# Design and Analysis of Robust Low Voltage Static Random Access Memories

by

Daeyeon Kim

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2012

Doctoral Committee:

Professor Dennis Michael Sylvester, Chair
Professor David Blaauw
Professor Trevor N. Mudge
Assistant Professor Kenn Richard Oldham

To my family

# ACKNOWLEDGEMENTS

During my graduate study, I have been supported by many people both inside and outside of the University of Michigan. I would like begin by expressing my gratitude to my advisor, Professor Dennis Sylvester, for his tremendous support and guidance. He has been a great mentor throughout my Ph.D. by showing directions in my works, encouraging me when I was down, and sharing me his passion as a scholar, as an innovator, and as an advisor. I would also like to thank Professor David Blaauw for practically being a co-advisor. He always has been enthusiastic and supportive in discussing and sharing new ideas and implementing them. I would like to thank Professor Trevor Mudge and Professor Kenn Oldham for being a member of my dissertation committee members and helping review my dissertation.

Inside the university, I have worked with many intelligent and smart people. My special appreciation goes to Jae-sun Seo and Mingoo Seok. They have helped me in countless ways since I started my graduate study, even after they graduated from the university. Gregory Chen and Michael Wieckowski helped me to learn fundamental ideas related to SRAM design and tape-outs. Matthew Fojtik worked in designing CPU and preparing testing environments in Chapter III. Sudhir Satpathy and Bharan Giridhar helped me in developing ideas and designing circuits in Chapter IV. Yoonmyung Lee and I worked together in analyzing a new device in Chapter V. David Fick and Nathaniel Pinckney worked as student administrators of a server pool (vlsipool) for simulations and tape-outs. I also enjoyed having discussions with members in our research group: Mohammad Hassan Ghaed, Dongsuk Jeon, Yen-Po Chen, Suyoung

Bang, Prashant Singh, Nurrachman Liu, Zhi Yoong Foo, Gyouho Kim, Yejoong Kim, Inhee Lee, and Dongmin Yoon. Besides work, I would also like to thank all my friends in Ann Arbor for having a wonderful time and sharing memories.

I also have been supported from outside the university. During internships in the industry, co-works with many talented mentors deepened and extended my knowledge on state-of-the-art SRAM design. I would like to appreciate Dr. Robert Aitken, Dr. Vikas Chandra, and Mr. Cezary Pietrzyk at ARM, Inc. for mentoring me. I would also like to thank the Korea Foundation for Advanced Studies for funding my fellowship.

Last but not the least, I am grateful to my family for their love and support. I would like to express special gratitude to my wife, Mi Yon Kim, for her love, support, and patience during my graduate study.

# TABLE OF CONTENTS

DEDICATION . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ii

ACKNOWLEDGEMENTS . . . . . . . . . . . . . . . . . . . . . . . . . . iii

LIST OF FIGURES . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . viii

LIST OF TABLES . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xii

LIST OF ABBREVIATIONS . . . . . . . . . . . . . . . . . . . . . . . . . xiii

ABSTRACT . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xv

CHAPTER

I. Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.1 Technology Scaling and Low Power Design . . . . . . . . . . . 2
1.2 Challenges of Designing Low Voltage SRAM . . . . . . . . . . 4
1.3 Contributions of This Work . . . . . . . . . . . . . . . . . . . . 6

II. Writability Analysis for Bit-Interleaved 8-T SRAMs . . . . . 10

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
2.2 Background and Related Works . . . . . . . . . . . . . . . . . 13
2.2.1 Write Failure and Read Disturb . . . . . . . . . . . 13
2.2.2 Static Noise Margin Analysis . . . . . . . . . . . . 15
2.2.3 Soft Error and Half Select Disturb . . . . . . . . . . 19
2.2.4 Writability vs. Half Select Disturb in Bit-Interleaved 8-T SRAMs . . . . . . . . . . . . . . . . . . . . 20
2.3 Writability Analysis Method . . . . . . . . . . . . . . . . . . 21
2.3.1 SRAM Dynamic Writability Metric . . . . . . . . . 21
2.3.2 SRAM Worst Case Corner Simulation . . . . . . . . 22
2.3.3 Bit-Interleaved Writability Analysis . . . . . . . . . 26
2.4 Writability Analysis at Near-Threshold . . . . . . . . . . . . 27
2.4.1 Word-line Boosting . . . . . . . . . . . . . . . . . . 27

