Ultra Energy-Efficient Sub-/Near-Threshold Computing: Platform and Methodology

Zhao Wenfeng

(*B.Eng, Huazhong University of Science and Technology, 2007*)

(*M.Eng, Huazhong University of Science and Technology, 2009*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2013

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Zhao Wenfeng

30 December 2013

Acknowledgements

My postgraduate study would not have been so exciting and memorable without the guidance and help from many people in my life. With this opportunity, I would like to express my sincere gratitude to them.

First of all, I would like to express my appreciation to my supervisor Prof. Ha Yajun, for his guidance, encouragement and trust during my PhD study. I am so grateful that he gave me the valuable study chance and an interesting research topic for the past years. He has been always willing to offer his insights and suggestions whenever I encounter technical problems and I would not be able to finish my research without his encouragement and inspiration. I also thank him for involving me in several collaboration projects beyond my research topic, which also greatly extends my knowledge.

I would like to acknowledge the in-depth comments and feedback from my doctoral committee members, Prof. Heng Chun Huat, Prof. Xu Yongping and Prof. Lian Yong. I would like to appreciate the help from Prof. Massimo Alioto for his rigorous review and discussions for a journal paper.

During my PhD time, I have worked with several group members under Prof. Ha Yajun. I would like to thank Anastacia Alvarez first, for her contribution to several research projects related to my PhD research. I also enjoyed the collabo-

rations with Loke Wei Ting, Dr. Syed Riswan and Dr. Do Thi Thu Trang and discussions with Dr. Yu Heng, Dr. Wang Yi, Li Ang, Hoo Chin Hau, Luo Shaobo and Chen Yongzhen.

It is a pleasant experience in the VLSI lab with the presence of many fellows friends: Zhao Jianming, Liu Xu, Li Xuchuan, Liu Xiayun, Zhou Lianhong, Li Yong Fu, Mahmood Khayatzadeh, Chua Dingjuan and Jerrin Pathrose. Also, I will never forget the afternoon coffee group with Pan Rui and Wu Tong.

Finally, I would like to express my gratitude from the bottom of my heart to my family (my parents and my wife) for their unconditional love and constant support.

List of Abbreviations

ADC	Analog to Digital Converter
AES	Advanced Encryption Standard
AFE	Analog Front End
AOF	Area Overhead Free
ASIC	Application Specific Integrated Circuits
BB	Body Biasing
BSN	Body Sensor Network
CAD	Computer Aided Design
CCS	Composite Current Source
CF	Composite Field
CMOS	Complementary Metal Oxide Semiconductor
CORDIC	COordinate Rotation DIgital Computer
CPU	Central Processing Unit
DCVS	Differential Cascade Voltage Swtich
DIBL	Drain Induced Barrier Lowering
DoE	Design of Experiments
DRAM	Dynamic Random Access Memory
DSTA	Deterministic Static Timing Analysis
DVS	Dynamic Voltage Scaling

ECG	Electrocardiography
EDA	Electronic Design Automation
EDNM	Effective Diode Network Model
FDC	Frequency to Digital Converter
FDSOI	Fully Depleted Silicon on Insulator
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GF	Galois Field
GIDL	Gate Induced Drain Leakage
GPU	Graphic Processing Unit
IC	Integrated Circuits
INWE	Inverse Narrow Width Effect
IoTs	Internet of Things
IP	Intellectual Property
LD	Logic Depth
LS	Level Shifter
LVSB	Low Voltage Swapped Biasing
MC	Monte Carlo
MCU	Micro-Controller Unit
MEP	Minimum Energy Point
MOMCAP	Metal-Oxide-Metal Capacitor
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
MTCMOS	Multiple Threshold CMOS
MVT	Mixed Threshold Voltage
Near-V_{th}	Near-threhsold

PDF	Probability Density Function
PTAT	Proportional to Absolute Temperature
RDF	Random Dopant Fluctuation
RF	Radio Frequency
RSCE	Reverse Short Channel Effect
RVT	Regular Threshold Voltage
S-Box	Substitution Box
SFBB	SelF-Body Biasing
SIMD	Single Instruction Multiple Data
SMA	Surrogate Modeling Adjustment
SoC	System on Chip
SSTA	Statistical Static Timing Analysis
Sub-V$_{th}$	Sub-threshold
TDC	Time to Digital Converter
TSD	Temperature Sensitive Delay
TSRO	Temperature Sensitive Ring Oscillator
ULP	Ultra Low Power
ULV	Ultra Low Voltage
VLSI	Very Large Scale Integration
V$_{th}$	Threshold Voltage
V$_T$	Thermal Voltage
VTC	Voltage Transfer Characteristics
WSN	Wireless Sensor Network
ZBB	Zero Body Biasing

Contents

Acknowledgements **ii**

List of Abbreviations **iv**

Contents **vii**

Summary **xi**

List of Tables **xiv**

List of Figures **xv**

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Thesis Contributions 7

1.3 Organization of the Thesis 9

2 Literature Review **10**

2.1 Modeling and Technology Implications 10

2.2 Circuit Building Blocks 12

2.2.1 Logic 12

2.2.2 SRAM 13
2.2.3 Level Shifter 14
2.3 Circuit/Architecture Techniques and Design Automation Methodologies 15
2.3.1 Circuit Techniques 15
2.3.2 Design Automation Methodologies 16
2.4 SoC Designs 17

3 Near-V_{th} ASIC Design: Statistical Timing Analysis and Performance Boosting 19

3.1 Background and Motivation 20
3.1.1 ULV Timing Analysis Challenges 21
3.1.2 ULV Body Biasing Challenges 22
3.2 Proposed Surrogate Model Adjustment based SSTA (SMA-SSTA) . 23
3.3 Area-Overhead-Free Body-Biasing Techniques 26
3.3.1 Conventional AOF-BB Schemes and Limitations 27
3.3.2 Proposed SelF-Body-Biasing Scheme 28
3.4 Case Study: Advanced Encryption Standard 33
3.4.1 Low-Cost AES Architectures and S-Box Implementation . . 34
3.4.2 Automated and Detailed SMA-SSTA Design Flow 36
3.4.3 Runtime for Local Variation Characterization and Comparison with SSTA 39
3.4.4 Performance and Design Margin Recovery 40
3.4.5 Physical Implementation Considerations 41
3.5 Testchip Measurement Results 41

3.5.1 Performance Measurement and Energy Comparison 42
3.5.2 Static and Dynamic Robustness of the Body Voltage Bias Point in SFBB 45
3.6 Conclusion and Summary 48

4 A 65nm 30.7fJ/bit Subthreshold Level Shifter Design 52
4.1 Introduction 53
4.2 State-of-the-Art Implementations 55
4.3 Proposed Level Shifter Design 57
4.3.1 NMOS-Diode Current Limiter based Level Shifter 57
4.3.2 Level Shifter Optimization with MTCMOS and Subthreshold Sizing 59
4.3.3 Comparative Analysis to Previous Implementations 62
4.4 Measurement Results and Discussions 64
4.5 Conclusion and Summary 67

5 Robust and Energy-Efficient Ultra-Low Voltage Standard Cell Design with Intra-Cell Mixed-V_{th} Methodology 70
5.1 Introduction 71
5.2 Related Work 73
5.2.1 Subthreshold Logic Robustness 73
5.2.2 MVT-LM Design Technique 75
5.3 MVT-ULV: Robustness-Driven Mixed-V_{th} for ULV Operation . . . 78
5.4 Experimental Results: Iso-Area Constraint 81
5.5 Experimental Results: Iso-Yield Constraint 84
5.5.1 Cell Level Evaluation 85

5.5.2 Library Level Evaluation 87
5.6 Conclusion 88

6 Exploring Energy Efficiency in Embedded DRAM 90
6.1 Background 91
6.2 Hidden-Refresh Scheme 93
6.3 Circuit Design for Self-Refresh eDRAM 95
6.3.1 Bitcell Choice and Operation Principle 97
6.3.2 Write/Read Bitline Circuit Design 100
6.3.3 Wordline Driver Circuit Design 101
6.4 Power Metrics of the Hidden-Refresh eDRAM under Voltage Scaling 104
6.5 Conclusions 108

7 A 0.4V 280nW Nearly All-Digital Frequency Reference-less Hybrid Domain Temperature Sensor 110
7.1 Introduction 111
7.2 Ratioed-Current/Delay PTAT Sensor Core 114
7.3 Hybrid Domain Temperature Sensing Scheme 117
7.4 Circuit Implementation Details 118
7.5 Measurement Results 123
7.6 Conclusion 127

8 Conclusion and Future Work 129
8.1 Conclusion 129
8.2 Future Work 131

Bibliography 133

Summary

The demand of low power and energy-efficient computing platforms are the driving force of the modern integrated circuit technologies. Power consumption and energy efficiency are becoming the prior design considerations for mobile computing devices, wireless sensor network, wearable and implantable biomedical devices, etc., where devices are powered by either low volume batteries or energy scavenging devices.

Voltage scaling is one effective way to reduce power consumption. Especially, the aggressive voltage scaling, Ultra-Low Voltage (ULV) sub-/near-threshold operation has been demonstrated as a viable solution for extremely energy-constrained applications, which paves the way for the realization of the above-mentioned power-aware computing platforms. However, this brings about opportunities together with many design challenges that are inevitable for ULV circuits, comprising logic robustness issue, circuit functional yield, as well as excessive energy/area/throughput design margins. There are some existing solutions that want to achieve robust ULV operation, but they come with significant energy/area overhead.

This thesis presents our design methodology and circuit optimization to achieve energy and area efficiency in ULV designs beyond the state-of-the-art solutions.

First, we present a design flow with a novel Surrogate Model Adjustment based

Statistical Static Timing Analysis (SMA-SSTA) framework and a novel SelF-Body-Biasing (SFBB) scheme to reduce the excessive design margins as well as to boost the performance in ULV ASIC designs. A 65nm AES encryption engine test chip is designed under this framework and delivers 12.2Mbps with 1.65pJ/bit at 0.5V, which is 22× and 7.8× over the state-of-the-art AES design, respectively, while reducing silicon area by 28%.

Second, we propose several custom designed ULV circuit blocks. A NMOS-diode current limiter based level shifter is proposed with MTCMOS and INWE-aware sizing techniques. This level shifter achieves 25.1ns delay and 30.7fJ/bit energy in a 65nm technology, outperforming all previous level shifter implementations. For logic design, we propose a robust and energy-efficient intra-cell mixed-V_{th} standard cell design methodology. This methodology identifies the bottleneck devices causing robustness degradation in ULV logic cells and replaces these devices with low threshold voltage devices. Library level comparison shows that on average 30.1% energy efficiency improvement can be achieved through the proposed methodology. For memory design, we explore the eDRAM as a memory alternative to SRAM. We introduce a hidden-refresh scheme to cope with the refresh operation and several circuit techniques to mitigate the requirement of additional power supplies in conventional eDRAM designs. Experiment results confirm that the hidden-refresh eDRAM shows higher density and lower access energy when compared to the SRAM counterpart.

Finally, we demonstrate the design of a 65nm 0.4V 280nW nearly all-digital temperature sensor for wireless sensing platforms. A ratioed-current/delay PTAT sensor core and hybrid domain temperature sensing scheme are proposed to eliminate the dependence on the external frequency reference. The measured eight

temperature sensor test chips show maximum error of -1.6^{o}C/1°C across 0$\sim$100°C range after two-point calibration, with 40 samples/second sample rate at 0.4V.

List of Tables

3.1 Comparison of state-of-the-art Area-Efficient AES architectures . . 35

3.2 Comparison of state-of-the-art normalized S-Box design 36

3.3 Summary and comparison of state-of-art area-efficient AES designs 44

4.1 Summary of the transistor sizing 62

4.2 Comparison to state-of-the-art LS designs 67

5.1 Comparison of SNM (VDD=300mV, 25°C) and VDD$_{min}$ of several logic cells under different logic design methods 83

5.2 Synthesis results of the ITC'99 benchmark circuits. 88

6.1 Comparison among SRAM, eDRAM and Hidden-Refresh eDRAM . 108

7.1 Categories of the CMOS temperature sensor 111

7.2 Summary of state-of-the-art nW temperature sensor 127

List of Figures

1.1 Technology scaling trend of supply voltage. 2

1.2 Scope of the thesis. 7

3.1 Conceptual illustration of (a) feed-forward SSTA, (b) feed-back based SMA-SSTA and (c) detailed flow chart of SMA-SSTA. 23

3.2 Timing derating feature of logic datapath under local variations. . . 26

3.3 (a) Cross-section of the deep N-well technology and, (b) parasitic diode connections of a CMOS inverter under three AOF-BB schemes (ZBB, LVSB, proposed SFBB). 27

3.4 Illustration of the EDNM model of (a) an inverter cell and, (b) a 2-input NAND cell for SFBB scheme. 30

3.5 Illustration of the EDNM model for SFBB scheme of (a) an inverter cell, (b) a NAND cell, (c) a 3-stage ring oscillator, (d) simulation waveform of the self-bias node voltage fluctuation of the 3-stage RO SFBB and, (e) timing potentials of the LVSB and SFBB scheme over the ZBB scheme. 30

3.6 Illustration of the AOFBB well tap designs and body-biasing floor-plan under a tap-less standard cell library technology. 32

3.7 State-of-the-art AES designs: energy vs. throughput and area (scaled to 65 nm node for the different adopted technologies). 34
3.8 The AES engine architecture with native round and key expansion S-Box. 37
3.9 Pareto set plot of DoE for delay sensitivity analysis toward different variation parameters. 38
3.10 $+3\sigma$ setup time accuracy analysis of SMA-SSTA after model adjustment. 38
3.11 Statistical distribution of the delay (normalized to $\mu-3\sigma$ delay under SFBB at 0.6V) for different body biasing schemes and resulting clock frequency improvement over DSTA (top-right). 40
3.12 Die micrograph with annotated core and on-chip testing buffer and summary of the AES encryption engine at room temperature. . . . 42
3.13 Measurement results of operating frequency and energy per bit of the testchip. 43
3.14 Leakage measurements of ZBB, LVSB and SFBB versus V_{DD} under different temperatures in a single die. 47
3.15 Leakage measurements across 15 dice. 47
3.16 Measured self-bias node voltage at 0.5V supply. 49
3.17 Dynamic stability test of the self-bias node. 49

4.1 Conventional DCVS level shifter topology. 55
4.2 Proposed NMOS-diode current limiter based level shifter topology with reduced pull-down device size. 58

4.3 Simulated transient waveform of the MTCMOS DCVS level shifter and the proposed level shifter. 58
4.4 Schematic of the optimization to the Proposed LS with MTCMOS and INWE-aware sizing. 60
4.5 Transient simulation of the INWE effects on NMOS. 60
4.6 Normalized comparison of the delay, energy and energy-delay product of the proposed level shifter with adopted optimization techniques. 61
4.7 Transient simulation of the proposed LS and the previous designs in [57, 58]. 63
4.8 Monte Carlo simulation of the proposed LS and the previous designs in [57, 58]. 63
4.9 Die photo and layout view of the proposed level shifter. 65
4.10 Measured waveform of the proposed LS with a 60mV to 1.2V conversion. 66
4.11 Measured waveform of the proposed LS with a 60mV to 1.2V conversion. 66
4.12 Measured LS delay from a typical die. 68
4.13 Measured statistics of the proposed LS: delay, VDD_{min}, dynamic power and leakage power. 68

5.1 (a) schematics of commercial standard cells and subthreshold logic failure mechanism, (b) cross-coupled NAND/NOR pair and, (c) example of butterfly plot. 74
5.2 NAND/NOR pairs of (a) RVT, (b) LVT, (c) previous MVT-LM technique, and (d) butterfly plot of three pairs. 76

5.3 Previous MVT-LM flip-flop design, (a) Mixed-I, and (b) Mixed-II, and (c) flip-flop with reset function. 76
5.4 (a) MVT-ULV NAND cell design with other possible variant cells, and (b) Monte-Carlo simulation of the NAND-NOR Pair VTC curves. 79
5.5 MVT-ULV AOI21/OAI21 cell design with annotated bottleneck transistors. 80
5.6 MVT-ULV flip-flop cell with asynchronous reset. 80
5.7 Butterfly plots of four sets of 2-input NAND-NOR pairs. 82
5.8 Butterfly plots of variant design of 2-input NAND-NOR pairs. . . . 83
5.9 Temperature effects on logic output swing. 84
5.10 Output swing failure rate of three sets of standard cells, RVT, MVT-LM, MVT-ULV under 27°C and 125°C. 86
5.11 Delay and power distribution of three 10-stage NAND-NOR chains (commercial, upsized and MVT-ULV). 86

6.1 Conceptual illustration of the hidden-refresh scheme for eDRAM. . 94
6.2 Top-level block diagram of the hidden-refresh eDRAM. 96
6.3 Detailed timing diagram of the hidden-refresh eDRAM. 96
6.4 2T gain-cell eDRAM with simplified operation waveform. 98
6.5 Timing diagram of the eDRAM write, read and refresh operation. . 98
6.6 Write/read bitline design of the hidden-refresh scheme. 100
6.7 Bootstrapped WWL driver and simulated waveform. 101
6.8 Illustration of the worst case read disturbance issue of the 2T gain-cell eDRAM. 103
6.9 Proposed tri-state RWL driver. 103

6.10 Schematic of the 1K-bit hidden-refresh eDRAM. 104
6.11 Power consumption of the hidden-refresh eDRAM. 105
6.12 Retention time and static power of the hidden-refresh eDRAM. . . 106
6.13 Read/write power with duty cycled refresh power of the hidden-refresh eDRAM. 106

7.1 Power consumption versus frequency of the state-of-the-art ultra-low power frequency reference for illustration of power overhead due to frequency reference. 113
7.2 Schematic of the proposed ratioed-current/delay PTAT sensor core. 114
7.3 Mathematical background of the operation principles of the proposed current-ratioed PTAT. 115
7.4 Optimal ΔV_{GS} vs. Adjusted-R^2 coefficient. 116
7.5 Timing diagram of the ratioed-current/delay temperature sensor. . 118
7.6 Schematic of the hybrid-domain temperature sensor. 120
7.7 Simulated delay ratio the ratioed-current/delay temperature sensor. 121
7.8 Simulated temperature error of the ratioed-current/delay temperature sensor. 121
7.9 Illustration of time domain (top left), hybrid domain (top right) processing and the hybrid domain processing benefits on TDC bandwidth (bottom left) and dynamic frequency scaling (bottom right, simulated). 122
7.10 Hybrid domain digital processing data format. 123
7.11 Die micrograph with annotated floorplan. 124

7.12 Measured delay ratio (left) and adjusted-R^2 coefficient (right) of 8 chips. 125
7.13 Measured temperature error across 8 dies. 125

Chapter 1

Introduction

1.1 Background and Motivation

The advancements of the solid-state-circuit technologies have empowered the prosperity of the integrated circuit industries for the past 60 years. Technology scaling trend, also known as the Moore 's law [1], is the driving force for pushing the limits of high-performance computing capability and system integration level. In the meanwhile, energy efficiency has been simultaneously improved through successive CMOS technology generations [2], with great contributions from both supply voltage scaling and parasitic reduction due to the shrinking device geometries.

The benefits from both technology scaling and architecture/circuit innovations have profoundly contributed to the diverse application scenarios of the CMOS technologies and the future generations of computing concepts. These emerging applications and concepts will eventually accelerate the formation of the corresponding business models in the economic ecosystems. For instance, battery-operated mobile computing applications are enabled by the technology progress, thereafter

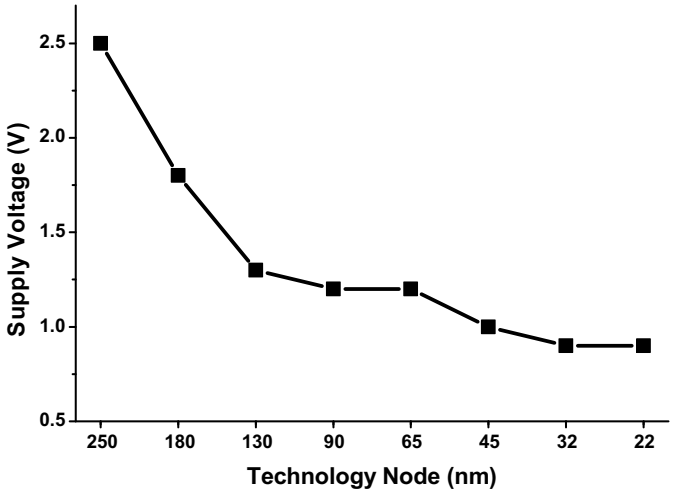


Fig. 1.1: Technology scaling trend of supply voltage.

leading to the rapid market growth of personal computing and mobile devices in recent years, such as smart phones, tablets and ultrabooks, etc.

However, as CMOS device feature size approaches nanometer regime, technology scaling encounters a challenging trend that the supply voltage almost ceases to scale beyond 90nm technology node, as shown in Fig. 1.1. The main reason is due to the dramatic leakage current increase in advanced technology nodes. As a consequence, the constant electric field scaling does not hold valid and the supply voltage scaling solely can hardly provide energy efficiency improvement. In the meanwhile, the transistor density doubles every generation, which leads to significant concerns for power density increase [3].

The intuition is to keep the supply voltage scaling trend all the way to the fundamental limit of the CMOS technology, as described in [4, 5]. This is an effective solution in reducing the dynamic energy due to the quadratic relationship to the supply voltage (CV_{DD}^2) and the leakage power reduction from Drain-Induced Barrier Lowering (DIBL) effect. On the other hand, power reduction of logic gates

also comes with the increased propagation delay due to aggressive voltage scaling. When the required energy for a given task is interested, the situation is altered as the leakage energy becomes remarkably increased at low voltages due to the increased logic delay. As a result, the total energy is dominated by the leakage energy when aggressive voltage scaling is applied, and this leads to the fundamental concept of Minimum Energy Point (MEP) in CMOS digital circuits, which is both analytically modeled [6] and silicon-verified [7]. The supply voltage of MEP point is often found to be around the transistor threshold in modern CMOS technologies, providing more than an order of magnitude energy reduction when compared to that of the nominal supply voltage. Therefore, this technique is also referred to as sub-/near-threshold (sub-/near-V_{th}) operation.

The sub-/near-V_{th} operation definitely provides a viable solution to achieve ultra-low-power consumption and improved energy efficiency beyond the technology scaling capability. Besides, it also provides great opportunities to refine previous design models for low power target or to resolve several challenging issues in emerging applications. For example, one direct observation from sub-/near-V_{th} operation is that further reducing the supply voltage beyond MEP point is less effective as the energy efficiency degrades. In this way, the system level task scheduling for energy minimizations needs to be revised accordingly when compared to the classic dynamic voltage scaling framework [8]. And it is strongly suggested to distinguish the role of energy and power, adopting the correct design strategies for different building blocks with varied duty cycle profiling [9].

Nowadays, several emerging applications are constrained by the power budgets and the energy-efficiency, while the sub-/near-V_{th} operation is a likely candidate for solving these design constraints. We categorize these emerging applications into

two broad scenarios, as listed below.

- Solutions to Dark Silicon with Improved Energy Efficiency

  Thanks to the doubled number of transistors in each technology generation, high performance computing platforms enter the multi-core era but with a nearly constant power budget. In view of this, today's multi-core computing architecture will eventually encounters the "multi-core apocalypse" (i.e., the dark silicon crisis) [10] in the near future. The dark silicon crisis indicates that future generations of multi-core CPUs/GPUs might have degraded performance improvement due to limited number of working cores.

  Near-V_{th} operation achieves a good compromise between the energy and performance to mitigate the dark silicon crisis, which turns the "dark silicon" into "dim silicon" [11] with improved energy-efficiency [12]. Several recent reported computing architectures also exploit the near-threshold operation, such as wide SIMD [13], 3-D many core [14], as well as future heterogeneous computer architecture with general-purpose processors and dedicated accelerators [15].

- Empowering WSN, BSN and Implantable Electronics

  The vision of Internet of Things (IoTs) advocates the connection of thousands of sensor-based computing platforms to form the wireless sensor networks [16], with encouraging future of smart cities/home, ubiquitous environmental monitoring and industrial control, etc. Further extending IoTs to body sensor networks [17] allows remote e-health services like healthcare assessment and diagnosis, fall detection and vital human body signal monitoring (e.g., ECG), etc. In addition, medical implantable electronics offer profound possibilities of

improving patients' life qualities [18]. Early devices like the pacemakers and the cochlear prosthesis devices have already been commercialized and newly-emerged devices like deep brain stimulation, intraocular pressure monitoring and retinal/neural prosthesis are in active research and prototyping period.

The above-mentioned computing platforms, which include the energy delivery subsystems, sensors/analog frond end, digital processing/control circuits and RF subsystems, are condensed into a vanishingly small volume (e.g., millimeter-scale) to cause negligible invasion to either the environment or to human body. The application nature also requires long lifetime, autonomous and even perpetual operation, whereas the form size of such systems brings about the practical challenges between the energy sources with limited power densities (e.g., Lithium batteries, solar-, piezo-/thermo-electric energy harvesting devices) and higher integration level of the system building blocks. Therefore, with the knowledge of the voltage scaling trend in modern IC technology, aggressive energy-aware and energy-efficient design considerations are particularly necessary for the above-mentioned applications.[19–22].

Being a viable solution for achieving ultra-low power/energy in CMOS circuits indeed, the robustness of the ultra-low voltage (ULV) CMOS circuits are however degraded. Therefore, the robust ULV operation becomes the top challenging task, which is generally not a serious issue for super-V_{th} circuit designs. This leads to the fact that the conventional super-V_{th} circuit topologies and design methodologies are not applicable in the ULV domain due to several practical challenges and limitations [9].

First, device characteristics are significantly compromised when the supply voltage approaches to the device threshold. The conduction (on) current degrades

exponentially and causes significant logic delay increase. Therefore, sub-V_{th} operation causes significant performance loss when compared to super-V_{th} counterpart. In addition, the on/off current ratio is also reduced and causes non-ideal, ratioed operation of both sub-V_{th} logic and memory design.

Second, process variabilities introduced by fabrication technologies become one of the major concerns for CMOS integrated circuits, and advanced technologies often comes with higher level intra-die variations due to the shrinking device dimensions. This essentially results in serious concerns for ULV designs as further reducing the supply voltage exacerbates the above-mentioned situations. Both the logic and memory suffer from higher failure rates due to device parameter variations. Meanwhile, with device parameter variations being modeled as Gaussian (i.e., normal) distribution, the logic delay shows non-Gaussian and non-symmetrical distribution with long tail towards the worst-case scenarios, as a consequence of exponential dependence between the on current and the transistor threshold voltage.

Third, the corresponding design methodologies and automated design flows, such as ultra-low voltage IPs and CAD supports are not available at the moment. As a result, efforts have to be made at all design entries, such as low-voltage logic/memory designs, low voltage liberty IPs for synthesis and timing closure as well as the physical design IPs for back-end design solutions. Especially, excessive design margins are inevitable in ULV domains if the conventional corner-based design flow is adopted and variation-aware design flows are especially interested.

During the past decade, the low-power communities have paid considerable attention and efforts to sub-/near-V_{th} designs. Sub-/near-V_{th} circuits have been widely investigated and silicon prototypes are reported from both academia and

industries. Given minimizing the energy as the fundamental target, optimal sub-/near-V_{th} designs are often the trade-off among robustness, energy, area and performance. For instance, ULV logic designs generally require transistor upsizing to mitigate logic failure caused by process variations, and ULV SRAM introduces more transistors into the basic bit-cells to achieve robust read/write operations. In addition, large design margins of the corner-based flow are inevitable as higher gate sizing effort is needed during technology mapping to meet the worst-case corner timing requirement. Eventually, all above facts lead to the increased energy and area, which largely offsets the benefits of sub-/near-V_{th} designs.

1.2 Thesis Contributions

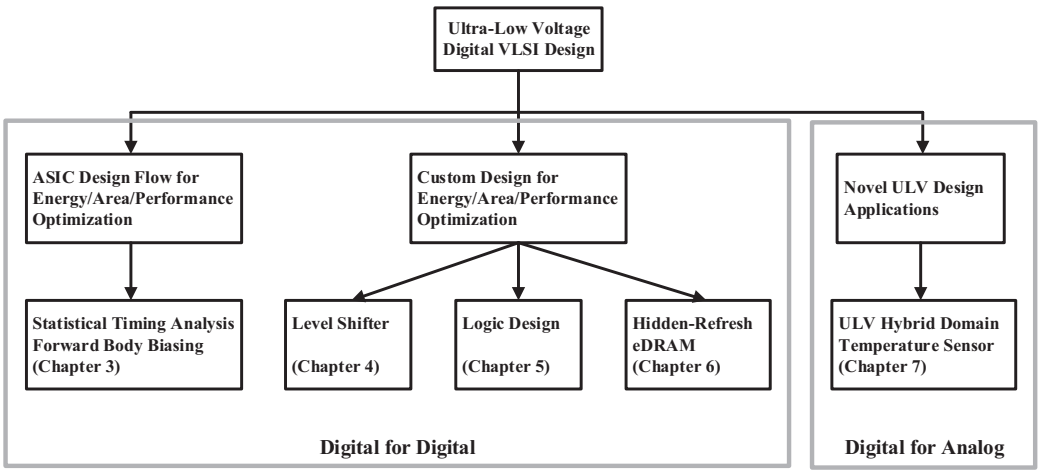


Fig. 1.2: Scope of the thesis.

This thesis presents the design methodologies for ultra-energy-efficient sub-/near-V_{th} computing, with special emphasis on further improving the energy efficiency beyond the state-of-the-art sub-/near-V_{th} designs. We aim at simultaneously

achieving robust ULV operation with energy/area/performance optimization, including ULV ASIC design flow for near-V_{th} timing closure and performance boosting through novel forward body biasing, customized circuit blocks (level shifter, logic design and memory design), and novel ULV design applications (nearly all-digital temperature sensor). The thesis scope is illustrated in Fig. 1.2 and the contributions of the thesis are elaborated as follows:

- Application-Specific Integrated Circuit (ASIC) Design Flow

  First, this thesis covers an ASIC design flow with near-V_{th} statistical timing analysis and design-time forward body biasing to reduce the excessive design margins and to improve the performance when compared with the conventional design flow. A novel computational efficient standard cell characterization method is proposed and an overhead-free forward body biasing scheme is introduced for performance boosting. As a result, we fabricated an Advanced Encryption Standard engine for proof-of-concept, demonstrating the effectiveness of improved performance and energy efficiency with minimum design overheads.

- Custom Design of ULV Circuit Building Blocks

  Second, this thesis focuses on the custom design and optimization of several key building blocks for ULV circuits, including level shifter, logic, and memory designs. Level shifter is an important interfacing circuit block for multiple voltage systems. In this thesis, a novel NMOS-diode current limiter based topology is proposed with improved propagation delay and energy efficiency. For logic design, we propose an energy-efficient intra-cell mixed-V_{th} methodology to enhance the logic cell robustness with minimum device upsizing and

energy overhead. For memory design, we investigate the logic-compatible hidden-refresh embedded DRAM as an alternative for the widely adopted SRAM with improved energy and area efficiency.

- ULV all digital temperature sensor

  Third, this thesis introduces a temperature sensor capable of ULV operation. Generally, ADC and TDC based temperature sensors are power hungry. In order to solve this, we propose a nearly all-digital hybrid domain temperature sensor. A ratioed-current PTAT sensor core and hybrid temperature sensing scheme are introduced. A main feature of this sensor topology is timing reference independent, eliminating the considerable energy required by a frequency reference.

1.3 Organization of the Thesis

The organization of the thesis is listed as follows. Chapter 2 reviews the state-of-the-art work related to the sub-/near-V_{th} designs. Chapter 3 describes the ASIC design flow with statistical timing analysis and design-time performance boosting through forward body biasing. Chapter 4 presents an energy-efficient subthreshold level shifter design. Chapter 5 addresses the energy overheads in logic design through a robustness-driven mixed-V_{th} standard cell design methodology. Chapter 6 demonstrates the design of logic-compatible hidden-refresh eDRAM as a viable memory alternative for SRAM. Chapter 7 covers the design of nearly all-digital hybrid domain temperature sensor. Chapter 8 concludes the thesis and discuss future directions of our research.

Chapter 2

Literature Review

State-of-the-art work related to the scope of this thesis is reviewed in this chapter, including analytical modeling framework and technology implications, circuit building blocks, circuit/architecture techniques and EDA design methodologies and complex ULV SoCs.

2.1 Modeling and Technology Implications

As confined by the classic Alpha-power law [23], the conventional Dynamic Voltage Scaling (DVS) [24] framework entails practical limitations due to the ignored contribution from leakage current, which continuously increases with technology scaling into the nanometer regime. Consequently, DVS technique has to be revisited for its fundamental benefits when aggressive voltage scaling is applied.

The energy consumption versus voltage scaling relationship is theoretically modeled in [6, 8] by taking into account of the portion of leakage energy. In contrast to the quadratic dynamic energy reduction, the leakage energy actually

experiences an exponential increase due to the increased logic delay under reduced supply. As a result, the total energy are found to be minimal when the supply scales to around the transistor threshold, while further supply voltage reduction is less energy-optimal. In addition, another important factor has to be taken into consideration is the device parameter variability in the CMOS fabrication process [25]. When process variations is considered for sub-V_{th} circuits, robust operation becomes challenging due to the exponential I_d-V_{gs} dependence. The threshold voltage variations are the most dominant variation sources, including both V_{th} corner shift (inter-die or global variation) and V_{th} mismatch due to random dopant fluctuations (RDF, intra-die or local variations).

In view of these, technology scaling is arguably favorable to energy minimization [26, 27]. Basically, the leakage current grows nearly an order of magnitude with each technology generation for several years [28]. The leakage increase leads to the on/off current ratio degradation, and such device characteristics degradation will be detrimental to the static noise margins, propagation delay as well as energy dissipation. In addition, scaled technologies suffer from even higher-level V_{th} mismatch when compared with older technologies.

Inspired by above-mentioned aspects, the optimal technology selection is investigated for achieving minimum energy consumption and variability in ULV applications [29]. Due to the vastly different duty cycle and maximum frequency among the ULV applications, the optimal technology tend to be application-specific as well. The basic rule of thumb is that advanced technologies are in favor of those applications with high duty cycle ratio and maximum operating frequency and vice versa.

2.2 Circuit Building Blocks

In order to achieve minimum energy operation, the CMOS circuits should be completely functional and reliable in the sub-V_{th} region. Therefore, building blocks, e.g., logic cells, memory and level shifter, are extensively studied in previous works.

2.2.1 Logic

Preliminary study on sub-V_{th} logic design in [7] suggests that static CMOS logic family with transmission-gate based cells are preferred for minimum voltage operation. Theoretically, standard cell libraries with minimum-sized devices (i.e., small strength logic cells) are energy-optimal in nominal corner [6], while this design choice may cause functional failure in other process corners.

Considering the performance loss in sub-V_{th} logic, several logic optimization methods are introduced to improve the propagation delay and variability as well. Sub-V_{th} logical effort [30] is proposed with a closed-form derivation of optimal stacking transistors sizing strategy to achieve performance optimization. Second-order geometrical effects are also explored for sub-V_{th} logic optimization, including both Reverse Short Channel Effect (RSCE) [31] and Inverse Narrow Width Effect (INWE) [32]. A statistical sizing methodology is also introduced based on RSCE [33].

Yield-enhanced and variation-aware sub-V_{th} logic designs are also explored in [34, 35]. By introducing practical functional yield evaluation criterions, such as butterfly plot and output voltage level, functional yield can be quantified during logic design/optimization phase. The major contributors of logic failure are the V_{th} mismatch and the unbalanced topologies in practical CMOS logic gates (e.g.,

NAND/NOR gates). The situation can be mitigated through device upsizing, while this incurs energy and area overheads.

Apart from the above mentioned work, Schmitt-Trigger logic family [36] is especially beneficial for those applications with optimal supply voltage below the minimum energy point supply with extremely low active duty cycle, such as always-on circuitries. However, the area, energy and performance penalties are inevitable for Schmitt-Trigger logic family.

2.2.2 SRAM

Static Random Access Memory (SRAM) is the most widely adopted embedded-memory type and on-chip SRAMs often play a dominant role in area and energy consumption in today's CMOS SoCs. Owing to this, keeping area and energy efficiency is essentially important for on-chip memories. Supply voltage scaling is also valid for SRAM to reduce the energy consumption as well as reducing the leakage power. However, similar to sub-V_{th} logic design, sub-V_{th} SRAM design is also challenging due to the critical design trade-off between the macro density and cell robustness. The situation is even worse when process variations (esp., the V_{th} mismatch) are incorporated, as the device mismatch is paramount in the extremely scaled SRAM bitcells [37, 38].

Sub-V_{th} SRAM has been an active research topic for the past decade. In order to mitigate the above-mentioned challenges, revising the bitcell topologies and exploring read/write assisting circuitries are viable choices to ensure low voltage functionality.

A revised 6T sub-V_{th} SRAM is demonstrated in [39] through a single-end transmission-gate access topology, with device upsizing and virtual VDD/VSS

write-assisting technique. Although this design increases no transistors to the bitcell, the upsized bitcell area is 2× larger than a traditional 6T cell.

Based on the conventional 6T cell design, novel read-decoupled topologies (7T [40], 8T [41–43], 9T [44], 10T [45–47]) with additional transistors or inserting feedback cutoff transistors are proposed to enable the correct read/write operation. Sense amplifier redundancy is also explored to achieve robust operation in dense SRAM designs [41]. In addition to this, a 10T bitcell with differential sensing scheme is proposed with bit-interleaving support [47].

In summary, the above-mentioned techniques can significantly improve the low voltage SRAM robustness, however, at the cost of increased area and energy overheads. One possible way is to seek for alternative memory circuits to replace the current SRAM. For example, Gain-cell based embedded DRAM (eDRAM) [48] can be a good candidate due to the small area and energy consumption. Nevertheless, the dynamic nature of eDRAM requires solutions for efficient refresh management.

2.2.3 Level Shifter

Level shifter are key building blocks of multiple supply voltage (multi-V_{DD}) designs. As the prospect of sub-/near-V_{th} blocks is promising, the level shifter is also desired to be operational in a wide dynamic range, from sub-threshold to above-threshold.

Unfortunately, the conventional level shifter based on Differential Cascade Voltage Switch (DCVS) topology is incapable of robust conversion from subthreshold to nominal due to the significant contention caused by the limited strength of the pull-down devices operated in the sub-V_{th} region.

