits. Since decisions taken at the architectural-level have a significant impact on both a computational power and an energy consumption, the architectural-level model is crucial to the voltage-scaled circuit design.

- **Area- and Energy-Efficient Standard-Cell Memory (SCM)**
  As an alternative to SRAM macros, Standard-Cell Memories (SCMs) have been widely studied over the past decade for the low voltage operation \[1, 15, 24–27\]. Since only standard-cells are used for SCMs, their custom design effort for SCMs can be reduced to the level of fully automated cell based design with keeping their stability even in sub-/near-threshold voltage operation, which leads to the drastic reduction of power and energy consumptions. However, SCMs typically consume several times larger area than SRAM macros. The more area on-chip memories consume, the less memory capacity LSI circuits have. The area overhead directly leads to the loss of computational power in LSI circuits. Therefore, designers suffer from severely degraded computational power in the sub-/near-threshold voltage operation if SCMs are simply implemented in LSI circuits.

This thesis proposes a cell-level area optimization approach to address this problem. More specifically, this thesis proposes Minimum Height Standard-Cells (MHSCs) which have a minimum possible cell height allowed by the logic design rule of a target technology. We describe a design method to determine the physical layout of MHSCs and the logics required in an MHSC library. As a result, the proposed SCM in a 65-nm process technology achieves area efficiency of 6.82 $\mu$m$^2$ per bit.
(682\(F^2\) per bit) which is 20% better than that of the state of the art SCMs.

- **Minimum Energy Point Operation**

SCMs enable LSI circuits to operate over a wide operating performance range down even to the sub-threshold voltage. Therefore, designers can enjoy further energy and power reductions by dynamic voltage tuning in the sub-/near-threshold voltage. Thus, it is natural that the focus of this thesis is then shifted to the method to dynamically and simultaneously tune the supply and threshold voltage over a wide operating performance range for energy-efficient LSI circuits. In this thesis, we refer to the pair of the supply voltage and the threshold voltage that minimizes the energy consumption of a circuit under a specific clock period as a minimum energy point (MEP). As described in Section 1.3, it is not trivial to find MEPs of even a simple inverter chain since MEPs depend on a number of parameters of a target circuit. Based on the simple performance models of a circuit, this thesis derives a simple necessary and sufficient condition for the MEP operation. The simple condition helps dynamically find the MEP of LSI circuits. Measurement results using SCMs fabricated in a 65-nm process technology also validate the condition derived in this thesis. This thesis also shows that the MEP operation using the simultaneous tuning of the supply voltage and the threshold voltage achieves less energy consumption than the conventional supply voltage scaling technique. The measurement results using SCMs show that the MEP operation achieves 44% less energy consumption at the best case than the conventional technique just scaling the supply voltage only.

### 1.5 Thesis Organization

This thesis is organized in the following way. In Chapter 2, the reviews of related works are presented in order to make the contribution of this thesis clear. The issues in the related works and the contribution of this thesis are presented. Chapter 3 presents the delay variation models over a wide operating performance range. In Chapter 3, this thesis shows that the optimum architectural-level design strategies change depending on the operating voltage. Chapter 4 presents area- and energy-efficient SCM structure. Chapter 5 presents a simple necessary and sufficient condition which helps designers dynamically find the MEPs. Chapter 6 concludes this thesis.
Chapter 2

Literature Review

The goal of this thesis is to minimize the energy consumption of LSI circuits using voltage scaling techniques. Designing voltage-scaled LSI circuits is associated with a considerable design effort due to several challenging issues such as performance variations and functional failures. To solve the problems, this thesis aims to develop (1) variability-aware performance models, (2) circuit design strategies overcoming the variability issues, and (3) optimum voltage selection techniques to minimize the energy consumption. There are a number of previous works related to the goals. In Chapter 2, reviews of related works are presented in order to make the contribution of this thesis clear. Section 2.1 presents the performance models for CMOS circuits for the better understanding of the related works. Sections 2.2 and 2.3 present the related works and their issues in order to understand the contribution of this thesis clearly.

2.1 Performance Models for LSI Circuits

Understanding performance models of LSI circuits is important to review the related work. This section firstly presents performance models which are widely used to design LSI circuits. Performance variability is becoming crucial in the aggressively scaled process technologies. Characteristics of variability and their impacts on the performance of LSI circuits are presented.

2.1.1 Delay and Energy Models

The energy consumption of a circuit consists of dynamic energy $E_d$ and static energy $E_s$ as shown in (2.1) [28]. The dynamic energy $E_d$ is consumed when logic gates charge or discharge load capacitors, and $E_d$ is a quadratic function of $V_{DD}$ as shown in (2.2). The
static energy $E_s$ is consumed by the leakage current. $E_s$ is exponential to $V_{th}$, and is linear to the delay $D$ and $V_{DD}$ as shown in (2.3).

\[
E = E_d + E_s, \quad (2.1)
\]
\[
E_d = k_1 V_{DD}^2, \quad (2.2)
\]
\[
E_s = k_2 D V_{DD} \exp\left(\frac{-V_{th}}{n_i \phi_T}\right). \quad (2.3)
\]

Here, $k_1$ and $k_2$ are the coefficients determined by the process technology. $n_i$ is ideal factor of MOSFETs, which is typically between 1 and 2, and $\phi_T$ is thermal voltage which is 26 mV at a room temperature.

A linear RC delay model is a simple but accurate delay model which assumes that propagation delay of a logic gate ($D$) is proportional to load capacitance ($C$) and inversely proportional to its ON current ($I_{on}$) \cite{18, 28}:

\[
D = k_f \frac{C V_{DD}}{I_{on}}, \quad (2.4)
\]

where $k_f$ is a fitting parameter determined by a process technology. When $V_{DD}$ is in the super-threshold region ($V_{DD} \gg V_{th}$), the delay can be accurately modeled using alpha power law MOSFET model \cite{29} shown in (2.5). The value of $\alpha$ is typically between 1 and 2. When $V_{DD}$ is in the near-threshold region ($V_{DD} \sim V_{th}$) and the sub-threshold region ($V_{DD} < V_{th}$), the delay can be approximated as exponential functions of $V_{DD}$ and $V_{th}$ as shown in (2.6) and (2.7), respectively \cite{18, 30}. The parameters from $k_3$ to $k_7$ are fitting coefficients.

\[
I_{on} = k_3 W \frac{V_{DD} - V_{th}}{n_i \phi_T}^{\alpha} \quad (V_{DD} \gg V_{th}), \quad (2.5)
\]
\[
I_{on} = k_4 W \exp\left(\frac{k_5 V_{DD} - V_{th}}{n_i \phi_T} + k_6 \frac{(V_{DD} - V_{th})^2}{n_i \phi_T}\right) \quad (V_{DD} \sim V_{th}), \quad (2.6)
\]
\[
I_{on} = k_7 W \exp\left(\frac{V_{DD} - V_{th}}{n_i \phi_T}\right) \quad (V_{DD} < V_{th}). \quad (2.7)
\]

Here, $W$ and $L$ are channel width and channel length of a transistor, respectively.

### 2.1.2 Performance Variability in CMOS Circuits

#### Characteristics of Variability and Delay Variation

Performance variability of LSI circuits is one of the most critical problems in the aggressively scaled process technologies. The parameter in MOSFETs widely fluctuates in every stage of the chip fabrications. They are generally categorized as Lot-To-Lot (L2L),
Wafer-To-Wafer (W2W), Die-To-Die (D2D), and Within-Die (WID) variations. Among them, the WID variation, which impacts the circuit performance, are getting more and more critical in the aggressively scaled process technologies. WID variations stem from several types of mismatches including:

- Random Dopant Fluctuation (RDF) caused by dopant density variations in the channel area [31],
- Metal Grain Granularity (MGG) resulting from the mismatch of work functions in metal gates [32], and
- Line Edge Roughness (LER) stemming from the atomic-level fluctuations of gate edges [32].

The Pelgrom model is a simple but accurate model to consider the variation impacts due to random WID variations [19]. In the Pelgrom model, threshold voltages ($V_{th}$) follows the Gaussian distribution and its standard deviation is inversely proportional to the square root of a channel area of a MOSFET:

$$V_{th} = A_{vt} \frac{1}{\sqrt{WL}}.$$  

Here, $A_{vt}$ is a fitting coefficient determined by a process technology. As (2.8) shows, $V_{th}$ significantly fluctuates in aggressively scaled process technologies. $V_{th}$ fluctuations cause delay variations of CMOS circuits according to (2.4), which may cause timing violations in the circuits. Since $\alpha$ in (2.5) is between 1 and 2, the circuit delay almost linearly fluctuates in accordance with the $V_{th}$ variation. On the other hand, the circuit delay is exponentially proportional to $V_{th}$ in (2.6) and (2.7), which makes it essential to design variability-aware circuits for the low voltage operation.

Statistical Static Timing Analysis (SSTA) is a popular solution for variability-aware timing analysis of digital circuits [33–35]. A typical approach of the SSTA is to perform SUM and MAX operations for the delay probability density functions (PDFs) of individual gates in the target circuit. Figure 2.1 (a) shows a concept of the SUM operation. Each logic gate has a delay variation (i.e., PDF) caused by the parameter variations in MOSFETs. The SUM operation gives designers a delay PDF of serially connected logic gates. In the same way, Fig. 2.1 (b) is a concept of the MAX operation which gives designers a delay PDF of parallelized logic gates. Generally, a delay PDF of LSI circuits can be obtained by iteratively performing SUM and MAX operations.
Chapter 2. Literature Review

Figure 2.1: Atomic operations in SSTA. (a) SUM operation. (b) MAX operation.

Figure 2.2: The definition of $k\sigma$ worst case delay.

**Worst Case Delay and Timing Yield**

For discussing the statistical timing, understanding the concept of timing yield is indispensable. The timing yield is defined as the probability that the critical path delay of the circuit is no more than a specific delay $x_k$. If the $x_k$ corresponds to a delay in the $k\sigma$ worst case condition, the timing yield for the $x_k$ in a normal distribution (i.e., the Gaussian distribution) can be calculated using the Cumulative Distribution Function (CDF) of the circuit delay in the following way:

$$\int_{-\infty}^{x_k} f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{k} \exp\left(-\frac{x^2}{2}\right)dx = \Phi(k).$$

(2.9)

where $\Phi(x)$ is a CDF of a standard normal distribution. $f(x)$ is a PDF of a critical path delay of the target circuit. The definition of the $k\sigma$ worst case delay is summarized in Fig. 2.2. The area of shaded part in Fig. 2.2 is $\Phi(k)$, and it corresponds to the timing yield for the $k\sigma$ worst case delay. Since it is required to keep a timing yield to a specific value in the circuit designs, accurate evaluation for the timing yield and the worst case delay is
essential for variability-aware circuit designs.

2.2 Techniques toward Energy-Efficient Circuit Design

Today, LSI circuits require a hierarchical design with multiple levels of abstractions. There are a number of techniques to improve performance and energy-efficiency of LSI circuits at each level. This section shows the several design techniques related to the low-voltage operation.

2.2.1 Variability-Aware Architectural Design

In this thesis, we refer to circuit designs such as logic depth tuning, parallelization tuning and gate sizing in each logic gate as the architectural design. Since decisions taken in the architectural design stage have large impacts on the performance, energy consumption and power dissipation, techniques for the optimum architecture have been extensively studied over several decades. The most practical techniques are parallelism and pipelining which are widely used in commercial processors. Figure 2.3 (a) is a normal architecture of LSI circuits. Typically, logic circuits (“A” and “B”) are inserted between registers. Figure 2.3 (b) shows a concept of parallelism. The parallelized logic circuits double the throughput of the target circuit at the cost of area. In the same way, Fig. 2.3 (c) shows a concept of pipelining. Critical path delays are reduced by inserted registers, which improve the maximum operating frequency (Fmax) of the circuit. Ref. [36] pointed out that combining these techniques and scaling down the supply voltage can reduce the energy consumption to about 20% at the best case without degrading throughput.

As described in Chapter 1, considering performance variations is essential in aggressively scaled process technologies. Therefore, techniques estimating the variation impacts on parallelism and pipelining have been extensively studied. Early studies that consider variability in architecture design were by Bowman et al. [33–35], which presented a statistical predictive model for the distribution of Fmax for a chip in the presence of process variations. The model provides insight into the impact of different components of variations on the distribution of Fmax. The WID delay distribution depends on the total number of independent critical paths (Ncp) for the entire chip. For a larger number of critical paths, the mean value of the maximum critical path delay increases. As the number of critical paths increases, the probability that one of them will be strongly affected by process variations becomes higher, and therefore, the mean of critical path delay also increases. On the other hand, the standard deviation (or delay spread) decreases with larger
$N_{cp}$. Another factor that affects the delay distribution is the logic depth per critical path. Random WID variations have an averaging effect on the overall critical path distribution, which reduces the relative delay variation $\frac{\sigma}{\mu}$ [34]. If the mean and the standard deviation of a single gate are $\mu$ and $\sigma$, respectively, the standard deviation of a path comprising the serially connected identical $n$ gates is $\sqrt{n}\sigma$. Since the mean delay of the path is $n\mu$, the relative delay variation is proportional to $\frac{1}{\sqrt{n}}$ [35].

Although these simple rules give designers intuitions toward LSI circuits with high energy-efficiency and a computational power, the situation becomes different in the low-voltage operation. As pointed out in Chapter 1, the delay distributions in the sub-/near-threshold voltage operation do not follow the Gaussian distribution, but follow lognormal distribution, which make statistical timing analysis more complicated. Several previous techniques present efficient statistical timing analysis approaches which can accurately predict Non-Gaussian delay distributions from realistic nonlinear gate and interconnect delay models [37–39]. All of those techniques are aiming at accurately reflecting the Non-Gaussianity when performing the two atomic operations of SSTA, SUM and MAX. None of them explicitly discusses the averaging effect or the effect of parallel critical paths on lognormal delay distributions at this point. In Chapter 3, this thesis proves several theorems that provide architectural insight into the impact of the lognormal timing variation on the performance of a target circuit. Based on the theorems, architectural-level design strategies such as gate upsizing, pipelining and memory readout structures are discussed in order to improve the operating speed of low voltage circuits.

Figure 2.3: (a) Normal architecture. (b) Parallelism. (c) Pipelining.
2.2. Techniques toward Energy-Efficient Circuit Design

2.2.2 On-Chip Memory for Low Voltage Operation

As described in Chapter 1, conventional 6T SRAM macros can no longer successfully operate in the aggressively voltage scaled operation. Therefore, the voltage scalability may be limited by on-chip memories. To realize the sub-/near-threshold voltage operation, several circuit-level approaches are extensively studied for low-voltage on-chip memories.

Low Voltage SRAM

One simple method to solve this problem is to increase the transistor size in 6T SRAM macros to cancel out the random variations in transistor performance. However, Refs. [40, 41] pointed out that 6T SRAM bit cells must be sized up by +400% in a 65-nm process technology in order to guarantee the stable operations, which is not an acceptable area overhead for the LSI circuits design. To solve this issue, designers have extensively studied specially designed SRAM structure and lowering its operating voltage to sub-/near-threshold region with keeping the stability of the operation. There is a large number of previous papers that present different bit cell topologies, appropriate transistor sizing, multiple voltage design, novel sensing schemes, peripheral assist circuits and statistical CAD methodologies for improving the readability and the writability of SRAM [41–43]. The specially designed SRAM macros typically require a full-custom circuit design to guarantee their operation, thus designing SRAM macros is associated with a considerable design effort.

Standard-Cell Memory (SCM)

Standard-Cell Memories (SCMs) have been studied over the past decade [1, 15, 24–27]. Since a standard-cell latch is used for a bit cell and only standard-cells are used in SCMs, their operating voltage can be scaled down to the same level as the common logic circuits, which leads to a drastic improvement of overall energy efficiency. It is also pointed out that their custom design effort can be reduced to the level of fully automated cell based design. Therefore, from the viewpoint of design effort, SCMs are a good choice compared with low-voltage SRAMs using full-custom techniques.

The SCM also achieves less energy consumption than the low-voltage SRAM structure even if the same supply voltage is applied to both memories. This is because the low-voltage SRAM structure still utilizes bit-line-based structure to improve the area-efficiency. Since all the bit cells on the same column share one bit line, the capacitance on the bit line is considerably large. Therefore, the low-voltage SRAM structure wastes the dynamic energy at the bit line in readout and write operations. On the other hand, we
can easily customize the readout and write structure of the SCM to reduce the dynamic energy consumption. For example, the dynamic energy consumption of the SCM can be effectively reduced by implementing a multiplexer tree to the readout circuit. The extreme example is to implement a 2-to-1 multiplexer tree to the readout circuit. The readout multiplexer tree effectively reduces the effective capacitance on a signal path in comparison with the simple bit-line-based structure, which leads to the reduction of the readout energy consumption. Note that a similar approach can be applied to the SRAM by splitting the bit line while aggressive splitting considerably degrades the operating speed of SRAMs. Like a multiplexer tree in SCMs, transmission gates or pass transistors are implemented to split the bit line, and the split bit lines are parallelly interconnected. Only the accessed bit line is precharged/discharged through transmission gates or pass transistors in order to perform the readout/write operation, which leads to the energy reduction at the cost of increasing the area. The extreme example is to split all the bit cells on the bit line. However, the critical path of aggressively split SRAMs consists of a number of logic gates without the signal repeating functions (i.e., transmission gates or pass transistors). As a result, the critical path delay of SRAMs considerably degrades if we aggressively split the bit line. Therefore, typical SRAMs still utilize bit-line-based structures without aggressive bit-line splitting in terms of delay and area. Summarizing the above, SCMs are a good choice for the energy-efficient memory for mobile computing applications. However, the total area of SCM is still several times larger than that of full-custom low-voltage SRAM macros. As described in Section 1.3, the area overhead directly leads to the loss of computational power. Therefore, area-efficient SCMs are required. To solve this problem, this thesis proposes a novel area-efficient SCM structure in Chapter 4. By specially designed area-efficient standard-cells, the area-overhead of SCMs is effectively reduced.

2.2.3 Transistor Sizing

Channel width ($W$) is a major tuning knob in a cell-level circuit design since $W$ has a large impact on energy-efficiency and performance of logic gates. Early studies related to gate sizing methodologies were by Linholm et al. [44, 45], which presented the optimum gate size of a buffer. A buffer is typically inserted to a node with large capacitance in order to optimize propagation delay. In high-speed applications, the optimum gate size is determined in the basis of the tapered topology, where the gate size of each inverter is a constant multiple of the previous one. The tapered topology is generalized to the theory of logical effort [46].

In sub-/near-threshold voltage, the delay characteristics of logic gates are significantly
different from those in the super-threshold voltage. Thus several works applied alternative delay models for sub-/near-threshold voltage operation. Keane et al. present sub-threshold logical effort and design optimization in complex gates with stacks of transistors. They present the optimal gate width of stacked transistors which maximizes drive current in a closed-form expression [47]. Lin et al. present an improved logical effort model for multiple voltage regimes [48] using a transregional model with high accuracy in sub-/near-threshold voltage [18]. Based on their proposed logical effort model, an accurate sizing of the transistors in a stack is discussed, and an optimization framework for digital circuits operating in both near- and super-threshold regions is presented. Liu et al. present an analytical framework for variability-aware sub-threshold cell sizing. The framework renders cells that have a narrower delay distribution as well as better active drive current [49]. Ref. [50] presents a variability-aware logical effort model considering a delay correlation between slew rates of adjacent logic gates. They point out that the correlation becomes significant and have a large impact on performance in near-threshold circuits. Once the propagation delay of a gate increases due to process variation, for example, the delay at the next stage also increases through a correlation of slew rates. Based on the model, they designed several buffer circuits with both higher energy efficiency and less worst case delay than those achieved based on the conventional method of logical effort.

Although a number of previous works proposed gate sizing techniques to optimize the propagation delay of LSI circuits. None of them presents that gate sizing in low voltage circuits is much stronger tuning knob to cancel out the random WID variability than that in nominal voltage circuits. This thesis analytically shows that gate sizing is a strong tuning knob to mitigate the performance variation in the voltage-scaled circuits. In Chapter 3, we analytically show that gate upsizing brings exponential delay improvement in the sub-/near-threshold voltage operation compared with that in the nominal voltage operation. The results imply that the previous works related to gate sizing play an important role in improving the performance of low voltage circuits in the architectural-level circuit design.

