

Extreme Power-Constrained Integrated Circuit Design

by

Mingoo Seok

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2011

Doctoral Committee:

Professor Dennis Michael Sylvester, Chair
Professor David Blaauw
Associate Professor Scott Malhke
Assistant Professor David Dale Wentzloff

© Minguo Seok 2011
All Rights Reserved

to my family

ACKNOWLEDGEMENTS

I have been extremely fortunate to interact with and learn from a great group of both professors and fellow students during my graduate study. This work would have been impossible without their contributions. My advisor, Dennis Sylvester and practically another advisor, David Blaauw have guided and inspired me throughout the last five year of graduate study. Jae-sun Seo and Youngmin Kim were also tremendously helpful to me when I first started the doctoral program.

I have been particularly lucky to have great collaborators in all of my projects. Scott Hanson, Yu-Shiang Lin and I worked together to design Phoenix Processor in Chapter II under the guidance of Dennis Sylvester and David Blaauw. Yoonmyung Lee, Zhiyoong Foo, Daeyeon Kim and Rach Liu also contributed to the project's success. Scott Hanson also offered valuable discussions in the study on power gating switches in Chapter IV. Jae-sun Seo and Scott Hanson helped the read-only-memory project in Chapter V. Gyouho Kim contributed to design and test of 2-Transistor voltage references in Chapter VI. Finally, Dongsuk Jeon and I closely worked together to create the world-record energy-efficient Fast Fourier Transform core in Chapter VII.

Not directly involved in the same projects, Bo Zhai, Carlos Tokunaga, Sanjay Pant, Gregory Chen, Michael Wieckowski, Vineeth Veetil, David Fick, Inhee Lee, Matthew Fojtik, Mohammad Hassan, Eric Karl, Youngmin Kim, Prashant Singh, Dongmin Yoon, Yejoong Kim, Jerry Kao, Wei-Shiang Ma, Bharan Girdidhar, Brian Cline, Cheng Zhou, and Sudhir Satpahy have been great discussion partners in the field of VLSI research. I also thank Youngmin Park, Sangwook Han, Daeyoung

Lee, Seunghyun Oh, Jaeyoung Joshua Kang, Jungkap Park, Dongjin Lee, Sangwon Seo, Razi Haque, Myungchul Kim, Hyo Gyuem Rhew, Changwook Min, David Lee, Youngjun Park, Junsun Park, Gwang-hyeon Baek, Youngjoon Song, Younghyun Shim, Jungkook Kim, Geonwook Yoo, Kyunghoon Lee, Seunghyun Lee, Juseop Lee, Sunghyun Cho, Hyungil Chae, Junseok Heo, Byungsoo Kim, Jooseuk Kim, Cheolwon Alexander Min, Seun Park, Deahyun Yoon, Daeyon Jung, Yoojin Choi and Jongwoo Lee for the discussions on the topics outside digital design. I would also like to thank the Korea Foundation of Advanced Studies and the Rackham Graduate School for funding my fellowship for total three years.

Last but not the least, I thank my family (father, mother, wife, partents in law, sister, sister's husband, sisters and a brother in law) and friends for their unconditional supports.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	ix
LIST OF TABLES	xv
LIST OF ABBREVIATIONS	xvi
ABSTRACT	xviii
CHAPTER	
I. Introduction	1
1.1 Emerging Sensing Applications	1
1.2 Challenges of Achieving Cubic Millimeter Scale Sensing Systems	3
1.3 Reducing Power Consumption in ICs	4
1.3.1 Power Consumption in Modern IC design	4
1.3.2 Ultra Low Voltage Operation	5
1.4 Challenges and Contributions	7
1.4.1 Minimizing Standby Power	7
1.4.2 Designing Ultra Low Power Analog Building Blocks	8
1.4.3 Improving Performance, Delay Variability and En- ergy Efficiency	9
1.4.4 Contribution Summary	11
1.5 Organization of this Work	13
II. Phoenix Processor: 35pW Standby and 226nW Active Power Sensor Platform	14
2.1 Motivation and Previous Work	14
2.2 Contribution	16
2.3 System Overview	16

2.4	System-level Optimization including Technology Selection . . .	18
2.5	Power Gating Under Relaxed Performance Constraints	22
2.6	CPU and Instruction Set Design for Standby Mode	24
2.7	DMEM Compression for Standby Mode	26
2.8	Ultra-Low Standby Power Memory Design	30
2.9	Test Chip Overview	34
2.10	Measured Results	35
2.10.1	Power and Performance Results	35
2.10.2	Power Gating Results	37
2.10.3	Memory Results	37
2.11	Summary	41
III. Technology Selections for Ultra Low Voltage Design		42
3.1	Motivation and Previous Work	42
3.2	Contribution	44
3.3	Ultra Low Voltage Application Spaces	44
3.4	Basic Optimal Technology Selection for Minimum Energy . . .	46
3.4.1	Modeling Logic and SRAM for Energy Comparison	46
3.4.2	Technology Choice for Minimum Energy	47
3.5	Impact of Standby Leakage Reduction	48
3.5.1	Leakage Reduction Methods in Ultra Low Voltage Regimes	49
3.5.2	Technology Choice with Leakage Reduction Schemes	50
3.6	Effect of Logic and SRAM Ratio on Technology Selection . . .	51
3.7	Variability and Technology Selection	53
3.7.1	Impact of Operating Point on Variability	54
3.7.2	Technology Selection for Min Variability	55
3.8	Summary	58
IV. Power Gating Switch Design for Ultra Low Voltage Operations		60
4.1	Motivation and Previous Work	60
4.2	Contributions	61
4.3	Impact of Sleep Energy on Total Energy Consumption	62
4.4	The Effects of Cutoff Structures on Total Energy Consumption	64
4.4.1	Theoretical Power Gating Switch	65
4.4.2	Practical Power Gating Switch	67
4.5	Strategy of Design Power Gating Switches	71
4.5.1	PGS Design Strategies in Ultra Low V_{DD} Regimes	71
4.5.2	Comparisons of the Optimization Methods	75
4.5.3	Case Study Using a Fabricated Microprocessor	77
4.6	Feasibility of Minimal-Sized PGSs	78
4.7	Beyond Basic PGSs	80
4.8	Summary	82

V. Robust Ultra Low Voltage ROM Design	83
5.1 Motivation and Previous Work	83
5.2 Contributions	84
5.3 Dynamic NAND Read-Only Memory (ROM) Design	85
5.3.1 Challenges of dynamic NAND ROM	85
5.3.2 On-current to off-current plot	87
5.3.3 A 32-stack dynamic NAND ROM with HVT bleeder	89
5.4 Static NAND AND NAND-NOR ROM	91
5.4.1 Investigating static ROM topologies	91
5.4.2 Static NAND ROM Monte Carlo Analysis	92
5.5 Measurement Results	93
5.6 Summary	96
VI. Pico-Watt 2-Transistor Voltage Reference with Digital Trimmability	97
6.1 Motivation and Previous Work	97
6.2 Contributions	100
6.3 Circuit Design	102
6.4 2T Reference Measured Results	106
6.5 Variability Analysis and Trimming Techniques	108
6.5.1 Statistical Measurement Results	108
6.5.2 Digitally Trimmable 2T Voltage References	109
6.5.3 Analysis and Minimization of Trimming Cost	110
6.6 Variant Designs of 2T Voltage References	113
6.7 Technology Portability	116
6.8 Summary	117
VII. 0.27V, 30MHz, 17.7nJ/transform, 1024-pt complex FFT Core with Super-Pipelining	118
7.1 Motivation and Previous Work	118
7.2 Contribution	118
7.3 Architecture Design	120
7.4 Circuit designs	120
7.4.1 Super-Pipelining Technique and FIFO Design	120
7.4.2 Two-Phase Latch for Less Delay Variability	123
7.4.3 Robust Clock Network Design	125
7.5 Measurement Results and Comparisons	127
7.6 Summary	128
VIII. Robust Clock Network Design for Ultra Low Voltage Operations	130

8.1	Motivation and Previous Work	130
8.2	Contribution	131
8.3	Clock Network Comparison at Ultra Low Voltage Regimes . .	132
8.3.1	Comparison Frameworks	132
8.3.2	Comparison at Nominal Conditions	134
8.3.3	Impact of MOSFET Process Variations	135
8.3.4	Impact of Interconnect Process Variations	137
8.3.5	Driving Interconnects at Ultra Low Voltage Regimes	138
8.4	Impact of Voltage and Technology Scaling	139
8.4.1	Supply Voltage Scaling	140
8.4.2	Technology Scaling	140
8.5	Clock Network Design for a 16b MSP430-like Microcontroller	142
8.6	Summary	145
IX. Conclusions		146
APPENDICES		148
BIBLIOGRAPHY		153

LIST OF FIGURES

Figure

1.1	An implantable intra-ocular pressure sensor (courtesy of Y-S. Lin) .	2
1.2	Supply voltage over technology scaling	5
1.3	V_{min}/E_{min} curve	6
1.4	Components that can directly benefit from the design methodologies of this work in cubic millimeter sensing systems (in gray)	12
1.5	Energy efficiency improvement in the blocks of cubic millimeter sensing systems	12
2.1	The Phoenix Processor.	17
2.2	(a)Energy optimal technology matrix (b)Optimal V_{DD} and energy over technologies	21
2.3	A typical power gating switch.	23
2.4	Footer allocation in the Phoenix Processor.	24
2.5	CPU diagram.	25
2.6	Distribution of temperature in Muskegon, MI in 2006 represented as the difference between temporally adjacent measurements [1]	27
2.7	Hardware support for compression.	28
2.8	Memory support for compression.	29
2.9	Proposed ultra low standby power SRAM cell.	31

2.10	Effectiveness of (a) stack forcing and (b) gate length biasing for leakage reduction	31
2.11	Memory column diagram showing completion detection.	32
2.12	SRAM array architecture for Data SRAM (DMEM)	33
2.13	Phoenix Processor die photo.	34
2.14	Measured frequency and energy consumption.	35
2.15	Measured (a) frequency distribution (b) active mode power distribution at 60 kHz (c) standby mode power distribution for 13 dies at $V_{DD}=0.5V$	36
2.16	Measured (a) frequency and (b) standby leakage as functions of Central Processing Unit (CPU) footer width.	38
2.17	Total energy consumption assuming 1000 instructions are executed every 10 minutes.	38
2.18	Measured (a) frequency and (b) power as functions of temperature.	40
2.19	Computed time profiles of (a) energy and (b) memory size for a temperature measurement routine.	40
3.1	Published sub- or near-threshold VLSI designs.	43
3.2	Application spaces in ultra low voltage operation.	45
3.3	Basic technology selection for (a) logic (b) SRAM.	48
3.4	E_{min} and standby leakage current over leakage reduction ratio (E_{min} shows small dip due to virtual ground bounce)	49
3.5	Result of co-optimization with sweeping (a) duty cycle (b) required performance	50
3.6	Optimal technology selection (a) for logic with co-optimized V_{DD}/PGS , (b) for SRAM with 10x leakage reduction	51
3.7	Energy density per operation and standby power density	52
3.8	Energy optimal technology selection with SRAM/logic area ratios of (a) 0.5 (b) 0.8	53

3.9	(a) Worst energy technology selection for SRAM ratio of 0.8 (b) energy saving ratio [\times] by moving to Figure 3.8(b)	53
3.10	(a) Log I_{ds} - V_{gs} curve, (b) change of variation depending on operating point	55
3.11	Optimal energy operation point over technologies	56
3.12	variability as a function of technology choice	57
3.13	(a) Mismatch-only (b) die-to-die and mismatch, energy per operation increase with less variability	58
4.1	Illustration of task scheduling at different deadlines	61
4.2	Basic PGS configuration	63
4.3	V_{min}/E_{min} curves with different K_{duty} considering sleep energy	63
4.4	V_{min}/E_{min} change with K_{leak} and K_{delay} , (a) V_{min}/E_{min} curves (b) $K_{leak} - V_{min}/E_{min}$, (c) V_{min}/E_{min} curves, (d) $K_{delay} - V_{min}/E_{min}$	66
4.5	K_{leak} and K_{delay} change with PGS width and V_{DD} , (a) width - K_{delay} (b) width - K_{leak} (c) $K_{leak} - K_{delay}$	68
4.6	V_{min}/E_{min} with different PGS sizes, $K_{duty}=100$	70
4.7	Comparison between raising V_{DD} and upsizing PGS in energy optimization	71
4.8	On/off-current of high V_{th} and regular V_{th} devices	72
4.9	Off-current vs. on-current as sweeping PGS width	72
4.10	New V_{min} and optimal PGS size at different K_{duty}	74
4.11	Comparison of three optimization strategies, (a) $K_{duty}V_{min}$ (b) K_{duty} optimal PGS width (c) $K_{duty}E_{min}$	76
4.12	Measured total energy consumption with two different PGS sizes from a test microprocessor	77
4.13	Measured minimal PGS size for functionality	79

4.14	Simulated virtual ground level over different workload and supply voltage	80
4.15	Generic, DTCMOS, and stack-forcing PGS	81
4.16	K_{leak} - K_{delay} curves with different PGSs	81
5.1	Power(left) and area(right) comparisons for SRAM-only Instruction SRAM (IMEM) (projected) and an SRAM/ROM hybrid IMEM (measured).	84
5.2	Schematics of three ROMs for ultra low voltage: (a) dynamic NAND, (b) static NAND, (c) static NAND-NOR.	85
5.3	Beta ratio and on- to off-current ratio (left), on-current reduction over number of stack (for minimum-sized FET) (right)	86
5.4	Current margin plot for 32-stack dynamic NAND ROM with HVT bleeder	88
5.5	Failure rate for the dynamic NAND with half-latches. (1000 Monte Carlo iterations with die-to-die and mismatch variations)	90
5.6	Current margin plot for 32-stack static NAND ROM	92
5.7	Current margin plot for 32-leg static NAND-NOR ROM	93
5.8	1000 Monte Carlo Simulation Program with Integrated Circuit Emphasis (SPICE) simulations for two ROM topologies considering mismatch and die-to-die	94
5.9	Histogram of operating frequency of static NAND ROM	94
5.10	Measured energy-per-operation and frequency	95
5.11	Die photo and dimensions	95
6.1	Power and minimal functional supply voltage comparisons	99
6.2	Schematics of a 2T voltage reference	101
6.3	Proper sizing of two transistors minimizes temperature dependency (simulated results).	103
6.4	A larger output capacitor provides better PSRR (simulated results)	104

6.5	Simulated output referred noise of a 2T voltage reference with different output capacitors	105
6.6	Simulated required V_{th} difference for proper operations	106
6.7	Measured (a) temperature coefficient (b) line sensitivity (c) power supply rejection ratio (d) current consumption of the 2T voltage reference	107
6.8	Measured (a) output voltage (b) temperature coefficient distribution of the 2T voltage referneces in two separate runs	109
6.9	Figure 13 Schematics of trimmable 2T voltage reference	111
6.10	Measured (a) TC and (b) output voltage change with trim settings	111
6.11	(a) Measured reductions of output voltage and temperature coefficient spreads (b) zoomed view	112
6.12	Measured PSRR of the trimmable 2T voltage reference	113
6.13	Simulated output referred noise of the trimmable 2T voltage reference	113
6.14	Schematics of 4T voltage reference	114
6.15	schematics of 2T voltage reference for lower output voltage	115
6.16	Measured temperature coefficients achieved by skewing the size of transistors	115
6.17	Die micrographs; (a) 1st $0.13\mu m$ run, (b) 2nd $0.13\mu m$ run, (c) $0.18\mu m$ run, (d) $65nm$ run	117
7.1	Effect of pipeline depth on energy consumption	119
7.2	A pipelined, $8\times 32b$ input, radix-4, 2-lane, 1024-pt, complex FFT architecture	121
7.3	Energy and throughput improvement by architecture modifications .	121
7.4	A 16b Baugh-Wooley multiplier is super-pipelined with 2-phase latches	122
7.5	Measured energy consumption of differently pipelined multipliers . .	123

7.6	Schematics of commutators and FIFOs	124
7.7	(a) Waveform of operation (b) FIFO energy consumption comparison.	124
7.8	Timing failure rates across Monte-Carlo simulations with random process variations fro two pipelined multipliers	125
7.9	Proposed clock network design with limited buffers and matched interconnects	126
7.10	Measured energy consumption and performance of the proposed FFT core	127
7.11	Die photo of the FFT core implemented in 65nm CMOS	129
8.1	Clock network topologies	132
8.2	Tradeoff between slew and clock network energy	133
8.3	Wire length and required buffer	134
8.4	Energy and skew comparison of clock networks.	135
8.5	(a) Skew with MOSFET variations, (b) Slew with MOSFET variations.	136
8.6	Skew contribution from interconnect variations	138
8.7	Driving a long interconnect without repeaters (a) delay comparison, (b) schematics.	139
8.8	Impact of voltage scaling on (a) skew, (b) slew.	140
8.9	Resistance scaling across technologies	141
8.10	Crossover voltages for $+2\sigma$ skew and slew	142
8.11	Layout view for the APR-ed microprocessor with 3-level buffered H-Tree clock network	143
8.12	Comparison for the clock networks of the 16b microcontroller on (a) skew, (b) slew, (c) energy consumption.	144
8.13	Optimal clock network across supply voltages	145

LIST OF TABLES

Table

2.1	Instruction set architecture overview.	26
2.2	Comparison to other low voltage microcontroller	36
6.1	Recent published designs of low power voltage references	99
6.2	Measurement summary of the proposed 2T voltage references	116
7.1	RC mismatch in clock network	126
7.2	Comparison of the proposed FFT core	127
7.3	Measurement summary of the proposed FFT core	128

LIST OF ABBREVIATIONS

A/D	Analog to Digital
CMOS	Complementary Metal-Oxide-Semiconductor
CPU	Central Processing Unit
DMEM	Data SRAM
DSP	Digital Signal Processing
FFT	Fast Fourier Transform
FO4	Fan-Out-of-4
ISA	Instruction Set Architecture
IROM	Instruction ROM
IMEM	Instruction SRAM
IC	Integrated Circuit
I/O	Input and Output
LS	Line Sensitivity
MDC	Multi-Path Delay Commutator
MEMS	Micro Electro Mechanical Systems
MOSFET	Metal-Oxide-Semiconductor Field Effect Transistor
NFET	Negative Channel Field Effect Transistor
PFET	Positive Channel Field Effect Transistor
PGS	Power Gating Switch
PSRR	Power Supply Rejection Ratio

PMU Power Management Unit

PVT Process, Voltage and Temperature

ROM Read-Only Memory

RF Radio Frequency

SRAM Static Random-Access Memory

SPICE Simulation Program with Integrated Circuit Emphasis

VLSI Very Large Scale Integration

V_{th} Threshold Voltage

V_t Thermal Voltage

ABSTRACT

Extreme Power-Constrained Integrated Circuit Design

by

Mingoo Seok

Chair: Dennis Michael Sylvester

Recently sensing systems of cubic millimeter scale have gained significant attention since they may be embedded virtually anywhere. Particularly, biomedical devices for implanting in human bodies to monitor critical vital signals, are one of the most promising applications of such small scale systems. For developing these systems, there are two challenging requirements; 1) long lifetime for minimal maintenance (e.g., additional surgeries) and 2) small volume for less invasive deployments. Ultra low power circuits are key enablers for these requirements since they ensure longer lifetime and minimize the volume of power sources, which often occupy the dominant portion of system volume.

Voltage scaling is one promising measure for reducing energy consumption of integrated circuits. Scaling supply voltage down to near, or below, the transistor threshold voltage has been shown to provide 10-20 \times energy savings compared to nominal voltage operations. However, supply voltage scaling is not a complete solution to developing a cubic millimeter sensing system since it brings up many challenges, including minimizing standby power, designing key analog building blocks with constrained power

budgets, and improving delay variability, performance, and energy efficiency beyond simple voltage scaling.

This thesis presents new circuit and architecture design approaches to overcome voltage scaling challenges and realize cubic millimeter sensing systems. Our proposed approaches yield record-setting energy efficiencies with numerous silicon demonstration vehicles to prove their efficacy.

We first describe the Phoenix Processor, a sensing platform for demonstrating standby power minimization techniques such as minimally-sized power gating switches and ultra-low leakage memory arrays. A test chip that includes an 8b microcontroller, embedded memories, a timer and temperature sensor, demonstrates a standby power of 35pW, which is 2-3 orders of magnitude lower than previously reported sensor processors.

As a key analog building block for use in many sensing systems, we also propose an ultra low power voltage reference. Test chips, extensively verified in 3 different semiconductor technologies across four different fabrication runs, typically consume down to 2.2pW, or $16000\times$ smaller than previous designs, without compromising temperature, supply voltage, and process stability.

Finally, a Fast-Fourier-Transform (FFT) core is designed to demonstrate techniques for improving performance, variability tolerance, and energy efficiency beyond conventional voltage scaling. A test chips in 65nm CMOS consumes a record-low 17.7nJ/FFT, at least $4\times$ lower than prior state-of-the-art, and operates at 30MHz at only 0.27V.

CHAPTER I

Introduction

1.1 Emerging Sensing Applications

Continued advances in electronics design enable wireless sensing systems, which source environmental data (e.g., temperature) and wirelessly transmit for further processing and data aggregation. Such designs, which can have volumes on the order of several cubic centimeters, have been proposed by both academics and industry. For example, a sensing system [28] can be used to monitor soil moisture for crop irrigation. Also, the pressure inside car tires can be measured by a sensor [32] to notify drivers of any problems. Another example is a pressure sensor [18] designed for monitoring the integrity of stent grafts in arteries. Along with companies, several universities have demonstrated sensing systems including wireless sensors for neural monitoring and stimulation, as well as [108, 38] Micro Electro Mechanical Systems (MEMS) sensors for gas detection [112].

With the success of cubic centimeter sensing systems, there is an emerging demand for smaller sensing systems at the cubic millimeter scale that can be embedded virtually anywhere due to its small volume. We envision a system that includes microprocessors, sensors, actuators, power sources, and radio within the volume constraint of several cubic millimeters. This size microsystem can be easily implanted inside human bodies, construction materials, and clothing, enabling a new set of applications.

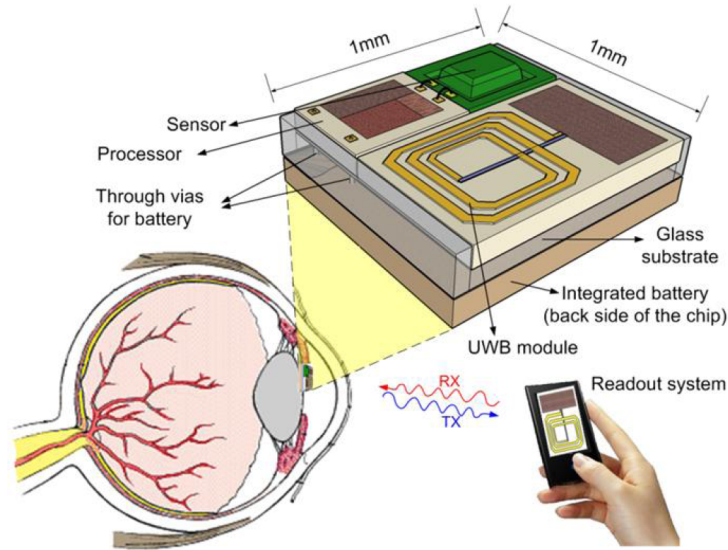


Figure 1.1: An implantable intra-ocular pressure sensor (courtesy of Y-S. Lin)

Aligned with this interest, at the University of Michigan there have been multi-year research projects to develop biomedical sensor systems for the diagnosis and treatment of Glaucoma [63], as illustrated in Figure 1.1. The diagnosis and treatment of Glaucoma disease is expected to help 60.5 million people worldwide by 2010 [34]. However, currently doctors must perform periodic measurements of pressure in the eye (intra-ocular pressure), requiring patients to make frequent trips to a doctor’s office to ensure sufficient temporal resolution of measurements [82]. A cubic millimeter scale sensing system that can be implanted inside a human eye can reduce cost while providing better patient care. In addition to Glaucoma, cubic millimeter sensing systems can potentially benefit a wide range of biomedical applications such as the monitoring of heart and brain signals.

While less volume constrained than biomedical applications, a wide range of general wireless sensing systems could leverage cubic millimeter sensing systems, including surveillance systems used in battlefield monitoring [104, 2]. In addition, small scale sensing systems can be mixed into construction materials (e.g., cement) without impacting their integrity in order to monitor the heat transfer within buildings

in a fine-grained fashion.

1.2 Challenges of Achieving Cubic Millimeter Scale Sensing Systems

To achieve such a small scale sensing system, it is critical to miniaturize all building blocks such as microprocessors, memories, radios, sensors, actuators, and power sources. Continuous reduction of minimum feature size of silicon technology, dictated by Moore's Law, enables building circuit components (e.g., microprocessors, memories, radios) on a cubic millimeter scale. Similarly, advances in MEMS technology have enabled small scale sensors and actuators for meeting the volume constraints of cubic millimeter sensing systems.

However, power source improvements, in both batteries and energy scavengers [55], are much less dramatic. Although recent advances in compact batteries [7] allow for a cubic millimeter power source, the amount of energy stored by the micro battery is also reduced roughly proportionally to the battery volume, imposing a challenge on overall power consumption of the sensing system.

Along with the miniaturization of each component, ensuring long lifetime is equally critical. Since the sensing systems are likely to be embedded in poorly accessible locations (e.g., inside human body or inside building walls), short lifetime lead to higher maintenance costs, reducing the feasibility of such systems.

To meet the volume and lifetime constraints of cubic millimeter sensing systems, ultra low power design of circuits and MEMS components is a key enabler. Ultra low power consumption may allow the use of millimeter scale power sources while enabling lifetime of several years. We can estimate the power budget of a sensing system for guaranteeing one year of lifetime when it is powered by a cubic millimeter power source. Based on a thin film zinc/silver oxide battery with capacity of $100\mu Ah/mm^2$

and output voltage of 1.55V [63], the entire sensing system should consume less than 177pW on average for a one year lifetime. This extremely small power consumption is several orders of magnitude lower than what most energy efficient commercial microprocessors, such as Texas Instruments MSP430, consume in their standby modes [94].

Given this power-limited design space, this thesis focuses on circuit and architectural techniques for extremely power constrained Integrated Circuit (IC) design. In the remainder of this chapter, ultra low voltage operation to minimize power consumption is reviewed. Then, the challenges to operating ICs in the ultra low voltage regime are identified and the contributions of this work are discussed.

1.3 Reducing Power Consumption in ICs

1.3.1 Power Consumption in Modern IC design

Power consumption has been a critical issue in integrated circuit design as it sharply increases in modern sub-micron device technologies. Ideal technology scaling reduces energy in the third order as shown in EQ 1.1. However, achieving such energy gains is not practical for two reasons. First, switching energy is no longer the dominant factor of total energy consumption due to the rapid increase of sub-threshold and gate leakage components in modern Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET)s. Second, much slower supply voltage scaling in recent technologies provides only a linear reduction in switching energy.

$$Delay = \frac{1}{f} = \frac{C_s V_{DD}}{I_{dsat}} \propto \frac{V_{DD}}{(V_{DD} - V_{th})^{1.3}} \quad Power \propto f V_{DD}^2 \quad Energy \propto C_s \cdot V_{DD}^2 \quad (1.1)$$

Supply voltage scaling requires Threshold Voltage (V_{th}) scaling to maintain the similar gate overdrive ($V_{DD} - V_{th}$) of MOSFETs, which dictates transistor perfor-

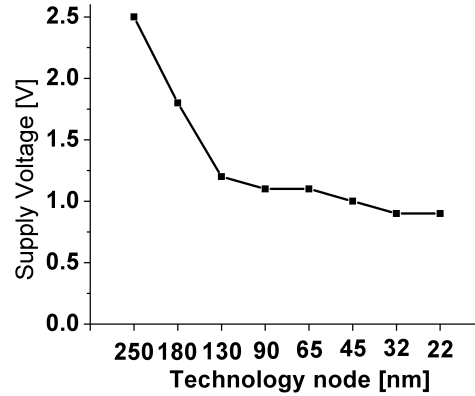


Figure 1.2: Supply voltage over technology scaling

mance. However, as shown in Figure 1.2, supply voltage has held nearly constant around 1V for the last several technology nodes since V_{th} scaling in those nodes leads to prohibitively large subthreshold leakage current. The stagnation of supply voltage scaling and increasing leakage energy result in worsening energy efficiency with technology scaling.

1.3.2 Ultra Low Voltage Operation

Voltage scaling technique has been a promising method to minimize energy consumed in circuits. As the supply voltage scales, quadratic to exponential energy savings in switching, subthreshold leakage, and gate leakage energy can be achieved. However, as pointed out in EQ 1.1, the scaled supply voltage directly degrades circuit performance. Therefore, dynamic voltage scaling is sometimes employed where supply voltage is lowered to a point where circuits can finish a task just before a known deadline. The lower limit of supply voltage with performance constraints usually falls well above the V_{th} of MOSFETs, as shown in [96, 109, 76].

However, we can scale the supply voltage further down to near or below the V_{th} (thereafter this is referred to as the ultra low voltage regime) to maximize energy efficiency since CMOS gates are known to be functional in this regime [66]. Recently

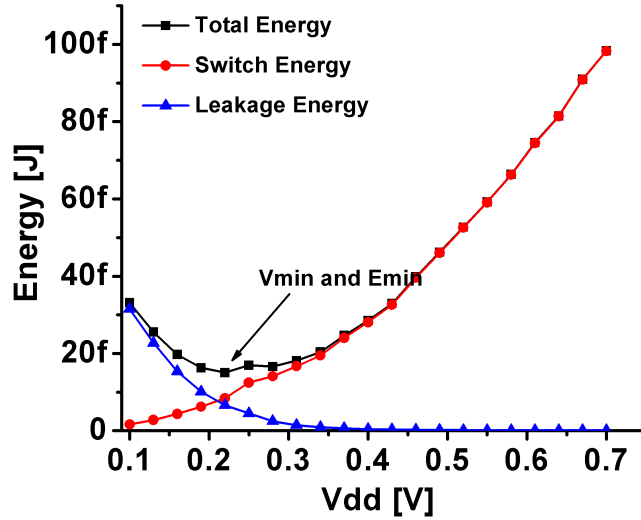


Figure 1.3: V_{min}/E_{min} curve

several researchers successfully demonstrate CMOS ICs operating at several hundreds of millivolts with several orders of magnitude better energy efficiency [102, 36, 41, 57].

One of the foundational concepts in ultra low voltage circuit design centers around a peak energy-efficient operating point. Zhai et al., [113] and Calhoun et al., [15] showed that energy-efficiency degrades when supply voltage is scaled too low since an increasingly slow circuit consumes substantial leakage energy, offsetting the quadratic savings in switching energy. Therefore, the total energy consumption begins to increase once the supply voltage scales down below a certain point, which we refer to as V_{min} . The optimal energy consumption, which occurs at V_{min} , is defined as E_{min} . This relationship is illustrated in Figure 1.3. The V_{min} often lies at 300-400mV for circuits in modern sub-micron CMOS technologies.

From these early studies on the energy optimal point, various aspects of design including technology, circuit, and architecture techniques have been actively researched for achieving better energy efficiency, robustness, and performance. At the technology level, several approaches for device design have been proposed [74, 35] since modern processes are sub-optimal for ultra low voltage operation as they are optimized for

high performance applications. Circuit-level studies include how to size MOSFETs for minimizing energy [17], reducing variability [56], and enhancing performance [48]. Additionally, Static Random-Access Memory (SRAM) has been intensively studied since its positive feedback structure suffers from degraded stability at the reduced on-current to off-current ratio found in the ultra low voltage regime. In particular, different topologies including a single-ended 6T [114], 8T [100], and 10T [16, 21] SRAM have been proposed to replace the conventional differential 6T structure. By employing design techniques for ultra low voltage operation, several computational cores [102, 41] and general microprocessors have been published that achieve performance of hundreds of kHz and energy consumption of several pJ per cycle [36, 57].

1.4 Challenges and Contributions

1.4.1 Minimizing Standby Power

It is exciting that a microprocessor consumes only several pico-joules per cycle or hundreds of nanowatt at ultra low voltage regimes, compared to 85pJ per cycle or tens of microwatt consumed by a state-of-art low power microprocessor [27]. Although the performance of hundreds of kHz is much slower, it is adequate for cubic millimeter sensing systems. However, simple voltage scaling is not sufficient to reduce the power consumption down to the level of hundreds of pico-watt that cubic millimeter sensing systems require as discussed in Section 1.2.

To meet the stringent power constraints, we find that it is critical to reduce standby power since cubic millimeter sensing systems spend most of their lifetime in waiting for periodic workloads to perform and hence the total energy consumption is dominated by standby mode. Our investigation with a 0.35V energy efficient microcontroller [36] indicates that it is necessary to reduce the standby power by 2-3 orders of magnitude for achieving an average power of several hundreds of picowatt. How-

ever, standby power has been largely overlooked in ultra low voltage design space despite of its importance. Therefore, we first focus on minimizing standby power consumption along with high energy efficiency in active mode.

We design a system platform called Phoenix Processor, in which we demonstrate the effectiveness of comprehensive standby power reduction techniques. These techniques include technology selection and power gating switch design methodology for ultra low voltage operations (Chapters III and IV, respectively). It also employs ultra low leakage SRAM design operating at ultra low voltage regimes (Chapter II), robust ultra low voltage ROM (Chapter V), and architecture optimizations with hardware compression support (Chapter II). All these techniques enables 35.4pW standby power and 226nW active power with approximately 100kHz performance at 0.5V. The average power consumption with a typical operation of 10^5 duty ratio is 37pW. The footprint of the system is less than $1mm^2$ in industrial $0.18\mu m$ CMOS technology, including 2kb SRAM, 640b ROM, 8b CPU, a watchdog timer, and a temperature sensor. The ultra low power consumption theoretically enables approximately 5 year lifetime with a single $1mm^2$ battery, which make it as a viable option for cubic millimeter sensor systems. This work is a result from multiple collaborators. I principally led the efforts on system-level design including technology selection, power gating switches, SRAM arrays, ROM arrays, and top-level integration.

1.4.2 Designing Ultra Low Power Analog Building Blocks

Along with digital components such as memory and CPU, it is also critical to design analog and mixed-signal modules with minimal power budgets since they often need to be integrated in cubic millimeter sensing systems for self-contained functionality. Such analog and mixed-signal functions include power conversion, analog-to-digital conversion, radio communication, sensing and time checking. It is an increasingly challenging task to minimize power consumption in those blocks since the

performance and robustness often degrade with reduced power budget more rapidly than digital components.

We focus on designing a ultra low power voltage reference (Chapter VI) as a key building block for such analog and mixed signal modules. The prototype design, named 2-Transistor voltage reference, shows $\sim 19ppm/^{\circ}C$ of temperature insensitivity, 0.033%/V of supply voltage insensitivity, -67dB of Power Supply Rejection Ratio (PSRR), and $\sim 2.2pW$ of power consumption at 0.5V. It has only two MOS-FETs and consume very small footprint of $1350\mu m^2$ in $0.13\mu m$ CMOS process. Power, area, line sensitivity, PSRR, and minimum functional voltage are dramatically improved and all the other metrics are compared favorable to the state-of-arts. We also investigate the effect of process variations on the performance of 2T voltage references and propose a digitally trimmable 2T voltage reference for mitigating process variations. We also demonstrate easy technology portability by implementing the same designs in three different CMOS technologies. Several variants of 2T voltage references are also proposed to tailor a specific needs such as different output voltages and temperature-dependent output voltages.

1.4.3 Improving Performance, Delay Variability and Energy Efficiency

Back to the digital design domain, there are still challenges to address before we widely use a ultra low voltage design as a design practice for cubic millimeter sensing systems. These challenges include improving degraded performance, mitigating heightened variability and improving energy efficiency beyond simple voltage scaling. The low performance is acceptable for control purposes; However, higher performance is often preferred, particularly for the systems with real-time process constraints [77]. Additionally, if designers see a ultra low voltage operation as a performance constrained design, delay variability must be adequately addressed. It is also important to improve energy efficiency beyond the limit of conventional voltage scaling as a

subset of cubic millimeter sensing systems often require a significantly large amount of computations like digital signal processing [101].

Therefore, we propose circuit and architecture techniques to mitigate those challenges and apply them to designing a Fast Fourier Transform (FFT) core (Chapters VII). Pipelining is a well-known method to improve performance or trade the gained performance for energy savings. However, less pipelining, 100-200 Fan-Out-of-4 (FO4) delays per stage, is often preferred in ultra low voltage designs for two benefits [103, 36, 83]. First, less pipelining has less energy overhead from sequential elements and clock distributions. Also, long paths per stage help to reduce delay variability from random process and environmental variations through averaging effects. Contrary to these common practices, we propose to use aggressive pipelining to improve energy efficiency and performance simultaneously. This can be achieved since shorter cycle time from pipelining can reduce leakage energy consumption from idling gates in circuits. Furthermore, since leakage energy consumption was stopping energy-saving voltage scaling, this reduction of leakage energy consumption extends useful voltage scaling, resulting in switching energy savings as well. This process is different from trading a gained performance for energy savings; it actually sets a new limit of better energy efficiency in ultra low voltage regimes. We utilize this super-pipelining for multipliers, serving delay-critical paths of the FFT core. It saves 18% energy consumption and improves performance by $3.6\times$ at the same V_{DD} , compared to a non-pipelined design. Architecture modifications are also proposed for reducing idling parts in the FFT core and increasing throughput, resulting in $2.86\times$ better energy efficiency and $6.2\times$ better throughput.

For mitigating delay variability, we use a latch-based design for delay-critical blocks (e.g. multipliers). It removes the hard boundaries of flip-flops and thus re-establish long paths through cycle borrowing capability, which average out process variations along the paths. Clock network design is also critical for circuit variability

since skew and slew variability can cause hold time violations. However, clock buffers is less effective to reduce the interconnect delay due to the negligible metal resistance in ultra low voltage regimes. Also buffers contribute a significant amount of skew and slew variability from Process, Voltage and Temperature (PVT) variations. Therefore, we propose a clock network design using a greatly reduced number of clock buffers and matched interconnects such as H-Tree. This method reduces skew and slew variability by several orders of magnitude with no energy overhead. More general framework for designing clock network in ultra low voltage operation is also discussed (Chapter VIII).

The FFT core is demonstrated in 65nm CMOS technology, consuming the lowest energy consumption of 17.7nJ per 1024-pt complex FFT while operating at the remarkable performance of 30MHz at 0.27V. The energy efficiency is improved by at least $4\times$ from state-of-arts. Also the performance is improved by several orders of magnitude, compared to tens to hundreds of kHz of typical ultra low voltage designs. This project is a collaboration work; I am directly responsible for super-pipelining technique, latch-based design, clock network design, and top-level integration.

1.4.4 Contribution Summary

Along with mitigating variability and improving performance, the primary contribution of this work is to reduce energy consumption below the level that cubic millimeter systems require. Figure 1.4 shows the building blocks that directly benefits by this work: microcontroller, memory (SRAM and ROM), reference circuits, and Digital Signal Processing (DSP) unit. Reference circuits can be a key building block for several analog and mixed-signal modules such as timer, radio, sensors, and power conversion. The energy savings for each component are summarized in Figure 1.5. The large energy savings confirm the viability to build a cubic millimeter sensing system with a power source of the same scale while ensuring multi-year lifetime.

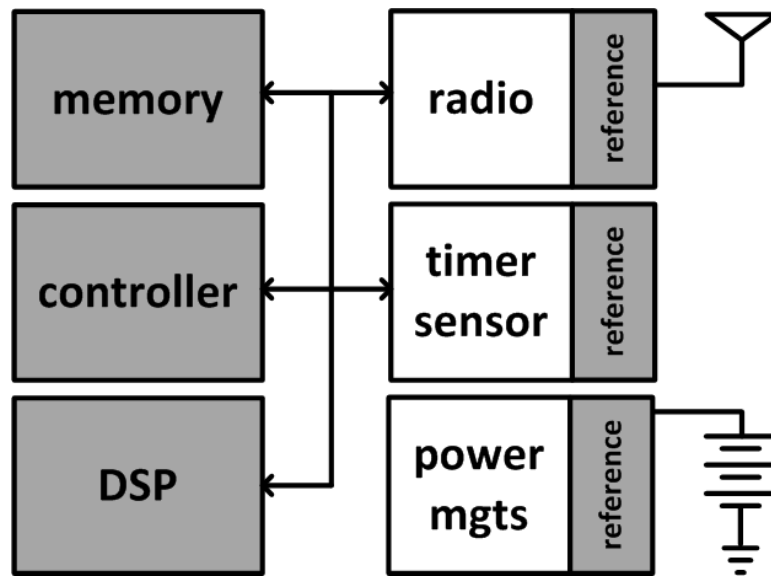


Figure 1.4: Components that can directly benefit from the design methodologies of this work in cubic millimeter sensing systems (in gray)

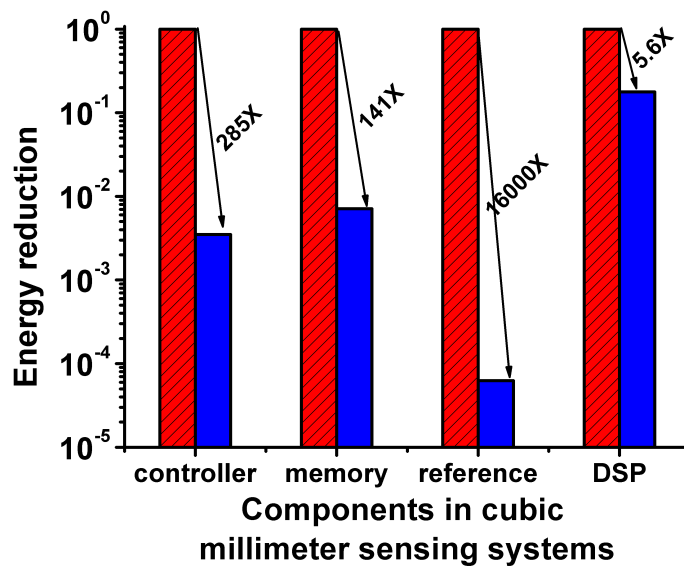


Figure 1.5: Energy efficiency improvement in the blocks of cubic millimeter sensing systems

1.5 Organization of this Work

The remainder of this thesis proposal is organized as follows. Chapter II focuses to present the the sensing platform, called Phoenix Processor, with ultra low power consumption both in standby and active mode. We emphasize one of the building block, ultra-low power custom SRAM array in this chapter. A brief discussion on micro-architecture design of Phoenix Processor is also presented. Chapter III, Chapter IV, and Chapter V in detail discuss three different standby strategies, namely technology selection, power gating switch design, and robust ROM array design, all employed in Phoenix Processor. After the discussions on Phoenix Processor, Chapter VI introduces a ultra low power voltage reference design as a key analog and mixed signal building block. We first focus on the basic topology of the proposed design. Then we discuss the impact of process variations, trimmable version for mitigating variability, technology portability and several variants of the proposed voltage reference. Finally, in Chapter VII, we introduce circuit and architecture technique to overcome other remaining issues of ultra low voltage designs, namely degraded performance, saturated energy efficiency, and heightened delay variability. A FFT core in 65nm is presented as an test vehicle to demonstrate the techniques. In Chapter VIII, we in detail discuss about robust clock network design, used in the FFT core design. Finally, Chapter IX concludes this work.

CHAPTER II

Phoenix Processor: 35pW Standby and 226nW Active Power Sensor Platform

2.1 Motivation and Previous Work

The prevalence of mobile computing has helped define a vision of complex computational resources in cubic centimeter scale [104, 72] and smaller scale. As the volumes of computing resources approach one cubic millimeter, active monitoring and actuation can be used to enrich a wide range of applications. Cubic millimeter computing will be particularly important in implantable medical devices, where reducing device volume helps minimize implant damage to the body. The diagnosis and treatment of Glaucoma, for example, requires periodic measurements of pressure in the eye (intra-ocular pressure). Intra-ocular pressure is currently monitored directly by a doctor, requiring frequent trips to the doctor's office to ensure sufficient temporal resolution [82]. An intra-ocular pressure sensor with a MEMS pressure sensor, microprocessor, memory, radio and power source small enough to be implanted in the eye would reduce both cost and time investment and would increase the temporal resolution of pressure measurements.