Multiple stage level shifter design [49] is a feasible solution as the contention

is minimized between consecutive conversion level using additional intermediate power rails (300mV, 400mV, 600mV and 1.2V). However, this introduces the overhead of generating the supply voltages via voltage regulators, which is costly in most cases. Increase the channel width of the pull-down devices is also helpful in achieving robust functionality, but the introduced area and power overheads are overwhelming as both the density and power consumption are important design metrics for subthreshold level shifter [50]. Therefore, area-/energy-efficient level shifter with small propagation delay are highly desired and a few attempts are demonstrated [50–58].

2.3 Circuit/Architecture Techniques and Design Automation Methodologies

2.3.1 Circuit Techniques

As the performance of sub-V_{th} circuits are strongly affected by the device threshold, threshold selection and body biasing are more effective circuit techniques to achieve optimal energy and performance. For instance, exploring the many-V_{th} devices in nanometer technologies can improve the energy-efficiency, performance and variability [111]. Body biasing is also adopted in [59–61] to mitigate the performance degradation and the variability. Similarly, bootstrapped techniques exploit the exponential dependence on the drain current and the gate overdrive voltage [62, 63]. Through simple insertions of both positive and negative charge pump circuits to the inverter, the gate overdrive voltage swing (0 to VDD) is expanded to -VDD to 2VDD, showing both improved delay as well as reduced leakage (due

to GIDL, Gate-Induced Drain Leaking effects).

Architecture level techniques like parallelism and pipelining are investigated for sub-/near-V_{th} designs. Efficient parallelism for medium-throughput applications with near-V_{th} operation are very popular for multimedia applications [61, 64, 65]. Optimal pipeline stage in sub-/near-V_{th} region is significantly affected by V_{th} mismatch. As a result, reducing the pipeline stage to have long critical path delay permits more averaging effects of device mismatch but at the cost of degraded throughput [25]. Latch-based pipelining [66] is therefore reclaimed in sub-V_{th} region for variation-tolerant pipeline design with significantly improved performance and energy efficiency.

2.3.2 Design Automation Methodologies

The dominance of process variations will also be detrimental to the low voltage circuit timing analysis. Basically, the two major timing violations (i.e., setup and hold time) are both affected by the process variations. Conventional Deterministic Static Timing Analysis (DSTA) is pessimistic as the corner models are extracted from extreme rare situations, thereby DSTA approach comes with unrealistic design margins (e.g., over-estimated setup/hold violations) [67]. In order to reduce such design margin, variation-aware approaches, such as Monte-Carlo (MC) simulation based [25] or Statistical Static Timing Analysis (SSTA) [101], are preferred. SSTA is more computational efficient than MC simulation.

Also, hold time violations are independent from the clock frequency but strongly correlated to the clock skew determined by the clock tree topologies. Unlike super-$_V th$ multi-stage clock-tree designs, ULV clock skews are dominated by the local variations. As a result, shallow clock tree with huge clock tree buffer is pro-

posed for efficient clock tree design [68].

Previous body biasing designs ignore the impact of body biasing at design time, therefore over dimensioning is inevitable during technology mapping. Sub-V_{th} body biasing driven synthesis flow is introduced [69]. However, the major drawback is that this body biasing driven synthesis flow is still corner-based, thereby still have overly conservative design margins.

2.4 SoC Designs

Due to the compelling energy benefits, sub-/near-V_{th} circuits and systems draw considerable attention from both academia and industries. During the past few years, several ULV SoC prototypes are demonstrated for various applications. A brief review of the ULV design progress is given below.

Embedded processors (e.g., micro-controller, MCU) with general-purposes instruction sets offer decent flexibility, therefore they are the most popular candidates for wireless sensor node applications [70–76]. These MCU-based platforms show widely spread performance spectrum, ranging from hundreds of kHz frequency for data sensing and logging applications [72], to tens of MHz frequency for future IoT applications demanding more computing power [76].

However, this architecture is less energy-efficient for biomedical applications with demanding digital signal processing requirements. Accelerator-based platforms [77–81] are therefore more preferred to achieve minimum energy for these specific applications, with dedicated hardware accelerators like FFT, FIR, CORDIC, etc. Novel computing architectures with coarse-grained reconfigurable processor is also demonstrated by IMEC and Samsung for emerging ambulatory biomedical

signal processing applications [81]. Industrial prospects of sub-/near-V_{th} circuits are very promising as efforts have also been made for several wide dynamic range, high-performance and energy-efficient hardware accelerators (motion estimation [82], SIMD vector processing [83], encryption [84], floating-point multiply-add unit [85], etc.,) by Intel. Also, an IA-32 Pentium processor [86] is demonstrated with encouraging energy efficiency benefits.

Finally, self-contained sensor node SoCs with fully integrated energy sources with power management units, sensors with AFEs, clocking, RF and ULV digital processing circuitries are demonstrated for wireless sensing [87, 88] and biomedical [89, 90] applications.

Chapter 3

Near-V_{th} ASIC Design: Statistical Timing Analysis and Performance Boosting

---

Near-threshold operation enables high energy efficiency at significant performance loss and increased sensitivity to process variations. In this chapter, we address both issues with two synergistic approaches. First, we introduce a novel statistical design methodology to efficiently and accurately evaluate the guardband, thereby keeping the near-threshold energy/performance cost of variations at its very minimum. Secondly, we introduce a novel body biasing technique to mitigate the performance loss at near-threshold voltages while not requiring any additional circuitry for the body biasing. Based on these ideas, a 65nm highly energy-efficient AES testchip is presented as a case study. Experimental results show maximum throughput up to 12.2Mbps with energy of 1.65pJ/bit at 0.5V, i.e. a 22× and 7.8×

improvement over previous designs, which was implemented in the same technology node. The proposed techniques also reduce area by 28% over a previous design in the same technology, and enable reliable operation over a wide voltage range (0.5V to 1.2V) without any additional post-silicon tuning or body bias feedback control, as opposed to traditional body biasing schemes.

The remainder of this chapter is organized as follows. Section 3.1 describes the background and motivation of this work. Section 3.2 introduces the Surrogate Model Adjustment-based Statistical Static Timing Analysis (SMA-SSTA) methodology. Section 3.3 details the Area-Overhead-Free Body-Biasing (AOF-BB) with novel SelF-Body-Biasing (SFBB) scheme for ULV design. Design considerations are given for area-/energy-efficient AES architectures in Section 3.4. Section 3.5 covers the testchip implementation and the detailed automated flow under the SMA-SSTA/SFBB framework. Section 3.6 reports the measurement results. The conclusions are drawn in Section 3.7.

3.1 Background and Motivation

Ultra-low voltage (ULV) operation is a popular approach to achieve high energy efficiency [91]. Minimum energy per operation is typically obtained when the supply voltage is close to the transistor threshold voltage, although this comes at the cost of significantly degraded performance and larger sensitivity to process variations [12]. The performance issue is determined by the low transistor speed, hence it can be addressed by making transistor faster (e.g., through forward body biasing). The variability issue is typically addressed by adding a design margin that permits correct operation even in the worst case. Such margin needs to be

accurately evaluated to avoid excessive pessimism (i.e., performance/energy/area degradation), or optimism [92] (i.e., yield degradation). Consequently, practical design challenges are inevitable for ULV timing analysis and performance boosting technique like forward body biasing.

3.1.1 ULV Timing Analysis Challenges

ULV timing analysis is significantly affected by local (within-die) and global (die-to-die) process variations [92]. The traditional Deterministic Static Timing Analysis (DSTA) is well known to be pessimistic due to the large design margin. Accurate evaluation of the required design margin can be achieved by Monte Carlo analysis [9, 25, 74], but it entails an extremely high computational effort. Thanks to its better computational efficiency, Statistical Static Timing Analysis (SSTA) is more practical [92]. SSTA requires a preliminary timing characterization of logic gates under global/local variations. Unfortunately, standard cell characterization for local variations is computationally intensive in ULV designs, and asymmetrical delay Probability Distribution Functions (PDFs) further complicates the characterization [25].

Previous solutions [101, 102] entail a significant computational effort for standard cell characterization due to the increased number of characterization points (e.g., using 0.5σ as interval between $\mu \pm 3\sigma$) to deal with the nonlinearities across the variation space. Also, the number of needed circuit simulations is proportional to the number of transistors in a standard cell and the number of parameters under process variations. This approach also increases the computational effort to perform statistical timing analysis due to the non-linear dependence between delay and slew.

3.1.2 ULV Body Biasing Challenges

Body biasing can directly modulate the threshold voltage of MOSFETs, which is an effective knob to boost the performance at near threshold. However, this technique is typically introduced as a post-silicon tuning technique or adaptive tuning solution [59–61]. Since these works do not include body biasing information during technology mapping, the gate strength and hence area are largely oversized. To partially limit such large area penalty, the works in [69, 93] include some limited body biasing information based on the corner analysis. However, these frameworks are still largely pessimistic since they are based on corner and DSTA analysis.

Another issue of body-biasing design is the overhead caused by the body biasing voltage generation circuitry and the related feedback control for post-silicon tuning. These additional circuits make SoC integration and verification more complicated, and are responsible for significant area/energy overhead. For instance, the body biasing circuitry takes up 18% of the total energy in [60].

In the remaining part of the chapter, we address the above issues by introducing two synergistic techniques that respectively estimate efficiently the statistical design margin and improve the transistor speed. As first technique, excessive design margin is suppressed through the novel Surrogate Model Adjustment Statistical Static Timing Analysis (SMA-SSTA). SMA-SSTA accurately evaluates the strictly required design margin with considerably reduced computational effort compared to previous statistical analysis. The second technique is a novel SelF-Body-Biasing (SFBB) scheme to boost performance with zero circuit overhead (i.e., it does not need any external component such as body bias generator/controller).

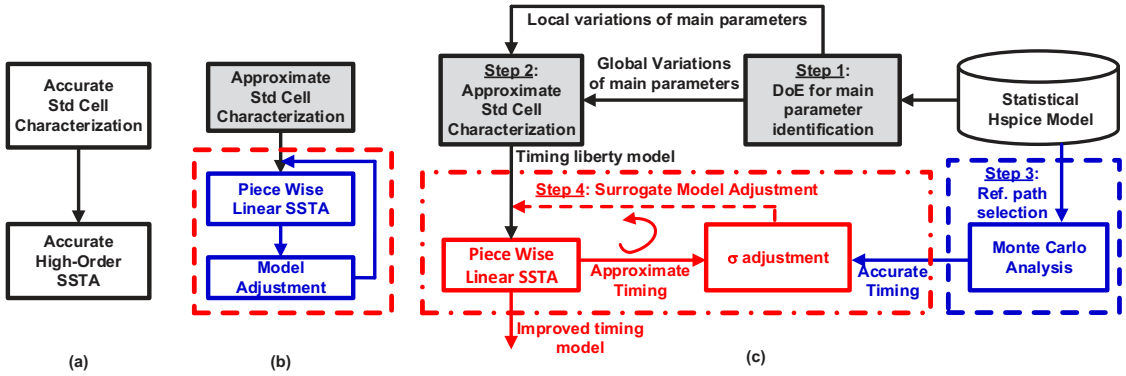


Fig. 3.1: Conceptual illustration of (a) feed-forward SSTA, (b) feed-back based SMA-SSTA and (c) detailed flow chart of SMA-SSTA.

3.2 Proposed Surrogate Model Adjustment based SSTA (SMA-SSTA)

Traditional SSTA flow involves two basic steps as shown in Fig. 3.1(a). The first step is the accurate characterization of standard cell libraries, and the second step is the statistical timing analysis based on the characterized libraries. The first step is well-known to be very computationally expensive [9].

To drastically reduce the computational effort of the traditional SSTA flow, we propose a Surrogate Model Adjustment-based SSTA (SMA-SSTA) framework, as shown in Fig. 3.1(b). It uses surrogate modeling to reduce the computational effort of both standard cell characterization (step one) and timing analysis (step two) in the traditional SSTA flow. Our contribution is on step one, whereas for step two we rely on the work in [103], which assumed gate-level timing variations to be known up front, without providing a solution for step one.

A detailed flow chart for the proposed SMA-SSTA framework is illustrated in Fig. 3.1(c). The flow starts from the statistical HSpice model and consists of the

following key steps:

1. To reduce the number of parameters under variations, we perform Design of Experiments (DoE) based sensitivity analysis to identify the most impactful variation parameters.

2. The standard cell timing is characterized under both global and local variations, considering only the impactful parameters identified in previous step. Global variations are treated normally. For simplicity, local variations are regarded as fully correlated within each standard cell (but the characterization results will be treated as uncorrelated during step 4).

3. A reference datapath is statistically characterized through Monte Carlo simulations (reference datapath selection will be discussed later on).

4. The pre-characterized surrogate timing models are fed to a linear SSTA tool to estimate the timing of the reference datapath in step 3 in a set of cases of interest (e.g., $\mu + 3\sigma$ for setup time check). Then, SSTA input parameters related to local variations (i.e., σ) are iteratively adjusted to match its timing predictions with the results from MC simulations obtained from step 3. This improves the surrogate model accuracy, despite of the inaccuracy introduced by the approximation introduced in step 2.

Being the above framework based on the adjustment of a surrogate model, it preserves the intrinsic correlations and statistical properties of timing as discussed in [103]. Once the surrogate model is calibrated at step 4, it can be used for timing analysis of arbitrary datapaths. In the following we will provide the details on the above steps.

In step 1, the most impactful variation parameters are identified through a delay sensitivity analysis of a reference logic gate. In detail, we evaluated the delay variation of a 1× inverter at different points of interest (i.e., μ-3σ to μ+3σ with 1-σ step) and selected only the variation parameters that led to the most significant delay variation.

In step 2, for simplicity we assumed that all transistors have the σ value pertaining to a single-finger device, regardless of their size. The inaccuracy introduced by this simplification is then mitigated through the calibration at step 4.

In step 3, based on [103], the reference datapath is selected as an appropriate portion of the critical path identified from a preliminary synthesis based on DSTA. The logic depth of the portion of the critical path that is adopted as reference datapath is chosen differently for setup and hold time check, as discussed in the following.

Fig. 3.2 shows the MC simulation results of the delay variability versus logic depth for two 40-stage inverter-chains (1× and 2× strength) at 0.5V. As the logic depth increases, σ/μ of global variations is basically constant regardless of cell strength and logic depth, while σ/μ of local variations is derating along the logic depth. As a consequence, when increasing the logic depth, σ/μ under both global and local variations also derate and converge to a lower bound set by global variations (see Fig. 3.2). From this plot, for setup time check, the logic depth of the reference datapath has to be small enough to capture a significant amount of local variations, but not unrealistically small to be representative of practical near-threshold designs. Accordingly, we set the logic depth to 15, which is certainly on the slow side of practically adopted logic depths [66].

For hold time check, in principle it would be possible to perform model ad-

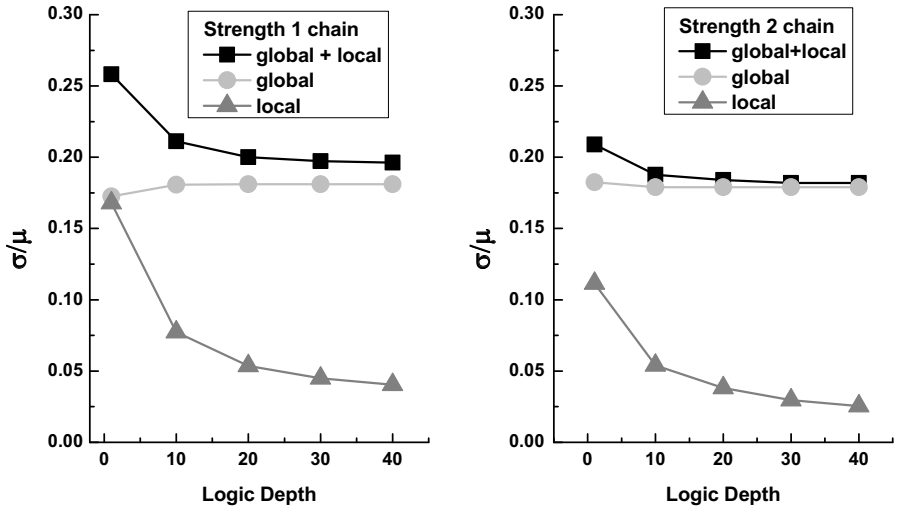


Fig. 3.2: Timing derating feature of logic datapath under local variations.

justment as is done for setup time check, but we need to consider that hold time violations are typically a major issue in ULV designs [9]. Accordingly, we introduced some slight pessimism by choosing a reference datapath with minimum-sized cells and large clock buffers, as actually desirable in ULV designs [102]. The resulting design margin is considerably smaller than DSTA methods. Details on the accuracy and runtime of the above discussed SMA-SSTA flow are reported in Section 3.4.1.

3.3 Area-Overhead-Free Body-Biasing Techniques

In this section, area-overhead-free body-biasing (AOF-BB) schemes are discussed. In such schemes, no additional circuitry (e.g., voltage generation/control) or routing resources are needed for body biasing. After highlighting the limitations of existing AOF-BB schemes, a novel biasing scheme to circumvent those

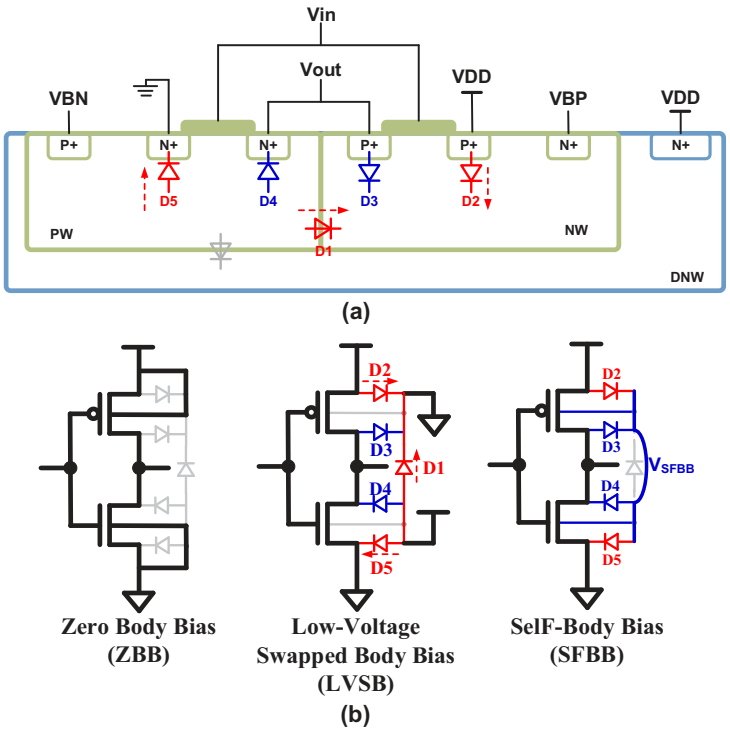


Fig. 3.3: (a) Cross-section of the deep N-well technology and, (b) parasitic diode connections of a CMOS inverter under three AOF-BB schemes (ZBB, LVSB, proposed SFBB).

limitations is introduced.

3.3.1 Conventional AOF-BB Schemes and Limitations

Fig. 3.3(a) shows the cross-section of an inverter cell in a triple-well process, where P-well (PW) and N-well (NW) are located in Deep N-well (DNW). Among the possible P- and N- well biasing strategies, the conventional Zero Body Biasing scheme (ZBB) ties NW to V_{DD} and PW to V_{SS}. Since all parasitic diodes are always reverse biased, ZBB circuits can reliably operate from super-threshold to sub-threshold.

When V_{DD} is constrained to be in the order of 0.5V, the Low-Voltage Swapped

Body Biasing (LVSB) can also be adopted [104], where the P- and N-well voltages are swapped compared to ZBB. This is an extreme forward body biasing case and can achieve significant performance boosting. The supply voltage limitation of LVSB is imposed by diodes D1, D2 and D5 in Fig. 3.3(b). Approximately, operation above 0.5V causes a large current flow due to the forward-bias diodes. The conduction of D3 and D4 also degrades the output voltage swing and remarkably increases the transistor junction capacitance, thereby further degrading energy efficiency.

3.3.2 Proposed SelF-Body-Biasing Scheme

To maintain the flexibility and the reliability of ZBB while boosting the transistor performance similarly to LVSB, we propose a novel SelF-Body-Biasing (SFBB) scheme, as shown Fig. 3.3(b). For SFBB, the body terminals of all NMOS and PMOS devices are directly connected to each other without being tied to external supplies, resulting in a self-biasing node that is applied to both NW and PW. SFBB scheme has two advantages:

1. The P-to-N well diode (D1) is shorted by the connection between the VBN and VBP node under SFBB scheme, thereby eliminating altogether its leakage current.

2. D2∼D5 form a diode network between V_{DD} and ground and lead to a deterministic self-bias voltage. This diode network forms a voltage divider and therefore ties the body to an intermediate voltage between ground and V_{DD}; hence, the resulting body biasing scheme can be used in a much wider range of supply voltages, compared to LVSB.

To understand how the body voltage is set in SFBB scheme, in the following we introduce the Equivalent Diode Network Model (EDNM) to evaluate the self-biased body voltage. We first consider the case of a single gate, in order to gain an insight into the impact of the cell topology. Then, we generalize to the case of multiple gates sharing the same substrate, to understand the impact of the gate-level topology.

Analysis of a single logic cell

Fig. 3.4 shows the detailed illustration of SFBB scheme on an inverter cell and a 2-input NAND cell. As shown in Fig. 3.4(a), the D2 and D5 diodes are series connected between the V_{DD} and ground, while the effect of D3 and D4 on the self-biasing node voltage is determined by the corresponding output voltage level. Assuming OUT is high, the equivalent diode-network is the parallel of D2-D3 in series with D5. D4 is reverse-biased and thus it can be regarded as an open circuit. Assuming for simplicity that all diodes are identical, the self-bias node voltage is $2/3V_{DD}$. Similarly, the self-bias body voltage when OUT is low results to $1/3V_{DD}$.

For a 2-input NAND cell, the resulting EDNM model is shown in Fig. 3.4(b). Analysis shows that the A high/B low input vector leads to the self-bias node voltage that is farthest from $V_{DD}/2$. In this case, the EDNM consists of four parallel diodes that are connected in series with single diode, thereby leading to a bias voltage of $4/5V_{DD}$. Similarly, a 2-input NOR cell has a body voltage of $1/5V_{DD}$. This simplified model was extensively validated through HSpice simulations, and was found to be accurate within 15%.

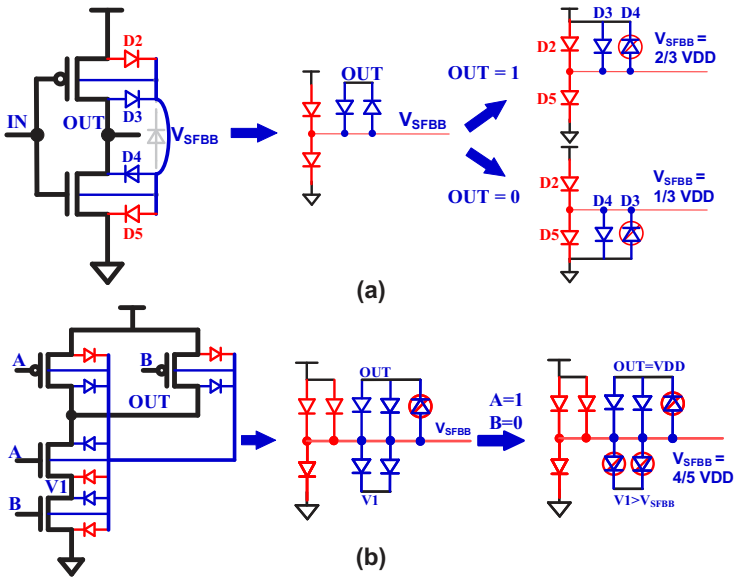


Fig. 3.4: Illustration of the EDNM model of (a) an inverter cell and, (b) a 2-input NAND cell for SFBB scheme.

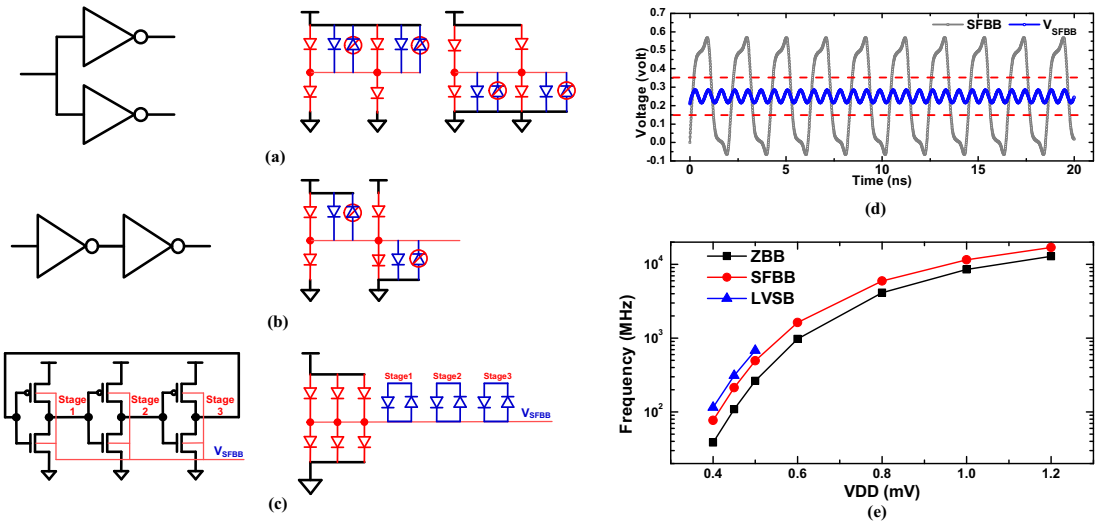


Fig. 3.5: Illustration of the EDNM model for SFBB scheme of (a) an inverter cell, (b) a NAND cell, (c) a 3-stage ring oscillator, (d) simulation waveform of the self-bias node voltage fluctuation of the 3-stage RO SFBB and, (e) timing potentials of the LVSB and SFBB scheme over the ZBB scheme.

Analysis of multiple cells

We use two inverters to illustrate the impact of the gate-level topology on self-bias node voltage. First, we consider the parallel connection in Fig. 3.5(a). Since the output nodes of two inverters are identical, the diode-network for two inverters are the same with a single inverter case, the voltage divider ratio is therefore unchanged. In the case with two cascaded inverters in Fig. 3.5(b), the two inverters have opposite output voltage levels and this results in a balanced diode-network. As a result, the self-biased body voltage is $V_{DD}/2$.

The above observations indicate that both the cell topology and the gate-level topology impact the self-bias node voltage. In actual designs, it is necessary to evaluate the practical range of the body voltage, so that cells are correctly characterized. For cells in low voltage timing libraries, as expected we found that the worst-case individual logic cells are the NAND/NOR gates. The corresponding voltage levels for isolated gates are around 80% V_{DD} and 20% V_{DD}, respectively. However, as suggested by intuition, averaging takes place in practical circuits consisting of a large number of different logic gates. As a consequence, the practical range of the body voltage is narrower, and it has a very limited fluctuation during the network switching. For example, the self-bias body voltage waveform in a 3-stage ring oscillator in Fig. 3.5(c-d) can fluctuate only by 40mV around $V_{DD}/2$. Extensive analysis of a wide range of circuits showed that self-bias node voltage are always well within 30% and 70% V_{DD} range. To characterize the cells accordingly, we adopt a strategy that is similar to FDSOI standard cell libraries [105]. In particular, the bias bounds of 30% to 70% V_{DD} are applied to VBN and VBP node for characterization, respectively.

As LVSB and SFBB are forward body biasing, they can inherently offer a

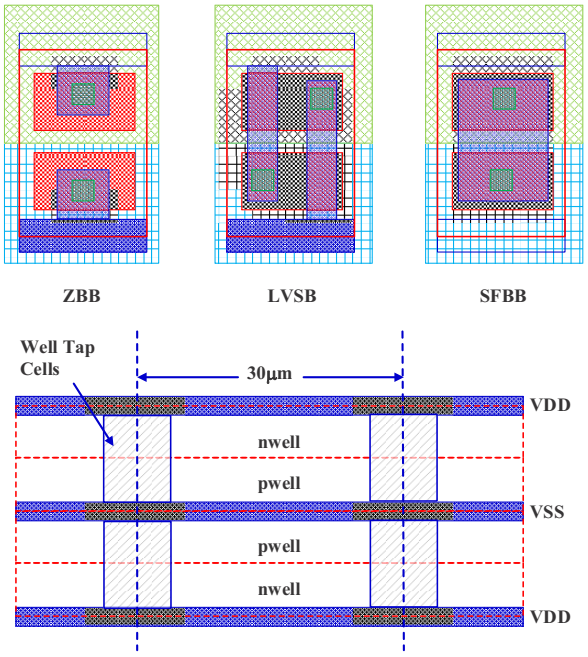


Fig. 3.6: Illustration of the AOFBB well tap designs and body-biasing floorplan under a tap-less standard cell library technology.

considerable performance improvement over ZBB. Also, as opposed to LVSB, the proposed SFBB can operate at supply voltages that are above 0.5V and up to the nominal voltage, since the forward body bias voltage is a fraction of V_{DD} rather than the entire V_{DD}. This is clearly shown in Fig. 3.5(e), which plots the simulated 3-stage ring oscillator frequency at 25°C for the three above AOF-BB schemes. From this figure, LVSB stops working beyond 0.5V due to the abrupt leakage increase and the above discussed logic swing degradation, while the proposed SFBB is able to operate correctly up to 1.2V and achieves up to 40% higher operating frequency than ZBB over the entire operation range.

In addition to the above advantages, SFBB maintains the desirable features of ZBB and LVSB that the body biasing does not require any area overhead, as the

body is biased through proper body tap cells. Such cells short the n-well and the p-well, and their layout is shown in Fig. 3.6. Similarly to LVSB and ZBB, body tap cells are regularly placed in the layout rows by the Place&Route tool.

3.4 Case Study: Advanced Encryption Standard

In this work, the above techniques are used to design a near-threshold Advanced Encryption Standard (AES) 65nm testchip. Fig. 3.7 summarizes the throughput-energy features of previously published AES prototypes, which focus on either high throughput (Gbps or more) [94–96] or very low power and low throughput [97–99]. Compared to low-throughput designs, the energy/bit of some high-throughput designs can be significantly smaller (down to pJ/bit or even slightly lower). On the other hand, the area of high-throughput designs is significantly larger than low-throughput designs (by 4-17×). Hence, a gap exists around medium data rate applications (tens of Mbps) with intermediate energy and compact area, compared to the existing designs.

The adoption of the proposed techniques permits to fill this gap, as our near-threshold[1] AES core in 65nm CMOS achieves 12.2Mbps (38.4Mbps) throughput with 1.65pJ/bit (2.34pJ/bit) energy at 0.5V (0.6V). At the same time, our AES engine occupies only 0.013 mm^2 silicon area. Hence, our techniques permit to achieve the lowest energy (by 5-10× or more) compared to previous designs for sensor nodes, while increasing the achievable throughput to tens of Mbps, and keeping area comparable or better.

[1]In the adopted technology, the NMOS threshold voltage is 0.6V.

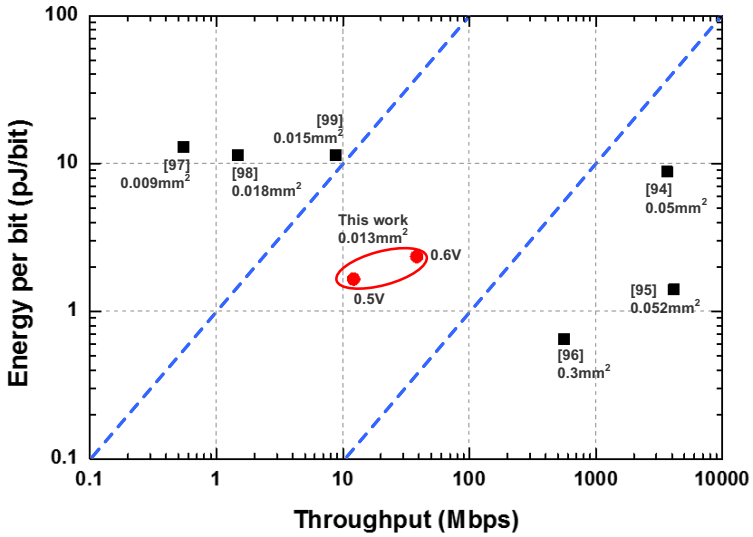


Fig. 3.7: State-of-the-art AES designs: energy vs. throughput and area (scaled to 65 nm node for the different adopted technologies).

3.4.1 Low-Cost AES Architectures and S-Box Implementation

In the following, the selection of the architecture organization and the S-BOX implementation are discussed with the goal of enhancing the efficiency of the energy-area-performance tradeoff.

Area-efficient AES core implementations typically adopt folded datapath microarchitecture to reduce the gate count [97–100], as opposed to very high-performance targets, whose area can be higher by an order of magnitude (see, e.g., [96]). In area-efficient designs, area tends to be dominated by memory, as shown by the area breakdown in Table 3.1 of the area-efficient designs in Fig. 3.7. Hence, the memory organization is a key lever to optimize the area-performance tradeoff.

In unified-RAM architectures (e.g., [97, 98]), data and key are fetched, processed and stored in a single RAM block. These architectures suffer from low

Table 3.1: Comparison of state-of-the-art Area-Efficient AES architectures

	Unified-RAM [97],[98]	Split-RAM [99]	Shift Register [100]
Memory	60%	55%	59%
S-Box	12%	7%	16%
MixColumn	9%	8%	9%
Control & Others	19%	30%	16%
Cycles/Block	1134	356	160
Gate Equivalent (GE)	3.4K	5.5K	4K
Silicon Prototype	Yes	Yes	No

performance due to the large number of clock cycles required for round computation and key expansion. Instead, split-RAM architectures (see, e.g., [99]) have two separate RAMs for data and key. This architecture enables key expansion interleaving during round computation while using a single S-BOX, thereby reducing the number of cycles (i.e., increasing throughput) by 2.9×. Such advantage comes at the price of more complex and larger-area control logic, as confirmed by Table 3.1. As an alternative, the architecture in [100] uses a shift-register based memory, which efficiently performs byte permutations in both data (round computation) and key processing (key expansion). The presence of an additional S-BOX permits to simultaneously perform round computation and key expansion [100], thereby improving the performance by 7.1× at only a 14.7% area penalty compared to unified-RAM architecture [98].

Another key choice to optimize the area-energy-performance tradeoff is the appropriate selection of the S-BOX implementation. Indeed, although S-BOX occupies a moderate fraction of the AES area (see Table 3.1), it lies on the critical path and hence it strongly impacts the overall performance-energy tradeoff (especially at near threshold, as the substantial contribution of leakage energy is lowered when

Table 3.2: Comparison of state-of-the-art normalized S-Box design

	Area (Norm. to NAND2)	Delay (Norm. to NAND2)	Norm. A-D Product
[106]	263.5	75.3	19842
[107]	335	62.9	21072
[108]	276	66	18216
[96]	324.5	45.3	14700

reducing the cycle time [66]). In our AES design, we adopt native Composite Field (CF) S-BOX design as in [96], which removes the overhead of mapping and inverse mapping from GF(2^8) to CF. In the adopted 65nm technology, this respectively reduces S-BOX delay and area by 28% and 20%, as shown in Table 3.2. However, the native CF AES design causes additional delay overheads in Mixcolumn unit (3 XOR delay). When considering the 5 XOR delay reduction obtained in native S-BOX, the overall critical path delay reduction is 2 XOR, which yields around 10% delay reduction. The detailed AES architecture with both native round and key S-BOX is depicted in Fig. 3.8.

3.4.2 Automated and Detailed SMA-SSTA Design Flow

A commercial standard cell library was pruned by discarding cells with fan-in larger than two [59]. Also, transmission gate structures connected at the input/output pins of cells were buffered to avoid sneaky leakage paths, and minimum-strength cells were eliminated to limit random variations. The resulting library comprises 25 different types of cells with different strengths, leading to 116 cells in total.

Regarding the step 1 of the SMA-SSTA flow in Section 3.2, the identification

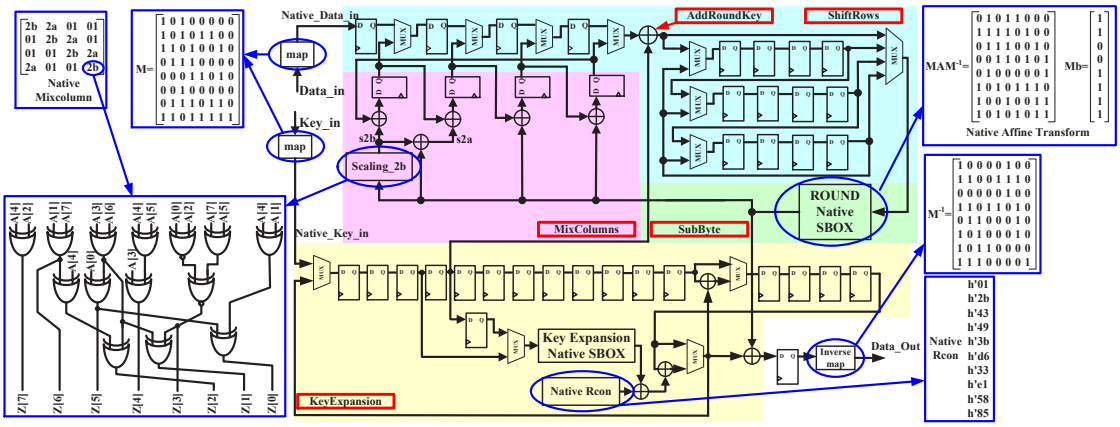


Fig. 3.8: The AES engine architecture with native round and key expansion S-Box.

of the main variation sources is performed through Design of Experiments (DoE) based on a delay sensitivity analysis on 1× inverter. The Pareto set plot pertaining to the variation parameters with delay normalized to its mean value is shown in Fig. 3.9. From this figure, in the considered 65nm technology, the dominant sources of variations are the threshold voltage and the oxide thickness. The oxide thickness variations are global, while threshold voltage variations are both global and local.

Regarding the step 2 of the SMA-SSTA flow, the cell characterization was performed at 0.5V for all the three AOF-BB schemes and at 0.6V for ZBB and SFBB, due to the 0.5V voltage limitation of LVSB. As the outcome of step 2, the resulting surrogate cell delay model of local variations is stored in the form of Composite Current Source (CCS) model, along with the global variations [110].