2.3 Post-Silicon Tuning for Energy-Efficiency Improvement

There are a number of techniques that tune performance of LSI circuits after the chip is fabricated. This section presents several dynamic performance tuning techniques related to this thesis. Note that techniques presented in the previous section and this section are
orthogonal, which indicates that both the techniques can be combined together.

### 2.3.1 Dynamic Voltage and Frequency Scaling

Typical processors handle time varying applications with different workload sizes. As shown in (2.2), the dynamic energy consumption can be quadratically reduced by scaling the supply voltage. Therefore, in such systems, designers can reduce a large amount of dynamic energy by (1) dynamically reducing the clock frequency to the minimum value that meets the application deadline, and then (2) scaling down the supply voltage. This Dynamic Voltage and Frequency Scaling (DVFS) is presented in 2000 as a promising system-level approach to effectively reduce energy and power dissipation [51]. DVFS is now regarded as a core technique for low-power and energy-efficient LSI circuits, and is implemented into a number of commercial processors [52, 53]. A key technique in DVFS is to find the optimum supply voltage that minimizes energy consumption without violating the application deadline. Based on the highly abstracted performance modeling of LSI circuits, a number of system-level task scheduling algorithms are thus proposed [54, 55].

As the supply voltage is aggressively scaled down, the propagation delay $D$ exponentially increases as shown in (2.6) and (2.7), which makes $E_s$ in (2.3) a dominant source of the total energy consumption. Therefore, DVFS is no longer effective to reduce the total energy consumption of low-voltage circuits in terms of energy-efficiency. Since the leakage current is a strong function of $V_{th}$, scaling of not only $V_{DD}$ but also $V_{th}$ is also an important method to reduce total energy consumption of LSI circuits.

### 2.3.2 Adaptive Body Biasing

Adaptive Body Biasing (ABB) is widely used to reduce the static energy consumption ($E_s$) [56, 57]. By tuning a transistor’s back-gate voltage, ABB enables designers to dynamically tune the threshold voltage ($V_{th}$). ABB can thus effectively reduce the static energy consumption by increasing $V_{th}$ at the cost of operating speed. The threshold voltage tuning by ABB can be modeled as follows:

$$V_{th} = V_{th0} + \gamma(\sqrt{\phi_s + V_{bb}} - \sqrt{\phi_s}).$$

(2.10)

Here, $V_{th0}$ is a threshold voltage when the back-gate voltage and the source voltage of the MOSFET are same. $\phi_s$ is the surface potential. $V_{bb}$ is $V_s - V_b$, where $V_s$ and $V_b$ are the source voltage and the back-gate voltage of the MOSFET, respectively. Historically,
2.4 Summary

the scalability of the back-gate voltage is limited by junctions between dopant regions and a substrate in planar bulk MOSFETs. Recently, a novel device called FD-SOI (Fully Depleted Silicon On Insulator) has drastically increased the scalability of $V_{BB}$ [9, 10], which makes the threshold voltage tuning a popular technique as well as DVFS.

2.3.3 Minimum Energy Point Tracking using Combined Supply and Threshold Voltage Scaling

The discussion in Subsection 2.3.1 implies that techniques to reduce both the dynamic energy consumption and the static energy consumption are required in order to achieve energy-efficient LSIs. Over the past 15 years, methods to reduce the energy consumption by simultaneously tuning $V_{DD}$ and $V_{th}$ under various dynamic workloads have been widely studied [58–60]. As described in Subsections 2.3.1 and 2.3.2, DVFS and ABB have different impacts on the total energy consumption while both the techniques also change the operating speed of LSI circuits. As a result, there are a number of pairs of $V_{DD}$ and $V_{th}$ that achieve the same operating speed but have different energy consumption. Among the pairs, there exists a pair that minimizes the total energy consumption. This thesis refers to the optimum pair as Minimum Energy Point (MEP). Ref. [22] pointed out that MEPs depend on not only the circuit architecture but also a number of parameters such as an activity factor, chip temperature and process variations. Therefore, it is not trivial to find MEPs of even a simple inverter chain.

Based on the performance models of CMOS circuits presented in this chapter, this thesis proposes a simple necessary and sufficient condition for the minimum energy point operation of a circuit in Chapter 5. The condition enables designers to dynamically find MEPs. Based on the condition, strategies to dynamically find MEPs over a wide performance range are also discussed.

2.4 Summary

In Section 2.1, performance modeling of LSI circuits is presented. Sections 2.2 and 2.3 show several conventional techniques for energy-efficient LSI circuits and their issues. From Chapter 3, this thesis proposes several techniques to address the issues.
Chapter 3

Statistical Timing Modeling for Voltage-Scaled Circuit Design

3.1 Introduction

With the aggressive transistor scaling, the performance variation of LSI circuits is no longer negligible in a circuit design. Variability-aware statistical timing models are thus required. The first goal of this chapter is to develop delay variation models in both the nominal voltage operation and the low voltage operation. More specifically, this thesis reveals how the delay variations of critical paths change according to the architectural-level circuit optimization. The architectural-level optimization includes the following parameters:

- Logic depth \((n)\),
- Degree of parallelism \((\rho)\), and
- Transistor size \((W \text{ and } L)\).

After that, based on the models, this thesis discusses architectural-level circuit design strategies for each operating voltage.

Chapter 3 is organized in the following way. Section 3.2 describes preliminaries. Section 3.3 explains fundamental characteristics of sub-/near-threshold voltage operation. Several theorems and their proofs are also presented in Section 3.3. Section 3.4 shows experimental results which demonstrate that the theorems proved in Section 3.3 hold for the actual CMOS circuits. Section 3.5 discusses how the theorems and corollaries presented in this paper can be effectively exploited in an architectural design phase. In Section 3.6,
several application examples are examined using theorems proved in Section 3.3. Section 3.7 concludes this chapter.

3.2 Preliminaries

3.2.1 Properties of Lognormal Distribution

Since a lognormal distribution is a key to understand the delay variations of the voltage-scaled LSI circuits, this subsection presents basic properties which hold for lognormal distributions. It is well known that if an Random Variable (RV) \( X \) has the Gaussian distribution \( N(\mu, \sigma^2) \), its PDF \( f(x) \) is represented as

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).
\]

(3.1)

Suppose we have an RV \( Y \) which is represented as \( Y = \exp(X) \). Then \( Y \) is a lognormal distribution function \( LN(\mu, \sigma^2) \). Its PDF \( g(y) \) is represented as

\[
g(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right).
\]

(3.2)

Note that \( \mu \) and \( \sigma^2 \) in (3.2) do not correspond to a mean and a variance of \( Y \), respectively. They correspond to a mean and a variance of \( \ln(Y) \). The shape of \( g(y) \) is asymmetric unlike the Gaussian distribution. The CDF of a lognormally distributed RV \( Y \) can be formulated using \( \Phi(x) \) in (2.9) as

\[
\text{(CDF of } Y) = \Phi\left(\ln\frac{y - \mu}{\sigma}\right).
\]

(3.3)

3.2.2 Delay Distribution for the Super-Threshold Voltage Operation

The alpha power law model [29] is commonly used for representing MOSFET’s ON current which is shown in (2.5). This model is accurate for the super-threshold voltage where the supply voltage is sufficiently higher than the threshold voltage. It is well known that the distribution of single gate delay in the super-threshold voltage operation follows the Gaussian distribution. This is because a logic gate delay in (2.4) for the super-threshold voltage operation is almost proportional to the threshold voltage \( V_{th} \), and \( V_{th} \) follows the Gaussian distribution by the Pelgrom’s law [19]. This makes it simpler to perform the SUM and the MAX operations for estimating the timing yield of the targeting circuit. A
number of papers thus presented statistical predictive models for CMOS circuits in the super-threshold voltage operations [33–35, 37–39].

3.2.3 Delay Distribution for the Near-Threshold Voltage Operation

If the supply voltage is scaled down to the near-threshold voltage, the situation becomes different. Using (2.4) and (2.6), the propagation delay $D$ of a logic gate can be expressed in the following way.

$$D = k_f C \frac{V_{DD}}{I_{on}} = \alpha \exp \left( \frac{\Delta V_{th}}{V_0} - k_0 \left( \frac{\Delta V_{th}}{n_{\phi T}} \right)^2 \right),$$

(3.4)

where $\Delta V_{th} = V_{th} - V_{th0}$ which represents a threshold voltage deviation induced by the random WID variations. $V_{th0}$ is the threshold voltage when there is no random WID variations. $\alpha$ is a constant, and $V_0$ is $\left( \frac{k_5}{n_{\phi T}} + 2k_6 \frac{V_{dd} - V_{sat}}{(n_{\phi T})^2} \right)^{-1}$. Since $\Delta V_{th}$ has an exponential impact on $D$, $D$ does not simply follow the Gaussian distribution. The following lemma indicates that $D$ follows the lognormal distribution if circuits operate in the near-threshold voltage.

**Lemma**

If $k_6 \left( \frac{\Delta V_{th}}{n_{\phi T}} \right)^2$ is sufficiently smaller than $\frac{\Delta V_{th}}{V_0}$ meaning that $k_6 \left( \frac{\Delta V_{th}}{n_{\phi T}} \right)^2$ can be ignored in comparison with $\frac{\Delta V_{th}}{V_0}$, $D$ can be exactly fit to a lognormal distribution function.

(pf.) From the assumption, $k_6 \left( \frac{\Delta V_{th}}{n_{\phi T}} \right)^2$ in (3.4) can be ignored in comparison with $\frac{\Delta V_{th}}{V_0}$. Therefore, we obtain

$$D \sim \alpha \exp \left( \frac{\Delta V_{th}}{V_0} \right) = D_0 \exp \left( \frac{V_{th}}{V_0} \right),$$

(3.5)

where we rename $D_0 = \alpha \exp \left( \frac{-V_{sat}}{V_0} \right)$ for simplicity. This work assumes that the threshold voltage ($V_{th}$) is only a variable that represents all impacts of device parameter variations on delay variation. It also assumes that the $V_{th}$ is a normally distributed RV with mean $\mu'$ and variance $\sigma'$ (i.e., $N(\mu', \sigma'^2)$). The PDF $f(v)$ of the $V_{th}$ is given as

$$f(v)dv = \frac{1}{\sqrt{2\pi}\sigma'} \exp \left( -\frac{(v - \mu')^2}{2\sigma'^2} \right) dv.$$  

(3.6)
Since $dV_{th} = V_0 \frac{dD}{D}$, delay’s PDF $g(d)$ is obtained as

$$g(d)dd = \frac{V_0}{d \sqrt{2\pi \sigma^2}} \exp\left(-\frac{(\ln(d) - \frac{\mu'}{V_0})^2}{2\left(\frac{\sigma'}{V_0}\right)^2}\right)dd,$$

(3.7)

which means that a variable $\frac{D}{D_0}$ has a lognormal distribution $LN\left(\frac{\mu'}{V_0}, \left(\frac{\sigma'}{V_0}\right)^2\right)$. (q.e.d.)

If we rename $\frac{\mu'}{V_0}$ to $\mu$ and $\frac{\sigma'}{V_0}$ to $\sigma$, the variable $D$ can be represented as a lognormal distribution $LN(\mu, \sigma^2)$.

### 3.2.4 Delay Distribution for the Sub-Threshold Voltage Operation

Like the near-threshold voltage operation, the delay distribution becomes a lognormal distribution when we scale down the supply voltage below the threshold voltage. From (2.4) and (2.7), the propagation delay of a logic gate can be modeled in the following way:

$$D = k_t \frac{LCV_{DD}}{W} \exp\left(-\frac{V_{DD} - V_{th}}{n \phi T}\right) = D'_0 \exp\left(\frac{V_{th}}{V_0'}\right),$$

(3.8)

where $D'_0 = k_t \frac{LCV_{DD}}{kT} \exp\left(-\frac{V_{th}}{n \phi T}\right)$ and $V_0' = n \phi T$. Therefore, from the discussions in Subsection 3.2.1, the propagation delay of a logic gate also follows the lognormal distribution $LN\left(\frac{\mu'}{V_0'}, \left(\frac{\sigma'}{V_0}\right)^2\right)$.

### 3.3 Lognormal Timing Model

From the discussions in Section 3.2, the propagation delay follows the Gaussian distribution and the lognormal distribution in the super-threshold voltage and the sub-/near-threshold voltage operation, respectively. Thus the focus of this chapter is shifted to finding how the architectural design strategies change between the Gaussian delay distribution (i.e., the super-threshold voltage operation) and the lognormal delay distribution (i.e., the sub-/near-threshold voltage operation).

As described in Subsection 3.2.2, the delay distributions of the circuits follow the Gaussian distribution in the super-threshold voltage operation, which makes easy to perform two atomic operations, SUM and MAX operations. To highlight the difference of these statistical operations between the Gaussian distribution and the lognormal distribution, this section firstly derives SUM and MAX operations for the lognormal delay distribution, which can more accurately represent the delay in the sub-/near-threshold voltage operation than the Gaussian delay distribution. After that, this section derives several
properties from these two SUM and MAX operations, which imply that the architectural-level design strategies are different depending on the operating voltage.

### 3.3.1 SUM Operation for Lognormal Distributions

Generally, the distribution of the sum of lognormal RVs does not have the closed form. However, the sum of lognormal RVs can be reasonably approximated as a lognormal RV [61]. Let $L$ be the sum of $n$ correlated lognormal RVs ($L_1, L_2, ..., L_n$),

\[
L = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} \exp(X_i) \sim \exp(Z),
\]

where $X_i$ and $Z$ are normally distributed correlated RVs as follows:

\[
X_i \sim N(\mu_i, \sigma_i^2), \quad Z \sim N(\mu_Z, \sigma_Z^2).
\]

Here, $\mu_i$ and $\sigma_i$ are the mean and the standard deviation of $X_i$, respectively. $\mu_Z$ and $\sigma_Z$ are the mean and the standard deviation of $Z$, respectively. The correlation coefficient of $X_i$ and $X_j$ is defined as

\[
r_{ij} = \frac{\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]}{\sigma_i \sigma_j}.
\]

Note that if $X_i$ and $X_j$ are independent, then $r_{ij} = 0$.

As a simple approximation method, this work uses Wilkinson’s method [61], which fits only the first two moments of the lognormal’s sum to those of another lognormal distribution function. The first two moments, $u_1$ and $u_2$, are as follows:

\[
\begin{align*}
\quad u_1 &= \mathbb{E}[L] = \mathbb{E}[\exp(Z)] = \sum_{i=1}^{n} \exp\left(\mu_i + \frac{\sigma_i^2}{2}\right), \\
\quad u_2 &= \mathbb{E}[L^2] = \mathbb{E}[\exp(2Z)] \\
&= \sum_{i=1}^{n} \exp\left(2\mu_i + 2\sigma_i^2\right) \\
&\quad + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \exp(\mu_i + \mu_j) \cdot \exp\left(\frac{1}{2} (\sigma_i^2 + \sigma_j^2 + 2r_{ij}\sigma_i\sigma_j)\right).
\end{align*}
\]
From (3.12) and (3.13), $\mu_Z$ and $\sigma_Z$ are derived as

$$\mu_Z = 2 \ln u_1 - \frac{1}{2} \ln u_2,$$

(3.14)

$$\sigma_Z^2 = \ln u_2 - 2 \ln u_1.$$  

(3.15)

Note that $\mu_Z$ and $\sigma_Z$ correspond to the $\mu$ and $\sigma$ of $L$ defined in (3.9), respectively. From (3.14) and (3.15), the following theorems can be derived.

**Theorem 1**

Let $L_1, L_2, \ldots, L_n$ be $n$ identical lognormally distributed RVs $\text{LN}(\mu, \sigma^2)$ and $\sigma \ll 1$ meaning that $\sigma$’s quadratic terms are negligible compared with the other terms, then $\mu_Z$ and $\sigma_Z$ can be represented as $\mu + \ln n$ and $\frac{\sigma}{\sqrt{n}} \sqrt{1 + \frac{2}{n} \sum_i \sum_j r_{ij}}$, respectively.

(pf.) $\mu_Z$ and $\sigma_Z$ can be derived from (3.14) and (3.15) as follows:

$$\mu_Z = \mu + \ln n + \frac{\sigma^2}{2} - \frac{1}{2} \ln \left( \exp\left(\frac{\sigma^2}{2}\right) + \frac{2}{n} \sum_i \sum_j \exp\left(r_{ij} \sigma^2\right) \right),$$

(3.16)

$$\sigma_Z = \sqrt{\ln \left( \frac{\exp\left(\sigma^2\right) - 1}{n} + \frac{2}{n} \sum_i \sum_j \exp\left(r_{ij} \sigma^2\right) + 1 \right)}.$$  

(3.17)

Since $\sigma \ll 1$, $\mu_Z$ and $\sigma_Z$ can be represented as follows by ignoring the $\sigma$’s quadratic terms:

$$\mu_Z \sim \mu + \ln n,$$

(3.18)

$$\sigma_Z \sim \frac{\sigma}{\sqrt{n}} \sqrt{1 + \frac{2}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r_{ij}}.$$  

(3.19)

(q.e.d.)

Theorem 1 indicates that the median ($\exp(\mu_Z)$) is proportional to the number of RVs summed up together ($n$):

$$(\text{Median of } L) \sim \exp(\mu + \ln n) = n \exp(\mu).$$  

(3.20)

Since $r_{ij}$ ranges from 0 to 1, $\sigma_Z$ ranges from $\frac{\sigma}{\sqrt{n}}$ to $\sigma$, which indicates that the parameter $\sigma$ decreases as the number of RVs summed up together ($n$) increases. Specifically, if $X_i$ and $X_j$ are mutually independent (i.e., $r_{ij} = 0$), then $\sigma_Z = \frac{\sigma}{\sqrt{n}}$. In this case the $L$’s $k\sigma$
3.3. Lognormal Timing Model

worst case can be expressed as follows:

\[(k\sigma \text{ worst case of } L) \sim \exp\left(\mu + \ln n + \frac{k\sigma}{\sqrt{n}}\right) = n \exp\left(\mu + \frac{k\sigma}{\sqrt{n}}\right). \quad (3.21)\]

Based on a model fitting result for our target process technology, if \(\sigma = 0.2, \mu = -20\) and \(r_{ij} = 0\), which roughly correspond to the propagation delay of a logic gate in the near-threshold region, approximation errors for \(\mu_Z\) and \(\sigma_Z\) introduced by ignoring the second order of \(\sigma\) in (3.16) and (3.17) are less than 0.1% and 1%, respectively, for \(n = 4\). Therefore, the assumption stated in Theorem1 (i.e., the second order of \(\sigma\) is negligible) is feasible.

3.3.2 Averaging Effect of Paths

In super-threshold voltage operation where the operating voltage is sufficiently higher than the threshold voltage, random WID variations have an averaging effect, which reduces the relative delay variation \(\frac{\sigma}{\mu}\) of a path as the number of logic gates chained in series increases. If the path consists of \(n\) identical gates chained in series, the variance of the path delay is \(n\sigma^2\). Since the mean delay of the path is \(n\mu\), the relative delay variation is proportional to \(\frac{1}{\sqrt{n}}\) [35]. Therefore, the relative delay variation decreases as \(n\) increases. In sub-/near-threshold voltage operation where the operating voltage is aggressively scaled down, a stronger averaging effect is observed than that observed in the super-threshold voltage operation. This means that the relative delay variation of an \(n\)-stage path is proportional to a value which is less than \(\frac{1}{\sqrt{n}}\).

Corollary1 (Averaging Effect)

For a given RV \(L\) which is the sum of \(n\) independent lognormal RVs, let \(x_{k\sigma,n}\) be the \(L\)'s \(k\sigma\) worst case as shown in Figure 3.1 and let variation \(v_n\) be defined as \(x_{k\sigma,n} - x_{0\sigma,n}\) which indicates the difference between the \(k\sigma\) worst case and the median of \(L\). Then the ratio of \(v_n\) to \(v_1\), which we refer to as averaging effect ratio is represented as:

\[(\text{Averaging effect ratio}) = \frac{v_n}{v_1} = n \cdot \frac{\exp\left(\frac{k\sigma}{\sqrt{n}}\right) - 1}{\exp(k\sigma) - 1}. \quad (3.22)\]

(pf.) Immediate from (3.20) and (3.21). (q.e.d.)