Although MEMS and circuit components easily meet the volume constraints of intra-ocular pressure sensing and other cubic millimeter computing applications, bat-

teries and energy scavenging power sources cannot be easily miniaturized while also serving the power demands of the MEMS and circuit components. Minimizing the power demands of each component is therefore one of the central challenges in designing a cubic millimeter computing system. Consider a system with a thin film zinc/silver oxide battery with a capacity of $100\mu Ah/cm^2$ and output voltage of 1.55V [63]. If the battery size is restricted to $1mm^2$, the average system current must be only 114pA (for power consumption of 177pW) to guarantee one year of battery life.

Early work proved that operation at extremely low voltage was possible [102] and later work demonstrated full subthreshold microprocessors operating at record low energy consumption [37, 116, 57]. Recent research has also shown that memory can be redesigned to operate robustly at low voltage [20, 16, 114]. However, the 177pW power budget is still several orders of magnitude less than most energy efficient microprocessor designs consume [36, 57, 101]. Therefore it is paramount to reduce the power of such systems.

In particular standby power consumption need to be minimized since a typical wireless sensing system may spend the majority of its lifetime in standby mode. For example, a typical wireless sensor data logger might take sensor measurements once every 10 minutes. Assuming a 100ms active period, the sensor spends $6000\times$ more time in standby mode than active mode. This disparity is of particular importance in low voltage systems, where leakage energy is comparable to dynamic energy in magnitude [113]. For example, the low voltage processor in [37] consumes only $5\times$ more power in active mode than it does in a clock gated standby mode. Total energy consumed in standby mode would exceed total energy consumed in active mode by $1200\times$ if this low voltage processor were used in the aforementioned wireless sensing system.

2.2 Contribution

Given the importance of standby mode power, we focus in this work on the development of an ultra low energy sensor processor designed for deep standby operation, called the Phoenix Processor [83]. The Phoenix Processor is a complete digital system for cubic millimeter computing that includes an 8-bit CPU, data and instruction memories, a watchdog timer, and a simple temperature sensor. In addition to aggressive voltage scaling, the Phoenix Processor leverages a comprehensive sleep strategy including system-level optimization for low voltage operation, a unique power gating approach, a ultra low leakage custom SRAM, a robust low V_{DD} ROM, an 8bit CPU with compact instruction set, and data memory compression.

My main contribution in this project includes the system-level optimization including technology selection, a unique power gating approach, a ultra low leakage custom SRAM and a robust low V_{DD} ROM along with a system integration. The 8 bit CPU design with optimized Instruction Set Architecture (ISA) and hardware compression support, the watchdog timer and the temperature sensor are primarily designed by my co-workers. Measurements of a $0.18\mu m$ test chip reveal that Phoenix consumes 226nW in active mode and only 35.4pW in sleep mode. Assuming 10^4 10^5 duty ratio, the average power is 58-37pW, which is below the required power budget for cubic millimeter scale sensing system.

2.3 System Overview

As shown in Figure 2.1, the Phoenix Processor is a modular system with a core unit consisting of an 8-bit CPU, a 52x40-bit data RAM (DMEM), a 64x10-bit instruction RAM (IMEM), a 64x10-bit Instruction ROM (IROM) and a Power Management Unit (PMU). The core serves as a parent to peripheral devices, including a watchdog timer and a temperature sensor. The core and peripheral devices communicate over a

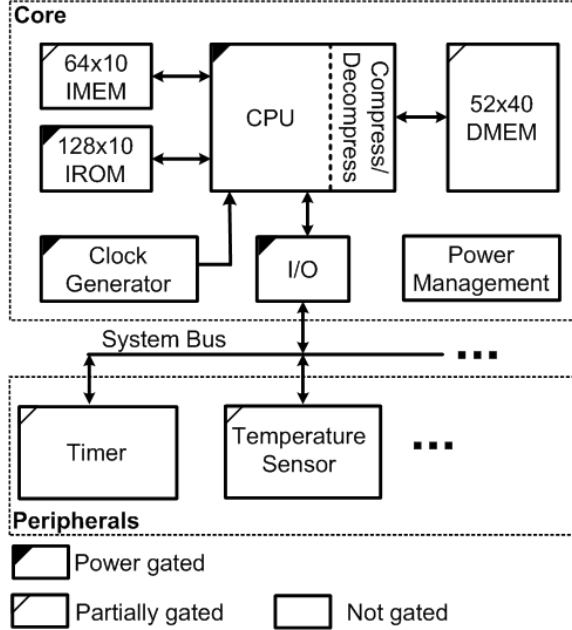


Figure 2.1: The Phoenix Processor.

system bus using a simple asynchronous protocol. The Input and Output (I/O) controller addresses up to 8 peripherals on the system bus for sensing systems requiring additional peripherals.

In typical operating conditions, the Phoenix Processor spends an extended period of time in standby mode (e.g., 10 minutes) and wakes up in response to an exception raised by the watchdog timer (a 0.9pW current-starved oscillator). Once awake, the Phoenix Processor polls the temperature sensor and runs a short routine to process and store the measurement. After completing the data processing routine, Phoenix returns to standby mode.

The power consumption in active mode is dominated by components with high switching activity, such as the CPU. To minimize this source of power consumption we scale voltage aggressively to 0.5V, a sub-threshold voltage (for high- V_{th} devices) or near-threshold voltage (for medium- V_{th} devices) in the target technology. The challenges of low voltage digital design have been covered extensively in recent literature [37, 116, 102] and will not be the focus of this work.

Instead, we place emphasis on accommodations made for standby mode operation. The Phoenix Processor was designed at the device, circuit and architecture levels with the primary goal of standby power minimization. In subsequent sections, we discuss each of the key components of this comprehensive standby mode strategy. We begin in Section 2.4 by discussing the design of "retentive gates," which are the gates that remain awake during standby mode. These include the PMU, timer, and portions of IMEM and DMEM. Without proper attention, retentive gates dominate standby mode power consumption. In Section 2.5 we discuss a unique power gating approach for power minimization in non-retentive gates (i.e., gates that sleep during standby mode). During standby mode, between 65% and 87% of all transistors are power gated, so efficient power gating is critical. The CPU architecture also plays an important role in determining standby mode power. We discuss the design of the CPU in Section 2.6 and pay particular attention to the ISA definition since this determines the footprint of IMEM, a significant source of leakage during standby mode. The CPU includes software and hardware support for DMEM compression, which is discussed in Section 2.7. As we will show using device measurements, the power consumption of retentive memory cells dominates the standby mode power consumption, so we devote Section 2.8 to discussing a custom SRAM cell that achieves ultra low leakage at a low operating voltage.

2.4 System-level Optimization including Technology Selection

Perhaps most important to minimizing power consumption in standby mode is technology selection. Despite its importance to both power and performance, there has been little investigation of technology selection for low voltage circuits. The requirements of sub-threshold and near-threshold circuits are different than those of

normal super-threshold circuits, and the optimal technology is therefore different. The required performance is much relaxed in typical low voltage sensing applications. Consequently, older technologies can easily meet performance requirements. Additionally, leakage becomes more important in low voltage operation than in super-threshold operation, making older technologies with high- V_{th} devices attractive. Near the energy-optimal supply voltage [113], leakage energy in active mode is comparable to switching energy. Furthermore, the long standby time observed in many sensor applications makes cumulative standby leakage energy significant as we observed earlier in this work. Advanced technology nodes have also been optimized exclusively for super-threshold operation. Consequently, inverse subthreshold swing degradation in scaled super-threshold devices leads to a reduction in noise margins and sub-optimal power and performance, as shown in [36]. The ideal technology would simultaneously offer small feature sizes and devices with ultra low leakage. Since no such technology was available for use in academic research, we investigated standard CMOS technologies from $0.25\mu m$ to 65nm using the method in [84] to determine the energy-optimal technology.

The optimization has been performed to minimize total energy consumption. We identify 5 high-level factors that affect the total energy of the system: technology, the size of power gating switch, supply voltage, duty cycle (defined as active time to total time ratio), and the ratio of memory to logic area. The memory to logic area ratio and duty cycle are determined primarily by application requirements while technology, the size of power gating switch, and supply voltage are chosen to minimize total energy consumption.

In order to find the ratio of memory to logic area, we need to estimate DMEM and IMEM size for applications. We examine periodic sensor data logging as a typical operation. We choose 512 words of DMEM since it will take approximately one year to fill the 512 words if one word of data is stored on daily basis. The IMEM size

is set relatively smaller at 64B since it is supplemented with 128B IROM. Adding more IROM can be done with negligible increase in standby power. However the size of DMEM and IMEM needs to be changed for different applications. Based on the DMEM and IMEM size, we estimate the area ratio of memory to logic as 1, which is backed by the actual die photo in Figure 2.13. Duty cycle is assumed to be 0.001.

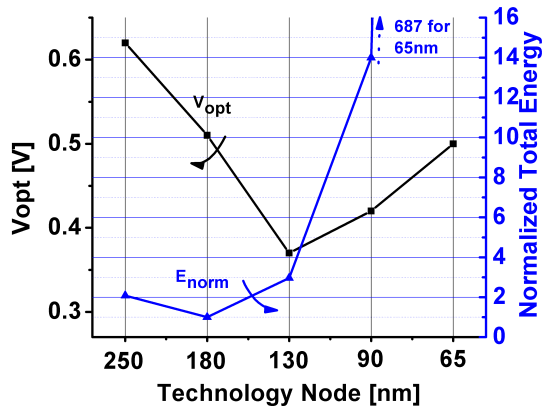
With the chosen area ratio of memory to logic and the duty cycle, we optimize the total energy consumption by sweeping technology, the size of power gating switch, and supply voltage based on the optimization framework in Chapter III. Five industrial CMOS technologies from 65nm to $0.25\mu m$ are considered. Supply voltage and the size of the power gating switch are selected within the range allowed by each technology. The Phoenix Processor is abstracted as a large collection of inverter chains with different activity ratios. The activity ratio for SRAM modules (DMEM and IMEM) is assumed to be 0.02 while that for CMOS logic (CPU, IROM) is set $10\times$ larger at 0.2. We include high- V_{th} devices in the optimization process since they can be used in the SRAM bitcell for leakage reduction. High- V_{th} I/O devices are also considered as a viable option for building bitcells.

The result of the optimization is shown in Figure 2.2. The matrix shows the energy optimal technology at the given performance and duty cycle requirement of each application. With the target duty cycle of 0.001, we have three different optimal technologies for different performance requirements. Since the Phoenix Processor has a relaxed performance requirement and higher performance causes more energy consumption, $0.18\mu m$ is selected as the energy optimal technology.

For the highlighted block whose duty cycle is 0.001 and performance is 5MHz (frequency of 40 FO4 delay), the optimal supply voltage and energy is plotted in the Figure 2.2. Old technologies are favored since they have lower leakage energy. The reason why $0.18\mu m$ technology gives lower energy consumption than $0.25\mu m$ is that the particular $0.18\mu m$ that we investigated has a higher- V_{th} I/O device than

Duty Cycle	0.001	180	180	180	130	90
	0.01	180	180	130	90	65
	0.1	180	180	130	90	65
	1	180	130	90	90	65
		50K	500K	5M	50M	500M
		40F04 Freq [Hz]				

(a)



(b)

Figure 2.2: (a)Energy optimal technology matrix (b)Optimal V_{DD} and energy over technologies

the $0.25\mu\text{m}$ technology. The $0.18\mu\text{m}$ technology includes a thin-oxide medium- V_{th} device with $V_{th} = 0.5\text{V}$ and a thick-oxide I/O device with $V_{th} = 0.7\text{V}$. All retentive gates are implemented using the high- V_{th} devices, which consume several orders of magnitude less leakage power per unit of gate width than the medium- V_{th} devices. If the $0.25\mu\text{m}$ technology offered a comparable high- V_{th} device, it could be the optimum choice unless the area overhead associated is intolerable. Note that we do not use high- V_{th} devices in non-retentive gates since the minimum dimension is larger than that of the thin-oxide device, which gives both area and active energy penalties. The optimum supply voltage estimated is 0.5V . The size of the power gating switch is determined proportionally to the relative size of actual modules to the inverter chains used in the optimization process.

In addition to the selection of an older technology, stack-forcing is used to reduce leakage power further. Leakage reduction due to the stack effect has been shown in previous work to be effective [68]. In our selected technology, stacking two transistors gives $2\times$ leakage reduction.

2.5 Power Gating Under Relaxed Performance Constraints

A power gating switch, as shown again in Figure 2.3, is often used in low power circuits to minimize leakage in non-retentive circuit blocks during standby modes. At normal super-threshold operating voltages (e.g., $> 1\text{V}$ in 45nm and 65nm designs), a high- V_{th} device is typically used as a power gating switch since it delivers comparable on-current to the nominal device with exponentially smaller off-current. Additionally, wide power gating switches are typically used to minimize the performance penalty of power gating.

For cubic millimeter computing applications with modest performance requirements, minimizing standby power is the most important goal. In such applications, performance can be sacrificed for lower leakage, which is in stark contrast to the typ-

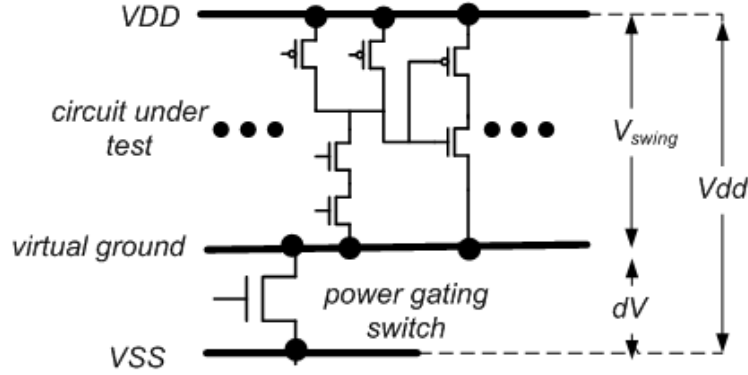


Figure 2.3: A typical power gating switch.

ical approach to power gating. In the Phoenix Processor, we leverage these modest performance requirements with a unique power gating approach.

Our power gating approach relies on a medium- V_{th} power switch rather than a high- V_{th} switch as in the typical approach. Since the on-current of the high- V_{th} device is exponentially smaller than that of the medium- V_{th} device at low voltage, a high- V_{th} device must be sized up $1000\times$ as compared to a medium- V_{th} device to meet the current demands of the primary circuit, which is implemented using medium- V_{th} devices. The area overhead as well as the power overhead of charging/discharging such a large switch is avoided by using a medium- V_{th} power switch.

In addition to using medium- V_{th} power switches, the strength of our power gating switch compared to the circuit under test is smaller than that of the typical power gating approach. A stronger power gating switch minimizes the performance penalty of power gating at the expense of additional leakage during standby mode. Given the modest performance demands for the Phoenix Processor, we choose to reduce standby mode leakage considerably by selecting a very weak power gating switch [84].

In the Phoenix Processor, the medium- V_{th} power switch is only $0.66\mu m$, which is 0.01% of total effective Negative Channel Field Effect Transistor (NFET) width and $3\times$ larger than the minimum width in the target technology. We increase the length from $0.18\mu m$ to $0.50\mu m$ to improve inverse subthreshold slope and consequently

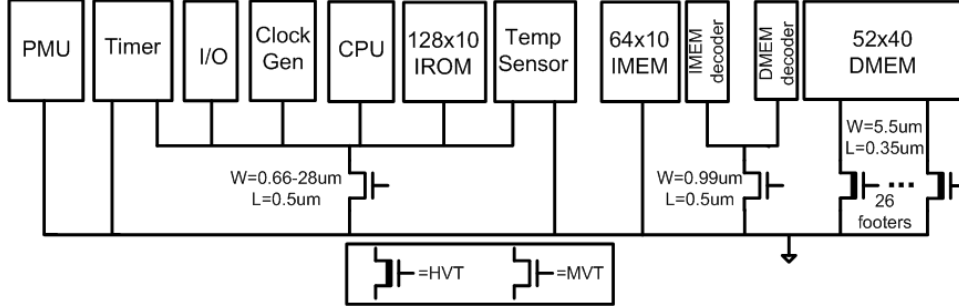


Figure 2.4: Footer allocation in the Phoenix Processor.

increase the on-current to off-current ratio. The $0.66\mu m$ power gating switch is connected to the CPU and several other logic blocks as shown in Figure 2.4. Simulations with a model of the CPU indicate that the virtual ground rail bounces by a maximum of 100mV, which is sufficient to guarantee correct logic operation. The non-retentive parts of IMEM and DMEM, such as decoders and output buffers, are connected to a separate power gating switch since the robustness of low voltage memory may be compromised by a voltage drop across the power gating switch. The measured energy and performance implications of our proposed power gating strategy will be discussed in Section 2.10.

2.6 CPU and Instruction Set Design for Standby Mode

In accordance with the conclusions of previous studies of subthreshold processor architectures [69], we have selected a simple CPU architecture with 2-stage pipeline, 8-bit data width, and 10-bit instruction width to reduce active mode power and standby mode power. The instruction set includes support for basic arithmetic computation in typical sensor logging applications. As shown in Figure 2.5, the first pipeline stage consists of instruction fetch and decode as well as a scratch memory with an 8-entry register file and 16-entry cache. The second pipeline stage includes a simple ALU, write-back logic, and a memory interface unit that compresses (decompresses) outgoing (incoming) memory traffic. The ALU includes hardware for

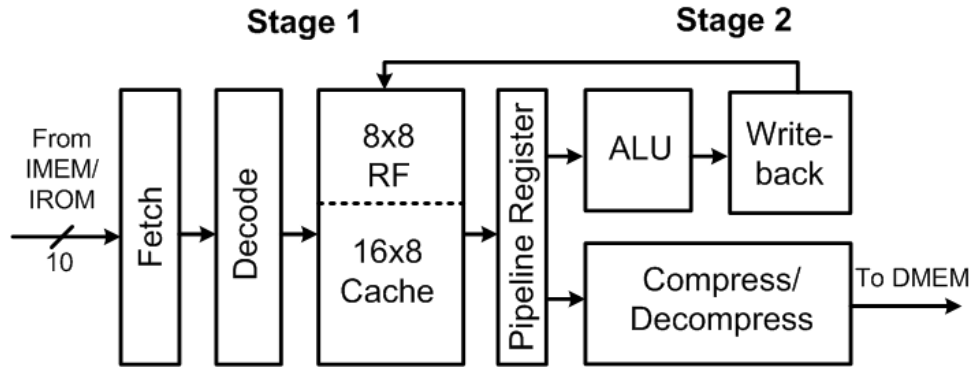


Figure 2.5: CPU diagram.

addition, subtraction, and shifting. The CPU has been designed to minimize energy in both active and standby modes, as shown in the remainder of this section.

Since the computational demands of cubic millimeter computing applications are typically modest, the CPU was simplified to support a minimum set of operations. Such simplicity reduces decode complexity and eliminates unnecessary switching activity, thus reducing active mode power. Furthermore, elimination of complex operations like multiplication eliminates large, leaky circuit blocks. Since leakage energy can be >30% of total energy in active mode for low voltage circuits [113], the resulting active mode power savings are significant.

ISA optimization also plays an important role in minimizing power consumption in standby mode. Since the contents of IMEM must be retained in standby mode, it is important to minimize the instruction width. The leakage penalty of instruction memory can alternatively be eliminated by using non-volatile memory, but this requires costly processing steps. The custom ISA for the Phoenix Processor was compressed to an instruction width of only 10 bits by selecting a minimum set of 18 instructions.

Table 2.1: Instruction set architecture overview.

Class	Members	Addressing Mode
Arithmetic	ADD, ADDI, SUB, MOVE, SHR	explicit
Flow Control	BEQZ, JUMPI, JUMPR	explicit
Compression	COMP, DECOMP	implicit
Compression	FREE	explicit
Load, Store	LOAD, STORE, STORE_OVER	implicit
Wake	GET_REQ, SEND_REQ, SEND_ACK	implicit
Sleep	HALT	

2.7 DMEM Compression for Standby Mode

While efficient instruction encoding helps minimize the footprint of IMEM, we use data compression to help minimize the footprint of DMEM. Along with fine-grained power gating in DMEM (to be discussed in Section 2.8, compression permits fewer DMEM entries to be retained and enables significant power reductions in standby mode. Compression of instruction and data memories has been explored previously [97, 59]. The IBM Memory Expansion Technology, for example, uses compression to more than double the size of main memory [97] but requires a complex memory management protocol targeted at server systems. To ensure that the energy overheads of compression do not surpass the standby mode reductions of a compressed DMEM, we adopt a simple compression architecture in the Phoenix Processor.

During compression, words from the 16-entry cache are sequentially converted to compressed words using a compression lookup table. The 512-byte virtual memory is divided into 16-byte blocks, and an entire 16-byte block from the cache must be compressed before being sent to the 266-byte physical memory.

The primary function of the Phoenix Processor is sensor data logging, which has two important consequences for compression. The first consequence is that access to memory is largely sequential (since temporally adjacent measurements are stored in spatially adjacent memory locations), thus limiting the compression/decompression overheads associated with random hopping among 16-byte blocks. The other im-

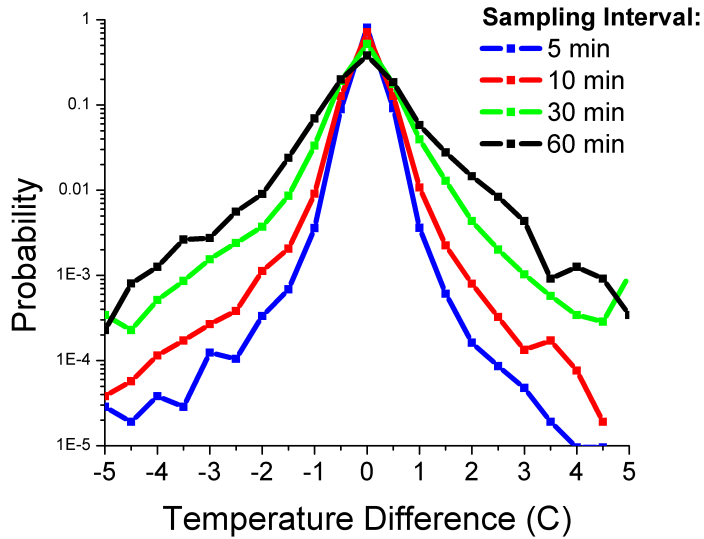


Figure 2.6: Distribution of temperature in Muskegon, MI in 2006 represented as the difference between temporally adjacent measurements [1]

portant consequence concerns compression dictionary selection. Typical sensor data is predictably compressible since two temporally adjacent points are likely to differ by only a small amount. The measurement for a particular time can be stored as the difference between the current and previous measurements. The resulting data distribution is tightly distributed around zero, making dictionary selection simpler. Since the Phoenix Processor includes an on-board temperature sensor, we consider a collection of ambient temperature measurements in Michigan as an example. Figure 2.6 shows the differences between temporally adjacent temperature measurements for a full year at different sampling intervals. In this difference format, 96% of the data falls in the range $-1^{\circ}C$ to $1^{\circ}C$ assuming a 10 minute sampling interval. We take advantage of this small range by using a fixed compression dictionary that uses short words to represent values in this range and longer words to represent the rare value outside of this range.

We use Huffman encoding to generate a lookup table-based dictionary using temperature measurements from [1] assuming a temperature precision of $1^{\circ}C$ and a sam-

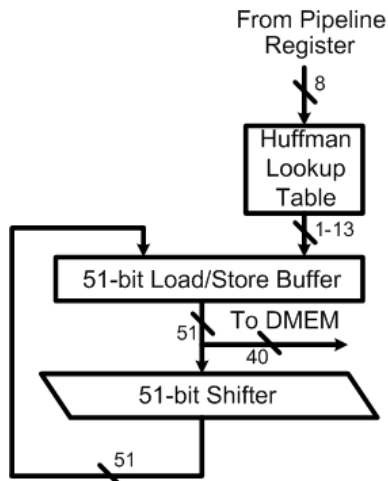


Figure 2.7: Hardware support for compression.

pling interval of 30 minutes (which was empirically determined to efficiently compress data sampled at intervals ranging from 5-60 minutes). The lookup table converts 8-bit uncompressed data words to compressed words with lengths between 1 and 13 bits. By using a fixed dictionary, the compression operation is simplified significantly, minimizing the active energy penalty. While the footprint of compressed data can grow by up to 60% if measured data is not sufficiently similar to the distribution in [1], an editable fixed dictionary could potentially be stored in DMEM to better match the needs of a specific application.

After using the Huffman lookup table to compress an 8-bit data word, the compressed word is shifted by a 51-bit shifter and then stored in a 51-bit load/store buffer (Figure 2.7). Once all entries in a 16-byte block have been loaded in the load/store buffer or the buffer is full, the compressed data is sent to DMEM using one of the 3 load/store instructions supported by the ISA.

Memory allocation is the primary challenge in implementing compression. Fixed length uncompressed blocks from virtual memory are translated to variable length compressed blocks in physical memory, and efficient placement of the variable length blocks within physical memory can be difficult. To address this problem, we divide DMEM into the two partitions shown in Figure 2.8: a statically allocated partition

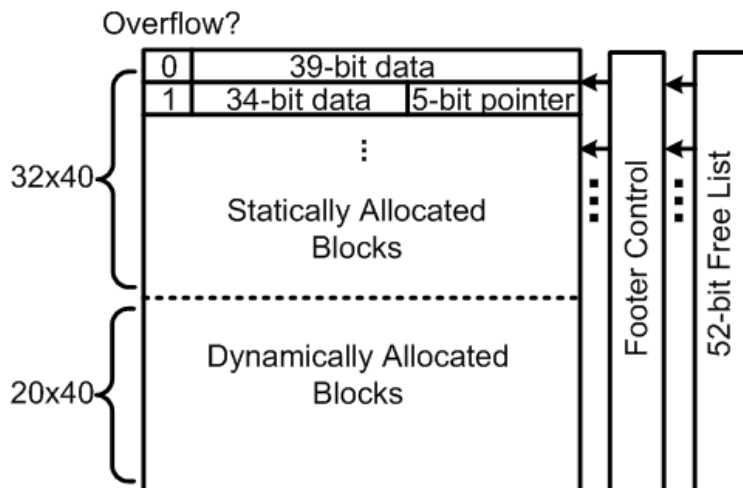


Figure 2.8: Memory support for compression.

and a dynamically allocated partition. Each 16-byte block in virtual memory is assigned a 40-bit entry in statically allocated memory. Data is normally stored in the statically allocated partition. However, if a 16-byte block does not fit within its statically allocated entry after compression, the overflow data is stored to an entry in dynamically allocated memory and a 5-bit pointer to the overflow data is stored in the statically allocated entry. A free-list is required to monitor which entries in dynamically allocated memory are available for storage. A priority encoder in the free-list returns the address of the first available entry in the event of an overflow. For compression purposes, the free-list need only monitor the dynamically allocated partition, but we monitor both memory partitions to permit fine-grained power gating (to be discussed in Section 2.8). Including the overhead of the free-list, the Phoenix Processor compression scheme represents 16-byte blocks with a minimum of 41 bits (a compression ratio of 32%). The effectiveness of the proposed compression scheme will be quantified using test-chip measurements in Section 2.10.

2.8 Ultra-Low Standby Power Memory Design

The power consumed by IMEM and DMEM dominates standby mode power since data must be retained in standby mode. In contrast, the CPU and other non-retentive logic can be fully power gated. Minimizing standby power in the IMEM and DMEM is therefore a critical design requirement for the Phoenix Processor. The memories must also be designed for robust operation at low supply voltage to avoid the overhead of a dual supply voltage system.

In the Phoenix Processor, the instruction memory is accessed every cycle by the CPU but does not need to be modified at runtime. Consequently, instruction memory is composed of a 64x10b SRAM (IMEM) and a 64x10b ROM. Commonly used procedures are stored in IROM while application-specific instructions are stored in IMEM. It is advantageous to put as many instructions in IROM as possible since ROM can be power gated during standby mode. In this work, we use the robust full static CMOS ROM implementation [83] described Chapter V in detail.

To minimize the standby leakage in retentive cells in IMEM and DMEM, we use the custom ultra low standby power SRAM cell shown in Figure 2.9. The bitcell transistors (cross-coupled inverters and access transistors) use the high- V_{th} I/O devices offered by the selected $0.18\mu m$ technology. Although the minimum dimensions of the I/O device are larger than those of the thin oxide device, the large leakage reduction justifies the use of the device. We further reduce leakage in the bitcell using stack forcing in the cross-coupled inverters as in other retentive gates. We use a stack height of two because the sensitivity of leakage to stack height becomes linear for larger stacks, as shown in Figure 2.10(a). Instead of further stack forcing, we find that increasing the length of the devices in the cross-coupled inverters gives a more area-efficient reduction in leakage. By increasing the length of the transistors from $0.35m$ to $0.50/mum$, the leakage is reduced by $2\times$, as shown in Figure 2.10(b). Stack forcing and gate length biasing are not applied to the access transistors to avoid

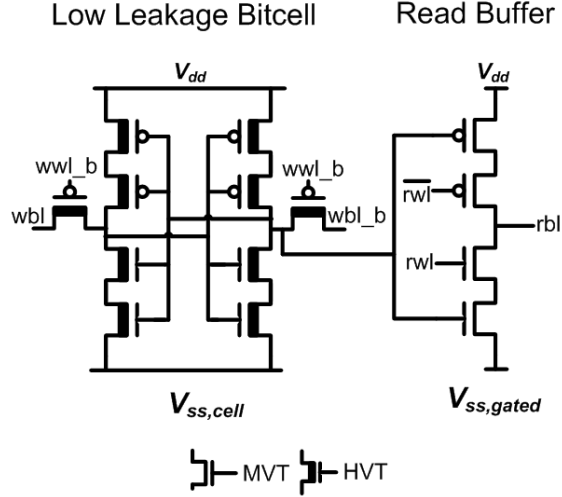


Figure 2.9: Proposed ultra low standby power SRAM cell.

upsetting write margins.

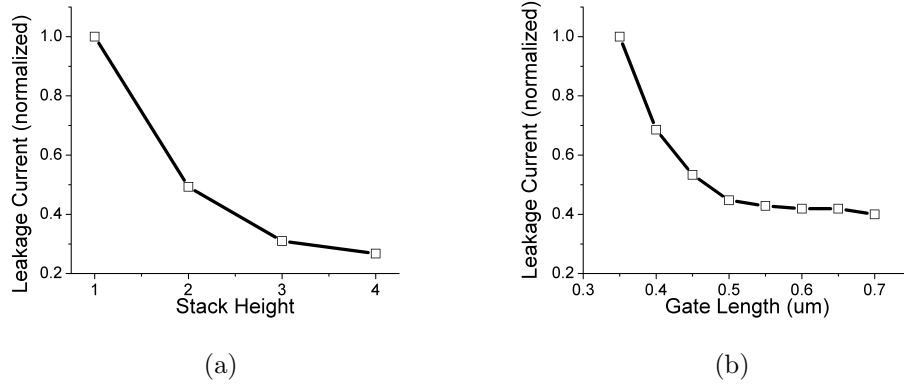


Figure 2.10: Effectiveness of (a) stack forcing and (b) gate length biasing for leakage reduction

To enable robust low voltage operation, the proposed SRAM cell in Figure 2.9 includes a full swing 4-transistor read buffer. Read buffers have been previously proposed to decouple read and write margins in low voltage SRAM cells [20, 16]. The full swing read buffer drives the bitline to both supply rails, ensuring a robust read. Since the read buffer can be power gated in standby mode without upsetting the bitcell data, it is implemented using medium- V_{th} devices for a negligible leakage penalty. The use of medium- V_{th} devices ensures a fast read time comparable to paths

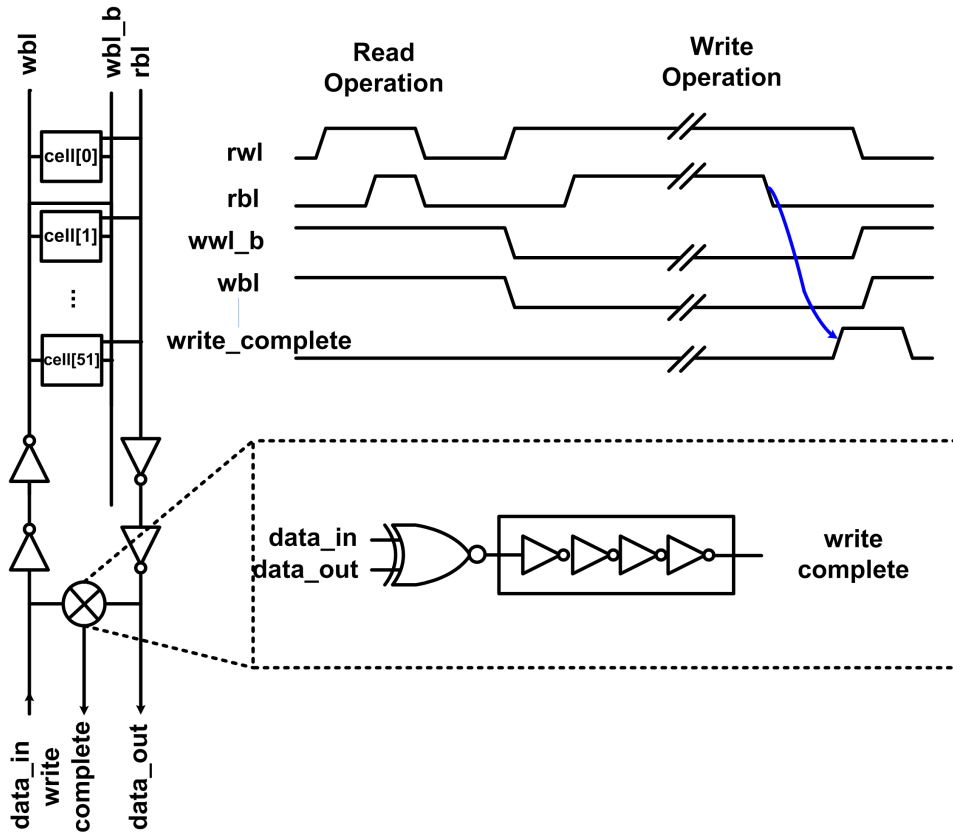


Figure 2.11: Memory column diagram showing completion detection.

in the CPU (implemented with medium- V_{th} devices). This is particularly important for the IMEM, which is accessed every cycle and lies on critical timing paths.

While the read delay is comparable to the delay of the CPU, the write operation through the high- V_{th} devices is slow. We therefore adopt an asynchronous write strategy in which the CPU stalls for 2-3 cycles during the write operation. The DMEM asserts a completion signal to alert the CPU when the write operation is finished. As shown in Figure 2.11, the write completion signal is generated by reading the contents of the row being written and comparing to the write data. Since read is single-ended, a replica delays the write completion signal to guarantee that both sides of the cell have been written correctly.

To permit further leakage reduction within DMEM, power gating switches are used to eliminate the standby mode power consumption in non-retentive entries. A

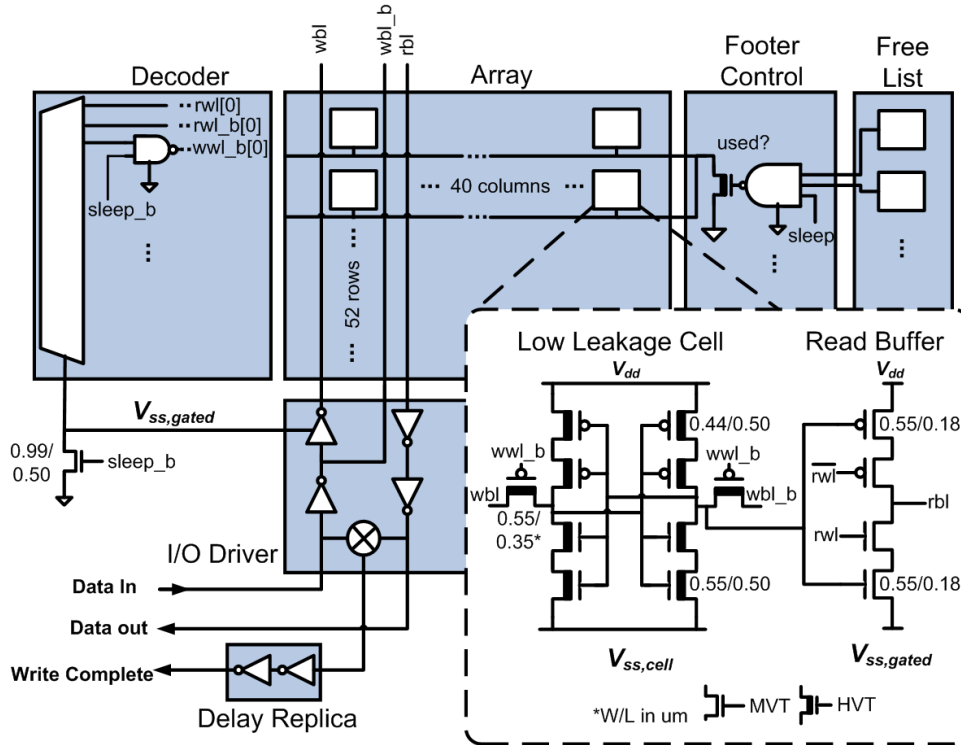


Figure 2.12: SRAM array architecture for DMEM

particular DMEM entry is power gated only if the free-list (described in Section 2.7) indicates that the entry is unused. Power gating granularity plays an important role in determining total power. Using a single power gating switch for each row allows the memory to grow to precisely match the footprint of the data. However, the width of a power gating switch is determined by the maximum current needed to read/write a single row, so the width of a single power switch changes minimally with power gating granularity. With one power switch allotted per DMEM row, the total leakage power is sub-optimal because the total power gate width is large. For example, the use of 52 switches for the 52 rows of DMEM would require a total power switch width approximately 52 times wider than the case where one footer is used for the entire DMEM. Higher footer granularity also leads to higher complexity in free-list management and a commensurate increase in standby power. We find that minimum standby power is achieved in the Phoenix Processor when two DMEM rows are grouped with a single footer. The entire SRAM array architecture is shown in

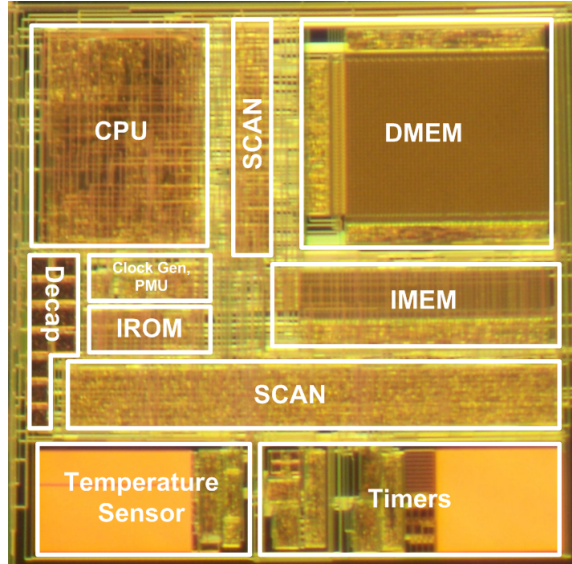


Figure 2.13: Phoenix Processor die photo.

Figure

2.9 Test Chip Overview

To demonstrate our standby mode strategy, we fabricated the Phoenix Processor in a $0.18\mu m$ process (die photo shown in Figure 2.13). The processor includes 60,332 medium- V_{th} devices and 32,167 high- V_{th} devices in an area of $915 \times 915 \mu m^2$. The memory, temperature sensor, and timer blocks were designed using a standard full-custom flow. The custom bitcell proposed in Section 2.8 was implemented in an area of $40\mu m^2$. The CPU and interfaces to memory, temperature sensor, and timer blocks were implemented using both synthesized and semi-custom blocks using a standard tool flow and a library limited to minimum-sized gates with maximum fan-in of three. As in previous low voltage processors, we routed signal, clock, and power wires using minimum width interconnect to reduce switching energy and improve routing density [37].

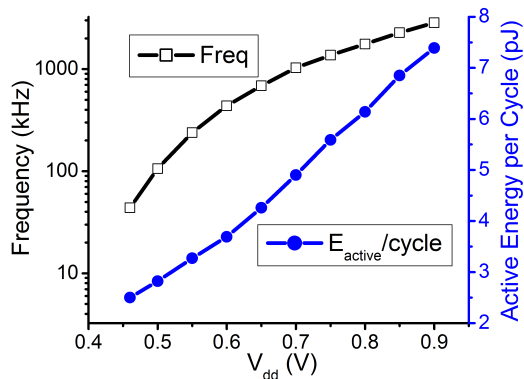


Figure 2.14: Measured frequency and energy consumption.

2.10 Measured Results

2.10.1 Power and Performance Results

The measured frequency and energy consumed per clock cycle are shown in Figure 2.14 as functions of V_{DD} for one test application. The frequency is determined by sweeping the clock frequency, running the test application, and noting the frequency above which the contents of memory are corrupted. The test application runs a short iterative sequence that writes a known list of numbers to DMEM, in the process exercising all timing critical instructions. Power is measured during execution using a high precision ammeter. At the target voltage of 0.5V, the die highlighted in Figure 2.14 operates at 106kHz with only 2.8pJ consumed per cycle, which corresponds to only 297nW. Figure 2.15(a) and Figure 2.15(b) show distributions of maximum operating frequency (mean of 121kHz) and active power at 60kHz (mean of 226nW) for 13 dies at $V_{DD}=0.5V$. The consequences of variability are particularly important at low operating voltages and have been covered extensively in previous work [37, 57].

Figure 2.15(c) shows that the mean standby mode power consumption for the same 13 dies at $V_{DD}=0.5V$ is 35.4pW with 50% of DMEM entries retained. The IMEM and DMEM consume 89% of standby power while the power gated CPU consumes

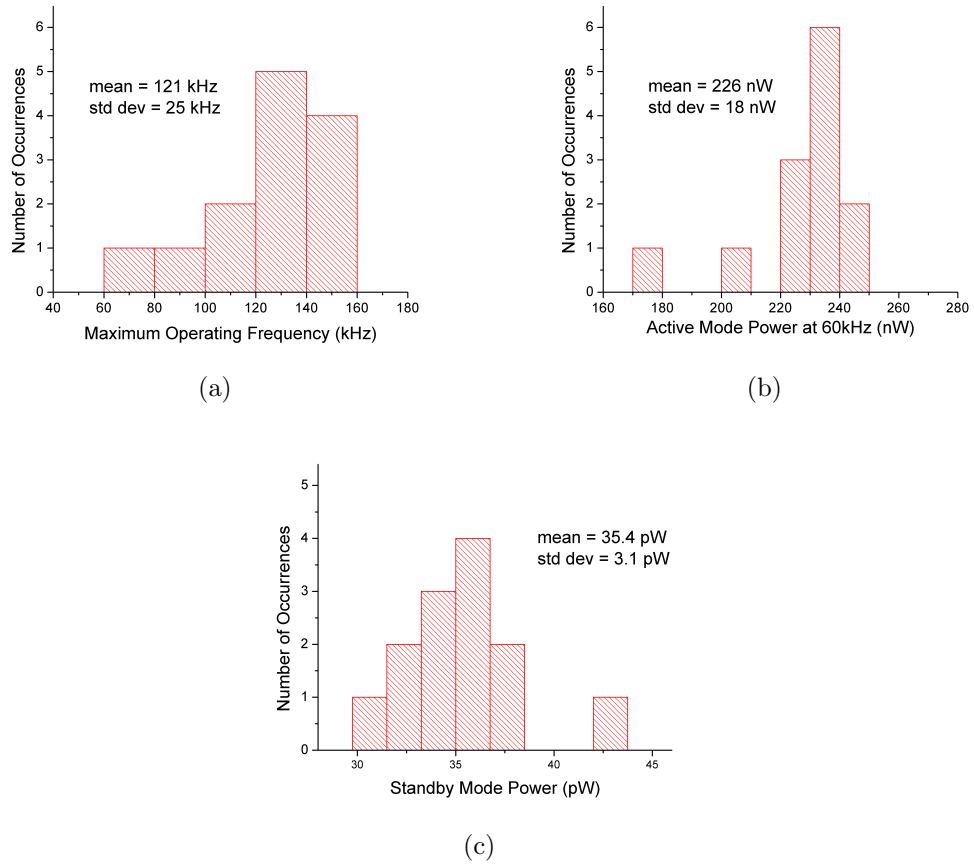


Figure 2.15: Measured (a) frequency distribution (b) active mode power distribution at 60 kHz (c) standby mode power distribution for 13 dies at $V_{DD}=0.5V$

Table 2.2: Comparison to other low voltage microcontroller

	Technology	Architecture	Frequency	Active Energy	Standby Power
[88]	0.13 μm	8b 8051	8MHz	18.8pJ/cycle	53.6 μW
[57]	65nm	16b MSP430	434kHz	27.3pJ/cycle	1 μW
[36]	0.13 μm	custom 8b	354kHz	3.5pJ/inst	153nW
[105]	0.25 μm	custom 8b	500kHz	12pJ/inst	13-20nW
Phoenix	0.18 μm	custom-8b	108kHz	2.8pJ/cycle	35.4pW

only 7% of the power. For a typical sensing application in which the sensor remains active for 1000 cycles every 10 minutes, these measurements indicate that the average power consumption is only 42pW, well within the limit of 177pW demanded by the battery described in Section 2.1. This ultra low power consumption is lower than any other work published as shown in Table 2.2.