The assumptions and approximations introduced in related step 3 and 4 in Section 3.2 were validated extensively. The model adjustment was performed on a 15-stage critical sub-circuit and the adjusted model was then applied to validate other datapath. This adjusted model was validated on several other randomly selected datapaths with logic depth ranging from 5 to 25. Fig. 3.10 depicts some

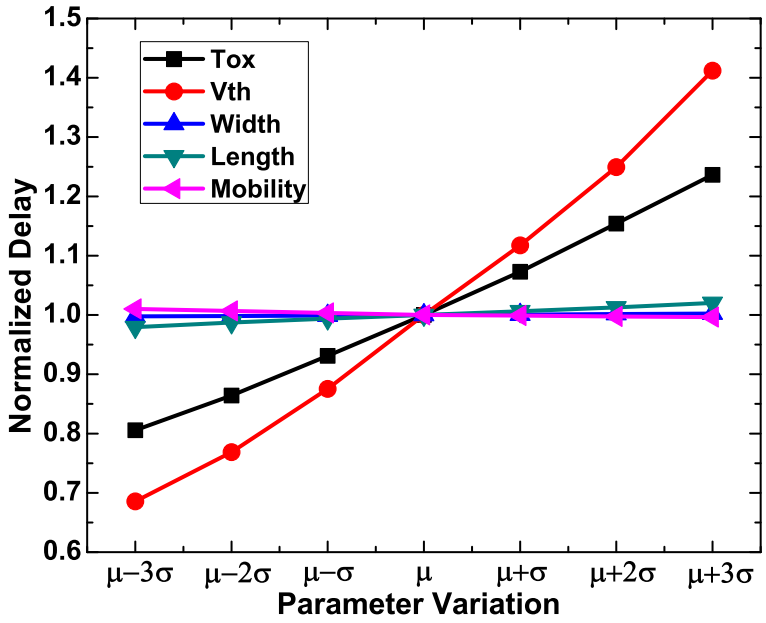


Fig. 3.9: Pareto set plot of DoE for delay sensitivity analysis toward different variation parameters.

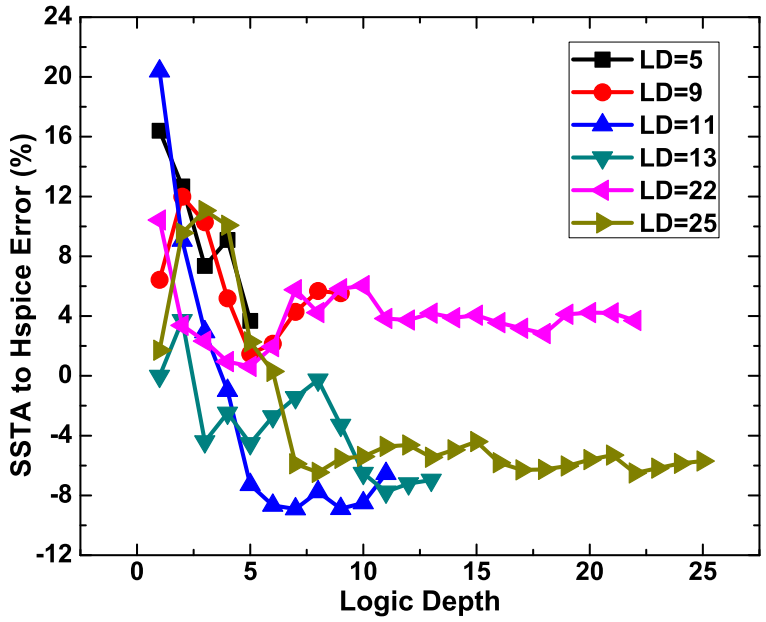


Fig. 3.10: +3σ setup time accuracy analysis of SMA-SSTA after model adjustment.

of the results, and shows that the error is moderate at low depths, and is below 7% for logic depths of ten or more, compared to Monte Carlo simulations. This confirms the correctness of the design considerations on the logic depth of the reference datapath in Section 3.2. Despite of the drastic simplification offered by the proposed models, this error is as small as that encountered in SSTA predictions [101]. Regarding the hold fix, the resulting area and power cost of buffer insertion for hold fix under SMA-SSTA was respectively found to be 4% and 6%, at 0.5V under ZBB scheme. Under the DSTA approach (using FF corner as the worst case as usual), the hold time fix overhead is 23% and 18% for area and power, respectively. This confirms that SMA-SSTA avoids over-design, as opposed to DSTA.

3.4.3 Runtime for Local Variation Characterization and Comparison with SSTA

The SMA-SSTA flow enables significant reduction in the runtime needed by the cell characterization under local variations. In our SMA-SSTA flow, only ($2N_{rv}+1$) spice simulations are needed for each delay/slew point in the proposed framework, where N_{rv} is the number of random variables [110]. For NLOPALV in [101], the number of SPICE simulations required to characterize each cell is on the order of $\mathcal{O}(N_{rv}N_{tr}N_{iter})$, where N_{tr} is the number of transistors within a logic cell and N_{iter} is the required iterations for operating point convergence. In addition, a total of 13 points (N_{char}) characterization (spacing = 0.5σ) is adopted from -3σ to +3σ. Accordingly, the speed-up offered by SMA-SSTA compared to NLOPALV can be

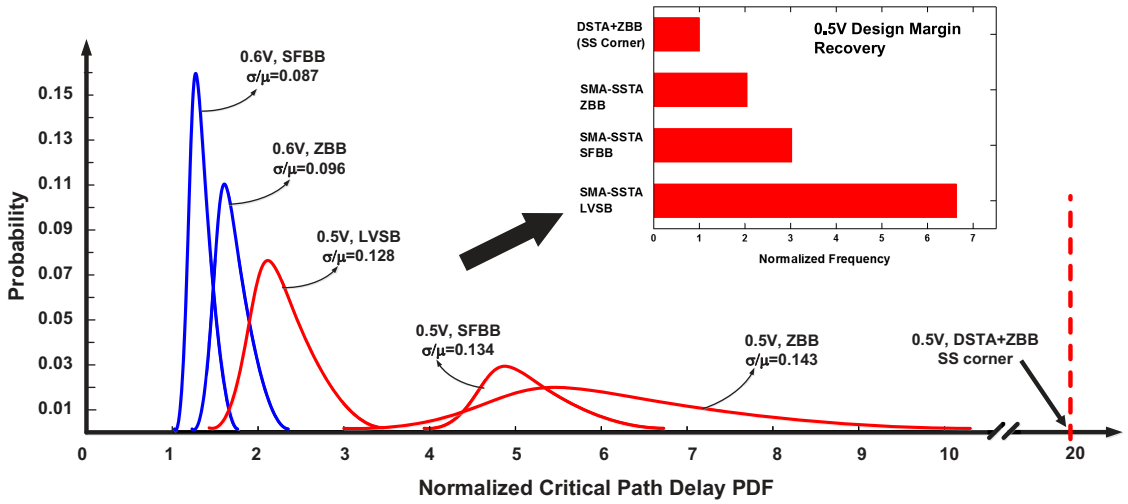


Fig. 3.11: Statistical distribution of the delay (normalized to $\mu-3\sigma$ delay under SFBB at 0.6V) for different body biasing schemes and resulting clock frequency improvement over DSTA (top-right).

expressed as

$$\frac{2N_{rv}+1}{(0.28N_{rv}N_{tr}+0.23N_{char}+1)N_{iter}} \tag{3.1}$$

where the coefficient in the denominator is the empirical value adopted from [101]. For example, for a 2-input NAND gate with N_{rv}=2, N_{tr}=4 and N_{iter}=3, the runtime of the proposed framework is only 26.8% of NLOPALV. The runtime advantage of SMA-SSTA is further improved for cell topologies with higher complexity.

3.4.4 Performance and Design Margin Recovery

The statistical distribution of the critical path delay obtained with SMA-SSTA is plotted in Fig. 3.11 for ZBB and SFBB for 0.5V and 0.6V, as well as for LVSB at 0.5V. For simplicity, the delay is normalized to the value achieved at $\mu-3\sigma$ by the SFBB scheme at 0.6V. From this figure, DSTA introduces a very large design margin, which translates into a significant clock frequency penalty. In particular,

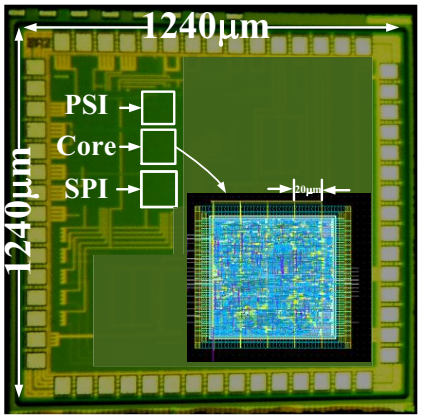
at 0.5V the clock frequency at $\mu + 3\sigma$ for DSTA is only 2.05MHz under ZBB scheme, whereas it increases to 4.2MHz when the SMA-SSTA framework is adopted. Accordingly, the SMA-SSTA methodology enables a 2× design margin recovery (i.e., performance), as shown in the upper-right portion of Fig. 3.11, which shows the clock frequency normalized to the DSTA approach. As expected, the adoption of LVSB and SFBB lead to a significant performance improvement compared to ZBB. Indeed, at 0.5V the clock frequency at $\mu + 3\sigma$ under LVSB and SFBB is respectively 13.6 and 6.2MHz, which represent a 3× and 6.6× improvement as shown in the upper-right part of Fig. 3.11. At 0.6V, SFBB has a higher clock frequency (17.4MHz), whereas LVSB cannot operate correctly.

3.4.5 Physical Implementation Considerations

The final design was placed and routed in SoC Encounter. In this design, a tap-less standard cell library is used and therefore the body tap cells are placed during floorplan phase. As usual, the spacing rules dictated by the technology are followed when placing tap cells throughout columns. PMOS and NMOS body taps are connected externally to easily reconfigure the same core to ZBB, LVSB and SFBB biasing scheme. This permits to fairly compare the three techniques in the presence of the same local variations.

3.5 Testchip Measurement Results

The AES encryption engine with AOF-BB schemes was implemented with the proposed design flow in a 65nm Low Leakage CMOS process. Fig. 3.12 shows the die photograph with annotated details and the test chip summary. Plain text and



Process	65 nm Low Leakage
Supply	0.5 - 1.2 V
Area	0.008 mm^2 (w/o power rings) 0.013 mm^2 (w power rings)
Maximum Throughput	12.2 Mbps (0.5V, SFBB) 38.4 Mbps (0.6V, SFBB)
Energy per bit	1.65 pJ/bit (0.5V, SFBB) 2.34 pJ/bit (0.6V, SFBB)

Fig. 3.12: Die micrograph with annotated core and on-chip testing buffer and summary of the AES encryption engine at room temperature.

key are inputted serially, and input (SPI, serial-to-parallel interface) and output buffers (PSI, parallel-to-serial interface) are provided for on-chip testing purposes. The serial scan of input/output data minimizes the number of testing pads (the chip is pad limited), but constrains the frequency of at-speed testing to about 100MHz. This sets a 0.7V voltage upper bound for at-speed testing, which fully covers the near-threshold region of interest.

3.5.1 Performance Measurement and Energy Comparison

Performance and energy were measured for the ZBB, LVSB and the proposed SFBB body biasing schemes. The supply voltage was varied from the minimum value that enables correct operation, which is in deep subthreshold, up to 0.7V. If not stated otherwise, in the following the temperature is set to 25°C. Fig. 3.13 shows the operating frequency and energy per bit of the three body biasing schemes. For a chip that is close to typical corner (see Fig. 3.13), the minimum operating voltage of ZBB is 230mV, and the minimum energy point occurs at 300mV. Around this

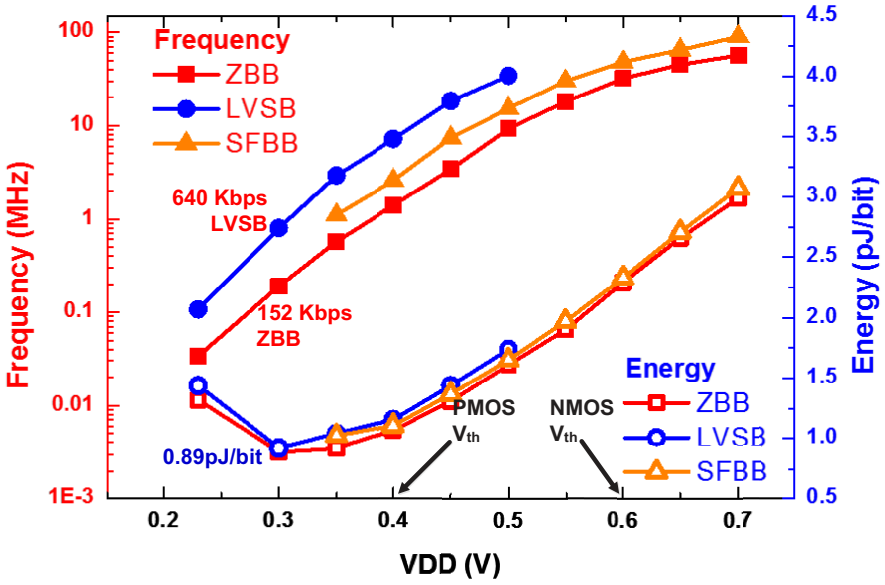


Fig. 3.13: Measurement results of operating frequency and energy per bit of the testchip.

point, the measured energy is 0.89pJ/bit and throughput is 152Kbps. The LVSB scheme has the same minimum operating voltage and the minimum energy point voltage of 300mV. At this voltage, LVSB increases the throughput to 640Kbps, while keeping 2% energy overhead of ZBB, which represent a 4.1× performance improvement at nearly the same energy. In the 0.23-0.5V range, LVSB boosts the performance by 3.7-5.4× compared to ZBB, with an energy that is 3.4%-13% higher than the latter. As a severe limitation of LVSB, operation above 0.5V is not allowed as explained in Section 3.3.

As expected, the proposed SFBB scheme does not have such voltage limitation, and can operate up to the nominal voltage (1.2V in this technology). Correct functionality of SFBB was verified experimentally in the 0.35-1.2V range. This is a considerable advantage of SFBB over LVSB in practical designs. Indeed, it strongly simplifies SoC integration since V_{DD} can be freely set during the op-

Table 3.3: Summary and comparison of state-of-art area-efficient AES designs

	[98]*	[99]	**This work, ZBB+SMA-SSTA**	**This work, LVSB+SMA-SSTA**	**This work, SFBB+SMA-SSTA**
AES Mode	ECB, enc/dec	ECB, enc	ECB, enc		
Technology	65nm LP	0.13µm	65nm LL		
Supply Voltage (V)	0.5	0.8	0.5/0.6	0.5	0.5/0.6
Area (mm^2)	0.018	0.06#	0.013		
Frequency (MHz)	4.9	12	9.3/32	34.3	15.3/48
Throughput (Mbps)	0.55	4.3	7.4/25.6	27.4	12.2/38.4
Energy (pJ/bit)	12.9	23	1.61/2.3	1.74	1.65/2.34
Cycle/Block	1134	356	160		
Latency (µs)	231.4	29.7	17.2/5	4.66	10.5/3.3
Near-threshold statistical design	No	No	Yes		

*0.5V data is extrapolated from [98]. #Area is estimated from [99] including power rings.

timization/aggregation of voltage domains and the related voltage selection. In addition, the performance of SFBB is much more scalable than LVSB, thanks to the wider voltage range. Because of the above mentioned limitation on the testing frequency, Fig. 3.13 reports the measured frequency up to 0.7V. From this figure, SFBB consistently offers 1.65× more performance over ZBB across the 0.35-0.7V range. Also, the frequency of SFBB can be scaled up to 90-MHz when V_{DD} is set to 0.7V, which is 2.6× higher than LVSB at 0.5V. Clearly, an even wider performance increase is achieved when pushing the voltage beyond the near-threshold region depicted in Fig. 3.13.

Table 3.3 shows the comparison of our AES core and state-of-the-art area-efficient AES implementations. From this table, the proposed SFBB body biasing scheme, the proposed SMA-SSTA methodology and the more efficient microarchitecture deliver a 22× throughput improvement and 7.8× reduction in energy per bit, when compared to [98]. Observe that [98] is implemented in the same technology node, and the performance improvement of our testchip is due to the

adoption of SFBB (1.65×), SMA-SSTA (2×), and microarchitecture (7.1×). The latter reduces the number of clock cycles per encryption from 1134 to 160.

Regarding the silicon area, the proposed SFBB and SMA-SSTA reduce the gate size (i.e., total area) for a given performance target, thereby improving area efficiency [109]. In this design, a 28% layout footprint reduction is achieved compared to [98], despite of the 14.7% area penalty associated with the selected high-performance architecture. Such area reduction is also partially responsible for the improved efficiency of the energy-performance tradeoff. Indeed, small area leads to shorter wires and reduced parasitics, thereby benefitting both energy and performance. Summarizing, the proposed SFBB scheme and SMA-SSTA methodology boost performance by 3.3× at no area overhead and insignificant energy penalty compared to a standard ZBB design. Compared to [98], our AES core achieves a 7.8× energy reduction and exhibits an additional 7.1× speed improvement thanks to a more efficient architecture.

3.5.2 Static and Dynamic Robustness of the Body Voltage Bias Point in SFBB

The stability of the body voltage bias point in the SFBB scheme was extensively assessed in a wide range of static and dynamic conditions. Regarding the static stability, leakage is an effective indicator of the stability of the body bias voltage, due to its exponential impact on the threshold voltage. Accordingly, the chip leakage current was measured under V_{DD} widely ranging from 0.5V to 1.2V, as well as for temperatures ranging from 25 to 75°C. Due to the very high leakage current of LVSB for V_{DD}>0.5V, a 1-kOhm current-limiting resistor was inserted

between the supply and the testchip.

Fig. 3.14 summarizes the leakage measurements versus V_{DD} at a temperature of 25°C, 50°C and 75°C (leakage is normalized to the measurement under ZBB scheme at 0.5V, 25°C). As expected, LVSB scheme suffers from a very large leakage current at 0.5V, which is more than one order of magnitude higher than SFBB and ZBB from Fig. 3.14. Furthermore, LVSB experiences a very rapid (exponential) increase in leakage current for V_{DD} over 0.5V at 25°C, which makes its adoption impractical for V_{DD} above 0.5V. At higher temperatures leakage further increases, and maintaining a targeted leakage current requires a 50mV reduction in V_{DD} for every 25°C temperature increase. In other words, the tight voltage constraint imposed by LVSB is further reduced at higher temperatures. On the other hand, as expected the proposed SFBB scheme and ZBB exhibit a moderate leakage increase at higher voltages, which confirms that they can operate at voltages up to the nominal V_{DD}.

For the sake of completeness, the leakage measurements were repeated in other 15 dice under the same environmental conditions, eliminating the current-limiting resistance altogether. Fig. 3.15 summarizes the measurement results across the dice at V_{DD} equal to 0.5, 0.8 and 1.2V, and 25°C temperature (leakage is normalized to the measurement under ZBB scheme at 0.5V, 25°C). As shown in this figure, the results are consistent with the observations made in Fig. 3.14. In particular, the leakage of SFBB scheme is consistently 5-7$\times$ the leakage of ZBB for the entire considered voltage range. This means that the SFBB body voltage is highly stable across voltages, temperatures and dice, thereby confirming the solid stability of its bias point.

The body voltage in SFBB scheme was also measured in dynamic through a

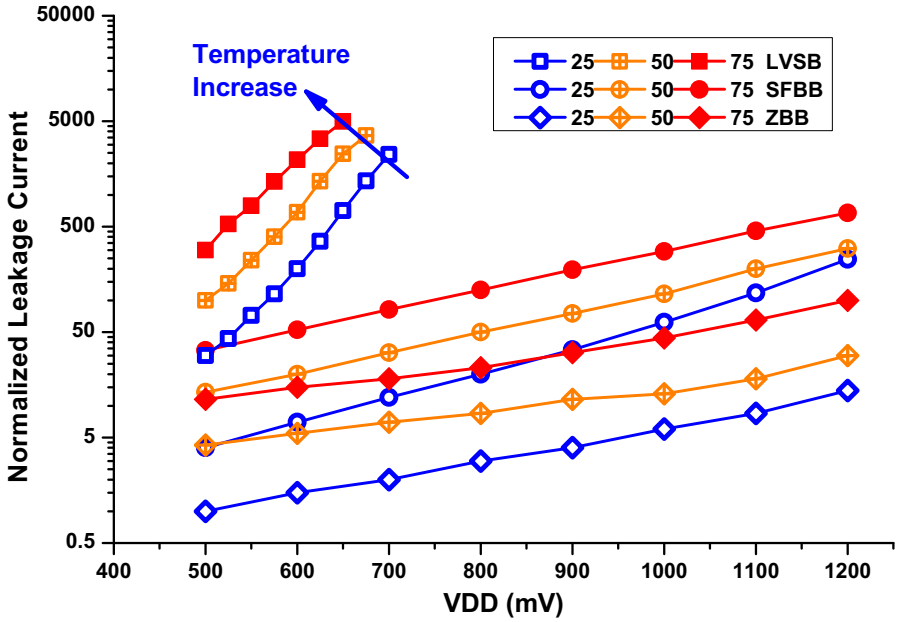


Fig. 3.14: Leakage measurements of ZBB, LVSB and SFBB versus V_{DD} under different temperatures in a single die.

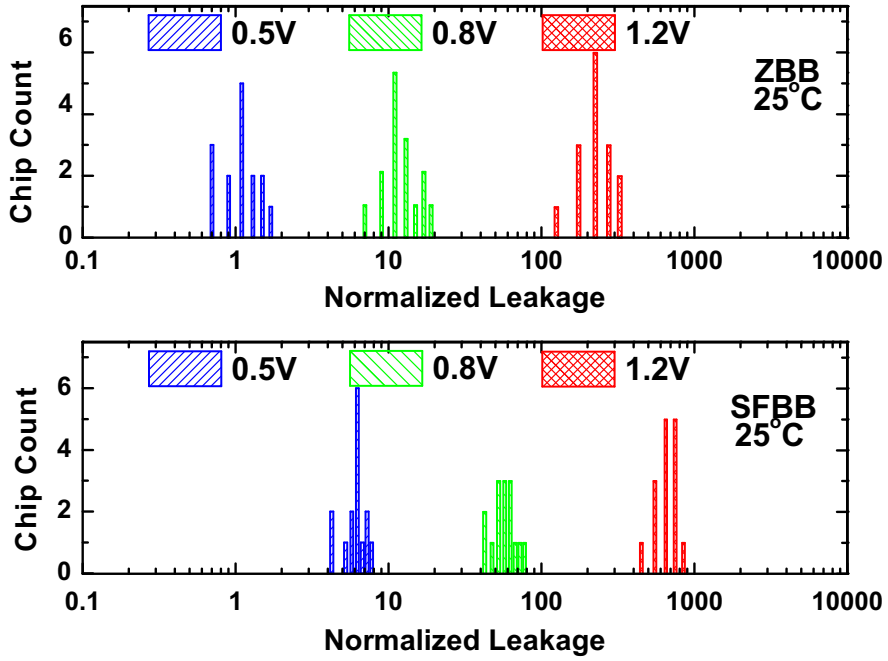


Fig. 3.15: Leakage measurements across 15 dice.

unity-gain CMOS analog buffer. As shown in Fig. 3.16, which refers to the case of supply voltage of 0.5V, the self-bias node voltage is close to $1/2V_{DD}$ and well within the sign-off range with VBN and VBP of $0.3V_{DD}$ and $0.7V_{DD}$. Small glitches with an amplitude of about 60mV are seen on the self-bias node due to the internal switching, as expected from Section 3.3.

The dynamic stability of the body voltage bias point was further analyzed by perturbing the steady-state value by forcing a large-signal voltage pulse and observing the transient response. An aggressive external voltage pulse with a width of 5 μs and amplitude of 0.25V was applied directly to the SFBB body voltage to mimic a dramatic voltage change. Such pulse was forced through a large coupling capacitance (0.1 μF). The measured SFBB body voltage response under V_{DD} of 0.5V is depicted in Fig. 3.17, which clearly shows that the body voltage rapidly goes back to its steady-state waveform after the transient. Extensive measurements in various conditions consistently confirmed that the body voltage bias point is highly stable even under extreme perturbations.

3.6 Conclusion and Summary

In this chapter, we have introduced two synergistic techniques that counteract two fundamental issues of near-threshold VLSI circuits: performance loss and large guardband due to process variations. A novel SelF-Body-Biasing scheme boosts the transistor speed while entailing zero area overhead. As opposed to existing area-overhead-free body-biasing schemes, our technique boosts the performance from near threshold to nominal voltage and ensures reliable operation, as opposed to LVSB, whose supply voltage is severely limited to 0.5V and below. This dramat-

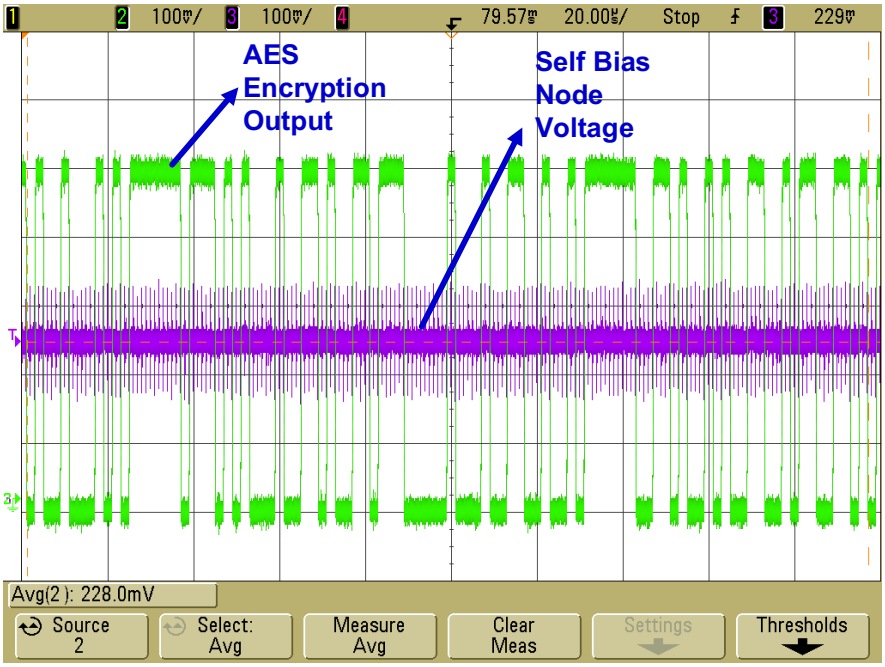


Fig. 3.16: Measured self-bias node voltage at 0.5V supply.

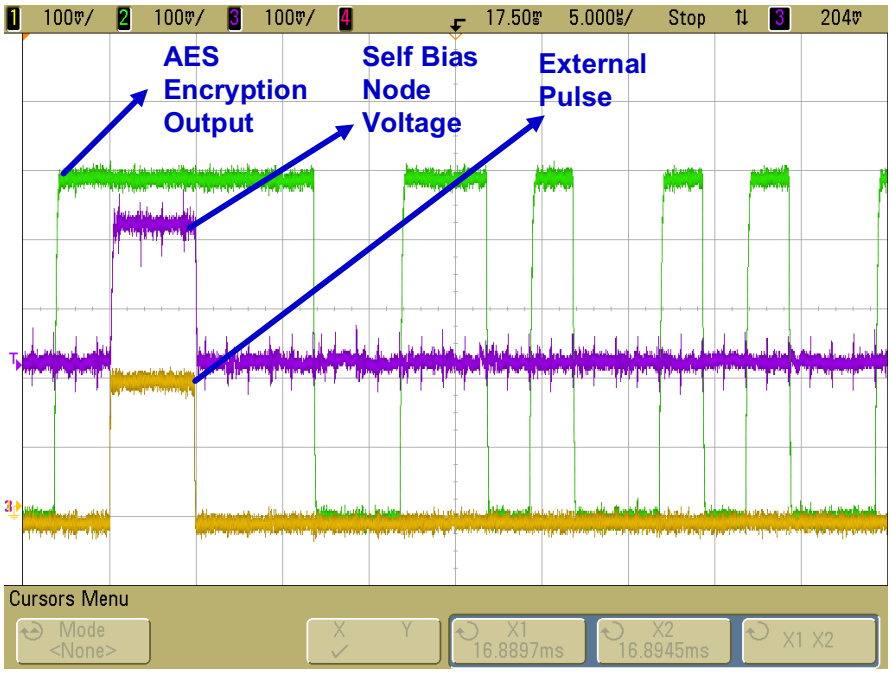


Fig. 3.17: Dynamic stability test of the self-bias node.

ically simplifies SoC integration, since SFBB does not set any constraint on the (static or transient) voltages that need to be distributed on chip.

A novel Surrogate Model Adjustment based Statistical Static Timing Analysis methodology has also been introduced for efficient cell library characterization and timing analysis at near threshold. The adoption of surrogate gate delay models for local variations permit to drastically reduce the computational effort compared to traditional SSTA, while maintaining a comparable accuracy. The joint adoption of these techniques enables for the first time the variation-aware body biased design, as opposed to previous body biasing design strategies that either completely ignore variations [59–61], or suffer from large guardband due to the simplistic timing analysis based on corners [69, 93].

Experimental results of a near-threshold AES core in 65nm CMOS have shown that the above techniques can be synergistically adopted to drive the throughput up to 12.2Mbps (38.4Mbps) at 0.5V (0.6V). This represents a 22.2× throughput improvement over [98], which is implemented in the same technology node. Also, the proposed techniques permit to achieve an energy of 1.65pJ/bit (2.34pJ/bit) at 0.5V (0.6V), i.e. a 7.8× (5.5×) energy reduction compared to [98]. As an additional benefit, the gate size reduction enabled by the improved transistor speed and the elimination of over-design, our AES engine occupies only 0.013 mm^2, i.e. 28% less than [98]. In general, our AES core enables 5-10× energy reduction over previous designs for sensor nodes [98, 99], and increases throughput to the range of tens of Mbps, while exhibiting comparable or better area.

In summary, the proposed techniques are very well suited for the design of near-threshold IPs with very high energy efficiency, while considerably boosting performance at reduced design effort and area. The proposed techniques enable

seamless SoC integration since they do not require any post-silicon tuning or external block to control the body bias voltage, and ensure reliable operation from near-threshold to nominal voltage.

Chapter 4

A 65nm 30.7fJ/bit Subthreshold Level Shifter Design

In this chapter, a novel static level shifter is proposed and measured in 65nm CMOS technology for robust and efficient level conversion in a wide input voltage range. Several circuit techniques are proposed to improve the energy efficiency, delay and area metrics. A novel level shifter topology with NMOS-diode based current limiter for current contention reduction is proposed for efficient level conversion through weakening the pull-up network strength. Second-order geometrical effect is also explored to increase the drivability of the devices in the pull-down network in the subthreshold region. Combining the popular MTCMOS technique in today's CMOS technology, the measured level shifter achieves robust conversion from deep subthreshold (sub-100mV) to nominal supply voltage (1.2V). For the target conversion from 0.3V to 1.2V, the proposed level shifter shows on average 25.1ns propagation delay with 30.7fJ/bit energy efficiency, and the average leakage

power is 2.2nW across 25 test chips.

4.1 Introduction

The growing demand of low-power systems are driven by the battery-powered or energy-scavenging applications, where power consumption becomes the prior design concern to prolong the operational lifetime [87]. As a result, voltage scaling is extensively studied to reduce the total power consumption and multi-V_{DD} technique is popular for today's low-power System-on-Chip (SoC) designs. For energy-constrained and performance-relaxed applications, aggressive voltage scaling into the subthreshold region shows even promising energy benefits [7] as the minimum energy point often exists in the subthreshold region [6].

For multiple supply voltage designs, level shifters are indispensable and ubiquitously inserted between different voltage domains, or directly used to drive the highly capacitive I/O devices or external loads. In view of the subthreshold operation, the level shifters are preferred to operate in a wide dynamic range, including the subthreshold input scenarios. Unfortunately, the conventional level shifter design based on the Differential Cascode Voltage Switch (DCVS) topology is challenging for robust up-conversion from subthreshold to superthrehsold supply voltage, which is due to the significant current contention caused by the limited strength of the subthreshold operated pull-down devices. Generally, when the input low signal scales below 500mV, the current contention between pull-up/pull-down networks leads to the conversion failure of the DCVS level shifter.

Multi-stage level shifter design [49] is one feasible solution as the contention in each stage is mitigated by using additional intermediate power rails with reduced

voltage difference (300mV, 400mV, 600mV and 1.2V). However, this introduces the overhead of generating the supply voltages via voltage regulators, which is costly in most cases. Also, increasing the channel width of the pull-down devices is also helpful to mitigate the contention, but the introduced area and power overheads are overwhelming as both the area and power consumption are important design metrics for subthreshold level shifters. Consequently, proper design techniques [50–58] for reliable and efficient level shifter operation are highly desired.

In this chapter, we demonstrate the design and measurement of a novel energy-efficient static level shifter in 65nm CMOS. Through the adoption of novel topology, MTCMOS and the subthreshold device sizing, the proposed level shifter achieves robust and efficient conversion from deep subthreshold voltage up to the nominal supply voltage of 1.2V. Measurement results show that the proposed level shifter can successfully up-convert a minimum supply voltage (on average 100mV) to the nominal voltage of 1.2V. For a 0.3V, 1MHz input signal, 25 measured level shifter achieves on average 25.1ns propagation delay and 30.7fJ/bit energy, which outperforms the state-of-the-art level shifter designs.

The rest of this chapter is organized as follows. Section 4.2 briefly reviews the state-of-the-art subthreshold level shifter implementations. Section 4.3 describes the proposed level shifter design. Section 4.4 presents the measurement results of the test chips and comparison to previous level shifter implementations. Section 4.5 concludes this chapter.

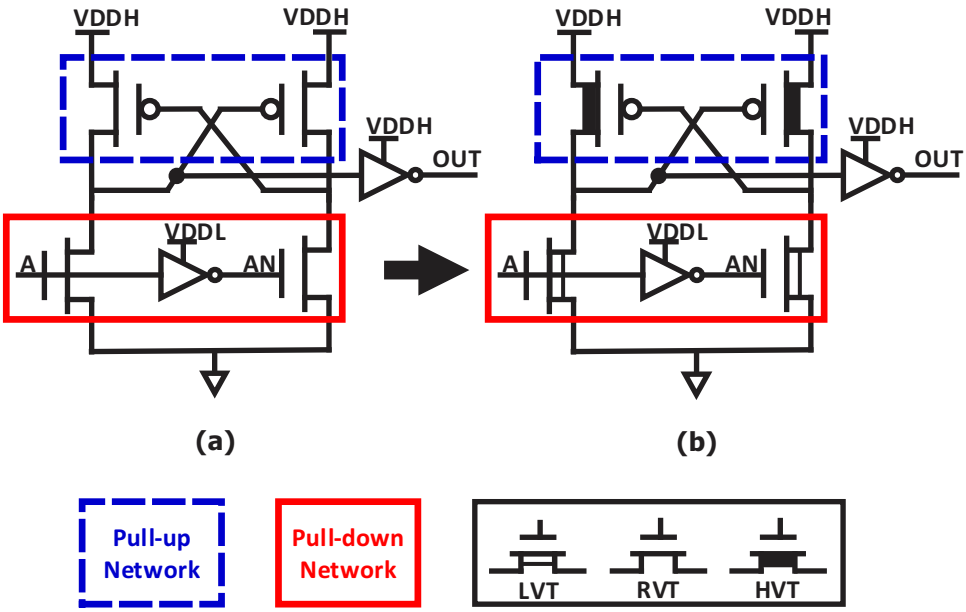


Fig. 4.1: Conventional DCVS level shifter topology.

4.2 State-of-the-Art Implementations

In order to achieve robust level conversion with subthreshold input, novel level shifter topologies and proper device choice and sizing are necessary to mitigate the prominent contention issue.

The conventional DCVS topology, as shown in Fig. 4.1, is less optimal because unreasonably large pull-down devices are required. In the chosen 65nm technology, assuming a full-RVT (regular threshold voltage device) implementation (Fig. 4.1(a)) is adopted for 300mV to 1.2V conversion, the NMOS size should be 830× of PMOS. With the availability of multiple-threshold devices, an MTCMOS implementation (Fig. 4.1(b)) can reduce the NMOS-to-PMOS ratio to around 60, however, which is still a significant overhead. To address this issue, several designs have been proposed in the past few years. And the key point to ensure the cor-

rect functionality is to balance the pull-up and pull-down strengths, while keeping minimal design overheads (e.g., area and energy, etc.).

PMOS current limiter based level shifter is proposed in [50]. By introducing a reference path to bias the current limiter devices, the pull-up network strength is weakened. However, the reference path contributes an always-on static current, resulting in high leakage power overheads.

In [51], a PMOS-diode current limiter based level shifter is proposed. In this work, the LS topology uses the drain node of the upper PMOS device as output, where the swing is limited between V_{DD} and V_{Tp}. In order to fix this issue, additional pull-down devices are inserted to pull this signal to ground. As a result, static current are introduced due to this pull-down requirements, leading to increased propagation delay and higher energy consumption.

A two stage level shifter with virtual supply topology is proposed in [52]. Through the insertion of an always-on NMOS diode gating transistor, the pull-up devices are weakened due to the virtual supply to the main conversion stage. Then the output is converted to full-swing through a latch comparator. Although this topology achieves proper functionality without additional power line, this design suffers from increased propagation delay.

A dynamic logic style subthreshold level shifter is proposed in [53]. This structure introduces additional clock synchronization circuitries between subthreshold and superthreshold voltage domains, which leads to significantly increased area and energy.

The level shifter designs in [54, 55] allows voltage conversion from 0.3V to 2.5V while still delivering decent delay and energy consumption. Thick oxide devices are used for pull-up strength reduction. However, thick oxide devices do not scale well

in advanced technologies, which might lead to area overhead and integration issues to other system blocks built with thin-oxide devices.

Recent works based on Wilson current mirror [57] and interrupted DCVS topology [58] with improved performance metrics are also demonstrated to be feasible for subthreshold operation. However, they have not been fully validated in silicon.

4.3 Proposed Level Shifter Design

This section describes the proposed level shifter design and the comparative analysis to previous designs. We first demonstrate the benefits of the proposed novel level shifter topology. Then, MTCMOS and subthreshold device sizing are considered to achieve the optimal design. At last, we show the comparative analysis to previous implementations through simulation results.