2.4.2 Device Sizing Optimization . . . . . . . . . . . . . . 28
2.4.3 Dual Writability Enhancement . . . . . . . . . . . 31
2.5 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . 32

**III. Design of Low Leakage SRAM** . . . . . . . . . . . . . . . . . 33

3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . 33
3.2 Background and Related Works . . . . . . . . . . . . . . . . 35
3.2.1 Leakage Mechanisms . . . . . . . . . . . . . . . . . 35
3.2.2 Leakage Reduction Techniques . . . . . . . . . . . . 37
3.2.3 Technology Selection for Leakage Reduction . . . . 44
3.3 Design of Low Leakage SRAM . . . . . . . . . . . . . . . . 44
3.3.1 SRAM Bitcell and Operation Modes . . . . . . . . . 44
3.3.2 Bit-line Boosting for Fast Write Operation . . . . . 46
3.3.3 Body Biasing with Boosted Supply . . . . . . . . . 47
3.3.4 Leakage Reduction during Standby Mode . . . . . . 48
3.3.5 Read Buffer Design . . . . . . . . . . . . . . . . . 51
3.4 Measurement Results . . . . . . . . . . . . . . . . . . . . 52
3.5 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . 56

**IV. Adaptive Write Architecture for Low Voltage 8-T SRAMs** . 58

4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . 59
4.2 Background and Related Works . . . . . . . . . . . . . . . . 60
4.2.1 Variation at Low Voltage . . . . . . . . . . . . . . 60
4.2.2 8-T SRAM Operations at Low Voltage . . . . . . . 60
4.2.3 8-T SRAM Write Time . . . . . . . . . . . . . . . . 62
4.2.4 8-T SRAM Static Failures and Half Select Disturb . 66
4.2.5 Previous Works . . . . . . . . . . . . . . . . . . . 66
4.3 Design of Adaptive Write Architecture . . . . . . . . . . . . 69
4.3.1 Write Word-Line Pulse Width Modulation . . . . . 69
4.3.2 Write Word-Line Voltage Level Modulation . . . . . 70
4.4 Measurement Results . . . . . . . . . . . . . . . . . . . . 72
4.4.1 Prototype Implementation . . . . . . . . . . . . . . 72
4.4.2 Measurement Data . . . . . . . . . . . . . . . . . . 74
4.5 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . 80

**V. Low Power Circuit Design Based on Heterojunction Tunneling Transistors** . . . . . . . . . . . . . . . . . . . . . . . . . 81

5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . 82
5.2 Background and Related Works . . . . . . . . . . . . . . . . 84
5.2.1 Necessity for Steep Subthreshold Swing . . . . . . . 84
5.2.2 Devices with Steep Subthreshold Swing . . . . . . . 86
5.3 HETT Device Physics and Modeling . . . . . . . . . . . . . 86

5.3.1 HETT Device Physics . . . . . . . . . . . . . . . . . . . 86
5.3.2 HETT Device Modeling . . . . . . . . . . . . . . . . . . . 88
5.4 HETT-Based Circuit Analysis . . . . . . . . . . . . . . . . . . . . . 90
5.4.1 Dynamic Power Reduction . . . . . . . . . . . . . . . 90
5.4.2 Limitations of HETT-Based Circuits . . . . . . . . . . 92
5.4.3 Increased Miller Capacitance . . . . . . . . . . . . . 97
5.5 HETT-Based SRAM Design . . . . . . . . . . . . . . . . . . . . . 100
5.5.1 Limitations in Standard 6-T SRAM . . . . . . . . . . 100
5.5.2 HETT-Based 7-T SRAM . . . . . . . . . . . . . . . . 104
5.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . 108