2.10.2 Power Gating Results

To further investigate our proposed power gating approach, we sweep the footer width on the CPU and measure the energy and performance implications. Figure 2.16(a) shows the maximum operating frequency of the CPU as a function of footer width. Frequency reduces by $5\times$ as the footer size approaches the minimum of $W = 0.66\mu m$ and $L = 0.5\mu m$. This performance penalty leads to greater active energy consumption per operation since leakage energy increases with clock period in active mode. The power consumption through the power gating switch results in an additional energy penalty. However, the standby leakage power savings from the narrow footer width, shown in Figure 2.16(a), easily offsets these penalties and reduces the total energy for the Phoenix Processor. Figure 2.17 confirms that the total energy consumption is $3.8\times$ lower for the small footer ($W = 0.66\mu m$) than the large footer ($W = 28\mu m$) assuming 1000 instructions are executed every 10 minutes. The small footer saves several orders of magnitudes of total energy compared to a design with no power switch.

2.10.3 Memory Results

The IMEM and DMEM consume 7.1fW/bitcell (not including the overhead of decoders and row drivers). The IMEM alone accounts for 39% of the total standby power (including the overhead of decoders and drivers), which underscores the importance of instruction set optimization. If the instruction width had been set to 15

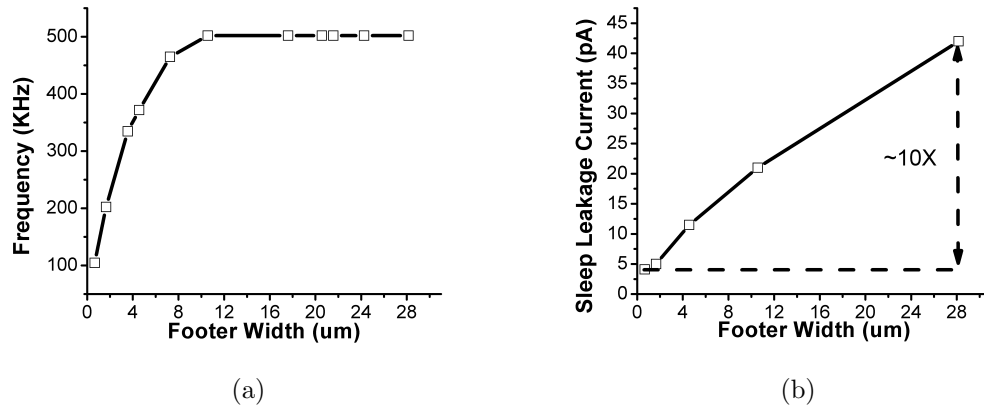


Figure 2.16: Measured (a) frequency and (b) standby leakage as functions of CPU footer width.

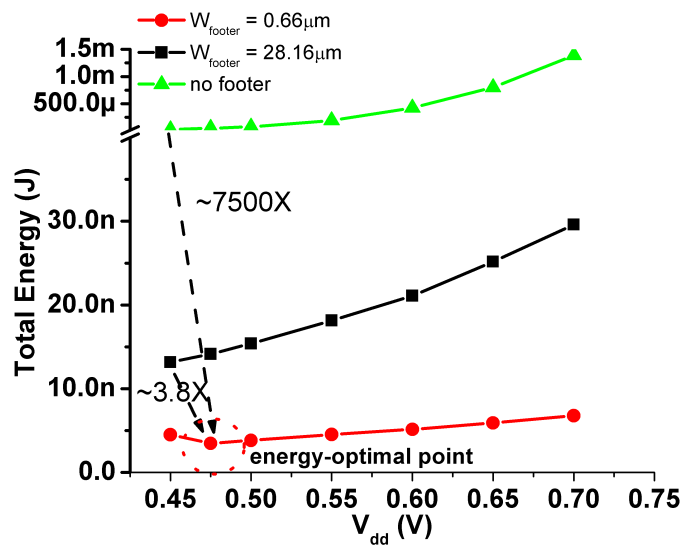
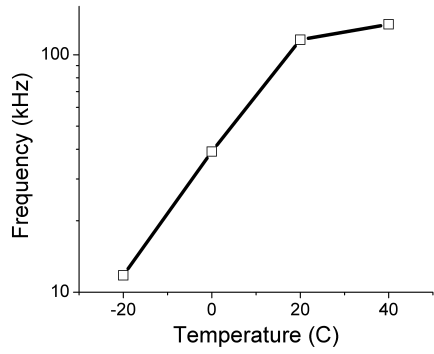


Figure 2.17: Total energy consumption assuming 1000 instructions are executed every 10 minutes.

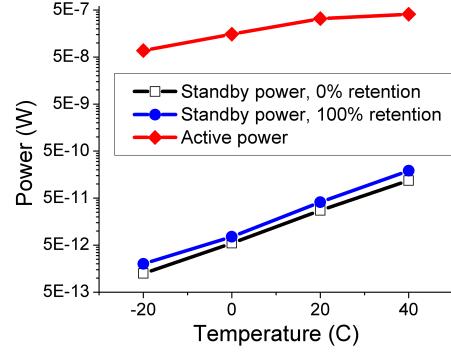
bits instead of 10 bits, measurements of a typical die show that the standby power of IMEM would have increased by 20%, which equates to 8% increase in total standby power.

Unlike IMEM, the power consumed by DMEM, which amounts to 51% of total standby power at 50% retention, can be reduced significantly by compressing the data and by changing the number of DMEM entries retained to match the footprint of compressed memory. For a typical die, the DMEM consumes 22pW with all entries retained and 7.5pW with all entries power gated due to the overhead of maintaining the free-list (i.e., the overhead of data compression).

To quantify the system-level benefits of compression and fine-grained power gating in memories, consider an ambient temperature sensing application in which the Phoenix Processor wakes up once every 10 minutes and runs a 1000 cycle routine to measure temperature and store the measured data. Using the measured temperature dependence of frequency and power consumption shown in Figure 2.18(a)(b) and a subset of the temperature profile from [1], we can compute the reduction in energy due to compression and power gating in DMEM over the lifetime of the chip. For this case study we assume that temperature is measured to a precision of $1^{\circ}C$. Figure 2.19(a) shows the total energy consumed over 37 hours (the time period over which uncompressed memory fills to capacity) for 3 cases. In Case 1, neither compression nor power gating are used in DMEM. In Case 2, power gating is used, but compression is not used. Finally, in Case 3, both power gating and compression are used. The use of power gating within DMEM (Case 2) reduces energy consumption by 7.3%, and the use of both power gating and compression (Case 3) reduces energy consumption by 14.7%. Compression also increases the effective size of DMEM and enables the processor to remain active for $7.8\times$ longer before memory fills to capacity, as shown in Figure 2.19(b).

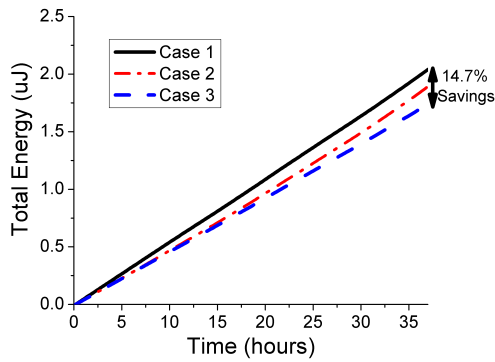


(a)

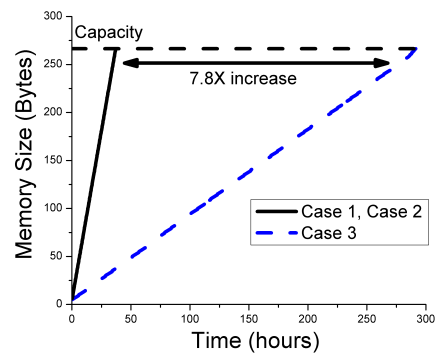


(b)

Figure 2.18: Measured (a) frequency and (b) power as functions of temperature.



(a)



(b)

Figure 2.19: Computed time profiles of (a) energy and (b) memory size for a temperature measurement routine.

2.11 Summary

In this work, we describe a sensor processor that operates at $V_{DD}=0.5V$ to minimize active mode energy and uses circuit, and architecture techniques to minimize standby mode energy. Measurements show that Phoenix consumes 226nW in active mode and only 35.4pW in standby mode. A thin film battery with the same form factor as Phoenix could provide a lifetime on the order of years, making Phoenix an attractive candidate for future cubic millimeter sensing systems.

CHAPTER III

Technology Selections for Ultra Low Voltage Design

3.1 Motivation and Previous Work

Figure 3.1 summarizes the range of recently reported Very Large Scale Integration (VLSI) circuits operating in the near or subthreshold voltage regime. The nominal supply voltages used during operation are denoted in the parenthesis. Figure 3.1 shows a wide performance spread for technologies ranging from 65nm to 180nm. However, the literatures of these designs provide little explanation on the rationales for their technology selection, implying that an in-depth consideration on choosing technology is lacking in the design phase. Since technology selection is expected to have a strong effect on leakage power and variability, the study on technology selection is clearly critical for successful ultra low voltage design.

In super-threshold operation where the nominal supply voltage for the technology is used, technology scaling has been a strong driver to enhance circuit performance, reduce area and improve power. However the benefit from using the latest technology for circuits operating in the sub- or near-threshold regime is unclear. The appropriate figure of merit in the ultra low voltage regime is "energy consumption per operation at a given performance" [113, 15]. In super-threshold operation, the target performance

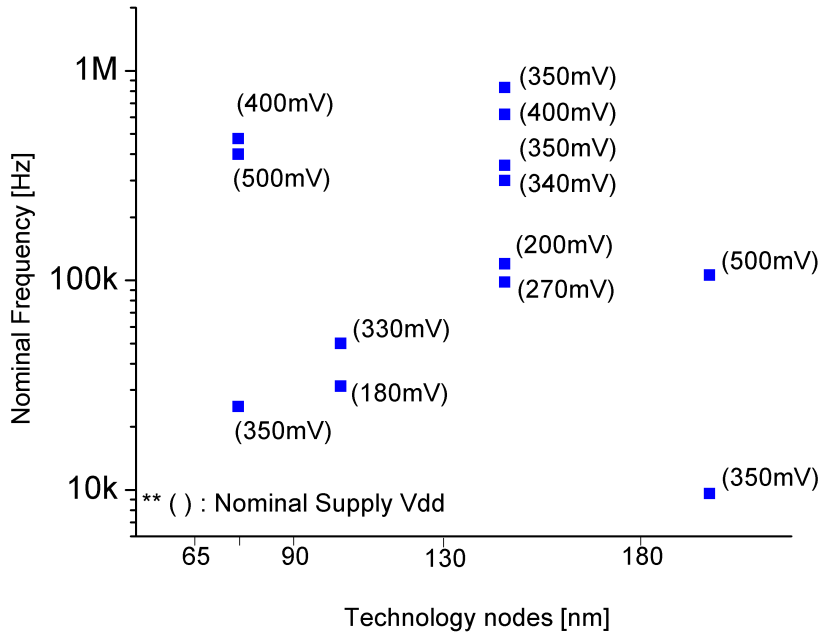


Figure 3.1: Published sub- or near-threshold VLSI designs.

is typically very high and hence only the most recent technologies meet the performance requirement, justifying the use of scaled technologies. On the other hand, the "given performance" for typical ultra low energy applications is substantially relaxed and in the range of 100 kHz to 1MHz, such that much older technology can fulfill the performance requirement as confirmed in Figure 3.1.

In addition, ultra low voltage applications often have low duty cycle (i.e., long standby times) [69]. Therefore, the energy consumption in standby should be emphasized in the design to ensure that the total energy will not be dominated by standby energy [84]. Since more recent technologies inevitably exhibit higher leakage, their use in ultra low voltage applications is further called into question.

Finally, the current scaling strategy [15] is tailored to traditional super-threshold operation. Therefore, the factors that become more important in ultra low voltage operation have often been given less emphasis as the technology scales. For example, [36] shows that subthreshold swing (S_S), which plays an important role in energy

per operation, degrades with technology scaling. The author of [74] calls for optimizing device technology for ultra low voltage operation however the practicality of developing a new technology solely for subthreshold and near-threshold operation is questionable given the resources required in technology development.

3.2 Contribution

We investigate the selection of optimal technology to achieve minimum energy and variability in ultra low voltage operation. We first define distinct application spaces according to their required performances and duty cycles. Then each technology is compared based on the minimum energy consumption for both CMOS logic and SRAM. This analysis also considers the use of leakage reduction technique as well as varying circuit compositions (logic to SRAM ratio) in the generic systems under consideration. We then investigate the optimal technology choice for minimum variability, which is a key concern in ultra low voltage operation. These optimal technology selections demonstrate that $1800\times$ energy saving and $4\times$ delay variation improvement can be achieved compared to poorly selected technologies. The Chapter concludes by analyzing the efficacy of circuit techniques to mitigate variability in the sub- and near-threshold regimes.

3.3 Ultra Low Voltage Application Spaces

Application space can be classified based on two metrics: required performance and duty cycle. In this study, required performance is defined as 40 FO4 inverter delays. It does not necessarily imply that the 40 FO4 delay is the exact number for clock frequency in super, near, or subthreshold operations. In [26], the typical clock frequency for a single pipeline stage in commercial microprocessor working in super-threshold regime lies in the range of 15 44 FO4 delay, depending on specific design

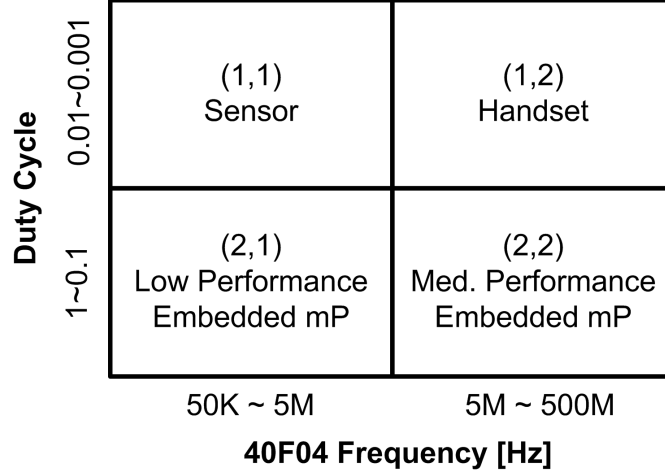


Figure 3.2: Application spaces in ultra low voltage operation.

differences. On the other hand, since microprocessors working in the subthreshold regime prefer shallow pipeline (more FO4 delay in a single pipeline stage) to mitigate variability[115], higher FO4 delay is observed in ultra low voltage microprocessor, for instance 400 FO4 delay of a single pipeline stage from]. Therefore, we consider the choice of 40 FO4 delay as a typical case that can be multiplied or divided according to a specific design needs.

The second metric, duty cycle, is defined as the ratio of active time to total time (sum of active time and standby time). Active time is the phase when a microprocessor has switching activity while standby time is an idling phase between active times.

By using these two metrics, we divide the whole application space of ultra low voltage into quadrants as shown in Figure 3.2. We briefly discuss each application space before turning to determining the technology preferences for each quadrant.

Application space (1,1), namely sensor type applications, has a very low duty cycle and low performance - this is the application space where subthreshold operation was originally of interest. Environmental sensor and implantable biomedical electronics is included in this category. In [69], the duty cycle of this type of application is estimated around 0.01 0.001. On the other hand, the relaxed required performance is likely to

set the supply voltage to the optimal energy point, denoted by V_{min} in [113]. Reducing the supply voltage below the V_{min} hurts the energy optimality because exponentially increased leakage energy offsets the quadratic saving in switching energy.

Application space (1,2) targeting handset application also has low duty cycle but higher performance requirement. We will shortly show that most of these applications require near-threshold as opposed to subthreshold operation to meet their target performance. Application spaces (2,1) and (2,2) have high duty cycles, and could be a good fit for embedded microprocessors for portable music players and home appliances.

Since each application space in the ultra low voltage regime has unique characteristics, it finds different technologies as optimal choice for minimum energy and variability, which is discussed in detail in the following sections.

3.4 Basic Optimal Technology Selection for Minimum Energy

In this section, we first investigate the minimum energy consumption for different technology nodes in the ultra low voltage application spaces. Five commercial technologies from 250nm to 65nm are considered.

3.4.1 Modeling Logic and SRAM for Energy Comparison

We model generic CMOS logic and SRAM by a simple inverter chain for energy analysis. Although RSCE (Reverse Short Channel Effect) has been reported to improve on/off current ratio in devices in the ultra low voltage regime [53], we do not consider this here to simplify the comparison across technologies. A chain of 40 of FO4 inverters are simulated in SPICE to determine propagation delay and energy consumption. Activity ratio for CMOS logic is assumed to be 0.2 while that for SRAM

is set to $10\times$ smaller at 0.02. In this section, energy reduction in standby mode is only achieved through clock gating while more efficient leakage saving techniques will be considered in the next section.

The supply voltage is set to meet the required performance in each application space. If the performance can be achieved at $V_{DD} = V_{min}$, the voltage is not scaled down below V_{min} . At the selected supply voltage, total energy consumption is modeled by EQ 3.1.

$$\begin{aligned} E_{total} &= E_{switch} + E_{leak} + E_{standby} \\ &= E_{active} + E_{standby} \end{aligned} \tag{3.1}$$

E_{switch} is defined as the energy consumed due to switching during active mode. E_{leak} is the leakage energy during active mode. E_{active} , energy consumption during active mode, is the sum of E_{switch} and E_{leak} . Finally $E_{standby}$ is the energy consumption during standby which is set by the duty cycle of the application space. Total energy consumption (E_{total} in EQ 3.1) for each technology will be compared to find the optimal technology selection in a given space.

3.4.2 Technology Choice for Minimum Energy

Figure 3.3(a) shows the energy optimal choice of technology for CMOS logic at each application space. There are two governing trends in the optimal choice. First, small duty cycle favors old technology due to its low standby power. Second, high performances prefers new technology because of its small active energy per operation from reduced parasitics and lower V_{DD} at which the technology can meet the required performance. For application spaces (1,1) and (2,2), the trends are most clear: (1,1) is dominated by old technologies while (2,2) finds new technologies as optimal choice for minimum energy.

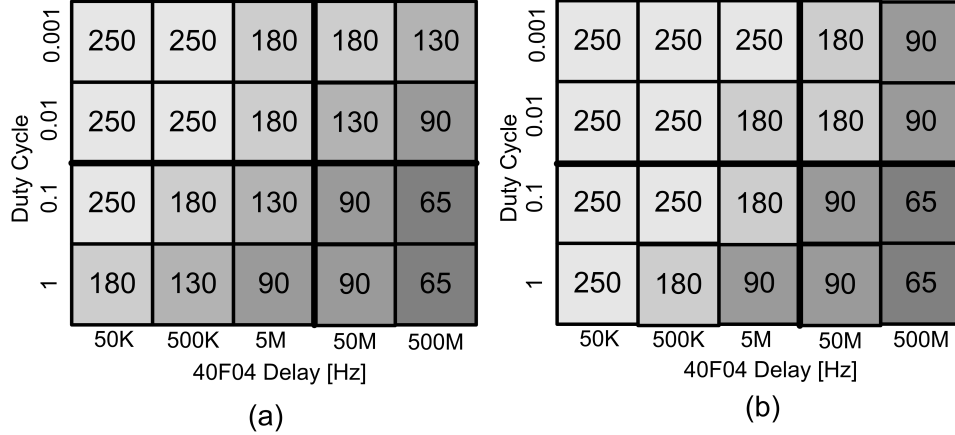


Figure 3.3: Basic technology selection for (a) logic (b) SRAM.

In application space (2,1), relatively old technologies are shown to be optimal. Due to the high duty cycle, we could expect newer technologies become energy optimal. However the higher performance of new technologies leads to task completion far ahead of required deadline, resulting in unwanted additional idle time and a waste of leakage energy. This energy penalty offsets the smaller active energy of the new technologies and makes older technologies a more preferred choice.

In the (1,2) quadrant, older technologies exhibit less leakage but higher active energy per operation since larger supply voltage is needed to meet the high performance requirements. Therefore the advantage of lower V_{DD} is offset by an increased active energy, leading to newer technologies becoming energy optimal. In the Figure 3.3(b), the same analysis is performed for SRAM. Slightly more preference is given to older technologies since total energy consumption from SRAM is more dominated by leakage power.

3.5 Impact of Standby Leakage Reduction

As pointed out in the previous section, standby energy plays an important role in optimal energy technology choice. Hence, we now consider leakage reduction techniques to supplement the results of Section 3.4.2.

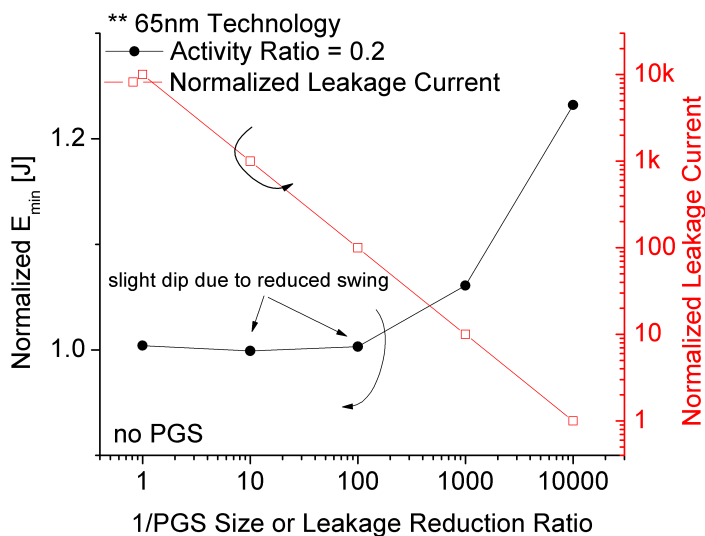


Figure 3.4: E_{min} and standby leakage current over leakage reduction ratio (E_{min} shows small dip due to virtual ground bounce)

3.5.1 Leakage Reduction Methods in Ultra Low Voltage Regimes

For CMOS logic several techniques, such as Power Gating Switch (PGS), have been proposed both in super-threshold [44] and ultra low voltage [84] operation to reduce standby leakage power. In [84] it was shown that both PGS size and supply voltage should be concurrently optimized for ultra low voltage operation. This co-optimization process depends on the application duty cycle: for high duty cycle, wide PGS that incur a small voltage drop are preferred to reduce E_{active} while sacrificing $E_{standby}$. For small duty cycles, very narrow PGS should be used to minimize $E_{standby}$ and hence total energy. While a small PGS reduces standby power, it also induces a longer delay due to the incurred drop across the PGS voltage and thus increasing E_{leak} . Figure 3.4 confirms that the stronger leakage reduction of a small PGS increases the E_{min} (Energy consumption at V_{min}) while saving leakage current.

In an SRAM it is more difficult to dramatically reduce standby leakage power due to its state retention requirement. Rather than traditional PGS, other techniques specific to SRAMs [52, 91, 51, 50] have been studied. The ability to reduce leakage

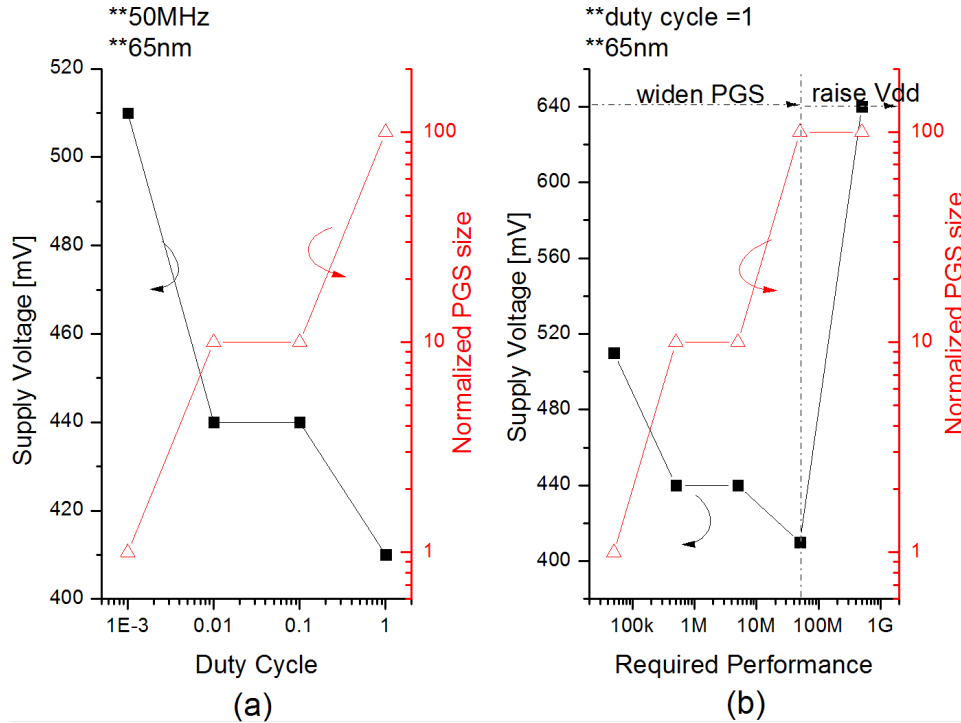


Figure 3.5: Result of co-optimization with sweeping (a) duty cycle (b) required performance

power, however, is much smaller than in the CMOS logic case with a PGS.

3.5.2 Technology Choice with Leakage Reduction Schemes

We re-investigate the energy optimal technology selection after applying the co-optimization of a PGS for CMOS logic. The co-optimization determines supply voltage and PGS size as a function of the required performance and duty cycle. In Figure 3.5(a), low duty cycle leads to a smaller PGS size and higher supply voltage. On the other hand, as the performance requirement becomes more stringent, a wider PGS is initially preferred followed by a higher supply voltage. (Figure 3.5(b)).

After this co-optimization process, the energy optimal technology choice for CMOS logic is determined as shown in Figure 3.6(a). Since more recent technologies exhibit large leakage currents, they benefit greatly from applying such a V_{DD} /PGS co-optimization. Consequently newer technologies become more desirable for logic in

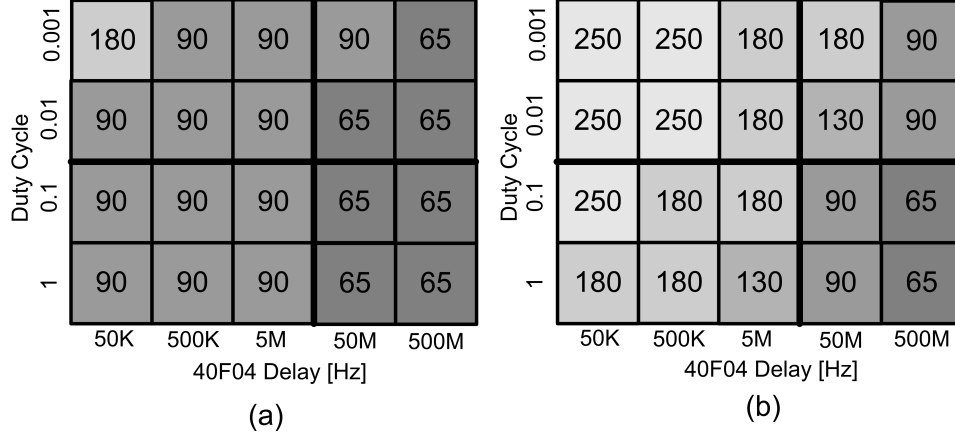


Figure 3.6: Optimal technology selection (a) for logic with co-optimized V_{DD}/PGS , (b) for SRAM with 10x leakage reduction

several application spaces than in Section 3.4.

However, the leakage power saving for SRAM is less effective due to the data retention issue. Instead of applying co-optimization, we fix the leakage saving for SRAM at 10 \times , a typical figure from [52, 91, 51, 50]. After applying this amount of leakage savings, the optimal choice for SRAM is determined as shown in Figure 3.6(b). The preference to older technologies is slightly reduced compared to Figure 3.3(b), however they still tend to dominate for SRAM except in highest performance scenarios.

3.6 Effect of Logic and SRAM Ratio on Technology Selection

In this section, we consider a generic system containing both CMOS logic and SRAM, which so far have only been considered separately. For this analysis, a single supply voltage for both logic and SRAM is assumed and is set to minimize the total energy. We first study energy density per operation and standby leakage density for both logic and SRAM, then investigate the effect of circuit composition ratio (SRAM area to total area) on the optimal technology selection.

Figure 3.7 shows the expected energy density per operation and leakage power

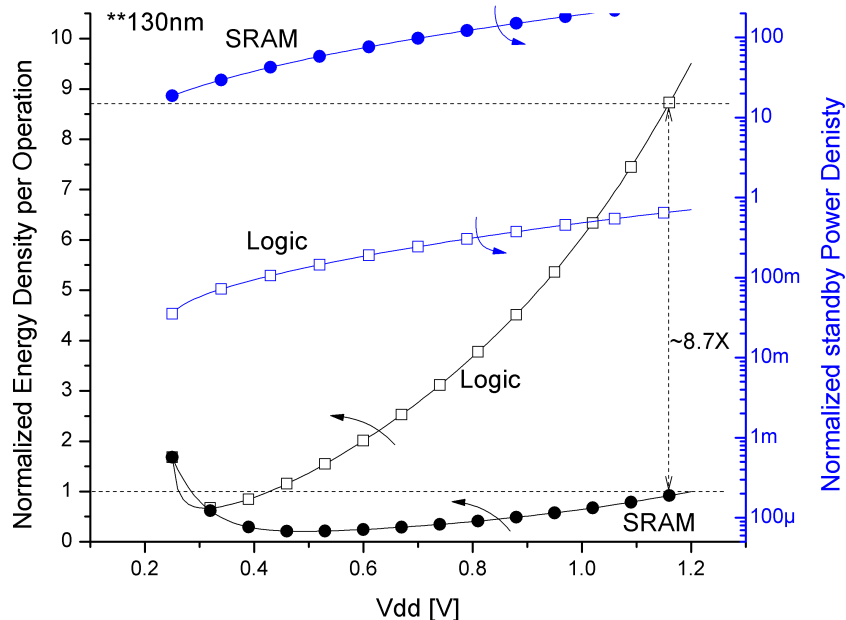


Figure 3.7: Energy density per operation and standby power density

density for CMOS logic and SRAM over supply voltage. In super-threshold operation, the energy density difference is $8.7\times$ which is similar to the number reported in [13]. However in ultra low voltage operation this difference is much reduced (i.e., $1.2\times$ at 300mV). In terms of standby power density, SRAM remains substantially higher than logic throughout the supply voltage range, which is noted in [99]. The net result is that logic does not have a significant impact on the choice of an optimal energy technology since SRAM tends to dominate the total energy.

Figure 3.8 shows the optimal technology choice at two different SRAM ratios in the system: 0.5 (i.e., 50% of total system area is SRAM) and 0.8. As we pointed out, SRAM strongly affects the technology choice for minimum energy as shown in 8(a) and 8(b). Comparing the two SRAM area ratios, the higher SRAM ratio clearly shifts the optimal technology choice toward the SRAM-only case.

Finally, the amount of energy saving possible in moving from the worst to the best case technology choice is shown in Figure 3.9(b). The savings can be as much as

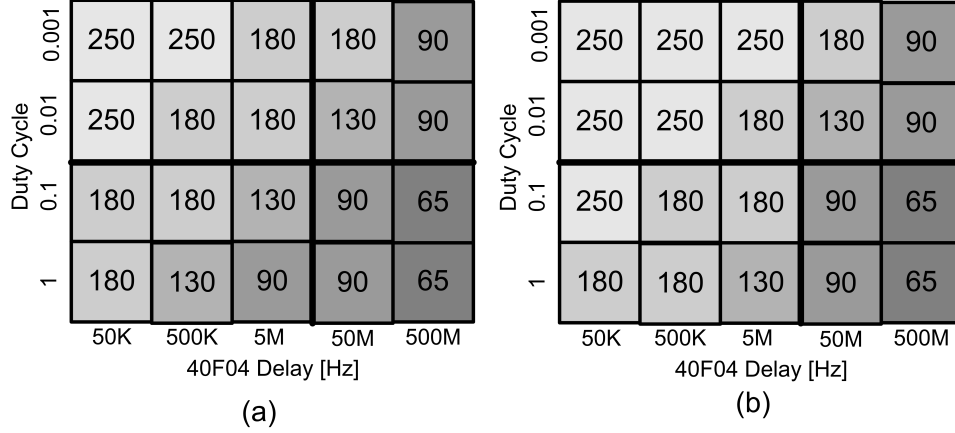


Figure 3.8: Energy optimal technology selection with SRAM/logic area ratios of (a) 0.5 (b) 0.8

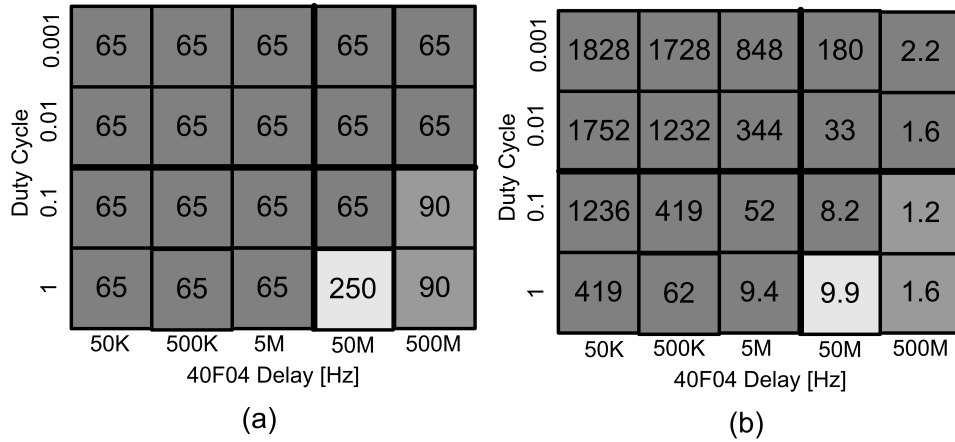


Figure 3.9: (a) Worst energy technology selection for SRAM ratio of 0.8 (b) energy saving ratio $[\times]$ by moving to Figure 3.8(b)

three orders of magnitude for the (1,1) application space due to the large differences in standby energy consumption. The savings become much smaller for the medium performance regime, as fewer technologies can meet the required performance.

3.7 Variability and Technology Selection

To this point we have investigated the optimal technology choice for minimum energy operation. This section changes focus to the role of process variability in technology selection since variability is among the most important concerns in the

ultra low voltage design space. In this section the measure of variability is defined as the σ/μ of 40 F01 inverter chain delay. Using the embedded data in SPICE models, either die-to-die (global shift) as well as with random within-die (mismatch) variation and mismatch-only variation are simulated, as discussed in the following sections.

3.7.1 Impact of Operating Point on Variability

Variability is known to be a key concern for subthreshold operation due to the exponential change of current in the subthreshold regime. There have been several studies on variability in the ultra low voltage regime. In [115], mismatch is analyzed, showing that "random dopant fluctuation (RDF)-induced V_{th} variation" becomes a single dominant variability source while both RDF induced V_{th} variation and critical dimension variation are equally important in super-threshold operation. In [24], the effect of variability on SRAM is analyzed concluding that ultra low voltage operation forces the size of SRAM cells to increase to mitigate variability.

Although previous studies have emphasized V_{th} variation within the topic of ultra low voltage variability, subthreshold swing (S_S) is also an important factor determining the magnitude of variability. Delay variability arises due to a two-stage process: First there exists some degree of V_{th} variation and second, this V_{th} variation is transformed to current variation (and hence delay variation) through the S_S as dictated by EQ 3.2. If S_S is very steep, a fixed amount of V_{th} change leads to an exponentially larger variation.

$$\Delta I \propto \exp\left(\frac{\Delta V_{th}}{S_s}\right) \quad (3.2)$$

One important observation is that the S_S of interest depends on the operating point, which can be either near-threshold or subthreshold. As shown in Figure 3.10(a), S_S is less steep in the near-threshold regime. Therefore less current variability is expected to result given the same amount of V_{th} variation. Figure 3.10(b) shows

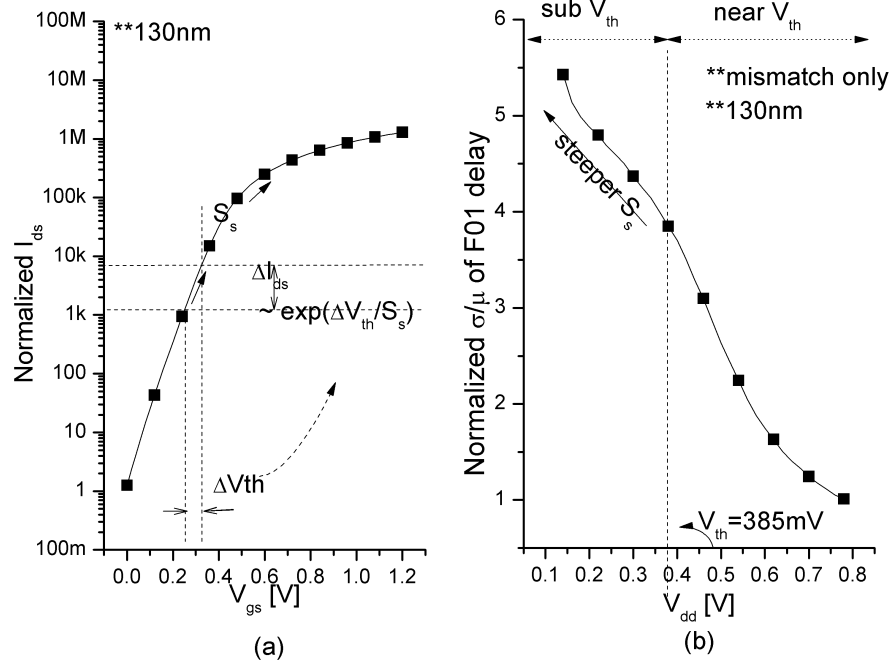


Figure 3.10: (a) Log I_{ds} - V_{gs} curve, (b) change of variation depending on operating point

the effect of operating point on variability from mismatch-only simulation, where a smaller variability is observed as the operating point moves toward the near-threshold regime.

3.7.2 Technology Selection for Min Variability

The operating point at which the transformation from V_{th} variation to delay variation occurs varies from technology to technology and across the defined application spaces. The key parameter is the difference between the chosen supply voltage for a given technology to minimize energy and the threshold voltage. In this section we focus on V_{min} operation since V_{DD} will typically be set to V_{min} for low performance applications. However the operations at $V_{DD} \neq V_{min}$ can be treated in the same analysis presented in this section. We first investigate the change of the optimal energy operation point over technology scaling and then analyze its effect on variability. Fig-

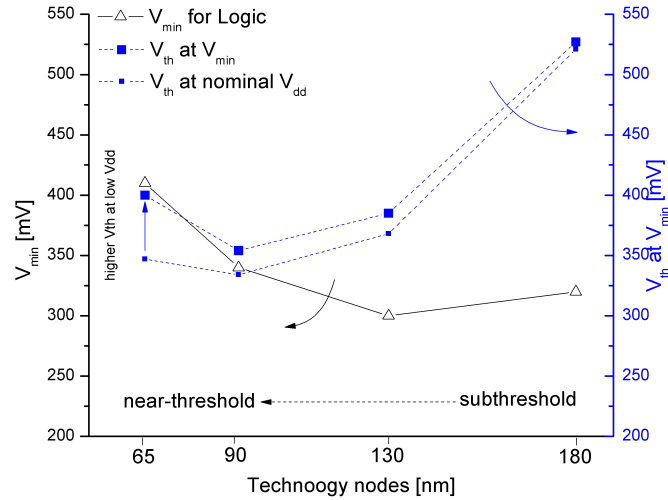


Figure 3.11: Optimal energy operation point over technologies

Figure 3.11 shows the change of the optimal energy point over technology scaling. Due to the reduced current on/off ratio with smaller technology, V_{min} tends to increase [113], which is confirmed in Figure 3.11. In addition, V_{th} has a decreasing trend with technology scaling. These trends combine to move optimal energy operation toward the near-threshold regime for scaled technologies.

This change of the optimal energy operating point directly affects the magnitude of the delay variability in the ultra low voltage regime. In Figure 3.12, we show a decreasing trend of delay variability with technology scaling at optimal energy operating points. Both global and mismatch in process parameters were considered. A $4\times$ improvement in total variability is observed moving from 180nm to 65nm node. While the V_{th} variation at ultra low voltage increases with technology scaling, overall variability is decreases, which confirms the importance of the operating point on variability in the ultra low voltage regime. In addition, mismatch-only analysis shows the same trend of reducing variability with process scaling. Therefore we can conclude that more recent technologies exhibit lower variability both in mismatch and die-to-die variation at their energy optimal operation point.