4.3.1 NMOS-Diode Current Limiter based Level Shifter

The proposed level shifter topology in this work is shown in Fig. 4.2, including the input inverter, the main voltage conversion stage and the output buffer. The key part in the main voltage conversion stage is the NMOS-diode based current limiter in the pull-up network, which is introduced on purpose to drastically reduce the current contention. Unlike the implementation of [51], we choose to use the drain of the pull-down NMOS as the output node, thereby reducing the additional pull-down devices in [51]. A full RVT implementation of the proposed level shifter is studied and we find out that the size of the pull-down devices in the proposed design can be identical to that of the pull-up devices owing to the introduction of

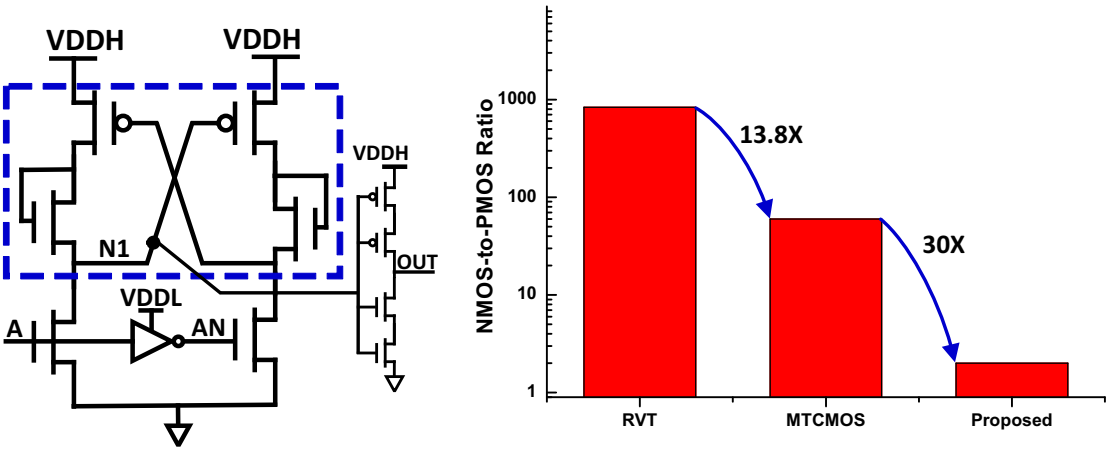


Fig. 4.2: Proposed NMOS-diode current limiter based level shifter topology with reduced pull-down device size.

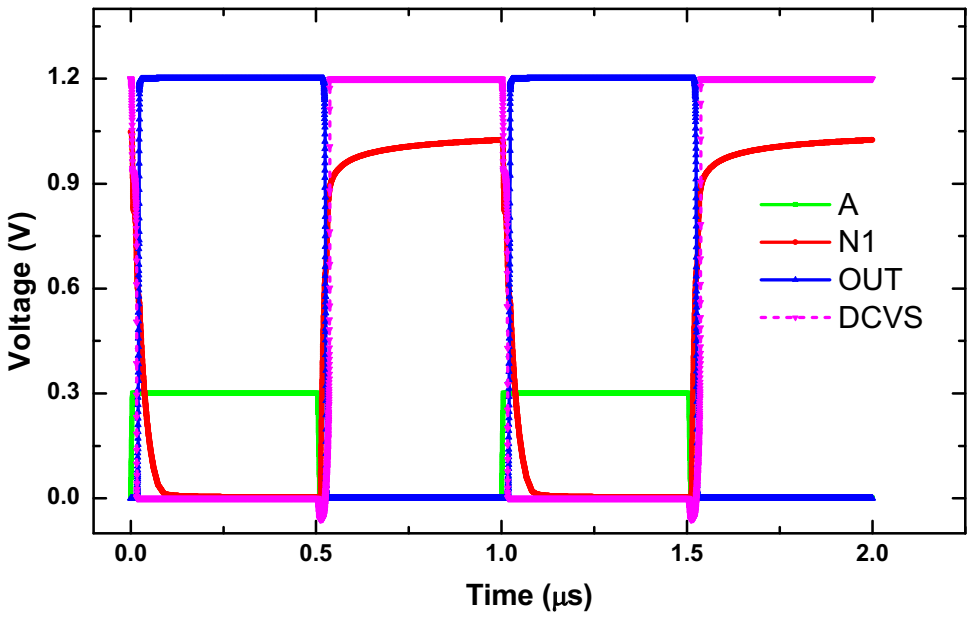


Fig. 4.3: Simulated transient waveform of the MTCMOS DCVS level shifter and the proposed level shifter.

the NMOS-diode current limiter. As a result, the NMOS-to-PMOS ratio can be reduced to 2, which is a 30× reduction in pull-down device size when compared to the MTCMOS DCVS implementation, as shown in right figure in Fig. 4.2.

The operation principle of the proposed level shifter is briefly described with the transient simulation of the level shifter. As shown in Fig. 4.3, when the input A is the low-to-high transition, the internal node N1 can be easily pulled down due to the reduced pull-up strength. In the meanwhile, the voltage droop across the NMOS diode is significantly larger, resulting in fast transition across the threshold. The node N1 is finally inverted to the output node with full swing (OUT). And the case of input A with high-to-low transition is similar. Note that the internal node N1 of the proposed design is not full swing, the output buffer may suffer from increased short circuit current. As a result, the output inverting buffer with stacking transistors should be adopted.

4.3.2 Level Shifter Optimization with MTCMOS and Subthreshold Sizing

Based on the proposed level shifter topology, the design can be further optimized through the MTCMOS and subthreshold device sizing. MTCMOS can further reduce the pull-up strength in the proposed topology. As shown in Fig. 4.4, with MTCMOS technique, the pull-up network are HVT transistors while the pull-down network are kept to be RVT transistors.

In addition, the inverse narrow width effect [32] is explored to improve the drivability of the pull-down devices. Fig. 4.5 shows the simulated NMOS threshold voltage and drain current versus the channel width in the chosen technology. As

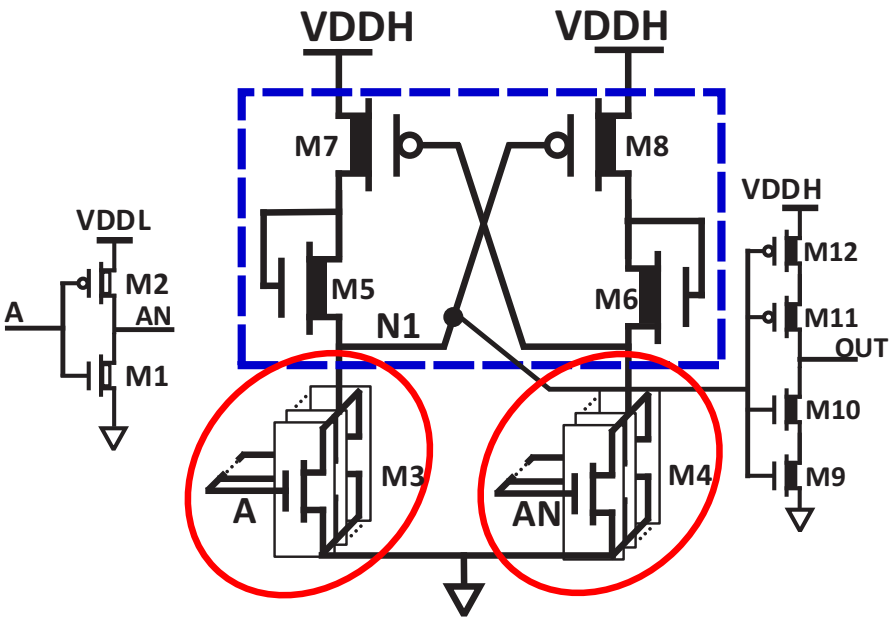


Fig. 4.4: Schematic of the optimization to the Proposed LS with MTCMOS and INWE-aware sizing.

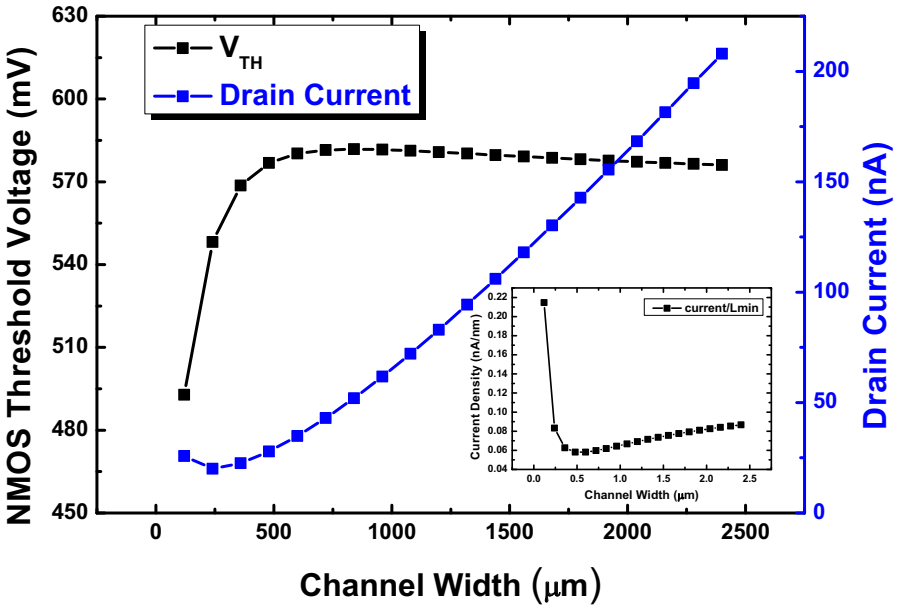


Fig. 4.5: Transient simulation of the INWE effects on NMOS.

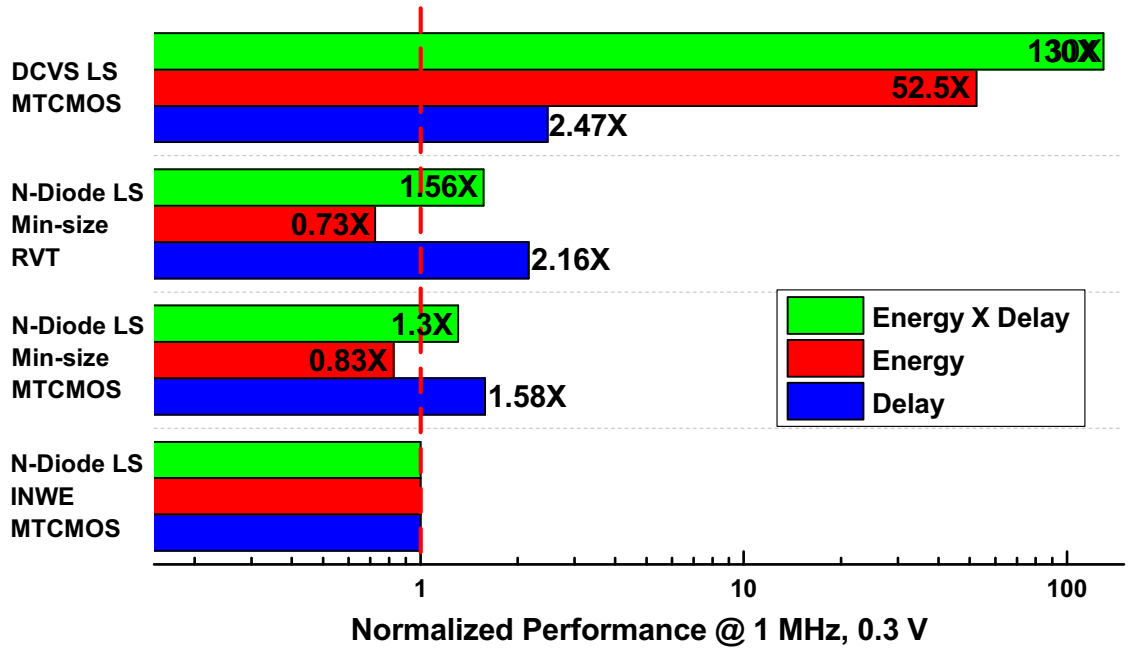


Fig. 4.6: Normalized comparison of the delay, energy and energy-delay product of the proposed level shifter with adopted optimization techniques.

can be observed, the threshold voltage experiences a deep roll-off when the channel width reduces, and the corresponding drain current shows an increase with narrow width transistor. Actually, the current density of the NMOS transistor (nA/nm) peaks at the minimum transistor width, and this can be utilized to significantly improve the level shifter performance. As a result, the pull-down network design can achieve both improved drivability and reduced device size. This also leads to reduced energy consumption due to minimal parasitic capacitance.

The proposed level shifter design and optimization strategies can be further elaborated with the following comparison. Fig. 4.6 shows the delay, energy and energy-delay product of the MTCMOS DCVS LS topology and the proposed level shifter with different optimization technique. We apply a 300mV, 1MHz and 50% duty cycle input signal with 5ns rise and fall time at 25°C, and the load of all

Table 4.1: Summary of the transistor sizing

Transistor	W/L (μm)	Transistor	W/L (μm)
M1	0.36/0.06	M2	0.6/0.06
M3	(0.12/0.06)*5	M4	(0.12/0.06)*5
M5	0.18/0.06	M6	0.18/0.06
M7	0.27/0.1	M8	0.27/0.1
M9	0.18/0.1	M10	0.18/0.1
M11	0.27/0.06	M12	0.27/0.1

LSs are a 1× strength inverter in the chosen technology. The remaining part will also use this default simulation setup if not stated otherwise. All performance metrics are normalized to the final LS design with all optimization techniques adopted. As shown in Fig. 4.6, the MTCMOS DCVS LS is efficient in neither energy nor delay metrics, while the baseline NMOS-diode based LS (built with minimum-sized RVT transistors) shows significant energy reduction due to the minimized pull-down device size. However, this incurs limited delay improvement. Further incorporating MTCMOS in the pull-up network and INWE-aware sizing in the pull-down network, total reduction compared to the DCVS LS in delay, energy and energy-delay product are 2.47×, 52.5× and 130×, respectively. Indeed, the energy/bit of the minimum sized N-Diode LS is optimal, the small delay and energy-delay product of the final design (with all optimization techniques applied) are more preferred. The transistor sizing of the final design are summarized in Table 4.1.

4.3.3 Comparative Analysis to Previous Implementations

In order to demonstrate the benefits of the proposed design, we perform a comparative analysis to previous LS designs in [57, 58]. For the sake of fair com-

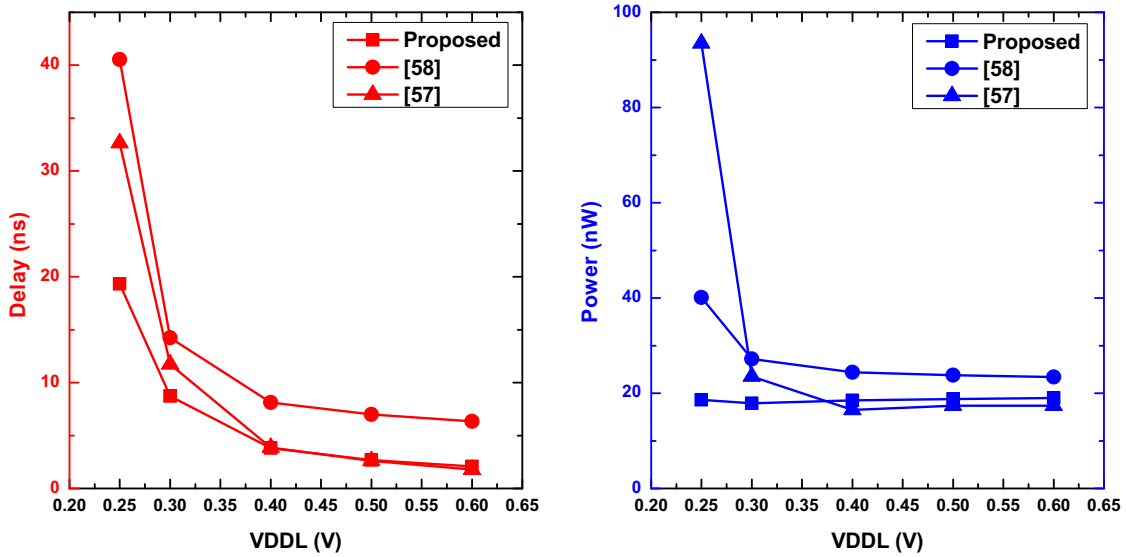


Fig. 4.7: Transient simulation of the proposed LS and the previous designs in [57, 58].

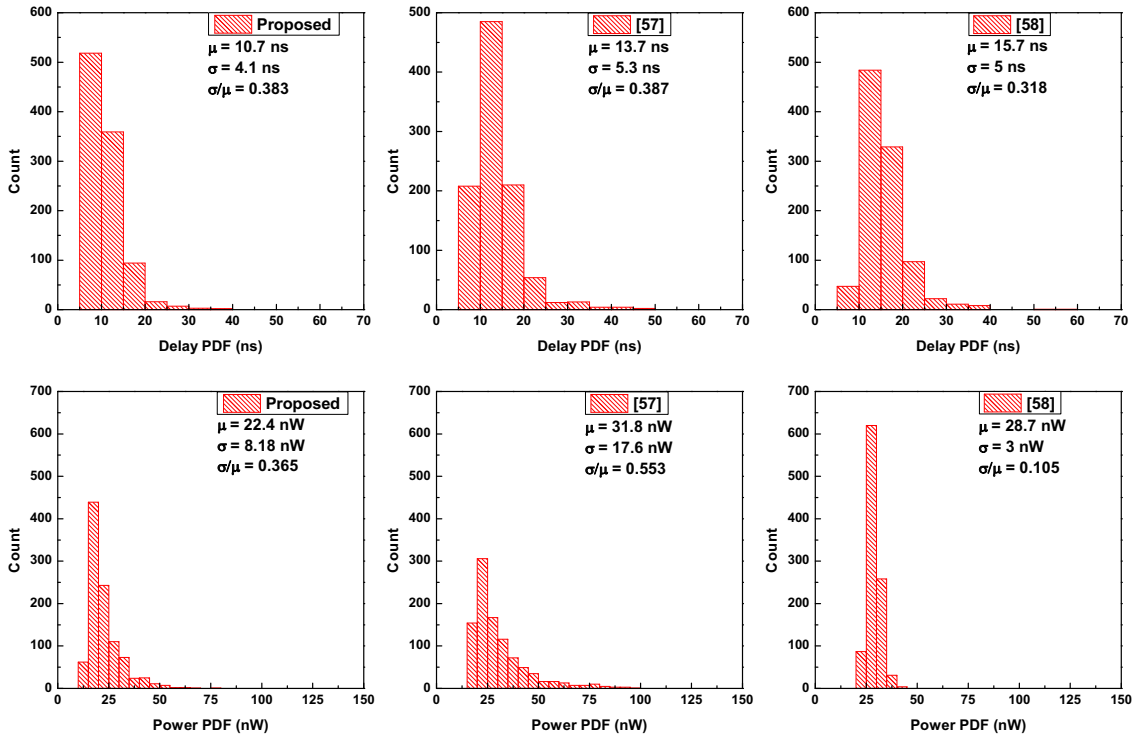


Fig. 4.8: Monte Carlo simulation of the proposed LS and the previous designs in [57, 58].

parison, we reconstruct and optimize the two implementations in the chosen technology. Due to the optimal topology discussed above, the proposed level shifter shows smaller propagation delay when compared to previous designs. As shown in Fig. 4.7, the propagation delay at 0.3V (0.25V) are 1.63× (2.1×) ([58]) to and 1.34× (1.69×) ([57]) of the proposed level shifter, respectively. Contemporarily, the power consumption of the proposed LS are reduced by 1.52× (2.16×) ([58]) and 1.31× (5×) ([57]), respectively. Compared to the design in [58] (which has a similar topology in the main shifting stage), our proposed level shifter achieves smaller propagation delay due to the fact that the pull-up network in [58] is overly weakened by the inverting input stage. As a result, the input stage of [58] has to be LVT devices. The proposed design uses only the RVT devices and the delay metric can be further improved if LVT devices are used, however, at the cost of increased leakage current.

1K-point Monte Carlo simulation at 0.3V is also performed to investigate the statistical performance of the proposed LS circuits. As shown in Fig. 4.8, the proposed level shifter shows on average delay and energy benefits over the previous designs. The mean delay of the proposed level shifter shows 1.47× ([58]) and 1.28× ([57]) delay reduction, respectively and the mean energy is reduced by 1.28× ([58]) and 1.42× ([57]), respectively.

4.4 Measurement Results and Discussions

The proposed NMOS-diode current limiter based level shifter was implemented in a 65nm Low Leakage CMOS process. Fig. 4.9 shows the die photo and the layout view of the level shifter, which occupied 16.3μm^2 (5.5μm×3.2μm) silicon area. The

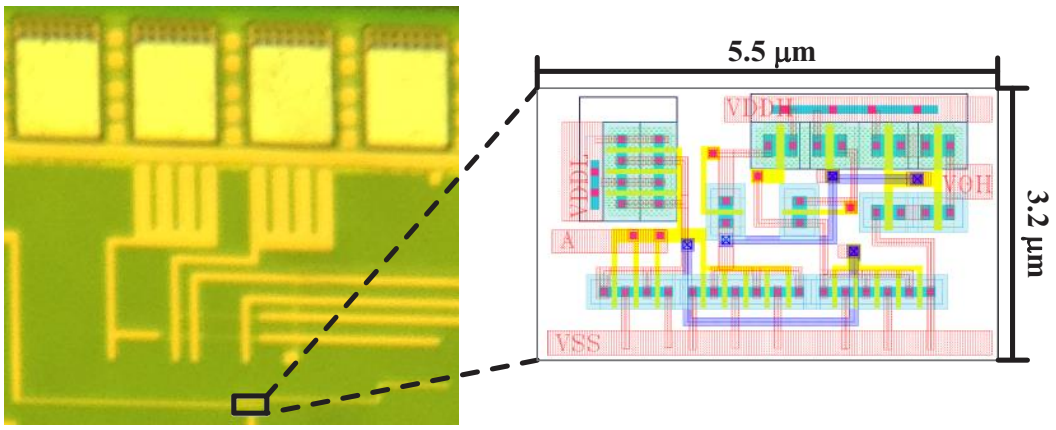


Fig. 4.9: Die photo and layout view of the proposed level shifter.

output of the level shifter is buffered through an inverter chain designed for driving the large capacitive load of the external testing equipment (up to 20pF), and the power consumption of the LS circuit was measured excluding the power of the output buffer. Both the dynamic and leakage current are on nano-ampere scale so the Agilent HP34401A was used for current measurement.

Fig. 4.10 and Fig. 4.11 show the measured waveform of the proposed level shifter. With a relaxed input frequency (10KHz), the LS circuit can successfully convert a 60mV signal into a 1.2V signal. For a 300mV input signal, the proposed level shifter can be operated at 1MHz without obvious delay. Fig. 4.12 shows the measurement propagation delay of a typical chip with different VDDL. As can be observed, the level shifter experiences an exponential delay increase when VDDL scales into the subthreshold region (below 0.5V). On the other head, when VDDL exceeds 0.5V, the LS delay is gradually saturated to a few nano-seconds.

Fig. 4.13 shows the measured statistics of the proposed LS at 0.3V, 1MHz input. The mean measured delay is 25.1ns, with standard deviation of 8ns. The minimum input voltage of the proposed LS can be as low as 60mV, while the

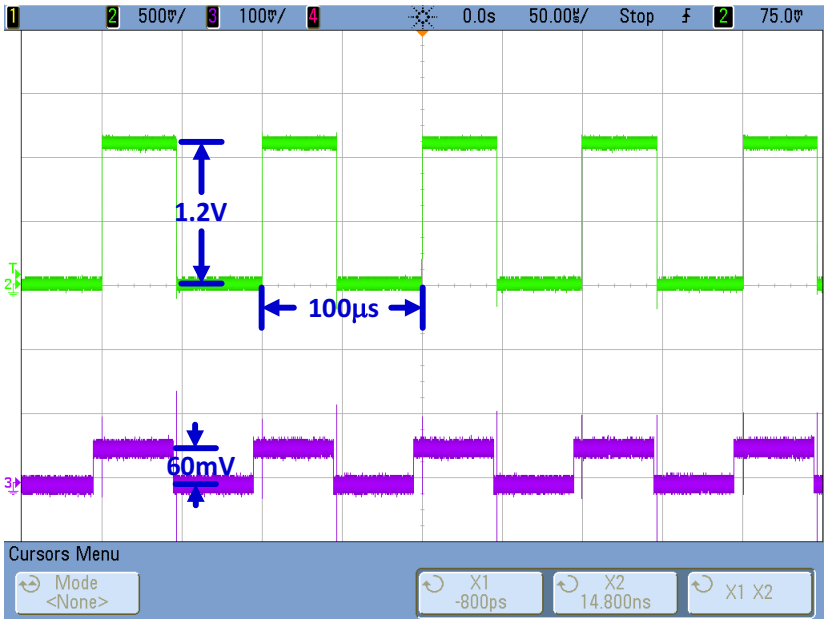


Fig. 4.10: Measured waveform of the proposed LS with a 60mV to 1.2V conversion.

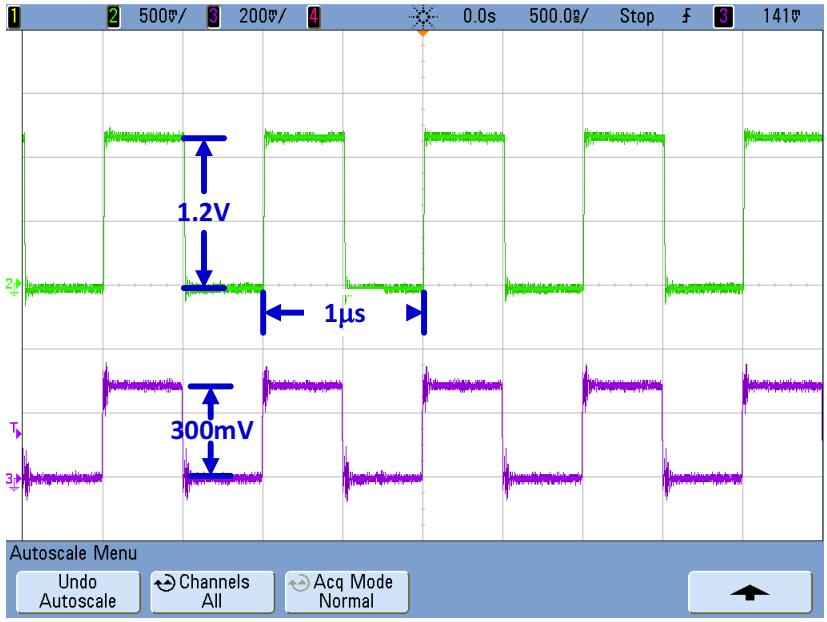


Fig. 4.11: Measured waveform of the proposed LS with a 60mV to 1.2V conversion.

Table 4.2: Comparison to state-of-the-art LS designs

	Tech. (μm)	Range (V)	Delay (ns)	Energy/bit	Leakage (nW)	Area (μm^2)
Proposed	0.065	0.1 - 1.2	25.1@0.3V	30.7fJ@0.3V, 1MHz	2.5	16.3
[50]	0.13	0.1 - 1.2	50@0.2V	25pJ@0.2V, 50KHz	8	NA
[51]	0.18	0.13 - 1.8	600@0.3V	20pJ@0.3V, NA	NA	NA
[52]	0.13	0.19 - 1.2	57.9@0.2V	NA	NA	NA
[53]	0.13	0.3 - 2.5	125@0.3V	1.7pJ@0.3V, 8MHz	NA	111800
[54]	0.13	0.3 - 2.5	41.5@0.3V	229fJ@0.3V, 5KHz	0.475	102.26
[55]	0.13	0.3 - 2.5	58.8@0.3V	191fJ@0.3V, 5KHz	0.724	71.9
[56]	0.35	0.23 - 3	10^4@0.4V	5.8pJ@0.4V, 10KHz	0.23	1880
[57]*	0.09	0.1 - 1.2	18.4@0.2V	94fJ@0.2V, 1MHz	6.6	NA
[58]*	0.09	0.18 - 1.2	21.8@0.2V	74fJ@0.2V, 1MHz	6.4	36.5

* Simulation results only.

worst case is 140mV among the 25 measured chips. The average dynamic and leakage power are 30.7nW and 2.5nW, respectively. These measurement results demonstrate that the proposed LS shows improved energy efficiency and delay, competitive area efficiency and static power when compared to the previous designs, as summarized in Table 4.2. To the author's knowledge, the measured sub-100mV minimum input is the lowest input signal level reported to date.

4.5 Conclusion and Summary

The level shifter circuits capable of converting subthreshold input are becoming essentially important. In this chapter, we proposed a novel NMOS-diode current limiter based level shifter topology and optimization techniques through MTCMOS and INWE-aware sizing. The proposed design was fabricated in a 65nm low leakage process and the measurement results across 25 dies validate the proposed design, which can convert an input signal as low as 100mV (on average) to 1.2V supply. In addition, the proposed level shifter shows on average 25.1ns propagation delay, 30.7fJ/bit energy and 2.5nW leakage power when converting a 300mV input to

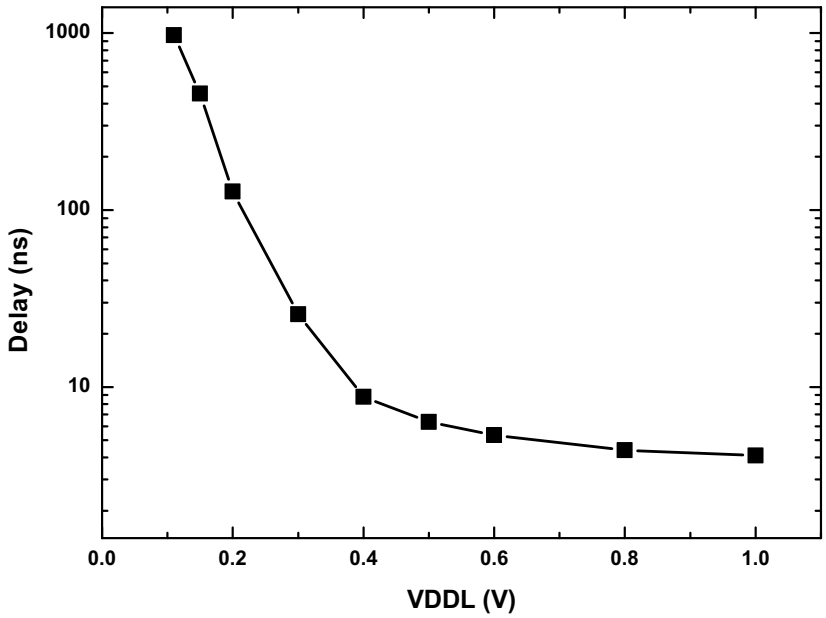


Fig. 4.12: Measured LS delay from a typical die.

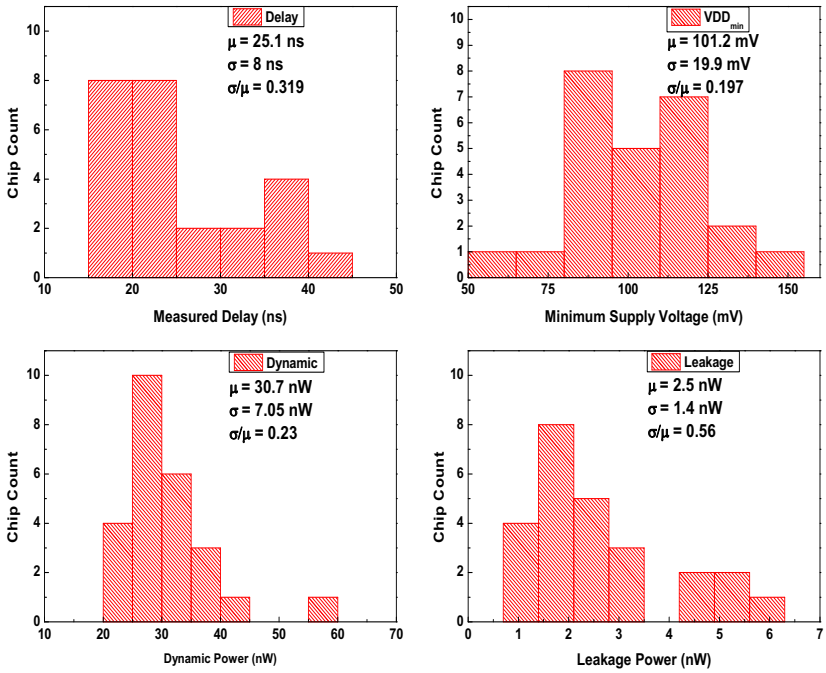


Fig. 4.13: Measured statistics of the proposed LS: delay, VDD_{min}, dynamic power and leakage power.

1.2V, occupying only 16.2μm^2 silicon area. In summary, the proposed LS circuit outperforms previous designs in several aspects, which is suitable for the ULV designs.

Chapter 5

Robust and Energy-Efficient Ultra-Low Voltage Standard Cell Design with Intra-Cell Mixed-V_{th} Methodology

---

High functional yield is one of the key challenges for subthreshold standard cell designs. Device upsizing is a commonly used but sub-optimal method due to its overheads in energy and area. In this chapter, we propose a robustness-driven intra-cell mixed-V_{th} design methodology (MVT-ULV) for the robust ultra-low voltage operation. It uses low threshold voltage transistors in the weak pulling network of logic gates to enhance the robustness. It guarantees the high functional yield with the minimum energy/area overheads. We demonstrate on a commercial 65nm CMOS process that, our proposed design methodology shows up to 60mV and

110mV robustness improvement at 300mV power supply voltage over the commercial library cells and the cells built with previous Leakage-Minimization mixed-V_{th} methods (MVT-LM) under the same cell area constraints, respectively. In addition, the proposed MVT-ULV library enables ITC'99 benchmark circuits to show on average 30.1% and 78.1% energy-efficiency improvement when compared to the libraries built with the device-upsizing methods and the previous MVT-LM methods under the same yield constraints, respectively.

5.1 Introduction

The compelling energy benefit of aggressive voltage scaling into the subthreshold region has been demonstrated on digital VLSI circuits in recent years for performance-relaxed applications. However, the on/off current ratio degradation and the sensitivity to process variations bring big challenges like the functional yield loss or logic failure to subthreshold logic circuits.

Prior researches on subthreshold logic design have been introduced in [6, 7, 30–35]. Several works focus on timing and energy optimization [6, 7, 30–33]. However, these works do not take functional yield into consideration. Therefore, a few works are proposed for yield enhancement for subthreshold logic design. Variation-aware logic designs are proposed for improving the logic cell robustness [34, 35] with the evaluation criterions like the butterfly plot or the logic gate output voltage swing level. Under a target functional yield constraint, device upsizing is an effective solution to reduce the functional failure for subthreshold logic. However, this design strategy performs trades-off with both the standard cell layout area and the power consumption overheads due to the increased device dimension.

Recently, a gate-level multi-V_{th} design technique is explored and demonstrated to be beneficial for subthreshold operation [111]. However, the robustness is still an open issue for multi-V_{th} logic library designs. The multi-V_{th} logic cells are built with monolithic threshold voltage devices. Therefore, the conventional device upsizing technique is still required and both energy and area overheads are inevitable.

In this chapter, we propose a novel and orthogonal subthreshold intra-cell mixed-V_{th} (MVT-ULV) logic design methodology and demonstrate its applicability on a commercial 65nm MTCMOS process for robustness enhancement with the minimum energy/area overheads. Previous intra-cell mixed-V_{th} design techniques [112, 113] have been proposed for the leakage optimization in super-threshold designs. However, they cannot be directly applied to the subthreshold region.

Our proposed MVT-ULV method aims at improving the cell robustness with the minimum layout area overhead. In other words, the proposed method can achieve the total energy reduction and area minimization under a target yield constraint. The basic idea of building MVT-ULV cells is to replace the selected regular-threshold-voltage (RVT) transistors with the low-threshold-voltage (LVT) ones in the weak pulling networks (either pull-up or pull-down), for instance, the stack transistors in NAND/NOR gates in the C^2MOS logic family. Modern CMOS processes generally support multi-V_{th} devices ($\Delta V_{th}\approx$100-150mV) and this replacement is feasible by only manipulating the CAD layers and proper device sizing. We demonstrate that logic cells built with a monolithic device choice, either RVT or LVT, require the similar upsizing strategy to improve functional yield. However, the proposed logic design can enhance robustness with the minimum upsizing overheads, which eventually reduces the standard cell footprint as well as energy

consumption.

The following part of this chapter is organized as follows. In Section 5.2, we cover the related works on the subthreshold logic robustness and the transistor-level mixed-V_{th} design. In Section 5.3, we introduce our proposed subthreshold intra-cell mixed-V_{th} design methodology for combinational logic gates and flip-flops. In Section 5.4 and Section 5.5 we validate the effectiveness of the proposed MVT-ULV library through same-area constraint and same-yield constraint, respectively. We conclude the chapter in Section 5.6.

5.2 Related Work

5.2.1 Subthreshold Logic Robustness

In [6], the authors claim that the minimum-sized commercial cell libraries are energy-optimal for subthreshold designs. However, the modern commercial cell libraries overlook the logical effort sizing strategy to maintain the compact layout footprint, as illustrated in Fig. 5.1(a). In this way, the robustness of the commercial logic cells can be problematic when operated in the subthreshold region.

Due to the degradation of I_{ON}/I_{OFF} current ratio and the increased sensitivity to process variations, subthreshold logic cells suffer from reduced output swings, which are the root cause of the subthreshold logic failure. The worst case of such failure mechanism can be seen in the logic gates (eg., NAND/NOR), which are composed of several stack transistors together with complementary parallel transistors. The leakage current from the parallel transistors degrades the output swing and causes the functional failure, as shown in Fig. 5.1(a).