**VI. Conclusion** . . . . . . . . . . . . . . . . . . . . . . . . . . . 110

**BIBLIOGRAPHY** . . . . . . . . . . . . . . . . . . . . . . . . . . . 112

# LIST OF FIGURES

**Figure**

1.1 Technology Scaling Trends of Supply Voltage . . . . . . . . . . . . . . 3

1.2 Die Photo of Intel High Performance 32nm Processor . . . . . . . . 5

1.3 Die Photo of 0.18$\mu$m Sensor System . . . . . . . . . . . . . . . . . . . . 5

1.4 Approaches for Analysis and Design of Robust Low Voltage SRAM 7

2.1 SRAM Write and Read Operations . . . . . . . . . . . . . . . . . . . 14

2.2 8-T SRAM Schematic . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

2.3 Static Noise Margin Schematic . . . . . . . . . . . . . . . . . . . . . . . 16

2.4 Butterfly Curves for Hold Margin . . . . . . . . . . . . . . . . . . . . . 17

2.5 Schematics for Measuring Read and Write SNM . . . . . . . . . . . 17

2.6 Butterfly Curves for Read/Write Margin . . . . . . . . . . . . . . . . . 18

2.7 Three Different Scenarios of Soft Errors . . . . . . . . . . . . . . . . . 19

2.8 Half Select Disturb in Bit-Interleaved Array . . . . . . . . . . . . . 20

2.9 The Definition of $T_{crit}$ . . . . . . . . . . . . . . . . . . . . . . . . . 22

2.10 Worst Case Corner Directions . . . . . . . . . . . . . . . . . . . . . . 24

2.11 Worst Case Corner Simulation Results . . . . . . . . . . . . . . . . . 25

2.12 Bit-interleaved Writability Analysis . . . . . . . . . . . . . . . . . . . 27

2.13 Bit-interleaved Writability as Voltage Scales . . . . . . . . . . . . . 28

2.14 Word-line Boosting Schematic . . . . . . . . . . . . . . . . . . . . . . 29

2.15 Bit-interleaved Writability with Word-line Boosting . . . . . . . . . 29

2.16 Bit-interleaved Writability with Device Sizing . . . . . . . . . . . . 30

2.17 Bit-interleaved Writability with Both Word-line Boosting and Device Sizing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