The averaging effect ratio represents the magnitude of the delay variation. If the averaging effect is stronger, the magnitude of the delay variation gets smaller. Therefore, if
the averaging effect is strong, the averaging effect ratio is small. Since, in the Gaussian distribution, the median is equal to the mean and the distance between the median and the $k\sigma$ worst case is $k\sigma$, the averaging effect ratio in the normal distribution is $\sqrt{n}$. Corollary 1 shows that the averaging effect ratio in lognormal delay distributions is smaller than that in Gaussian delay distributions (i.e., $\sqrt{n}$). In an actual circuit, the median corresponds to the delay in a TT condition where both the threshold voltages of pMOS and nMOS are typical values. The $\sigma$ in (3.22) is a parameter which is proportional to a standard deviation of $V_{th}$. Typical values for the $\sigma$ in the latest process technologies are ranging from 0.2 to 0.6. For example, the averaging effect ratio of a 16-stage logic path in a circuit operated with the near-threshold voltage operation ($V_{DD} = 0.4$ V) is 1.62 if $\sigma = 0.4$ and $k = 5$ while that in a circuit operated with the super-threshold voltage ($V_{DD} = 1.0$ V) is 4. This intuitively means that the performance degradation along with an increase of the logic depth in the low voltage operation is slower than that in the super-threshold voltage operation.

### 3.3.3 Delay Distribution of Parallel Paths

The impact of the number of independent parallel paths ($N_{cp}$, denoted $p$ for simplicity) on the maximum critical path delay is obtained by performing a MAX operation for the paths. It is well known that the maximum operating frequency ($F_{max}$) of a sequential circuit needs to be lowered as the number of critical paths increases to maintain the same timing yield [34]. In this thesis, we refer to the speed of the $F_{max}$ degradation as an $F_{max}$ degradation speed. This key word is used in the following theorem.
Theorem 2

Suppose we have a lognormally distributed RV $L_{\log}$ and a normally distributed RV $L_{\text{norm}}$. Let $\max_i(L)$ be defined as a result of the MAX operation for $i$ identical independent RVs $L$. Then the Fmax degradation speed along with the increase of the number of RVs $p$ for $\max_p(L_{\log})$ is exponential to that for $\max_p(L_{\text{norm}})$.

(pf.) Let $f(x)$ be a PDF for either normally or lognormally distributed RVs. The MAX of $p$ parallelly connected $f(x)$ is

$$G(x) = \left(\int_{-\infty}^{x} f(x)dx\right)^p = \left(1 - \int_{x}^{\infty} f(x)dx\right)^p.$$  \hfill (3.23)

For a large $x$ in (3.23), which means $\int_{x}^{\infty} f(x)dx$ is small, $G(x)$ can be approximated as

$$G(x) \sim 1 - p \int_{x}^{\infty} f(x)dx = 1 - p + pF(x),$$  \hfill (3.24)

where $F(x)$ is a CDF of $f(x)$.

Let $y_0$ and $d$ be defined as a timing yield and a target delay which satisfies the timing yield $y_0$, respectively. To keep this timing yield $y_0$ constant, the value of (3.24) should be kept at $y_0$ for different $p$ values. If $f(x)$ is a lognormal distribution function $LN(\mu, \sigma^2)$, with (3.3), the target delay $(d_{\log})$ for satisfying the yield $y_0$ is

$$d_{\log} = \exp\left(\mu + \sigma\Phi^{-1}\left(\frac{y_0 - 1 + p}{p}\right)\right).$$  \hfill (3.25)

In a similar way, if $f(x)$ is the Gaussian distribution function $N(\mu', \sigma'^2)$, the target delay $(d_{\text{norm}})$ for satisfying $y_0$ is

$$d_{\text{norm}} = \mu' + \sigma'\Phi^{-1}\left(\frac{y_0 - 1 + p}{p}\right).$$  \hfill (3.26)

Since the Fmax is inversely proportional to $d$, it is immediate from (3.25) and (3.26) that the Fmax degradation speed along with the increase of $p$ for $\max_p(L_{\log})$ is exponential to that for $\max_p(L_{\text{norm}})$. (q.e.d.)

Although $d_{\log}$ is exponential to $d_{\text{norm}}$ as presented in (3.25) and (3.26), the degradation speed of $d_{\log}$ along with the increase of $p$ is very slow. This property is summarized in the following corollary:
Corollary 2

The degradation of $d_{\log}$ is sublinear to the increase of $p$.

(pf.) The CDF of Gaussian distribution, $\Phi(x)$ is represented by an error function ($\text{erf}(x)$):

$$\Phi(x) = \frac{1}{2} \left( 1 + \text{erf} \left( \frac{x}{\sqrt{2}} \right) \right).$$  \hspace{1cm} (3.27)

In [62], $\text{erfc}(x) = 1 - \text{erf}(x)$ is approximated as the sum of two exponential terms for a positive number $x$:

$$\text{erfc}(x) \approx \frac{1}{6} \exp(-x^2) + \frac{1}{2} \exp(-4x^2/3).$$ \hspace{1cm} (3.28)

In this thesis, for simplicity, we further approximate (3.28) by ignoring a non-dominant term for a large $x$:

$$\text{erfc}(x) \approx \frac{1}{6} \exp(-x^2).$$ \hspace{1cm} (3.29)

The approximation (3.29) provides an inverse function of $\Phi(x)$ in a closed form:

$$\Phi^{-1}(x) \approx \sqrt{-2 \ln(12(1 - x))}.$$ \hspace{1cm} (3.30)

From (3.25) and (3.30), the degradation of $d_{\log}$ can be expressed by an elementary function as follows:

$$d_{\log} = \exp \left( \mu + \sigma \sqrt{2} \ln \left( \frac{p}{12(1 - y_0)} \right) \right).$$ \hspace{1cm} (3.31)

Since $\exp \left( \sqrt{\ln p} \right)$ is sublinear to $p^a$ for any positive number $a$ \footnote{Since $\lim_{p \to \infty} \exp \left( \sqrt{\ln p} \right)/p^a = 0$ for $a > 0$, $\exp \left( \sqrt{\ln p} \right)$ is sublinear to $p^a$.}, the degradation of $d_{\log}$ is sublinear to the increase of $p$.

(q.e.d.)

Based on a model fitting result for our target process technology, the approximation errors for a single stage buffer delay $d_{\log}$ introduced by the approximation (3.29) are less than 2.7% and 0.2% for $3\sigma$ timing yield when $p = 1$ and $p = 256$, respectively. Hence the approximation (3.29) is feasible.
3.3. Lognormal Timing Model

3.3.4 Impact of Gate Sizing on Delay Variation

A smaller transistor gate width causes a larger threshold voltage variation, resulting in the larger high-$\sigma$ worst case delay. According to (2.8), the standard deviation of threshold voltage ($V_{th}$), $\sigma_{V_{th}}$, is inversely proportional to the square root of channel area ($\sqrt{WL}$). Let us define the term of gate sizing effect on the worst case delay as a ratio of the reduction in the worst case delay to the increase in gate size. Then we obtain the following corollary by (2.8), (3.5) and (3.8).

**Corollary 3**

The gate sizing effect on the worst case delay in the sub-/near-threshold voltage operation is approximately exponential to that in the super-threshold voltage operation if load capacitance of the gate is proportional to its gate width. (pf.) Let us suppose a normally distributed RV $V_{th} \sim N(V_{th0}, \sigma_{V_{th}}^2)$, then, from (2.8), (3.3), (3.5) and (3.8), the $k\sigma$ worst case delay of a single gate in the sub-/near-threshold voltage operation $D_{log}$ is as follows:

$$D_{log} = D'_0 \exp\left(\beta \sigma_{V_{th}k}\right), \quad (3.32)$$

where $k$ is a parameter which determines the target timing yield of the gate. $D'_0$ and $\beta$ are constants. It is well known that the distribution of single gate delay $D_{norm}$ in the super-threshold voltage operation has the Gaussian distribution if $V_{th}$ is a normally distributed RV. From (2.8), the $k\sigma$ worst case delay of a single gate in the super-threshold voltage operation $D_{norm}$ can be represented:

$$D_{norm} = \beta'_1 V_{th0} + \beta'_2 k\sigma_{V_{th}} + \text{const.}, \quad (3.33)$$

where $\beta'_1$ and $\beta'_2$ are constants. From (3.32) and (3.33), it is immediate that the gate sizing effect on the worst case delay in the sub-/near-threshold voltage operation is exponential to that in the super-threshold voltage operation. (q.e.d.)

This corollary indicates that gate sizing for a logic gate operated in the sub-/near-threshold voltage exponentially affects the performance of the logic gates.
3.4 Validation with a 28-nm Process Technology Model

This section validates the properties described in the theorems and corollaries by comparing them with circuit simulation results obtained using a commercial 28-nm process technology. We perform transistor-level circuit simulation [63] using a foundry provided Monte Carlo simulation package. We use a 1.0 V as the super-threshold voltage. As a representative of a lognormal delay distribution, we use a 0.4 V as a supply voltage. The 0.4 V supply voltage corresponds to the near-threshold voltage in this process technology. We examined the delay of circuits having different logic depths and different numbers of parallel paths. $L_i$ ($i = 1, 2, ..., k$) shown in Fig. 3.2 is a buffer of $i$th logic stage in a circuit. All the buffers have an identical lognormal distribution $LN(\mu, \sigma^2)$ if they operate in the sub-/near-threshold voltage. The delay value $D_k$ is obtained from a delay of an intermediate part of a sufficiently long buffer chain as shown in Fig. 3.2.

3.4.1 Delay Distribution of a Buffer Chain

Figure 3.3 shows delay distributions of a buffer chain having different logic depths shown as $n$. Through an input slew, buffer delays are mutually correlated. If the delay of a buffer is large, the output transition time (i.e., input slew of the next buffer) of the buffer is also large, which results in an increase of an input slew rate in the next buffer. This causes an increase in the propagation delay of a buffer in the next stage. This is the mechanism of the correlation between the two consecutive buffers. For reflecting this correlation in our analytical model, we assume that the adjoined buffers are correlated from each other as follows:

$$r_{ij} = \begin{cases} 1 & (i = j) \\ r & (i = j \pm 1) \\ 0 & \text{(otherwise)} \end{cases}$$  \hspace{1cm} (3.34)

Solid lines in Fig. 3.3 are obtained using a model derived from Theorem1 by fitting the parameter values $\mu$, $\sigma$ and $r$. Dots show Monte Carlo simulation results. As can be
3.4. Validation with a 28-nm Process Technology Model

Figure 3.3: Buffer chain simulation result \((V_{DD} = 0.4 \text{ V})\). \(\mu = -21, \sigma = 0.21\) and \(r = 0.32\).

seen from Fig. 3.3, the delay distributions estimated using our analytical model are well matched with the results of Monte Carlo simulation using a commercial process technology model. It demonstrates that the delay distributions of actual multiple stage buffer chains closely fit the lognormal distribution.

### 3.4.2 Validation of Averaging Effect

We examine the averaging effect ratio in the \(4\sigma\) worst case of the \(n\)-stage buffer chain for various logic depths \(n\). Note that in order to obtain a \(3\sigma\) timing yield per chip, we have to obtain more than the \(3\sigma\) timing yield per critical path. Hence, as a representative value, we use a \(4\sigma\) to evaluate the worst case delay per critical path, which corresponds to a \(3\sigma\) timing yield per chip and there are about 40 independent critical paths in it. Figure 3.4 shows the results. The dots show the results of Monte Carlo simulation. The solid line labeled “Gaussian Model” shows the averaging effect ratio for normally distributed independent RVs. The graph labeled “Lognormal Model” shows the averaging effect ratio for lognormally distributed independent RVs. The graph for lognormal RVs can be obtained using (3.22). For \(n = 256\), a 27% smaller averaging effect ratio (i.e., stronger averaging effect) can be observed in a lognormal delay distribution compared with that in a Gaussian delay distribution.
Chapter 3. Statistical Timing Modeling for Voltage-Scaled Circuit Design

3.4.3 Validation of Fmax Degradation Speed

Since Fmax is an inverse of the worst case delay, we evaluate the worst case delay for analyzing the Fmax degradation speed. First, we examine the speed of the worst case delay degradation in a buffer chain along with the increase of a logic depth. We suppose that the buffer chain represents a critical path in a processor. Figure 3.5 shows the 4σ worst case delay for different logic depths. The number of parallel paths is one in this case. Dots show the results of Monte Carlo simulation and the solid line shows the results obtained with an analytical model presented in Theorem1. Each dot represents the 4σ
3.4. Validation with a 28-nm Process Technology Model

The Number of Critical Paths \( N_{cp} \)

The worst case delay normalized by the delay of a single stage buffer operated with a corresponding supply voltage. The results show that the degradation speeds for the worst case delay of 16-stage and 256-stage buffer chains in a lognormal delay distribution are 39% and 49% smaller than those in a Gaussian delay distribution, respectively. This demonstrates that the averaging effect in the low voltage operation is stronger than that in the super-threshold voltage operation as described in Theorem1.

Let us consider to reduce the number of pipeline stages of a processor from 4 to 2. If the number of pipeline stages is halved, the logic depth in a critical path is roughly doubled. Therefore, if the logic depth of the 4-stage processor is 32, that of the 2-stage processor is 64. According to Fig. 3.5, when the logic depth is doubled from 32 to 64, the 4\( \sigma \) worst case delay becomes 91% larger in the low voltage operation while it becomes 98% larger in the super-threshold voltage operation. Therefore, the performance degradation of a processor incurred by reducing the number of its pipeline stages in the low voltage operation is smaller than that incurred in the super-threshold voltage operation.

Next we evaluate the degradation speed of the worst case delay in parallel buffer chains in order to see the performance of a processor with many parallel critical paths. We use a circuit where the 8-stage buffer chains are connected in parallel as shown in Fig. 3.6. We consider the parallel paths as a representative of a chip for the evaluation. If the chip requires a 3\( \sigma \) timing yield, the parallel paths also require the 3\( \sigma \) timing yield. Therefore, we evaluate the 3\( \sigma \) worst case delay of the parallel paths in this experiment. The overall delay in this circuit is obtained by performing MAX operation for all critical paths. Figure 3.7 shows the worst case delays for different numbers of critical paths \( N_{cp} \). Dots show
Figure 3.7: The number of critical paths $N_{cp}$ vs. the $3\sigma$ worst case delay. The logic depth is 8.

Figure 3.8: The number of critical paths $N_{cp}$ vs. the $3\sigma$ worst case delay. The logic depth is 1.

the results of Monte Carlo simulation and the solid line shows the results obtained with an analytical model presented in Theorem2 and Corollary2. Each dot represents the $3\sigma$ worst case delay normalized by the delay of a single-path buffer chain operated with a corresponding supply voltage. The results demonstrate that the speed of the worst case delay degradation for parallel critical paths against an increase of $N_{cp}$ in the low voltage operation is faster than that in super-threshold voltage operation. However, the $3\sigma$ worst case delay in the low voltage operation is sublinear to the number of parallel critical paths, and
the degradation speed is less than 11% in both cases, which indicates the increase of the number of parallel paths has weak impact on the delay degradation. Figure 3.8 shows the worst case delays of $N_{cp}$-parallel single stage buffers. Although the 3σ worst case delay in the low voltage operation is still sublinear to the number of parallel critical paths, the degradation speed in Fig. 3.8 is faster than that in Fig. 3.7. The degradation for $N_{cp} = 64$ is increased to 26% in the sub-/near-threshold voltage operation while it is still less than 5% in the super-threshold voltage operation. This fact implies that designers should be careful in designing a circuit with short logic depths and many critical paths such as SRAM readout circuits. This is because they will suffer from delay degradations in the low voltage operation although the delay degradations in the super-threshold voltage operation are small.

### 3.4.4 Validation of Gate Sizing Effect

We evaluate the 4σ worst case delays in single-stage buffers (i.e., $D_1$ in Fig. 3.2) for different gate sizes. Figure 3.9 shows PDFs of different sized single-stage buffers whose delay distribution is lognormal and Gaussian, respectively. “X” represents the drive strength of the gate, which is proportional to the gate width. The figure shows that PDFs of buffers with a large $X$ have a smaller relative variation although their median does not move widely. This is because the typical value of transistor’s ON current and load capacitance are roughly proportional to the gate width, and the delay variation is strongly dependent on the gate width according to Pelgrom’s model (2.8). Between the lognormal delay dis-
Chapter 3. Statistical Timing Modeling for Voltage-Scaled Circuit Design

Figure 3.10: Buffer size $X$ vs. $4\sigma$ worst case delay.

Figure 3.11: The $4\sigma$ worst case delay for different gate sizes.

distribution and the Gaussian delay distribution, for the same gate size, the relative variation (i.e., $\sigma/\mu$) in the lognormal delay distribution is larger than that in the Gaussian delay distribution. Figure 3.10 shows the $4\sigma$ worst case delay of single buffers with different gate sizes. Each delay is normalized by the delay with $X = 1$. If we look at the $4\sigma$ worst case delay, a 90% delay degradation for a $0.5X$ buffer and a 46% delay improvement for a $4X$ buffer can be observed in the lognormal delay distribution compared with those in the Gaussian delay distribution, respectively. Figure 3.11 shows single buffer delays for different gate sizes. The vertical axis shows the $4\sigma$ worst case delay in the lognormal delay distribution in a logarithmic scale, and the horizontal axis shows the counterpart in the Gaussian delay distribution in a linear scale. Since dots roughly line up in a straight line, Corollary3 holds even for the commercial process technology model. Main reasons for
the nonlinearity are that a buffer’s load capacitance is not exactly proportional to the gate width and that the ON current is not linear to the gate width due to the narrow channel effect.

### 3.4.5 Validation of Properties in Other Logic Cells

We validate that the properties derived for the buffer chain hold for 2-input NAND (NAND2 in short) or 2-input NOR (NOR2 in short) with stacked transistors. In order to evaluate rise and fall delays simultaneously, we evaluate the propagation delay of two serially connected logic gates as the delay of NAND2/NOR2 as shown in Fig. 3.12.

Figure 3.13 shows averaging effect ratios for NAND2 and NOR2 chains. The line labeled “Gaussian Model” shows the averaging effect ratio for normally distributed independent RVs (i.e., $\sqrt{n}$). The other lines show the results obtained with the analytical model presented in Theorem1. Although the worst case delay of the chains increases be-
Figure 3.14: Logic depth vs. $4\sigma$ worst case delay for NAND2/NOR2 chains.

cause of stacked transistors in NAND2/NOR2, we confirmed that the stronger averaging effect in the lognormal delay distribution is still remained.

Figure 3.14 shows the $4\sigma$ worst case delay for different logic depths for NAND2/NOR2 chains. The lines show the results obtained with the analytical model presented in Theorem1. The results show that the averaging effect ratio of the NAND2/NOR2 chains in the low voltage operation is always smaller than that in the super-threshold voltage operation. Figure 3.14 shows the $4\sigma$ worst case delays with the different number of stages normalized by the delay of the single stage chain operated with the corresponding supply voltage. The normalized worst case delay in the lognormal delay distribution is smaller than that in the Gaussian delay distribution. This demonstrates that there is a stronger averaging effect in the low voltage operation than that in the super-threshold voltage operation.

Figure 3.15 shows the $3\sigma$ worst case delays for different numbers of critical paths $N_{cp}$. Each dot represents the $3\sigma$ worst case delay normalized by the delay of a single-path NAND2/NOR2 chain operated with a corresponding supply voltage. The lines show the results obtained with the analytical model presented in Theorem2 and Corollary2. Like buffer chains, the $3\sigma$ worst case delay in the lognormal delay distribution is sublinear to the number of parallel critical paths as described in Corollary2, and the increase of the delay over the delay of the single path is less than 10% in both cases.