The butterfly plot together with the corresponding output voltage swing [34]

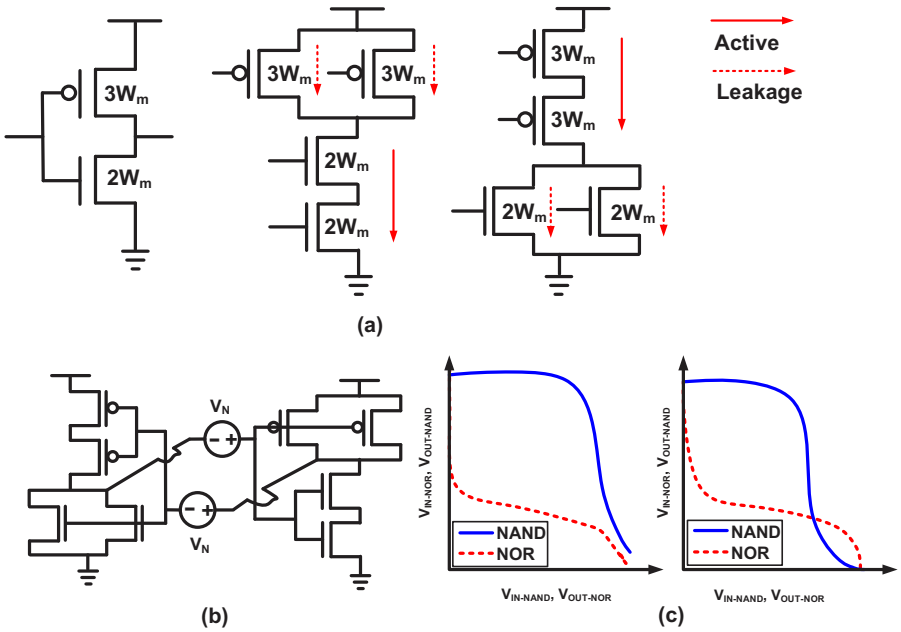


Fig. 5.1: (a) schematics of commercial standard cells and subthreshold logic failure mechanism, (b) cross-coupled NAND/NOR pair and, (c) example of butterfly plot.

can be used to indicate the functional failure rate. The butterfly plot is derived via the Monte-Carlo analysis of the voltage transfer characteristic (VTC) curves of the cross-coupled inverters (eg., NAND/NOR pair), as shown in Fig. 5.1(b). The worst case VTC curves are used to form the butterfly plot and to evaluate whether the two bi-stable points are observed (Fig. 5.1(c)). Both the global and local variations cause the VTC curve shifts and degrade the static noise margin (SNM) of the butterfly plot. Practical solution to enhance the functional yield is to upsize the devices to mitigate the process variations. However, the device upsizing solution contradicts the energy minimization goal due to the increased device size. Also, area increase is inevitable.

5.2.2 MVT-LM Design Technique

Gate-level multi-V_{th} design technique [111] is proposed for ULV circuits to achieve enhanced performance and energy efficiency. They still stick to the traditional design strategy with monolithic device choice at the gate-level. Further going beyond the gate-level multi-V_{th} design techniques, transistor-level (intra-cell) mixed-V_{th} designs are proposed for Leakage Minimization (MVT-LM) with the minimum overhead in critical path delay. However, these designs haven't been validated for ULV applications. Therefore, we will elaborate several related works [112, 113] and discuss the applicability of MVT-LM to subthreshold logic designs.

Combinational Cell Design

The previously proposed MVT-LM cell design in [112] achieves leakage reduction through the RVT device assignment while preserving the worst-case timing arc delay. Transistors are assigned with the RVT devices when it is not on the critical timing arcs. Examples of 2-input MVT-LM NAND/NOR gates are illustrated in Fig. 5.2(c). The design strategy in [112] improves leakage reduction with the minimum delay overhead. However, the cell robustness is further deteriorated as the stacking network of the MVT-LM cell is even weaker when compared to the RVT/LVT cell with monolithic devices. Butterfly plots of three NAND/NOR pairs are shown in Fig. 5.2(d) with supply voltage of 300-mV in a 5k-point Monte-Carlo (3σ) simulation. As illustrated, the MVT-LM pair yields diminishing SNM (approaching 0-mV) at 300-mV, which is even worse than the RVT/LVT pair.

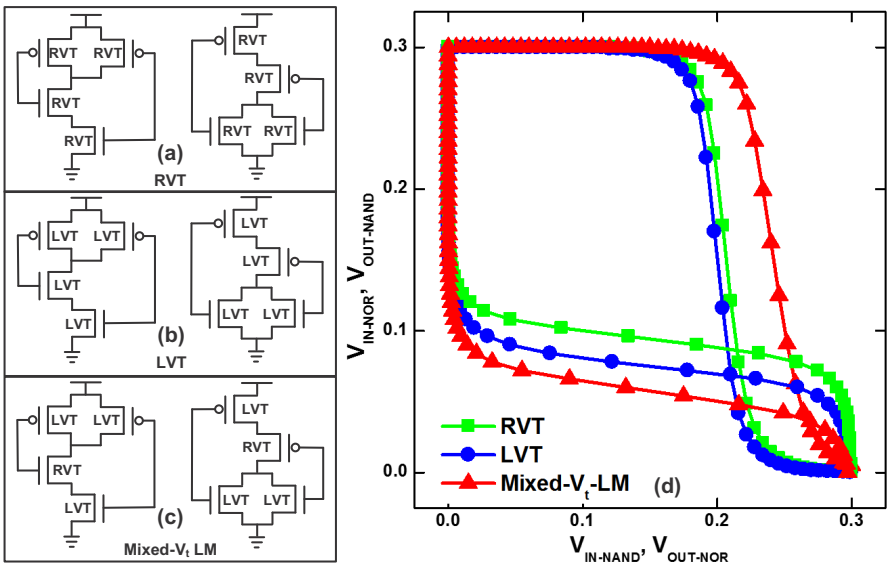


Fig. 5.2: NAND/NOR pairs of (a) RVT, (b) LVT, (c) previous MVT-LM technique, and (d) butterfly plot of three pairs.

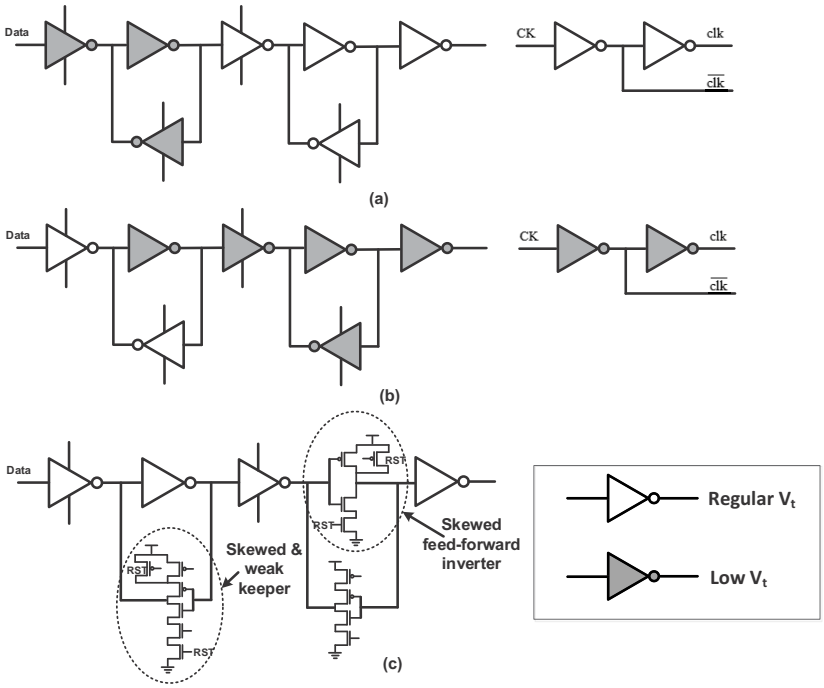


Fig. 5.3: Previous MVT-LM flip-flop design, (a) Mixed-I, and (b) Mixed-II, and (c) flip-flop with reset function.

Sequential Cell Design

MVT-LM flip-flops are also investigated in the superthreshold region for leakage reduction due to the significant leakage contribution from sequential cells. The idea in [113] is to assign RVT devices in either the master or the slave stage of the master-slave flip-flops. Thus, the delay is increased in either the setup time or the clock-to-q delay. This technique avoids the abrupt timing change and still retains the good leakage reduction purpose. Fig. 5.3(a-b) show the original design of [113].

Nevertheless, these two sets of designs do not contribute to better robustness of the ULV flip-flops. Moreover, flip-flops with reset/set functions are extensively used in digital VLSI designs. Therefore, it is more practical if we take a flip-flop with the asynchronous reset function (Fig. 5.3(c)) for demonstration. As annotated in Fig. 5.3(c), the logic failure of the reset flip-flop is due to:

- Skewed Feed-Forward (FF) inverter in slave stage latch, and
- Skewed and minimum-sized weak Feed-Back (FB) keeper in the master stage latch.

Similar to the previous discussed method in section 5.2.1, the assignment based on [113] cannot recover the above-mentioned failure because every single inverter in the reset flip-flop is still built with monolithic RVT/LVT devices. Thus, the robustness of reset flip-flop cannot be optimized.

5.3 MVT-ULV: Robustness-Driven Mixed-V_{th} for ULV Operation

As discussed in previous sections, the logic design based on the monolithic device, either RVT or LVT, is more prone to functional failure due to the weak stack devices and the process variations in the subthreshold region. Besides, the previous intra-cell MVT-LM techniques are optimized for leakage reduction, which are demonstrated to be less beneficial to the logic robustness in subthreshold operation. In this section, we will introduce our proposed robustness-driven MVT-ULV logic design methodology.

Fig. 5.4(a) shows the diagram of the proposed MVT-ULV 2-input NAND cell design for robustness enhancement. The idea is to assign LVT devices to the weak pulling-down network in NAND gates. The similar concept is also applicable to the NOR gates as well. Through this assignment, the VTC curves under process variations can be tightened and eventually lead to enhanced logic robustness. As can be observed in Fig. 5.4(b), the VTC curves (NAND-NOR pair) of MVT-ULV cells under 3σ Monte-Carlo simulation (5k-point) are significantly tightened when compared to the RVT cells.

It is worth noting that there are other design variants under the proposed design concept. Although several other intra-cell mixed-V_{th} variants exist (24 variants for 2-input NAND/NOR gates), MVT-ULV restricts LVT devices to the parallel structures due to the robustness consideration. Thus, there are two other variants that can also improve the robustness in the NAND/NOR logic design, as shown in Fig. 5.4(a). The choice can be determined by choosing the design variants with maximum SNM improvement because it is mostly influenced by the threshold

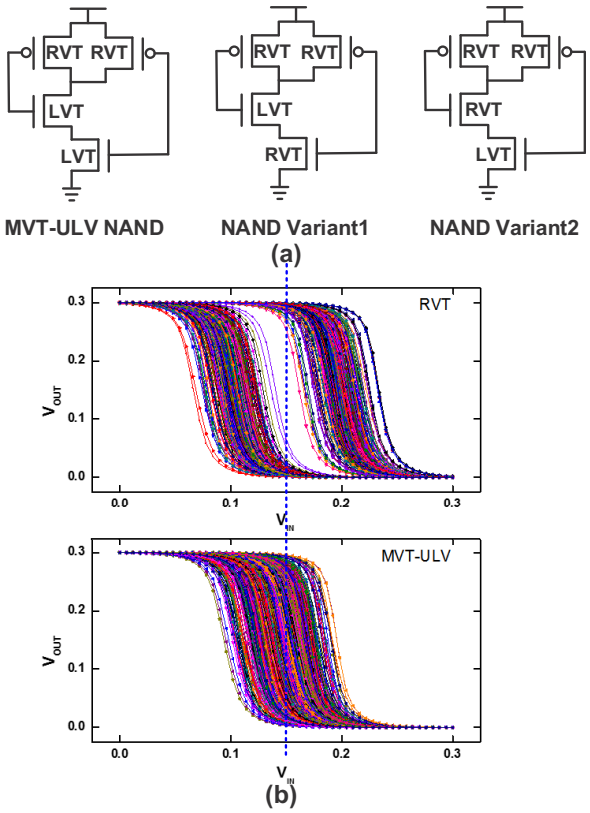


Fig. 5.4: (a) MVT-ULV NAND cell design with other possible variant cells, and (b) Monte-Carlo simulation of the NAND-NOR Pair VTC curves.

voltage difference between the RVT and LVT devices in the chosen process.

The design of complex logic gates, such as AOI or OAI cells, does not completely follow the previous design strategy. For such cells, both their stack and parallel structures co-exist in either the pull-up or the pull-down networks, making the relative strength less different when compared to the cells like NAND/NOR. Previous strategies on NAND-NOR pairs may have negative effects on AOI/OAI cell robustness. Therefore, the LVT devices should be assigned only to the bottleneck transistors. For example, in AOI21/OAI21 cells, the pulling network is dominated by the single stack transistors annotated in Fig. 5.5.

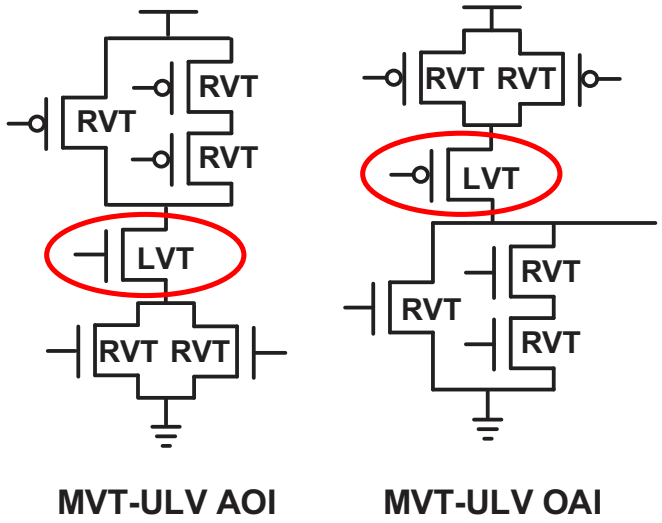


Fig. 5.5: MVT-ULV AOI21/OAI21 cell design with annotated bottleneck transistors.

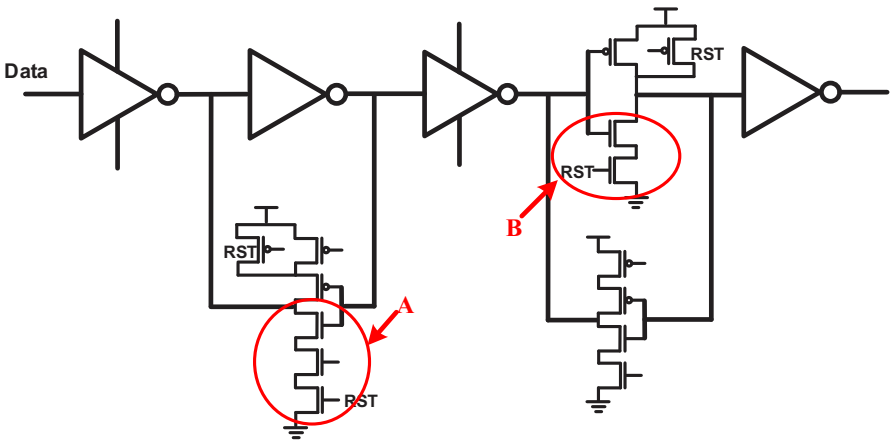


Fig. 5.6: MVT-ULV flip-flop cell with asynchronous reset.

With the previous knowledge, the MVT-ULV method can also be applied to flip-flop cells with the reset functions, as shown in Fig. 5.6. For the case of a slave stage latch with the FF inverter, it is similar to the NAND-NOR logic cells. While for the master stage latch with the FB inverter, the AOI/OAI design strategy can be adopted.

5.4 Experimental Results: Iso-Area Constraint

In this section, we provide more simulation results to demonstrate the effectiveness of the proposed robustness-driven MVT-ULV methodology. The pre-assumption is that all cells with the same logic functions are limited with the iso-area for the fair comparison of robustness. We select several logic cells from a 65nm commercial RVT standard cell library. For the iso-area layout footprint, we only replace RVT stack devices into LVT. This can be achieved through an additional device layer, which means no area overheads. Compared to the weak current in the stack transistors in monolithic (RVT/LVT) logic cells, the degraded on/off current ratio is recovered by using LVT stack devices. In addition, we also built LVT and MVT-LM cells for comparison.

5k-sample Monte-Carlo (MC) simulations are run for each complementary logic cells to obtain the worst-case VTC curves and the corresponding butterfly plots. Fig. 5.7 shows the four sets of butterfly plot of NAND-NOR pairs with the supply voltage of 300mV at room temperature of 25°C. As can be observed, the proposed MVT-ULV design shows approximately 2X SNM improvement over the RVT/LVT design, while up to 11X SNM improvement over the MVT-LM cells. More data of SNM and VDD_{min} of other logic cells are summarized in Table 5.1. The proposed

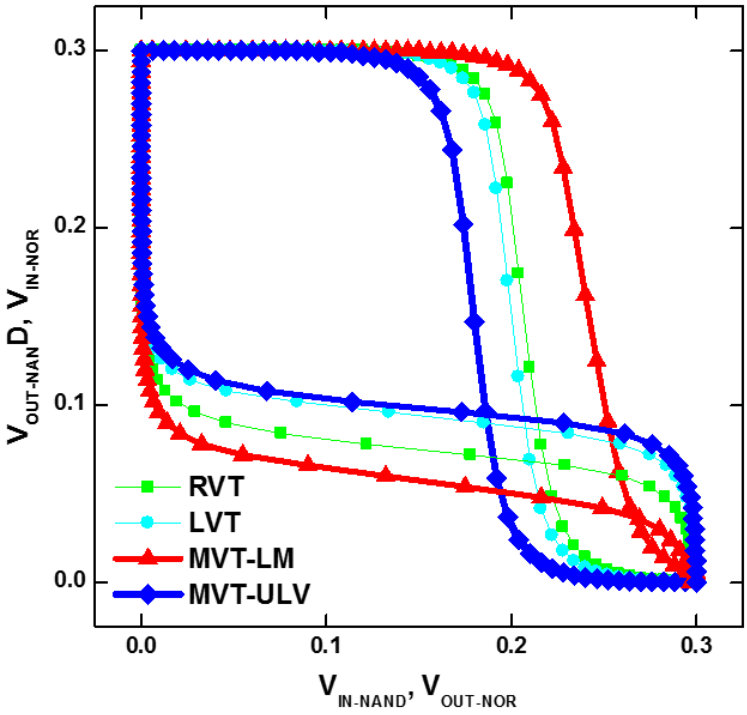


Fig. 5.7: Butterfly plots of four sets of 2-input NAND-NOR pairs.

MVT-ULV cells are consistently better in term of SNM, which indicates a lower minimum supply voltage.

We also investigate the variant designs in this process. As shown in Fig. 5.8, both variant designs show improved SNM when compared to the RVT designs. However, the improvement is less significant when compared to the chosen MVT-ULV design in the chosen process. It is possible that for another technology with large differences in the threshold voltage between RVT/LVT devices, variants may be better alternatives.

In addition, temperature is another dominant factor for subthreshold logic robustness. Commonly, the increased temperature will cause exponential increase in gate leakage and the I_{ON}/I_{OFF} ratio is further degraded. This makes the logic

Table 5.1: Comparison of SNM (VDD=300mV, 25°C) and VDD_{min} of several logic cells under different logic design methods

		NAND2-NOR2 X1	NAND2-NOR2 X2	NAND3-NOR3 X1	AOI21-OAI21 X1	Async reset DFF[α]
SNM	RVT/LVT	58mV/69mV	95mV/100mV	30mV/18mV	55mV/75mV	FB[β]: 52mV/48mV
	MVT-LM	11mV	68mV	-2mV (failed)	28mV	FF[χ]: 58mV/53mV
	MVT-ULV	**120mV**	**165mV**	**75mV**	**90mV**	**FB: 62mV** **FF: 82mV**
VDD_{min}	RVT/LVT	240mV/230mV	203mV/194mV	270mV/280mV	245mV/225mV	FB: 248mV/252mV
	MVT-LM	290mV	230mV	305mV	270mV	FF: 242mV/247mV
	MVT-ULV	**180mV**	**136mV**	**225mV**	**210mV**	**FB: 240mV** **FF: 218mV**

[α] The MVT-LM FF is built based on [14], therefore the latch designs are either RVT or LVT.

[β] FB: master stage latch containing feed-back keeper

χFF: slave stage latch containing feed-forward inverter

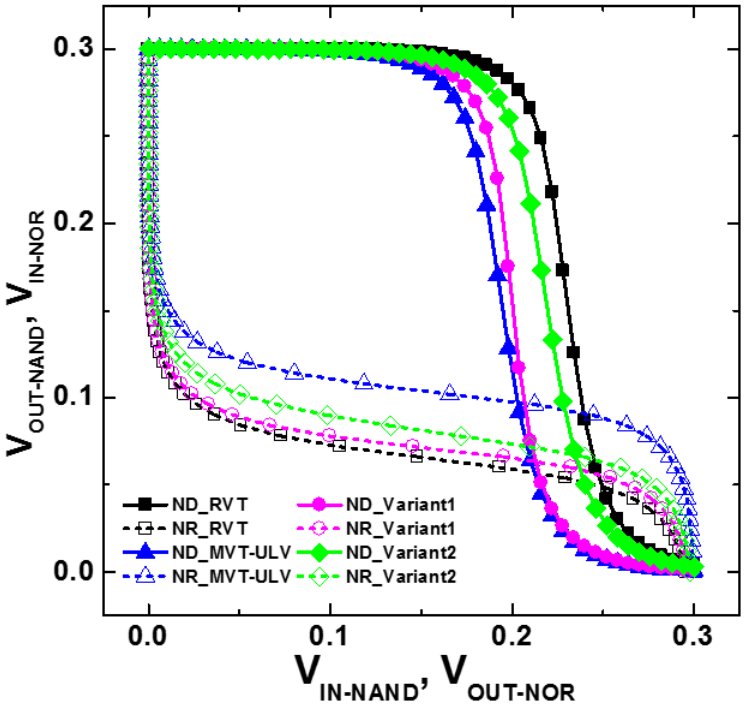


Fig. 5.8: Butterfly plots of variant design of 2-input NAND-NOR pairs.

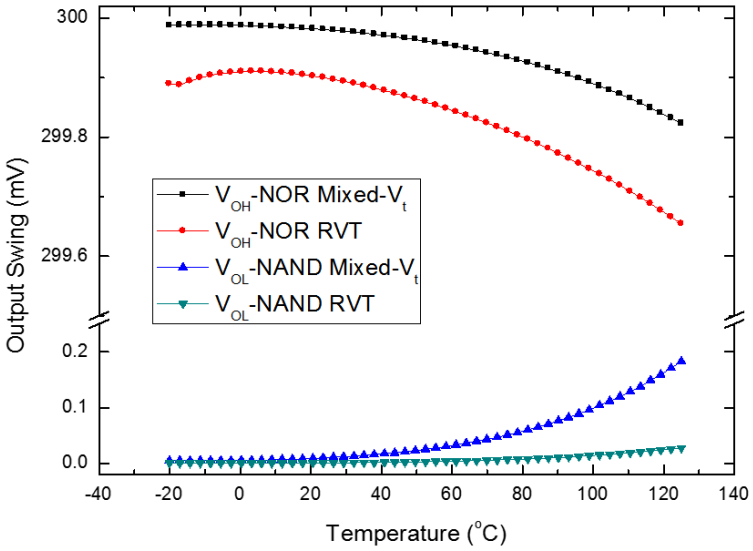


Fig. 5.9: Temperature effects on logic output swing.

cell functionality most vulnerable at high temperature corners. The output swing voltage levels of V_{OL} (NAND) / V_{OH} (NOR) gates across the -20 to 125 ^{o}C range are plotted in Fig. 5.9. As shown in Fig. 5.9, the robustness of the commercial cells is severely degraded at high temperature corners while the proposed MVT-ULV cells still show better robustness across a wide operation temperature range.

5.5 Experimental Results: Iso-Yield Constraint

The previous discussions indicate that the proposed MVT-ULV logic cells show a significant robustness improvement over the commercial libraries cells with the same area constraint. This indicates the potential energy benefits when a target functional yield constraint is applied. In this section, we compare the proposed MVT-ULV design with the conventional device upsizing solution.

5.5.1 Cell Level Evaluation

With the purpose to achieve the target functional yield constraint, the conventional logic cells are upsized to mitigate process variations and process skews. The output swing failure rate is used as a quantitative indicator for a certain functional yield [34, 35]. Under this framework, the proposed MVT-ULV requires the minimum upsizing to achieve the same functional yield.

Fig. 5.10 shows the output failure rate versus the normalized device width of 2-input NAND cells in three sets of designs (RVT, MVT-LMMVT-ULV) at both room temperature and the high temperature. As shown, with 0.1% failure rate at 27^oC and 0.5% failure rate at 125^oC, no extra device upsizing is required for the proposed MVT-ULV cells. However, significant device upsizing up to 4X and 6X are required for RVT and MVT-LM cells to achieve the target failure rate, respectively. In this case, the proposed MVT-ULV NAND gate gains around 4X and 2X reduction in active device area and layout area over the upsized RVT cells, respectively. Similar situations can also be found during the design of other logic cells. Under the same yield constraint in this work, for 1X strength AOI21/OAI21 cells, up to 1.6X/2X active device area reduction can be observed, respectively. Compared to the upsized flip-flops with approximate 2X area and higher clock loading, the proposed MVT-ULV flip-flop achieves the same robustness with no overheads in area and clock loading.

Due to the increased device capacitance, the circuit delay and power consumption under the conventional upsizing scheme increase. The FO4 delay and power consumption of ten-stage NAND-NOR-inverter-chain are examined at 0.3V with a 5k-point Monte-Carlo simulation. As shown in Fig. 5.11, the commercial cell shows the worst balance between the rise and fall time, indicating the worst robustness

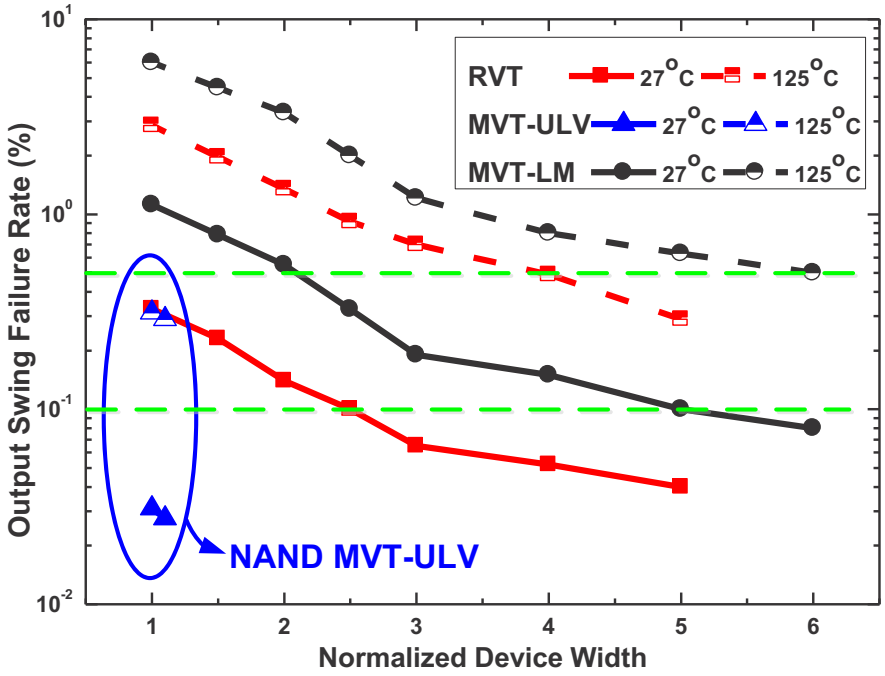


Fig. 5.10: Output swing failure rate of three sets of standard cells, RVT, MVT-LM, MVT-ULV under 27°C and 125°C.

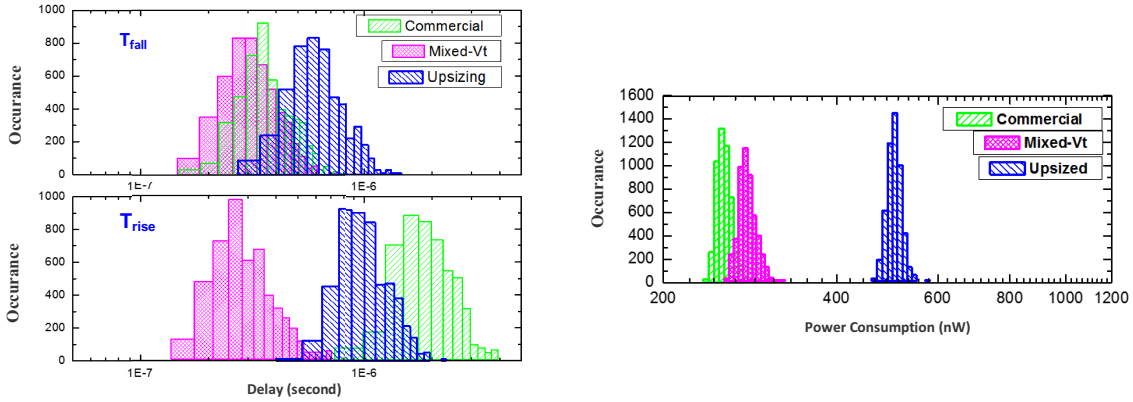


Fig. 5.11: Delay and power distribution of three 10-stage NAND-NOR chains (commercial, upsized and MVT-ULV).

for the subthreshold operation. The MVT-ULV design shows improved mean delay over the other two designs and statistically balances the rising and falling timing. Under the same functional yield constraint, the MVT-ULV design shows up to 40% switching power reduction when compared to the upsized design.

5.5.2 Library Level Evaluation

In order to demonstrate the effectiveness of the proposed MVT-ULV methodology for subthreshold operation, we designed four sets of standard cell libraries (RVT, LVT, MVT-LM, MVT-ULV) with the same functional yield constraint (0.1% @ 27°C and 0.5% @ 125°C) in a 65nm commercial CMOS process at 0.3V. Both RVT and LVT libraries are designed with the conventional upsizing (weak pulling network upsizing) technique according to [34]. The MVT-LM library is designed based on [112, 113] with the similar upsizing strategy with RVT/LVT libraries, and the MVT-ULV library is designed according to our proposed methodology. For all libraries, 12 logic functions (inverters, nand2, nand3, nor2, nor3, mux2, xor2, DFF with reset function, etc.,) with 3 driving strengths are included. The standard cells are designed with 9-track compact layout templates and the cells are characterized at 0.3V at TT corner of 27°C.

ITC'99 benchmark circuits are selected with six different designs (from small to large circuits). Synthesis results are listed in Table 5.2. As shown in Table 5.2, the maximum frequency of both MVT-LM and the proposed MVT-ULV library are similar, and are on average 51.6% improvements over the RVT library. The LVT library still shows to be best in term of the operating frequency.

For the energy efficiency metric, the proposed MVT-ULV library outperforms the other three library sets by 33.3%, 30.1% and 78.1% over RVT, LVT and MVT-

Table 5.2: Synthesis results of the ITC'99 benchmark circuits.

	Maximum Frequency (MHz)				Area (um^2)				Total Power Consumption (nW)				Energy Efficiency (nW/MHz)			
	RVT	LVT	MVT ULV	MVT LM	RVT	LVT	MVT ULV	MVT LM	RVT	LVT	MVT ULV	MVT LM	RVT	LVT	MVT ULV	MVT LM
B01	0.71	3.3	**0.9**	0.9	140.8	127.8	**126.7**	166	2.5	10.9	**2.82**	4	3.5	3.3	**3.1**	4.4
B03	0.52	2.5	**0.77**	0.79	582.8	550.8	**503.6**	594	5.8	27.7	**7.1**	9.3	11.1	11.1	**9.2**	11.8
B04	0.43	1.85	**0.68**	0.6	1828	1758	**1488**	3343	12.1	53.6	**14.5**	33.8	28.1	29	**21.3**	56.3
B12	0.35	1.8	**0.6**	0.62	2998	2999	**2682**	3661	12.6	62	**14.9**	22.7	36	34.4	**24.8**	36.6
B14	0.15	0.8	**0.26**	0.25	18494	18954	**12454**	23102	60.1	303	**64.6**	143	401	379	**248**	550
B15	0.21	0.96	**0.28**	0.26	20142	19146	**15303**	20330	58.5	265	**61.2**	93.9	279	276	**219**	361

LM, respectively. This is expected according to previous discussions. Due to the upsized devices to achieve the target functional yield, the MVT-ULV requires the minimum upsizing and eventually achieves better energy efficiency. In addition, the area benefits of MVT-ULV library are 23.6%, 19.5% and 54.7% over RVT, LVT and MVT-LM library, respectively. These results indicate that our proposed MVT-ULV logic design methodology is promising to achieve the robust and energy-efficient designs for the subthreshold operations.

5.6 Conclusion

In this chapter, we propose a novel intra-cell mixed-V_{th} design methodology for robust subthreshold standard cell library based designs. The proposed MVT-ULV design methodology shows up to 60mV and 110mV SNM improvement at 300mV supply voltage over the commercial library cells and the cells with the previous Leakage-Minimization mixed-V_{th} method (MVT-LM) under the same cell area constraints, respectively. In addition, the proposed MVT-ULV library enables on average 30.1% (over RVT/LVT library) and 78.1% (over MVT-LM

library) energy-efficiency improvement under the same yield constraints, respectively. These promising results demonstrate the robustness and energy efficiency benefits of the proposed intra-cell MVT-ULV standard cell library. Future work will extend this design methodology for processes (45nm and below) with significant second order geometrical effects.

Chapter 6

Exploring Energy Efficiency in Embedded DRAM

In this chapter, we explore the feasibility of using logic-compatible gain-cell based embedded DRAM (eDRAM) as a potential memory alternative targeting for ULV/ULP systems. The higher density provided by the eDRAM allows higher memory capacity to be integrated to the volume-limited ULV/ULP systems. However, the dynamic nature of eDRAM requires the memory cells to be periodically refreshed for data integrity, and this refresh operation reduces the memory access availability as well as increases the energy overheads due to the refresh operation.

To solve these, we first focus on a hidden-refresh scheme to allow the eDRAM to operate with a SRAM-like interface, without dedicated refresh cycles for data retention. This ensures 100% availability memory access, thereby reducing the design effort for system integration. Also, low-voltage eDRAM behavior is not fully exploited at the moment. Therefore, we explore the voltage scaling behaviors

of the eDRAM under the proposed hidden-refresh scheme.

6.1 Background

SRAM is one of the indispensable circuit building blocks in today's System-on-Chips (SoC) designs and SRAM generally occupies large silicon area and power. 6T-SRAM is the most prevalent memory type, which provides fast operation speed and high density. However, the stringent density requirements push SRAM bitcells to be extremely small in each technology generation, therefore the 6T-SRAM is very sensitive to local variations and prone to functional failure due to the significant reduced read/write margins. Voltage scaling will further worsen the situation and generally the 6T-SRAM fails to be fully functional below 0.6V.

The emerging SRAM designs capable of ULV operation are also proposed for aggressive power reduction purposes. The contradictory read/write optimization strategies complicate the ULV SRAM designs and ULV SRAM bitcells generally adopt more transistors (7T to 10T) [37, 38, 40–47] to maximize the read/write margins. As a result, this incurs reduced density which is a disadvantage when large memory capacity is needed for enhanced computational power.

Recent circuit [48, 114–119] and architecture [120] researches in memory/cache designs explore the gain-cell based eDRAM as a potential alternative to the SRAM. When compared with SRAM, the gain-cell based eDRAM can provide improved memory density. In addition, since no direct leakage path exists in the gain-cell eDRAM. Therefore, the leakage current of eDRAM is smaller than SRAM.

However, eDRAM suffers from the data retention issue due to the electrical charge leaking through the access transistors. As a result, the periodical refresh,

which includes a dummy read and write-back operation, is indispensable for data integrity consideration. In conventional 1T1C (1-transistor and 1-capacitor) e-DRAM, trench capacitor is introduced through additional mask layers to maintain high density and to increase the retention time. Consequently, the 1T1C eDRAM is generally incompatible with the standard digital process.

Gain-cell based eDRAMs rely on the MOSFET gate capacitors (∼fF) as the main storage node. Therefore, gain-cell based eDRAM suffers from shorter retention time than 1T1C eDRAM. Despite of this, the gain-cell eDRAM is fully compatible to the standard digital process, which is an advantage over the 1T1C eDRAM when targeting for low cost ULV system.

When considering the replacement of SRAM by eDRAM in ULV/ULP systems, it is very crucial to provide an efficient refresh scheme to allow easy integration. The previous eDRAM designs [48, 114–119] all need a dedicated period for data refresh, which causes reduced memory availability. This situation is not a significant issue for high-performance applications, but it is considerably an serious overhead in practical ULV/ULP designs. For example, a low-cost ULP MCU has to be re-designed to include a refresh-controller for data refresh and pipeline stall operation.

In this chapter, we propose a hidden-refresh scheme for 100% availability, gain-cell based eDRAM. This scheme ensures the gain-cell eDRAM with a SRAM-like interface without dedicated refresh cycles for data retention, therefore it relaxes the design efforts for the easiness of system integration. Several circuit techniques and design considerations are explicitly applied to enable the hidden-refresh scheme and to ensure robust read/write/refresh operation.

Also, we explore the voltage scaling effects of the eDRAM based on the above-mentioned scheme. The intuition is that voltage scaling will definitely reduce the

access/refresh power, but the retention time is also reduced at the same time, which may cause increased refresh rate and energy. As a result, the power/energy of voltage scaled eDRAM becomes an interesting trade-off, which is another focus of this chapter.

The rest of this chapter is organized as follows. Section 6.2 describes the proposed hidden-refresh scheme. Section 6.3 presents circuit design considerations for the self-refresh eDRAM design Section 6.4 evaluates the voltage scaling effects on the hidden-refresh eDRAM. Section 6.5 concludes this chapter.

6.2 Hidden-Refresh Scheme

Conventional eDRAM designs have to refresh periodically to retain the data storage and a dedicated period of time for refresh has to be considered. During the refresh period, the eDRAM design is not available for any access (read/write) operation, which indicates reduced memory availability. Note that the conventional eDRAM designs are targeting for high-performance operation, which are optimized with reduced cycle time for hundreds MHz and near-GHz operation. This fast clock results in a short period for refresh, which leads to inevitable system overhead for refresh control.

When eDRAM is considered for low-power systems with MHz range operation, the conventional refresh scheme is not applicable as the eDRAM availability is further deteriorated with reduced operating frequency while the data retention time is not changed. In addition, system level refresh control is generally expensive in most low-cost low-power systems.