3.1 Low Leakage 10-T SRAM Layout . . . . . . . . . . . . . . . . . . . 35

3.2 Leakage Source in MOS devices . . . . . . . . . . . . . . . . . . . . 37

3.3 Transistor Stacking Schematic . . . . . . . . . . . . . . . . . . . . . 38

3.4 Leakage Reduction via Transistor Stacking . . . . . . . . . . . . . . 39

3.5 Power Gating Concept . . . . . . . . . . . . . . . . . . . . . . . . . 40

3.6 Supply Voltage Ramping Timing Diagram . . . . . . . . . . . . . . 42

3.7 Leakage Reduction by Supply Voltage Ramping . . . . . . . . . . . 43

3.8 Low Leakage 10-T SRAM Schematic . . . . . . . . . . . . . . . . . 45

3.9 Comparison Between Bit-line Boosting and Word-line Boosting . . . 46

3.10 Bit-line Boosting Simulation Results . . . . . . . . . . . . . . . . . 47

3.11 Comparison of Reverse Body Biasing and Stack Forcing . . . . . . . 48

3.12 SRAM Leakage Paths during Standby Mode . . . . . . . . . . . . . 49

3.13 The Effect of Bit-line Floating on Leakage Reduction . . . . . . . . 50

3.14 SRAM Leakage and PMOS Reverse Body Biasing . . . . . . . . . . 50

3.15 Design of Word-line Keeper . . . . . . . . . . . . . . . . . . . . . . 51

3.16 Two Different 4-T Read Buffers . . . . . . . . . . . . . . . . . . . . 52

3.17 Read Buffers Performance Comparison . . . . . . . . . . . . . . . . . 53

3.18 Measured SRAM Performance Improvement . . . . . . . . . . . . . 54

3.19 Measured SRAM Leakage . . . . . . . . . . . . . . . . . . . . . . . 55

3.20 Temperature Variation of Measured Leakage . . . . . . . . . . . . 56

3.21 Chip Micrograph and Dimension . . . . . . . . . . . . . . . . . . . 57

4.1 FO4 Delay over Supply . . . . . . . . . . . . . . . . . . . . . . . . 61

4.2 Read vs. Write at Low Voltage . . . . . . . . . . . . . . . . . . . . 61

4.3 SRAM Write Time . . . . . . . . . . . . . . . . . . . . . . . . . . 62

4.4 SRAM Write Time as Supply Scales . . . . . . . . . . . . . . . . . 63

4.5 SRAM Write Time Distribution at 0.65V and 1.0V . . . . . . . . . 65

4.6 SRAM Write Time Distribution at 0.4V with and without Boosting 67

4.7 SRAM Half Select with WWL Boosting . . . . . . . . . . . . . . . 68

4.8 High Level Timing Diagram of Adaptive Write Architecture . . . . 70

4.9 High Level Block Diagram of Adaptive Write Architecture . . . . . 71

4.10 Block Diagram of Prototype Implemation . . . . . . . . . . . . . . 73

4.11 A cascaded bit-line Structure . . . . . . . . . . . . . . . . . . . . . 75

4.12 Bit-Interleaved Bitcells Structure . . . . . . . . . . . . . . . . . . . 76

4.13 Chip Micrograph . . . . . . . . . . . . . . . . . . . . . . . . . . . 76

4.14 A Shmoo Plot to Find the $V_{min}$ with Normal Write/Read Operation 77

4.15 WWL Underdrive to Fix Half Select Disturb . . . . . . . . . . . . . 77

4.16 Failing Bits at 500mV with WWL Underdrive . . . . . . . . . . . . 78

4.17 WWL Shmoo at 0.5V . . . . . . . . . . . . . . . . . . . . . . . . . 79

5.1 Motivation on Developing Devices with Steep Subthreshold Swing . 85

5.2 Tunneling FET Device Concept . . . . . . . . . . . . . . . . . . . 87

5.3 CMOS-Compatible Implementation of Complementary Tunneling FETs 88

5.4 Device Symbols for HETTs . . . . . . . . . . . . . . . . . . . . . . . . 91

5.5 BSIM vs. Verilog-A . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 91

5.6 Comparison of Dynamic Power Consumption . . . . . . . . . . . . . 93

5.7 Forward and Reverse Bias Drain Current of HETT . . . . . . . . . . 95

5.8 Limitation of HETT-Based Pass-Gates . . . . . . . . . . . . . . . . . 96

5.9 HETT-Based Transmission-Gate . . . . . . . . . . . . . . . . . . . . . 97

5.10 Miller Capacitance of HETT . . . . . . . . . . . . . . . . . . . . . . . 98

5.11 $C_{gd}$ Comparison . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99

5.12 Overshoot effects in HETT . . . . . . . . . . . . . . . . . . . . . . . 100

5.13 CMOS 6-T Read and Write Operations . . . . . . . . . . . . . . . . . 101

5.14 HETT 6-T Read and Write Operations with Inward Direction Access Transistors . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102

5.15 Static Noise Margins of HETT 6-T SRAM . . . . . . . . . . . . . . . 103

5.16 HETT 6-T Read and Write Operations with Outward Direction Access Transistors . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104