Figure 3.16 shows the $4\sigma$ worst case delay of single stage NAND2/NOR2 gates with different gate sizes. Each delay is normalized by the delay with $X = 1$. This result demonstrates the sensitivity of the normalized $4\sigma$ worst case delay to gate size $X$ is also larger in the lognormal delay distribution than that in the Gaussian delay distribution for
3.4. Validation with a 28-nm Process Technology Model

Figure 3.15: The number of critical paths $N_{cp}$ vs. the $3\sigma$ worst case delay for NAND2/NOR2 chains.

Figure 3.16: Gate size $X$ vs. $4\sigma$ worst case delay for NAND2/NOR2 chains.

NAND2/NOR2 chains.

Figure 3.17 shows delays of single stage NAND2/NOR2 chains for different gate sizes. The vertical axis shows the delay for the low voltage operation in a logarithmic scale, and the horizontal axis shows the delay for the super-threshold voltage operation in a linear scale. Since dots roughly line up in a straight line, Corollary3 holds even for the logic gates with stacked transistors.
3.5 Summary of Properties

Table 3.1 summarizes the theorems and corollaries presented in previous sections. In this section, we show how these theorems and corollaries are mutually dependent in each voltage region.

Let us consider $p$-parallel $n$-stage buffer chains where individual buffers have the same gate size $X$ from each other as shown in Fig. 3.18. We suppose to keep the timing yield $y_0$ constant. A buffer whose drive strength $X = 1$ is supposed to have a delay distribution $LN(\mu, \sigma^2)$ in the sub-/near-threshold voltage. From (3.32) in Corollary3, we can derive each buffer in chains has a delay distribution $LN\left(\mu, \left(\sigma / \sqrt{X}\right)^2\right)$. Then, from Theorem1,
3.5. Summary of Properties

Worst Case Delay: $D_{y_0}(n, p, X)$

Figure 3.18: $p$-parallel $n$-stage buffer chains where all buffers in chains have the same gate size $X$.

an $n$-stage chain composed of $X$-drive buffers has the following delay distribution:

$$LN\left(\mu + \ln n, -\frac{\sigma}{\sqrt{nX}} \sqrt{1 + \frac{2}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r_{ij}}\right). \quad (3.35)$$

Finally, from Theorem 2 and Corollary 2, the worst case delay of the circuit $D_{y_0}(n, p, X)$ can be expressed as:

$$D_{y_0}(n, p, X) = n \exp\left(\mu + \frac{\sigma}{\sqrt{nX}} \sqrt{1 + \frac{2}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r_{ij}} \sqrt{2} \sqrt{\ln \frac{p}{12(1-y_0)}}\right). \quad (3.36)$$

For better understanding, we consider (3.36) in two conditions according to the degree of $n$ as follows:

**When $n$ is sufficiently large**

The factor $\frac{1}{\sqrt{nX}} \sqrt{1 + \frac{2}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r_{ij}} \sqrt{2} \sqrt{\ln \frac{p}{12(1-y_0)}}$ in (3.36) becomes ignorable. This factor corresponds to the averaging effect of the buffer chains presented in Corollary 1.

In this condition, the sensitivity of $p$ and $X$ to $D_{y_0}(n, p, X)$ is very small, which means that we can take the same design strategy for the optimization of the sub-/near-threshold circuit as taken in the super-threshold circuit design. For example, in case of $n = 64$, $p = 256$ and $y_0 = \Phi(4)$, doubling the value of $X$ from 1 to 2 reduces the worst case delay of the circuit only by 5.0%. Similarly, doubling the value of $p$ from 256 to 512 with $X = 1$ increases the worst case delay by 0.4%. These trends are quite similar to those in
the super-threshold voltage operation.

When $n$ is small

The factor \( \frac{1}{\sqrt{n}} \sqrt{1 + \frac{2}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r_{ij} \sqrt{2 \ln \frac{p}{12(1-y_0)}}} \) in (3.36) is relatively large in comparison with $\mu$. Hence, the sensitivity of $p$ and $X$ to $D_{y_0}(n, p, X)$ is relatively large in this case. From Corollary 2 and Corollary 3, we can conclude that for short critical paths, tuning $X$ has a stronger impact on the performance than tuning the degree of parallelism in the sub-/near-threshold voltage operation. Similar trend is also observed in the super-threshold voltage operation but it is weaker than the sub-/near-threshold voltage operation.

For example, in case of $n = 8$, $p = 256$ and $y_0 = 0.4$, doubling the value of $X$ from 1 to 2 reduces the worst case delay of the circuit by 13% in the sub-/near-threshold voltage operation while the delay is reduced by 2% in the super-threshold voltage operation. Similarly, doubling the value of $p$ from 256 to 512 with $X = 1$ increases the worst case delay by 1.2% in the sub-/near-threshold voltage operation while the delay increases by 0.16% in the super-threshold voltage operation. Both gate sizing effect and parallelization effect in the sub-/near-threshold voltage are much stronger than those observed in the super-threshold voltage operation.

3.6 Application Example

The aim of this section is to show that the optimum design strategy changes depending on the supply voltage through several application examples. First, the performance gain by pipelining is discussed. After that, we show that the optimum memory readout structure changes depending on the supply voltage.

3.6.1 Pipelining

This chapter revealed that a stronger averaging effect is observed in the lognormal distribution than in the Gaussian distribution. This averaging effect intuitively implies that the performance gain of a processor obtained by increasing the number of its pipeline stages in the sub-/near-threshold voltage operation is smaller than that obtained in the super-threshold voltage operation. For better understanding, let us consider the $4\sigma$ worst case delay of buffer chains. Figure 3.19 shows the $4\sigma$ worst case delay of buffer chains with different logic depths $n$. Each worst case delay is normalized by the $4\sigma$ worst case delay of a buffer chain with $n = 128$. Each point represents the degree of pipelining. For example, if we move leftward in Fig. 3.19, the logic depth of a critical path decreases, and
therefore the degree of pipelining increases. By reducing \( n \) from 128 to 4, the \( 4\sigma \) worst case delay decreases to 1/30 in the super-threshold voltage operation while it decreases to 1/20 in the low voltage operation. Therefore, the performance gain by the pipelining is less influential in the low voltage operation than in the super-threshold voltage operation. Ref. [36] pointed out that pipelining requires several extra registers and control circuits, which leads to overhead in terms of delay, energy and area. Therefore, designer should be careful in designing deeply pipelined circuits for the sub-/near-threshold voltage operation since a large overhead may be involved to achieve the same performance gain to pipelined circuits with the super-threshold voltage.

Section 3.5 revealed that gate upsizing effectively reduces the worst case delay of sub-/near-threshold circuits with short critical paths. Therefore, in designing deeply pipelined circuits for the sub-/near-threshold voltage operation, designers should consider gate upsizing in order to mitigate the impact of the delay variation.

### 3.6.2 Standard-Cell Memory Readout Circuit

The performance variations in memory readout circuits are examined using properties derived in this chapter. We show that in the sub-/near-threshold voltage operation, a multiplexer based readout circuit in a standard-cell memory can achieve comparable perfor-
mance to the conventional bit-line-based SRAM readout circuit if we look at the high-\( \sigma \) worst case delay.

**Readout Circuit in On-Chip Memories**

In [27], Meinerzhagen et al. examined various SCM readout structures. They showed that a multiplexer based readout logic leads to smaller area and lower power consumption than a tri-state buffer based readout logic. The tri-state buffer based readout logic uses exactly one tri-state buffer per storage cell, where all tri-state buffer outputs in the same column are connected to a bit line like the readout structures in SRAM macros. If column length becomes large, stronger buffers are required in a tri-state buffer based readout logic to reduce the readout delay. On the other hand, a multiplexer based readout logic utilizes multiple 2-to-1 multiplexers interconnected to operate as a binary selection tree. Since the load capacitance of each multiplexer is relatively small, the high drive strength in each multiplexer is not required. The multiplexer based readout logic thus achieves lower energy and area than those in the tri-state buffer based readout logic. A number of SCMs thus employ the multiplexer based readout logic [15, 24–26].

On the other hand, full-custom SRAM macros including variability-aware low-voltage SRAMs still utilize a bit-line-based readout structure with a strong sense amplifier, which results in large load capacitance on a bit line. Therefore, the readout energy consumption of SRAM macros is larger than that of multiplexer based SCMs since the load capacitance is charged and discharged in every readout cycle.

**Averaging Effect in Readout Multiplexer Tree in Standard-Cell Memory**

From Corollary1, we can prove that the worst case readout delay of SCMs can be smaller than that of SRAMs if we consider the WID variation in sub-/near-threshold voltage. The readout delay of SRAM macros strongly depends on the delay for sense amplifiers to sense the bit line swing by an access transistor in a bit cell. As shown in (2.4) in Chapter 2, the delay is proportional to the load capacitance on a bit line. Therefore the \( k\sigma \) worst case delay of highly parallelized SRAM macros is considerably large. On the other hand, the critical path delay of the multiplexer based readout structure is the sum of serially connected 2-to-1 multiplexers’ delay. Therefore, the \( k\sigma \) worst case delay of the multiple tree is relatively small from the averaging effect.

To validate the consideration mentioned above, let us consider the delay distribution between different readout circuits shown in Fig. 3.20. Figure 3.20 (a) shows a traditional bit-line-based memory readout circuit with a strong sense amplifier on bit lines. Fig-
3.6. Application Example

Figure 3.20: Memory readout structure. (a) SRAM (b) SCM.

Figure 3.20 (b) shows an SCM readout circuit consisting of multi-stage MUX cells. Each memory structure has 256 word memory, and all transistors used for access transistors and MUX cells have the minimum size transistor for a fair comparison. We perform transistor-level circuit simulation [63] with a foundry provided Monte Carlo simulation package targeting a 28-nm process technology for the circuit simulation. We use a 0.4 V supply as the near-threshold voltage, and we do not consider the wire load in the readout circuit. Figure 3.21 shows the CDF versus readout delay per bit cell of a traditional SRAM and an SCM. “SRAM Bit Line Delay” shows the minimum delay for the sense amplifier to sense the minimum identifiable bit line swing, and “SCM MUX Tree Delay” shows the 8-stage MUX delay. We set the voltage of the bit line swing to be 50 mV so that 30,000 sense amplifiers can read values of bit cells correctly. Note that in order to obtain a 3σ yield per chip and the word size of the chip is 32, about 24,000 sense amplifiers must operate correctly. The horizontal axis in Fig. 3.21 is a logarithmic scale, which means that lognormal CDF becomes a straight line in this scale. Dots in Fig. 3.21 show the results of the Monte Carlo simulation. Since the dots line up roughly in a straight line, the property that the gate delay has the lognormal distribution as discussed in this chapter. Thus, we predict the 5σ worst case delays for every bit cell of both SRAM and the SCM by fitting them to the lognormal distribution. Note that in order to keep 3σ yields per memory circuit with 1 KB capacity, we have to keep 5σ yields per bit cell. Solid lines in Fig. 3.21 are the fitting results. From the fitting results, the two delay lines get closer and eventually intersect when we look at the high-sigma worst case. Fitting results show that,
in the 0σ condition, where there is no WID variation, SRAM is faster than SCM by 20 ns. However, in the 5σ worst case condition per bit cell, the SCM is faster than the SRAM by 3.5 ns. The results show that readout structure based on a multiplexor (MUX) cell can achieve comparable performance to the conventional SRAM readout circuit if we look at a high-σ (e.g., 5σ per bit cell) the worst case delay although the conventional SRAM is much faster than the MUX-based readout structure in a typical case condition.

On the other hand, the delay variation in the super-threshold voltage operation is much smaller than that in the near-threshold voltage operation, which makes the two lines in Fig. 3.21 steep. Therefore, SRAM readout circuits in the super-threshold voltage operation still operate faster than MUX-based readout structures.

### 3.7 Summary

Voltage scaling is one of the most promising approaches for achieving high performance and energy efficient computation of microprocessors. This thesis derived delay variation models in both the nominal voltage operation and the low voltage operation. This thesis showed that the propagation delay of circuits follow the lognormal distribution in the sub-/near-threshold voltage operation while they follow the Gaussian distribution in the super-threshold voltage operation. Based on this fact, this thesis proves several theorems that help consider architectural design strategies for both the operating voltages. Corollary1 shows that sub-/near-threshold voltage operation has a stronger averaging effect
than the super-threshold voltage operation where the standard deviation is proportional to the square root of the number of chained gates in series. With Monte Carlo simulation using a commercial 28-nm process technology model, we show that the averaging effect of a buffer chain in the sub-/near-threshold voltage operation is 27% stronger than that in the super-threshold voltage operation. Theorem 2 shows that the maximum operating frequency (F_{\text{max}}) of a circuit operated with the sub-/near-threshold voltage is more widely degraded than the same circuit operated with the super-threshold voltage when the number of critical paths increases. However, Corollary 2 shows that the F_{\text{max}} degradation along with the increase of the number of parallel paths (N_{\text{cp}}) is negligibly slow in both cases. This means that the impact of the parallelization on the F_{\text{max}} degradation is negligible even in the sub-/near-threshold voltage operation. Corollary 3 shows the gate upsizing exponentially improves the worst case delay in the sub-/near-threshold voltage operation compared with the delay improvement in the super-threshold voltage operation. Section 3.6 showed that several optimum architectural-level design strategies change depending on the supply voltages. Especially, Subsection 3.6.2 shows that the multiplexer based readout structure also achieves less worst case delay than the bit-line-based SRAM readout structure in low voltage operation. Therefore, instead of SRAM macros, using SCMs as on-chip memories is a good choice in terms of operating speed in the sub-/near-threshold voltage operation. However, area-overhead in SCMs is considerably large compared with SRAM macros. In Chapter 4, a design method to reduce the area-overhead of SCMs are proposed.
Chapter 4

Area- and Energy-Efficient Standard-Cell Memory using Minimum Height Standard-Cells

4.1 Introduction

As described in Chapter 2, a major drawback of SCMs is area-efficiency. Therefore, this thesis firstly aims to develop area-efficient SCM structure. The dynamic energy consumption of SCMs can be effectively reduced by the architectural-level optimization. This thesis secondly proposes energy-efficient SCM structure in order to achieve area- and energy-efficient on-chip memories.

Several papers have already proposed area-efficient SCM structure [1, 25, 26]. In [26], SCMs integrating a area-optimized standard-cell are proposed. Two latch cells and two NAND2 gates, which are the dominant cells that occupy the SCMs’ area, are implemented to one standard cell in a 65-nm process technology in order to improve the SCM’s area-efficiency. The implementation results show that the SCM achieves the area-efficiency of 8.5 $\mu$m$^2$ per bit, which is the best area-efficiency in the prior-art SCMs in a 65-nm process technology. Refs. [1, 25] propose a design methodology which optimizes the physical placement of SCMs. By the controlled placement of the SCMs, a placement density approaching 100% is achieved in a 28-nm process technology. As a result, an average area reduction of 25% is achieved in comparison with non-controlled SCMs. The controlled SCMs also achieve less area than full-custom SRAMs for memories of up to approximately 1 kb. However, the total area of SCM with a large capacity is still several times larger than that of full-custom SRAM macros. To solve this problem, this thesis proposes a novel area-efficient SCM structure targeting the capacities of up to several
Chapter 4. Area- and Energy-Efficient Standard-Cell Memory using Minimum Height Standard-Cells

4.2 Minimum Height Standard-Cell

4.2.1 Concept

A conventional standard-cell library includes complex logic gates such as FADD, XOR and DFF. In order to keep their routability, a conventional standard-cell has the height of 6, 9 or 12 wire tracks. In SCM designs, however, those complex logic gates are not required since SCMs only require simple logic functions. For example, typical SCM has the following functions: i) reading values of the storage elements specified by address signals, ii) writing some values to the storage elements specified by address signals, and iii) just keeping values at storage elements. Therefore, SCMs require only simple logic functions such as address decoders and readout multiplexers (MUXes) and no complex logic function such as exclusive-or (XOR) or majority function (MAJ) is required. Therefore, this thesis proposes a minimum height standard-cell library (MHSC library in the
4.2. Minimum Height Standard-Cell

Figure 4.2: An inverter cell with minimum cell height.

following) with only simple logic gates. A concept of the MHSC is illustrated in Fig. 4.1. The MHSC has the minimum height to construct complementary CMOS logics, which leads to area reduction of SCMs. It is still possible to use a commercial place and route tool to design the proposed SCM with MHSCs. Thus the design cost of the proposed SCM is still lower than that of full-custom SRAM macros.

It is important to clarify the dominant factors that determine the physical structure of the MHSCs since the height of MHSCs strongly depends on the mask design rule. In the rest of this section, we present process-independent methods to design the MHSC library.

4.2.2 Physical Design

Figure 4.2 shows a layout example of an inverter cell with the minimum height to construct complementary CMOS logics. As depicted in Fig. 4.2, the height of MHSCs is determined by the following four parts:

- Part 1: Poly extension from diffusion layer,
- Part 2: pMOS transistor gate width or nMOS transistor gate width,
- Part 3: Poly extension to put contact on poly layer, spacing between poly and diffusion, and
- Part 4: Metal spacing.

In MHSC design, each part is designed to be the minimum value. Part 1 can be easily determined by mask design rules. We can also easily determine parts 3 and 4. However, the maximum value among above four, which determines the height of MHSCs depends on the mask design rules. The gate widths of pMOS transistors and nMOS transistors (part 2) are set as the same size to minimize the height of MHSCs.
It is hard to determine the value of the part 2 (i.e., gate width of transistors). This is because gate width of standard-cells is a major tuning knob that affects the performance of SCMs. Especially, the stability of standard-cells strongly depends on the gate width in the aggressively voltage-scaled region. Moreover, the gate widths of pMOS transistors and nMOS transistors are the same. Therefore, pull-up/down networks are unbalanced in MHSCs, which further degrades the stability of MHSCs. In this thesis, we clarify the factors that determine the minimum value of the part 2 in the following way.

**Yield-driven width** $W_Y$

Standard-cells with small transistors are vulnerable to process variation. We define the yield-driven gate width $W_Y$ as the minimum gate width that keeps specific yields (e.g. 5σ per cell). Note that $W_Y$ also depends on the supply voltage since yields of standard-cells are degraded in low voltage operation. An analytical approach for determining $W_Y$ is discussed in Section 4.4.

**Minimum width** $W_{\text{DRmin}}$ **in mask design rules**

There is a minimum gate width that is allowed by the mask design rules. We denote the minimum gate width as $W_{\text{DRmin}}$.

**Contacted-minimum** $W_{\text{CM}}$

$W_{\text{CM}}$ is defined as the sum of i) the minimum height of contact layer, and ii) the minimum extension of diffusion layer from contact layer. For example, the value of part 2 in Fig. 4.2 is determined by the contacted-minimum $W_{\text{CM}}$.

In this thesis, we define the minimum gate width in MHSCs as $W_{\text{MH}}$. We can obtain $W_{\text{MH}}$ by taking a MAX of the three gate widths ($W_Y$, $W_{\text{DRmin}}$ and $W_{\text{CM}}$):

$$W_{\text{MH}} = \max \{W_Y, W_{\text{DRmin}}, W_{\text{CM}}\}. \tag{4.1}$$

Reducing cell heights degrades routability within standard-cells. To solve this problem, MHSC libraries have only simple logic cells such as NAND2 and NOR2. Metal 2 layer or more are also used within standard-cells, which leads to routing congestion problems in the place-and-route phase. However, SCMs require only simple logic function such as address decoders and readout multiplexers. Therefore, routing congestion problems in SCM design is not so curial as in conventional logic circuits design.
As described above, the MHSC library has only simple logic cells in comparison with conventional standard-cell libraries. The simplified MHSC library is discussed at the next subsection.