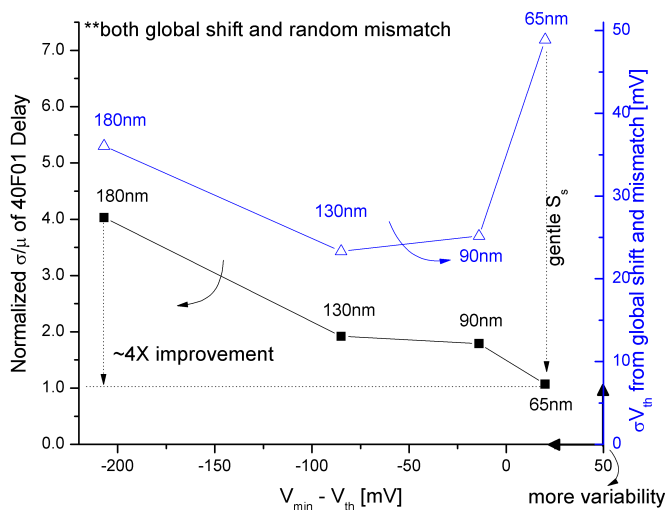


Figure 3.12: variability as a function of technology choice

3.7.2.1 Mitigating Variability

Despite the conclusion of Section 3.7, mitigating variability remains a significant challenge in the ultra low voltage regime. In particular, V_{min} is a function of not only technology parameters but also circuit topology including the strength of PGS [84] and the activity ratio [113]. Hence the operating point could inevitably occur in the subthreshold regime depending on design choices. Furthermore, the absolute amount of variability at any V_{min} may still be unacceptable for a given application even in the near-threshold regime.

energy per operation increase with less variability Here, we briefly study the effectiveness of two techniques to reduce variability: Upsizing devices to suppress mismatch variability [113, 56] and increasing supply voltage, which trades off a global energy penalty for a reduced sensitivity to underlying process variation sources (i.e., S_S at the operating point flattens). Figure 3.13(a) focuses on the case where mismatch is the only source of variation and shows that increasing transistor sizes can achieve a better tradeoff between energy penalty for variability improvement, depending on the operating regime. After some point, however, raising supply voltage becomes more

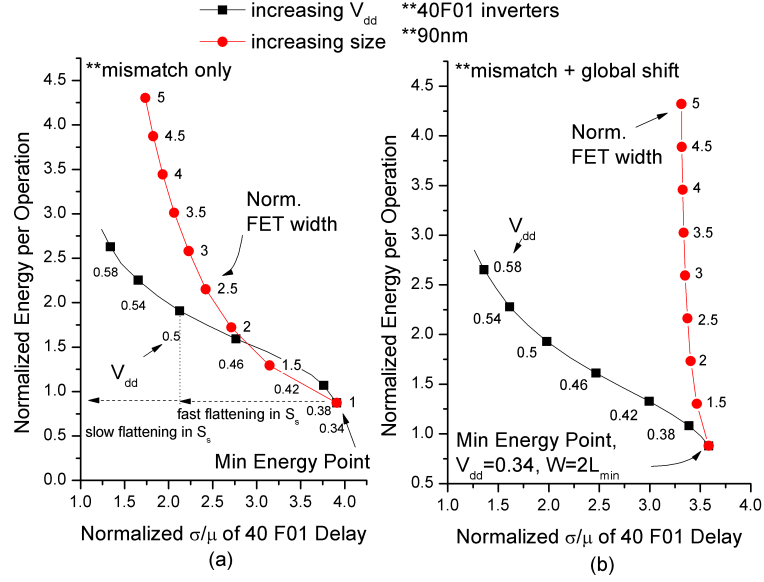


Figure 3.13: (a) Mismatch-only (b) die-to-die and mismatch, energy per operation increase with less variability

advantageous since the efficacy of the upsizing decreases with the larger device size as shown in EQ 3.3 [75].

$$\Delta V_{th} \propto \frac{1}{\sqrt{WL}} \quad (3.3)$$

When adding global shift to the analysis along with mismatch, increasing V_{DD} becomes the more effective knob to suppress variability with lower energy cost, as shown in Figure 3.13(b) since global variation is unaffected by transistor sizing.

3.8 Summary

We present one of the first systematic studies of technology selection for the ultra low voltage operation. Technology selection is analyzed and optimized for two criteria: minimum energy and minimum variability. For energy minimization, mature technologies are often the optimal choice depending on application space and can save up to $1800\times$ in total energy. It is also shown that, counter to intuition, newer

technologies actually reduce delay variability in the ultra low voltage regime by up to 4×.

CHAPTER IV

Power Gating Switch Design for Ultra Low Voltage Operations

4.1 Motivation and Previous Work

Standby energy, which has become important in modern CMOS processes due to the increasing contributions of subthreshold and gate leakage current, becomes more significant in ultra low voltage operations for two reasons. First, the reduced switching energy consumption from scaled supply voltages renders the sleep energy a more significant portion of total energy consumption. Second, ultra low power applications often have low duty cycles. Although they run slowly at V_{min} , there is a considerable amount of sleep time between the moment of completing a task (T_{min}) and the start of a new task ($T_{deadline}$), as defined in Figure 4.1. Although ultra low voltage computational cores [103, 46] and general microprocessors [116, 36, 83] have been designed and tested showing that ultra low voltage designs achieve the active energy consumption of several pJ per cycle, these designs have often overlooked the importance of sleep energy consumption. Since there is a considerable amount of sleep energy consumption during the period, an optimization method that considers sleep energy consumption is vital to an energy-optimal design.

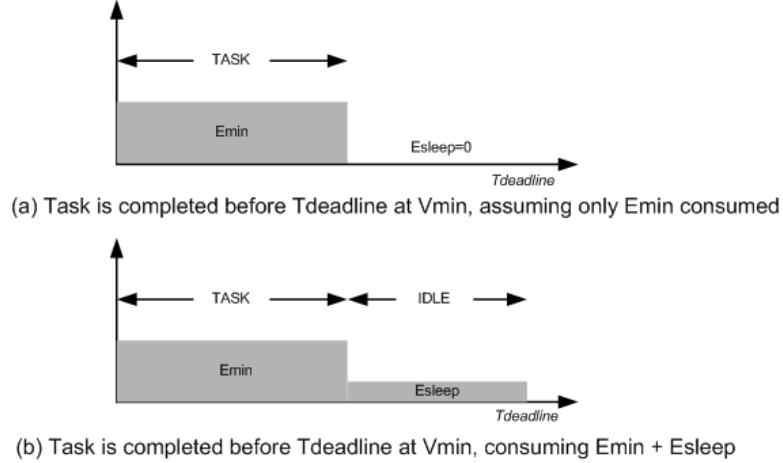


Figure 4.1: Illustration of task scheduling at different deadlines

4.2 Contributions

In this Chapter, we start by proving the importance of sleep energy for reducing total energy consumption. Then, we discuss the effects of PGS [67], a well-known sleep energy reduction scheme, on energy consumption in ultra low voltage regimes. Our proposed optimization, which modulates PGS size and supply voltage simultaneously, suggests using very small PGSs with a supply voltage higher than V_{min} , unlike conventional practices in which a large PGS is often used (typically 10% of total NFET width). In SPICE simulations of generic circuits, the optimization method achieves $125\times$ reduction in total energy consumption and $50\times$ savings in PGS area. The effectiveness of this proposed optimization is also confirmed by measurement results from a fabricated microprocessor. We also discuss the functional feasibility of using minimal PGSs with SPICE simulations and silicon measurements. Finally, other approaches to perform power gating are quantitatively compared for energy optimal designs.

4.3 Impact of Sleep Energy on Total Energy Consumption

We first investigate the case in which circuits experience non-zero sleep time. In other words, T_{min} , the time when circuits complete a task at the traditional $V_{DD}=V_{min}$ comes earlier than $T_{deadline}$, the moment when the circuit begin a new task. In this case, the total energy is the sum of sleep energy (E_{sleep}) and active energy ($E_{switch} + E_{leak}$). We define duty cycle K_{duty} as $T_{deadline} / T_{min}$, which represents the ratio of maximum allotted-time to actual used-time (i.e., circuit delay at V_{min}). If $K_{duty} > 1$, then circuits experience sleep time and consume additional energy.

For this scenario, we run SPICE simulations using inverter chains to estimate the contribution of sleep energy consumption to total energy consumption. In this paper, SPICE simulations are performed using a commercial $0.13\mu m$ CMOS technology. Unless mentioned explicitly, a 99-stage inverter chain is used. Inverters use regular V_{th} devices while PGSs uses high V_{th} device. The $V_{th,high-V_{th}}$ is 560mV and $V_{th,regular}$ is 350mV at nominal conditions. At $V_{DD}=V_{min}$ (220mV), E_{min} of the inverter chain is simulated as 15.4fJ/cycle at a delay of 5.66s (176 kHz). NFET and Positive Channel Field Effect Transistor (PFET) in inverters are sized at $0.32\mu m$. The logic depth of the inverter chain is equivalent to 25 Fan-Out-of-4 (FO4) delays, which is shorter than most ultra low voltage designs. For a single inverter chain, the circuit activity rate is 1. The chosen logic depth and switching activity approximate the worst case voltage drop scenario across power gating switches, and provide conservatism in the results.

We initially assume that there is no cutoff technique applied in sleep mode. The total energy consumption for inverter chains can be expressed as EQ 4.3, which is derived from EQ 4.2. EQ 4.3 shows that nearly the same amount of leakage current exists for both sleep and active time. Therefore, a significant increase in total energy consumption is expected. Figure 4.3 shows that sleep energy contributes a large amount of energy consumption at lower duty cycles or higher K_{duty} (i.e., circuits

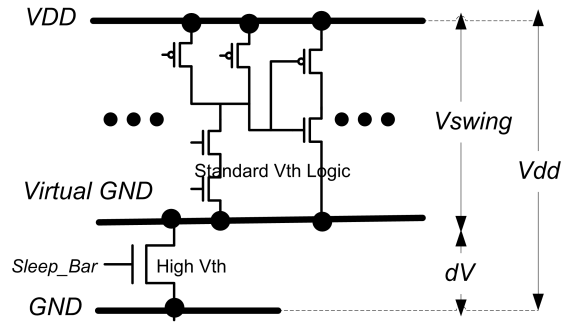


Figure 4.2: Basic PGS configuration

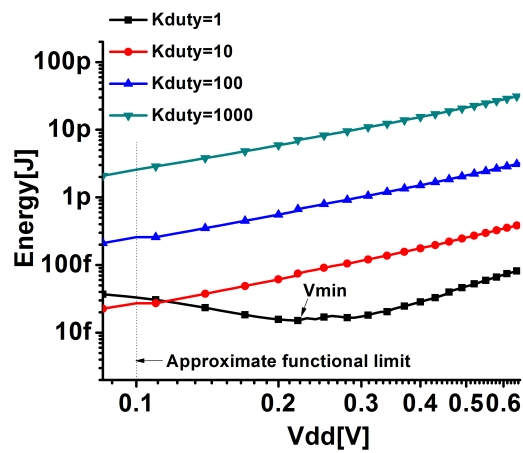


Figure 4.3: V_{min}/E_{min} curves with different K_{duty} considering sleep energy

spend more time in sleep mode). Since ultra low power applications often have low duty cycles, it is paramount to consider sleep energy in total energy optimization frameworks.

$$E = E_{switch} + E_{leak} = \frac{1}{2}n \cdot C \cdot V_{DD}[\alpha + \eta \cdot n \cdot e^{\frac{-V_{DD}}{mV_T}}] \quad (4.1)$$

$$\begin{aligned} E_{total} &= E_{switch} + E_{leak} + E_{sleep} \\ &= E_{switch} + t_{delay} \cdot P_{leak} + (T_{deadline} - t_{delay}) \cdot P_{leak} \\ &= E_{switch} + T_{deadline} \cdot P_{leak} \end{aligned} \quad (4.2)$$

Another interesting observation is that both E_{switch} and $T_{deadline}P_{leak}$ in EQ 4.3 are proportional to V_{DD} , resulting in lower energy-optimal supply voltage than conventional V_{min} , as shown in Figure 4.3. The optimal supply voltage can be scaled down until CMOS gates fail to function, while it is often bounded by the contribution of leakage energy in the conventional analysis. The minimal functional voltage for CMOS gates is assumed to be 100mV, although this assumption has little impact on the results of this work.

4.4 The Effects of Cutoff Structures on Total Energy Consumption

Given the significant contribution of sleep energy to total energy consumption, PGSs are attractive for improving overall energy efficiency. While several other methods can be used in sleep mode, such as reverse body-biasing, PGSs are considered the most effective measure to reduce leakage energy consumption [10, 9]. However, PGS design in ultra low voltage regimes differs from conventional practices. Therefore, in this section, we first study the effects of PGSs on energy consumption of circuits

operating in ultra low voltage regimes. Section 4.5 then lays out a strategy for using PGSs to minimize total energy consumption based on our findings in this section.

The purpose of employing PGSs in circuits is to reduce sleep power by strongly shutting off leakage paths during sleep modes. However, the benefit of reducing sleep energy consumption comes with performance degradation due to the voltage drop across PGSs [67]. In ultra low voltage regimes, the performance degradation can induce extra active energy consumption since circuits consume extra leakage energy for longer periods of active time. Therefore, designers should be aware of the effects of PGSs on sleep and active energy consumption in ultra low voltage regimes.

To capture the effects of PGSs on energy consumption, we propose two parameters in EQ 4.4. The first parameter, denoted by K_{leak} , sleep energy reduction factor, is the ratio of sleep power with PGSs to sleep power without such structures. The second parameter, the delay degradation factor, denoted by $1/K_{delay}$, is the ratio of circuit delay with PGSs to delay without them.

$$\begin{aligned} \frac{1}{K_{delay}} &= \frac{t_{delay_w_cutoff_circuitry}}{t_{delay_wo_cutoff_circuitry}} \\ K_{leak} &= \frac{I_{leak_w_cutoff_circuitry}}{I_{leak_wo_cutoff_circuitry}} \end{aligned} \tag{4.3}$$

4.4.1 Theoretical Power Gating Switch

This section investigates E_{min} assuming that circuits use a theoretical PGS having independent controls on K_{leak} and $1/K_{delay}$. For example, EQ 4.5 shows the total energy consumption of circuits with the PGS of K_{leak} and $1/K_{delay}$, where T_{min} denotes the delay of main circuits at V_{min} without the PGS; P_{leak} denotes the leakage power without the PGS; and t_{delay} expresses the delay of main circuits at a specific V_{DD} with the PGS. In EQ 4.5, E_{switch} is technically affected by the PGS due to the change of the voltage swing. However, this can be ignored without sacrificing much

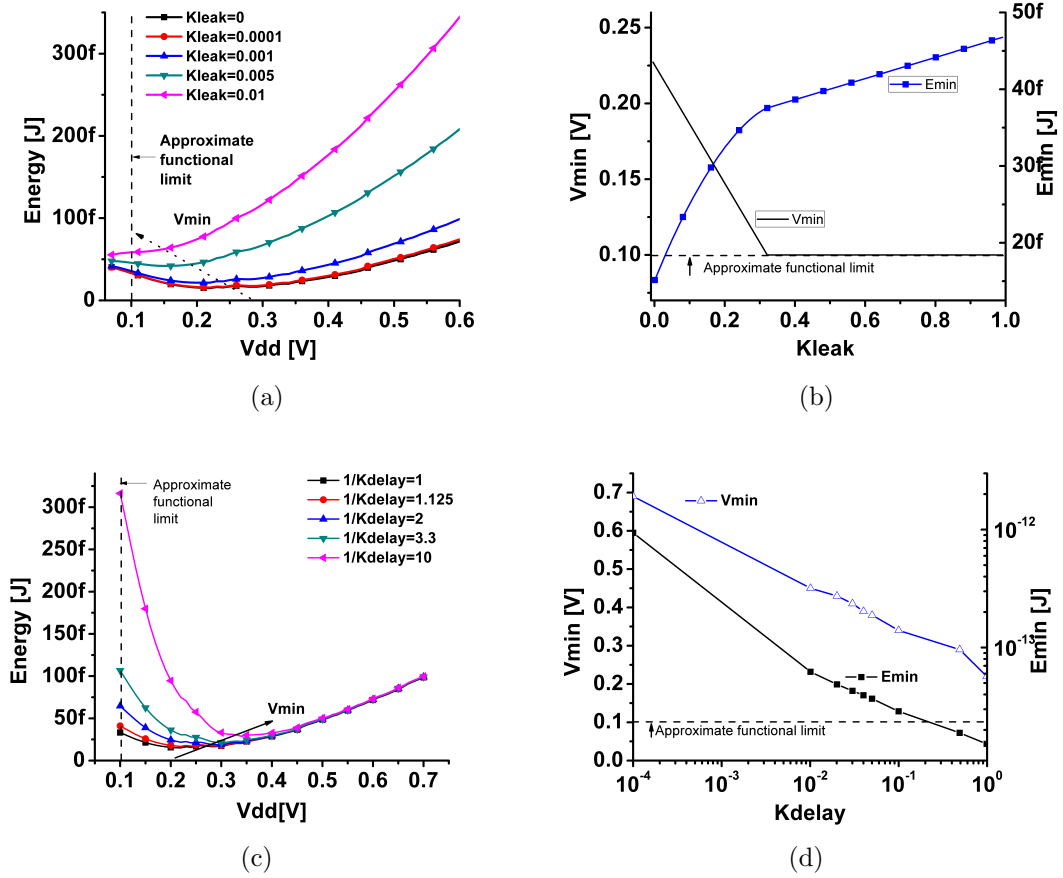


Figure 4.4: V_{min} / E_{min} change with K_{leak} and K_{delay} , (a) V_{min}/E_{min} curves (b) $K_{leak} - V_{min}/E_{min}$, (c) V_{min}/E_{min} curves, (d) $K_{delay} - V_{min}/E_{min}$

accuracy. However, we include the changes of E_{switch} after this section for a more complete analysis.

$$\begin{aligned}
 E_{total} &= E_{switch} + E_{leak} + E_{sleep} \\
 &= E_{switch} + \frac{1}{K_{delay}} t_{delay} P_{leak} + (K_{duty} T_{min} - \frac{t_{delay}}{K_{delay}}) \cdot K_{leak} P_{leak}
 \end{aligned} \tag{4.4}$$

We investigate the changes of V_{min} and E_{min} while sweeping either K_{leak} , as shown in Figure 4.4(a) and 4.4(b), or $1/K_{delay}$ as shown in Figure 4.4(c) and 4.4(d). Figure 4.4(a) and 4.4(b) show that small values of K_{leak} can reduce E_{sleep} and push V_{min} to a conventional V_{min} . On the other hand, Figure 4.4(c) and 4.4(d) show that large values

of $1/K_{delay}$ increases E_{leak} due to the longer delay. Since higher V_{DD} can alleviate the performance degradation, higher V_{min} is preferred to offset the increase of E_{leak} for this case.

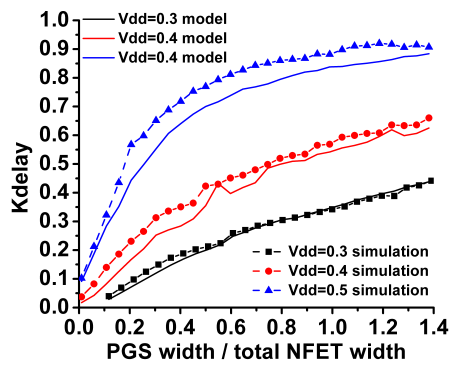
4.4.2 Practical Power Gating Switch

While we assume PGSs with independent controls on K_{leak} and $1/K_{delay}$ in the previous section, they are actually co-related in practical PGS designs. For a simple PGS (Figure 4.2), we can derive $1/K_{delay}$ and K_{leak} in ultra low voltage regimes, as shown in EQ 4.6 and EQ 4.7. In the derivation, it is assumed that the voltage across the PGS in sleep mode is V_{DD} , due to the very high resistance of the PGS when it is off. V_{swing} , which is a highly non-linear function of PGS width and technology parameters, reduces for wider PGSs and lower threshold voltages.

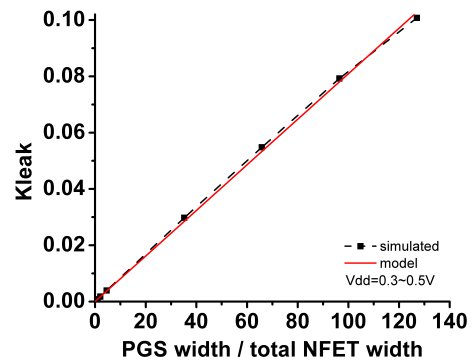
$$\begin{aligned}
\frac{1}{K_{delay}} &= \frac{t_{delay_w_mtcmos}}{t_{delay_wo_mtcmos}} \\
&= \frac{CV_{swing}}{CV_{DD}} \cdot \frac{\mu C_{ox} \frac{W}{L} V_T^2 (1-m) e^{\frac{-V_{swing}}{mV_T}} (1 - e^{\frac{-V_{swing}}{mV_T}})}{\mu C_{ox} \frac{W}{L} V_T^2 (1-m) e^{\frac{-V_{DD}}{mV_T}} (1 - e^{\frac{-V_{DD}}{mV_T}})} \\
&= \frac{V_{swing}}{V_{DD}} e^{\frac{V_{DD} - V_{swing}}{mV_T}}
\end{aligned} \tag{4.5}$$

$$\begin{aligned}
K_{leak} &= \frac{I_{leak_w_sleep_structure}}{I_{leak_wo_sleep_structure}} \\
&= \frac{\mu_{eff,hvt,n} C_{ox} \frac{W_{hvt,n}}{L_{eff,hvt,n}} \exp\left(\frac{-V_{th,hvt,n}}{mV_T}\right) (1 - \exp(-\frac{V_{DD}}{V_T}))}{\mu_{eff,svt,n} C_{ox} \frac{W_{svt,n}}{W_{eff,svt,n}} \exp\left(\frac{-V_{th,svt,n}}{mV_T}\right) (1 - \exp(-\frac{V_{DD}}{V_T}))} \\
&= k \cdot W_{hvt,n}
\end{aligned} \tag{4.6}$$

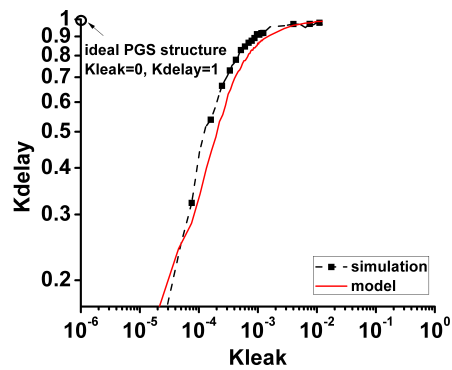
Both $1/K_{delay}$ and K_{leak} are functions of PGS width and supply voltage, as shown in EQ 4.6 and EQ 4.7. $1/K_{delay}$ can quickly approach 1 by increasing the width of PGSs at high supply voltages, while it slowly increases at low supply voltages.



(a)



(b)



(c)

Figure 4.5: K_{leak} and K_{delay} change with PGS width and V_{DD} , (a) width - K_{delay} (b)width - K_{leak} (c) K_{leak} - K_{delay}

On the other hand, K_{leak} is a linear function of the width of PGSs at a wide range of supply voltages. Figure 4.5(a) and 4.5(b) compare the derived equations against SPICE simulations, demonstrating acceptable accuracy. Figure 4.5(c) shows the inter-relationship between K_{leak} and $1/K_{delay}$ as the width of PGSs is swept. The ideal cutoff structure point is at the point where $K_{leak} = 0$ and $1/K_{delay} = 1$. This figure also provides a means to quantitatively compare the efficacy of different PGSs for ultra low voltage regimes, as discussed further in Section 4.7.

As shown in EQ 4.7, total energy consumption can be derived from EQ 4.5, 4.6. The change of E_{switch} from PGS is included here for higher accuracy. EQ 4.7 shows that the total energy is a function of V_{DD} , K_{leak} , K_{delay} and technology parameters. T_{min} is the circuit delay without PGSs evaluated at its own V_{min} , and is thus constant.

$$\begin{aligned}
E_{total} &= E_{switch} + E_{leak} + E_{sleep} \\
&= \frac{1}{2} \cdot n \cdot CV_{DD}V_{swing}\alpha \\
&+ \frac{1}{2} \cdot n \cdot CV_{DD}^2\eta \cdot ne^{(-\frac{V_{DD}}{mV_T})} \cdot \left(\frac{V_{swing}}{V_{DD}}e^{\frac{V_{DD}-V_{swing}}{mV_T}}\right) \\
&+ (K_{duty}t_{norm} - e^{\frac{V_{DD}-V_{swing}}{mV_T}} \cdot t_{min}) \cdot k \cdot W_{hvt,n} \cdot k \cdot W_{hvt,n} \cdot I_{leak_wo_mtcmos} \cdot V_{DD} \\
E_{total} &\approx K_1V_{DD}V_{swing} + K_2V_{DD}^2e^{\frac{-V_{DD}}{mV_T}} \left(\frac{V_{swing}}{V_{DD}}e^{\frac{V_{DD}-V_{swing}}{mV_T}}\right) + K_3K_{duty}kW_{hvt,n}V_{DD}
\end{aligned} \tag{4.7}$$

Since K_{leak} and K_{delay} are functions of supply voltage and PGS width, we investigate energy consumption by sweeping both of these parameters. Sleep energy consumption is roughly proportional to both supply voltage and PGS width. Here, the effect of tdelay on sleep energy consumption is ignored since for large K_{duty} the tdelay term in E_{sleep} is much smaller than $K_{duty} \times T_{min}$ while for small K_{duty} the sleep energy consumption itself becomes small and less important in E_{total} . Additionally, subthreshold leakage current, the dominant source of sleep energy consumption, is

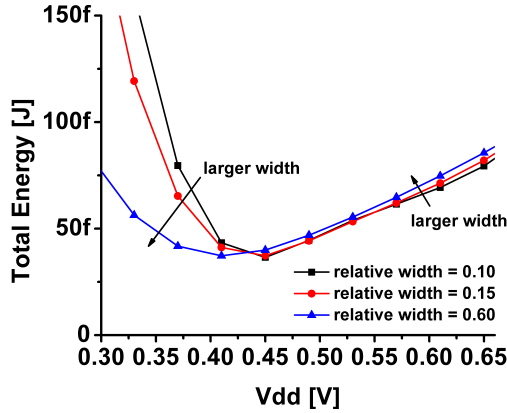


Figure 4.6: V_{min} / E_{min} with different PGS sizes, $K_{duty}=100$

nearly constant with supply voltage in the ultra low voltage regime while it often increases in super-threshold regimes due to short-channel effects. Therefore, we use a lumped coefficient, K_3 , for simplicity in EQ 4.7.

On the other hand, active energy consumption has a complex relationship with supply voltage and PGS width. First, PGS width affects the performance of circuits. For example, small PGSs (i.e., larger $1/K_{delay}$) induce longer delay, resulting in higher E_{leak} consumption in circuits. In near-threshold regimes ($V_{DD} > 450\text{mV}$ for this technology), the increase in E_{leak} is relatively small, while it can significantly increase total energy consumption in sub-threshold regimes due to the importance of E_{leak} , as shown in Figure 4.6.

The effect of supply voltage on active energy consumption is similar to the traditional analysis [113]. Lowering V_{DD} causes performance degradation and thus leads to higher E_{leak} consumption (i.e. higher $1/K_{delay}$), while it quadratically reduces E_{switch} .

One interesting observation is that large values of $1/K_{delay}$ or E_{leak} can be alleviated by either using larger PGSs or raising supply voltages. However there is a difference between these approaches. Using larger PGSs reduces the voltage drop across PGSs, leading to lower active energy consumption compared to raising supply voltage. However, raising supply voltage is more effective in improving performance

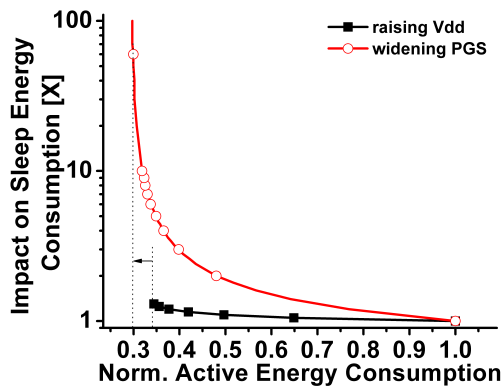


Figure 4.7: Comparison between raising V_{DD} and upsizing PGS in energy optimization

with a smaller increase in sleep energy consumption. To confirm these trends, we perform SPICE simulations where circuits initially have excessive E_{leak} consumption that must be alleviated using either of the discussed methods. Figure 4.7 shows that both raising V_{DD} and widening PGS can reduce active energy consumption but with differing impacts on sleep energy consumption. The larger PGS increases sleep energy consumption by $30\times$ while raising supply voltage incurs only a 25% penalty. Given the advantage of widening PGS is improved active energy consumption compared to raising V_{DD} , this approach should be used in cases of small K_{duty} , where active energy is more important than E_{sleep} , which will be confirmed in Section 4.5.1.

4.5 Strategy of Design Power Gating Switches

4.5.1 PGS Design Strategies in Ultra Low V_{DD} Regimes

This section presents a strategy for using PGSs in ultra low voltage regimes based on the findings in Section 4.4. We first review the conventional methods of designing PGSs. Then, we propose our PGS design method employing co-optimization in ultra low voltage regimes. In this method, supply voltage and PGS width are simultaneously chosen to achieve full energy savings at a given duty cycle. For the designs

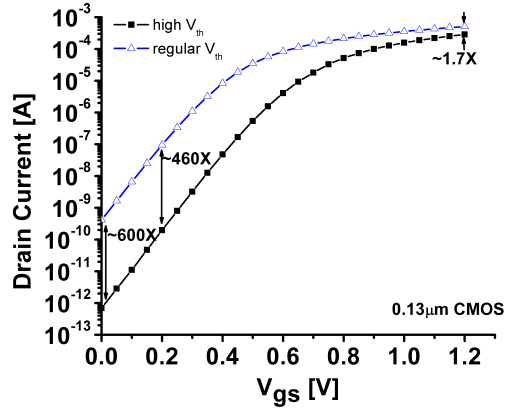


Figure 4.8: On/off-current of high V_{th} and regular V_{th} devices

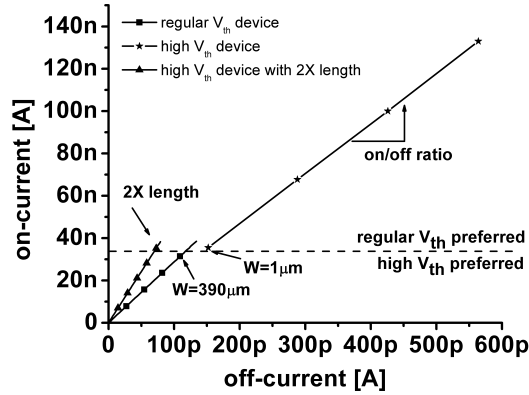


Figure 4.9: Off-current vs. on-current as sweeping PGS width

targeted at nominal V_{DD} operations the performance degradation is often constrained by less than 5-10%. Therefore, the width of PGSs needs to be large enough to supply proper current and minimize virtual ground bounces. Often, the constraints lead to large PGS width, often 10% of total NFET width of main circuits [67, 45].

Also, high V_{th} devices have been a popular choice for PGSs since they have similar on-current but much smaller off-current than regular V_{th} devices. Figure 4.8 shows that in this technology, high V_{th} devices have 600 \times smaller off-current, while they have only 1.7 \times smaller on-current at $V_{DD}=1.2V$. Therefore, high V_{th} PGSs can provide 352 \times reduction in off-current at the same on-current.

In ultra low voltage regimes, PGS design can be different. High V_{th} devices become

less attractive since they have the similar on-current to off-current ratio as regular V_{th} devices in ultra low voltage regimes. Here on-current is defined as saturation current since the V_{ds} required for device saturation is only $3-4V_t$ in ultra low voltage regime. Using high V_{th} PGSs is beneficial only for the case where circuits draw a current smaller than what a minimum-sized regular V_{th} PGS can deliver. Figure 4.9 shows that circuits with current of less than 30nA can exploit high V_{th} PGSs for the targeted technology. For the higher current draw, regular V_{th} devices are preferred due to an unnecessary use of area by high V_{th} PGSs. The crossover point between regular V_{th} and high V_{th} PGS is technology-dependent, thus requiring careful evaluations for each technology.

In this sense, devices with a large on-current to off-current ratio are preferred for PGSs in ultra low voltage regimes. One way of improving the ratio is to use longer channel devices [53], as shown in Figure 4.9. Note that in this particular technology, high V_{th} devices exhibit a slightly better on-current to off-current ratio than regular V_{th} devices. However, since the ratio is technology-dependent, a careful evaluation is needed for each technology.

Another important factor to consider is that the conventional practices of sizing PGSs for maintaining performance is no longer valid since minimizing total energy consumption is a more important goal for ultra low power applications. Therefore, PGSs should be optimized for minimizing total energy consumption. Since both PGS width and supply voltage affect total energy consumption, as we discuss in Section 4.4, we propose an optimization method, called co-optimization, for designing PGSs. In this proposed method, PGS width and supply voltage are simultaneously selected for minimizing total energy consumption.

We investigate total energy consumption at different duty cycles by sweeping all combinations of PGS widths and supply voltages in the SPICE simulations using inverter chains. If K_{duty} is equal to one, then the optimal energy consumption can

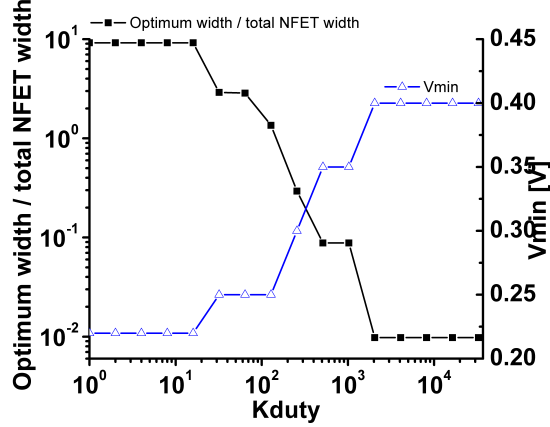


Figure 4.10: New V_{min} and optimal PGS size at different K_{duty}

be achieved by supplying the conventional V_{min} without PGSs. This is because PGSs induce extra delay and more E_{leak} consumption. Since there is no sleep time, i.e. $K_{duty} = 1$, the sleep leakage reduction is of no use in this case. The results are shown at the left end of Figure 4.10.

When K_{duty} falls roughly between 1 and 100, the optimal V_{DD} is similar to the conventional V_{min} and the optimal PGS width becomes large. These relatively small values of K_{duty} imply that E_{sleep} is small. Therefore, the increase in E_{sleep} caused by the use of larger PGSs is a negligible part of total energy consumption. This is well matched to the idea expressed in Section 4.4, that increasing PGS width is more energy-efficient than raising V_{DD} when sleep time is small. This is well supported by SPICE simulations using inverter chains, as shown in Figure 4.10. If the large PGS causes too much area overhead, it can be omitted with a relatively small sleep energy penalty.

When $K_{duty} > 100$, small PGSs and $V_{DD} > V_{min}$ are preferred for minimizing total energy consumption since raising V_{DD} imposes a lower penalty on E_{sleep} , as discussed in Section 4.4. This is confirmed by SPICE simulations using inverter chains, as shown in Figure 4.10. The small PGSs force the effective voltage between virtual rails to approach conventional V_{min} .

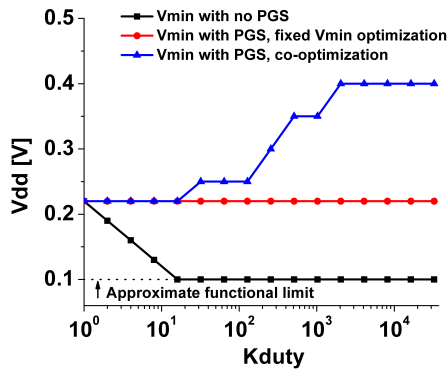
Typical sensor-type applications have K_{duty} of 104 [70]. Therefore, to achieve optimal energy consumption, the regular V_{th} PGS can be downsized to 0.01% of total NFET width of main circuits, as shown in Figure 4.10. However, since 0.01% of total NFET width is smaller than the minimum width of device in this technology, a high V_{th} PGS is instead used. For the same on-current, the high V_{th} PGS should be sized at 1% of total NFET width of the main circuits. As stated earlier the logic depth and switching activity of the test circuits incur worst case voltage drop across PGSs. Since longer length or less activity reduces the current delivery requirement, optimized PGSs can be made even smaller in many practical settings.

4.5.2 Comparisons of the Optimization Methods

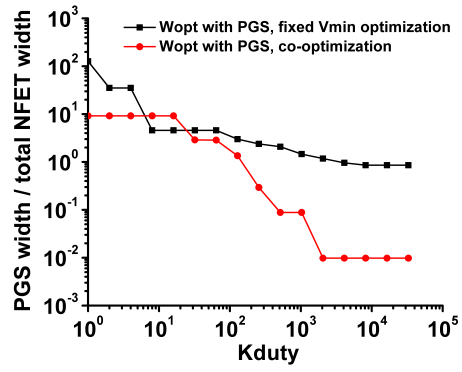
We run SPICE simulations using inverter chains to compare our proposed co-optimization with two baseline approaches for designing PGSs. The first baseline approach is to use no cutoff structure and optimize supply voltage only. The second baseline approach, referred to as fixed- V_{min} -optimization, uses PGSs at a conventional fixed V_{min} . Figure 4.11(a) shows the change of V_{min} for each strategy. It illustrates that the co-optimization calls for a higher V_{DD} than the conventional V_{min} for large values of K_{duty} . However, the V_{min} is scaled down to the functional limit of supply voltage that allows the task to be completed in a given time (K_{duty}) for the no-cutoff approach.

On the other hand, Figure 4.11(b) illustrates the optimal PGS width for each optimization approach. The co-optimization suggests the use of extremely small PGSs for energy optimization. However, the fixed V_{min} -optimization cannot suggest such small PGSs since they degrade performance and thus consume extra active energy at the fixed V_{min} .

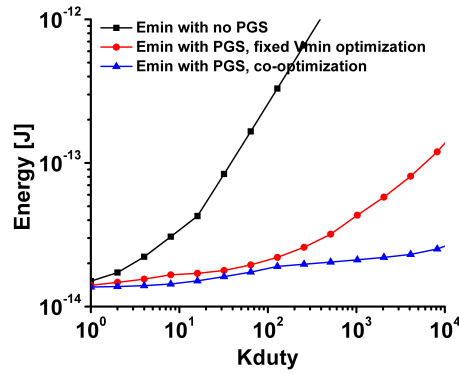
Finally, the total energy consumption of these strategies is compared in Figure 4.11(c). Even at relatively small values of K_{duty} , the no-cutoff strategy con-



(a)



(b)



(c)

Figure 4.11: Comparison of three optimization strategies, (a) $K_{duty}V_{min}$ (b) K_{duty} optimal PGS width (c) $K_{duty}E_{min}$.

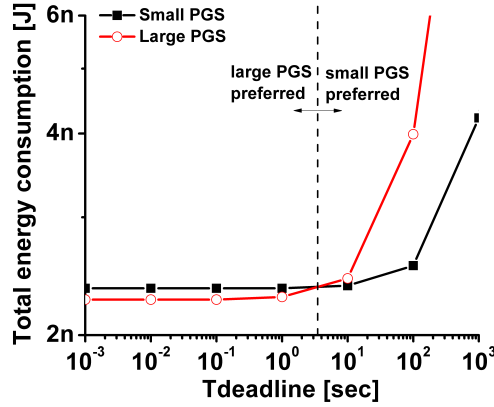


Figure 4.12: Measured total energy consumption with two different PGS sizes from a test microprocessor

sumes a significantly large amount of energy. The fixed- V_{min} optimization and co-optimization exhibit comparable energy consumption for small values of K_{duty} . However, co-optimization saves a considerable amount of total energy consumption when $K_{duty} > 1000$. Note that sensor applications often have K_{duty} larger than 1000. For these high K_{duty} applications, the co-optimization can save up to 99% of total energy consumption, compared to the other approaches.

4.5.3 Case Study Using a Fabricated Microprocessor

We apply the proposed design method to a microprocessor designed for ultra low power applications [83]. It is fabricated in $0.18\mu m$ CMOS and consists of 4000 gates. The total NFET width is $6000\mu m$. The microprocessor has tunable PGSs with widths ranging from $0.66\mu m$ to $28\mu m$ for mitigating the effects of process variations on PGSs. Using the smallest PGS, E_{active} is measured as 2.35pJ/cycle with $I_{sleep}=2\text{pA}$. The processor operates at 60 kHz with $V_{DD}=0.475\text{V}$. For the smallest and the largest PGSs, we measure active energy consumption and sleep energy consumption. We estimate that 1000 instructions are executed during active mode. We then calculate the total energy consumption at several values of K_{duty} .

Figure 4.12 shows that as sleep time become small (i.e., larger K_{duty}) the ideal

strategy transitions from using the widest ($28\mu m$) PGS to employing the $0.66\mu m$ PGS. The large PGS is slightly more energy efficient at high duty cycles due to less performance degradation and smaller voltage drop across the PGS. However, the small PGS becomes energy-optimal at low duty cycles since sleep energy consumption represents a large portion of total energy consumption. These strategies cross over when $T_{deadline}$ is 4 seconds. Since $T_{deadline}$ for most ultra low power systems is larger than 4 second [70], small PGSs are energy-optimal for these applications. If 1000 seconds (16 minutes) sleep time is assumed, the small PGS provides $4.6\times$ lower total microprocessor energy consumption compared to the large PGS. We cannot measure T_{min} of the microcontroller (i.e., the microprocessor delay at V_{min} without PGSs), therefore we approximate it as the delay at V_{min} with the large PGS. With the estimated T_{min} , the K_{duty} for 10 seconds is 106.

4.6 Feasibility of Minimal-Sized PGSs

Even if performance degradation is ignored, designers are unlikely to view extremely small PGSs as viable options since the voltage drop across PGSs may cause functional robustness problems. In super-threshold regimes, it is true that the small PGSs cause functional failures. Figure 4.13 shows that the microprocessor discussed in Section 4.5 is not functional with the small PGSs at $V_{DD} > 0.8V$. However, in ultra low voltage regimes, the microprocessor with the small PGSs is functional. Therefore, it is important to understand the different feasibilities of small PGSs in ultra low voltage regimes.

One reason that the small PGS functions well in ultra low voltage regimes can be found in the relationship of V_{ds} and subthreshold current. As shown in EQ 4.1, subthreshold current becomes insensitive to V_{ds} once V_{ds} is larger than $3.4 V_t$. In other words, even if the microprocessor attempts to draw a large current, for example, because of many simultaneous internal node switches, the V_{ds} or voltage drop across

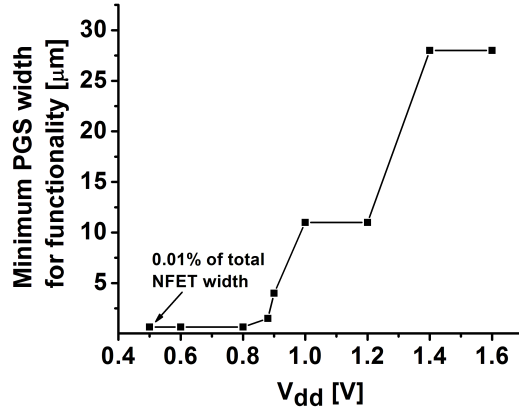


Figure 4.13: Measured minimal PGS size for functionality

the PGS changes only by a small amount. Instead, the current draw is limited and the microprocessor is slowed. However, linear and saturated current of devices in super-threshold regimes have a linear relationship with V_{ds} . Therefore, the V_{ds} of the PGS quickly rises to the point at which the PGS can supply a large current. This V_{ds} increase appears as a large virtual ground bounce, making the minimal PGS less robust in super-threshold regimes.

To confirm these concepts, we perform SPICE simulations with two different sets of inverter chains. The first set has one inverter chain that is switching and four chains that are not switching. The second set has five inverter chains that are switching. Each inverter chain is identical, thus the second set draws $5\times$ higher current draw. We investigate voltage drops across PGSs for these circuits PGSs are sized at 0.05% of total NFET width for each set.

Figure 4.14 illustrates that relative virtual ground levels are smaller for ultra low voltage regimes for both low and high work load cases, which is expected, given the different relationships of V_{ds} with drain current in two different voltage regimes. Additionally, in ultra low voltage regimes, the relative increase of the virtual ground level from low to high work load is smaller. The final observation is that the relative virtual ground level goes up at $V_{DD} < 0.4V$. This is because the V_{ds} of the PGS gets

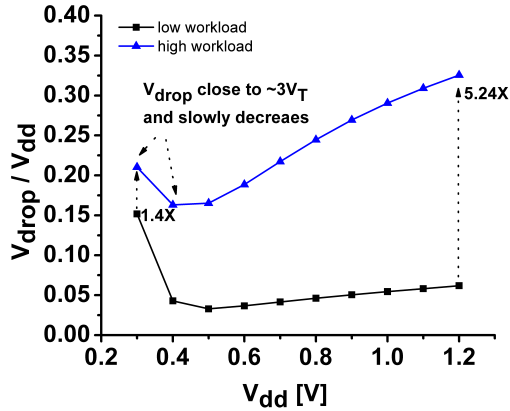


Figure 4.14: Simulated virtual ground level over different workload and supply voltage close to $3.4V_t$ and then decreases only slightly.