Fortunately, with the awareness of the reduced clock frequency in low power

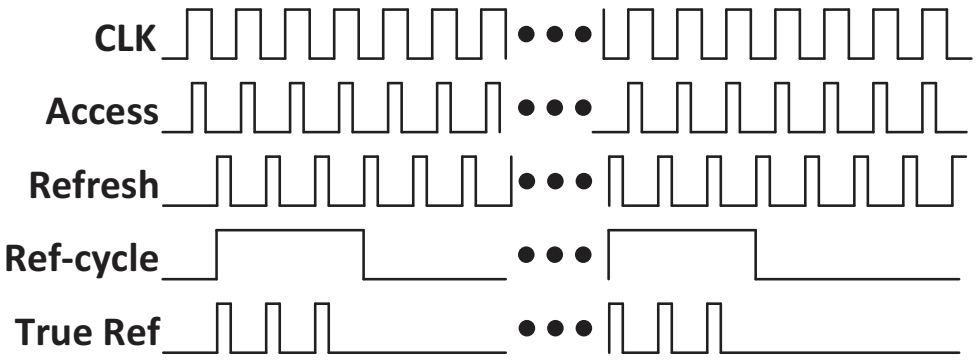


Fig. 6.1: Conceptual illustration of the hidden-refresh scheme for eDRAM.

systems, we can explore the long clock cycle for smart refresh control at memory circuit design level. In order to reduce the system design overhead, we propose a hidden-refresh scheme for eDRAM. This allows the eDRAM to have a SRAM-like interface with 100% memory availability while refresh operation is handled through the refresh operation between two consecutive access operation.

Fig. 6.1 shows the timing diagram of the proposed hidden-refresh scheme for eDRAM. Since a long clock cycle is generally available in low power systems (e.g., below tens of MHz), it is possible to utilize this long clock cycle by inserting refresh operation between two consecutive access operations.

One straightforward way of implementing this scheme is to use both the clock levels as a natural schedule. When CLK signal is high, the normal access operation (i.e., read or write operation) is scheduled to either write in (read out) the data to (from) the memory macro. When CLK signal is low, the refresh operation (i.e., dummy read followed by write-back) is scheduled.

Here, we need to clarify several power/energy metrics of the proposed hidden-refresh eDRAM operation. Although the refresh operation is "hidden", the refresh power still exists. As a result, refresh power will also contribute to both the dynamic

and static power of the hidden-refresh eDRAM. For the normal access mode, the dynamic power (read/write power) has to take the refresh power into consideration. When the eDRAM is in the data retention mode without being accessed, the total static power includes both the leakage power and the refresh power, as the refresh operation still takes place during the retention mode.

Indeed, it is possible to enable the refresh operation all the time. However, considering the power/energy overhead of the refresh operation, it is less energy efficient. Assuming a refresh duty cycle profile is known up-front for an eDRAM macro, as shown in Fig. 6.1, the true refresh operation can be performed only according to this duty cycle to maintain the data integrity of the whole memory and to minimize the refresh power/energy overhead.

6.3 Circuit Design for Self-Refresh eDRAM

In this section, we will cover the circuit design aspects for the self-refresh eDRAM design. Also, this design will be used for simulation to further investigate the voltage scaling effects on the hidden-refresh eDRAM.

The top level eDRAM block diagram of the hidden-refresh eDRAM is shown in Fig. 6.2. The external interfaces of the eDRAM is identical to a synchronous SRAM macro.

The major architectural modification to enable the hidden-refresh scheme can be realized with additional building blocks, which is highlighted using dashed lines. As shown in Fig. 6.2, except for the main building blocks similar to normal SRAM, an additional address multiplexor and a refresh counter is added. The address multiplexor is used for access and refresh address selection while the refresh counter

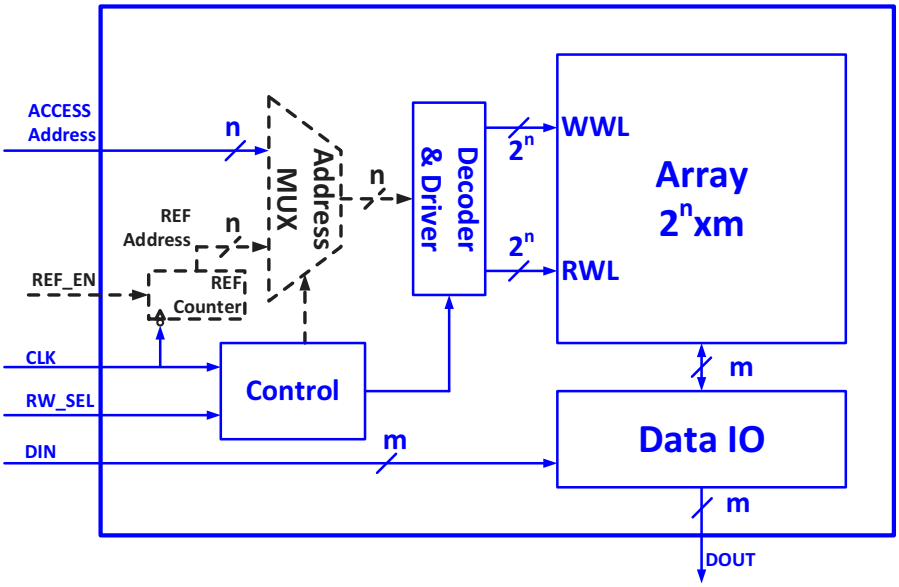


Fig. 6.2: Top-level block diagram of the hidden-refresh eDRAM.

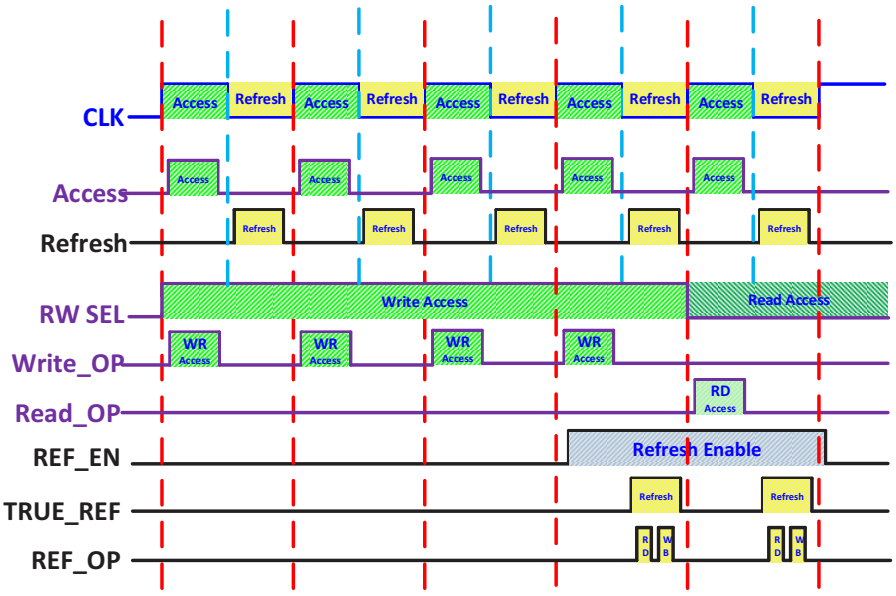


Fig. 6.3: Detailed timing diagram of the hidden-refresh eDRAM.

is used to generate the refresh address when refresh enable signal is applied. A more detailed timing diagram based on Fig. 6.1 is shown in Fig. 6.3. As shown, the access operation is further classified into write (Write_OP) and read (Read_OP) access, which will be used as the control signal for wordline and bitline control. In addition, only when refresh operation is enabled (REF_EN = 1), the refresh operation is regarded to be valid and the corresponding dummy read and write-back (REF_OP) operation is performed when CLK is low.

Note that a clock signal is required for refresh operation, which is used for the counter to function periodically. The pulse generation for the both Access and Refresh signal in Fig. 6.3 can be achieved through the classical Address Transition Detection (ATD) circuit [28]. The access address is applied from external signals, while the refresh address is generated from the refresh counter.

6.3.1 Bitcell Choice and Operation Principle

Previous gain-cell eDRAM bitcells consists of either two (2T) or three transistors (3T) with an optional MOSFET capacitors or diodes for different design purposes. In this chapter, we stick to the basic 2T bitcell implementation [116], which includes a PMOS access transistor and a NMOS storage transistor, as shown in Fig. 6.4.

For the considered commercial 65nm technology, the write transistor (M_W) is an HVT (high threshold voltage) PMOS device with lowest available leakage in the process to increase the retention time, while an RVT (regular threshold voltage) NMOS device is chosen for a moderate read sensing speed.

The operating principles and timing diagram of the 2T eDRAM are shown Fig. 6.4 and Fig. 6.5, as described below:

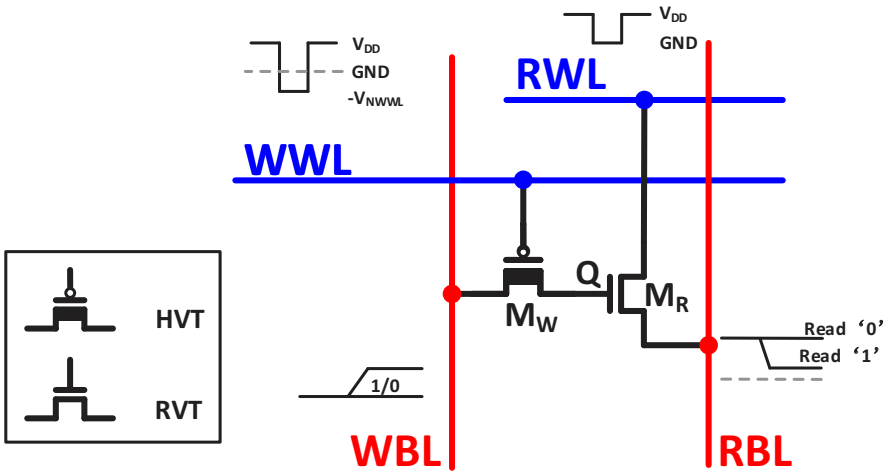


Fig. 6.4: 2T gain-cell eDRAM with simplified operation waveform.

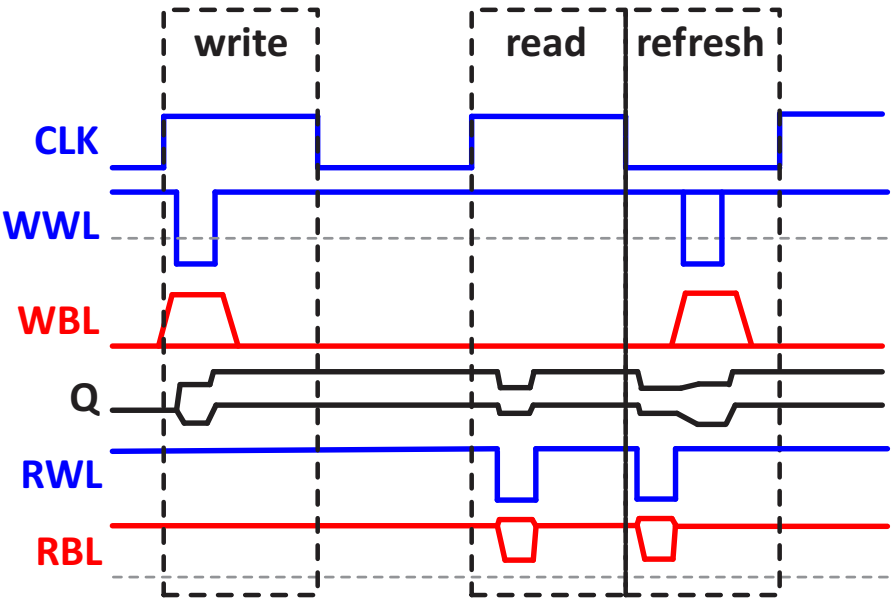


Fig. 6.5: Timing diagram of the eDRAM write, read and refresh operation.

For write operation, due to the voltage droop caused by the write transistor M_W, the write wordline (WWL) of the selected row is applied with a negative pulse to enhance the drivability of the write transistor during write operation. Also, a proper timing control is needed to ensure that the input data fed to the write bitline (WBL) should be kept stable until the WWL signal goes back to V_{DD}, which is due to the level-sensitive write mechanism. In the meantime, the internal storage node Q will also be coupled to a higher voltage level due to WWL switching. Furthermore, it is worth noting that the WBL voltage level (during un-accessed time) has significant effects on the retention time of the storage node in the gain-cell based eDRAM. For the chosen 2T eDRAM design with PMOS write device, WBL should be always discharged to VSS when the WBL is un-accessed.

For read operation, the read bitline (RBL) is needed to be precharged first. During sensing period, the precharge is disable and the selected read wordline (RWL) is pulled from V_{DD} to GND to sense the storage node value. When the storage node value is "0", the read transistor M_R is off and the RBL is kept to be V_{DD}. When the storage node value is "1", the read transistor M_R is on and ideally the RBL is discharged from V_{DD} to a lower voltage level.

Refresh operation is a combination of the read and write-back operation, as shown in Fig. 6.5, with half clock-cycle when CLK is low. As a result, the proposed hidden-refresh scheme requires a clock cycle at least two times longer than a refresh operation and equivalently, the hidden-refresh scheme halves the maximum operating cycle when compared to the previous eDRAM design. However, this is still tolerable for ULP system required maximum frequency up to only tens of MHz. And the benefit of the hidden-refresh scheme is that the eDRAM has a SRAM-like interface for the easiness of system integration.

6.3.2 Write/Read Bitline Circuit Design

Fig. 6.6 shows the write (WBL) and read (RBL) bitline circuit design of the proposed hidden-refresh scheme. As mentioned earlier, the write bitline should be precharged to GND (through WBL_PC) for maximizing the retention time when it is not accessed. Considering both the write/write-back requirement of the same WBL port, two enable signals (WR and WB) are used to distinguish the write operation and write-back operation for correct data/refresh data input. The WBL_PC should be disabled during both write and write-back operation, which can be realized through a simple OR operation of WR and WB.

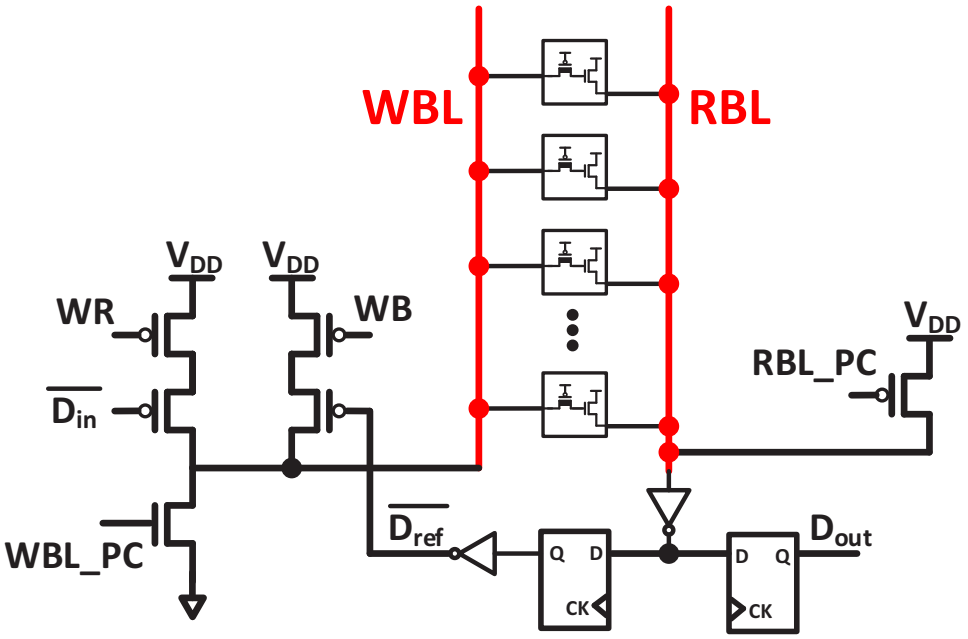


Fig. 6.6: Write/read bitline design of the hidden-refresh scheme.

For read bitline circuit, the RBL is first precharged to V_{DD}, then the RBL is inverted and sampled by two flip-flops for read out and refresh operation, respectively. This read out scheme is significantly different from the SRAM design as

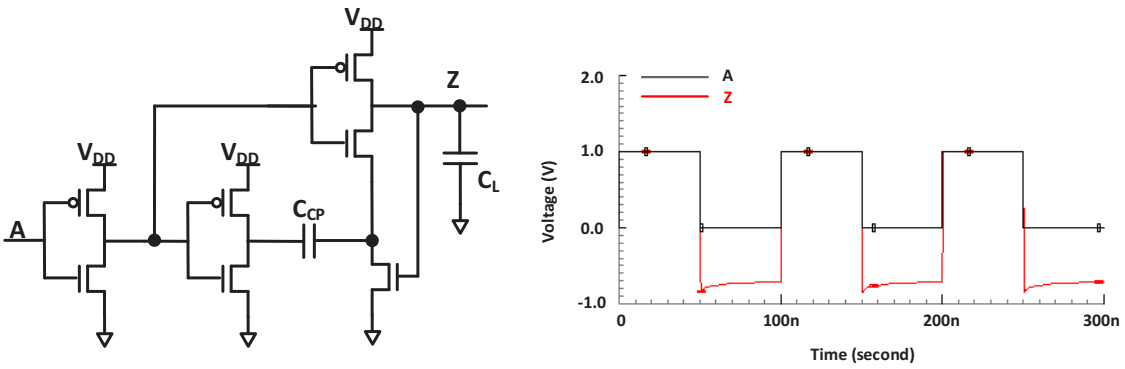


Fig. 6.7: Bootstrapped WWL driver and simulated waveform.

the SRAM address/control/data signals will be consistent within one clock cycle during operation. On the contrary, the address/control/data of the hidden-refresh eDRAM are altering every half clock cycle during the refresh period. As a result, two flip-flops (for both read and refresh operation) in one column are necessary for the correct hidden-refresh operation. The sample clock signals for both flip-flops can be generated through proper delay control to ensure the correct data to be stored during read out operation.

6.3.3 Wordline Driver Circuit Design

Wordline drivers are very important for the correct eDRAM operation. Several design issues in wordline drivers are encountered in previous eDRAM designs. As mentioned in Section 6.3.1, the write wordline requires a negative voltage to enhance the drivability, and additional supply is inevitable for providing this negative voltage.

In order to minimize the additional power supply overhead, we propose a charge-pump WWL driver, as shown in Fig. 6.7. A negative V_{DD} charge pump driver based on the bootstrapped inverter [62] is adopted. The conversion efficiency,

which is also equivalent to the bootstrapped negative voltage level, can be satisfied through a reasonable ratio between the charge pump capacitor (C_{CP}) and the load capacitance of all write access transistors in one row (C_L). The charge pump capacitor can be implemented with MOSCAP to minimize the design overhead.

Another practical design concern is the read disturbance issue in the 2T e-DRAM design with multiple cells sharing one read bitline, as shown in the left figure in Fig. 6.8. During read operation, the selected wordline driver is pulled to GND and the RBL voltage is determined by the data on the storage node. When data "1" is stored, the selected wordline driver should be able to pull down the RBL through the read device. However, the worst case for the read "1" disturbance arises when all cells store data "1". The simplified equivalent read out circuit is shown in the right figure in Fig. 6.8. For multiple cells sharing one RBL, the un-selected read wordline (V_{DD}) delivers contention current to the RBL through the storage transistors with data "1". As a result, this significant read disturbance might cause read "1" failure. And existing solutions suggests to use additional power rails [117] to lower the contention current (set UNSEL_RWL to a voltage smaller than V_{DD}) or a compromised design choice with reduced cells per RBL [119].

Fig. 6.9 shows the proposed tri-state RWL buffer to mitigate the read disturbance issue, with the purpose of reducing the contention current from the un-selected row. As shown in Fig. 6.9, the tri-state enable signal allows the selected RWL to be functional as usual, while the un-selected RWLs are floating as the un-selected tri-state buffers are disabled. Consequently, the RBL can be easily pull-down to a low voltage level.

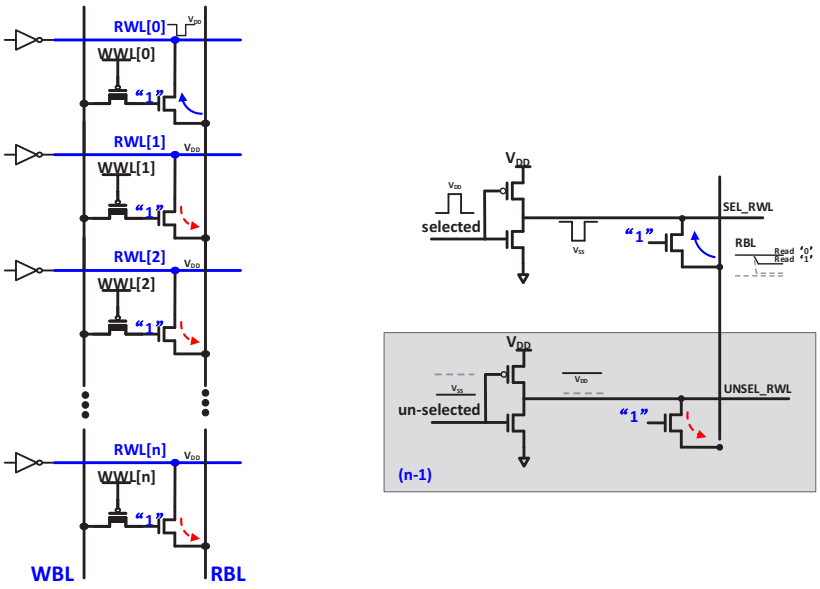


Fig. 6.8: Illustration of the worst case read disturbance issue of the 2T gain-cell eDRAM.

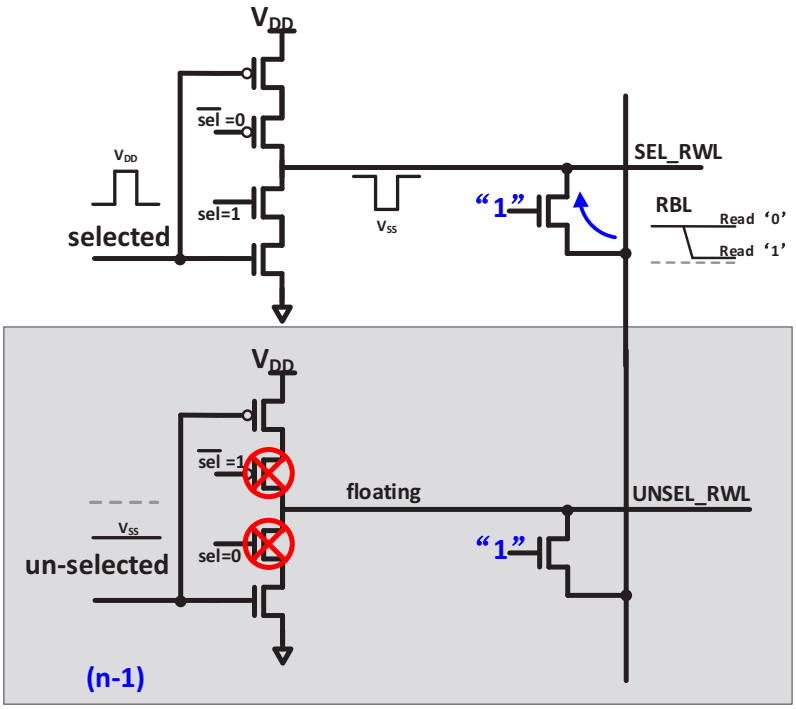


Fig. 6.9: Proposed tri-state RWL driver.

6.4 Power Metrics of the Hidden-Refresh eDRAM under Voltage Scaling

In this section, the energy metrics of the hidden refresh eDRAM is investigated under supply voltage scaling from nominal voltage of 1.2V down to 0.6V. Based on the previous mentioned techniques in this chapter, we design a 1-Kbit (64-row by 16-bit) 2T gain-cell based hidden-refresh eDRAM in a commercial 65nm technology, as shown in Fig. 6.10.

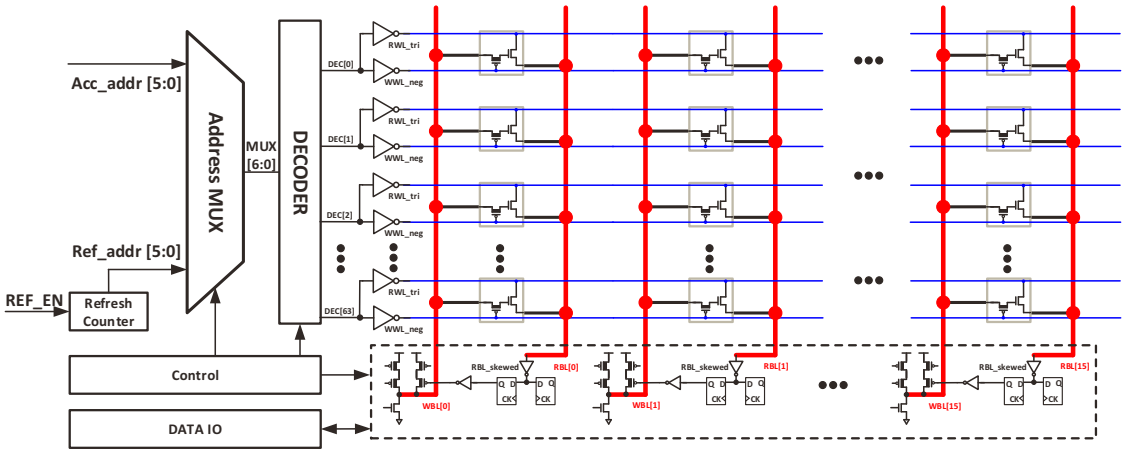


Fig. 6.10: Schematic of the 1K-bit hidden-refresh eDRAM.

Simulation results show that the worst case read and write operation at 0.6V can be finished within 10ns. This implies that a target operating frequency of 10MHz can be achieved for the 2T eDRAM, where a clock cycle of 100ns is available. The simulated power consumption under different supply voltages at room temperature 25°C are plotted in Fig. 6.11. As can be observed, the refresh operation shows largest power consumption as both read and write-back operation are performed. Read power is significantly larger than write power, which is due

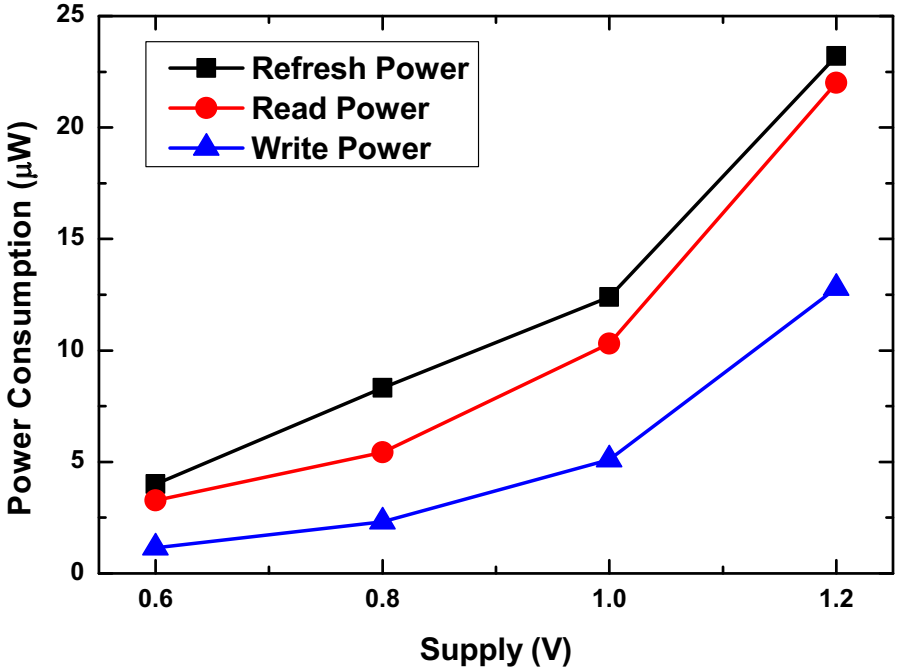


Fig. 6.11: Power consumption of the hidden-refresh eDRAM.

to the switching power of the flip-flops in the RBL sensing circuits. The power reduction trends are obvious. When the supply voltage scales from 1.2V to 0.6V, approximate 5× power reduction can be achieved for all three operation types.

Fig. 6.12 shows the retention time of the 2T eDRAM and the static power consumption, which confirms the retention time reduction due to supply voltage scaling [117]. When supply voltage scales from 1.2V to 0.6V, the retention time reduces from 120μs to 40μs, which is a 3× reduction. As a result, the power benefit from voltage scaling (5×) dominates the retention time loss (3×). For the 1-Kbit memory with a 10MHz clock, the 64-address memory can be refreshed within 6.4μs. Therefore, the eDRAM is fully functional with a duty cycled refresh operation. This results in a static power consumption of 640nW at 0.6V, as shown in Fig. 6.12.

Since the hidden-refresh eDRAM is designed with an SRAM-like interface,

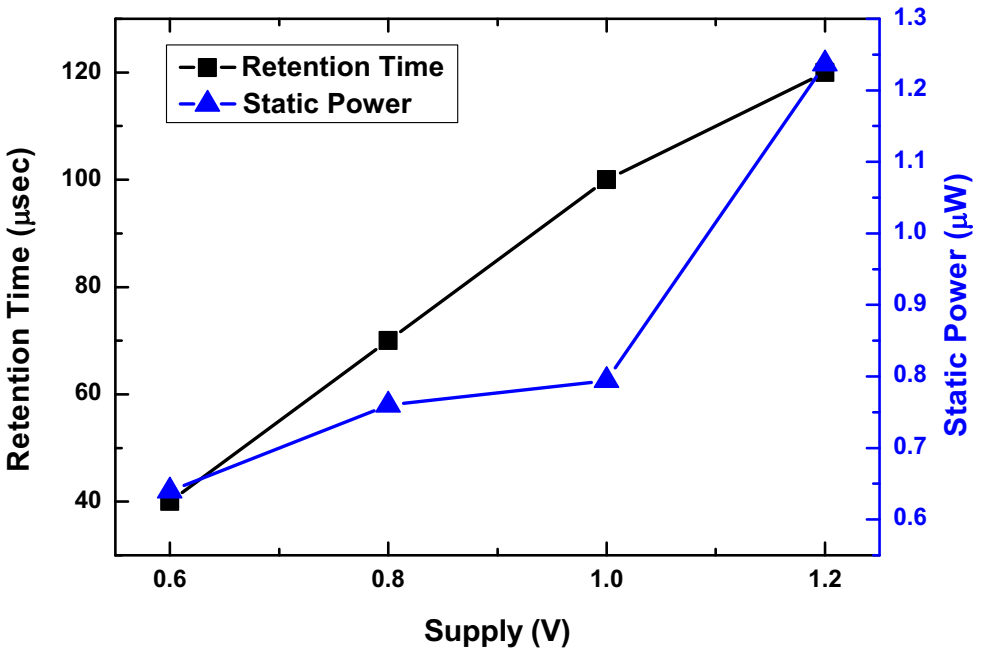


Fig. 6.12: Retention time and static power of the hidden-refresh eDRAM.

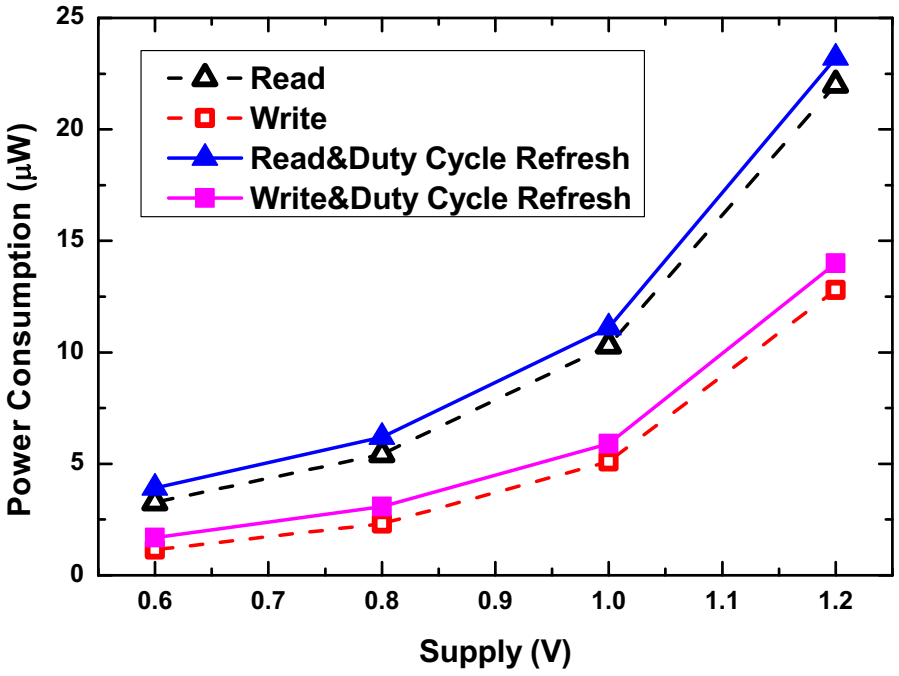


Fig. 6.13: Read/write power with duty cycled refresh power of the hidden-refresh eDRAM.

when considering the actual read/write power, the duty cycled refresh power should also be included, as depicted in Fig. 6.13. At 0.6V, the hidden-refresh eDRAM shows 3.9μW read power (including the duty cycled refresh power) with 10MHz operating frequency, and the write power (including the duty cycled refresh power) is 1.7μW.

Based on the above results, voltage scaling is still beneficial for power reduction in the hidden-refresh eDRAM down to 0.6V. Although the refresh duty cycle is reduced by 3× at lower supply voltage, the power reduction of 5× from voltage scaling is dominating. However, we found that further reducing the supply below 0.5V would cause serious concerns to the eDRAM functionality. This is because the refresh duty cycle below 0.6V is approaching to or even smaller than the required refresh period (6.4μs). Indeed, it is still possible to revise the hidden-refresh scheme to enable more than one refresh address within one clock cycle. However, this modification is beyond the scope of this chapter.

Table 6.1 shows the comparison of the proposed hidden-refresh eDRAM to previous SRAM/eDRAM designs. As can be observed, the 2T eDRAM shows higher density than the SRAM design. When normalized the area of SRAM designs [39, 137] to 65nm, the 2T eDRAM still shows around 2.5× reduction in bitcell area compared to SRAM designs. When operating at 0.5V, the access energy of 16-bit operation is 2.33pJ [39]. While for eDRAM with 0.6V supply, the 16-bit operation access energy (read & duty cycled refresh) of hidden-refresh eDRAM is 0.4pJ, which is 5.8× energy reduction. These promising results show that the hidden refresh eDRAM might be a viable option for replace SRAM. However, the static power of the 2T eDRAM is 3.2× of that of SRAM, which is mainly caused by the refresh power overhead. Further enhancing the retention time will

Table 6.1: Comparison among SRAM, eDRAM and Hidden-Refresh eDRAM

	[39]	[137]	[116]	This work
Type	6T SRAM	8T SRAM	2T-eDRAM	2T-eDRAM
Access	No	No	External	Hidden
Technology	0.13µm	40nm	65nm	65nm
Power Supply	Single	Single	Multiple	Single
Capacity	2-kb	512-kb	192-kb	1-kb
Bitcell Area	4.788µm^2	0.706µm^2	0.48µm^2	0.48µm^2
Frequency	5.62MHz, 0.5V	6.25MHz, 0.5V	500MHz, 1.2V	10MHz, 0.6V
Access Energy (16b)	2.33pJ	12.9pJ	NA	0.4pJ (include ref.)
Retention Time	NA	NA	20µs, 0.8V	40µs, 0.6V
Static Power	200nW, 0.5V	72.8 µW, 0.5V	109 µW/Mb, 85°C	640nW, 600mV

be helpful in reducing the static power. In addition, 65nm process has higher leakage when compared to the 0.13μm process, therefore eDRAM design in 0.13μm will have longer retention time for static power reduction. Compared to eDRAM without hidden-refresh scheme [116], this work achieves lower voltage operation. In addition, the proposed design uses only one power supply, which is an advantage compared to the conventional eDRAM designs.

6.5 Conclusions

In this chapter, we explore the energy efficiency in eDRAM design as a memory alternative to SRAM. The eDRAM provides higher density than conventional SRAM, however, the refresh operation complicates the system-level design due to the overhead of the control blocks. In order to remove the system level integration effort, we propose a hidden-refresh scheme to enable the eDRAM to have a SRAM-like interface. In addition, existing eDRAM implementations requires additional supply voltages for correct functionality. Several circuit techniques, including

bootstrapped WWL driver and tri-state RWL driver, are proposed to minimize this overhead for achieving true single-V_{DD} operation. Finally, a 1-Kbit hidden-refresh eDRAM is designed and simulated in a commercial 65nm technology. Through comparison to a SRAM counterpart, eDRAM shows promising benefits in term of memory density and access energy. This result is especially compelling for those memory-intensive ULP systems with stringent area and energy limits.

Chapter 7

A 0.4V 280nW Nearly All-Digital Frequency Reference-less Hybrid Domain Temperature Sensor

---

This chapter covers a case study of applying ULV digital-assist circuits for emerging sensor designs. We present the design of a sub-μW nearly all-digital hybrid domain temperature sensor for wireless temperature sensing applications. A subthreshold-biased ratioed-current/delay based PTAT sensor core is proposed. In addition, we propose a hybrid domain digital processing technique based on the proposed sensor core to relax the requirement of an external accurate frequency reference, which is generally power hungry for energy-constrained systems. The proposed design was fabricated in a 65nm CMOS process and measured with 0.4V power supply. The eight measured chips show -1.6/1°C error across the 0$\sim$100°C range after 2-point calibration, and the power consumption is on average 280nW.

Table 7.1: Categories of the CMOS temperature sensor

References	Type	Sensor Core
[121, 122]	Voltage Domain	BJT-based
[123–127]	Time Domain	Delay line based
[128–131]	Time Domain	Current to delay converter based
[132–135]	Frequency Domain	Frequency to digital converter based

7.1 Introduction

Smart temperature sensors based on CMOS technologies are popular due to the low cost and easiness for the on-chip integration purposes. As a result, CMOS temperature sensors are widely used for on-chip dynamic thermal management in high-performance microprocessors, environment sensing/food monitoring in wireless sensor nodes or RFID tags and temperature compensation in MEMS resonators. Due to the different performance requirements for various applications, temperature sensor designs are vastly different and challenging to achieve high accuracy/resolution, high data rate and low power consumption as well.