4.2.3 Simplified Cell Library

In [17], Jain et al. pointed out that standard-cells with four stacked transistors exhibit significantly larger delay than logic gates with three or less stacks if we consider the $6\sigma$ WID variation. Therefore, it is better to prune complex logic gates with four stacked transistors from the viewpoint of the variability. As described at the previous subsections, it is also true that SCMs require only simple logic function such as address decoders and readout multiplexers. Therefore, the MHSC library has only simple logic gates with one or two stacked transistors. All the possible complementary CMOS logic gates with one or two stacked transistors are summarized below:

- 1-input cell: INV $\bar{a}$,
- 2-input cell: NAND2 $ab$, NOR2 $\bar{a} + \bar{b}$,
- 3-input cell: AOI21 $ab + c$, OAI21 $(a + b)c$,
- 4-input cell: AOI22 $ab + cd$, OAI22 $(a + b)(c + d)$.

Storage elements are also required to design SCMs. In this thesis, we present D-Latch cells as storage elements. Energy- and area-efficient latch cells are described at the next subsection.

4.2.4 Simplified Latch for Energy- and Area-Efficiency

Energy- and area-efficient D-Latch cells are presented since latch cells are dominant cells that occupy the area of SCMs. Figure 4.3 (a) shows simplified latch architecture that is used as storage elements. Note that the proposed latch cells are only used as bit cells and their fanout is always one. Therefore, the input/output drivers in a conventional latch can be removed, which leads to less area and dynamic energy consumption. The integration of clock drivers also reduces area and dynamic energy consumption as depicted in Fig. 4.3 (b).

It is important to estimate the yield of latch cells since latch cells are the most vulnerable cells to process variation. In this thesis, the yield is estimated using the analytical stability model proposed in [64, 65]. We determine the yield-driven width $W_Y$ by the stability model in Section 4.4.
4.3 Energy-Efficient Memory Architecture

We consider a dual port SCM with one read port and one write port which operates in a single clock cycle. We assume that the SCM has \( R \)-words and each word has \( C \)-bits with an address width \( m \). A block diagram of the proposed SCM is depicted in Fig. 4.4. As described in Section 4.2, latch cells are implemented as bit cells. In readout operation, the outputs of latch cells are selected at the R-to-1 MUX by one-hot signals from the read
address decoder. In write operation, latch cells labeled “Write Latch” act as maser latches and latch cells selected by the write address decoders act as slave latches. Therefore, they form master-slave registers in write operation.

In this section, energy-efficient SCM architecture is discussed. In order to reduce the area-overhead of SCMs, a bit-line-based readout/write structure is typically used, where all of input or output ports in a column of the memory array share one bit line. Since the bit-line-based structure discharges or charges a large capacitive load on a bit line, it consumes large dynamic energy consumption in each operation as described in Subsection 3.6.2. This thesis presents energy-aware SCM architecture. The key is to reduce the effective capacitance of the SCM in read or write operation.

### 4.3.1 Write Scheme

Figure 4.5 shows an energy-aware write scheme. Gating circuits in “Write Gating” prevent glitches in one-hot signals from an address decoder. Since the write gating circuit is required for each word, a huge clock tree is required, which leads to considerably large
energy consumption. To reduce the energy consumption, several previous works implement the clock gating circuits to the clock tree [1, 24, 25]. This thesis also implements a clock gating scheme in “Second Level Write Clock Gating” which is also presented in [1, 25]. g-bit address signals from the MSB are used for the enable signal in the gating circuit, which reduces the dynamic energy consumption at the clock tree up to $1/2^g$ in comparison with the simple structure.

The dynamic energy consumption at the bit-line also occupies a large portion of the energy consumption in write operation. This is because the capacitance on the bit-line is approximately proportional to the input capacitance of a latch cell and the number of rows ($R$) of the memory. To reduce the dynamic energy consumption, this thesis newly implements a write demultiplexer for a write data path depicted in “DEMUX.” The $2^g$ AND gates in “DEMUX” reduce dynamic energy consumption up to $1/2^g$ of the energy consumption in the simple bit-line-based write structure.

### 4.3.2 Readout Scheme

As described in Subsection 3.6.2, a multiplexer readout logic achieves less energy consumption than a tri-sate buffer based logic and bit-line-based readout structure. Therefore, it is better to implement a multiplexer readout logic to the proposed SCM to reduce energy consumption. The readout data is logically implemented as the $2^m$ OR operations between ANDed latch outputs and one-hot signals from an address decoder. Refs. [1, 25] implement the readout structure by firstly taking a NAND2 of the output of each latch and an one-hot signal from a read address decoder. After the NAND2 gates, a 2-input NAND-NOR tree is iteratively interconnected to form a readout multiplexer. To implement the multiplexer described above, $(2^{m+1} - 1)$ NAND2/NOR2 gates are required for each column, which leads to the area-overhead of the SCM in comparison with the simple bit-line-based readout structure. To reduce the area overhead, this thesis proposes the AOI22-NAND2-NOR2 tree depicted in Fig. 4.6. The first two stages of the readout multiplexer in [1, 25] is replaced with AOI22 cells. Since three NAND2 gates are replaced with one AOI22 cell, the number of transistors required at the first two stages is reduced by 33%, which leads to less area overhead and less energy consumption.

Since the proposed readout scheme activates only logic gates on a readout signal path, the dynamic energy consumption in readout operation is drastically reduced in comparison with the simple bit-line-based readout structure. For example, the readout scheme in Fig. 4.6 activates only four logic gates in each readout operation.

It is also true that the energy-efficient memory architecture needs several extra pe-

In this section, the proposed MHSCM is implemented in a 65-nm FD-SOI process technology as a case study. We firstly analyze the stability of standard-cells to determine the yield-driven width $W_Y$ in Section 4.2. Based on the design method in Section 4.2, we design a MHSC library in the 65-nm FD-SOI process technology. After that, we design MHSCMs with several configurations using the MHSC library and evaluate their performance. The performance is compared with the prior-art SCMs and SRAMs designed using 65-nm process technologies.

4.4.1 Stability Analysis for Latch Cells and Standard-Cell Design

Storage elements are the most vulnerable to process variations in on-chip memories. The WID process variation has a large impact on storage elements’ stability since transistors...
are aggressively scaled in storage elements. In [24], it is pointed out that failure modes of conventional 6T SRAM macros are explained as: i) readout failure, ii) write failure, or iii) hold failure. Readout failures arise from the direct access to storage elements via access transistors which is not presented in standard-cell latches. It is also true that a feedback loop in a latch cell is isolated in write operation by a clocked-inverter. Therefore, standard-cell latches basically do not suffer from the readout failures or the write failures. The only failure to be considered is a hold failure that latch cells can not hold their values. There are several existing techniques which analytically estimate the hold stability of latch cells [64–66]. In this thesis, we use an analytical stability model to estimate the hold stability of latch cells [64, 65]. Based on the stability model, we determine the yield-driven width $W_Y$ described in Section 4.2.

Refs. [64, 65] developed the analytical stability model of cross-coupled inverters operating in sub-threshold voltage. The schematic of the cross-coupled inverters are depicted in Fig. 4.7 (a). One of the inverters is a clock inverter to isolate a feedback loop in write operation. As shown in Fig. 4.7 (a), we set ‘0’ and ‘1’ signals to the pMOS transistor and the nMOS transistor of the clock inverter to make a feedback loop, respectively. The stability of the inverters is graphically understood as a butterfly curve. Figure 4.7 (b) shows voltage transfer characteristics (VTCs) of the two cross-coupled inverters. As depicted
4.4. Implementation of the Proposed SCM in a 65-nm FD-SOI Process Technology

In Fig. 4.7 (b), the two VTCs form a butterfly curve. When the cross-coupled inverters successfully hold their value, the butterfly curve has two eyes. On the other hand, if the cross-coupled inverters fail to hold their value due to WID variations, the shape of the two VTCs fluctuates and one of the eyes closes. It is hard to analytically derive the probability that one of the eyes closes due to process variations. Refs. [64, 65] developed the approximation function of the probability with acceptable estimation errors. The approximated probability $Y$ is obtained by integrating a bivariate Gaussian function as shown in (4.2):

$$Y = \int_{L_{\text{const}}}^{H'_{\text{const}}} \int_{L'_{\text{const}}}^{H_{\text{const}}} f(\Delta u, \Delta v) d\Delta ud\Delta v,$$

where $\Delta u$ and $\Delta v$ are random variables that follow a correlated bivariate Gaussian distribution. $H_{\text{const}}, H'_{\text{const}}, L_{\text{const}},$ and $L'_{\text{const}}$ are parameters that depend on a process technology, supply voltage ($V_{\text{DD}}$) and gate size ($W$).

The analytical model (4.2) is derived by I-V characteristics of transistors which operate in sub-threshold voltage. The characteristics are summarized below:

$$I = \frac{k W}{L} \exp\left(\frac{V_{\text{GS}} + \eta V_{\text{DS}} - V_{\text{th}}}{n_i v_T}\right) \left(1 - \exp\left(\frac{-V_{\text{DS}}}{v_T}\right)\right),$$

where $n_i$ is an ideal factor, $v_T$ is the thermal voltage, $\eta$ is a DIBL (Drain-Induced Barrier Lowering) coefficient, $W$ is a transistor gate width, $L$ is a transistor gate length, and $V_{\text{th}}$ is a threshold voltage of a transistor. By sweeping gate-to-source voltage ($V_{\text{GS}}$) and drain-to-source voltage ($V_{\text{DS}}$), all the parameters are fitted through transistor-level DC simulation where the supply voltage is in the sub-threshold region ($V_{\text{DD}} < 200$ mV in the 65-nm FD-SOI process technology). Based on the fitted parameters, $L_{\text{const}}, L'_{\text{const}}, H_{\text{const}},$ and $H'_{\text{const}}$ are derived. The analytical model is verified by Monte Carlo simulation in the 65-nm FD-SOI process technology. Figure 4.8 shows yield of the latch cells for $W = W_{\text{CM}}$ and $W = W_{\text{DRmin}}$. The horizontal axis is supply voltage and the vertical axis is yield per latch cell. Solid line is the estimated yield of latch cells using the analytical model (4.2). Dots are results obtained by 100,000 Monte Carlo trials using transistor-level simulation. Note that threshold voltages of transistors follow the Gaussian distribution and they fluctuate in each Monte Carlo trial. The results show that the analytical model (4.2) is feasible in the 65-nm FD-SOI process.

Figure 4.9 shows the yield per bit cell estimated by the analytical model (4.2). The horizontal axis is gate width of latch cells and the vertical axis is yield per latch cell $Y$. The supply voltage is set as 200 mV. $W_Y$, $W_{\text{DRmin}}$ and $W_{\text{CM}}$ are plotted to determine $W_{\text{MH}}$ in (4.1). In this thesis, we define $W_Y$ as the minimum gate width to keep $5\sigma$ yield per latch
Chapter 4. Area- and Energy-Efficient Standard-Cell Memory using Minimum Height Standard-Cells

Figure 4.8: Verification of the analytical stability model of latch cells (4.2).

Figure 4.9: Yields of latch cells for various gate widths.

cell. Note that in order to keep $3\sigma$ yield per memory macro, we have to keep more than $3\sigma$ yield per latch cell. If we keep $5\sigma$ yield per latch cell, we can keep about $3\sigma$ yield per 16 kb memory. Note that a 16 kb capacity corresponds to the largest target capacity
4.4. Implementation of the Proposed SCM in a 65-nm FD-SOI Process Technology

Table 4.1: 5.5-track minimum height standard cell library in the target 65-nm FD-SOI process technology.

<table>
<thead>
<tr>
<th>Logic</th>
<th>Drive Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>INV</td>
<td>1X, 2X, 4X, 8X, 16X, 32X</td>
</tr>
<tr>
<td>NAND2</td>
<td>1X, 2X</td>
</tr>
<tr>
<td>NOR2</td>
<td>1X, 2X</td>
</tr>
<tr>
<td>AOI22</td>
<td>1X</td>
</tr>
<tr>
<td>OAI22</td>
<td>1X</td>
</tr>
<tr>
<td>Delay Cell (2-stacked INV)</td>
<td>1X</td>
</tr>
<tr>
<td>4-bit D-LATCH</td>
<td>1X</td>
</tr>
<tr>
<td>1-bit D-LATCH for peripheral circuits</td>
<td>1X</td>
</tr>
</tbody>
</table>

of SCMs in this thesis. The results show that \( W_{CM} \) is larger than \( W_Y \) and \( W_{DR_{min}} \) in the 65-nm FD-SOI process. Since the yield of latch cells improves as we increase the supply voltage, \( W_Y \) decreases as we increase the supply voltage. Therefore if we design SCMs which operate in more than a 200 mV supply voltage in the 65-nm FD-SOI process, \( W_{MH} \) in (4.1) is determined by contacted-minimum \( W_{CM} \):

\[
W_{MH} = W_{CM} \text{ (in 65 nm FD-SOI).} \tag{4.4}
\]

Based on (4.4), a MHSC library is designed in the 65-nm FD-SOI. By the design method in Section 4.2 and \( W_{MH} \), the height of MHSCs is derived as 5.5 wire tack. As described in Section 4.2, the logic gates in the MHSC library have only simple logic functions with one or two stacked transistors. Among the simple logic gates, logic gates in Table 4.1 are designed. “Delay Cell” is an inverter. However, it has two stacked transistors in the pull-up network and the pull-down network. It is used to fix hold violations in an automated place-and-route process. Figure 4.10 shows layout examples of NAND2 and AOI22. Metal 1, 2 and 3 are used design MHSCs. In addition to the three layers, Metal 4 and 5 are used for signal routing in the place-and-route process. Metal 6 and 7 are used as the power mesh in the place-and-route process.

4.4.2 Comparison with Prior-Art SCMs

The performance of the proposed SCM is compared with SRAM macros and SCM macros in [1, 24, 26, 67–70]. The proposed SCMs with several configurations are designed with the 65-nm FD-SOI technology. The configurations designed in this work are 32×32 (1 kb), 128×32 (4 kb), 512×32 (16 kb) and 256×64 (16 kb). The core utilization of
all the SCMs is kept around 80% in order to reduce the routing congestion. The SCM is implemented with a standard digital circuit design flow which utilizes a commercial place and route tool. Therefore, unlike full-custom SRAM macros, flexible floorplanning is available when we design SCMs. For example, Fig. 4.11 shows several layouts of the
4.4. Implementation of the Proposed SCM in a 65-nm FD-SOI Process Technology

Figure 4.12: Area-comparison between the proposed SCMs, prior-art SCMs and SRAMs. The area of the SCMs in [1] is multiplied by \((100 \text{ nm}/50 \text{ nm})^2 = 4\).

The proposed SCM with a \(512\times32\) configuration. Unlike full-custom SRAM macros, the proposed SCMs enable designers to make on-chip memories with various aspect ratios. Note that SCMs in Fig. 4.11 are designed to achieve almost the same area-efficiency under the similar design constraint except for floor plans. For the subsequent experiments, the performance of square-shaped SCMs is evaluated for the four configurations described above.

Figure 4.12 shows the area of various SCMs and SRAMs. The total area in Fig. 4.12 is normalized by the capacity of the memory. Therefore each point represents the area per bit. The area of the SCMs in [1] is extrapolated based on the feature size \(F\) since they are designed in a 28-nm process technology while the others are designed in 65-nm process technologies. Note that we assume that \(F\) is 100 nm and 50 nm in the 65-nm process technology and the 28-nm process technology, respectively. Therefore the SCMs’ area in [1] is multiplied by \((100 \text{ nm}/50 \text{ nm})^2 = 4\). Note that Ref. [1] evaluated various configurations (i.e., various \(R\)s and \(C\)s) that has the same memory capacity (i.e., \(R \times C\)). Among the configurations, the smallest area is plotted to Fig. 4.12. The results show that the proposed SCM achieves better area-efficiency than prior-art SCMs. If we look at \(128 \times 32\) configurations, the proposed SCM achieves 6.82 \(\mu\text{m}^2/\text{bit}\) which is 20% smaller
than that of the most area-efficient SCM in prior works. 6.82 $\mu m^2$/bit corresponds to $682F^2$ in this process technology ($F = 100$ nm). The results show that MHSCs which have a minimum possible cell height allowed by the logic design rule effectively improve area-efficiency of the SCM. Although the proposed SCM achieves better area-efficiency than the prior-art SCMs, full-custom SRAM macros are still more area-efficient than the proposed SCM. The area overhead of the proposed SCM with a 512x32 capacity, which is the closest configuration to prior-art SRAMs, is up to +132% compared with full-custom SRAM macros (CACTI, SRAM1 and SRAM2).

The operating frequency is evaluated using a commercial static timing analysis (STA) tool. The maximum operating frequency of the SCMs is evaluated using post-layout STA. Figure 4.13 shows the maximum operating frequency of the proposed SCMs. The proposed SCMs operate at the wide supply range from 1.2 V to 0.35 V. The maximum operating frequency is more than 12 MHz at $V_{DD} = 0.35$ V. The maximum operating frequency is summarized in Table 4.2. The proposed SCMs achieve at least ten times faster speed than that of prior-art SCMs in 65-nm process technologies (SCM1, SCM2 and SCM3) and sub-threshold SRAM macros (SRAM1 and SRAM2).

This thesis also evaluates the energy consumption using post-layout gate-level simulation. We evaluated the energy consumption when the operating frequency is the maximum operating frequency. The following energy consumption are evaluated:

- Read energy: energy consumption when the SCM is accessed sequentially.
- Write energy: energy consumption when we write random values.
Table 4.2: Comparison between Prior-Art SCMs and SRAMs.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>65-nm</td>
<td>512×32</td>
<td>SKAM Model</td>
<td>150(1) (1.2 V)</td>
<td>7.9 pW/bit</td>
<td>1.01 GHz</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>256×128</td>
<td>AM</td>
<td>154(1) (0.4 V)</td>
<td>98 pW/bit</td>
<td>475 kHz</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>1024×32</td>
<td>AM</td>
<td>155(1) (0.35 V)</td>
<td>-</td>
<td>500 kHz</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>28-nm</td>
<td>PLS</td>
<td>32.7(1) (0.35 V)</td>
<td>-</td>
<td>1 MHz</td>
<td>3.9</td>
</tr>
<tr>
<td>65-nm</td>
<td>128×32</td>
<td>AM</td>
<td>14(1) (0.5 V)</td>
<td>0.5 pW/bit</td>
<td>110 kHz</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>65-nm</td>
<td>AM</td>
<td>35(1) (0.35 V)</td>
<td>6 pW/bit</td>
<td>200 kHz</td>
<td>12.7</td>
</tr>
<tr>
<td>65-nm</td>
<td>28-nm</td>
<td>PLS</td>
<td>-</td>
<td>0.32 nW/bit</td>
<td>-</td>
<td>8.5</td>
</tr>
<tr>
<td>65-nm</td>
<td>512×32</td>
<td>PLS</td>
<td>Sleep: 2.7</td>
<td>Sleep: 14</td>
<td>Sleep: 37</td>
<td></td>
</tr>
<tr>
<td>28-nm</td>
<td>256×128</td>
<td>AM</td>
<td>Read: 6.6</td>
<td>Read: 21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-nm</td>
<td>1024×32</td>
<td>AM</td>
<td>Write: 11</td>
<td>Write: 37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-nm</td>
<td>128×32</td>
<td>AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28-nm</td>
<td>28-nm</td>
<td>AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(2): Area estimated from die photograph by [24].
(3): The value is extrapolated using the feature size (F) for a fair comparison. The value is multiplied by (100 nm/50 nm)². The original area before the extrapolation is 3.01 µm²/bit.


- Sleep energy: energy consumption when the SCM is not in read operation or write operation. The sleep energy almost corresponds to leakage energy consumption.