The 0.13 and $0.18\mu m$ technologies considered in this paper exhibit less process variations than leading-edge scaled technologies. In such cases robustness can be improved by using a wider PGS at the cost of sleep energy consumption [10]. To further mitigate process variations, trimmable PGSs such as those in [83] can be used for selecting appropriate width PGSs post-silicon to minimize sleep energy. Since robust operation is of critical importance, statistical simulations across PVT variations should be considered.

4.7 Beyond Basic PGSs

So far, we have discussed only the basic PGS topology. However there are many variations for PGSs to improve the fundamental tradeoff between performance degradation and sleep energy reduction. In this section, we quantitatively compare different flavors of PGSs and provide guidelines for choosing energy-optimal PGSs in ultra low voltage regimes.

Figure 4.15 shows three well-known PGS topologies: basic PGS, DTCMOS PGS, and stack-forcing PGS. In DTCMOS PGSs, the gate and the body of the PGSs are tied to increase on-current. Therefore, DTCMOS PGSs are expected to have a smaller

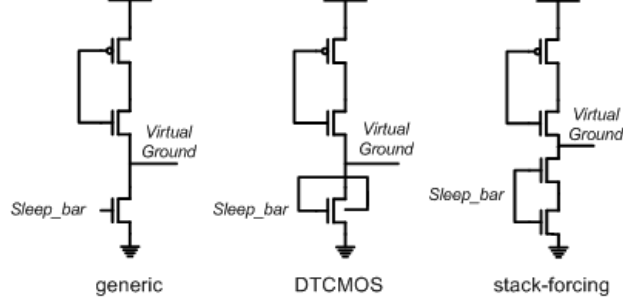


Figure 4.15: Generic, DTCMOS, and stack-forcing PGS

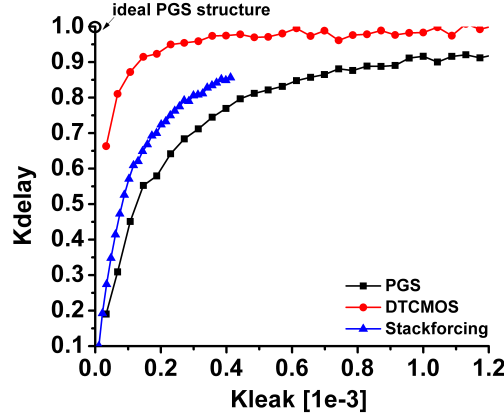


Figure 4.16: $K_{leak} - K_{delay}$ curves with different PGSs

$1/K_{delay}$, compared to the basic PGS. The stack-forcing PGS uses two FETs in series to reduce off-current using the stack effect [68]. These series-connected FETs induce negative V_{gs} at the upper FET, which exponentially decreases off-current. Therefore, it exhibits smaller K_{leak} than the basic PGS. However, $1/K_{delay}$ can be worse. At $V_{DD}=0.5V$, the $K_{leak}-K_{delay}$ curves of these structures are shown in Figure 4.16. For the same K_{leak} , the DTCMOS structure provides the smallest $1/K_{delay}$, and thus the smallest E_{leak} , followed by stack-forcing PGS.

Super-cutoff PGS [47] is not considered in comparisons since the penalty of generating bias voltages is difficult to quantify. However, it can be a promising design choice due to the exponential relationship between subthreshold current and supply voltage in ultra low voltage regimes. In [58], a detailed analysis on the tradeoff between generating bias voltages and sleep energy reduction is presented for ultra low

voltage operations.

4.8 Summary

This Chapter investigates the interaction of optimal energy, supply voltage and PGS for ultra low voltage designs. The results show that ignoring sleep leakage energy in ultra low voltage regimes can significantly degrade energy efficiency. Therefore, we propose several approaches for designing PGSs including co-optimization, which seeks to achieve optimal energy by simultaneously adjusting both PGS size and V_{DD} . Unlike typical practices in higher voltage regimes, in which large PGSs and nominal supply voltage are often chosen, our proposed optimization suggests using minimal PGS and higher V_{DD} for those applications with long sleep time. This reduces energy by $125\times$ in SPICE simulations. The effectiveness of the proposed method is confirmed by the silicon measurements from an ultra low power microprocessor. Finally, the feasibility of using minimal-sized PGSs in ultra low voltage regimes is studied with the focus of functional robustness using SPICE simulations and silicon measurements.

CHAPTER V

Robust Ultra Low Voltage ROM Design

5.1 Motivation and Previous Work

From the Phoenix Processor, we observe that SRAM for both data and instructions is the dominant source of total standby energy consumption in a typical sensing platform. It is therefore paramount to minimize the standby power consumption of the memories. Although most data memory (DMEM) must be both read and written, instruction memory (IMEM) can be re-optimized to take advantage of its read-only nature. For example, by storing common procedures in ROM with a power gating switch designed for ultra low voltage operation, both standby power and area can be reduced. Figure 5.1 shows that standby power can be reduced by 43% and area reduced by 10.7% in a sensing platform by replacing 128 out of 192 SRAM words with power-gated ROM.

However, there are four key challenges for designing robust ROM at ultra low voltages: 1) The reduced on-current to off-current ratio causes robustness problems, 2) there is potentially a large skew in beta ratio (relative strength between NFET and PFET) at low voltage, 3) for dynamic ROM styles, conventional keepers (half-latches) are likely to lose state and 4) significant variability further complicates each of the previous three issues.

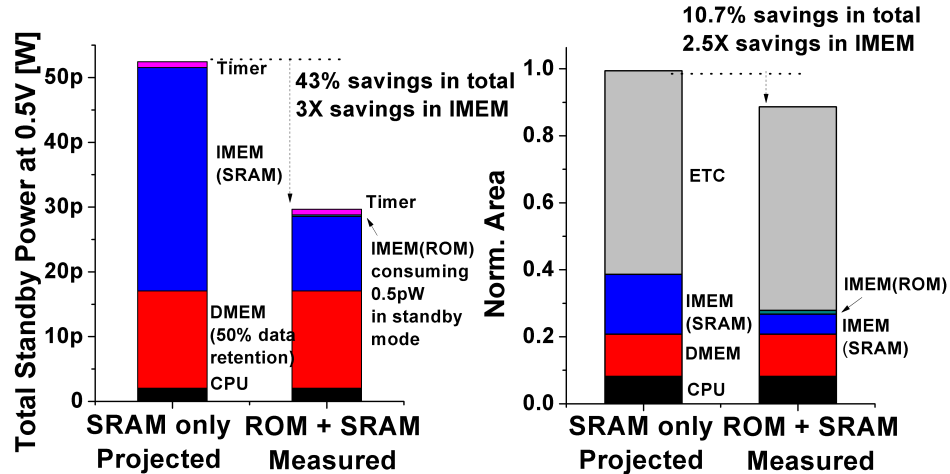


Figure 5.1: Power(left) and area(right) comparisons for SRAM-only IMEM (projected) and an SRAM/ROM hybrid IMEM (measured).

5.2 Contributions

We explore the design of ultra low voltage ROM. First we investigate the challenges of designing conventional dynamic NAND ROM at ultra low voltages and propose circuit techniques to overcome these challenges. We also propose a back-of-the-envelope method, referred to as a current margin plot, which estimates the theoretical functionality of ROM at ultra low voltages and provides guidelines for design decisions. We then propose two alternative ROM topologies, static NAND ROM and static NAND-NOR ROM, that improve robustness, performance, and energy-per-operation compared to dynamic NAND ROM. The current margin plot is used to estimate robustness for the two static ROM topologies. We conclude by describing a $0.18\mu\text{m}$ test chip that includes structures for each of the three ROM topologies discussed. The $0.18\mu\text{m}$ technology is chosen due to a superior balance between switching and leakage energy relative to more recent technologies. Measurements show that the static NAND ROM improves performance by 26X, energy by 3.8X, and minimum functional supply voltage by 100mV over a conventional dynamic NAND ROM.

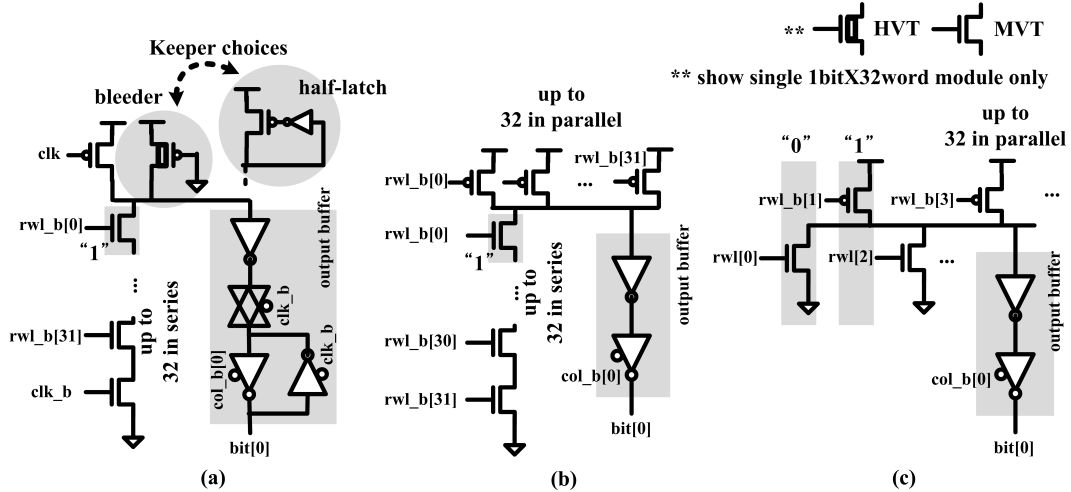


Figure 5.2: Schematics of three ROMs for ultra low voltage: (a) dynamic NAND, (b) static NAND, (c) static NAND-NOR.

5.3 Dynamic NAND ROM Design

This Section first investigates the challenges of ultra low voltage ROM design with a particular focus on the dynamic NAND ROM topology (Figure 5.2(a)), which is commonly used in superthreshold operation. Although dynamic NOR is also commonly used in superthreshold regime, it is not considered in this study due to the reason discussed in the Section 5.3.3. Then a method called a current margin plot is proposed to show theoretical robustness at ultra low voltages, which can be applied to any ROM topology. Using this method we describe the design of a dynamic NAND ROM targeting ultra low voltage operation.

5.3.1 Challenges of dynamic NAND ROM

The dynamic NAND ROM (Figure 5.2(a)) operates in two phases: precharge and evaluation. In precharge when clock is low, the dynamic node is charged up to V_{DD} . In evaluation when the clock is high, the dynamic node is either discharged to V_{SS} by stacked NFETs or held at V_{DD} by a half-latch keeper depending on the read word line signals. Having an NFET for a specific read word line means a high output value

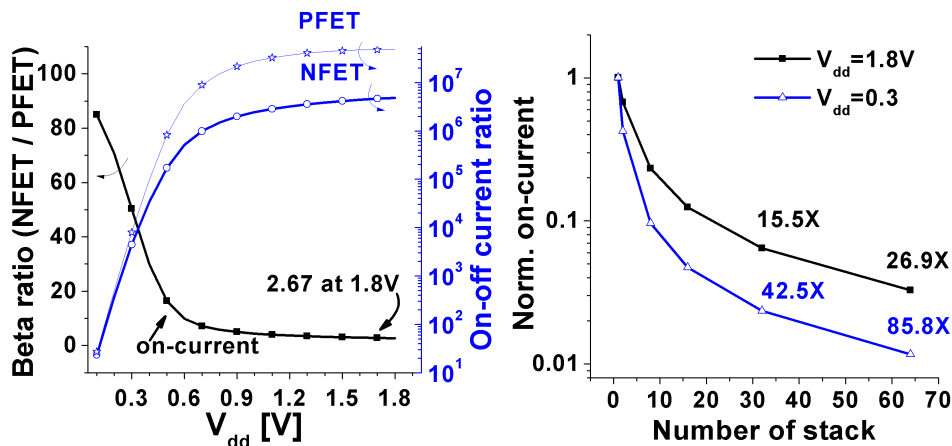


Figure 5.3: Beta ratio and on- to off-current ratio (left), on-current reduction over number of stack (for minimum-sized FET) (right)

since the NFET for the word line is turned off.

Operation becomes less robust in the ultra low voltage regime for several reasons. First, the on- to off-current ratio is reduced (Figure 5.3), resulting in robustness problems since on-current becomes less distinguishable from off-current. This problem is exacerbated in ROM design since ROM usually has a large number of FETs in series for NAND, or in parallel for NOR styles [39]. The FETs in series limit on-current while FETs in parallel increase the worst-case leakage current. As shown in Figure 5.3, the on-current decreases super-linearly as more FETs are connected in series, resulting in a worse on- to off-current ratio. In the technology used in this work, the on- to off-current ratio of 32-stacked NFETs is only 152X at 0.5V, which is several orders of magnitude smaller than at nominal voltage.

Additionally, the large skew in beta ratio can further aggravate this on- to off-current ratio problem. Beta ratio can be dramatically different between subthreshold and superthreshold since on-current is exponentially dependent on V_{th} and devices are typically optimized for superthreshold operation. In this technology, the min-sized NFET is 20X stronger than the min-sized PFET at $V_{DD}=0.5V$, compared to 2.7X at $V_{DD}=1.8V$, as shown in Figure 5.3. Therefore the ratio of PFET on-current to NFET

off-current is reduced by roughly 20X in addition to the already-reduced on-current to off-current ratio of ultra low voltage operation.

The functionality of the half-latch keeper (Figure 5.2(a)) in the ultra low voltage regime is another problem for dynamic ROM design. The half-latch becomes more important at low voltages since the charge on dynamic nodes is reduced linearly with scaled supply voltage while leakage current stays almost constant. The half-latch is not able to maintain its state for two reasons. First, its retention ability is reduced. When we view the half-latch keeper as broken back-to-back inverters in SRAM, its static noise margin is known to degrade in ultra low voltage regime. Second, the same amount of charge sharing has a more destructive effect at low voltages. Finally, large variability further complicates low voltage dynamic NAND ROM design. Simulations show that if NFET off-current is larger than nominal by 1σ and PFET on-current smaller than nominal by 1σ due to process variations, the total on- to off-current ratio is reduced by 4.8X at 0.5V.

5.3.2 On-current to off-current plot

In this Section a back-of-the envelope method, a "current margin plot", is described to estimate the theoretical robustness of ROM in the ultra low voltage regime. All the factors mentioned in the previous Section are accounted for in the method. Here we apply it to a 32-stack dynamic NAND ROM with a high- V_{th} PFET bleeder with length of $0.45\mu\text{m}$ and width of $0.33\mu\text{m}$ (Figure 5.2(a)). Figure 5.4 shows the margin of on- to off-current ratio in the 32-stack dynamic NAND ROM with bleeder for two different operations at different voltages: 1) evaluation of a one (eval-1) and 2) evaluation of a zero (eval-0). The eval-1 (left side) is the case where the output is maintained by the bleeder. Here the worst case off-current through the NFET stack should be smaller than the current that the bleeder provides to guarantee functionality. A guardband equal to the standard deviation of off-current is included to

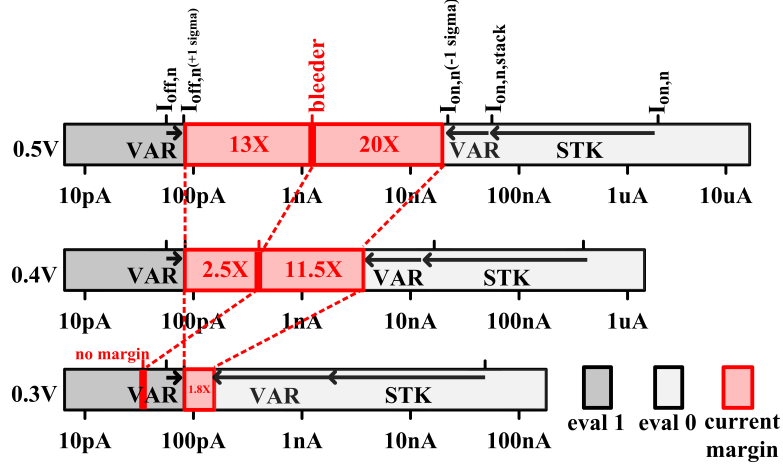


Figure 5.4: Current margin plot for 32-stack dynamic NAND ROM with HVT bleeder estimate worst-case off-current and is denoted as VAR in Figure 5.4. On the other hand, complete discharge through stacked NFETs is required for eval-0 (right side of Figure 5.4). Here the discharging current through NFETs should be larger than the bleeder current. However the discharging current is reduced by the series connection (STK) and guardband for process variation (VAR), resulting in a range of just 20X between the stack and the bleeder at 0.5V.

The magnitude of the guardbands for FET current due to process variation (VAR) is determined based on the standard deviation taken from 1000 Monte Carlo simulations. Here, the log of current is assumed as a Gaussian distribution, which is based on the fact that subthreshold current is an exponential function of normally distributed V_{th} , which is the dominant factor for current variation [verma08] at ultra low voltage. One thing to note here is that variation in off-current is almost constant while on-current variation is increased with smaller supply voltage, which is considered in the current margin plot. We set VAR based on the current at $\mu \pm \sigma$ of V_{th} . The current reduction due to the large stack (STK) is based on simulation results (Figure 5.3). The margin of on- to off-current ratio provides information about two circuit metrics: robustness and performance. Clearly larger margins offer more robustness in light of substantial process variation. In addition to robustness, the margin dictates circuit

performance. For instance, a large margin for the eval-1 case implies fast recovery from signal degradation of dynamic nodes. A large margin between discharging current and bleeder current is preferred for reduced contention. The margin of the on- to off-current ratio is diminished as the supply voltage is scaled down. Since process variation in the bleeder current can further degrade the margin, robustness at ultra low voltage for this ROM topology is questionable.

5.3.3 A 32-stack dynamic NAND ROM with HVT bleeder

In the previous Section, a 32-stack dynamic NAND ROM with a bleeder is used to illustrate the current margin plot. This Section discusses the design decisions regarding ROM topology, stack height and keeper style, which collectively point to a 32-stack dynamic NAND ROM with bleeder as a reasonable design choice.

First, dynamic NAND is chosen over dynamic NOR ROM due to the large skew in beta ratio. Consider a 32-stack dynamic NOR ROM. The off-current through 31 parallel-connected NFETs is only 20X smaller than the on-current of a single PFET serving as keeper at $V_{DD}=0.5V$, even without considering process variation. In addition, this reduced on- to off-current ratio degrades performance, which is one of the primary advantages of the NOR topology. Larger leakage power and larger footprint compared to the NAND topology are other drawbacks. Therefore the dynamic NAND structure is chosen over a NOR topology in this study. Note that this decision is motivated primarily by technology limitations. A different technology may make the NOR structure more attractive.

Second, given a dynamic NAND structure, we select a stack height of 32. A taller stack reduces area but can cause robustness problems due to small discharging currents and significant charge sharing. A stack of 64 devices would reduce the margin of on- to off-current ratio between the bleeder current and the worst case on-current by 2X compared to a stack of 32 devices.

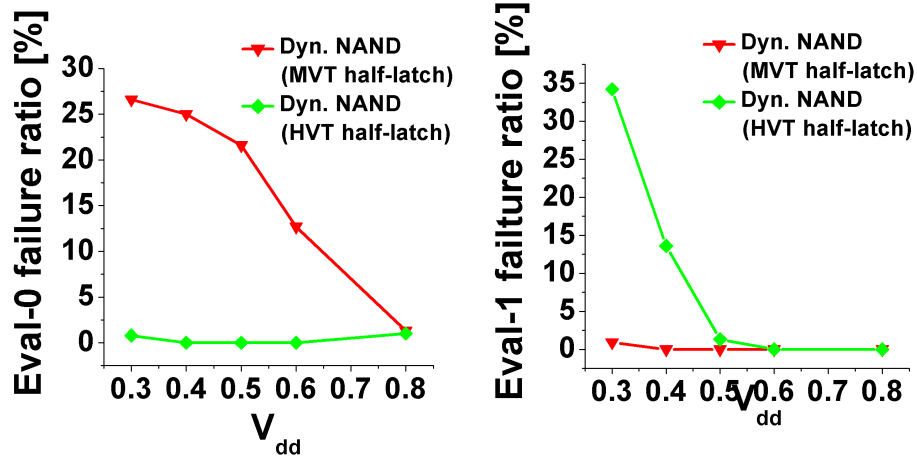


Figure 5.5: Failure rate for the dynamic NAND with half-latches. (1000 Monte Carlo iterations with die-to-die and mismatch variations)

Finally, a HVT bleeder is chosen over the half-latch keeper configuration. Monte Carlo simulation with die-to-die and mismatch variations shows that two half-latches with different strengths, (a medium V_{th} (MVT) device with $W/L=0.33\mu\text{m}/1.8\mu\text{m}$ and an HVT device with $W/L=0.33\mu\text{m}/0.45\mu\text{m}$) fail to discharge (eval-0) or hold (eval-1) dynamic nodes in the ultra low voltage regime as shown in Figure 5.5. The MVT half-latch often becomes so strong that series-connected NFETs are unable to discharge the dynamic node while the HVT half-latch is often unable to supply enough charge to overcome charge sharing. However the HVT bleeder operates more robustly than the half-latch in the ultra low voltage regime. An HVT bleeder provides the same on-current as the HVT half-latch for eval-0, so the series connected NFETs are able to discharge the dynamic node. Additionally, the bleeder constantly provides current even after the dynamic node accidentally pulls low, so the correct value will eventually be restored in contrast to a half-latch.

Setting the appropriate strength of the bleeder is important due to the tradeoff between recovery and contention. The margin of on- to off-current in Figure 5.4 specifies the available strength that the bleeder can have. If the bleeder strength resides outside the margin, it can cause incomplete discharge for eval-0 or poor recovery for

eval-1. However setting bleeder strength is a nontrivial task. While keeper strength is adjusted at nominal V_{DD} through gate sizing, it cannot be applied in an area efficient manner in the ultra low voltage regime due to the exponential variability in current. As shown in Figure 5.4, if a MVT device is used as a bleeder, the on-current is reduced by 3 orders of magnitude to reside in the allowed margin, requiring infeasible length biasing. Therefore other knobs such as using different V_{th} devices or applying body bias should be considered. We use a different V_{th} in this work to avoid generating an extra body bias.

5.4 Static NAND AND NAND-NOR ROM

Although we have shown that dynamic NAND ROM can operate at very low voltages, the performance, energy-per-operation and minimum functional voltage are unsatisfactory due to the small current margin. The bleeder is among the most important components of this design style, which gives rise to a challenging sizing tradeoff. Therefore, new topologies without a bleeder are worth investigating. In this Section, we describe the design of 2 full-static ROMs: 32-stack NAND and 32-leg NAND-NOR as shown in Figures 5.2(b) and 5.2(c). Since these structures have no bleeder, a larger on- to off-current margin is expected.

5.4.1 Investigating static ROM topologies

This Section applies the current margin plot to the two static ROM topologies to investigate theoretical robustness at ultra low voltages. Figure 5.6 shows the margin of on- to off-current ratio for the 32-stack static NAND ROM. Only the eval-0 case is considered here since it is the most stringent for large NFET stacks. Larger margin is still observed after incorporating the effect of parallel connection of PFETs (denoted as PAR), series connection of NFETs, and guardbands for process variation. The avoidance of the bleeder also helps increase the margin and ease design. Overall a

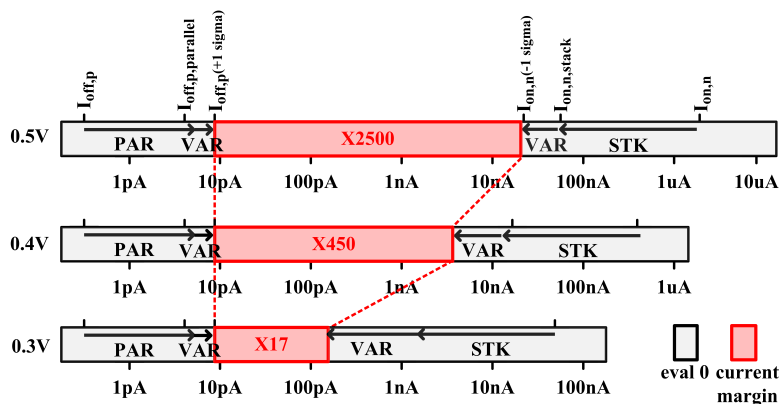


Figure 5.6: Current margin plot for 32-stack static NAND ROM

17X margin is maintained at a very aggressive V_{DD} of 0.3V, compared to zero margin for the 32-stack dynamic NAND ROM described earlier.

If NFET and PFET strengths are balanced, the static NAND ROM can be improved by replacing the long stack with parallel legs using inverted input signals as shown in Figure 5.2(c). However since the technology used in this study has a large beta ratio at low voltages, this topology is less robust than the static NAND ROM. Figure 5.7 shows the much reduced on-off margin in the eval-1 case where a single PFET contends with 31 NFETs, which is same as in a dynamic NOR topology. The current margin disappears at 0.3V, as in the dynamic NAND ROM. However, better robustness is expected over dynamic NAND ROM since no bleeder is present. Although the NAND-NOR ROM topology is not ideal in the technology used in this study, it may be a good candidate for technologies with more balanced beta ratios at low voltage.

5.4.2 Static NAND ROM Monte Carlo Analysis

Since the current margin plot is a first-order method to estimate robustness, Monte Carlo simulations considering all sources of variation are performed to investigate the effectiveness of the plot as well as the robustness of the ROM topologies. As shown in

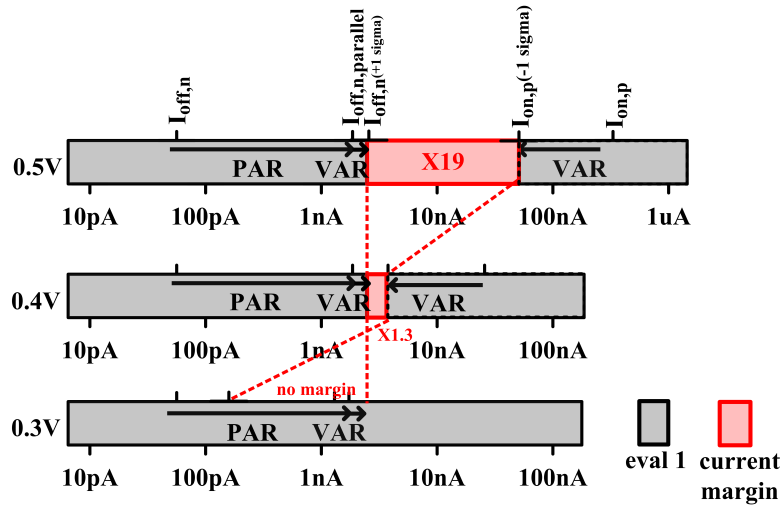


Figure 5.7: Current margin plot for 32-leg static NAND-NOR ROM

Figure 5.8, the static NAND ROM starts to fail at 0.3V in the eval-0 case due to the large NFET stack. The failure voltage is higher than that estimated by the current margin plot since the latter considers only one standard deviation of variation. In comparing topologies, the minimum functional voltage for the static NAND ROM is larger than that for the dynamic NAND ROM by nearly 200mV, confirming that the current margin plot is able to track the trends as well as the static NAND ROM is more robust.

5.5 Measurement Results

A 10x128bit dynamic NAND ROM with HVT bleeder, a 10x128bit static NAND ROM, and a 10x128bit static NAND-NOR ROM were fabricated in a 0.18 μ m CMOS technology. Each ROM contains an identical set of random data patterns as well as patterns causing worst case charge sharing. The worst case pattern for the dynamic NAND ROM and the static NAND ROM is 31 series-connected NFETs while the worst case for the static NAND-NOR structure is a single PFET connected to 31 parallel-connected NFETs. Relevant silicon measurements and dimensions are shown

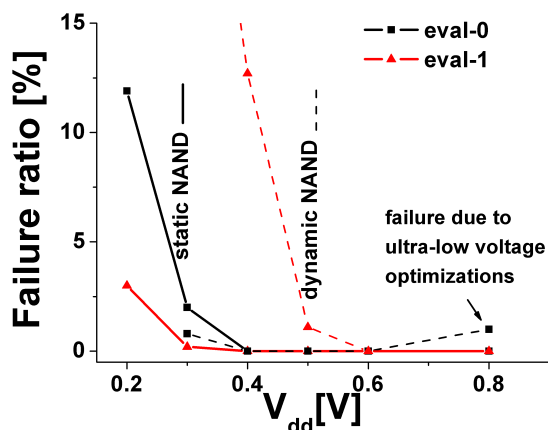


Figure 5.8: 1000 Monte Carlo SPICE simulations for two ROM topologies considering mismatch and die-to-die

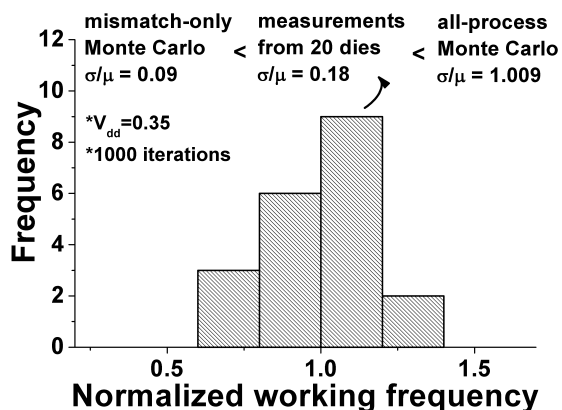


Figure 5.9: Histogram of operating frequency of static NAND ROM

in Figure 5.10 and 5.11.

The two static topologies show dramatically improved maximum operating frequency, energy-per-operation, and minimum functional voltage compared to the dynamic NAND ROM. The small on- to off-current ratio degrades performance in the dynamic NAND ROM, leading to substantially lower performance. The static NAND ROM and the static NAND-NOR ROM show similar energy, performance and minimum operating voltage numbers, though the static NAND ROM has a small advantage over the static NAND-NOR ROM, as predicted in previous Sections. Figure 5.9 shows the effect of variability on the performance of the static NAND ROM. The

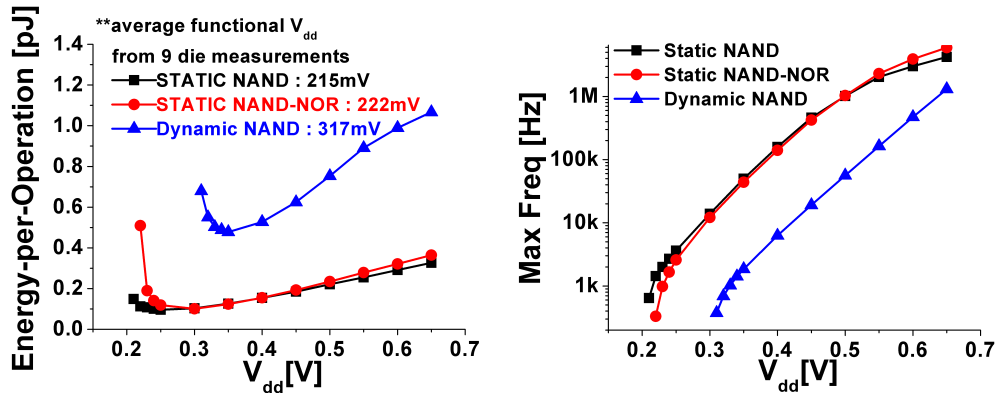


Figure 5.10: Measured energy-per-operation and frequency

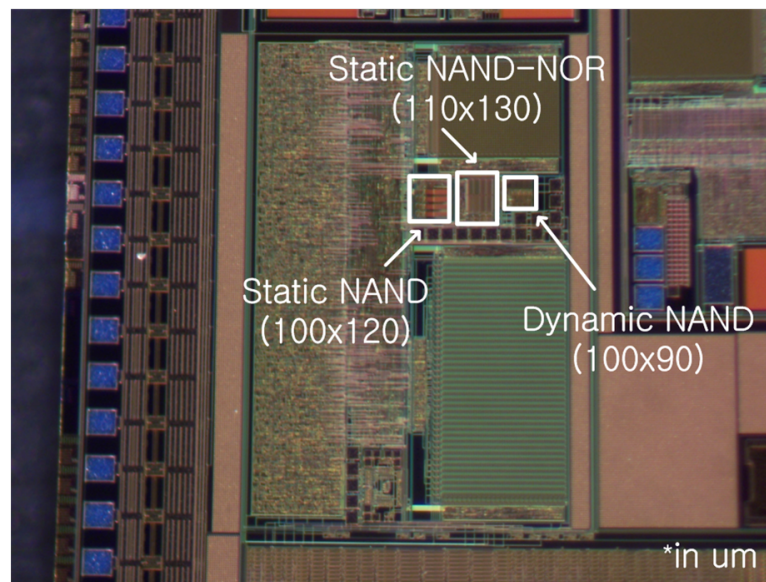


Figure 5.11: Die photo and dimensions

variation in maximum operating frequency (σ/μ) across 20 dies at 0.35V is just 18%. This number falls between the bounds set by a Monte Carlo simulation that includes die-to-die and mismatch variation and another Monte Carlo simulation considering mismatch only. Since all the 20 dies come from a single wafer, it is not surprising that the relatively small measured variability is much closer to the variability predicted by the mismatch-only simulation.

5.6 Summary

Three different ROM topologies for ultra low voltage operation are investigated with the test chip fabrication in an industrial $0.18\mu\text{m}$ CMOS technology. The challenges in ultra low voltage design are analyzed and incorporated in the current margin plot which is devised for estimating theoretical low voltage robustness. Silicon measurements shows that the static NAND ROM shows 26X faster performance, 3.8X smaller energy-per-operation and 100mV smaller minimum working voltage than the dynamic NAND ROM with 33% area penalty. The static NAND-NOR ROM is also studied as a potential candidate for other technologies.

CHAPTER VI

Pico-Watt 2-Transistor Voltage Reference with Digital Trimmability

6.1 Motivation and Previous Work

Sensor systems like Phoenix Processor often need to include analog and mixed-signal modules such as linear regulators, Analog to Digital (A/D) converters, and Radio Frequency (RF) communication blocks for self-contained functionality. Voltage references are key building blocks for these modules. In particular, linear regulators require a voltage reference to supply a constant voltage level to the entire system. Also, amplifiers in A/D converters employ several bias voltages. Therefore, it is often necessary to incorporate multiple voltage reference circuits in a system as a key building block.

However, integrating voltage references in sensing systems has posed new design challenges. Since sensing systems often need to consume less than hundreds of nano-watts due to limited energy sources, voltage references, as well as other modules, need to consume as little energy as possible. The footprint should be minimized as well, particularly for biomedical applications for less invasive surgery. Additionally, it is preferable that voltage references be able to operate across a wide range of supply voltages, in particular near or below 1V, since power sources such as energy scavenging

units provide only low output voltages. The restrictions on power consumption and area can easily be doubled, tripled, or more, since many voltage references are likely to be integrated in a system.

There are currently several approaches to designing voltage references in Complementary Metal-Oxide-Semiconductor (CMOS) technology [89, 4, 14, 106, 60]. The most common method is a bandgap voltage reference using parasitic BJTs (Bipolar Junction Transistors) [89, 4, 14, 106]. In order to generate temperature insensitive output voltage, bandgap voltage references linearly combine two voltages of opposing temperature characteristics: a complementary-to-absolute-temperature (CTAT) voltage and a proportional-to-absolute-temperature (PTAT) voltage. Another method is to combine PTAT and CTAT currents, rather than voltages, to generate a temperature-independent output voltage [60, 30, 12, 6]. We can also design voltage references by employing two devices of different threshold voltages, which can be implemented by distinct gate doping [71] or selective channel implanting [98, 8]. Alternatively, we can achieve stable output voltage based on the finding that the weighted difference between gate-source voltages of two complementary MOS transistors is temperature insensitive [61]. Another approach is to use subthreshold-biased transistors to lower minimum functional supply voltage and power consumption [29, 33, 3]. Finally, we can store and refresh reference voltage at a floating node to design voltage references [40, 92].

However, these voltage references hardly meet the demanding requirements of low power consumption, low functional supply voltage, and small footprint. Recently published ultra low power designs consume only tens of pico-watts during standby mode, while they often consume hundreds of nano-watts in active mode at $V_{DD}=0.3-0.5V$ [57, 83, 23, 42]. Therefore, voltage references for these systems should consume less power than the systems and also be functional at a similar range of supply voltages. However, current voltage reference designs are off the requirements, as

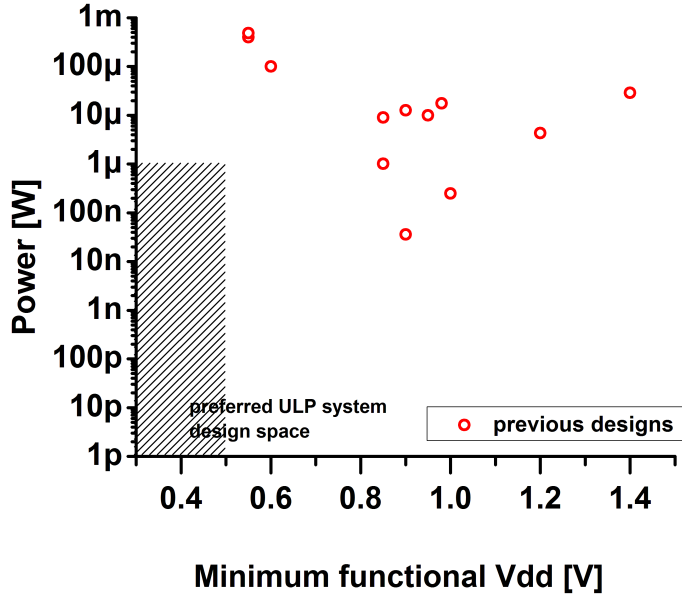


Figure 6.1: Power and minimal functional supply voltage comparisons

shown in Table 6.1 . Only one of them consumes less than tens of nano-watts, while most of them consume more than a micro-watt. Additionally, most of the designs fail to be functional below 0.8V. Figure 6.1 illustrates the large gap that exists between the requirements and the measured performance of the existing voltage references.

The less scalability of power consumption and minimum functional supply voltage comes from the fact that most of the design including bandgap references, resort

Table 6.1: Recent published designs of low power voltage references

Design	Min V_{DD} [V]	Power [μ W]	TC [$ppm/^{\circ}C$]	LS [%/V]	Area [mm^2]
[4]	0.9	12.6	962	n/a	0.016
[60]	0.98	17.6	15	3.6	0.24
[30]	0.95	10	17 (post trim)	n/a	1.09
[12]	0.85	9	33	4.4	n/a
[6]	2.4, 0.84 (sim)	4.6 at 2.1V	116	0.1	0.1
[71]	2	<2	300, 30 (post trim)	n/a	n/a
[98]	0.6	100 at 1V	37.7	n/a	0.06
[61]	1.4	29.1 at 3V, $100^{\circ}C$	36.9	0.012	0.055
[29]	0.9	0.036	10	0.27	0.045
[33]	1.2	4.3	119	n/a	0.23
[3]	0.85	1.02	57	n/a	0.063
[40]	1.0	0.25	16.9	0.76	0.049
[92]	3.3	3.3 (standby)	372	8.3	0.044
[54]	0.55	398	270	12.1	0.019
[54]	0.55	482	150	20.7	0.07

to amplifiers for error correction [89, 4, 14, 106, 60, 30, 12, 6, 71, 98, 8, 3, 40, 54]. Although the amplifier provides good temperature and supply voltage insensitivity, the associated power overhead is significant. Some of recent designs avoid amplifiers [61, 29, 33, 92]; however they often rely on MOSFETs in saturation mode, which consumes a significant amount of power. Both amplifiers and saturated devices require headroom, limiting supply voltage scalability. Additionally, many designs use integrated resistors for stable output voltages across temperature, which limits the scalability of the footprint of designs [89, 4, 14, 106, 60, 30, 12, 6, 71, 61, 33, 3, 92, 54].

6.2 Contributions

Therefore, we propose a voltage reference with no amplifier, no saturated device, and no resistor to meet the requirements of power, area, and minimum functional V_{DD} [86]. Figure 6.2 shows the proposed voltage reference using only two transistors, therefore called 2-Transistor (2T) voltage reference, which consumes as little as 2.2pW at $V_{DD}=0.5V$ and $25^{\circ}C$. This voltage reference offers improved power efficiency by 3-4 orders of magnitude compared to the previous state-of-the-art design [29]. The design uses subthreshold-biased devices with distinct V_{th} levels, i.e. one regular thick oxide and one native device for achieving a stable output voltage. In this case, the number of fabrication masks remains the same as the normal fabrication option. Since process variations can widen the spread in temperature coefficient (TC) and output voltage, we collected statistical results from the 2T voltage reference prototypes in two different runs of $0.13\mu m$ CMOS technology. Measurement results indicate that 2T voltage references exhibit considerable spread in TC and output voltage due to die-to-die and run-to-run process variations. [85]

Process sensitivity is a common problem for most voltage references and is typically addressed through trimming. However, trimming is often a time/cost intensive process, particularly if it involves laser trimming of resistors for bandgap voltage

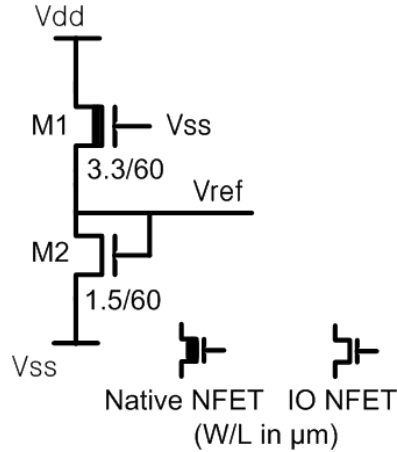


Figure 6.2: Schematics of a 2T voltage reference

references. Therefore, we propose a digitally trimmable version of the 2T voltage reference design to improve TC and output voltage accuracy across dies while reducing trimming time and cost [85]. Measurements from the prototype chip in $0.13\mu\text{m}$ show that trimming enables tighter distributions of TC and nominal output voltage across 25 dies. After economical one-temperature point digital trimming, the TCs lie between $13.5\text{ppm}/^\circ\text{C}$ and $47\text{ppm}/^\circ\text{C}$ while the nominal output voltage varies by 0.35% from the mean value. The voltage reference consumes 29.5pW at 0.5V and 25°C .

We also propose several variants from the 2T voltage reference. First, a variant to generate a specific output voltage, higher or lower than nominal values is implemented. Additionally, a version with temperaturedependent, either PTAT or CTAT, voltage reference is demonstrated.

Technology portability is also investigated. While most of the conventional voltage references requires considerable amount of modifications and characterizations in order to be ported to other technologies, the proposed designs involve only resizing of two transistors for technology porting.

6.3 Circuit Design

We observe that the amplifier and/or saturated MOSFET keep the scalability of power consumption and minimum functional V_{DD} . Therefore, if we can eliminate amplifiers and any saturated devices while maintaining voltage insensitivity against temperature and supply voltage in designing voltage references, it can not only enable supply voltage scalability but also significantly reduce power consumption and area. In this respects, we proposed a 2T voltage reference as shown in Figure 6.2. Two different device types are used: a native device for M1 and a thick oxide I/O device for M2. The native device is identical to a normal MOSFET with a near-zero V_{th} . Both devices have thick gate oxides to support high V_{DD} . Native mode devices are widely available in modern foundry technologies [22, 111]. In a process in which a threshold adjustment implant step is separately masked, the native device can be manufactured with no additional masking requirements [29]. One common usage of the native device is to limit Vds of thin oxide device by connecting in series, as shown in [63]. It also has been used in designing bandgap type voltage reference circuits [89, 6]. Although we have used a native device for M1, any combination of two devices with a considerable V_{th} difference can be used for the 2T voltage reference. The required V_{th} difference will be discussed shortly.