Basically, the existing temperature sensors can be categorized into three major types based on their distinguished operating principles, as shown in Table 7.1. The first type is the voltage-domain BJT-based temperature sensors [121, 122]. The temperature-dependent voltage is then converted to digital codes through the integrated high-resolution ADCs, which are preferred in applications with high accuracy/resolution requirements, such as temperature compensation in MEMS resonators. Second, the time-domain temperature sensors explores the temperature-dependent delay with integrated TDCs are also proposed [123–131]. The temperature-dependent delay is then converted to digital codes through time-to-digital converters. In order to generate a delay large enough to achieve a decent resolution, either

hundreds of cascaded inverter chains [123–127] or the current-to-delay converter based sensor core (diodes, MOSFETs, etc.) are used [128–131]. The third type is to utilize the temperature dependent frequency of the ring oscillator as the sensing element [132–135].

With the growing interest of integrating smart temperature sensors into the ultra-low-power wireless sensor platforms and RFID tags, where compact nW temperature sensors with moderate sensing error are attractive due to the limited system power budgets (i.e., a few μW) [136]. However, due to the use of ADCs, long delay lines and ring oscillators with super-threshold supply, the power consumption of the above-mentioned temperature sensors are normally well exceeding the μW power budgets, making them unsuitable for wireless sensing applications.

Efforts have been made for different temperature sensor types to reduce the total power consumption. And several prototypes are demonstrated with sub-μW power consumption [130, 131, 134, 135]. However, current-to-delay converter [130, 131] based temperature sensors are highly dependent upon the availability of the frequency reference for TDC operation. Therefore, additional clock reference is needed but it is generally power hungry for ultra-low-power platforms, as shown in Fig 7.1. In order to mitigate this, temperature sensors with integrated on-chip frequency references into the FDC based sensors [134, 135] are demonstrated. However, these designs rely heavily on the analog blocks and techniques through iterative design efforts, which is a disadvantage for technology migration.

In order to resolve the above-mentioned challenges, this chapter presents a 0.4V 280nW nearly all-digital hybrid domain temperature sensor for wireless sensing applications. A ratioed-current/delay PTAT sensor core realized with two subthreshold biased MOSFETs is proposed. Based on the proposed sensor core,

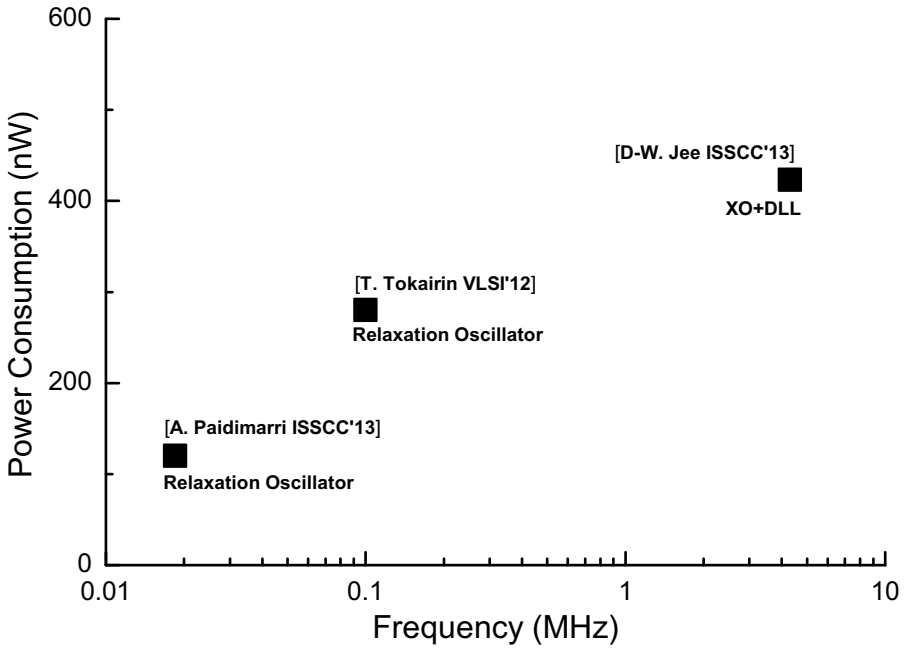


Fig. 7.1: Power consumption versus frequency of the state-of-the-art ultra-low power frequency reference for illustration of power overhead due to frequency reference.

we propose a hybrid domain processing technique. As a result, the requirement for accurate frequency reference is eliminated and the nearly all-digital implementation makes this design more technology scaling friendly.

The rest of this chapter is organized as follows. Section 7.2 describes the proposed ratioed-current/delay PTAT sensor core. Section 7.3 presents the proposed hybrid domain temperature sensing scheme and its benefits. Section 7.4 shows the measurement results of the hybrid domain temperature sensor. Section 7.5 concludes this chapter.

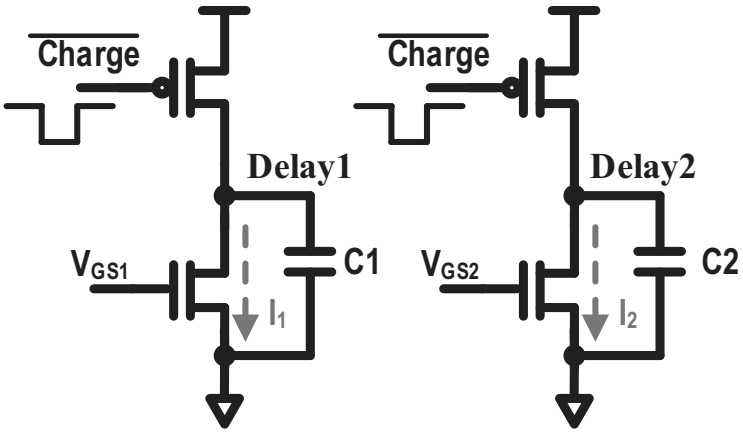


Fig. 7.2: Schematic of the proposed ratioed-current/delay PTAT sensor core.

7.2 Ratioed-Current/Delay PTAT Sensor Core

In this work, we explore the ratioed-current/delay to be the PTAT sensor core. Fig. 7.2 portraits the schematic of the ratioed-current/delay PTAT sensor core. Two identical NMOS devices are biased in the subthreshold region with the gate overdrive voltage of V_{GS1} and V_{GS2}, respectively. The corresponding subthreshold drain current (I_1 and I_2) of the two NMOS devices are expressed in Eq. 7.1 and Eq. 7.2,

$$I_1 = \mu C_{ox} \frac{W}{L} {V_T}^2 exp\left(\frac{V_{GS1} - V_{th}}{nV_T}\right)\left[1 - exp\left(-\frac{V_{DS1}}{V_T}\right)\right] \tag{7.1}$$

$$I_2 = \mu C_{ox} \frac{W}{L} {V_T}^2 exp\left(\frac{V_{GS2} - V_{th}}{nV_T}\right)\left[1 - exp\left(-\frac{V_{DS2}}{V_T}\right)\right] \tag{7.2}$$

where μ is the NMOS mobility, C_{ox} is the oxide capacitance, V_T is the thermal voltage, V_{th} is the NMOS threshold voltage, n is the subthreshold slope, V_{GS} and V_{DS} are the gate over-drive voltage and drain-source voltage, respectively. Note that the two NMOS devices are identically and can be sized large enough to

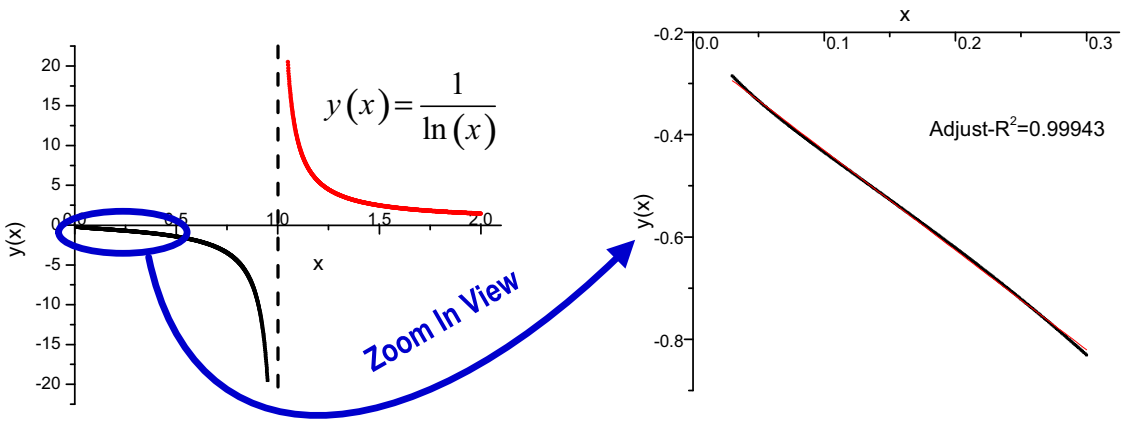


Fig. 7.3: Mathematical background of the operation principles of the proposed current-ratioed PTAT.

minimize the V_{th} mismatch, thus the current ratio can be expressed in Eq. 7.3.

$$\frac{I_2}{I_1} = exp\left(\frac{V_{GS2} - V_{GS1}}{nV_T}\right) \tag{7.3}$$

As can be seen in Eq. 7.3, the current ratio is determined by the difference of the overdrive voltage ΔV_{GS} (= V_{GS2}-V_{GS1} < 0, assuming $V_{GS1} > V_{GS2}$), the subthreshold swing n and the thermal voltage V_T. Equivalently, the absolute temperature are proportional to ΔV_{GS} and $1/\ln(I_2/I_1)$, as shown in Eq .7.4.

$$T = \frac{V_{GS2} - V_{GS1}}{nk/q} \frac{1}{ln\,(I_2/I_1)} \tag{7.4}$$

Then we make observations of the basic function: y(x) = 1/ln(x), as shown in Fig. 7.3. Notice that there are two lobes of the function plot, and we are interested in the circled region of the left lower lobe, and a zoom view of the region is also plotted in Fig. 7.3 on the right. As can be observed, a portion of this curve ($x \in (0.05, 0.3)$) shows decent linearity. The linear fitting adjusted-R^2 coefficient is

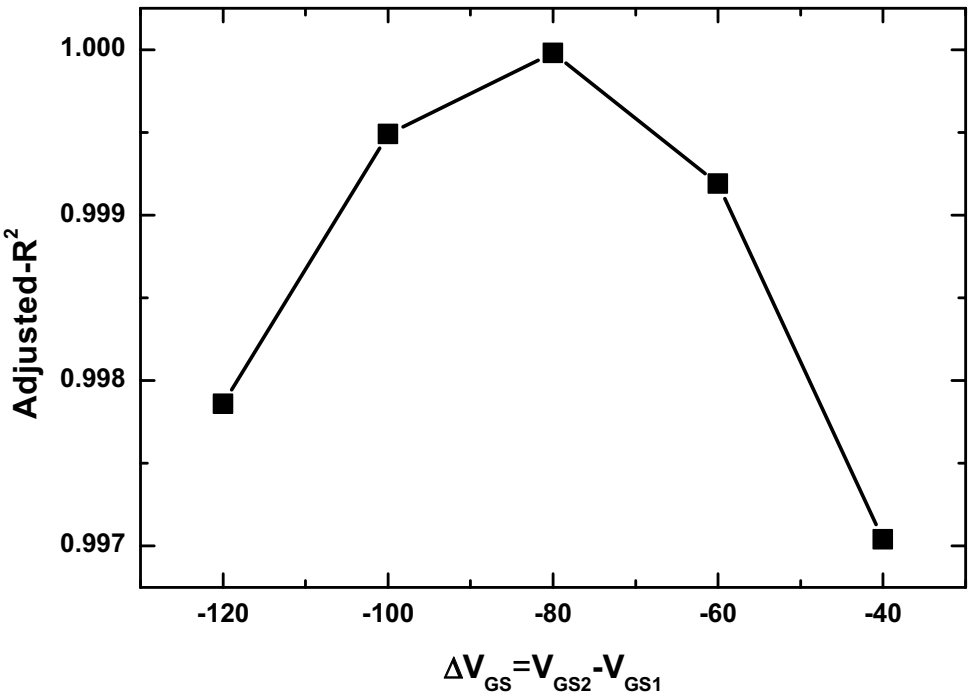


Fig. 7.4: Optimal ΔV_{GS} vs. Adjusted-R^2 coefficient.

0.99943 for this specific region. It turns out that if a proper ΔV_{GS} is applied to the temperature sensor core to meet the desired current ratio, the absolute temperature is proportional to the current ratio, as shown in Eq. 7.5.

$$T \propto \frac{I_2}{I_1} \tag{7.5}$$

A theoretical calculation of the optimal ΔV_{GS} versus the Adjusted-R^2 coefficient across 0-100°C temperature range is performed, as shown in Fig. 7.4. For ΔV_{GS} within -100 to -60 mV range, the Adjusted-R^2 coefficient is over 0.999 and the coefficient peaks at ΔV_{GS} of -80mV, which is the chosen value in this work. As a result, the current radio can be functional like a PTAT.

With the linear dependence between the absolute temperature and the current

ratio, we can further explore the delay ratio through simple current-to-delay converter. The output node is charged to VDD through PMOS and then the PMOS is turned off and the output node slowly discharges through the subthreshold biased NMOS to the switching threshold. After several simple logic gates, we can get two Temperature-Sensitive-Delays (TSDs) pulses, where the delay ratio is proportional to the absolute temperature, as described in the equivalent equations in Eq. 7.6.

$$t_{d1} \propto \frac{C_1}{I_1}, t_{d2} \propto \frac{C_2}{I_2} \Leftrightarrow \frac{t_{d1}}{t_{d2}} \propto \frac{I_2}{I_1} \tag{7.6}$$

7.3 Hybrid Domain Temperature Sensing Scheme

The conventional time domain temperature sensors use only one temperature-sensitive delay [123], or use the delay difference of two temperature-sensitive delays [130, 131]. As a result, the absolute delay information has to be digitized through the use of time-to-digital converters (TDCs), where an accurate external frequency reference is needed.

As a result, with the proposed ratioed-current/delay sensor core, the requirement of the external frequency reference is relaxed through the hybrid domain processing technique, as shown in Fig. 7.5. Since the ratio of the two TSDs generated from the sensor core is interested, we can use a free-running temperature-sensitive ring-oscillator (TSRO) as the main clock of the TDCs to sample both TSDs simultaneously. And the ratio can be then calculated through digital arithmetic circuits. Since we use a frequency-domain information (from TSRO) to quantize the time-domain information (from TSDs), then the TDC values and ratio calculation are therefore represented by the hybrid domain (time-frequency domain), as opposed

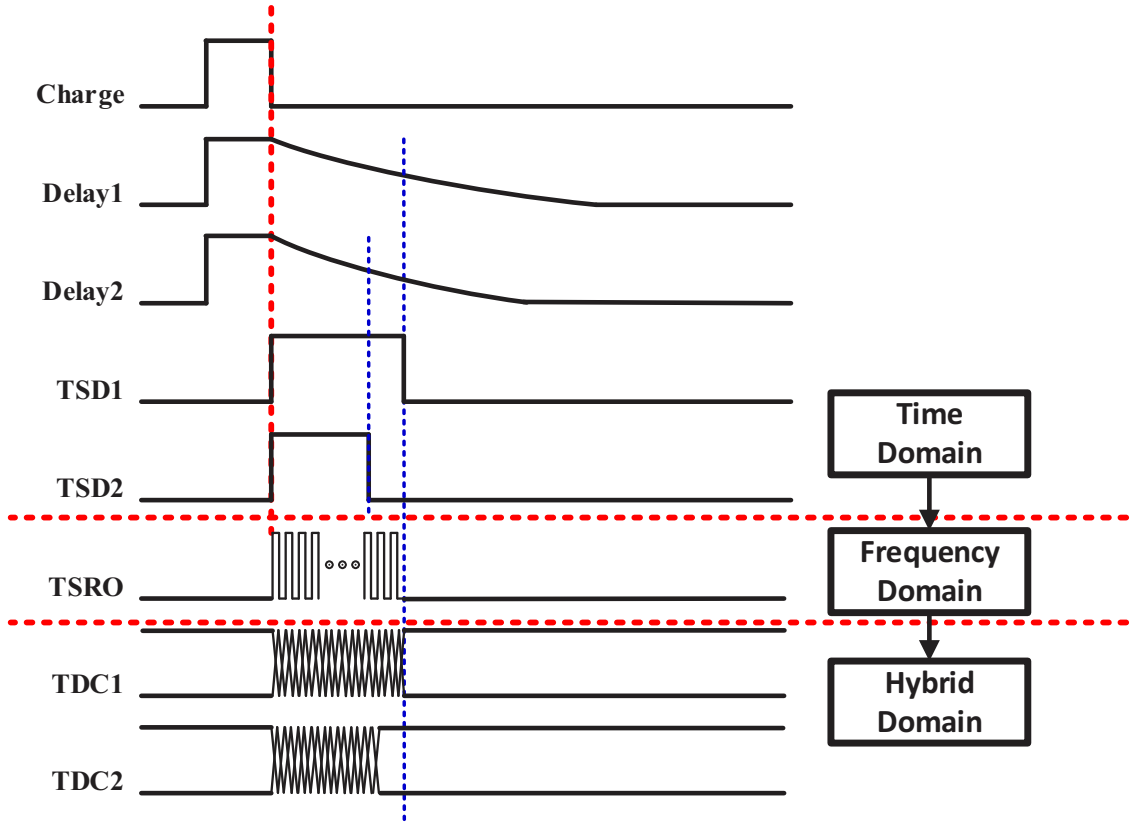


Fig. 7.5: Timing diagram of the ratioed-current/delay temperature sensor.

to the previous time-domain value sampled by a frequency reference. It is worth noting that the hybrid-domain TDC values should be large enough to minimize the quantization error for delay ratio calculation.

7.4 Circuit Implementation Details

The circuit block diagram of the hybrid domain temperature sensor is shown in Fig. 7.6. A 5-tap PMOS resistor ladder is implemented to provide the temperature-insensitive overdrive voltage. With 0.4V supply, the difference of the overdrive voltage is 80mV. The current-to-delay converters converts the NMOS discharge

current into delay information, and the capacitor is used to increase the delay pulse width. The capacitor ratio of 27 is adopted to balance the two delays across a wide temperature range as they will be sampled by the same Temperature Sensitive Ring Oscillator (TSRO) to satisfy the trade-off between the TSRO frequency and overflow of the TDCs. As a result, 31-stage TSRO is implemented to digitize the Temperature Sensitive Delays (TSDs) with binary counter based 14-bit TDCs. The 31-stage TSRO is selected to ensure that there is enough time slack for the 14-bit TDCs with long carry-chain as its critical path.

We can still refer to Fig. 7.5 for the detail operation principle of the hybrid domain temperature sensor. PMOS devices are first turned on to charge the capacitors. From the onset of the switching-off of the PMOS devices, the current-to-delay converters start to discharge the capacitors and convert the currents into the TSDs. The two TSDs are then digitized by the TSRO and the delay ratio is calculated, both in the hybrid domain. The two TSDs are also used as the clock-gating signals for the two TDCs (GCLK1 and GCLK2) and the enable signal of the TSRO to reduce dynamic power. The hybrid domain ratio calculation is performed using a 15-bit single-precision floating point format.

The discharging NMOS devices are sized with longer gate length to minimize the device mismatch. Also, this will lead to smaller capacitors. In the chosen technology, NMOS device with 2μm/1μm geometry is used, and the total capacitance is 1.1pF(unit capacitor is around 39fF). The PMOS device is sized to have proper I_{on} and I_{off} to ensure the correct charging and discharging of the capacitors in both current-to-delay converters.

Fig. 7.7 shows the simulation results of the delay ratio based on the hybrid domain sensor described above. The process variation effects are also considered to

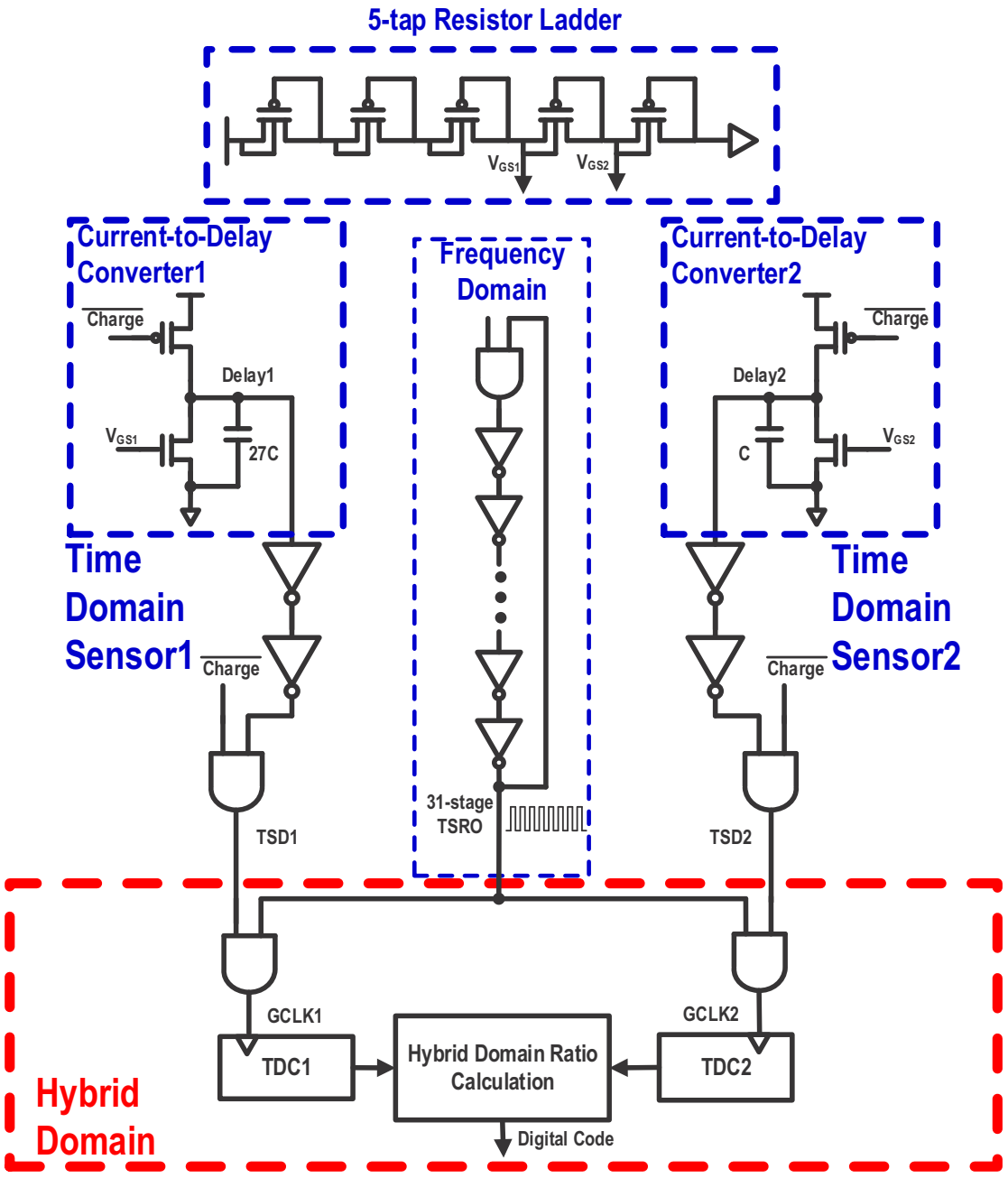


Fig. 7.6: Schematic of the hybrid-domain temperature sensor.

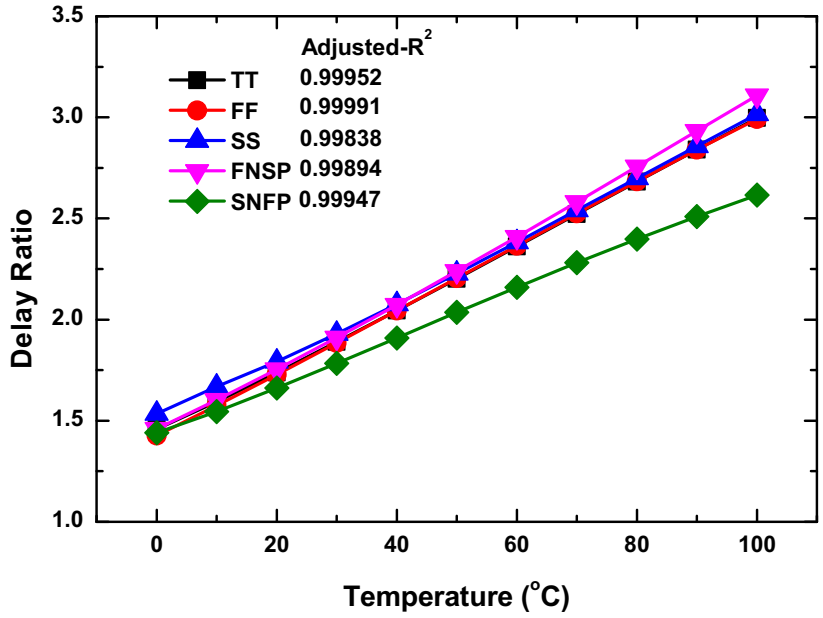


Fig. 7.7: Simulated delay ratio the ratioed-current/delay temperature sensor.

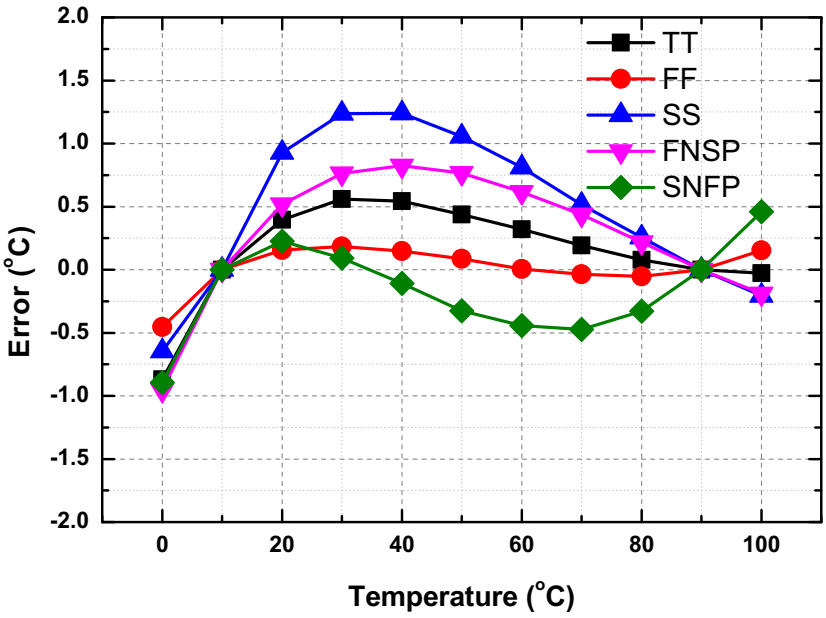


Fig. 7.8: Simulated temperature error of the ratioed-current/delay temperature sensor.

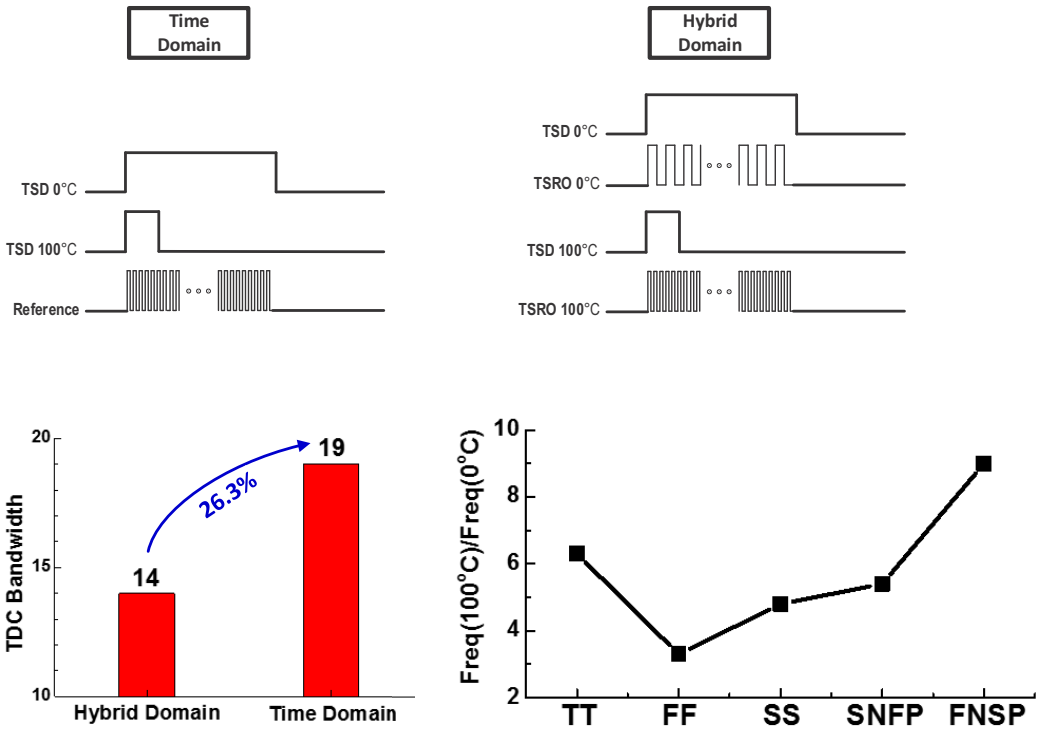


Fig. 7.9: Illustration of time domain (top left), hybrid domain (top right) processing and the hybrid domain processing benefits on TDC bandwidth (bottom left) and dynamic frequency scaling (bottom right, simulated).

verify the sensor performance. As can be observed, the proposed delay ratio shows good PTAT behaviors under all process corners with worst-case at SS corner with Adjusted-R^2 coefficient of 0.99838. However, the slope can be slightly varied, which is due to the relatively imbalanced driving strength between the charging (PMOS) and discharging devices (NMOS) with different process corners. Fig. 7.8 shows the simulation error after two-point calibration at 10 and 90°C. For the interested temperature range, the maximum error across process corners is -1$\sim$1.3^{o}C.

In addition to removing the frequency reference, the hybrid domain processing has extra benefits. The conceptual illustration is shown in Fig. 7.9. For the time domain case, the worst-case frequency is needed to sample the delay at

Sign	5-Bit Based Exponent	9-Bit Normalized Fraction
[14]	[13:9]	[8:0]

Fig. 7.10: Hybrid domain digital processing data format.

high temperature/FF corner and this fast clock will significantly increase the TDC bandwidth and power consumption at low temperature/SS corners. This will incurs large TDC bandwidth and higher power consumption due to the high frequency of the reference. On the contrary, for the hybrid domain case, the TSRO provides a process-temperature-tracking frequency. As long as the TDC values are large enough to minimize the quantization error, the TSRO can dynamically reduce the TDC bandwidth and dynamic power. In this way, the hybrid domain technique provides 26.3% TDC bandwidth reduction, as indicated in Fig. 7.9. In addition, 3.3× worst case dynamic power reduction (FF corner) can be observed for hybrid domain technique at 0^oC due to the dynamic frequency scaling of the TSRO.

The hybrid domain digital processing uses a 15-bit single-precision floating point format revised from the IEEE 754 standard, as shown in Fig. 7.10.

7.5 Measurement Results

The temperature sensor is realized in UMC 65nm 1P6M low leakage process. Metal-oxide-metal capacitors (MOMCAP) is adopted in this prototype. Fig. 7.11 shows the die micrograph with annotated floorplan. The sensor takes $0.022mm^2$ silicon area in total and sensor core plus two TDCs takes only $0.0054mm^2$.

The sensor core, TDCs and the hybrid domain processing circuit are operated with an off-chip regulated 0.4V supply. The ESPEC SU-240 temperature chamber

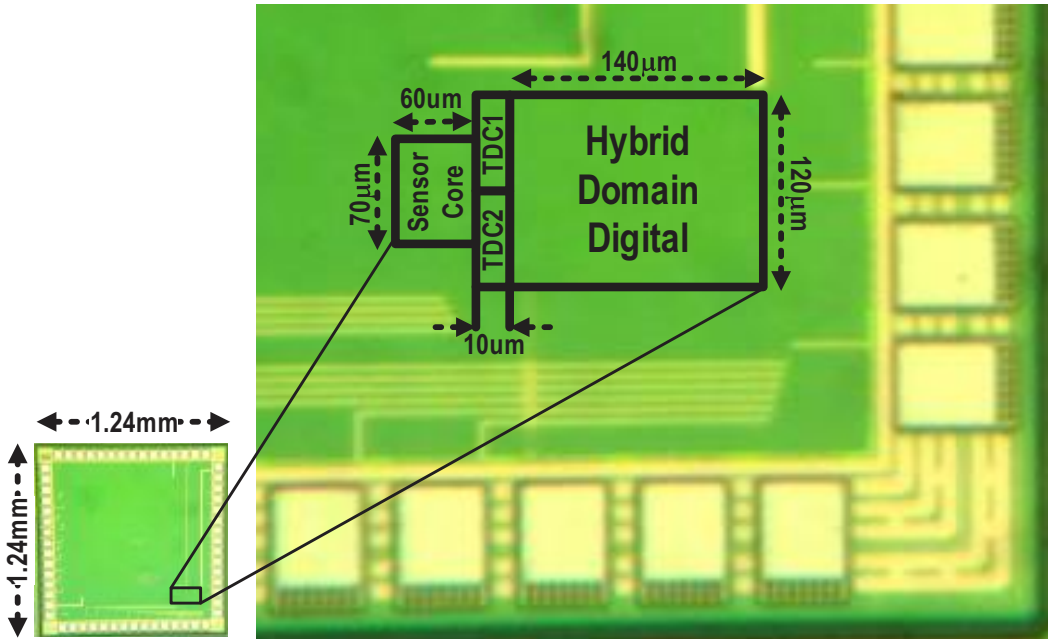


Fig. 7.11: Die micrograph with annotated floorplan.

is used and eight test chips packaged in QFN footprint are measured over the temperature range from 0 to 100°C, with 10°C per step.

The measured delay ratio, which is represented as digital code output of the hybrid domain processing unit, is shown in Fig. 7.12 left, and the statistics of the corresponding adjusted-R^2 coefficient is shown in Fig. 7.12 right. As shown, the measured delay ratio shows larger offset/slope variations when compared to the simulated results in Fig. 7.7. This might be caused by the less accurate subthreshold models. Despite of this, the measured mean value of Adjust-R^2 is 0.9993 and the standard deviation is only 0.0004, showing good linearity. Fig. 7.13 shows the measured sensor inaccuracy over 8 test chips. After 2-point calibration at 10°C and 90°C, the measured error of the hybrid domain temperature sensor is -1.6^{o}C/1°C across 0$\sim$100°C.

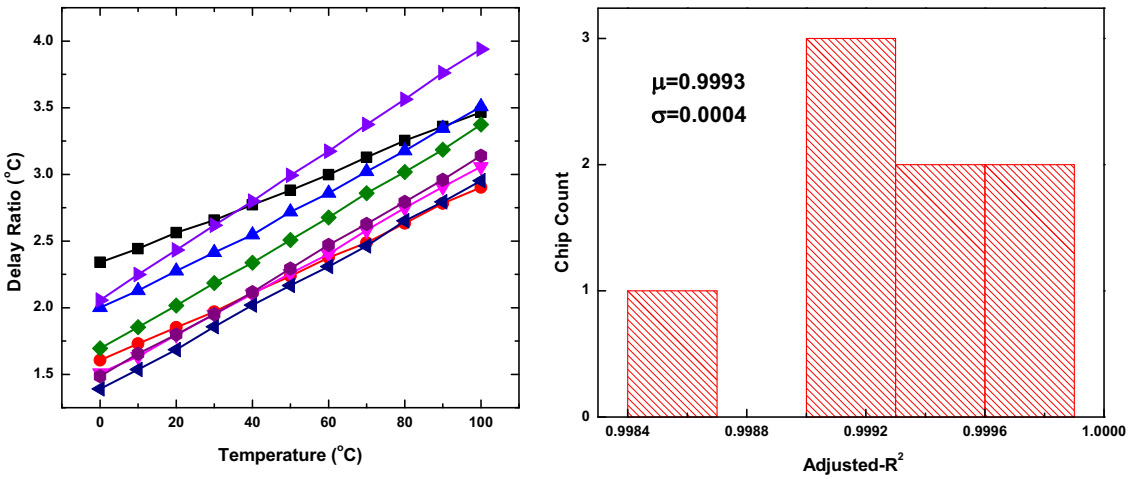


Fig. 7.12: Measured delay ratio (left) and adjusted-R^2 coefficient (right) of 8 chips.

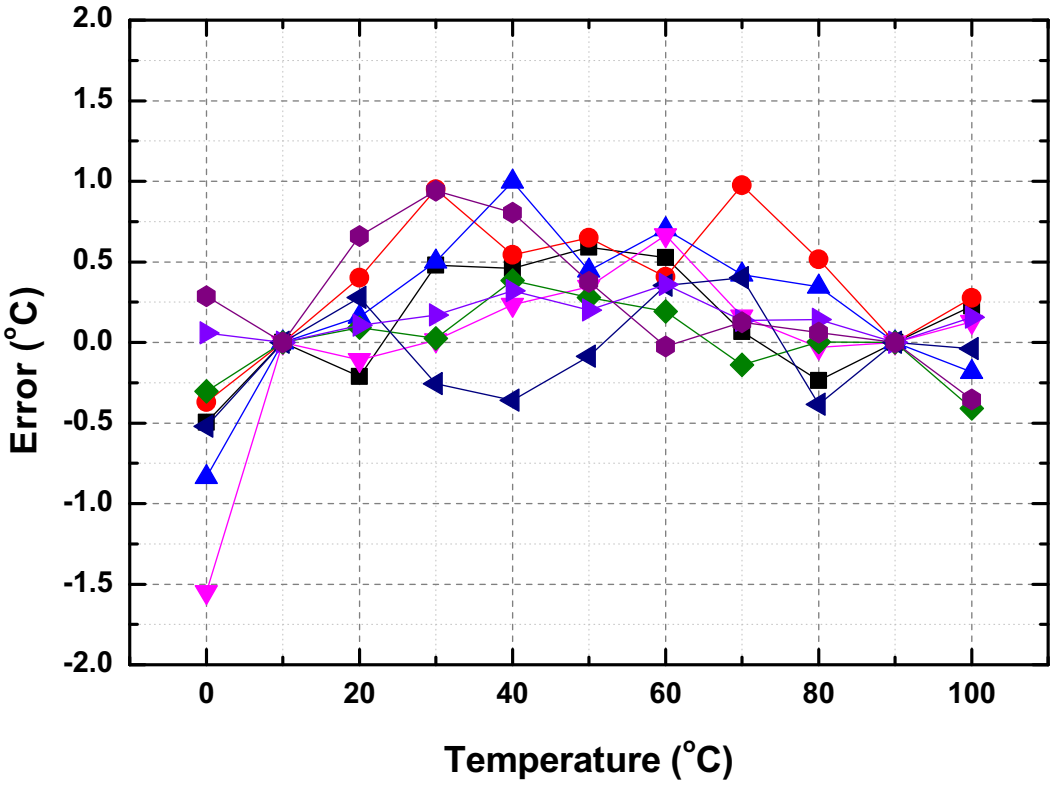


Fig. 7.13: Measured temperature error across 8 dies.