The proposed SCM has random values in the initial state. Figures 4.14, 4.15 and 4.16 show the evaluation results. Energy consumption is normalized by the word width (C) and the number of clock cycles. The normalized energy is equivalent to “energy per accessed bit” of prior-art memories. The results show that write, read and sleep energy decreases as we lower V_{DD}. If we look at the SCM with a 512×32 configuration, it achieves energy consumption of 36.9 fJ, 19.6 fJ and 7.02 fJ, respectively. Table 4.2 summarizes the performance of the proposed SCMs and the prior-art SCMs and SRAMs. If we look at the proposed SCMs with a 128×32 configuration, the proposed SCM exhibits comparable energy to prior-art SCMs with a same memory capacity (SCM2 and SCM3). In the same way, the proposed SCMs with a 512×32 configuration, which is the closest configuration to prior-art SRAMs, shows less energy than prior-art SRAMs (CACTI, SRAM1 and SRAM2). Since SRAM macros employ the bit-line-based structure which has large load capacitance at the bit line, its energy consumption is large. The proposed SCM, on the other hand, employs a low activity readout structure with signal gating. Therefore, the energy efficiency of the proposed SCM is better than that of SRAM macros. Figures 4.14, 4.15 and 4.16 also show that the optimum supply voltages that minimize the each energy consumption are different. This is because the ratios of dynamic energy consumption and
leakage energy consumption are different in each operation. Dynamic energy consumption quadratically decreases with the $V_{DD}$ scaling whereas leakage energy consumption increases in low supply voltage. Since clock trees are activated only in write operation, dynamic energy consumption in write operation is thus larger than those of the others, resulting in the lower optimum supply voltage than those of the other operations in order to reduce the dynamic energy consumption. Therefore, further energy reduction in write operation will be observed if we scale the supply voltage below 0.35 V. Figure 4.17 shows the leakage power of the proposed SCM with a 512×32 capacity. The leakage power is normalized by the number of the total bit cells. Note that the proposed SCM with the other configurations exhibits a comparable or less leakage power since the leakage power is roughly proportional to the number of logic gates. The leakage power decreases to 317 pW per bit when we scale the supply voltage. We can further reduce the leakage power by applying the reverse body bias. Simulation results show that 0.3 V reverse body bias effectively decreases the leakage power to 82 pW per bit. However its value is still much larger than that of the prior-art SCMs in Table 4.2. Further aggressive reverse body biasing will reduce the leakage power. Note that reverse body biasing also decreases the maximum operating speed of the SCM. For example, if we apply 0.3 V reverse body bias at a 0.35 V supply voltage, the maximum operating speed decreases to 5.8 MHz. If we apply further aggressive reverse body biasing, the maximum operating speed will be comparable to the prior-art SCMs.

4.5 Summary

In this chapter, an energy- and area-efficient SCM structure which is aimed at sub-/near-threshold operation is proposed. Minimum height standard-cells with simplified latches are designed, addressing area-overhead of SCMs. Energy-aware readout/write scheme is then presented. By utilizing AOI22-NAND2-NOR2 readout scheme with signal gating, the proposed SCM reduces dynamic energy consumption. Evaluation results using a 65-nm FD-SOI process technology show that the proposed SCM with a 4 kb capacity achieves the energy consumption of less than 11 fJ per bit with a 20 MHz clock frequency and area efficiency of 6.82 $\mu$m$^2$ per bit cell (682$F^2$ per bit cell). The area of the proposed SCM is 20% smaller than that of the most area-efficient SCM in prior works. The proposed SCM with a 16 kb capacity achieves 75% and 31% less energy consumption than nominal-voltage SRAMs and low-voltage SRAMs, respectively. The area overhead of the proposed SCM is up to +132% compared with the SRAMs.
Figure 4.14: Estimated write energy consumption per bit with a scaled $V_{DD}$.

Figure 4.15: Estimated read energy consumption per bit with a scaled $V_{DD}$.

Figure 4.16: Estimated sleep energy consumption per bit with a scaled $V_{DD}$.
Figure 4.17: Leakage power per bit with a scaled $V_{DD}$. 
Chapter 5

Minimum Energy Point Operation using Supply and Threshold Voltage Scaling

5.1 Introduction

This chapter discusses a post-silicon tuning technique which dynamically scales $V_{DD}$ and $V_{th}$ for energy-efficient LSI circuits. The focus of this chapter is to find a Minimum Energy Point (MEP) that minimizes the energy consumption of circuits under a specific clock frequency. The MEP can be defined as a solution of the following optimization problem:

$$\min_{V_{DD}, V_{th}} E$$

s.t. $D \leq D_0$. \hspace{1cm} (5.1)

Here, $E$ is the total energy consumption of a circuit defined as (2.1). $D$ and $D_0$ are the critical path delay and the delay constraint of the circuit, respectively. It is not trivial to find its solution (i.e., MEP) of even a simple inverter chain since MEPs depend on a number of parameters of a target circuit. Among these parameters, an activity factor and a chip temperature are important parameters that determine MEPs of a circuit. In [22], based on the simple transregional models of a CMOS circuit, Nose et al. showed a simple equation that holds when a circuit operates on MEPs at the highest temperature and at the lowest $V_{th}$ corner in the process condition. The equation is summarized in (5.2).

$$\frac{P_{\text{LEAK, max}}}{P_D} = \frac{2N_c\alpha}{\alpha - 1} \hspace{1cm} (\alpha > 1.1).$$ \hspace{1cm} (5.2)
Chapter 5. Minimum Energy Point Operation using Supply and Threshold Voltage Scaling

Note that the right-hand side of (5.2) is normalized by 1 V. $P_{\text{LEAK,max}}$ is the leakage power dissipation at the highest temperature and at the lowest $V_{\text{th}}$ corner in the process condition. $P_{\text{D}}$ is the dynamic power dissipation. $N_s$ is $n_i \cdot \phi_T$, where $n_i$ is the ideal factor of MOSFETs, and $\phi_T$ is the thermal voltage. $\alpha$ is a fitting parameter of the alpha power law MOSFET model [29]. Based on (5.2), they pointed out that the ratio of maximum leakage power to total power dissipation is about 30% if the circuit operates on an MEP, which implies that balance between dynamic energy consumption and leakage energy consumption is the key to energy efficiency. They also pointed out that MEPS strongly depend on an activity factor and a chip temperature of a circuit. For example, if an activity factor of a circuit is considerably small, its leakage energy is the dominant factor that determines the total energy dissipation. Therefore, increasing both $V_{\text{th}}$ and $V_{\text{DD}}$ effectively reduces leakage energy and thus total energy consumption without degrading its operating speed. Although (5.2) is a simple formula, they use several approximations by the Taylor expansion to derive (5.2) in [22]. Unlike Ref. [22], this thesis derives the ratio of the dynamic energy consumption to the static energy consumption ($E_d/E_s$) when a circuit operates on MEPS in a closed form function. The formula is summarized in (5.29). Based on the alpha power law MOSFET model [29] and the simple transregional models of a CMOS circuit, (5.29) is derived without any approximations. This thesis also gives a proof that (5.29) is a necessary and sufficient condition for the minimum energy point operation of a circuit. The condition indicates that the pair of $V_{\text{DD}}$ and $V_{\text{th}}$ determines an MEP if and only if it satisfies (5.29). Since there are several existing techniques to monitor $E_d$, $E_s$ (i.e., leakage current) and the chip temperature ($T$) which appear in (5.29), we can easily find MEPS by finding $V_{\text{DD}}$ and $V_{\text{th}}$ satisfying (5.29) under a specific delay constraint. For example, Ref. [71] proposed a dynamic power monitor of circuits. They showed that the dynamic power dissipation can be accurately estimated by several performance counters of key signals which reflect the dynamic power dissipation. Ref. [72] proposes a leakage current sensor using an analog approach. The leakage current and the on-chip temperature can also be monitored using the fully digital leakage monitor presented in [73]. The on-chip temperature monitors are also proposed in [74, 75]. Critical path replicas which estimate the critical path delay of a circuit are also proposed to check whether or not the circuit meets the specific delay constraint [76, 77]. Based on the monitored values, simple but effective algorithms to find MEPS are proposed in [23, 78]. In [23], based on (5.2), the ratio of switching current to leakage current ($I_{\text{SW}}/I_{\text{LEAK}}$, i.e., $E_d/E_s$) is kept to a specific value so that a circuit operates on MEPS. They proposed the simple algorithm to change $V_{\text{DD}}$ or $V_{\text{BB}}$ iteratively based on the monitored value $I_{\text{SW}}/I_{\text{LEAK}}$. Similarly, the MEP tracking algorithm is proposed in [78]. They show the simple algorithm that iter-
5.2 Modeling for Minimum Energy Point Operation

Let us consider to minimize the energy consumption of a circuit under a specific time constraint by tuning \( V_{\text{DD}} \) and \( V_{\text{th}} \). We target a 50-stage fanout-4 inverter chain designed to reflect the behavior of a microprocessor pipeline as shown in Fig. 5.1. The energy consumption of the circuit can be accurately modeled by (2.1), (2.2) and (2.3). For better understanding, we show again the total energy consumption, the dynamic energy consumption and the static energy consumption of the circuit in (5.3), (5.4) and (5.5), respectively.

\[
E = E_d + E_s, \quad (5.3)
\]
\[
E_d = k_1 V_{\text{DD}}^2, \quad (5.4)
\]
\[
E_s = k_2 D V_{\text{DD}} \exp \left( -\frac{V_{\text{th}}}{N_s} \right), \quad (5.5)
\]

Figure 5.1: Test circuit: 50-stage fanout-4 inverter chain.
Chapter 5. Minimum Energy Point Operation using Supply and Threshold Voltage Scaling

Figure 5.2: Energy and performance contours for a 50-stage inverter chain. Solid line: energy contour. Dashed line: performance contour. Bold line: minimum energy curve.

Here, \( N_s \) is \( n_i \cdot \phi_T \), where \( n_i \) is ideal factor of MOSFETs, which is typically between 1 and 2, and \( \phi_T \) is thermal voltage which is proportional to the chip temperature \( T \). \( \phi_T \) is 26 mV at a room temperature. Since \( n_i \) is less than 2, \( N_s \) in (5.5) is less than 52 mV at a room temperature. The critical path delay \( D \) of the circuit can be modeled using the propagation delay of a single logic gate shown in (2.4), (2.5), (2.6) and (2.7). The critical path delay models are summarized in (5.6), (5.7) and (5.8).

\[
D = \frac{k_3 V_{DD}}{(V_{DD} - V_{th})^\eta}, \quad (V_{DD} \gg V_{th}) \quad (5.6)
\]
\[
D = k_4 V_{DD} \exp \left( -k_5 \frac{V_{DD} - V_{th}}{N_s} - k_6 \left( \frac{V_{DD} - V_{th}}{N_s} \right)^2 \right), \quad (V_{DD} \approx V_{th}) \quad (5.7)
\]
\[
D = k_7 V_{DD} \exp \left( -\frac{(V_{DD} - V_{th})}{N_s} \right), \quad (V_{DD} \leq V_{th}) \quad (5.8)
\]

Parameters from \( k_3 \) to \( k_7 \) are the fitting coefficients determined by the process technology and the architecture of the target circuit. Note that although parameters \( k_3, k_4 \) and \( k_7 \) in (5.6), (5.7) and (5.8) are similar to parameters \( k_3, k_4 \) and \( k_7 \) in (2.5), (2.6) and (2.7), they are different.

5.2.1 Necessary Condition for Minimum Energy Point Operation

Figure 5.2 shows energy and performance contours as a function of \( V_{DD} \) and \( V_{th} \) for the inverter chain in a 28-nm process technology. The performance (i.e., frequency) value is the inverse of the delay of the circuit. The minimum energy point is determined by a pair
of $V_{DD,\text{opt}}$ and $V_{\text{th,\text{opt}}}$, where the energy consumption of the circuit is minimized under a specific performance constraint. It can be found at a tangent point of an energy contour and a performance contour. Thus, at the minimum energy point, the following equation holds as a necessary condition.

\[
\frac{\partial E_d}{\partial V_{DD}} + \frac{\partial E_s}{\partial V_{DD}} = -\frac{\partial D}{\partial V_{DD}}, \tag{5.9}
\]

The left-hand side and the right-hand side of (5.9) represent gradients of an energy contour and a performance contour at the minimum energy point, respectively.

### 5.2.2 Minimum Energy Curve in Sub-threshold Region

If we look at a sub-threshold region where $V_{DD}$ is less than or equal to $V_{\text{th}}$, the ON current increases exponentially as $V_{\text{th}}$ decreases. Thus the delay $D$ decreases exponentially as shown in (5.8). The leakage current exponentially increases as $V_{\text{th}}$ decreases as shown in (5.5), and offsets the decrease in the delay $D$. Therefore $E_s$ does not change in the sub-threshold region [79]. This characteristic is also seen in (5.10) which can be obtained by substituting (5.8) into (5.5).

\[
E_s = k_2 k_7 V_{DD}^2 \exp \left( -\frac{V_{DD}}{N_s} \right), \tag{5.10}
\]

Multiplying (5.9) by $\left( \frac{-\partial E_d}{\partial V_{\text{th}}} - \frac{\partial E_s}{\partial V_{\text{th}}} \right) \frac{\partial D}{\partial V_{\text{th}}}$ gives us the following relationship.

\[
\left( \frac{\partial E_d}{\partial V_{DD}} + \frac{\partial E_s}{\partial V_{DD}} \right) \frac{\partial D}{\partial V_{\text{th}}} = \left( \frac{\partial E_d}{\partial V_{\text{th}}} + \frac{\partial E_s}{\partial V_{\text{th}}} \right) \frac{\partial D}{\partial V_{DD}}. \tag{5.11}
\]

Since $E_d$ and $E_s$ do not depend on $V_{\text{th}}$ in the sub-threshold region, $\frac{\partial E_d}{\partial V_{\text{th}}}$ and $\frac{\partial E_s}{\partial V_{\text{th}}}$ are zero in the sub-threshold region. On the other hand, the delay $D$ depends on the threshold voltage ($V_{\text{th}}$). Therefore, $\frac{\partial D}{\partial V_{\text{th}}} \neq 0$ holds for all $V_{\text{th}}$s in the sub-threshold region. Thus, (5.11) can be converted into (5.12).

\[
\left( \frac{\partial E_d}{\partial V_{DD}} + \frac{\partial E_s}{\partial V_{DD}} \right) \frac{\partial D}{\partial V_{\text{th}}} = 0 \iff \frac{\partial E_d}{\partial V_{DD}} + \frac{\partial E_s}{\partial V_{DD}} = 0. \tag{5.12}
\]

By partially differentiating (5.4) and (5.10) with respect to $V_{DD}$, the following relationship
Chapter 5. Minimum Energy Point Operation using Supply and Threshold Voltage Scaling

\[ y = \frac{k_1}{k_2 k_7} \exp \left( \frac{V_{DD}}{N_s} \right) \]

\[ y = \frac{V_{DD}}{2N_s} - 1 \]

\[ 2N_s \]

\[ V_{DD, opt} \]

\[ y = V_{DD} \]

\[ y = \frac{V_{DD}}{2N_s} - 1 \]

\[ 2N_s \]

\[ V_{DD} \]

\[ V_{DD, opt} \]

Figure 5.3: Minimum energy points in sub-threshold region.

is obtained.

\[ \frac{\partial E_d}{\partial V_{DD}} = 2k_1 V_{DD} = \frac{2E_d}{V_{DD}}, \quad (5.13) \]

\[ \frac{\partial E_s}{\partial V_{DD}} = 2k_2 k_7 V_{DD} \exp \left( -\frac{V_{DD}}{N_s} \right) - \frac{k_2 k_7 V_{DD}^2}{N_s} \exp \left( -\frac{V_{DD}}{N_s} \right) \]

\[ \frac{2E_s}{V_{DD}} - \frac{E_s}{N_s}. \quad (5.14) \]

Therefore, (5.12), (5.13) and (5.14) give us the following equation which holds as a necessary condition for the minimum energy point operation in the sub-threshold region.

\[ \frac{E_d}{E_s} = \frac{V_{DD}}{2N_s} - 1. \quad (5.15) \]

On the other hand, the \( E_d/E_s \) ratio is also simply calculated as \( k_1/k_2 k_7 \cdot \exp (V_{DD}/N_s) \) from (5.4) and (5.10). Therefore, the following equation holds in the sub-threshold region.

\[ \frac{E_d}{E_s} = \frac{k_1}{k_2 k_7} \exp \left( \frac{V_{DD}}{N_s} \right) = \frac{V_{DD}}{2N_s} - 1. \quad (5.16) \]

Since the middle and the right-hand side of (5.16) are exponential and linear to \( V_{DD} \), respectively, there are two \( V_{DD}s \) which satisfy (5.16) as shown in Fig. 5.3. However, the lower one is about \( 2N_s \) which is less than 102 mV in the target process technology at a room temperature. This is too small for a supply voltage to ensure a sufficient noise
5.2. Modeling for Minimum Energy Point Operation

Figure 5.4: Minimum energy curve of a circuit designed with a 28-nm process technology.

margin. If we do not adopt this low $V_{\text{DD}}$, there is only one practical $V_{\text{DD}}$ which satisfies (5.16). Therefore, the $V_{\text{DD,opt}}$ in the sub-threshold region is stacked at a specific value. This is the reason why the minimum energy curve in the sub-threshold region is horizontal (see Fig. 5.4). The energy contour and the performance contour are convex and linear in a practical voltage region as shown in Fig. 5.2, respectively. Thus, there is only one tangent point for the two contours in the practical voltage region. Therefore, (5.16) is a necessary and sufficient condition for the minimum energy point operation in the sub-threshold region.

5.2.3 Minimum Energy Curve in Near-threshold Region

The curve in the near-threshold region where $V_{\text{DD}}$ is slightly higher than $V_{\text{th}}$ is similar to that in the sub-threshold region, which is almost horizontal (see Fig. 5.4). This is because both the leakage power and the delay are exponential to $V_{\text{th}}$ as shown in (5.5) and (5.7). As a result, the effects of the leakage power and the delay on the static energy consumption are mostly canceled each other. Therefore, the characteristics of the energy consumption and the delay in the near-threshold region are similar to those in the sub-threshold region. However, as the overdrive voltage (i.e., $V_{\text{DD}} - V_{\text{th}}$) increases, the effect of strong inversion in MOSFETs gets stronger, which makes the minimum energy curve steeper as shown in Fig. 5.4.
5.2.4 Minimum Energy Curve in Super-threshold Region

By partially differentiating (5.6) with respect to $V_{DD}$ and $V_{th}$, respectively, the following relationship can be obtained.

$$
\frac{\partial D}{\partial V_{DD}} = \frac{k_3}{(V_{DD} - V_{th})^\alpha} - \frac{\alpha k_3 V_{DD}}{(V_{DD} - V_{th})^{\alpha+1}}
= \frac{-\alpha V_{DD} + (V_{DD} - V_{th})}{V_{DD} (V_{DD} - V_{th})} \cdot k_3 V_{DD}
= \frac{-\alpha V_{DD} + (V_{DD} - V_{th})}{V_{DD} (V_{DD} - V_{th})} D,
$$

(5.17)

$$
\frac{\partial D}{\partial V_{th}} = \frac{\alpha k_3 V_{DD}}{(V_{DD} - V_{th})^{\alpha+1}}
= \frac{\alpha}{V_{DD} - V_{th}} \cdot k_3 V_{DD}
= \frac{\alpha}{V_{DD} - V_{th}} D.
$$

(5.18)

From (5.18), $D = \frac{V_{th} - V_{th}}{\alpha} \frac{\partial D}{\partial V_{th}}$ can be obtained. Therefore, substituting $D = \frac{V_{th} - V_{th}}{\alpha} \frac{\partial D}{\partial V_{th}}$ into (5.17) gives us the following relationship.

$$
-\frac{\partial D}{\partial V_{DD}} = \frac{\alpha V_{DD} - (V_{DD} - V_{th})}{\alpha V_{DD}} \frac{\partial D}{\partial V_{th}}.
$$

(5.19)

Like the discussion in Subsection 5.2.2, $\frac{\partial E_s}{\partial V_{DD}}$, $\frac{\partial E_s}{\partial V_{th}}$, $\frac{\partial E_d}{\partial V_{th}}$ and $\frac{\partial E_s}{\partial V_{th}}$ can be obtained using (5.4) and (5.5).

$$
\frac{\partial E_d}{\partial V_{DD}} = 2k_1 V_{DD} = \frac{2E_d}{V_{DD}},
$$

(5.20)

$$
\frac{\partial E_d}{\partial V_{th}} = 0,
$$

(5.21)

$$
\frac{\partial E_s}{\partial V_{DD}} = k_2 \frac{\partial D}{\partial V_{DD}} \exp\left(-\frac{V_{th}}{N_s}\right) + k_2 D \exp\left(-\frac{V_{th}}{N_s}\right)
= \frac{\partial D}{\partial V_{DD}} E_s + \frac{E_s}{V_{DD}},
$$

(5.22)

$$
\frac{\partial E_s}{\partial V_{th}} = k_2 \frac{\partial D}{\partial V_{th}} V_{DD} \exp\left(-\frac{V_{th}}{N_s}\right) - k_2 \frac{V_{DD}}{N_s} \exp\left(-\frac{V_{th}}{N_s}\right)
= \frac{\partial D}{\partial V_{th}} E_s - \frac{E_s}{N_s}.
$$

(5.23)

By substituting (5.19) into the right-hand side of (5.9), the right-hand side can be expressed using $\alpha$, $V_{DD}$ and $V_{th}$ only. In the same way, from (5.17), (5.18), (5.20), (5.21),
(5.22) and (5.23), the left-hand side can also be expressed using $\alpha$, $V_{DD}$, $V_{th}$, $N_s$, $E_d$ and $E_s$ only. By the substitutions, (5.9) can be converted to the following equation which holds as a necessary condition for the minimum energy point operation in the super-threshold region.

$$\frac{E_d}{E_s} = \frac{\alpha V_{DD} - (V_{DD} - V_{th})}{2N_s\alpha} - \frac{1}{2^2}. \tag{5.24}$$

Note that (5.24) is a similar expression to (5.2) presented in [22]. However, unlike (5.2), (5.24) indicates that the ratio $E_d/E_s$ is not constant, but depends on $V_{DD}$ and $V_{th}$ (i.e., the coordinate of MEPs). $E_d/E_s$ also depends on $N_s$ which is proportional to the chip temperature ($T$). Based on (5.24), for typical parameters of CMOS circuits, the $E_d/E_s$ ratio changes in a range between 2 and 7 in the super-threshold region. For example, if we set the performance constraint of the target circuit as 500 ps, $V_{DD,opt}$ and $V_{th, opt}$ are 1.0 V and 0.18 V, respectively (see Fig. 5.2). In this condition, the ratio of the static energy consumption to the total energy consumption ($E_s/E$) is about 20%. On the other hand, if we increase the constraint from 500 ps to 10 ns, $V_{DD, opt}$ and $V_{th, opt}$ shift to 0.35 V and 0.25 V, respectively. As a result, the value of $E_s/E$ increases to about 27%.