The output voltage Vref can be modeled by EQ 6.1, the well-known subthreshold current equation. Setting the current through M1 and M2 equal, EQ 6.2 will hold given that 1) both devices are in weak inversion, 2) Vds for M1 and M2 is greater than $3-4V_t$ (Thermal Voltage (V_t)), and 3) M1 follows the subthreshold current equation at Vgs down to -Vref.

From EQ 6.2, we obtain an analytical solution for Vref in EQ 6.3, where both the first term and the second term are either proportional or complementary to absolute temperature. Note that the V_{th} of MOSFETs is roughly complementary to the temperature [93]. By selecting the width and length of the two devices appropriately,

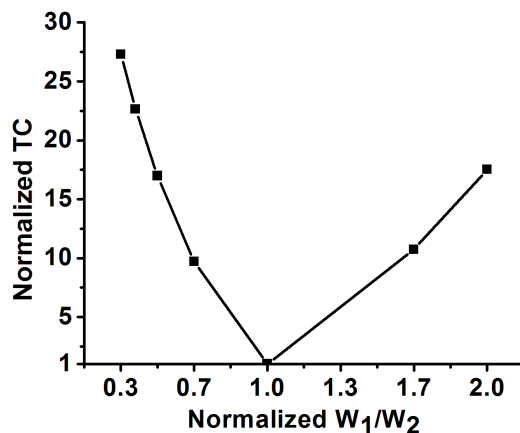


Figure 6.3: Proper sizing of two transistors minimizes temperature dependency (simulated results).

the temperature dependence of the two terms can be made to cancel out and temperature insensitivity can be obtained. The lack of a V_{DD} term in EQ3 leads to good line sensitivity and PSRR (Power Supply Rejection Ratio) without any amplifiers or active current sources. M1 decouples the output from the supply voltage, acting as a subthreshold cascode.

Sizing M1 and M2 in the 2T voltage reference aims to minimize both power consumption and temperature sensitivity. The longest gate length ($L_1=L_2=60\mu m$) allowed by the process design rules is used for both devices for ultra low power consumption, although shorter gate length can be used to reduce footprint and noise sensitivity if energy budget for voltage references is relaxed. The widths ($W_1=3.3\mu m$, $W_2=1.5\mu m$) are chosen to minimize temperature sensitivity. In selecting widths, the different characteristics (: mobility and m: subthreshold slope factor) of two devices must be considered. As shown in Figure 6.3, the optimal W_1 and W_2 balances out the temperature-dependent parts in EQ 6.3, resulting in little temperature coefficient.

We add a 0.8pF output capacitor for improving PSRR since coupling through the parasitic MOSFET capacitance can degrade PSRR. Simulated behavior in Figure 6.4 shows that larger output capacitance improves PSRR. The output capacitor also

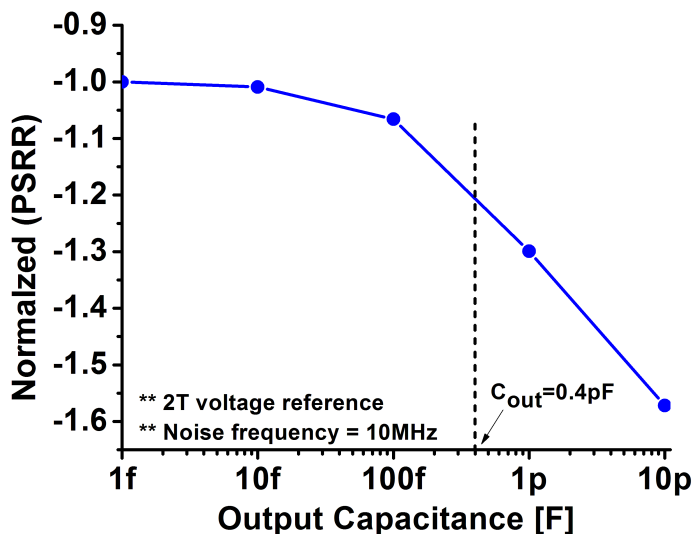


Figure 6.4: A larger output capacitor provides better PSRR (simulated results)

reduces the noise on output voltage. Since the subthreshold-biased devices exhibit large resistance, an output capacitor should be added to suppress thermal noise. The simulated results on output referred noise with different output capacitance values are shown in Figure 6.5. Both results from SPICE simulation and RC-filter noise power, ($P_{total}=kT/C$) [79] are plotted to confirm a good consistency.

The minimum supply voltage is limited by whether V_{ds} of M2 is larger than $3-4V_t$. If not, EQ 6.2 does not hold since the final V_{ds} term in EQ 6.1 cannot be neglected. On the other hand, the maximum supply voltage is set by reliability issues such as oxide breakdown. If necessary, diode connected transistors can be added between V_{DD} and M1 to increase the maximum supply voltage.

EQ 6.3 implies a design constraint on the required difference of V_{th} between the two devices (M1 and M2). Assuming typical subthreshold swing (90mV/dec) for the two devices (i.e., $m_1=m_2=1.5$), the minimum V_{th} difference is approximately $1.33V_{ref}$ from EQ 6.3 if we can neglect the second log term to the first order. Note that V_{ref} is equivalent to V_{ds} of M1, which should be larger than $4V_t$ as shown above (i.e. to neglect the final V_{ds} term in EQ 6.1). Hence, the minimum V_{th} difference is

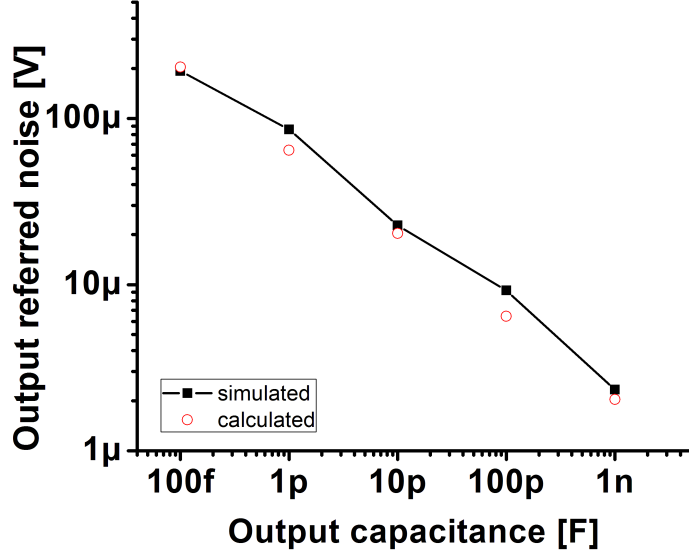


Figure 6.5: Simulated output referred noise of a 2T voltage reference with different output capacitors

approximately $5.3V_t$ for this type of voltage reference.

We also run SPICE simulations to confirm the required minimum difference in V_{th} . After changing the V_{th} of the M1 in the 2T voltage reference (Figure 6.2), we measure the degradation of temperature coefficient by sweeping the device widths. As shown in Figure 6.6, the temperature coefficient across -20 to $80^\circ C$, degrades quickly as the V_{th} difference becomes smaller than $250mV$. Since the V_t at $80^\circ C$ is $30.6mV$, the practical minimum required V_{th} for a good operation is about $8V_t$, larger than the first order estimation.

$$I_{sub} = \mu C_{ox} \frac{W}{L} (m - 1) V_T^2 \exp\left(\frac{V_{gs} - V_{TH}}{mV_T}\right) \cdot \left(1 - \exp\left(\frac{V_{ds}}{V_T}\right)\right) \quad (6.1)$$

$$\begin{aligned} I &= \mu_1 C_{ox1} \frac{W_1}{L_1} (m_1 - 1) V_T^2 \exp\left(\frac{0 - V_{ref} - V_{TH1}}{m_1 V_T}\right) \\ &= \mu_2 C_{ox2} \frac{W_2}{L_2} (m_2 - 1) V_T^2 \exp\left(\frac{V_{ref} - V_{TH2}}{m_2 V_T}\right) \end{aligned} \quad (6.2)$$

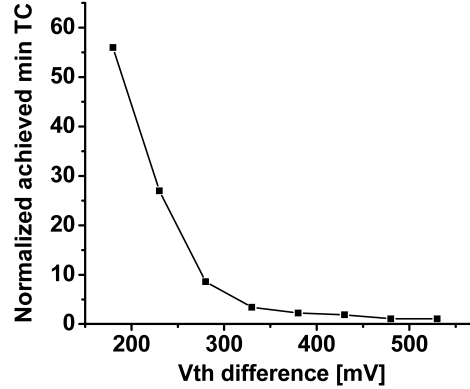


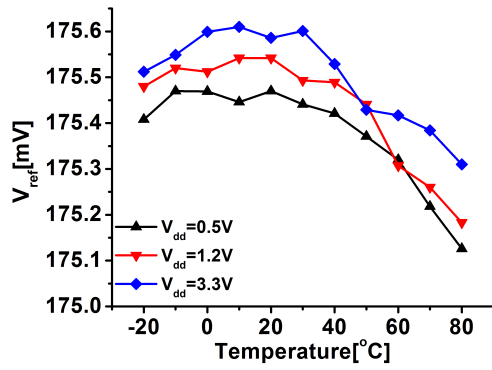
Figure 6.6: Simulated required V_{th} difference for proper operations

$$V_{ref} = \frac{m_1 m_2}{m_1 + m_2} \cdot (V_{TH2} - V_{TH1}) + \frac{m_1 m_2}{m_1 + m_2} V_T \ln\left(\frac{\mu_1 C_{ox1} W_1 L_2}{\mu_2 C_{ox2} W_2 L_1}\right) \quad (6.3)$$

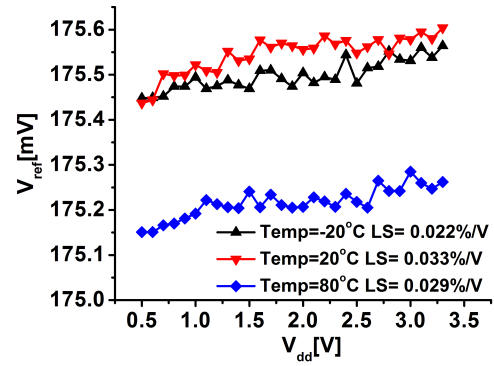
6.4 2T Reference Measured Results

We fabricated a test chip in a standard $0.13\mu\text{m}$ CMOS technology with no process options (i.e. thick oxide and native device are offered in a standard option). As shown in Figure 6.7(a), the 2T voltage reference exhibits a TC of $20\text{ppm}/^\circ\text{C}$ at three different supply voltages. In an absolute term, it is only $35\mu\text{V}/^\circ\text{C}$. Line sensitivity measurements are shown in Figure 6.7(b), where the output voltage changes only $0.033\%/V$. We also measure PSRR up to 100kHz where the PSRR is about -67dB as shown in Figure 6.7(c). We use a 0.8pF fingered, metal-to-metal output capacitor. Although PSRR for higher frequency is not measured, it should have minimal effects since the 2T voltage reference acts as a low pass RC filter. Figure 6.7(d) shows the current consumption for the 2T voltage reference. At 20°C and $V_{DD}=0.5\text{V}$, it consumes extremely low power of 2.22pW . Even at the worst case of 80°C and $V_{DD}=3.3\text{V}$, it still consumes less than nW , realizing a sub-nW voltage reference design for the first time.

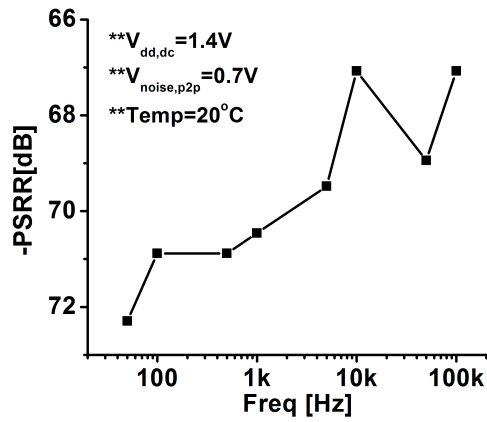
We compare the proposed 2T voltage reference to recently published low power



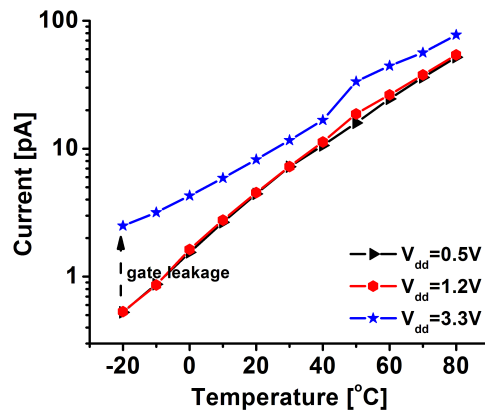
(a)



(b)



(c)



(d)

Figure 6.7: Measured (a) temperature coefficient (b) line sensitivity (c) power supply rejection ratio (d) current consumption of the 2T voltage reference

voltage references [4, 14, 106, 60, 30, 12, 6, 71] in Figure 6.1. First of all, power consumption is reduced by $16000\times$. This ultra low power consumption is particularly critical for building sensor type systems which consume typically hundreds of nW at active mode and tens of pW at standby mode [8, 61, 29, 33]. As shown in Figure 6.1, other voltage references already consume a comparable amount of power to what the entire sensing system consumes at active mode, adding a significant power overhead. On contrary, the proposed 2T voltage reference consumes so minimal power that it causes negligible power overhead at not only active but even standby mode.

In addition, the 2T voltage reference can operate at 0.5V, lower than the previous designs work as shown in Figure 6.1 since it requires no saturated devices and consequently less headroom. Although higher supply voltage is often available from batteries or I/O pads, the lower limit of functional supply voltage can facilitate system design. Also, it is extremely useful if there is no source for higher voltages, for example, in the system depending on only energy scavenging units which often generates low voltage level. Otherwise, the system requires voltage conversion scheme, which induces power overhead and design complexity.

6.5 Variability Analysis and Trimming Techniques

6.5.1 Statistical Measurement Results

Process variation is important to consider since it significantly affects the performance of circuits in modern CMOS technologies. In particular, since voltage references require high precisions in the output voltage and temperature coefficient, it is clear that a large number of measurements are required before designers become confident in the designs.

In order to evaluate the tolerance of the 2T voltage reference to run-to-run and die-to-die process variations, we measure 49 prototype dies of the 2T voltage refer-

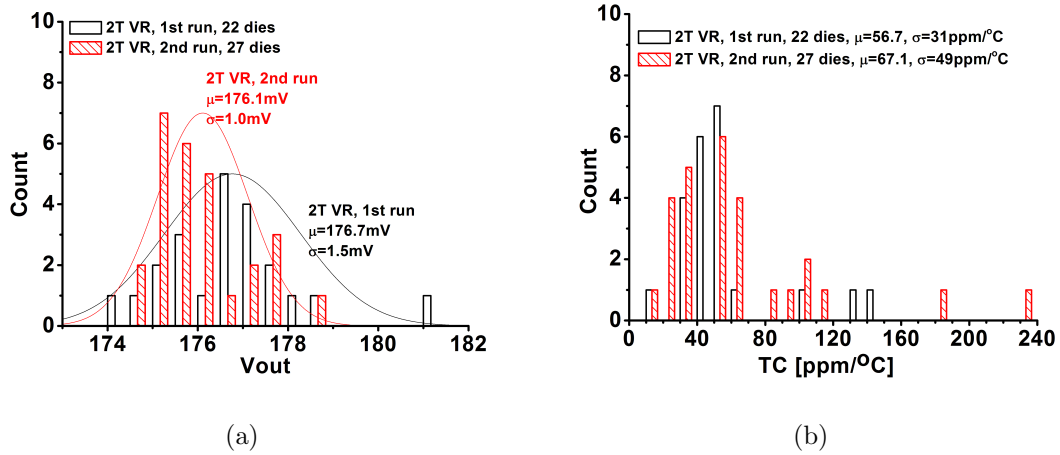


Figure 6.8: Measured (a) output voltage (b) temperature coefficient distribution of the 2T voltage references in two separate runs

ence from two separate runs in a $0.13\mu\text{m}$ CMOS technology. Although the voltage reference operates in the sub- V_{th} regime, where small process variations exponentially modulate subthreshold current, process variation impact on the 2T voltage reference is expected to be small due to 1) the linear effect of V_{th} on output voltage and 2) large device dimensions, suppressing both V_{th} variation due to random dopant fluctuations and geometric variations. Measured results for output voltage and temperature coefficient are plotted in Figure 6.8(a) and Figure 6.8(b). As shown in Figure 6.8(a), the average output voltage changes by a small amount (0.3%) between the two runs. However, both runs show considerable spread in TC and output voltage from die-to-die variations. The standard deviations of output voltage for each run are 1.5 and 1.0mV. The temperature coefficient is affected in a larger degree, as shown in Figure 6.8(b). In particular, the worst-case temperature coefficient is 3-4 times larger than the average value, requiring a measure to tighten the spread.

6.5.2 Digitally Trimmable 2T Voltage References

To minimize the TC and output voltage spread, we design a 2T voltage reference with digital trimming, as shown in Figure 6.9. The ratio of top-to-bottom device

widths is critical to TC and output voltage, as shown in Section 6.3. However, the optimal width ratio at design time may not be ideal for each chip due to process variations. Therefore, it is beneficial to be able to change the width ratio post-silicon. This voltage reference design can selectively turn on and off the four top and four bottom devices using associated switches. Bottom devices are sized as powers of 2 for range and granularity, while top devices are sized up gradually from the minimum width of native devices ($3\mu m$). By applying control signals *bmod* and *tmod* to the switches, the top-to-bottom width ratio varies from 0.52 to 3.75 with 256 different settings. Control signals swing from 0 to V_{DD} , requiring no extra supply voltage. One-Time-Programmable memories such as fuses can be used to provide the signals with minimal power overhead [31]. Once the switches are turned off, any top and bottom devices connected to them have negligible effect on the output voltage, acting as a dangling capacitor. Finally, a 0.8pF output capacitor is added to suppress the effect of noise on output voltage. Figure 6.10 illustrates the measured TC and output voltage for different settings in the trimmable voltage reference. Figure 6.10 shows that for a given total width of top devices, for example $22\mu m$, setting the bottom device total width to $10\mu m$ minimizes TC. A clear trend is observed where a specific width ratio leads to minimum TC, forming a diagonal line in the matrix. Likewise, output voltage changes at different settings, and depends directly on the width ratio. This is again confirmed by the diagonal line in Figure 6.10.

6.5.3 Analysis and Minimization of Trimming Cost

Minimizing trimming time is critical to reduce the cost per die, and can be achieved by reducing the number of temperature points, reducing the number of settings, and avoiding laser-trimming. Since there is no laser trimming required in the proposed voltage reference due to the digital trimming capability, it is important to develop a trimming process that uses minimum temperature points and control settings while

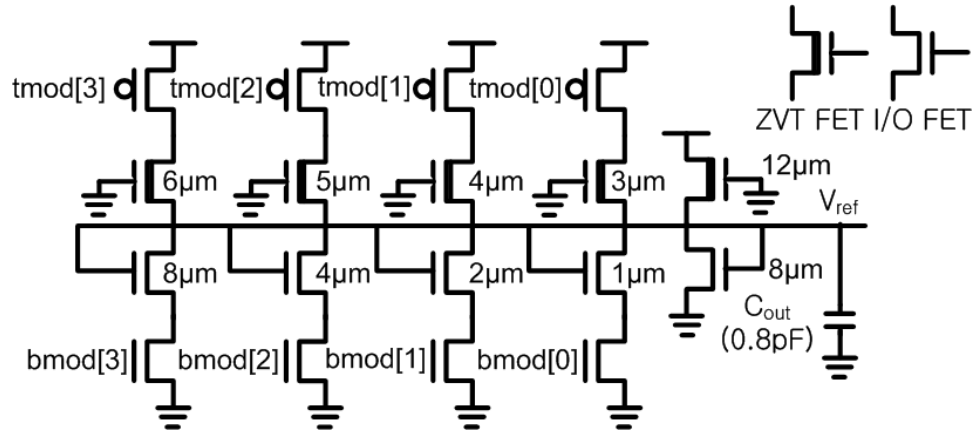


Figure 6.9: Figure 13 Schematics of trimmable 2T voltage reference

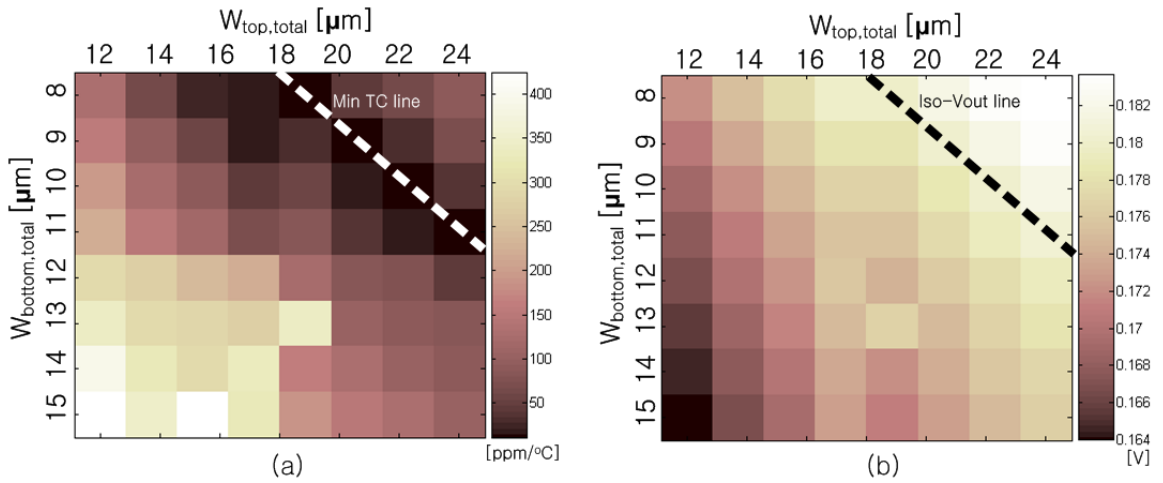


Figure 6.10: Measured (a) TC and (b) output voltage change with trim settings

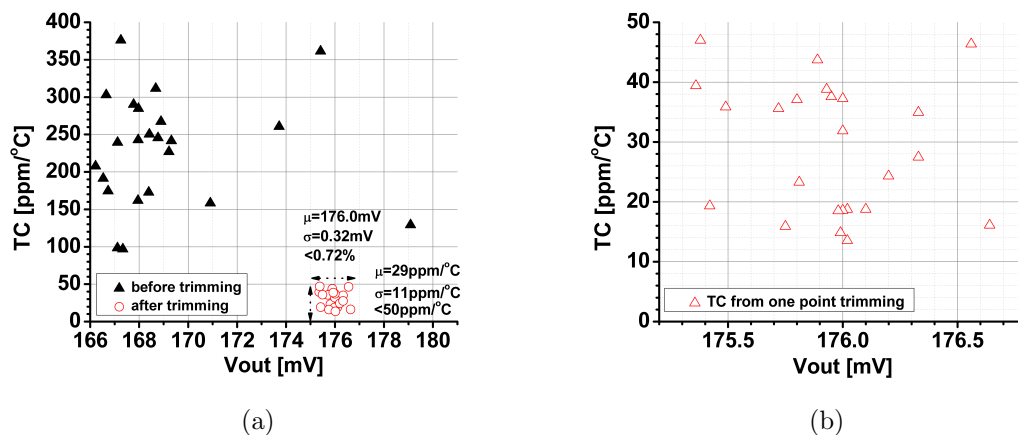


Figure 6.11: (a) Measured reductions of output voltage and temperature coefficient spreads (b) zoomed view

maintaining good post-trim performance.

More likely the design goal will be to meet a specified TC constraint with minimum deviation from the desired output voltage. Our objective for the trimming process is to minimize output voltage spread subject to TC being less than $50\text{ppm}/^{\circ}\text{C}$.

We try to trim with only one temperature point (80°C) to minimize the trimming cost for 25 dies. Since we cannot measure the TC with one temperature point, the trimming process entirely relies on the output voltage. Using the measurement at a single temperature point, we find the settings that minimize the output voltage spread. We test the voltage reference at the chosen setting at a finer temperature granularity (every 10°C from -20 to 80°C) and observe that both one temperature trimming reduces the spread of TC and output voltage by $9.6\times$ and $9.8\times$ compared to pre-trim results for the 25 dies as shown in Figures 6.11(a) and 6.11(b).

PSRR, Line Sensitivity (LS), and power consumption are also measured for the trimmable voltage references. Figure 6.12 shows that PSRR is measured as -51 to -64dB , which tracks simulation results well. Typical power consumption is 29.5pW at 0.5V , 25°C and 2.5nW at 3V , 80°C . Output referred noise is investigated with SPICE simulations, as shown in Figure 6.13. Together with a 0.8pF output capacitor, the

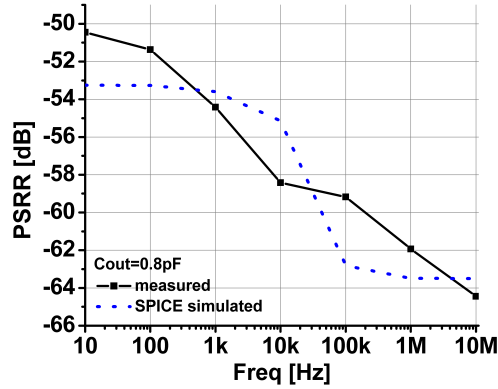


Figure 6.12: Measured PSRR of the trimmable 2T voltage reference

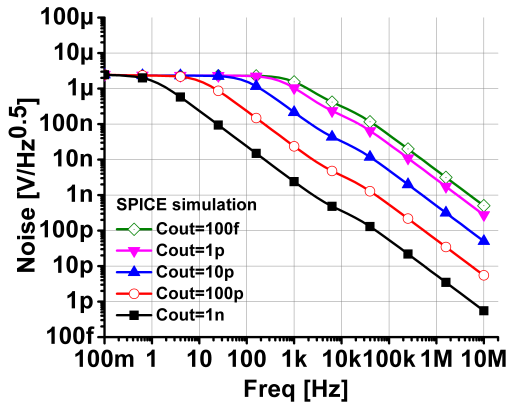


Figure 6.13: Simulated output referred noise of the trimmable 2T voltage reference

trimmable voltage reference effectively suppresses noise, showing $20nV/Hz^{1/2}$ with 1nF output capacitor at 100Hz. As a reference point, [61] exhibits $152nV/Hz^{1/2}$ with 100nF output capacitor at 100Hz.

6.6 Variant Designs of 2T Voltage References

In this section, we describe several variants of the 2T voltage reference. First variant is a 4T voltage reference to produce a higher output by stacking two 2T voltage reference, as shown in Figure 6.14. The measurement results from prototype design fabricated in $0.13\mu m$ CMOS shows a TC of $98.8ppm/^{\circ}C$, line sensitivity of $0.036\%/V$,

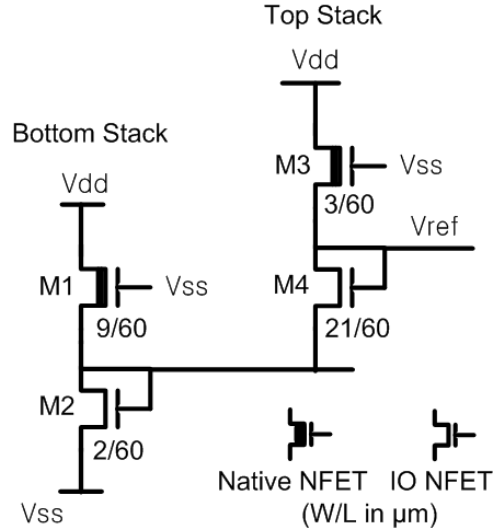


Figure 6.14: Schematics of 4T voltage reference

PSRR of -59dB at 100kHz, and power consumption of 10.85pW at $V_{DD}=0.5V$. The design requires $3500\mu m^2$ including the 0.8pF output capacitor. Measurement results are also summarized in Table 6.2. The die photo is shown in Figure 6.17.

We can also generate lower output voltage by replacing the bottom device (M1 in Figure 6.2) with multiple devices in the original 2T voltage reference as shown in Figure 6.15. This design has been implemented in a low power micro system [23], to achieve a specific bias voltage to analog circuitry.

By skewing the transistor size of the bottom and top devices (M1 and M2 in Figure 6.2), we can make CTAT or PTAT voltage reference. The temperature coefficient can be configured post-silicon by using the same topology as the trimmable voltage reference as shown in Figure 6.9. Output voltage linearly increases or decreases with temperature, as shown in Figure 6.16. The measured temperature coefficients for PTAT and CTAT are $145ppm/^{\circ}C$ and $-550ppm/^{\circ}C$.

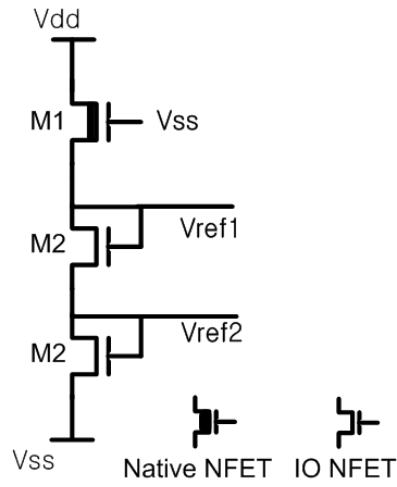


Figure 6.15: schematics of 2T voltage reference for lower output voltage

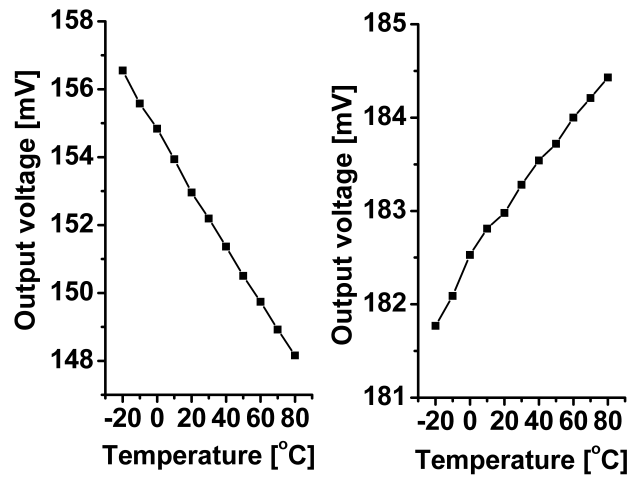


Figure 6.16: Measured temperature coefficients achieved by skewing the size of transistors

Table 6.2: Measurement summary of the proposed 2T voltage references

	2T			Trimmable	4T
Process	0.13 μ	0.18 μ	65nm	0.13 μ	0.13 μ
V_{DD}	0.5-3.0V	0.5-3.6V	0.5-2.5V	0.5-3.0V	0.5-3.0V
Vout (min)	174.9mV	326.8mV	327.2mV	175.2mV	341.5mV
Vout (max)	178.7mV	330.0mV	333.0mV	176.5mV	348.1mV
TC (min)	16.87ppm/ $^{\circ}C$	54.1ppm/ $^{\circ}C$	89.13ppm/ $^{\circ}C$	5.3ppm/ $^{\circ}C$	80.2ppm/ $^{\circ}C$
TC (max)	231ppm/ $^{\circ}C$	176.4ppm/ $^{\circ}C$	118.2ppm/ $^{\circ}C$	47.4ppm/ $^{\circ}C$	142.5ppm/ $^{\circ}C$
LS	0.033%/V	0.044%/V	0.33%/V	0.036%/V	0.036%/V
PSRR	-53/62dB (100Hz/10Mhz)	-49/55dB	-40/79dB	-51/64dB	-58/59dB
P(norm)	4.4pA (0.5V, 25 $^{\circ}C$)	11pA (0.5V, 25 $^{\circ}C$)	0.48nA (0.5V 20 $^{\circ}C$)	59pA (0.5V 25 $^{\circ}C$)	21.7pA (0.5V 25 $^{\circ}C$)
P(max)	81pA (3V, 80 $^{\circ}C$)	139pA (3V, 80 $^{\circ}C$)	8.13nA (2.5V 80 $^{\circ}C$)	847pA (3V 80 $^{\circ}C$)	400pA (3V 80 $^{\circ}C$)
Size	1350 μm^2	1425 μm^2	900 μm^2	9300 μm^2	3500 μm^2
Comment	2 runs, 49 dies	1 run, 14 dies,	1 runs 17 dies	post trim, 1 run, 25 dies	1 run, 30 dies

6.7 Technology Portability

Good technology portability is useful to integrate voltage references in the same die with a system for reducing footprint and cost. It can be particularly useful for a sensing system which often area and cost limited. In order to demonstrate the technology portability of the proposed voltage reference design, we implement 2T voltage references (Figure 6.2) in 0.18 μm and 65nm CMOS technologies other than 0.13 μm CMOS that we have discussed in the previous sections. Given the simple topology, porting the design only involves sizing of the two transistors. One thing to note is that footprint increase at long channel technologies is minimal since the same very long devices are employed for low power.

Measurements results from multiple dies in two different technologies are shown in Table 6.2, which shows similar performance as the design in the 0.13 μm CMOS. For the design in the 65nm technology, we intentionally use shorter channel lengths, which are $L_1=0.6\mu m$ and $L_2=1.2\mu m$, to demonstrate the tradeoff between area and power consumption. The die photos for all implementations are included in Figure 6.17.

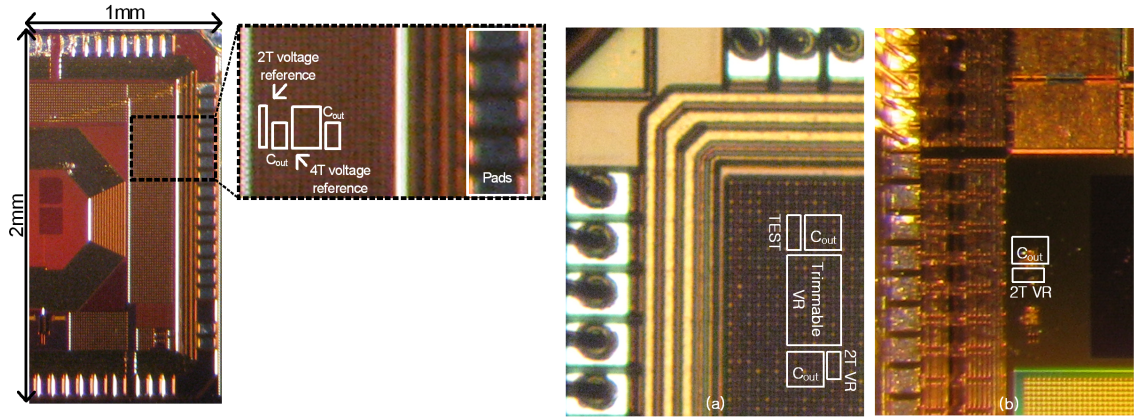


Figure 6.17: Die micrographs; (a) 1st $0.13\mu m$ run, (b) 2nd $0.13\mu m$ run, (c) $0.18\mu m$ run, (d) $65nm$ run

6.8 Summary

In this Chapter, we propose a 2T voltage reference and its variant which consumes sub-nanowatt and work at $0.5V$ of supply voltage while still maintaining a good temperature coefficient, line sensitivity, and power supply rejection ratio. Digital trimming ability is also added to the design, effectively tightening the spread of temperature coefficient and output voltage with an economical one-temperature point trimming. Easy technology portability is confirmed as well to successfully implement the design in the three different CMOS technologies. In particular, the ultra low power consumption enables integrating multiple voltage reference circuits with little overhead for cubic millimeter sensing systems.

CHAPTER VII

0.27V, 30MHz, 17.7nJ/transform, 1024-pt complex FFT Core with Super-Pipelining

7.1 Motivation and Previous Work

Voltage scaling has several limitations, including significant performance degradation and heightened delay variability due to large driving current sensitivity to PVT variations in the ultra low voltage regime. In addition, energy efficiency degrades below a certain voltage, V_{min} , due to rapidly increasing leakage energy consumption. This paper proposes a new approach to ultra-energy efficient design that uses circuit and architectural methods to further reduce the minimum energy point, or E_{min} , while simultaneously improving performance and robustness. The approaches are demonstrated on an FFT core in 65nm CMOS.

7.2 Contribution

Pipelining is a well-known method to improve performance, typically at the expense of energy consumption due to added sequential elements. However, in this paper, we make the counterintuitive observation that inserting additional pipeline latches improves both energy efficiency and performance in the ultra low voltage operating regime. Since pipelining shortens the clock period, it limits leakage energy

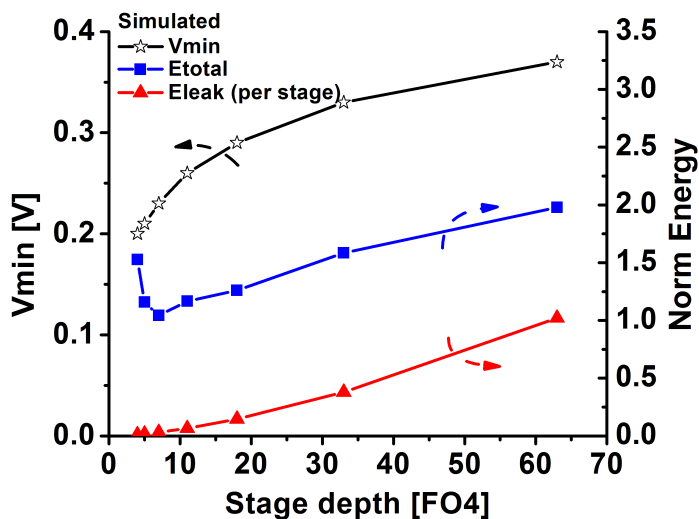


Figure 7.1: Effect of pipeline depth on energy consumption

consumed by idling gates, which reduces energy consumption and allows further voltage scaling. Simulations of inverter chains show that reducing stage depth from 65 to 11 fanout-of-four (FO4) delays yields 36% energy savings and a V_{min} reduction from 0.37V to 0.26V, as shown in Figure 7.1. By applying this super-pipelining approach to the multipliers in an FFT core, we find that it consumes minimum energy when pipelined in 6 stages at a stage depth of 17 FO4 delay. This design approach differs radically from conventional ultra low voltage designs, which tend to use limited pipelining and typically have cycles times in the 50-200 FO4 range [23, 102]. In this paper, we also show how clocking overhead can be reduced through circuit techniques for facilitating super-pipelining while process variation is addressed through the use of latch-based design. Additionally, architecture modifications are proposed to improve energy efficiency and throughput. Measurements show that the FFT core consumes only 17.7nJ per 1024-pt complex FFT while operating at 30MHz at $V_{DD}=0.27V$, demonstrating best reported FFT energy efficiency [102, 25, 90].

7.3 Architecture Design

An important principle driving the proposed ultra low voltage design methodology is to suppress leakage energy, allowing for larger potential energy savings by enabling voltage scaling. We first address this by architectural modifications through minimizing idling modules as shown in Figure 7.2. In a traditional memory-based FFT (Figure 7.2, bottom right), most memory cells idle while a single butterfly unit processes data word by word over many clock cycles. These idling cells waste leakage energy, harming energy efficiency and voltage scalability. On the other hand, conventional pipeline architectures such as Multi-Path Delay Commutator (MDC) have high memory utilization but low butterfly unit activity [43]. We therefore modify MDC to accept 4 inputs concurrently by proposing a new commutator configuration, enabling full utilization of both butterflies and memory elements. Additionally, we use two of the modified MDC lanes to double throughput and halve memory counts per lane, reducing leakage energy consumption from commutators. As shown in Figure 7.3, these modifications improve energy efficiency and throughput by $2.8\times$ and $6.2\times$, respectively, compared to a radix-4 memory-based FFT core.

7.4 Circuit designs

7.4.1 Super-Pipelining Technique and FIFO Design

Multipliers in the FFT are super-pipelined as shown in Figure 7.4. To successfully employ super-pipelining, sequential element overhead must be limited. Six latches share a local clock driver to reduce clock load. The drivers also use minimum-width fingers that enhance drivability at iso-input capacitance due to smaller V_{th} from inverse narrow width effects. Additionally, two latches are embedded in a mirror adder to save two transistors per latch. Latches are upsized from min-width for robustness such that they pass corners and 2 million Monte-Carlo mismatch simulations, pro-

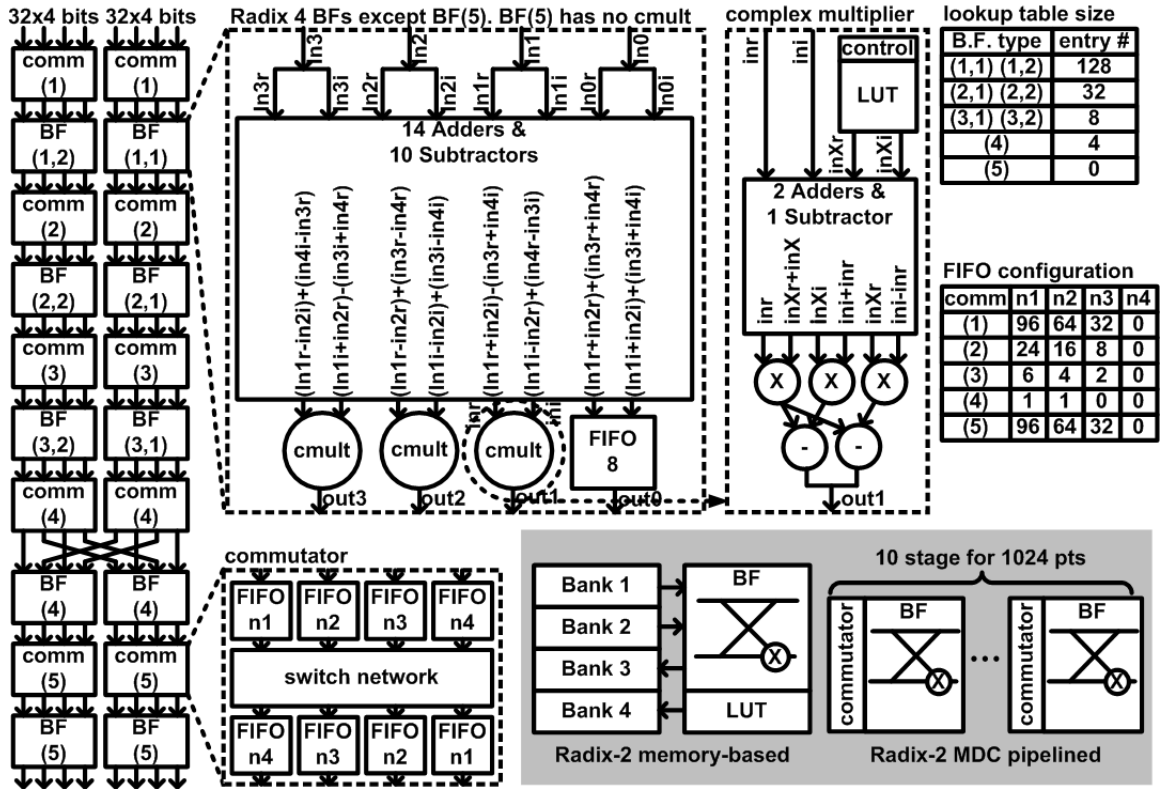


Figure 7.2: A pipelined, $8 \times 32b$ input, radix-4, 2-lane, 1024-pt, complex FFT architecture

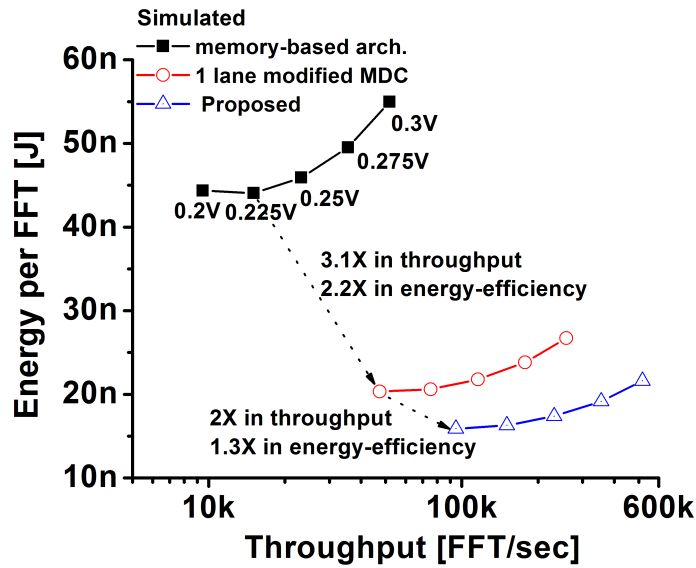


Figure 7.3: Energy and throughput improvement by architecture modifications

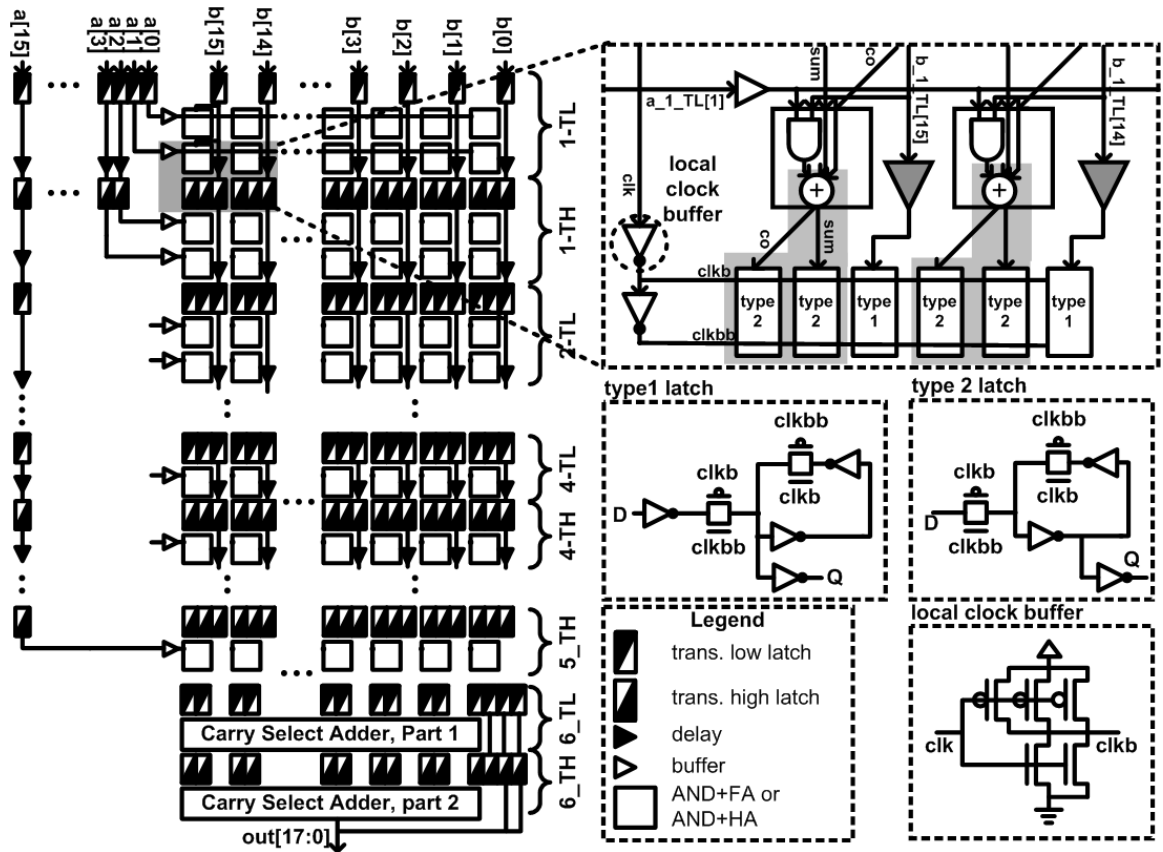


Figure 7.4: A 16b Baugh-Wooley multiplier is super-pipelined with 2-phase latches providing an estimated 99% chip-level yield with 10k latch instances per chip at 0.2V. We implement the proposed multiplier along with an unpipelined baseline multiplier, separately from the FFT core. Measured results in Figure 7.5 shows that the super-pipelined multiplier operates at 18MHz at 0.225V. It is $1.6\times$ faster while consuming 30% less energy than an unpipelined multiplier. It operates $3.6\times$ faster at iso- V_{DD} .