The worst case sampling rate of the hybrid domain temperature sensor can be operated over 800Hz 10% duty-cycle clock (used for charging the capacitor, not for reference) at 0^oC. We take 20 samples to obtain the average output to reduce the read-out error and this results in 40 conversion/sec. With this rate, the test chips dissipate on average 280nW power consumption supply at room temperature. For $0\sim100^oC$ sensing range, the 9 bit fractional digits lead to average resolution of 0.25^oC across eight chips.

Table 7.2 shows the summary of the state-of-the-art nano-watt temperature sensors. Previous nW implementations generally require 1V/1.2V supply, or even dual supply voltages. Also, analog blocks, such as resistors and op-amps, are extensively used. For TDC-based sensors, if external frequency reference power is considered, the total power will exceed the μW limit if a wide sensing range and a moderate sample rate is needed. For FDC-based sensors with integrated frequency reference, the sensor accuracy and power consumption are largely affected by the available resistor types to achieve the desired temperature dependence. As a result, improving area and accuracy with nW power budget is challenging in scaled technologies.

Hybrid domain temperature sensor does not require external accurate frequency reference, and the only used analog block is capacitor. Also, it achieves 0.4V operation with nearly all-digital implementation, which is preferred for wireless sensing systems and technology scaling friendly. Simulation results indicate that over 90% power consumption of the hybrid domain temperature sensor are leakage power. Therefore, our design in 0.18μm will decrease below 100nW. It is also worth noting that the proposed sensor can further reduce power if system building block like MCU is available for ratio calculation, and only sensor core with TDCs are

Table 7.2: Summary of state-of-the-art nW temperature sensor

	This Work	CICC'13 [135]	CICC'08 [134]	TCS II'09 [130]	JSSC'10 [131]
Technology	**65nm**	0.18µm	0.18µm	0.18µm	0.18µm
Area (mm^2)	**0.0054 (Core + TDC) 0.022 (Total)**	0.09	0.05	0.0324	0.042
Analog Blocks	**Cap**	Res, op-amp, etc	Res, op-amp, etc	Res, cap, op-amp, etc	Res, cap, op-amp, etc
Supply Voltage	**0.4V**	1.2V	1V	1V	0.5V, 1V
Frequency Reference Dependence	**No**	Integrated	Integrated	Required	Required
Power Consumption	**280nW**	65nW	220nW	405nW	119nW
Range (°C)	**0~100**	0~100	0~100	0~100	-10~30
Calibration Method	**2-point**	2-point	2-point	2-point	2-point
Inaccuracy (°C)	**-1.6~+1**	-1.4~+1.3	-1.6~+3	-0.8~+1	-1~+0.8
Resolution (°C)	**0.25**	0.3	0.3	0.3	0.2
Sample/second	**40**	32	10	1000	33

needed.

The proposed temperature sensor design proves that ultra-low voltage (down to 0.4V), nearly all-digital implementation with decent linearity and inaccuracy are feasible in nanometer technologies. Comparable performance can be achieved when compared to the state-of-the-art designs. Especially, the hybrid domain processing concept achieves no dependence on the frequency reference.

7.6 Conclusion

In this chapter, we present the design and implementation of a 65nm 0.4V 280nW hybrid domain temperature sensor for wireless sensing platforms. A ratioed-

current/delay sensor core and hybrid domain temperature sensing scheme are proposed to eliminate the generally required frequency reference and the measured inaccuracy after 2-point calibration is -1.6^{o}C/1°C across 0$\sim$100°C with 40 samples/second rate. Due to the nearly all-digital implementation, the proposed design shows reduced design effort and is suitable for scaled technologies.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

Sub-/near-V_{th} operation is a compelling circuit design strategy for reducing the power consumption in CMOS digital integrated circuits, which also provides more than an order of magnitude energy reduction compared to the nominal supply voltage operation. However, the compromised device characteristics, together with the increased sensitivity to process variations under sub-/near-V_{th} operation, bring along great challenges in achieving a proper design trade-off between robustness, energy/area efficiency, and performance. This thesis presents our research work for designing robust sub-/near-V_{th} circuits with minimum energy/area overhead and performance boosting target on different design entries.

First, a near-V_{th} ASIC design flow with statistical timing analysis incorporating design-time forward body-biasing to reduce the excessive design margin and boost the performance is introduced in Chapter 3. We proposed the Surrogate Model Adjustment based Statistical Static Timing Analysis (SMA-SSTA) for re-

ducing the runtime cost of standard cell characterization for local variations and statistical timing analysis. In addition, a novel SelF-Body-Biasing (SFBB) scheme is proposed for overhead free performance boosting purposes. The two synergistic approaches enable the variation-aware body-biased design for near-V_{th} ASIC design for the first time. The benefit of the proposed design flow is experimentally verified on a near-threshold AES encryption engine fabricated in a commercial 65nm low leakage process. Through the co-design of architecture/design flow, the measured testchip delivers 12.2Mbps, with 1.65pJ/bit at 0.5V, which is 22× and 7.8× over a state-of-the-art AES design, while still reducing 28% silicon area.

Second, customized designs for several key ULV building blocks are demonstrated, as listed below:

- Energy-efficient level shifter design

  In Chapter 4, we proposed the design of an NMOS-diode current limiter based level shifter. Several circuit techniques, such as MTCMOS and inverse-narrow-width-effect (INWE) aware sizing, are explored for further improving the energy efficiency of the level shifter. Measurement results shows the proposed level shifter design achieves on average 25.1ns propagation delay and 30.7fJ/bit operation when converting a 300mV input to 1.2V.

- Robust and energy-efficient intra-cell mixed-V_{th} standard cell design methodology

  In Chapter 5, we proposed a novel intra-cell mixed-V_{th} standard cell design methodology for robust ULV operation with improved energy-efficient. The proposed solution replace the bottleneck devices in ULV logic cells with LVT devices to maintain the cell area and the energy efficiency is maintained due

to the reduced parasitics. Library level experiments validate that the proposed methodology achieves on average 30.1% energy-efficiency improvements compared with the previous device upsizing techniques.

- Energy/Area-Efficient Hidden-Refresh eDRAM

  In Chapter 6, we explored the potential energy benefits of eDRAM as an alternative to SRAM. The eDRAM provides higher density but the refresh operation complicates the system-level design due to the reduced availability during refresh. We proposed a hidden-refresh eDRAM design to have a SRAM-like interface and several circuit techniques are introduced to realize true single-VDD operation. Simulation results show that eDRAM has promising density and energy benefits over the SRAM counterpart, which is a promising design choice for memory-intensive ULP systems.

Finally, we propose a 0.4V 280nW nearly all-digital hybrid domain sub-V_{th} temperature sensor capable of ULV operation. A ratioed-current/delay PTAT sensor core and hybrid domain temperature sensing scheme are proposed to eliminate the dependence on frequency reference. After 2-point calibration, the eight testchips show measured inaccuracy of -1.6^{o}C/1°C across 0$\sim$100°C range, with 40 samples/second sample rate.

8.2 Future Work

We will further explore the effectiveness of the proposed intra-cell mixed-V_{th} standard cell design methodology in real silicon designs. It will be applied to a wide variety of designs, including both dedicated hardware accelerators and general-purpose micro-controller. The application of the hidden-refresh eDRAM will be

further investigated with novel bitcell design for improving retention and reducing the static power. In addition, it will be more beneficial to incorporate architectural exploration of both SRAM/eDRAM in ULP systems to find out the optimal combination. For the temperature sensor design, further extending this sensor type to multiple application scenarios require efforts to reduce the calibration workload.

In addition, low power clocking circuit such as efficient KHz/MHz frequency reference are necessary for self-contained mm^3-scale sensor platforms as crystal oscillators are too bulky. Also, power management circuits are essentially important as sensor node computing platforms will be largely dependent upon the energy harvesting devices. A future sensory platform with above-mentioned techniques/blocks are planned and will be implemented in the near future.

Bibliography

[1] G. Moore, "No exponential is forever: but 'forever' can be delayed!" in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2003, pp. 20 - 23.

[2] J. G. Koomey, S. Berard, M. Sanchez, and H. Wong, "Implications of historical trends in the electrical efficiency of computing," *IEEE Annuals of the History of Computing*, vol. 33, no. 3, pp. 46 - 54, 2011.

[3] S. Borkar, "Design challenges of technology scaling," *IEEE MICRO*, vol. 19, no. 4, pp. 23 - 29, 1999.

[4] R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circuits*, vol. 7, no. 4, pp. 146 - 153, Apr. 1972.

[5] J. D. Meindl and J. A. Davis, "The fundamental limit on binary switching energy for terascale integration (TSI)," *IEEE J. Solid-State Circuits*, vol. 35, no. 10, pp. 1515 - 1516, Oct. 2000.

[6] B. Calhoun, A. Wang, and A. P. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1778 - 1786, Sept. 2005.

[7] A. Wang, and A. P. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310 - 319, Jan. 2005.

[8] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "The limit of dynamic voltage scaling and insomniac dynamic voltage scaling," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, no. 11, pp. 1239 - 1252, Nov. 2005.

[9] M. Alioto, "Ultra-Low power VLSI circuit design demystified and explained: a tutorial," *IEEE Trans. Circuit and Syst. I: Reg. Papers*, vol. 59, no. 1, pp. 2 - 29, Jan, 2012.

[10] H. Esmaeilzadeh, *et al.*, "Dark silicon and the end of multicore scaling," *IEEE MICRO*, vol. 32, no. 3, pp. 122 - 134, 2012.

[11] L. Wang, K. Skadron, and B. H. Calhoun, "Dark vs. dim silicon and near-threshold computing," *In Dark Silicon Workshop in conjunction with ISCA*, 2012.

[12] R. Dreslinski, *et al.*, "Near-threshold computing: reclaiming moore's law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253 - 266, Feb. 2010.

[13] E. Krimer, *et al.*, "Synctium: a near-threshold stream processor for energy-constrained parallel applications," *IEEE Computer Architecture Letters*, pp. 21 - 24, 2010.

[14] D. Fick, *et al.*, "Centip3de: a 3930 DMIPS/W configurable near-threshold 3D stacked system with 64 ARM Cortex-M3 cores," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2012, pp. 190 - 192.

[15] S. Borkar, *et al.*, "The future of microprocessors," *Communications of the ACM*, vol. 54, no. 5, pp. 67 - 77, 2011.

[16] B. Warneke, M. Last, B. Liebowitz, and K. S. J. Pister, "Smart dust: communicating with a cubic-millimeter computer," *Computer*, vol. 34, pp. 44 - 51, 2001.

[17] Yang, G.Z., Ed. *Body Sensor Networks*; Springer: London, UK, 2006.

[18] R. Sarpeshkar, *Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications, and Bio-Inspired Systems*. Cambridge, U.K. Cambrige Univ. Press, 2010.

[19] L. Doherty, B. A. Warneke, B. E. Boser, and K. S. J. Pister, "Energy and performance considerations for smart dust," *Int. J. Parallel Distrib. Syst. Networks*, vol. 4, no. 3, pp. 121 - 133, 2001.

[20] B. H. Calhoun, *et al.*, "Design considerations for ultra-low energy wireless microsensor nodes," *IEEE Trans. Computers*, vol. 54, no. 6, pp. 727 - 740, June. 2006.

[21] A. P. Chandrakasan, N. Verma, and D. C. Daly, "Ultralow-power electronics for biomedical applications," *Annual Review of BioMed. Eng.*, pp. 247 - 274, Apr. 2008.

[22] G. Chen, S. Hanson, D. Blaauw, and D. Sylvester, "Circuit design advances for wireless sensing applications," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1808 - 1827, Nov. 2011.

[23] T. Sakurai and A. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584 - 594, Apr. 1990.

[24] K. Flautner, S. Reinhardt, and T. Mudge, "Automatic performance setting for dynamic voltage scaling," *in Proc. 7th Annu. Int. Conf. Mobile Computing and Networking (MobiCom'01)*, May. 2001, pp. 260 - 271.

[25] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *proc. Int'l Symp. Low Power Electro. and Design*, Aug. 2005, pp. 20 - 25.

[26] S. Hanson, M. Seok, D. Sylvester, and D. Blauw, "Nanometer device scaling in subthreshold logic and SRAM," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 175 - 185, Jan. 2008.

[27] D. Bol, R. Ambroise, D. Flandre, and J.-D. Legat, "Interests and limitations of technology scaling for subthreshold logic," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.,* vol. 17, no. 10, pp. 1508 - 1519, Oct. 2009.

[28] J. Rabaey, "*Digital Integrated Circuits: A Design Perspective*," Prentice Hall, 2003.

[29] M. Seok, D. Sylvester, and D. Blaauw, "Optimal technology selection for minimizing energy and variability in low voltage applications," in *proc. Int'l Symp. Low Power Electro. and Design*, 2008, pp. 9 - 15.

[30] J. Keane, H. Eom, T. H. Kim, S. Sapatnekar, and C. Kim, "Subthreshold logical effort: a systematic framework for optimal subthreshold device sizing, " in *proc. Design Automation Conference*, 2006, pp. 425 - 428.

[31] T. H. Kim, H. Eom, J. Keane, and C. Kim, "Utilizing reverse short channel effect for optimal subthreshold circuit design," in *proc. Int'l Symp. Low Power Electro. and Design*, 2006, pp. 127 - 130.

[32] J. Zhou, *et al.*, "A 40 nm inverse-narrow-width-effect-aware sub-threshold standard cell library," in *proc. Design Automation Conference*, 2011. pp. 441 - 446.

[33] B. Liu, M. Ashouei, J. Huisken, and J. P. de Gyvez, "Standard cell sizing for subthreshold operation," in *proc. Design Automation Conference*, 2012, pp. 962 - 967.

[34] J. Kwong and A. Chandrakasan, "Variation-Driven device sizing for minimum energy sub-threshold circuits," in *proc. Int'l Symp. Low Power Electro. and Design*, Aug 2006, pp. 8 - 13.

[35] Y. Pu, J. P. de Gyvez, H. Corporaal, and Y. Ha, "V_T balancing and device sizing towards high yield of sub-threshold static logic gates," in *proc. Int. Symp. on Low Power Electronics and Designs*, pp. 355 - 358, Aug 2007.

[36] N. Lotze, and Y. Manoli, "A 62 mV 0.13 μm CMOS standard-cell-based design technique using schmitt-trigger logic," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 47 - 60, Jan. 2012.

[37] N. Verma, J. Kwong, and A. P. Chandrakasan, "Nanometer MOSFET variation in minimum energy subthreshold circuits," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 163 - 174, Jan. 2008.

[38] B. Calhoun and A. P. Chandrakasan, "Static noise margin variation for subthreshold SRAM in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 42, no. 3, pp. 680 - 688, Mar. 2007.

[39] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "A variation-tolerant sub-200 mV 6-T subthreshold SRAM," *IEEE J. Solid-State Circuits*, vol. 43, no. 10, pp. 2338 - 2348, Oct. 2008.

[40] M.-F. Chang, *et al.*, "A sub-0.3 V area-efficient L-shaped 7T SRAM with bitline swing expansion schemes based on boosted read-bitline, asymmetric-V_{TH} read-port, and offset cell VDD biasing techniques, " *IEEE J. Solid-State Circuits*, vol. 48, no. 10, pp. 2558 - 2569, Oct. 2013.

[41] N. Verma, and A. P. Chandrakasan, "A 256 kb 65 nm 8T subthreshold SRAM employing sense-amlifier redundancy," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 141 - 149, Jan. 2008.

[42] M. E. Sinangil and A. P. Chandrakasan, "A reconfigurable 8T ultra-dynamic voltage scalable (U-DVS) SRAM in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 3163 - 3173, Nov. 2009.

[43] T.-H. Kim, J. Liu, and C. H. Kim, "A voltage scalable 0.26 V, 64 kb 8T SRAM with Vmin lowering techniques and deep sleep mode," *IEEE J. Solid-State Circuits*, vol. 44, no. 6, pp. 1785 - 1795, 2009.

[44] M.-F. Chang, S.-W. Chang, P.-W. Chou, and W.-C. Wu, "A 130 mV SRAM with expanded write and read margins for subthreshold applications," *IEEE J. Solid-State Circuits*, vol. 46, no. 2, pp. 520 - 529, Feb. 2011.

[45] B. H. Calhoun and A. P. Chandrakasan, "A 256 kb sub-threshold SRAM in 65 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006, pp. 2592 - 2601.

[46] T.-H. Kim, J. Liu, J. Keane, and C. H. Kim, "0.2 V, 480 kb subthreshold SRAM with 1 k cells per bitline for ultra-low-voltage computing," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 518 - 529, 2008.

[47] I. J. Chang, J. J. Kim, S. P. Park, and K. Roy, "A 32 kb 10T subthreshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2008, pp. 388 - 622.

[48] W. K. Luk and R. H. Dennard, "A novel dynamic memory cell with internal voltage gain," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 884 - 894, April. 2005.

[49] B. Zhai, *et al.*, "Energy-efficient subthreshold processor design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 8, pp. 1127 - 1137, Aug. 2009.

[50] T.-H. Chen, J. Chen, and L. T. Clark, "Subthreshold to above threshold level shifter design," *J. Low Power Electron.*, vol. 2, no. 2, pp. 251 - 258, Aug. 2006.

[51] H. Shao and C. Tsui, "A robust, input voltage adaptive and low energy consumption level converter for sub-threshold logic," in *proc. ESSCIRC,* pp. 312 - 315, 2007.

[52] S. N. Wooters, B. H. Calhoun, and T. N. Blalock, "An energy-efficient subthreshold level converter in 130-nm CMOS," *IEEE Trans. Circuits. Syst. II, Exp. Briefs*, vol. 57, no. 4, pp. 290 - 294, Apr. 2010.

[53] I. J. Chang, J. Kim, K. Kim, and K. Roy, "Robust level converter for subthreshold/superthreshold operation: 100 mV to 2.5 V," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 8, pp. 1429 - 1437, Aug. 2011.

[54] Y. Kim, D. Sylvester, and D. Blaauw, "LC^2: limited-contention level converter for robust wide-range voltage conversion," in *Symp. VLSI Circuits Dig. Tech. Papers*, pp. 188 - 189, 2011.

[55] Y. Kim, Y. Lee, D. Sylvester, and D. Blaauw, "SLC: split-control level converter for dense and stable wide-range voltage conversion," in *proc. ESSCIRC,* pp. 478 - 481, 2012.

[56] Y. Osaki, T. Hirose, N. Kuroki, and M. Numa, "A low-power level shifter with logic error correction for extremely low-voltage digital CMOS LSIs," *IEEE J. Solid-State Circuits*, vol. 47, no. 7, pp. 1776 - 1783, Jul. 2012.

[57] S. Lütkemeier and U. Rückert, "A subthreshold to above-threshold level shifter comprising a wilson current mirror," *IEEE Trans. Circuits. Syst. II, Exp. Briefs*, vol. 57, no. 9, pp. 721 - 724, Sep. 2010.

[58] M. Lanuzza, P. Corsonello, and S. Perri, "Low-Power Level Shifter for Multiple-Supply Voltage Designs," *IEEE Trans. Circuits. Syst. II, Exp. Briefs*, vol. 59, no. 12, pp. 922 - 926, Dec. 2012.

[59] S. Hanson, *et al.,* "Exploring variability and performance in a sub-200-mV processor," *IEEE J. of Solid-State Circuits*, vol. 43, no. 4, pp. 881 - 891, April, 2008.

[60] M. Hwang and K. Roy., "ABRM: adaptive β-ratio modulation for process-tolerant ultradynamic voltage scaling," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 2, pp. 281 - 290, Feb, 2010.

[61] Y. Pu, J. P. de Gyvez, H. Corporaal, and Y. Ha, "An ultra-low-energy multi-standard JPEG co-processor in 65 nm CMOS with sub/near threshold supply voltage," *IEEE J. of Solid-State Circuits*, vol. 45, no. 3, pp. 668 - 680, Mar, 2010.

[62] Y. Ho, and C. Su, "A 0.1-0.3 V 40-123 fJ/bit/ch on-chip data link with ISI-suppressed bootstrapped repeaters," *IEEE J. of Solid-State Circuits*, vol. 47, no. 5, pp. 1242 - 1251, May, 2012.

[63] Y. Ho, Y.-S. Yang, C. C. Chang, and C. Su, "A near-threshold 480 MHz 78 μW all-digital PLL with a bootstrapped DCO," *IEEE J. of Solid-State Circuits*, vol. 48, no. 11, pp. 2805 - 2814, May, 2012.

[64] R. Rithe, C.-C. Cheng, and A. P. Chandrakasan, "Quad full-HD transform engine for dual-standard low-power video coding," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 11, pp. 2724 - 2736, Nov. 2012.

[65] M. Tikekar, C.-T. Huang, C. Juvekar, V. Sze, and A. P. Chandrakasan, "A 249-Mpixel/s HEVC video-decoder chip for 4K Ultra-HD applications," *IEEE Journal of Solid-State Circuits*, preprint.

[66] D. Jeon, M. Seok, C. Chakrabarti, D. Blaauw, and D. Sylvester,"A super-pipelined energy efficient subthreshold 240MS/s FFT core in 65nm," *IEEE J. of Solid-State Circuits*, vol. 47, no.1, pp. 23 - 34, Jan. 2012.

[67] A.R. Sadeghi, D. Naccache (Eds.), *Towards Hardware-Intrinsic Security*, Springer, 2010.

[68] M. Seok, D. Blaauw, and D. Sylvester, "Clock network design for ultra-low power applications," in *proc. Int'l Symp. Low Power Electro. and Design*, 2010, pp. 271C276.

[69] M. Meijer, J. P. de Gyvez, and A. Kapoor, "Ultra-low-power digital design with body biasing for low area and performance-efficient operation", *ASP J. of Low Power Electro.*, vol. 6, no. 4, pp. 1 - 12, 2011.

[70] L. Nazhandali, *et al.*, "Energy optimization of subthreshold-voltage sensor network processors," in *proc. Int. Symp. Comput. Archit.*, 2005, pp. 197 - 207.

[71] B. Zhai, *et al.*, "2.60 pJ/Inst subthreshold sensor processor for optimal energy efficiency," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2006, pp. 154 - 155.

[72] S. Hanson, *et al.*, "A low-voltage processor for sensing applications with picowatt standby mode," *IEEE J. of Solid-State Circuits*, vol. 44, no. 4, pp. 1145 - 1155, April. 2009.

[73] S. C. Jocke, *et al.*, "A 2.6-μW sub-threshold mixed-signal ECG SoC ," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2009, pp. 60 - 61.

[74] J. Kwong, Y.K. Ramadass, N. Verma, and A. Chandrakasan, "A 65 nm Sub-V_t microcontroller with integrated SRAM and switched capacitor DC-DC converter," *IEEE J. of Solid-State Circuits*, vol. 44, no. 1, pp.115 - 126, 2009.

[75] S. Lütkemeier, *et al.*, "A 65 nm 32 b subthreshold processor with 9T multi-V_t SRAM and adaptive supply voltage control," *IEEE J. of Solid-State Circuits*, vol. 48, no. 1, pp. 8 - 19, Jan. 2013.

[76] D. Bol, *et al.*, "Sleepwalker: A 25-MHz 0.4-V sub-mm^2 7-μW/MHz microcontroller in 65-nm LP/GP CMOS for low-carbon wireless sensor nodes," *IEEE J. of Solid-State Circuits*, vol. 48, no. 1, pp. 20 - 32, Jan. 2013.

[77] N. Ickes, D. Finchelstein, and A. Chandrakasan, "A 10-pJ/instruction, 4-MIPS micropower DSP for sensor applications," in *proc. ASSCC*, Nov. 2008, pp. 289 - 292.

[78] M. Ashouei, *et al.*, "A voltage-scalable biomedical signal processor running ECG using 13 pJ/cycle at 1 MHz and 0.4 V," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2011, pp. 332 - 334.

[79] J. Kwong and A. P. Chandrakasan, "An energy-efficient biomedical signal processing platform," *IEEE J. of Solid-State Circuits*, vol. 46, no. 7, pp. 1742 - 1753, Jan. 2011.

[80] S. R. Sridhara, *et al.*, "Microwatt embedded processor platform for medical system-on-chip applications," *IEEE J. of Solid-State Circuits*, vol. 46, no. 4, pp. 721 - 730, April. 2011.

[81] M. Konijnenburg, *et al.*, "Reliable and Energy-Efficient 1MHz 0.4V Dynamically Reconfigurable SoC for ExG Applications in 40nm LP CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2013.

[82] H. Kaul, *et al.*, "A 320 mV 56 μW 411 GOPS/Watt ultra-low voltage motion estimation accelerator in 65 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 316 - 317.

[83] H. Kaul, *et al.*, "A 300 mV 494 GOPS/W reconfigurable dual-supply 4-way SIMD vector processing accelerator in 45 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 260 - 261.

[84] S. Mathew, *et al.*, "53 Gbps native GF $(2^4)^2$ composite-field AES-encrypt/decrypt accelerator for content-protection in 45 nm high-performance microprocessors," in *Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2010, pp. 169 - 170.

[85] H. Kaul, *et al.*, " A 1.45GHz 52-to-162GFLOPS/W variable-precision floating-point fused multiply-add unit with certainty tracking in 32nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2012, pp. 182 - 184.

[86] G. Ruhl, *et al.*, "IA-32 processor with a wide-voltage-operating range in 32-nm CMOS," *IEEE MICRO*, vol. 33, no. 2, pp. 28 - 36, 2013.

[87] G. Chen, *et al.*, "Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2010, pp. 288 - 289.

[88] Y. Lee, *et al.*, "A modular 1mm^3 die-stacked sensing platform with optical communication and multi-modal energy harvesting," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2010, pp. 288 - 289.

[89] F. Zhang, *et al.*, "A batteryless 19 μW MICS/ISM-band energy harvesting body area sensor node SoC," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2012, pp. 298 - 300.

[90] G. Chen, *et al.*, "A cubic-millimeter energy-autonomous wireless intraocular pressure monitor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2011, pp. 310 - 311.

[91] M. Alioto, "Guest editorial for the special issue on ultra-low-voltage VLSI circuits and systems for green computing," *IEEE Trans. Circuits. Syst. II, Exp. Briefs*, vol. 59, no. 12, pp. 849 - 852, Dec. 2012.

[92] M. Orshansky, S. Nassif, D. Boning, *Design for Manufacturability and Statistical Design*, Springer, 2008.

[93] M. Meijer and J. P. de Gyvez, "Body-bias-driven design strategy for area- and performance-efficient CMOS circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 1, pp. 42 - 51, Jan, 2012.

[94] S.-F. Hsiao, M.-C. Chen, and C.-S. Tu, "Memory-free low-cost designs of advanced encryption standard using common subexpression elimination for subfunctions in transformations," *IEEE Trans. Circuits Syst. I, Reg. Papers,* vol. 53, no. 3, pp. 615 - 626, Mar. 2006.

[95] P.-C. Liu, J.-H. Hsiao, H.-C. Chang, and C.-Y. Lee, "A 2.97 Gb/s DPA-resistant AES engine with self-generated random sequence," in *proc. ESSCIRC*, Spet. 2011, pp. 71 - 74.

[96] S. K. Mathew, *et al.,* "53 Gbps native GF$(2^4)^2$ composite-field AES-encrypt/decrypt accelerator for content-protection in 45 nm high-performance microprocessors," *IEEE J. of Solid-State Circuits*, vol. 46, no. 4, pp. 767 - 776, April, 2011.

[97] M. Feldhofer, J. Wolkerstorfer and V. Rijmen, "AES implementation on a grain of sand," *IEE Proc. on Inf. Secur.*, vol. 152, no. 1, pp. 13 - 20, Oct, 2005.

[98] C. Hocquet, *et al.,* "Harvesting the potential of nano-CMOS for lightweight cryptography: an ultra-low-voltage 65 nm AES coprocessor for passive RFID tags," *Springer J. of Crypto. Eng.*, vol. 1, no. 1, pp. 79 - 89, 2011.

[99] T. Good and M. Benaissa, "692-nW advanced encryption standard (AES) on a 0.13-μm CMOS," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 12, pp. 1753 - 1757, Dec, 2010.

[100] P. Hämäläinen, T. Alho, M. Hännikäinen, and T. Hämäläinen, "Design and implementation of low-area and low-power AES encryption hardware core," in *proc. DSD*, pp. 577 - 583, 2006.

[101] R. Rithe, *et al.*, "The effect of random dopant fluctuations on logic timing at low voltage," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 5, pp. 911 - 924, May, 2012.

[102] N. Ickes, *et al.*, "A 28 nm 0.6 V low power DSP for mobile applications," *IEEE J. of Solid-State Circuits*, vol. 47, no. 1, pp.35 - 46, 2012.

[103] L. Xie, A. Davoodi, J. Zhang, and T-H. Wu, "Adjustment-based modeling for timing analysis under variability," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 28, no. 7, pp. 1085 - 1095, July, 2009.

[104] S. Narendra, *et al.*, "Ultra-low voltage circuits and processor in 180nm and 90nm technologies with a swapped-body biasing technique," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 156 - 158, Feb, 2004.

[105] S. K. Mathew, *et al.*, "Sub-500ps 64b ALUs in 0.18μm SOI/bulk CMOS: design and scaling trends," *IEEE J. of Solid-State Circuits*, vol. 36, no. 11, pp. 1636 - 1646, Nov, 2001.

[106] D. Canright, "A very compact rijndael s-box," Naval Postgraduate School, Monterey, CA, *Tech. Rep. NPS-MA-04-011*, 2005.

[107] M. M. Wong, M. L. D. Wong, A. K. Nandi, and I. Hijazin, "Construction of optimum composite field architecture for compact high-throughput AES S-boxes," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 6, pp. 1151 - 1155, June, 2012.

[108] X. Zhang and K. K. Parhi, "High-speed VLSI architectures for the AES algorithm," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 9, pp. 957 - 967, Sep, 2004.

[109] D. Markovic, and R. Brodersen, *DSP Architecture Design Essentials*, Springer, 2012

[110] Synopsys. Liberty NCX User Guide Version F-2011.06, June 2011.

[111] M. Seok, "A fine-grained many V_T design methodology for ultra low voltage operations," in *proc. Int'l Symp. Low Power Electro. and Design*, pp. 161 - 166, Aug 2012.

[112] C. S. Nagarajan, L. Yuan, G. Qu, and B. G. Stamps, "Leakage optimization using transistor-level dual threshold voltage cell library," in *Int'l Symp. on Quality Electronic Design (ISQED)*, pp. 62 - 67, Mar 2009.

[113] J. Kim, and Y. Shin, "Minimizing leakage power in sequential circuits by using mixed V_t flip-flops," in *Int'l Conf. on Computer-Aided Design (ICCAD)*, 2007, pp. 797 - 802.

[114] D. Somasekhar, *et al.*, "2 GHz 2 MB 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 174 - 185, 2009.

[115] K. Chun, *et al.*, "A sub-0.9V logic-compatible embedded DRAM with boosted 3T gain cell, regulated bit-line write scheme and PVT-tracking read reference bias," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2009, pp. 134 - 135.

[116] K. Chun, P. Jain, T. Kim, and C. H. Kim, "A 1.1 V, 667 MHz random cycle, asymmetric 2T gain cell embedded DRAM with 99.9 percentile retention time of 110 μsec," in *Symp. VLSI Circuits Dig. Tech. Papers*, Jun.2010, pp. 192 - 193.

[117] K. Chun, W. Zhang, P. Jain, and C. Kim, "A 700 MHz 2T1C embedded DRAM macro in a generic logic process with no boosted supplies," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2011, pp. 506 - 507.

[118] W. Zhang, K. Chun, and C. Kim, "A write-back-free 2T1D embedded DRAM with local voltage sensing and a dual-row-access low power mode," in *proc. Custom Integr. Circuits Conf.*, 2012, pp. 1 - 4.

[119] Y. Lee, M.-T. Chen, J. Park, D. Sylvester, and D. Blaauw, "A 5.42nW/kB retention power logic-compatible embedded DRAM with 2T dual-V_t gain cell for low power sensing applications," in *proc. ASSCC*, 2010, pp. 1 - 4.

[120] X. Liang, R. Canal, G. Wei, and D. Brooks, "Process variation tolerant 3T1D-based cache architectures," in *proc. MICRO*, 2007, pp. 15 - 26.

[121] F. Sebastiano, *et al.*, "A 1.2V 10μW NPN-based temperature sensor in 65nm CMOS with an inaccuracy of $\pm 0.2^o$C (3σ) from -70°C to 125°C," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2010, pp. 312 - 313.

[122] K. Souri, Y. Chae, and K. Makinwa, "A CMOS temperature sensor with a voltage-calibrated inaccuracy of $\pm 0.15^o$C (3σ) from -55°C to 125°C," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2012, pp. 208 - 210.

[123] P. Chen, C.-C. Chen, C.-C. Tsai, and W.-F. Lu, "A time-to-digital-converter-based CMOS smart temperature sensor," *IEEE J. Solid-State Circuits*, vol. 40, no. 8, pp. 1642 -1648, Aug. 2005.

[124] P. Chen, *et al.*, "A fully digital time-domain smart temperature sensor realized with 140 FPGA logic elements," *IEEE Trans. Circuit and Syst. I: Reg. Papers*, vol. 54, no. 12, pp. 2661 - 2668, Dec. 2007.

[125] P. Chen, *et al.*, "A time-domain SAR smart temperature snesor with curvature compensation and a 3σ inaccyracy of -0.4^{o}C$\sim$+0.6^{o}C over a 0°C to 90°C range," *IEEE J. Solid-State Circuits*, vol. 45, no. 3, pp. 600 - 609, Mar. 2010.

[126] P. Chen, *et al.*, "All-digital time-domain smart temperature sensor with an inter-batch inaccuracy of -0.7^{o}C - +0.6^{o}C after one-point calibration," *IEEE Trans. Circuit and Syst. I: Reg. Papers*, vol. 58, no. 5, pp. 913 - 920, Jun. 2011.

[127] K. Woo, *et al.*, "Dual-DLL-based CMOS all-digital temperature sensor for microprocessor thermal monitoring," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 68 - 69.

[128] E. Saneyoshi, K. Nose, M. Kajita, and M. Mizuno, "A 1.1V 35μm$\times$35μ thermal sensor with supply voltage sensitivity of 2°C/10%-supply for thermal management on the SX-9 supercomputer," *Symp. VLSI Circuits Dig. Tech. Papers,* pp. 152 - 153, 2008.

[129] G. R. Chowdhury and A. Hassibi, "An on-chip temperature sensor with a self-discharging diode in 32-nm SOI CMOS," *IEEE Trans. Circuits. Syst. II, Exp. Briefs,* vol. 59, no. 9, pp. 568 - 572, Dec. 2012.

[130] M. K. Law and A. Bermak, "A 405-nW CMOS temperature sensor based on linear MOS operation," *IEEE Trans. Circuits. Syst. II, Exp. Briefs,* vol. 56, no. 2, pp. 891 - 895, Dec. 2009.

[131] M. K. Law, A. Bermak, and H. C. Luong, "A sub-μW embedded CMOS temperature sensor for RFID food monitoring application," *IEEE J. Solid-State Circuits*, vol. 45, no. 6, pp. 1246 - 1255, Mar. 2010.

[132] K. Kim, H. Lee, S. Jung, and C. Kim, "A 366kS/s 400μW 0.0013mm^2 frequency-to-digital converter based CMOS temperature sensor utilizing multiphase clock," in *proc. CICC*, pp. 203 - 206, Sept. 2009.

[133] S. Hwang, *et al.*, "A 0.008 mm^2 500 μW 469 kS/s frequency-to-digital converter based CMOS temperature sensor with process variation compensation," *IEEE Trans. Circuit and Syst. I: Reg. Papers*, vol. 60, no. 9, pp. 2241 - 2248, Sep, 2013.

[134] Y.-S. Lin, D. Sylvester and D. Blaauw, "An ultra low power 1V, 220nW temperature sensor for passive wireless applications," in *proc. CICC*, pp. 507 - 510, Sept. 2008.

[135] S. Jeong, J. Sim, D. Blaauw, and D. Sylvester, "65nW CMOS temperature sensor for ultra-low power microsystems," in *proc. CICC*, pp. 1 - 4, Sept. 2013.

[136] Y. W. Li, and H. Lakdawala, "Smart integrated temperature sensor - mixed-signal circuits and systems in 32-nm and beyond," in *proc. CICC*, pp. 1 - 8, Sept. 2011.

[137] S. Yoshimoto et al., "A 40-nm 0.5-V 12.9-pJ/access 8T SRAM using low-power disturb mitigation technique," *Symp. VLSI Circuits Dig. Tech. Papers,* pp. 77 - 78, 2013.

List of Abbreviations

1. W. Zhao, A. B. Alvarez, Y. Ha, "A 65-nm 30-fJ/bit subthreshold level sonverter for robust and wide range voltage conversion," submitted under review.

2. W. Zhao and Y. Ha, "A 65-nm 12.2-Mbps 1.65-pJ/bit near-threshold AES engine based on novel self-body-biasing and statistical design methodology", submitted to IEEE TVLSI, under revision.

3. W. Zhao, Y. Ha, C. H. Hoo and A. B. Alvarez, "Robustness driven energy-efficient ultra-low voltage standard cell design with intra-cell mixed-Vt methodology", in Proc. of ISLPED, pp. 323-328, 2013.

4. W. Loke, W. Zhao and Y. Ha, "Criticality-based routing for FPGAs with reverse body bias switch box architectures", in Proc. of FPL, pp. 1-6, 2013.

5. W. Loke, Y. Ha and W. Zhao, "A Power and Cluster-Aware Technology Mapping and Clustering Scheme for Dual-VT FPGAs," in Proc. IPDPSW, pp. 221-226. 2012.