Multiplying (5.24) by $E_s/V_{DD}$ gives us the following relationship.

$$\frac{E_d}{V_{DD}} = \left(\frac{\alpha - (1 - V_{th}/V_{DD})}{2N_s\alpha} - \frac{1}{2V_{DD}}\right)E_s. \tag{5.25}$$

According to (5.4) and (5.5), the left-hand side of (5.25) is a linear function of $V_{DD}$ while the right-hand side is an exponential function of $V_{th}$. For satisfying the converted necessary condition (5.25) for all the MEPs in super-threshold region, the change of $V_{DD}$ should be much smaller than that of $V_{th}$. This is the reason why the minimum energy curve in the super-threshold region is vertical.

Similar to the discussion in Subsection 5.2.2, (5.24) is a necessary and sufficient condition for the minimum energy point operation in the super-threshold region. If we set the performance constraint as $D_0$, (5.24) can be converted to (5.26).

$$\frac{k_1 V_{DD}}{k_2 D_0 \exp\left(-\frac{V_{th,D_0}}{N_s}\right)} = \left(\frac{\alpha - 1}{2N_s\alpha} + \frac{V_{th,D_0}}{V_{DD}}\right) - \frac{1}{2^2}. \tag{5.26}$$

where $V_{th,D_0}$ is a threshold voltage when the circuit delay $D$ is $D_0$. Note that we can obtain $V_{th,D_0}$ as shown in (5.27) and can thus approximate that $V_{th,D_0}$ is a linear function of $V_{DD}$.

$$V_{TH,D_0} = V_{DD} - \left(\frac{k_3}{D_0} V_{DD}\right)^{\frac{1}{b}}. \tag{5.27}$$
Chapter 5. Minimum Energy Point Operation using Supply and Threshold Voltage Scaling

Similar to the discussion in Subsection 5.2.2, the left-hand side of (5.26) shows the exponential trends while the right-hand side of (5.26) is a linear function. Therefore, there are two $V_{DD}$s which satisfy (5.26) as shown in Fig. 5.5. However, the lower one is not a practical value since it is typically in the range between $2N_s$ and $4N_s$. Therefore, if we do not adopt this low $V_{DD}$, there is only one practical $V_{DD}$ which satisfies (5.26). Therefore, (5.24) is a necessary and sufficient condition for the minimum energy point operation in the super-threshold region.

### 5.2.5 Impact of DIBL on Minimum Energy Curve

If we simply follow (5.24) expressing the necessary and sufficient condition for minimum energy points in super-threshold region, $V_{th,\text{opt}}$ should always increase as $V_{DD,\text{opt}}$ decreases as shown in dashed curve in Fig. 5.4. However, as the effect of Drain-Induced Barrier Lowering (DIBL) becomes stronger in nanometer technologies, the situation becomes different. DIBL is a short-channel effect in MOSFETs referring to a reduction of threshold voltage of the transistor at higher drain voltages. The effective threshold voltage $V_{th,\text{eff}}$ which is shifted by the DIBL effect can be modeled as (5.28).

$$V_{th,\text{eff}} = V_{th0} - \eta V_{ds},$$  \hspace{1cm} (5.28)

where $V_{th0}$ is the zero bias threshold voltage, $V_{ds}$ is the drain voltage, and $\eta$ is the DIBL coefficient which is typically on the order of 0.1. Since the horizontal axis of Fig. 5.4
5.2. Modeling for Minimum Energy Point Operation

Figure 5.6: Minimum energy curves for different temperature and activity.

shows \( V_{th0} \), the minimum energy curve shifts rightward from the optimal positions of the effective threshold voltage \( (V_{th,eff}) \) at a high \( V_{DD} = V_{ds} \) region. This is the reason why the \( V_{th,\text{opt}} \) slightly decreases as \( V_{DD,\text{opt}} \) decreases at a high supply voltage region contrary to the expected minimum energy curve in the super-threshold region derived from the necessary and sufficient condition (5.24).

5.2.6 Impacts of Temperature and Activity Factor

As described in [22], the location of MEPs depends on a temperature and an activity factor. As shown in Fig. 5.6, the locations of the minimum energy curve also depend on a temperature and an activity factor of the circuit. The activity factor is the probability that the circuit node transitions from 0 to 1. The dynamic energy consumption is proportional to the activity factor and the static power is exponential to the temperature. For example, when the activity factor decreases from 0.1 to 0.01, the minimum energy points shift upper right along the performance contours. Similarly, when the temperature increases from 25 to 125 degree, the minimum energy points also shift upper right along the performance contours. Although the location of the minimum energy curve may be shifted depending on the activity factor and the temperature, the shape of the curves are quite similar to each other.

From (5.24), we can show that the MEPs in the super-threshold region shift upper right when the activity factor decreases or a chip temperature increases. As described above, the
$E_d/E_s$ ratio decreases when the activity factor decreases or a chip temperature increases, which makes the value of $k_1 V_{DD}/k_2 D_0 \exp (-V_{TH,D0}/N_s)$ in Fig. 5.5 smaller. Therefore the intersection point of the two lines in Fig. 5.5 (i.e., MEP) shifts rightward. This fact indicates that $V_{DD}$ and $V_{th}$ increase and thus the MEP shifts upper right. Similarly, we can show that the MEPs in the sub-threshold region shift upper right from (5.16) and Fig. 5.3.

### 5.2.7 Summary of Properties

The analytical properties derived in the previous subsections are summarized as follows.

**Property1:** Minimum energy curves are

- vertical in the super-threshold region ($V_{DD} \gg V_{th}$), and
- horizontal in the sub-threshold region ($V_{DD} \leq V_{th}$).

**Property2:** Minimum energy curves shift upper right when a chip temperature increases or an activity factor decreases.

**Property3** (Necessary and Sufficient Condition for MEP Operation): The pair of $V_{DD}$ and $V_{th}$ determines an MEP if and only if it satisfies (5.29).

$$
\frac{E_d}{E_s} = \begin{cases} 
\frac{\alpha V_{DD} - (V_{DD} - V_{th})}{2N_s \alpha} & (V_{DD} \gg V_{th}) \\
\frac{V_{DD}}{2N_s} - 1 & (V_{DD} \leq V_{th})
\end{cases}
\tag{5.29}
$$

These three properties give the designers intuitions for energy-efficient CMOS circuits. For example, Property1 implies that the number of $V_{DD}$s or $V_{th}$s can be reduced for the minimum energy point operation. In the super-threshold region, a minimum energy curve of a circuit is vertical. Therefore it implies that tuning only $V_{DD}$ with a fixed $V_{th}$ maximizes the energy efficiency. Property2 implies that if a chip has several macros with different activity factors, designing them using transistors with different $V_{th}$s can effectively improve the energy efficiency. Property3 implies that i) a dynamic energy consumption, ii) leakage current, and iii) a chip temperature are important parameters for finding an MEP. By monitoring these parameters and tuning $V_{DD}$ or $V_{th}$ to satisfy (5.29), we can guarantee the minimum energy point operation of a target chip.

Note that (5.29) does not show the optimum $E_d/E_s$ ratio for the near-threshold region since it is hard to derive the optimum $E_d/E_s$ ratio in a closed form function of $V_{DD}$ and $V_{th}$. Therefore, another approach is required to guarantee the minimum energy point
operation in the near-threshold region. Note that MOSFETs in the near-threshold region have the intermediate characteristics between those in the sub-threshold region and the super-threshold region. Therefore, the optimum $E_d/E_s$ ratio in the near-threshold region can be estimated with an acceptable error using “(5.29) ($V_{DD} \leq V_{th}$)” and “(5.29) ($V_{DD} \gg V_{th}$).”

One of the most practical approaches is to estimate the optimum $E_d/E_s$ ratio by extrapolating “(5.29) ($V_{DD} \gg V_{th}$)” to the near-threshold region. However, as $V_{DD}$ is scaled down and approaches $V_{th}$, the effect of weak inversion in MOSFETs gets stronger, which leads to a modeling error of the optimum $E_d/E_s$ ratio. As a result, the discontinuity of two equations in (5.29) gets 1/2 at $V_{DD} = V_{th}$. This can be observed by substituting $V_{DD} = V_{th}$ into “(5.29) ($V_{DD} \gg V_{th}$)” and “(5.29) ($V_{DD} \leq V_{th}$).” The discontinuity at $V_{DD} = V_{th}$ (i.e., 1/2), however, does not have a significant impact on the minimum energy operation. For example, if the circuit finds the MEP based on the estimated $E_d/E_s$ ratio with the error of 1/2, it suffers from the energy overhead compared with the actual minimum energy point operation. Subsection 5.3.2 in this thesis shows that an energy overhead is +1% at the worst case even if the $E_d/E_s$ error is more than 1/2.

Another approach is to extrapolate both “(5.29) ($V_{DD} \gg V_{th}$)” and “(5.29) ($V_{DD} \leq V_{th}$)” to the near-threshold region. The two $E_d/E_s$ ratios have the same value when $V_{DD}$ is $V_{th} + N_s\alpha$. This can be observed by substituting $V_{DD} = V_{th} + N_s\alpha$ into “(5.29) ($V_{DD} \gg V_{th}$)” and “(5.29) ($V_{DD} \leq V_{th}$).” Therefore, the extrapolated optimum $E_d/E_s$ ratio can be derived as a continuous function of $V_{DD}$ and $V_{th}$ by extending the two domains of “(5.29) ($V_{DD} \gg V_{th}$)” and “(5.29) ($V_{DD} \leq V_{th}$)” to $V_{DD} \geq V_{th} + N_s\alpha$ and $V_{DD} \leq V_{th} + N_s\alpha$, respectively. Since there are no discontinuities in the extrapolated $E_d/E_s$ ratio, the modeling error when $V_{DD}$ approaches $V_{th}$ is mitigated compared with the previous extrapolation method.

5.3 Silicon Measurements

In this section, the analytical properties described in Section 5.2 are validated through measurement results of the RISC processor fabricated in a 65-nm process technology.

5.3.1 Test Chip Architecture

The test chip is a 32-bit, 5-stage pipelined RISC processor designed with a 65-nm process technology. The photograph of the processor is shown in Fig. 5.7. The following on-chip memories are implemented in the processors.

- 4 kB instruction cache (I-Cache),
Figure 5.7: Chip photograph of the fabricated RISC processor in a 65-nm process technology.

- 8 kB instruction scratch pad memory (I-SPM), and
- 16 kB data scratch pad memory (D-SPM).

Standard-Cell Memories (SCMs) utilizing minimum height standard-cells proposed in Chapter 4 are utilized to design these on-chip memories. Figure 5.8 shows the structure of the SCM. The SCM structure is similar to that presented in Chapter 4, but it is not identical. Different points are summarized below:

- The SCM is not a dual port memory but a single port memory.
- Clock gating circuits and demultiplexers are not implemented in write circuitry.
- 2-to-1 MUX trees utilizing clocked-inverter-based MUX2 cells are implemented in readout circuitry instead of AOI22-NAND2-NOR2 readout trees.
- A 1-bit D-Latch cell is implemented to each bit cell. 4-bit clock-shared latches are not implemented.
- The SCM still utilizes minimum height standard-cells with a 5.5-track height. The logic cells utilized in the SCM are summarized in Table 5.1.
5.3. Silicon Measurements

![Diagram of SCM structure]

Figure 5.8: The SCM structure.

Table 5.1: 5.5-track minimum height standard cell library in the target 65-nm FD-SOI process technology.

<table>
<thead>
<tr>
<th>Logic</th>
<th>Drive Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>INV</td>
<td>1X, 2X, 4X, 8X, 16X, 32X</td>
</tr>
<tr>
<td>NAND2</td>
<td>1X, 2X</td>
</tr>
<tr>
<td>NOR2</td>
<td>1X, 2X</td>
</tr>
<tr>
<td>Clocked-inverter-based MUX2</td>
<td>1X</td>
</tr>
<tr>
<td>1-bit D-LATCH for peripheral circuits and bit cells</td>
<td>1X</td>
</tr>
</tbody>
</table>

The memories implemented in the processor consist of several subblocks of SCMs with a 2 kB capacity. Each SCM with a 2 kB capacity achieves the area-efficiency of $8.46 \, \mu m^2$ per bit including peripheral circuits. Note that $8.46 \, \mu m^2$ corresponds to $846 F^2$ in this process technology ($F = 100 \, \text{nm}$), which is almost the same area-efficiency to that in Ref. [26] (i.e., “SCM3” in Table 4.2).

The target circuit in this thesis is the SCM in the processor. Based on the measurement results of the SCM, this thesis validates the analytical properties described in Section 5.2. In order to measure the SCM’s power consumption over a wide performance region, the following structure is implemented.

- The supply voltages of SCMs and the other circuits in the processor (logic circuits in the following) are separated. Therefore the SCMs’ power consumption can be directly measured. In this thesis, the SCMs’ supply voltage is set as the same value to that of logic circuits.
Chapter 5. Minimum Energy Point Operation using Supply and Threshold Voltage Scaling

- In SCM macros, the back gate voltage of nMOS and pMOS transistors can be tuned independently of those of logic circuits. In this thesis, the back gate voltage $V_{BB}$ is defined as the voltage condition that the back gate voltage of nMOS and pMOS transistors are set as $V_{BB}$ and $V_{DD} - V_{BB}$, respectively.

- For logic circuits, only back gate voltage of pMOS transistors can be tuned due to the design simplicity of the processor. In this thesis, the same back gate voltage is applied to pMOS transistors in both logic circuits and SCMs.

The processor is designed so that SCMs occupy about 95% of the critical path delay of the processor. Therefore the maximum operating frequency (Fmax) of the processor can be approximated as the Fmax of the SCMs with acceptable approximation errors. Thus, the SCMs’ performance is evaluated through the conventional Fmax test of a processor. As a testbench program, a Discrete Cosine Transform (DCT) loop program is used. A main memory with the DCT program is interconnected with the processor. As for the energy consumption, the SCM’s energy consumption is measured in the following way.

- The static energy $E_s$ can be obtained by measuring the power dissipated in the circuit when all the clock activities are disabled.

- By measuring the power dissipation when the processor repeatedly executes a DCT program at the Fmax clock frequency, the total energy consumption which corresponds to $E_s + E_d$ is measured. By subtracting $E_s$ from the total energy consumption, the dynamic energy consumption $E_d$ of the SCM is measured.

5.3.2 Minimum Energy Curves of Standard-Cell Memories

Energy Reduction by Minimum Energy Point Operation

Figure 5.9 shows the minimum energy curve of the SCM. The horizontal axis is the back gate voltage ($V_{BB}$) of the SCM. Note that $V_{th}$ increases if we move rightward in Fig. 5.9. The solid line is the energy contour of the SCM. Note that its unit is nanojoule per clock cycle. The dashed line is an Fmax contour of the SCM. The bold line labeled “Measured MEP” represents MEPs for various performance constraints (i.e., minimum energy curve). The area labeled “Fmax < 100 kHz or fail” is where the processor operates below a 100 kHz clock frequency or fails to operate. The measurement results show that the processor can operate in the wide supply range down to a 0.3 V single supply voltage with a 1.13 V reverse body bias. Therefore the processor can operate on MEPs in the wide performance range between 391 kHz and 47.5 MHz. If we use typical SRAM
5.3. Silicon Measurements

macros as on-chip memories, the MEPs where a processor can operate are limited due to the SRAM’s stability issue in the low voltage operation. For example, Ref. [17] proposes the voltage-scalable processor with 10T SRAM macros. The minimum operating voltage of the SRAM macros is 0.55 V.

According to Property1, MEP curves are vertical in the super-threshold region while they are horizontal in the sub-threshold region. The measurement results of the SCMs show that the minimum energy curve shown in Fig. 5.9 is vertical in the high-\(V_{\text{DD}}\) region while it is horizontal in the low-\(V_{\text{DD}}\) region, which validates Property1 described in Subsection 5.2.7. The results also indicate that simultaneous tuning of \(V_{\text{DD}}\) and \(V_{\text{BB}}\) is essential for the MEP operation if the target LSI circuits are required to operate over a wide operating performance range from the super-threshold region to the sub-/near-threshold region. If only DVFS is applied to LSI circuits, designers suffer from the energy overhead compared with the MEP operation. To check the fact mentioned above, let us compare the energy consumption when we apply the following three voltage scaling strategies to the SCMs:

1. Only DVFS is applied to the SCMs. \(V_{\text{BB}}\) is fixed at \(-0.4\) V. The SCMs operate at the MEP only when \(F_{\text{max}}\) is 20 MHz.

![Figure 5.9: Minimum energy curve of the SCM. Solid line: energy contour [nJ/cycle]. Dashed line: \(F_{\text{max}}\) contour. Bold line: minimum energy curve.](image-url)
Chapter 5. Minimum Energy Point Operation using Supply and Threshold Voltage Scaling

Figure 5.10: Energy consumption for various operating frequencies. \(V_{BB}\) is fixed at \(-0.4\) V for the conventional DVFS technique.

2. Only DVFS is applied to the SCMs. \(V_{BB}\) is fixed at \(-0.97\) V. The SCMs operate at the MEP only when \(F_{max}\) is around 2 MHz.

3. Both DVFS and ABB are applied to the SCMs. The SCMs always operate at the MEPs over a wide operating performance range.

If we apply strategies 1 and 2 to the SCMs, the SCMs operate on the two dotted lines in Fig. 5.9, respectively. Figure 5.10 shows the measurement results when we use strategies 1 and 3. The horizontal axis is \(F_{max}\) of the target SCMs. The vertical axis is the energy consumption per clock cycle. If \(F_{max}\) is more than 20 MHz, the dotted line on the left side in Fig. 5.9 is quite close to the MEP curve. Therefore, the energy overhead using DVFS only is considerably small compared with the MEP operation. However, as \(F_{max}\) decreases, the shape of the MEP curve becomes horizontal. As a result, the distance between the two lines increases. If we look at the 2 MHz \(F_{max}\), the energy overhead increases to 44%. In the same way, Fig. 5.11 shows the measurement results when we use strategies 2 and 3. If \(F_{max}\) is around 2 MHz, the two strategies achieve almost the same energy-efficiency. However, if \(F_{max}\) increases to 40 MHz, the energy overhead increase to 26%.