The FIFOs in the commutators contribute as much as 29% of the total FFT energy consumption in this architecture. To address this, we replace the address decoder with a cyclic address generator for reduced energy and use logic-based readout paths for improved performance, as shown in Figure 7.6. FIFO operation is described in Figure 7.7(a). Simulation results in Figure 7.7(b) show that the proposed FIFO design consumes 12% lower energy while improving performance by 20% over a memory with

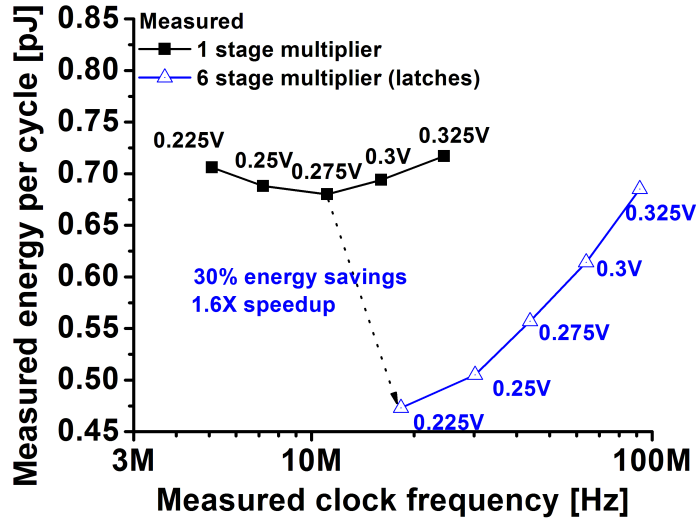


Figure 7.5: Measured energy consumption of differently pipelined multipliers

MUX-based readout.

7.4.2 Two-Phase Latch for Less Delay Variability

Although the above techniques improve energy efficiency and performance, we must pay attention to delay variability and overall design robustness given the ultra low voltage design point. We propose the use of 2-phase latches rather than flip-flops. Although the stage depth is drastically reduced in super-pipelined designs, time borrowing removes hard boundaries in the pipeline, re-establishing averaging of process variations along long paths that are present in unpipelined designs. Figure 7.8 shows Monte Carlo simulations on latch and flip-flop pipelined multipliers indicating that a latch pipelined multiplier can absorb delay variations, leading to higher performance yield. In addition, variability-induced hold time violations must also be avoided to ensure functionality. We identify short paths, aided by the regular structure of multipliers, and add delay elements that incur a marginal energy overhead of 2.4% per multiplier. Padded short paths were verified to satisfy hold times using 150k Monte-Carlo simulations under random process variations and corners.

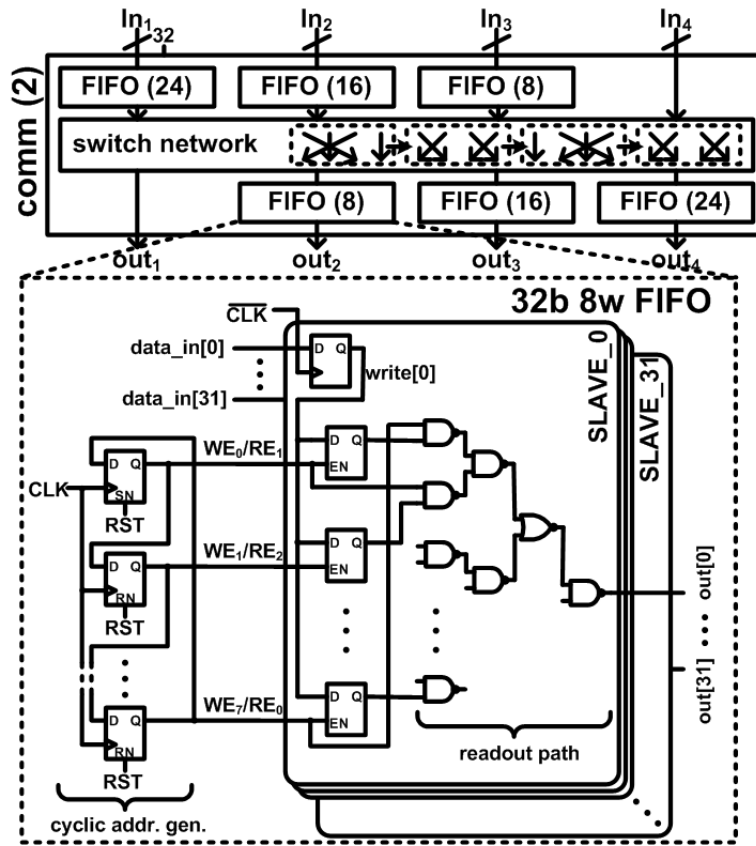


Figure 7.6: Schematics of commutators and FIFOs

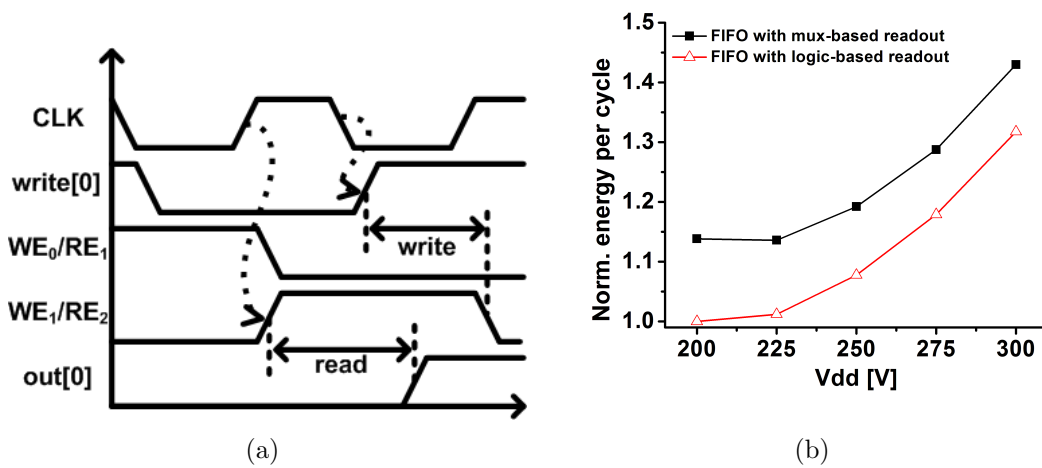


Figure 7.7: (a) Waveform of operation (b) FIFO energy consumption comparison.

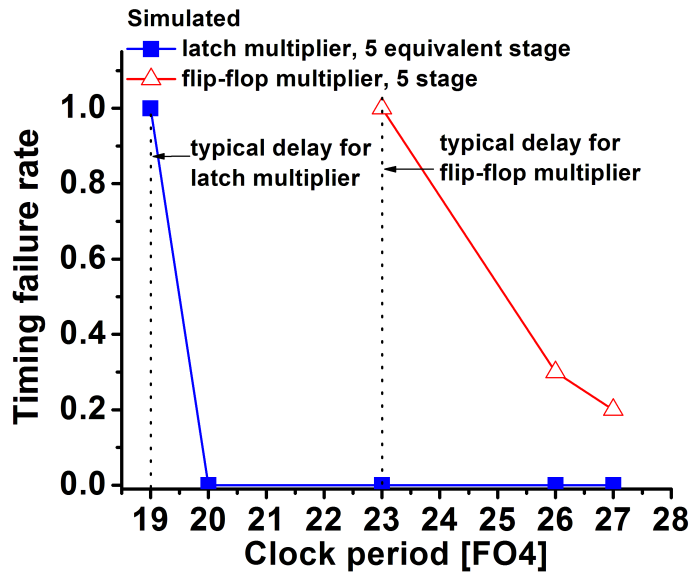


Figure 7.8: Timing failure rates across Monte-Carlo simulations with random process variations from two pipelined multipliers

7.4.3 Robust Clock Network Design

The clock distribution network is uniquely designed to suppress process variation induced skew and resulting hold time violations. Conventionally, many clock buffers are used to mitigate RC mismatch. However, at low V_{DD} the mismatch in these buffers is exacerbated and contributes significant skew, while RC delay is small compared to gate delay. Therefore, we design a 3-level clock network where reduced number of large buffers and matched RC interconnect are used. The lowest and middle levels of clock network are implemented with minimum width thin interconnect while the top level uses thick metal interconnect for lower RC delay and better slew. Figure 7.9 shows that the simulated worst-case RC mismatch is less than 0.15ns ($0.2 \times \text{FO4}$ at $V_{DD}=0.3\text{V}$).

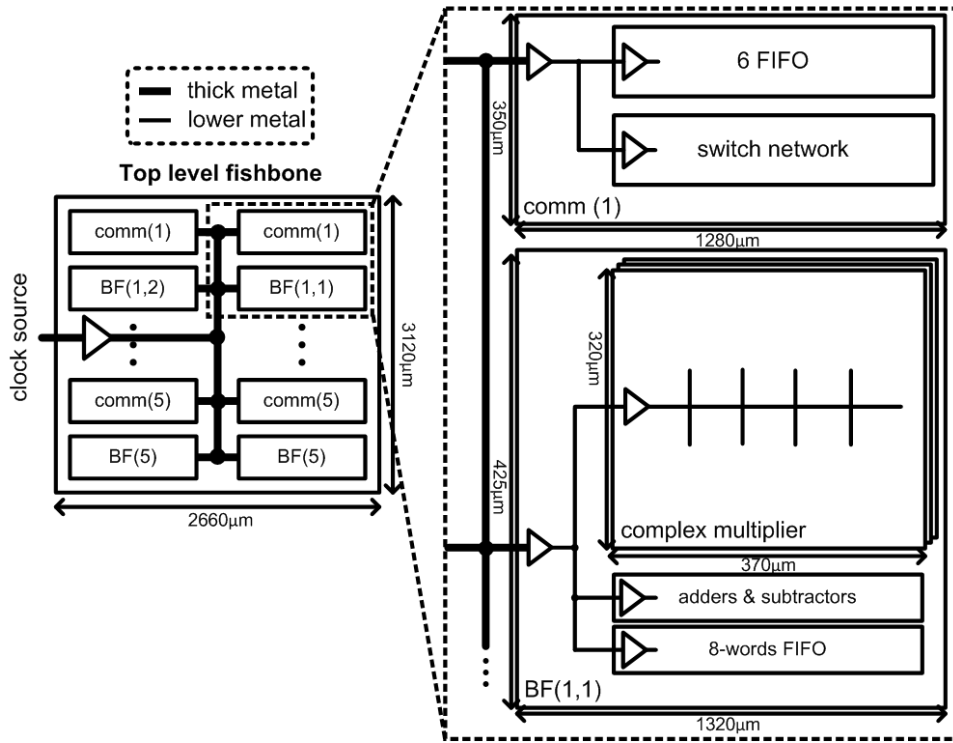


Figure 7.9: Proposed clock network design with limited buffers and matched interconnects

Table 7.1: RC mismatch in clock network

RC mismatch [ns]		
LV1	complex multiplier	0.072n
LV2	BF	0.086n
LV2	COMM	0.142n
LV3	FFT	0.036n

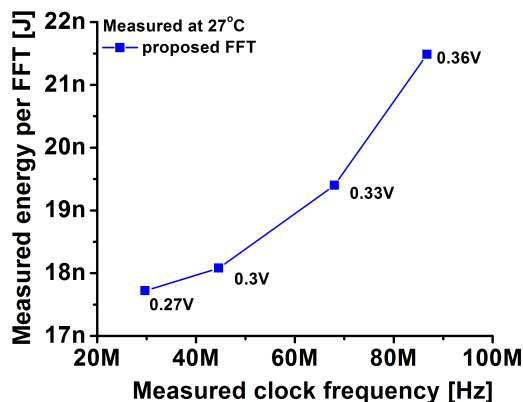


Figure 7.10: Measured energy consumption and performance of the proposed FFT core

Table 7.2: Comparison of the proposed FFT core

	Proposed	[102]	[25]	[90]
Technology	65nm	180nm	90nm	130nm
Type	1024 CV	1024 RV	256 CV	256 RV
Bit width	16b	16b	10b	16b
Area	2.71x3.15mm ²	2.6x2.1mm ²	2.26x2.26mm ²	n/a
V_{DD}	0.27V	0.35V	0.85V	0.5V
Clock Freq.	30MHz	10kHz	300MHz	7kHz
Energy/FFT	17.7nJ	155nJ	12.8nJ	100nJ
Norm. energy/FFT	17.7nJ	111.9nJ	71.0nJ	400nJ

7.5 Measurement Results and Comparisons

The FFT core is fabricated in 65nm CMOS using the above circuit and architectural techniques. Measurements in Figure 7.10 show that it computes 234k 16b 1024-pt complex FFT per second. The clock frequency is measured as 30MHz with $V_{DD}=0.27V$ compared to frequencies of 10s of kHz for typical ultra low voltage designs at the same supply voltage. The FFT consumes 17.7nJ/transform, which is 4× smaller than prior work when scaled for FFT size and technology [25]. Table 7.2 shows the energy consumption and performance comparisons with the existing low power FFT core designs. A die photograph is shown in Figure 7.11. Measurements are summarized in Table 7.3.

Table 7.3: Measurement summary of the proposed FFT core

Type	1024-pt complex
V_{DD}	0.27V
Clock Freq.	30MHz
Energy / FFT	17.7nJ
Throughput	234k FFT / sec
Area	2.66x3.12 mm^2
Technology	65nm CMOS

7.6 Summary

The FFT core implementation demonstrate the circuit and architecture technique to mitigate several issues in ultra low voltage operation, which include low performance, delay variability, and leakage power waste. Super-pipelining techniques reduce leakage power waste, enabling higher energy efficiency beyond the limit of traditional voltage scaling technique. It also improves the performance. Additionally, architecture modifications improve energy efficiency by minimizing leakage power waste as well as increasing throughput. For less delay variability, latch-based design is employed to average out the effect of process variations throughout long paths which are re-established by the cycle borrowing ability. We also design a robust clock network to eliminate skew and slew variability. The silicon measurements show that the core consumes record-low 17.7nJ per 1024-pt complex FFT and operates at the remarkable performance of 30MHz at 0.27V.

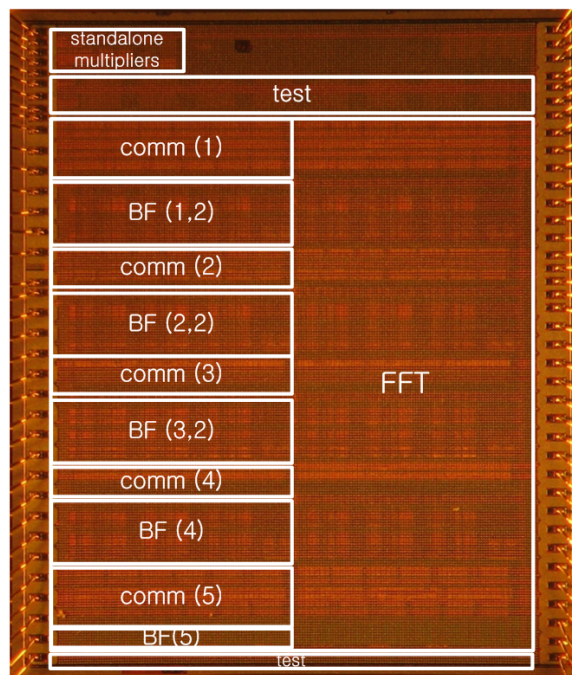


Figure 7.11: Die photo of the FFT core implemented in 65nm CMOS

CHAPTER VIII

Robust Clock Network Design for Ultra Low Voltage Operations

8.1 Motivation and Previous Work

As pointed out for several times, the scaled supply voltage makes design less robust due to the reduced on-current to off-current ratio of transistors. The robustness can be further compromised by process and environmental variations since the subthreshold current varies exponentially with variations such as random V_{th} mismatch [115]. However, attempts to improve robustness often lead to higher energy consumption: for example, MOSFET upsizing to mitigate random V_{th} mismatch. In that sense, achieving low power and robustness together poses a challenge for designing ultra low voltage circuits.

$$T_{cq,reg}(T_{clk,slew}) + T_{min,logic} \geq T_{hold}(T_{clk,slew}) + T_{clk,uncertainty} \quad (8.1)$$

In order to achieve an ultra low power and robust system, clock network design is critical. With the highest switching activity, the clock network consumes up to 40% of total dynamic power [65]. With similar trends in ultra low voltage regimes,

clock networks make a large impact on total energy consumption, requiring additional design efforts. Along with the low power requirements, the clock network should be designed for robustness. As shown in EQ 8.1, skew should be minimized and well-defined against process and environment variations, otherwise the design can have short paths and functional failures [57]. Additionally, slew needs to be well-controlled since it degrades the setup and hold time of registers.

There have been several works on designing clock networks in ultra low voltage regimes. In [49] and [62] the authors designed charge-pump based clock buffers to enhance robustness. Although the robustness is improved, both designs incur energy overhead and require custom clock buffers. The authors in [95] seek to tighten slew variations at ultra low voltage regimes by constraining slew differently at each clock tree level. However, they did not consider skew, a key metric for clock network design.

8.2 Contribution

Therefore, in this Chapter we investigate a low power and robust clock network design methodology that avoids custom gates while considering energy, skew, slew and their variability at ultra low voltage regimes. We start by comparing various clock networks for a generic design. Several levels of buffered and un-buffered H-Trees and a simple signal-route clock network are studied. Then, device and interconnect process variations are analyzed for their impacts on clock networks. In addition, the impact of supply voltage and technology scaling on clock network is investigated.

From these studies, we find that the design methodology of clock network in ultra low voltage regimes should be radically different due to the negligible interconnect resistance. Typically, in super-threshold regimes designers add buffers to mitigate interconnect delay. However it becomes disadvantageous in ultra low voltage regimes since buffer delay varies with process, temperature, and supply voltage variations and degrades skew/slew robustness, while reducing already negligible skew contributions

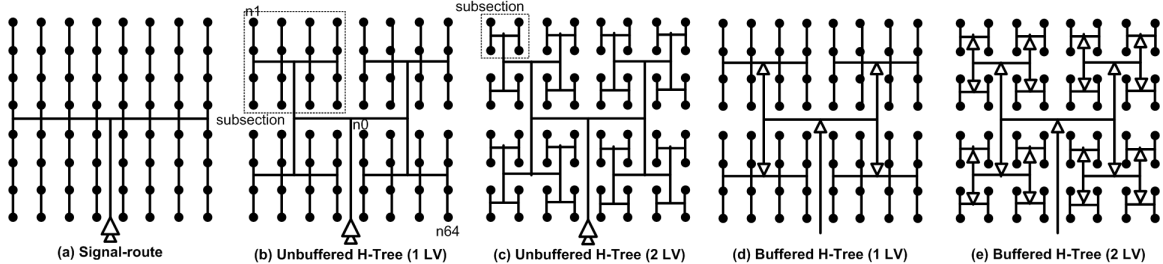


Figure 8.1: Clock network topologies

of interconnects. Therefore, we propose a different method using no buffers inside clock networks for minimizing skew/slew variations and energy consumption. As a case study, several clock networks for a 16b MSP430-compatible microprocessor [73] are implemented and simulated in SPICE. We confirm that an optimally-selected clock network greatly outperforms other typical clock networks in skew/slew variations and energy consumptions.

8.3 Clock Network Comparison at Ultra Low Voltage Regimes

8.3.1 Comparison Frameworks

Figure 8.1 shows clock networks for a simplified design where 4096 master-slave flip-flops or sinks are placed regularly in 1.4 x 1.4mm² area. (Only 64 sinks are shown in Figure 8.1 for clarity) These are used in the simulations throughout Sections 8.3 and 8.4. The candidate networks for comparison are signal-route, and 1-4 level unbuffered/buffered H-Trees (3- and 4-level H-Trees are not shown in Figure 8.1). The signal-route clock network routes the clock like an ordinary signal with no balancing attempted. At the bottom of the H-Tree, sinks are also routed as signals. The signal-route network can be considered as a 0-level un-buffered H-tree to simplify plotting. Grid and grid-tree hybrid clock networks are not considered in this work since they often incur large power penalties. The chosen sink density is based on a survey of two microprocessors: 32b ARM Cortex M3 microcontroller [5] and 16b MSP430-like

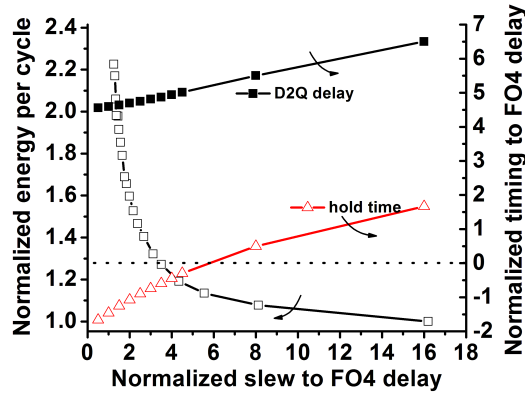


Figure 8.2: Tradeoff between slew and clock network energy

microcontroller, which is used in the case study of Section 8.5.

We assume 4FO4 as slew constraint (i.e., clock signal transitions from 10% to 90% in 4 fanout-4 delays) since this represents a balance between energy consumption of clock buffers and slew of clock signals as shown in Figure 8.2. At this slew rate, the D-Q delay and hold time remain in reasonable ranges. However other slew rates can be chosen based on application requirements. Clock drivers are sized up to achieve the same slew rates at sinks for all topologies considered. For the signal-route and un-buffered H-Trees, a large central driver (usually consisting of several cascaded buffers of increasing driving strengths) is sized to switch the entire clock network at the slew constraint. For the buffered H-Trees, many small buffers, sized for the iso-slew constraint, are distributed inside the clock network. Here, we assume a fanout-4 ratio to drive buffers. Higher level H-Trees require more buffers although individual buffers are smaller at the iso-slew constraint (Figure 8.3).

Since interconnect resistance is negligible in ultra low voltage designs, minimum width metal is used for clock networks, reducing energy consumption. The clock net is shielded by supply voltage nets to minimize crosstalk noise. Since the clock network is shielded, the wire capacitance can be well-defined regardless of surrounding wires and their switching activities.

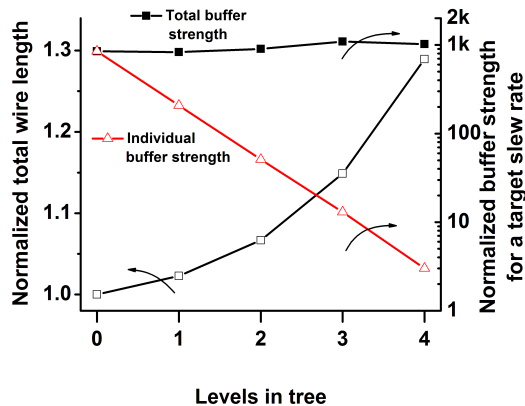


Figure 8.3: Wire length and required buffer

We use a 0.3V supply voltage and a $0.18\mu\text{m}$ CMOS technology, which is a typical technology and supply voltage combination for energy-optimal ultra low power designs [87, 11]. However, we also consider the impact of higher supply voltages and a scaled technology later in Section 8.4.

8.3.2 Comparison at Nominal Conditions

Given the framework of Section 8.3.1, we compare the energy consumption and global skew for the clock networks with SPICE simulations. In this work we consider energy dissipated in clock buffers and interconnects. Energy consumed internal to registers including local clock drivers that sharpen clock signal edges are not included as these will be constant across network topologies. Figure 8.4 shows that higher level H-trees consume more energy due to longer interconnect. For higher level trees, un-buffered networks consume slightly more energy than buffered counterparts due to the iso-slew constraint. Since wire RC increases quadratically with the length of the wire, distributed buffers in the buffered H-Trees are more energy-efficient for achieving the same slew than central drivers in un-buffered H-Trees.

Skew is improved exponentially as we increase the tree level since the area of the subsection, which is proportional to the amount of skew, becomes $4\times$ smaller per level

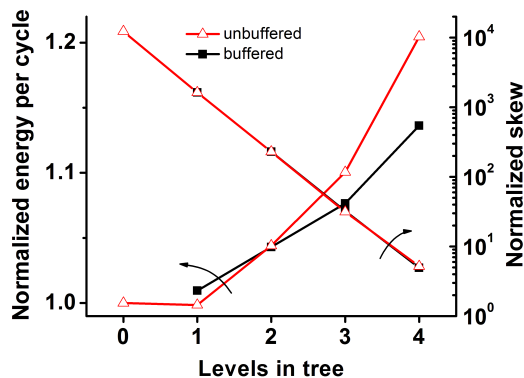


Figure 8.4: Energy and skew comparison of clock networks.

(Figure 8.1(b) and (c)). Theoretically, a 6-level H-Tree eliminates any path mismatch for the 4096 sinks. The signal-route or 0-level H-tree exhibits the largest skew due to the longest path mismatch as expected.

8.3.3 Impact of MOSFET Process Variations

It is well known that MOSFET parameter variations, such as random V_{th} mismatch, have an exponential effect on gate delay at ultra low voltage regimes [115]. In a clock network, delay variation degrades skew and slew from the expected values, causing both performance degradation and functional failure. Although clock buffers use relatively large MOSFETs, they still show considerable delay variations from random V_{th} mismatch due to the high sensitivity of subthreshold current. Therefore, it is critical to consider the effect of process variations on clock networks for robust operation at ultra low voltage regimes. We do not include the effects of temperature and supply voltage variations for simplicity since these affects skew and slew in the similar fashion as process variations do, only worsening skew and slew variations further.

In this section, we consider the impact of MOSFET process variations on clock network designs. Monte Carlo simulations with random MOSFET mismatch on the clock networks are performed. We use SPICE model with embedded statistical data

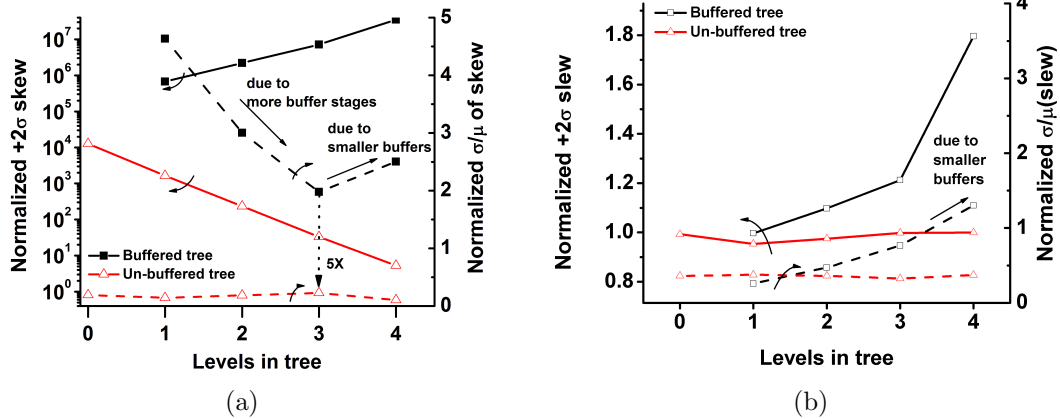


Figure 8.5: (a) Skew with MOSFET variations, (b) Slew with MOSFET variations.

from foundries. Global variation is ignored since it has a negligible impact on skew and can be tuned out using global parameters such as body biasing and voltage scaling at a reasonable overhead [23, 78, 107, 86].

Figure 8.5(a) shows the $+2\sigma$ value of skew across different clock network topologies. Compared to the case with no process variation, buffered trees exhibit several orders of magnitude larger degradation in skew. This is because the buffer delay which used to be cancelled among buffers starts to contribute to skew. Another interesting observation is that the $+2\sigma$ skew increases for higher level buffered-H-Tree while the opposite trends are observed without process variations. It implies that adding buffers in ultra low voltage regimes has no contribution in mitigating path RC mismatch but only degrades the total path delay. We will discuss the issue of driving interconnects in Section 8.3.5. The σ/μ of skew for the buffered H-Trees is also at least $5\times$ worse than un-buffered topologies. Figure 8.5(b) shows the slew having similar trends to the skew. The un-buffered topologies show a good robustness on slew control while buffered trees have degraded and more variable slew as we increase tree level.

The σ/μ for skew and slew shows different trends with tree level. Figure 8.5(a) shows that the skew variability first reduces since clock signals travel through more

stages of buffers and thus delay variations are averaged. However it starts to increase at level 3 due to the smaller and thus more process-sensitive buffers. However, slew variability is mostly determined by the final buffers which directly drive sinks. Therefore, it has no averaging effect, different from the skew case.

8.3.4 Impact of Interconnect Process Variations

Interconnect variation is another source of performance variability in scaled CMOS technologies. However, it can be considered as a secondary effect in ultra low voltage regimes since its impact on delay is roughly linear, while device variations have exponential effects. Therefore, we apply the worst case interconnect variation to the studied clock networks, and evaluate whether their skew contribution is significant compared to the contribution of MOSFET variations.

Finding the worst case corner for interconnects is difficult and requires detailed information from physical design since a fixed process variation (e.g., thinner interconnect) might cause two opposite effect on delay, depending on whether the particular wire delay is capacitance dominated or resistance dominated [64]. However, at ultra low voltage regimes, the worst case for interconnects is better defined since interconnect delays are always capacitance dominated due to the negligible interconnect resistance. The worst interconnect corner for skew can be defined between two non-overlapping paths experiencing min and max interconnect capacitance (provided by the foundry design kit). For example, in 8.1, if the path from n0 to n1 has max capacitance and the path from n0 to n64 has min capacitance, two sinks at n1 and n64 will experience the largest skew.

With the worst interconnect corner, we run SPICE simulations to evaluate the contribution of interconnect variations on skew, compared to MOSFET variations. Simulations show that it takes only 10-15% of total skew across the 1-4 level buffered H-Trees. For un-buffered trees, the interconnect variations might seem to be con-

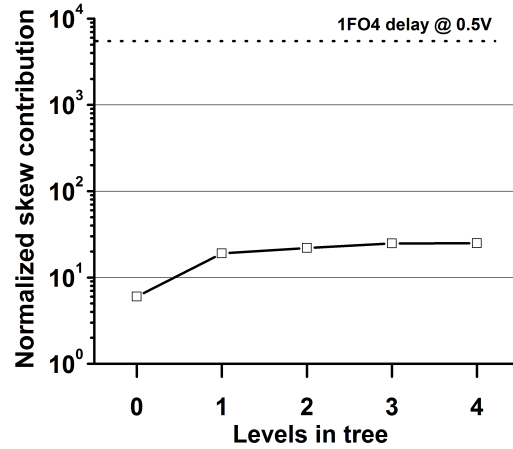


Figure 8.6: Skew contribution from interconnect variations

tributing non-negligible skew. However, this is mainly because the large central drivers are little affected by process variations. As shown in Figure 8.6, the absolute amount of skew contribution from interconnect variations is much smaller than gate delays in ultra low voltage regimes. Additionally, the worst case corner for interconnects is highly pessimistic. Therefore, we can simply ignore interconnect variations without too much loss of accuracy.

8.3.5 Driving Interconnects at Ultra Low Voltage Regimes

At super-threshold regimes, repeaters are commonly added in the middle of a long interconnect, which gives better performance [19]. The benefit comes from shorter interconnect segments (i.e. quadratically smaller wire RC) and sharper slew rate to the input of a following buffer. As shown in Figure 8.7(a), adding one buffer in the middle of a long interconnect improve performance at $V_{DD}=1.8V$ for wires $> 3mm$.

However, these benefits are no longer valid at ultra low voltage regimes. First the delay penalty of adding buffers is often much larger than the reduction of wire RC. Figure 8.7(a) shows that adding buffers cannot reduce delay even for interconnects longer than 20mm. EQ 8.2 using the results of [81] can easily verify the results. Slew rate is also negligibly affected by interconnects since the total resistance ($R_{fet}+R_{wire}$)

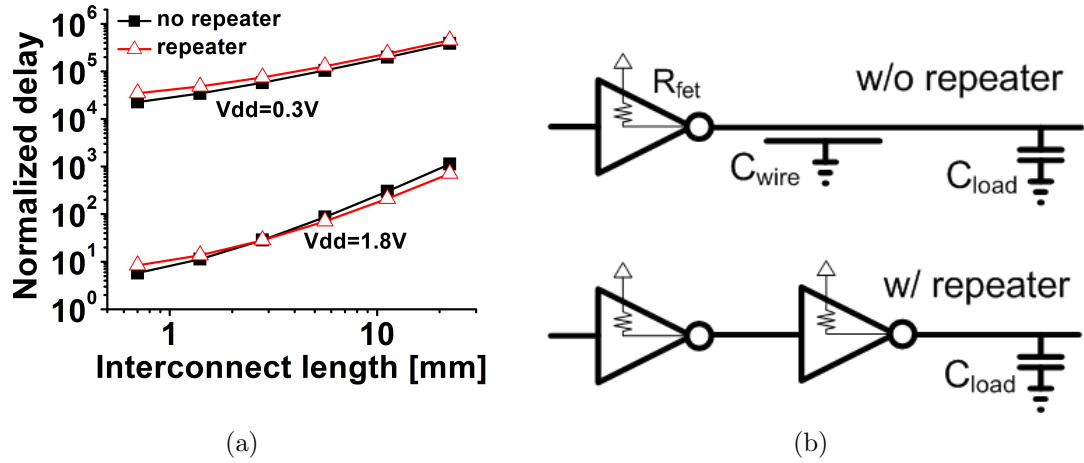


Figure 8.7: Driving a long interconnect without repeaters (a) delay comparison, (b) schematics.

is dominated by MOSFET resistance.

$$\begin{aligned}
 t_{w/repeater} &\approx 0.693 \cdot R_{FET} \cdot (C_{wire} + C_{load} + C_{inv}) \\
 t_{w/orepeater} &\approx 0.693 \cdot R_{FET} \cdot (C_{wire} + C_{load})
 \end{aligned}
 \tag{8.2}$$

Technically, adding buffers to drive a long interconnect is only harmful at ultra low voltage regimes since they act as another source of variation. It also consumes more energy.

8.4 Impact of Voltage and Technology Scaling

In Section 8.3, we considered 0.3V supply voltage and a $0.18\mu m$ technology. While this represents the optimal choice [87] for most energy-constrained systems, other application spaces may require higher performance and therefore prefer different supply voltages and technologies. In this section we discuss the impact of the supply voltage and technology on the optimal selection of clock networks.

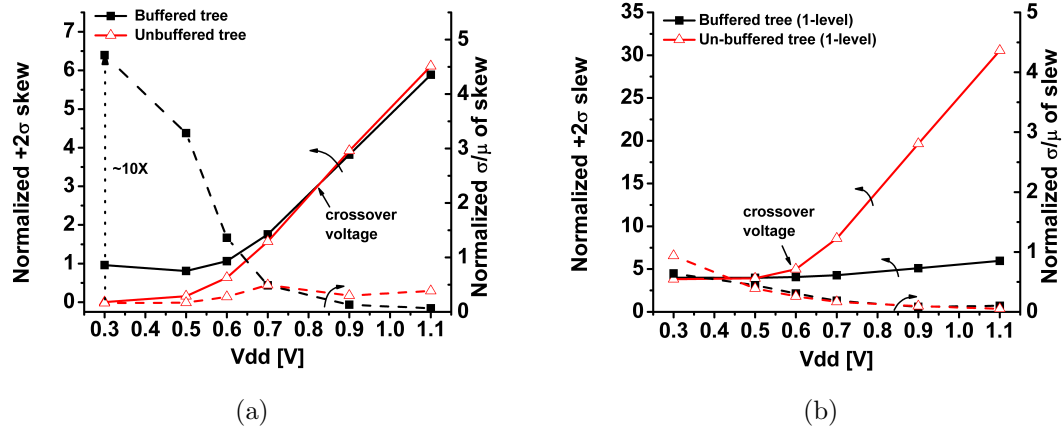


Figure 8.8: Impact of voltage scaling on (a) skew, (b) slew.

8.4.1 Supply Voltage Scaling

Figure 8.8(a) shows the results of Monte-Carlo iterations with random MOSFET variations on 1-level buffered and un-buffered H-Trees over supply voltages. One interesting observation is that there is a crossover voltage at 0.85V in Figure 8.8(a). At $V_{DD} < 0.85$ V, the un-buffered network outperforms in $+2\sigma$ skew and σ/μ of skew. However, the buffered tree performs better at $V_{DD} > 0.85$ V. This is because the buffers in the buffered H-Tree become less sensitive to process variations at higher supply voltage. Additionally, buffers start to drive interconnects strong enough to mitigate some of path RC mismatches, resulting in improved skew-related metrics.

Slew also has a crossover voltage at 0.6V in Figure 8.8(b). At $V_{DD} > 0.6$ V, a degradation in $+2\sigma$ slew is observed for the un-buffered H-Tree since interconnect resistance is no longer negligible compared to the MOSFET resistance of the clock drivers. However, the buffered H-Tree maintains the similar slew across the supply voltages due to shorter interconnect.

8.4.2 Technology Scaling

In Section 8.4.1, we observed crossover voltages in skew and slew. Technology scaling also acts in the similar way since scaled technology has more resistive inter-

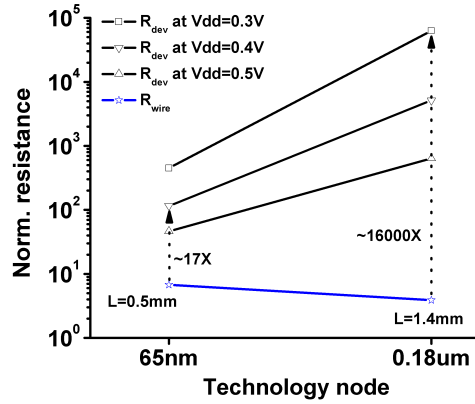


Figure 8.9: Resistance scaling across technologies

connects and less resistive MOSFETs with lower V_{th} . Figure 8.9 shows the MOSFET and interconnect resistance trends in two different technologies. We assume that the interconnect length is scaled with the channel length of technology for the same design. Still, increase in wire resistance is observed. The difference between wire and device resistance reduces from $16000\times$ at $V_{DD}=0.3V$ and $0.18\mu m$ technology to only $17\times$ at $V_{DD}=0.5V$ and $65nm$ technology.

We additionally run the Monte-Carlo simulations on the 1-level buffered and unbuffered H-Trees to identify the crossover voltages of $+2\sigma$ skew and slew in a $65nm$ General Purpose (GP). GP process is chosen as a more pessimistic option for unbuffered topology, compared to Low Power (LP) process CMOS technology. We use the statistical data supplied by the foundry design kits.

Figure 8.10 shows the trends of crossover voltages over two technology nodes. Both skew and slew crossover voltages appear at lower voltage for scaled technology due to the reduced difference of resistance between devices and interconnects. Slew might be the limiter for un-buffered clock networks since its crossover voltage appears earlier than the skew counterpart. One might want to move the slew crossover voltage to higher voltage regime to exploit less skew and skew variability from un-buffered clock networks. Since we use minimum width interconnects for low power, thick top-level

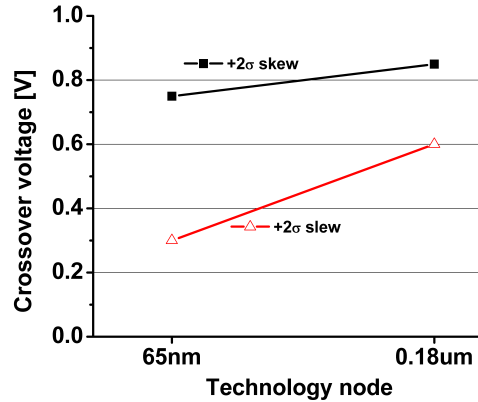


Figure 8.10: Crossover voltages for $+2\sigma$ skew and slew

metals and wider metals can be considered as an option to improve slew. However it might have energy overhead, requiring a careful evaluation.

8.5 Clock Network Design for a 16b MSP430-like Microcontroller

In Sections 8.3 and 8.4, the simplified design, where sinks are regularly placed, is used to study on designing robust yet low power clock networks. In this section, we will continue our investigations on clock networks using more practical design, a 16b MSP430-compatible microprocessor.

We first characterize standard cells at $V_{DD}=0.3V$ in a $0.18\mu m$ CMOS technology with 6 metal stack. The core of the microcontroller is synthesized and APR-ed (Automatic Placement & Route) with industrial tools. Then, 7 different clock networks including signal-route and 1-3 level buffered/un-buffered H-Trees clock networks are implemented. Fourth and fifth metal layers are used to implement clock networks. It is shielded with V_{DD} net. One example APR-ed design employing 3-level buffered H-Tree is shown in Figure 8.11. The traces for H-Trees are highlighted for visibility. The total footprint is $0.6\times 0.6mm^2$. Interconnects, buffers and flip-flops in clock net-

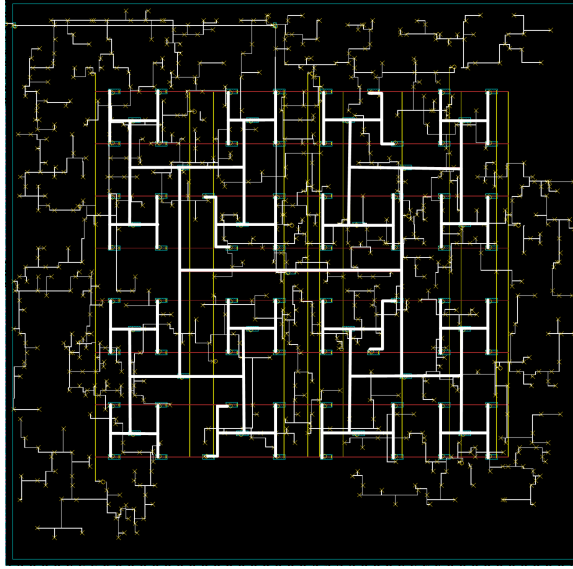


Figure 8.11: Layout view for the APR-ed microprocessor with 3-level buffered H-Tree clock network

works are extracted with parasitic capacitance and resistance in a SPICE format for simulations. Mismatch Monte-Carlo iterations are performed to evaluate skew, slew and energy for each clock network.

As shown in Figure 8.12(a), 1-level un-buffered H-Trees can improve 4 orders of magnitude in $+2\sigma$ skew and $36\times$ in σ/μ of skew, compared to the worst case clock network. The 1-3 level buffered clock networks can produce up to $5\times$ worse skew from the values of design phase, which can cause functional failures after fabrications. Note that the worst clock network in the comparison, which is the 3-level buffered H-Tree, might be chosen as an optimal network in super-threshold regimes [110, 80], confirming the importance of the clock network selection at ultra low voltage regimes.

The $+2\sigma$ slew and its variability are plotted in Figure 8.12(b), which has the similar trends to skew. The 3-level buffered H-Tree can have 28% higher slew from variation than the design values, resulting in less robust design. Energy consumption for each clock network is also compared in Figure 8.12(c). Higher level trees consume more energy. Although the 1-level un-buffered consumes the second least energy after the signal-route clock network (Signal-route clock network consumes 4% less

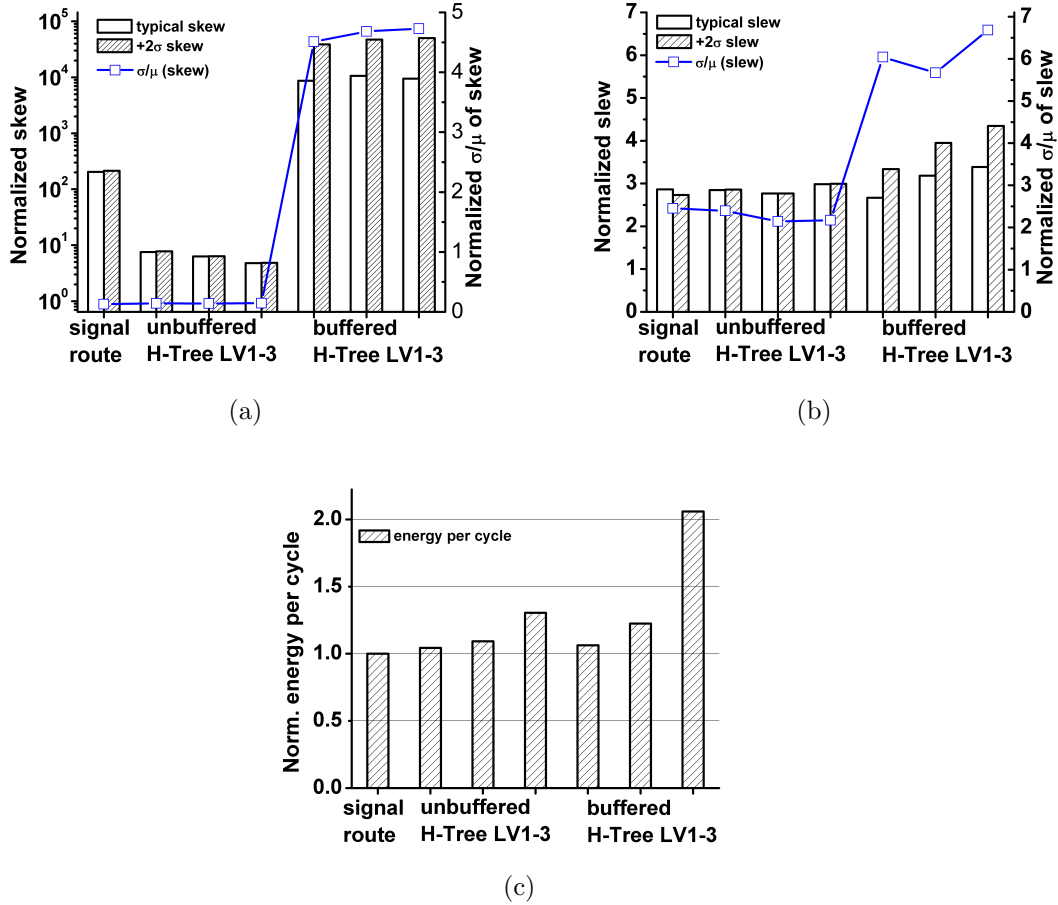


Figure 8.12: Comparison for the clock networks of the 16b microcontroller on (a) skew, (b) slew, (c) energy consumption.

energy than 1-level un-buffered H-Tree), it consumes 49% less energy than the 3-level buffered H-Tree. The 3-level buffered H-Tree consumes relatively larger energy than expected from the simplified analysis in Section 8.3 since the individual buffer strength scales more slowly than the simplified design at the slew constraint.

In Section 8.4, we observed that a clock network which used to be optimal at low V_{DD} loses optimality when supply voltage goes up to a certain point, which we define as a crossover voltage. Here we also sweep the supply voltage to find the crossover voltages. As shown in Figure 8.13, the 1-level un-buffered H-Tree is skew-optimal choice at $V_{DD}=0.3-1.0V$. At $V_{DD} > 1.0V$, the 1-level buffered H-Tree becomes skew-optimal. The crossover voltage for $+2\sigma$ slew appears at $V_{DD}=0.6V$, which is

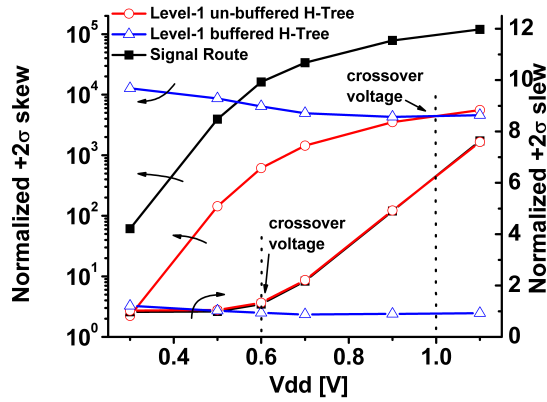


Figure 8.13: Optimal clock network across supply voltages

lower than the skew crossover voltage. Thick metal layers or non-minimum width interconnects might be considered to alleviate slew degradation at the cost of energy consumption

8.6 Summary

In this Chapter, we investigate on designing a low power yet robust clock network at ultra low voltage regimes. After comparing several clock networks in energy consumption, skew, and slew in the contexts of both simplified and practical designs, we find that a radically different methodology of using no buffers inside clock network is beneficial at ultra low voltage design. A case study with a 16b microcontroller shows that the optimally-chosen clock network at an energy-optimal operating point can improve $+2\sigma$ skew by 4 orders of magnitude, skew variability by $36\times$, and energy consumption by 49%, compared to the clock network which can be considered a typical practice in super-threshold voltage designs. Impact of process variation and supply voltage and technology scaling on ultra low voltage clock network design are also investigated.

CHAPTER IX

Conclusions

This work focused on extremely power constrained IC design for cubic millimeter sensing systems. Ultra low power consumption is a key enabler to scale the volume of power sources and increase lifetime of cubic millimeter systems. The use of very low voltages has gained favor recently due to the large potential energy savings, however there are several challenges to be addressed before it is widely accepted as a design practice. These challenges include standby power reduction, performance improvement, variability mitigation, and analog design under severe power constraints.

This work proposed a range of circuit and architecture techniques to overcome these challenges, building upon ultra low voltage operations. The effectiveness of the proposed techniques is confirmed with numerous silicon demonstrations. Chapter II discusses the design of a picowatt sensing platform, called Phoenix Processor, with the emphasis of minimizing standby power in microprocessors and embedded memory. Chapters III, IV, and V discuss in detail the standby power minimization techniques employed in the Phoenix Processor. Chapter VI describes the design and optimization of a ultra low power voltage reference, a key analog building block for cubic millimeter sensing systems. Finally, Chapters VII and VIII discuss circuit and architecture techniques to improve performance, variability, and energy efficiency beyond conventional voltage scaling, which are successfully demonstrated in a FFT

core in a 65nm CMOS technology.

The work presented in this thesis brings the vision of cubic millimeter sensing systems closer to reality. However, a significant amount of efforts and challenges still remains for developing cubic millimeter sensing systems. In the digital design domain, performance scalability, adaptability to cope with variability, reconfigurability, and robustness improvements are areas of interest. The analog domain requires further innovations in designing components such as low power radios, sensors, and sensor interfaces. Power sources are another area requiring breakthroughs. Along with advances in power sources such as micro-batteries and energy scavenging devices, smart management of multiple power sources and efficient voltage conversion techniques will greatly benefit cubic millimeter sensing systems.

APPENDICES

APPENDIX A

Related Publications

- Mingoo Seok, Dongsuk Jeon, Chaitali Chakrabarti, David Blaauw, Dennis Sylvester, "A 0.27V, 30MHz, 17.7nJ/transform 1024-pt Complex FFT Core with super-pipelining," *International Solid-State Circuits Conferences*, 2011, in press
- Gregory Chen, Hassan Ghaed, Razi-Ul Haque, Michael Wieckowski, Yejoong Kim, Gyouho Kim, David Fick, Daeyeon Kim, Mingoo Seok, Kensall Wise, David Blaauw, Dennis Sylvester, "A 1 Cubic Millimeter Energy-Autonomous Wireless Intraocular Pressure Monitor," *International Solid-State Circuits Conferences*, 2011, in press
- Mingoo Seok, Gyouho Kim, David Blaauw, Dennis Sylvester, "Variability Analysis of a Digitally Trimmable Ultra-Low Power Voltage Reference," *European Solid-State Circuits Conference*, Sep, 2010
- Daeyeon Kim, Gregory K. Chen, Matthew Fojtik, Mingoo Seok, Dennis Sylvester, David Blaauw, "A Femtowatt-Scale Ultra-Low Leakage 10T SRAM with Speed Compensation Scheme," *International Symposium on Circuits and Systems*, 2011, submitted

- Mingoo Seok, David Blaauw, Dennis Sylvester, "Clock Network Design for Ultra-Low Power Applications," *International Symposium on Low Power Electronics and Design*, Aug, 2010
- Mingoo Seok, Scott Hanson, Michael Wieckowski, Gregory K. Chen, Yu-Shiang Lin, David Blaauw, Dennis Sylvester, "Circuit Design Advances to Enable Ubiquitous Sensing Environments," *International Symposium on Circuits and Systems*, invited, 2010
- Gregory K. Chen, Matthew Fojtik, Daeyeon Kim, David Fick, Junsun Park, Mingoo Seok, Mao-Ter Chen, Zhiyoong Foo, Dennis Sylvester, David Blaauw, "A Millimeter-Scale Near-Perpetual Sensor System with Stacked Battery and Solar Cells," *International Solid-State Circuits Conference*, 2010
- Mingoo Seok, Gyouho Kim, Dennis Sylvester, David Blaauw, "A 0.5V 2.2pW 2-Transistor Voltage Reference," *Custom Integrated Circuit Conference*, 2009
- Michael Wieckowski, Gregory K. Chen, Mingoo Seok, Dennis Sylvester, David Blaauw, "A Hybrid DC-DC Converter for Nanoampere Sub-1V Implantable Applications," *Symposium on VLSI Circuits*, 2009
- Mingoo Seok, Scott Hanson, Yu-Shiang Lin, Zhiyoong Foo, Daeyeon Kim, Yoonmyung Lee, Nurrachman Liu, Dennis Sylvester, David Blaauw, "Phoenix: an Ultra-Low Power Processor for Cubic Millimeter Sensor Systems," *ACM/IEEE Design Automation Conference*, 2009 [DAC/ISSCC Student Design Contest Winner]
- Dennis Sylvester, Scott Hanson, Mingoo Seok, Yu-Shiang Lin, David Blaauw, "Designing Robust Ultra-Low Power Circuits," *International Electron Device Meetings*, invited, 2008
- Mingoo Seok, Scott Hanson, Jae-sun Seo, Dennis Sylvester, David Blaauw "Robust Ultra-low Voltage ROM Design," *Custom Integrated Circuit Conference*, 2008

- Yoonmyung Lee, Mingoo Seok, Scott Hanson, David Blaauw, Dennis Sylvester "Standby Power Reduction Techniques for Ultra-Low Power Processors," *European Solid-State Circuits Conference*, 2008
- Mingoo Seok, Dennis Sylvester, David Blaauw, "Optimal Technology Selection for Minimizing Energy and Variability in Low Voltage Applications," *International Symposium on Low Power Electronics and Design*, 2008
- Mingoo Seok, Scott Hanson, Yu-Shiang Lin, Zhiyoong Foo, Daeyeon Kim, Yoonmyung Lee, Nurrachman Liu, Dennis Sylvester, David Blaauw, "The Phoenix Processor: A 30pW Platform for Sensor Applications," *IEEE Symposium on VLSI Circuits*, 2008
- Mingoo Seok, Scott Hanson, Dennis Sylvester, David Blaauw, "Analysis and Optimization of Sleep modes in Subthreshold Circuit Design," *ACM/IEEE Design Automation Conference*, 2007
- Scott Hanson, Mingoo Seok, David Blaauw, Dennis Sylvester, "Nanometer Device Scaling in Subthreshold Circuits," *ACM/IEEE Design Automation Conference*, 2007
- Mingoo Seok, Scott Hanson, David Blaauw, Dennis Sylvester, "Sleep Mode Analysis and Optimization with Minimal-Sized Power Gating Switch for Ultra-low Vdd Operations," *Transactions on VLSI systems*, submitted, 2010
- Scott Hanson, Mingoo Seok, Yu-shiang Lin, Zhiyoong Foo, Daeyeon Kim, Yoonmyung Lee, Nurrachman Liu, Dennis Sylvester, David Blaauw, "A Low-Voltage Processor for Sensing Applications With Picowatt Standby Mode," *Journal of Solid State Circuits*, Apr., 2009
- Scott Hanson, Bo Zhai, Mingoo Seok, Brian Cline, Kevin Zhou, Meghna Singhal, Michael Minuth, Javin Olson, Leyla Nazhandali, Todd Austin, Dennis Sylvester,

David Blaauw, "Exploring Variability and Performance in a Sub-200mV Processor," *Journal of Solid State Circuits*, Apr., 2008

- Scott Hanson, Bo Zhai, Mingoo Seok, Brian Cline, Kevin Zhou, Meghna Singhal, Michael Minuth, Javin Olson, Leyla Nazhandali, Todd Austin, Dennis Sylvester, David Blaauw, "Performance and Variability Optimization Strategies in a 150mV processor," *IEEE Symposium on VLSI Circuits*, 2007
- Scott Hanson, Mingoo Seok, Dennis Sylvester, David Blaauw, "Nanometer Device Scaling in Subthreshold Logic and SRAM," *Transactions on Electron Devices*, 2007

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] 'National Oceanic Atmospheric Administration'. Muskegon Meteorological Data. Aug. 2008.
- [2] I.F. Akyildiz, Weilian Su, Y. Sankarasubramaniam, and E. Cayirci. A survey on sensor networks. *Communications Magazine, IEEE*, 40(8):102 – 114, aug. 2002.
- [3] A.-J. Annema. Low-power bandgap references featuring dtmosts. *Solid-State Circuits, IEEE Journal of*, 34(7):949 –955, jul. 1999.
- [4] A.J. Annema, P. Veldhorst, G. Doornbos, and B. Nauta. A sub-1v bandgap voltage reference in 32nm finfet technology. pages 332 –333, feb. 2009.
- [5] ARM. ARM CPUs. <http://www.arm.com/products/CPUs>, 2010.
- [6] H. Banba, H. Shiga, A. Umezawa, T. Miyaba, T. Tanzawa, S. Atsumi, and K. Sakui. A cmos bandgap reference circuit with sub-1-v operation. *Solid-State Circuits, IEEE Journal of*, 34(5):670 –674, may. 1999.
- [7] J. B. Bates, N. J. Dudney, B. Neudecker, A. Ueda, and C. D. Evans. Thin-film lithium and lithium-ion batteries. *Solid State Ionics*, 135(1-4):33 – 45, 2000.
- [8] R.A. Blauschild, P.A. Tucci, R.S. Muller, and R.G. Meyer. A new nmos temperature-stable voltage reference. *Solid-State Circuits, IEEE Journal of*, 13(6):767 – 774, dec. 1978.
- [9] D. Bol, R. Ambroise, D. Flandre, and J.-D. Legat. Analysis and minimization of practical energy in 45nm subthreshold logic circuits. pages 294 –300, oct. 2008.
- [10] D. Bol, C. Hocquet, D. Flandre, and J.-D. Legat. Robustness-aware sleep transistor engineering for power-gated nanometer subthreshold circuits. pages 1484 –1487, may. 2010.
- [11] David Bol, Denis Flandre, and Jean-Didier Legat. Technology flavor selection and adaptive techniques for timing-constrained 45nm subthreshold circuits. In *ISLPED '09: Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, pages 21–26, New York, NY, USA, 2009. ACM.

- [12] A. Boni. Op-amps and startup circuits for cmos bandgap references with near 1-v supply. *Solid-State Circuits, IEEE Journal of*, 37(10):1339 – 1343, oct. 2002.
- [13] S. Borkar. Low power design challenges for the decade. *Design Automation Conference, 2001. Proceedings of the ASP-DAC 2001. Asia and South Pacific*, pages 293–296, 2001.
- [14] A.P. Brokaw. A simple three-terminal ic bandgap reference. *Solid-State Circuits, IEEE Journal of*, 9(6):388 – 393, dec. 1974.
- [15] B.H. Calhoun and A. Chandrakasan. Characterizing and modeling minimum energy operation for subthreshold circuits. *Low Power Electronics and Design, 2004. ISLPED '04. Proceedings of the 2004 International Symposium on*, pages 90–95, 2004.
- [16] B.H. Calhoun and A. Chandrakasan. A 256kb sub-threshold sram in 65nm cmos. *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, pages 2592–2601, Feb. 2006.
- [17] B.H. Calhoun, A. Wang, and A. Chandrakasan. Modeling and sizing for minimum energy operation in subthreshold circuits. *Solid-State Circuits, IEEE Journal of*, 40(9):1778–1786, Sept. 2005.
- [18] CardioMEMS. EndoSure Wireless AAA Pressure Sensor. <http://www.cardiomems.com>, 2008.
- [19] Anantha Chandrakasan. Design of High Performance Microprocessor Circuits. *IEEE Press*, 2000.
- [20] L. Chang, D.M. Fried, J. Hergenrother, J.W. Sleight, R.H. Dennard, R.K. Montoye, L. Sekaric, S.J. McNab, A.W. Topol, C.D. Adams, K.W. Guarini, and W. Haensch. Stable sram cell design for the 32 nm node and beyond. *VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium on*, pages 128–129, June 2005.
- [21] Leland Chang, Y. Nakamura, R.K. Montoye, J. Sawada, A.K. Martin, K. Kinoshita, F.H. Gebara, K.B. Agarwal, D.J. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek. A 5.3ghz 8t-sram with operation down to 0.41v in 65nm cmos. pages 252 –253, jun. 2007.
- [22] M.H. Chang, J.K. Ting, J.S. Shy, L. Chen, C.W. Liu, J.Y. Wu, K.H. Pan, C.S. Hou, C.C. Tu, Y.H. Chen, S.L. Sue, S.M. Jang, S.C. Yang, C.S. Tsai, C.H. Chen, H.J. Tao, C.C. Tsai, H.C. Hsieh, Y.Y. Wang, R.Y. Chang, K.B. Cheng, T.Y. Chu, T.N. Yen, P.S. Wang, J.W. Weng, J.H. Hsu, Y.S. Ho, C.H. Ho, Y.C. Huang, R.Y. Shiue, B.K. Liew, C.H. Yu, S.C. Sun, and J.Y.C. Sun. A highly manufacturable 0.25 μm multiple-vt dual gate oxide cmos process for logic/embedded ic foundry technology. pages 150 –151, jun. 1998.

- [23] G. Chen, M. Fojtik, Daeyeon Kim, D. Fick, Junsun Park, Mingoo Seok, Mao-Ter Chen, Zhiyoong Foo, D. Sylvester, and D. Blaauw. Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells. pages 288–289, feb. 2010.
- [24] G.K. Chen, D. Blaauw, T. Mudge, D. Sylvester, and Nam Sung Kim. Yield-driven near-threshold sram design. *Computer-Aided Design, 2007. ICCAD 2007. IEEE/ACM International Conference on*, pages 660–666, Nov. 2007.
- [25] Yuan Chen, Yu-Wei Lin, Yu-Chi Tsao, and Chen-Yi Lee. A 2.4-gsample/s dvfs fft processor for mimo ofdm communication systems. *Solid-State Circuits, IEEE Journal of*, 43(5):1260–1273, may. 2008.
- [26] D.G. Chinnery and K. Keutzer. Closing the gap between asic and custom: an asic perspective. *Design Automation Conference, 2000. Proceedings 2000. 37th*, pages 637–642, 2000.
- [27] CoretexM0. ARM Cortex M0. 2008.
- [28] Crossbow-Technology. eKo Pro Series System for Environmental Monitoring. <http://www.xbow.com/Eko/index.aspx>, 2008.
- [29] G. De Vita and G. Iannaccone. A sub-1-v, 10 ppm/°c, nanopower voltage reference generator. *Solid-State Circuits, IEEE Journal of*, 42(7):1536–1542, July 2007.
- [30] J. Doyle, Young Jun Lee, Yong-Bin Kim, H. Wilsch, and F. Lombardi. A cmos subbandgap reference circuit with 1-v power supply voltage. *Solid-State Circuits, IEEE Journal of*, 39(1):252–255, jan. 2004.
- [31] Elodie Ebrard, Bruno Allard, Philippe Candelier, and Patrice Waltz. Review of fuse and antifuse solutions for advanced standard cmos technologies. *Microelectron. J.*, 40(12):1755–1765, 2009.
- [32] Freescale-Semiconductor. MPXY8300; Microcontroller, Pressure Sensor, X-Z Accelerometer and RF Transmitter. <http://www.freescale.com/>, 2008.
- [33] G. Giustolisi, G. Palumbo, M. Criscione, and F. Cutri. A low-voltage low-power voltage reference based on subthreshold mosfets. *Solid-State Circuits, IEEE Journal of*, 38(1):151–154, jan. 2003.
- [34] A. Broman H. Quigley. The number of people with glaucoma worldwide in 2010 and 2020. pages 262–267, 2006.
- [35] S. Hanson, Mingoo Seok, D. Sylvester, and D. Blaauw. Nanometer device scaling in subthreshold circuits. *Design Automation Conference, 2007. DAC '07. 44th ACM/IEEE*, pages 700–705, June 2007.

- [36] S. Hanson, Bo Zhai, Mingoo Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw. Performance and variability optimization strategies in a sub-200mv, 3.5pj/inst, 11nw sub-threshold processor. *VLSI Circuits, 2007 IEEE Symposium on*, pages 152–153, June 2007.
- [37] S. Hanson, Bo Zhai, Mingoo Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw. Exploring variability and performance in a sub-200-mv processor. *Solid-State Circuits, IEEE Journal of*, 43(4):881–891, April 2008.
- [38] R.R. Harrison. The design of integrated circuits to observe brain activity. *Proceedings of the IEEE*, 96(7):1203–1216, jul. 2008.
- [39] S.K. Hsu, A. Agarwal, S.K. Mathew, R.K. Krishnamurthy, M. Hansson, and A. Alvandpour. A 9ghz 320x80bit low leakage microcode read only memory in 65nm cmos. *Solid-State Circuits Conference, 2006. ESSCIRC 2006. Proceedings of the 32nd European*, pages 299–302, Sept. 2006.
- [40] Hong-Wei Huang, Chun-Yu Hsieh, Ke-Horng Chen, and Sy-Yen Kuo. A 1v 16.9ppm/ $^{\circ}$ c 250na switched-capacitor cmos voltage reference. *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, pages 438–626, Feb. 2008.
- [41] Myeong-Eun Hwang, A. Raychowdhury, Keejong Kim, and K. Roy. A 85mV 40nW Process-Tolerant Subthreshold 8x8 FIR Filter in 130nm Technology. *VLSI Circuits, 2007 IEEE Symposium on*, pages 154–155, June 2007.
- [42] S. C. Jocke, J. F. Bolus, S. N. Wooters, A. D. Jurik, A. C. Weaver, T. N. Blalock, and B. H. Calhoun. A 2.6- μ w sub-threshold mixed-signal ecg soc. pages 60–61, jun. 2009.
- [43] Yunho Jung, Hongil Yoon, and Jaeseok Kim. New efficient fft algorithm and pipeline implementation results for ofdm/dmt applications. *Consumer Electronics, IEEE Transactions on*, 49(1):14–20, feb. 2003.
- [44] James Kao, Anantha Chandrakasan, and Dimitri Antoniadis. Transistor sizing issues and tool for multi-threshold cmos technology. *DAC '97: Proceedings of the 34th annual conference on Design automation*, pages 409–414, 1997.
- [45] J.T. Kao and A.P. Chandrakasan. Dual-threshold voltage techniques for low-power digital circuits. *Solid-State Circuits, IEEE Journal of*, 35(7):1009–1018, jul. 2000.
- [46] H. Kaul, M.A. Anders, S.K. Mathew, S.K. Hsu, A. Agarwal, R.K. Krishnamurthy, and S. Borkar. A 320 mv 56 uw 411 gops/watt ultra-low voltage motion estimation accelerator in 65 nm cmos. *Solid-State Circuits, IEEE Journal of*, 44(1):107–114, Jan. 2009.

- [47] H. Kawaguchi, K.-I. Nose, and T. Sakurai. A cmos scheme for 0.5 v supply voltage with pico-ampere standby current. pages 192 –193, 436, feb. 1998.
- [48] John Keane, Hanyong Eom, Tae-Hyoung Kim, Sachin Sapatnekar, and Chris Kim. Subthreshold logical effort: a systematic framework for optimal sub-threshold device sizing. *DAC '06: Proceedings of the 43rd annual conference on Design automation*, pages 425–428, 2006.
- [49] Jonggab Kil, Jie Gu, and Chris H. Kim. A high-speed variation-tolerant interconnect technique for sub threshold circuits using capacitive boosting. In *ISLPED '06: Proceedings of the 2006 international symposium on Low power electronics and design*, pages 67–72, New York, NY, USA, 2006. ACM.
- [50] Chris H. Kim, Jae-Joon Kim, Saibal Mukhopadhyay, and Kaushik Roy. A forward body-biased low-leakage sram cache: device and architecture considerations. *ISLPED '03: Proceedings of the 2003 international symposium on Low power electronics and design*, pages 6–9, 2003.
- [51] Chris H. Kim and Kaushik Roy. Dynamic vt sram: a leakage tolerant cache memory for low voltage microprocessors. *ISLPED '02: Proceedings of the 2002 international symposium on Low power electronics and design*, pages 251–254, 2002.
- [52] Nam Sung Kim, K. Flautner, D. Blaauw, and T. Mudge. Circuit and microarchitectural techniques for reducing cache leakage power. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 12(2):167–184, Feb. 2004.
- [53] Tae-Hyoung Kim, J. Keane, Hanyong Eom, and C.H. Kim. Utilizing reverse short-channel effect for optimal subthreshold circuit design. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 15(7):821–829, July 2007.
- [54] P. Kinget, C. Vezyrtzis, E. Chiang, B. Hung, and T.L. Li. Voltage references for ultra-low supply voltages. *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, pages 715–720, Sept. 2008.
- [55] H. Kulah and K. Najafi. An electromagnetic micro power generator for low-frequency environmental vibrations. pages 237 – 240, 2004.
- [56] J. Kwong and A.P. Chandrakasan. Variation-driven device sizing for minimum energy sub-threshold circuits. *Low Power Electronics and Design, 2006. ISLPED'06. Proceedings of the 2006 International Symposium on*, pages 8–13, Oct. 2006.
- [57] Joyce Kwong, Yogesh Ramadass, Naveen Verma, Markus Koesler, Korbinian Huber, Hans Moormann, and Anantha Chandrakasan. A 65nm sub-vt micro-controller with integrated sram and switched-capacitor dc-dc converter. *IEEE International Solid-State Circuits Conference*, pages 318–616, 2008.

- [58] Yoonmyung Lee, Mingoo Seok, S. Hanson, D. Blaauw, and D. Sylvester. Standby power reduction techniques for ultra-low power processors. pages 186–189, sep. 2008.
- [59] Charles Lefurgy, Peter Bird, I-Cheng Chen, and Trevor Mudge. Improving code density using compression techniques. *MICRO 30: Proceedings of the 30th annual ACM/IEEE international symposium on Microarchitecture*, pages 194–203, 1997.
- [60] Ka Nang Leung and P.K.T. Mok. A sub-1-v 15-ppm/deg;cmos bandgap voltage reference without requiring low threshold voltage device. *Solid-State Circuits, IEEE Journal of*, 37(4):526–530, apr. 2002.
- [61] Ka Nang Leung and P.K.T. Mok. A cmos voltage reference based on weighted vgs for cmos low-dropout linear regulators. *Solid-State Circuits, IEEE Journal of*, 38(1):146–150, Jan 2003.
- [62] Saihua Lin, Yu Wang, Rong Luo, and Huazhong Yang. A capacitive boosted buffer technique for high-speed process-variation-tolerant interconnect in udvs application. In *ASP-DAC '08: Proceedings of the 2008 Asia and South Pacific Design Automation Conference*, pages 304–309, Los Alamitos, CA, USA, 2008. IEEE Computer Society Press.
- [63] Yu-Shiang Lin, S. Hanson, F. Albano, C. Tokunaga, R.-U. Haque, K. Wise, A.M. Sastry, D. Blaauw, and D. Sylvester. Low-voltage circuit design for widespread sensing applications. *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, pages 2558–2561, May 2008.
- [64] Ying Liu, Sani R. Nassif, Lawrence T. Pileggi, and Andrzej J. Strojwas. Impact of interconnect variations on the clock skew of a gigahertz microprocessor. In *DAC '00: Proceedings of the 37th Annual Design Automation Conference*, pages 168–171, New York, NY, USA, 2000. ACM.
- [65] Nir Magen, Avinoam Kolodny, Uri Weiser, and Nachum Shamir. Interconnect-power dissipation in a microprocessor. In *SLIP '04: Proceedings of the 2004 international workshop on System level interconnect prediction*, pages 7–13, New York, NY, USA, 2004. ACM.
- [66] J.D. Meindl and J.A. Davis. The fundamental limit on binary switching energy for terascale integration (tsi). *Solid-State Circuits, IEEE Journal of*, 35(10):1515–1516, Oct 2000.
- [67] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, and J. Yamada. 1v high-speed digital circuit technology with 0.5 μm multi-threshold cmos. pages 186–189, sep. 1993.
- [68] Siva Narendra, Vivek De, Dimitri Antoniadis, Anantha Chandrakasan, and Shekhar Borkar. Scaling of stack effect and its application for leakage reduction.

- ISLPED '01: Proceedings of the 2001 international symposium on Low power electronics and design*, pages 195–200, 2001.
- [69] L. Nazhandali, M. Minuth, and T. Austin. Sensebench: toward an accurate evaluation of sensor network processors. *Workload Characterization Symposium, 2005. Proceedings of the IEEE International*, pages 197–203, Oct. 2005.
- [70] L. Nazhandali, B. Zhai, A. Olson, A. Reeves, M. Minuth, R. Helfand, Sanjay Pant, T. Austin, and D. Blaauw. Energy optimization of subthreshold-voltage sensor network processors. *Computer Architecture, 2005. ISCA '05. Proceedings. 32nd International Symposium on*, pages 197–207, June 2005.
- [71] H.J. Oguey and B. Gerber. Mos voltage reference based on polysilicon gate work function difference. *Solid-State Circuits, IEEE Journal of*, 15(3):264 – 269, jun. 1980.
- [72] G. Ono, T. Nakagawa, R. Fujiwara, T. Norimatsu, T. Terada, M. Miyazaki, K. Suzuki, K. Yano, Y. Ogata, A. Macki, S. Kobayashi, N. Koshizuka, and K. Sakamura. 1-cc computer: Cross-layer integration with 3.4-nw/bps link and 22-cm locationing. *VLSI Circuits, 2007 IEEE Symposium on*, pages 90–91, June 2007.
- [73] OpenCores. OpenCores. <http://www.opencores.org>.
- [74] B.C. Paul, A. Raychowdhury, and K. Roy. Device optimization for digital sub-threshold logic operation. *Electron Devices, IEEE Transactions on*, 52(2):237–247, Feb. 2005.
- [75] M.J.M. Pelgrom, A.C.J. Duinmaijer, and A.P.G. Welbers. Matching properties of mos transistors. *Solid-State Circuits, IEEE Journal of*, 24(5):1433–1439, Oct 1989.
- [76] IBM PowerPC. IBM Power PC. <http://www.chips.ibm.com/products/powerpc>.
- [77] Hao Qu and J. Gotman. A patient-specific algorithm for the detection of seizure onset in long-term eeg monitoring: possible use as a warning device. *Biomedical Engineering, IEEE Transactions on*, 44(2):115 –122, feb. 1997.
- [78] Y.K. Ramadass and A.P. Chandrakasan. Minimum energy tracking loop with embedded dc-dc converter delivering voltages down to 250mv in 65nm cmos. pages 64 –587, feb. 2007.
- [79] B. Razavi. Design of Analog CMOS Integrated Circuits. *McGraw-Hill*, 2001.
- [80] P.J. Restle and A. Deutsch. Designing the best clock distribution network. pages 2 –5, jun. 1998.
- [81] T. Sakurai. Closed-form expressions for interconnection delay, coupling, and crosstalk in vlsis. *Electron Devices, IEEE Transactions on*, 40(1):118 –124, jan. 1993.

- [82] U. Schnakenberg, P. Walter, G. vom Bogel, C. Kruger, H.C. Ludtke-Handjery, H.A. Richter, W. Specht, P. Ruokonen, and W. Mokwa. Initial investigations on systems for measuring intraocular pressure. *Sensors and Actuators*, 85(1–3):287–291, 2000.
- [83] Mingoo Seok, S. Hanson, Yu-Shiang Lin, Zhiyoong Foo, Daeyeon Kim, Yoonmyung Lee, Nurrachman Liu, D. Sylvester, and D. Blaauw. The phoenix processor: A 30pw platform for sensor applications. *VLSI Circuits, 2008 IEEE Symposium on*, pages 188–189, June 2008.
- [84] Mingoo Seok, Scott Hanson, Dennis Sylvester, and David Blaauw. Analysis and optimization of sleep modes in subthreshold circuit design. *DAC '07: Proceedings of the 44th annual conference on Design automation*, pages 694–699, 2007.
- [85] Mingoo Seok, Gyouho Kim, D. Blaauw, and D. Sylvester. Variability analysis of a digitally trimmable ultra-low power voltage reference. sep. 2010.
- [86] Mingoo Seok, Gyouho Kim, D. Sylvester, and D. Blaauw. A 0.5v 2.2pw 2-transistor voltage reference. pages 577–580, sep. 2009.
- [87] Mingoo Seok, Dennis Sylvester, and David Blaauw. Optimal technology selection for minimizing energy and variability in low voltage applications. *ISLPED '08: Proceeding of the thirteenth international symposium on Low power electronics and design*, pages 9–14, 2008.
- [88] M. Sheets, F. Burghardt, T. Karalar, J. Ammer, Y.H. Chee, and J. Rabaey. A power-managed protocol processor for wireless sensor networks. pages 212–213, 2006.
- [89] B.S. Song and P.R. Gray. A precision curvature-compensated cmos bandgap reference. *Solid-State Circuits, IEEE Journal of*, 18(6):634–643, Dec 1983.
- [90] S.R. Sridhara, M. DiRenzo, S. Lingam, Seok-Jun Lee, R. Blazquez, J. Maxey, S. Ghanem, Yu-Hung Lee, R. Abdallah, P. Singh, and M. Goe. Microwatt embedded processor platform for medical system-on-chip applications. pages 15–16, jun. 2010.
- [91] Y. Takeyama, H. Otake, O. Hirabayashi, K. Kushida, and N. Otsuka. A low leakage sram macro with replica cell biasing scheme. *Solid-State Circuits, IEEE Journal of*, 41(4):815–822, April 2006.
- [92] H. Tanaka, Y. Nakagome, J. Etoh, E. Yamasaki, M. Aoki, and K. Miyazawa. Sub-1- μ a dynamic reference voltage generator for battery-operated drams. *Solid-State Circuits, IEEE Journal of*, 29(4):448–453, apr. 1994.
- [93] Y. Taur and T. H. Ning. Fundamentals of Modern VLSI Devices. *Cambridge*, 1998.

- [94] Texas-Instruments. MSP430 Ultra-Low Power Microcontrollers. <http://focus.ti.com/mcu/>, 2008.
- [95] Jeremy R. Tolbert, Xin Zhao, Sung Kyu Lim, and Saibal Mukhopadhyay. Slew-aware clock tree design for reliable subthreshold circuits. In *ISLPED '09: Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, pages 15–20, New York, NY, USA, 2009. ACM.
- [96] Transmeta. Transmeta Crusoe. <http://www.transmeta.com>.
- [97] R.B. Tremaine, P.A. Franaszek, J.T. Robinson, C.O. Schulz, T.B. Smith, M.E. Wazlowski, and P.M. Bland. IBM Memory Expansion Technology (MXT). *IBM Journal of Research and Development*, 45(2):271–286, 2001.
- [98] M. Ugajin and T. Tsukahara. A 0.6-v voltage reference circuit based on sigma- ν architecture in cmos/simox. pages 141–142, 2001.
- [99] N. Verma, J. Kwong, and A.P. Chandrakasan. Nanometer mosfet variation in minimum energy subthreshold circuits. *Electron Devices, IEEE Transactions on*, 55(1):163–174, Jan. 2008.
- [100] Naveen Verma and A.P. Chandrakasan. A 65nm 8t sub-vt sram employing sense-amplifier redundancy. *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pages 328–606, Feb. 2007.
- [101] Naveen Verma, Ali Shoeb, John V. Guttag, and Anantha P. Chandrakasan. A micro-power eeg acquisition s^ocwith integrated seizure detection processor for continuous patient monitoring. pages 62–63, jun. 2009.
- [102] A. Wang and A. Chandrakasan. A 180mV FFT processor using subthreshold circuit techniques. *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*, pages 292–529 Vol.1, Feb. 2004.
- [103] A. Wang and A. Chandrakasan. A 180-mV subthreshold FFT processor using a minimum energy design methodology. *Solid-State Circuits, IEEE Journal of*, 40(1):310–319, Jan. 2005.
- [104] B. Warneke, M. Last, B. Liebowitz, and K.S.J. Pister. Smart dust: communicating with a cubic-millimeter computer. *Computer*, 34(1):44–51, Jan 2001.
- [105] B.A. Warneke and K.S.J. Pister. An ultra-low energy microcontroller for smart dust wireless sensor networks. pages 316 – 317 Vol.1, feb. 2004.
- [106] R.J. Widlar. New developments in ic voltage regulators. *Solid-State Circuits, IEEE Journal of*, 6(1):2 – 7, feb. 1971.
- [107] Michael Wieckowski, Gregory K. Chen, Mingoo Seok, David Blaauw, and Dennis Sylvester. A hybrid dc-dc converter for sub-microwatt sub-1v implantable applications. pages 166–167, jun. 2009.

- [108] K.D. Wise, A.M. Sodagar, Ying Yao, M.N. Gulari, G.E. Perlin, and K. Najafi. Microelectrodes, microelectronics, and implantable neural microsystems. *Proceedings of the IEEE*, 96(7):1184–1202, jul. 2008.
- [109] Intel XScale. Intel XScale. <http://www.intel.com/design/intelxscale>.
- [110] C. Yeh, G. Wilke, H. Chen, S. Reddy, H. Nguyen, T. Miyoshi, W. Walker, and R. Murgai. Clock distribution architectures: A comparative study. In *ISQED '06: Proceedings of the 7th International Symposium on Quality Electronic Design*, pages 85–91, Washington, DC, USA, 2006. IEEE Computer Society.
- [111] K.K. Young, S.Y. Wu, C.C. Wu, C.H. Wang, C.T. Lin, J.Y. Cheng, M. Chiang, S.H. Chen, T.C. Lo, Y.S. Chen, J.H. Chen, L.J. Chen, S.Y. Hou, J.J. Law, T.E. Chang, C.S. Hou, J. Shih, S.M. Jeng, H.C. Hsieh, Y. Ku, T. Yen, H. Tao, L.C. Chao, S. Shue, S.M. Jang, T.C. Ong, C.H. Yu, M.S. Liang, C.H. Diaz, and J.Y.C. Sun. A 0.13 μm cmos technology with 193 nm lithography and cu/low-k for high performance applications. pages 563–566, 2000.
- [112] E.T. Zellers, S. Reidy, R.A. Veeneman, R. Gordenker, W.H. Steinecker, G.R. Lambertus, Hanseup Kim, J.A. Potkay, M.P. Rowe, Qiongyan Zhong, C. Avery, H.K.L. Chan, R.D. Sacks, K. Najafi, and K.D. Wise. An integrated micro-analytical system for complex vapor mixtures. pages 1491–1496, jun. 2007.
- [113] Bo Zhai, David Blaauw, Dennis Sylvester, and Krisztian Flautner. Theoretical and practical limits of dynamic voltage scaling. *DAC '04: Proceedings of the 41st annual conference on Design automation*, pages 868–873, 2004.
- [114] Bo Zhai, David Blaauw, Dennis Sylvester, and Scott Hanson. A sub-200mv 6t sram in 130nm cmos. *IEEE International Solid-State Circuits Conference*, 2007.
- [115] Bo Zhai, Scott Hanson, David Blaauw, and Dennis Sylvester. Analysis and mitigation of variability in subthreshold design. *ISLPED '05: Proceedings of the 2005 international symposium on Low power electronics and design*, pages 20–25, 2005.
- [116] Bo Zhai, L. Nazhandali, J. Olson, A. Reeves, M. Minuth, R. Helfand, Sanjay Pant, D. Blaauw, and T. Austin. A 2.60pj/inst subthreshold sensor processor for optimal energy efficiency. *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, pages 154–155, 0-0 2006.