Validation of the Necessary and Sufficient Condition

In order to validate Property3, let us consider the \(E_d/E_s\) ratio of the SCMs. Figure 5.12 shows the \(E_d/E_s\) ratio on MEPs. Horizontal axis is the performance constraint of the processor. Vertical axis is the \(E_d/E_s\) ratio when the SCM operates on an MEP. The solid
5.3. Silicon Measurements

Figure 5.11: Energy consumption for various operating frequencies. $V_{BB}$ is fixed at −0.97 V for the conventional DVFS technique.

Figure 5.12: $E_d/E_s$ ratio on MEPs.

line labeled “Measured $E_d/E_s$” is the measured $E_d/E_s$ ratio on MEPs. The dashed line labeled “(5.29) ($V_{DD} \gg V_{th}$)” is the estimated $E_d/E_s$ using (5.29) ($V_{DD} \gg V_{th}$). Note that we substitute the coordinate of measured MEP (i.e., crosses in Fig. 5.9) to (5.29) in order to obtain the dashed curve. In order to obtain $V_{th}$ in (5.29), we use the constant current method by the transistor-level simulation. The parameters $\alpha$ and $N_s$ in (5.29) are extracted through DC simulation of MOSFETs in the target process technology. Similarly, the dashed line labeled “(5.29) ($V_{DD} \leq V_{th}$)” is the estimated $E_d/E_s$ ratio using (5.29) ($V_{DD} \leq V_{th}$). The dotted line labeled “(5.2) (Ref. [22])” represents a result when we approximate the optimum $E_d/E_s$ ratio as a constant. Note that several previous voltage
tuning methods such as [23] guarantee the minimum energy point operation by keeping the $E_d/E_s$ ratio to a fixed value. Therefore, “(5.2) (Ref. [22])” represents a conventional method to find MEPs. As a representative value, the reciprocal number of (5.2) is used.

Figure 5.12 shows that the measured $E_d/E_s$ is not a constant over a wide range of Fmax. As a result, if we estimate the $E_d/E_s$ ratio as a constant like “(5.2) (Ref. [22]),” an estimation error is observed. For example, the estimation error between “(5.2) (Ref. [22])” and the actual measured results is 36% at the worst case. On the other hand, “(5.29) ($V_{DD} >> V_{th}$)” is not a constant and exhibits the same trends to the measured $E_d/E_s$ ratio on various MEPs. As a result, the estimation error is 19% at the worst case, which shows that (5.29) achieves better estimation accuracy than the previous works. The results also show that “(5.29) ($V_{DD} \geq V_{th}$),” which shows the optimum $E_d/E_s$ ratio in low voltage operation, approaches the measured $E_d/E_s$ values as the Fmax decreases. These results indicate that (5.29) is a necessary condition for the minimum energy point operation.

Figures 5.13, 5.14 and 5.15 show $E_d/E_s$ ratios where the SCM operates on the various points over a performance contour of 391 kHz, 8 MHz and 28.57 MHz, respectively. The horizontal axis represents points on each performance contour. Note that although only $V_{DD}$s of these points are plotted, $V_{th}$ also changes as $V_{DD}$ changes to meet the fixed performance constraint. Like the discussion in Subsection 5.2.2 and 5.2.4, $E_d/E_s$ exponentially increases as $V_{DD}$ increases while the right-hand side of (5.29) increases linearly. As a result, there exists one practical intersection point corresponding to the MEP, which vali-
5.3. Silicon Measurements

Figure 5.14: $E_d/E_s$ ratio on 8 MHz Fmax contour.

Figure 5.15: $E_d/E_s$ ratio on 28.57 MHz Fmax contour.

dates Property3 in Subsection 5.2.7. The vertical line labeled “Measured MEP” shows the $V_{DD}$ of an actual MEP of the SCM. The results show that the difference in $V_{DD}$ between the actual MEP and the MEP estimated from (5.29) is 9 mV at the worst case, which shows that we can find MEPs from (5.29) with an acceptable estimation error. For example, if we set the performance constraint as 8 MHz, the MEP estimated from (5.29) (i.e.,
Chapter 5. Minimum Energy Point Operation using Supply and Threshold Voltage Scaling

Energy consumption during a DCT loop
Energy consumption when all the clock activities are disabled

Total energy

\[ \alpha M \]

\[ \alpha M E_d \]

\[ E_s \]

\[ \alpha M E_d E_s \]

Figure 5.16: Definition of the parameter \( \alpha M \).

the intersection point) is about 519 mV (see Fig. 5.14). If the chip operates in this voltage condition, the energy overhead is less than 1% in comparison with the actual minimum energy point operation for the 8 MHz performance constraint. The results show that the parameters \( E_d \) and \( E_s \) are an important parameter to find MEPs of a circuit.

5.3.3 Impacts of Activity Factor

Since clock gating circuits are not implemented to the fabricated SCM, the clock tree in the SCM consumes a dynamic energy in every clock cycle. The clock tree occupies a large part of the total dynamic energy consumption (\( E_d \)) of the SCM. Therefore, the activity factor of the SCM does not strongly depend on the benchmark program. Note that typical on-chip memories have clock gating circuits. Therefore their activity factor strongly depends on the benchmark program. For example, when a memory-intensive application is running, their activity factor is considerably higher than that for just running a NOP loop.

In order to evaluate the activity factor dependency of MEPs, this thesis introduces the coefficient \( \alpha M \) which intuitively represents the activity factor of the SCM. In this experiment, we synthetically modify the dynamic energy by multiplying the original \( E_d \) by \( \alpha M \). For example, if we analyze the MEP when the SCM is running with an activity factor of 10%, we use a 0.1 for the \( \alpha M \). Note that \( k_1 \) in (5.4) is just a parameter depending on an activity factor. The definition of \( \alpha M \) is summarized in Fig. 5.16. The parameter \( \alpha M \) roughly corresponds to the probability that the SCM with clock gating circuits is accessed.
5.4 Voltage Scaling Strategy for Minimum Energy Point Operation

In this section, the voltage scaling strategy for minimum energy point operation is discussed from the derived properties. Property2 shows that the MEP of a circuit depends on a chip temperature and an activity factor. This fact implies that MEPs of microprocessors depend on the target application program. It is also true that MEPs of microprocessors gradually change due to the temperature variation even if they execute the

Figure 5.17: Minimum energy curve of the SCM for \( M = 0.1 \). Solid line: energy contour [nJ/cycle]. Dashed line: Fmax contour. Bold line: minimum energy curves.

Figure 5.17 shows the minimum energy curves of the SCM when \( M = 0.1 \). The bold line labeled “Measured MEP” corresponds to the actual MEPs of the SCM. If we look at “Measured MEP,” the minimum energy curve shifts upper right in comparison with the minimum energy curve in Fig. 5.9. Note that Fig. 5.9 corresponds to the result when \( M = 1 \). Therefore, the results partly validate Property2 in Subsection 5.2.7 since \( M \) represents the activity factor of the SCM. Figure 5.17 also validates the Property1 when the activity factor of the SCM is low.
same program. Therefore, Property2 implies that tracking MEPs dynamically is essential for energy-efficient LSI circuits. Property3 shows the simple equation (5.29) to check their minimum energy point operation. Property3 also shows that the pair of $V_{DD}$ and $V_{th}$ satisfying (5.29) determines the MEP. Therefore, like the simple algorithm presented in [23, 78], the MEP of microprocessors can be dynamically tracked by utilizing the monitor circuits of i) dynamic energy consumption, ii) leakage current, iii) a chip temperature and iv) a critical path delay. As described in Section 5.1, several papers proposed monitors of these parameters. From iv), the pair of $V_{DD}$ and $V_{th}$ that meets the performance constraint can be found. i), ii) and iii) give us the value of $E_d$, $E_s$ and the chip temperature ($T$). By substituting all the values to (5.29), we can check the minimum energy point operation of the microprocessors. If the pair is not the MEP, the MEP of the microprocessors for the specific performance constraint can be found by iteratively changing $V_{DD}$ and $V_{th}$.

### 5.5 Summary

This thesis derived several analytical properties for the minimum energy points of a CMOS circuit. Property1 shows the shape of the minimum energy curves, which gives designers intuitions to reduce the number of $V_{DD}$s and $V_{th}$s with keeping good energy efficiency of a circuit. Property2 shows that the minimum energy curves move upper right when a chip temperature increases or an activity factor increases. Property3 presents a necessary and sufficient condition of $V_{DD}$ and $V_{th}$ for the minimum energy point operation. The condition shows that the pair of $V_{DD}$ and $V_{th}$ satisfying a simple equation of $E_d$, $E_s$ and $T$ determines an MEP of a circuit and vice versa. Based on these properties, the voltage scaling strategy to find MEPs of a circuit is discussed. Measurement results using SCMs fabricated in a 65-nm process technology validate these properties. This thesis also shows that a processor with SCMs can operate over a wide supply range down to a 0.3 V single supply voltage, enabling it to track MEPs for various performance constraint between 391 kHz and 47.5 MHz. The measurement results show that the MEP operation using simultaneous tuning of $V_{DD}$ and $V_{BB}$ achieves 44% less energy consumption at the best case than the conventional DVFS technique if the target LSI circuits are required to operate over a wide operating performance range.
Chapter 6

Conclusion

6.1 Summary of This Thesis

Energy-efficient LSI circuits are highly required to keep the growth of our information society. Scaling the supply voltage and the threshold voltage is a promising approach to reduce the energy consumption of LSI circuits. The goal of this thesis is thus to minimize the energy consumption of the LSI circuits by voltage scaling technologies. The aggressive voltage scaling has posed designers several severe problems such as the performance variation and functional failures, which makes circuit design strategies different between the nominal voltage operation and the low voltage operation.

In Chapter 3, we discussed the architectural-level circuit design strategies for both the nominal voltage operation and the low voltage operation. Based on the simple performance models of LSI circuits, this thesis showed that random variations in MOSFETs have different impacts on the circuit delay depending on the operating voltage. As a result, this thesis showed that the delay variation follows the Gaussian distribution in the super-threshold voltage while it follows the lognormal distribution in the sub-/near-threshold voltage. Based on the fact, this thesis derives the basic statistical operations including the SUM and MAX operations for estimating the variation impact on the architectural-level circuit design. After that, this thesis derived several theorems and properties that help consider architectural design strategies for both the operating voltages. This thesis analytically derived the averaging effect which insists that the relative delay variation of a path decreases as the number of chained gates in series increases. This thesis also showed that sub-/near-threshold voltage operation has a stronger averaging effect than the super-threshold voltage operation, which implies that the performance gain by pipelining of sub-/near-threshold circuits is less influential than that in the super-threshold voltage. Therefore, designer should be careful in designing deeply pipelined circuits for the low
voltage operation. This thesis also showed that gate upsizing of sub-/near-threshold circuits has an exponential impact on the delay variation compared with the super-threshold circuits, which imply that gate upsizing is a major tuning knob in the low-voltage circuit design. It is also shown that the multiplexer based readout structure achieves less worst case delay than the bit-line-based SRAM readout structure in the sub-/near-threshold voltage operation. Therefore, instead of SRAM macros, utilizing Standard-Cell Memories (SCMs) as on-chip memories for sub-/near-threshold voltage operation is effective in terms of operating speed.

Functional failure of on-chip memories is crucial in the aggressively voltage scaled region. To solve problem, several papers proposed SCM structure that consists of digital standard-cells only. The SCM can thus reduce the custom design effort to the level of fully automated cell based design with keeping their stability even in sub-/near-threshold voltage operation. However, SCM’s area still occupies several times larger than that of full-custom SRAM macros, which directly leads to the loss of the computational power of LSI circuits. In Chapter 4, this thesis proposed Minimum Height Standard-Cells (MH-SCs) which have a minimum possible cell height allowed by the logic design rule of a target technology. A systematic approach to design MHSCs is presented. As a result, the proposed SCM in a 65-nm process technology achieves area efficiency of $6.82 \mu m^2$ per bit ($682 F_2^2$ per bit) which is 20% better than that of the state of the art SCMs. This thesis also presents energy efficient readout and write schemes for reducing dynamic energy consumption. We show that exploiting a low activity readout structure with signal gating instead of the bit-line-based structure is the key to achieve both high area- and energy-efficiency with keeping its operation speed satisfiable for applications such as mobile computing and HPC computing. Post-layout simulation results show that the proposed SCMs achieve 31% less and 21% less energy consumption than the prior-art sub-threshold SRAMs and SCMs in 65-nm process technologies. After that, this thesis shows that an embedded processor integrating SCMs in a 65-nm process technology can operate without functional failures over a wide operating performance range down to the sub-threshold voltage by silicon measurement.

In Chapter 5, a technique dynamically tuning the supply voltage and the threshold voltage over a wide operating performance range is discussed as a post-silicon tuning method. Based on the simple transregional performance modeling of LSI circuits, this thesis derived a simple necessary and sufficient condition for Minimum Energy Point (MEP) operation where the energy consumption of the circuit is minimized under a specific clock frequency. The condition indicates that the pair of the supply voltage and the threshold voltage satisfying a simple equation of the energy consumption and the chip
temperature determines the MEP of the circuit. By monitoring these parameters, designers can dynamically find the MEP of the circuit over a wide performance operating range. The condition enables for LSI circuits to always operate on the MEP which dynamically changes depending on environment and applications. Measurement results using SCMs in a 65-nm process technology validate the necessary and sufficient condition. They also show that simultaneous tuning of $V_{\text{DD}}$ and $V_{\text{BB}}$ effectively reduces the energy consumption if target LSI circuits are required to operate over a wide operating performance range. The results show that the MEP operation achieves 44% less energy consumption at the best case than the conventional DVFS technique.

### 6.2 Energy Reduction by Standard-Cell Memories with the Minimum Energy Point Operation

In order to demonstrate how the contribution in this thesis is mutually related and how much energy consumption can be reduced by the proposed techniques, let us consider the following embedded processor as a toy example:

- The embedded processor has logic circuits (e.g., program counters, registers and arithmetic logic units) and on-chip memories with sub-threshold SRAMs such as SRAMs presented in [68, 69].
- Supply voltage scaling (i.e., DVFS) is the most practical approach to reduce the energy consumption of the LSI circuits. The supply voltage of the processor is dynamically scaled to improve the energy-efficiency.
- Sub-threshold SRAMs improve the scalability of LSI circuits down to the sub-threshold region. We assume that both the logic circuits and the on-chip memories consume the same amount of energy in the low voltage operation.

The energy consumption of on-chip memories can be reduced by the proposed SCM structure. Chapter 4 revealed that the dynamic energy consumption can be reduced by exploiting low activity readout/write structures instead of the bit-line-based structure. The results show that replacing the sub-threshold SRAMs with the proposed SCMs in Chapter 4 reduces the energy consumption of the on-chip memories by 31%. As a result, the total energy consumption of the embedded processor can be reduced by 16%. Note that $V_{\text{th}}$ optimization (i.e., MEP operation) is not applied to both the sub-threshold SRAMs and SCMs. Chapter 3 revealed that the worst case delays of SRAMs and SCMs are comparable if we target the high-\(\sigma\) worst case delay, which insists that the two on-chip memories
achieve the same operating speed. Chapter 5 revealed the necessary and sufficient condition for the MEP operation. Base on the condition, we can tune $V_{DD}$ and $V_{BB}$ so that the processor operates on MEPs over a wide operating performance range down to the sub-threshold voltage. Measurement results show that the MEP operation can reduce the energy consumption of digital circuits (i.e., the logic circuits and the SCMs in the embedded processor) by 44% at the best case with keeping the operating frequency in comparison with the conventional DVFS technique. Combined with the 16% energy reduction obtained using the SCMs as presented above, the total energy consumption of the embedded processor can be reduced by 53% at the best case. Although the discussion is a toy example, the results show that the proposed techniques presented in this thesis play an important role on reducing the energy consumption of the LSI circuits.

6.3 Future Work

There are still remaining challenges in the proposed techniques. One of the problems is that the proposed SCMs still occupy several times larger area than the conventional 6T SRAMs while the SCMs achieve less energy consumption than the 6T SRAMs. In the HPC platform and the mobile computing platform, for example, the area overhead may no longer be acceptable since the area overhead directly leads to the loss of the computational power. To address the problem, the memory hierarchy design with SCMs and SRAMs is essential for the future LSI circuits. Like the discussion about the cache memory, both the high area-efficiency and the high energy-efficiency can be obtained by implementing the SCMs with a small capacity as a lower level cache than the conventional 6T SRAMs with a large capacity. There are several approaches to model area, delay and energy of 6T SRAMs and SCMs [1, 67]. Based on the models, finding the best combination of the SCM capacity and the SRAM capacity is a key to minimize the energy consumption with an acceptable area overhead.

Although this thesis targeted the MEP operation of digital circuits, designers also suffer from the energy overhead of peripheral circuits such as on-chip DC-DC converters, PLLs, I/O cells, other analog circuits and off-chip modules. Since the performance of DC-DC converters (e.g., conversion efficiency, transition time and output voltage resolution) has a large impact on the energy efficiency of the target voltage-scaled digital circuits, considering the energy overhead introduced by the DC-DC converters is essential. Several papers such as Ref. [80] pointed out that the energy overhead of the on-chip voltage regulators is not negligible compared with the energy consumption of the core logic circuits. As a result, the position of MEPs which minimize the energy consumption
of a chip may change depending on whether or not we consider the energy overhead of the voltage regulators. If the chip operate on wrong MEPs without considering the energy overhead of the voltage regulators, the chip suffers from the energy overhead compared with the actual MEP operation. Therefore, developing MEP tracking technique considering peripheral analog circuits further improves the energy-efficiency of LSI circuits.
Bibliography


Threshold Computing: Reclaiming Moore’s Law Through Energy Efficient Inte-

M. Srinivasan, A. Kumar, S. Gb, R. Ramanarayanan, V. Erraguntla, J. Howard,
S. Vangal, S. Dighe, G. Ruhl, P. Aseron, H. Wilson, N. Borkar, V. De, and S. Borkar,
“A 280mV-to-1.2V Wide-Operating-Range IA-32 Processor in 32nm CMOS,” in

CMOS Circuits Operating Near Threshold,” IEEE Transactions on Very Large Scale

Transistors,” IEEE Journal of Solid-State Circuits, vol. 24, no. 5, pp. 1433–1439,

Near-Threshold Chip Multi-Processing,” in International Symposium on Low Power

Scheme with Fine-Grain Body Biasing for Robust and Energy-Efficient Operations,”

[22] K. Nose and T. Sakurai, “Optimization of VDD and VTH for Low-power and High
Speed Applications,” in Asia and South Pacific Design Automation Conference, Jan

lay and Power Monitoring Schemes for Minimizing Power Consumption by Means
of Supply and Threshold Voltage Control in Active and Standby Modes,” IEEE Jour-

Standard-Cell Based Memories in the Sub-VT Domain in 65-nm CMOS Technol-
ogy,” IEEE Transactions on Emerging and Selected Topics in Circuits and Systems,


Publication List

Journal


International Conference


Domestic Conference (in Japanese)


114

BIBLIOGRAPHY


Misc.


5. 塩見 準, 小野寺 秀俊, “ロジック部とメモリ部で独立して電圧制御可能なエネルギー最小点追跡プロセッサ,” in 第 6 回 IEEE SSCS Japan Chapter VDEC Design Award, 2016 年 8 月 (presentation and demonstration).
Awards


2. IPSJ Computer Science Research Award for Young Scientists, “ロジック部およびメモリ部の独立電圧制御によるプロセッサの消費エネルギー最小化,” in 組込み技術とネットワークに関するワークショップ, 2017 年3月.


7. IEEE SSCS Japan Chapter VDEC Design Award, “ロジック部とメモリ部で独立して電圧制御可能なエネルギー最小点追跡プロセッサ,” in IEEE SSCS Japan Chapter VDEC Design Award, 2016 年8月.