5.17 Proposed HETT 7-T SRAM Structure . . . . . . . . . . . . . . . . . 105

5.18 Read Operation in 7-T SRAM Array . . . . . . . . . . . . . . . . . . 106

5.19 HETT 7-T Layout . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107

5.20 Read/Write Margin Comparison . . . . . . . . . . . . . . . . . . . . . 108

5.21 Standby Power Comparison . . . . . . . . . . . . . . . . . . . . . . . 109

# LIST OF TABLES

**Table**

1.1 Tradeoffs of SRAM Solutions . . . . . . . . . . . . . . . . . . . . . . . 6

3.1 Sleep Power Breakdown of a Sensor Application . . . . . . . . . . . 34

3.2 Body Biasing Operation . . . . . . . . . . . . . . . . . . . . . . . . 42

3.3 Low Leakage 10-T SRAM Operation Modes . . . . . . . . . . . . . 45

3.4 SRAM Failures According to Speed . . . . . . . . . . . . . . . . . . 53

4.1 Static Write Failures and Yield below 0.65V . . . . . . . . . . . . . 64

4.2 Dynamic Write Failures Increases as WWL Pulse Width Decreases at 0.65V . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 64

4.3 Voltage Level Modulation . . . . . . . . . . . . . . . . . . . . . . . 72

4.4 Supplies for Each $V_{min}$ Cases . . . . . . . . . . . . . . . . . . . . . . 79

4.5 Adaptive Write Architecture Power Measurement . . . . . . . . . . 80

# LIST OF ABBREVIATIONS

**BDW** Bit-interleaved Dynamic Writability

**BSIM** Berkeley Short-channel IGFET Model

**BIST** Built In Self Test

**BSW** Bit-interleaved Static Writability

**CMOS** Complementary Metal-Oxide-Semiconductor

**CPU** Central Processing Unit

**CSV** Comma-Separated Values

**DC** Direct Current

**DUT** Design Under Test

**DMEM** Data Memory

**DW** Dynamic Writability

**FO4** Fan-Out of 4

**HETT** Heterojunction Tunneling Transistor

**HVT** High Threshold Voltage

**IMEM** Instruction Memory

**LER** Line Edge Roughness

**LDO** Low-Dropout Regulator

**LVT** Low Threshold Voltage

**MOS** Metal-Oxide-Semiconductor

**MOSFET** Metal-Oxide-Semiconductor Field-Effect Transistor

**MTCMOS** Multithreshold-Voltage CMOS

**NAND** Negated AND

**NHETT** N-channel HETT

**NMOS** N-channel MOS

**NTC** Near-Threshold Computing

**PC** Personal Computer

**PDN** Pull-Down Network

**PMOS** P-channel MOS

**PHETT** P-channel HETT

**PUN** Pull-Up Network

**PVT** Process, Voltage, and Temperature

**SCN** Switched-Capacitor Networks

**SECDED** Single Error Correction/Double Error Detection

**SNM** Static Noise Margin

**SoC** System on Chip

**SRAM** Static Random Access Memory

**SW** Static Writability

**SVT** Standard Threshold Voltage

**RBL** Read Bit-Line

**RDF** Random Dopant Fluctuation

**RO** Ring Oscillator

**T-CAD** Technology Computer Aided Design

**VLSI** Very-Large-Scale Integration

**WBL** Write Bit-Line

**WBLB** Write Bit-Line Bar

**WL** Word-Line

**WWL** Write Word-Line

# ABSTRACT

Design and Analysis of Robust Low Voltage Static Random Access Memories

by

Daeyeon Kim

Chair: Dennis Michael Sylvester

Static Random Access Memory (SRAM) is an indispensable part of most modern VLSI designs and dominates silicon area in many applications. In scaled technologies, maintaining high SRAM yield becomes more challenging since they are particularly vulnerable to process variations due to 1) the minimum sized devices used in SRAM bitcells and 2) the large array sizes. At the same time, low power design is a key focus throughout the semiconductor industry. Since low voltage operation is one of the most effective ways to reduce power consumption due to its quadratic relationship to energy savings, lowering the minimum operating voltage ($V_{min}$) of SRAM has gained significant interest.

This thesis presents four different approaches to design and analyze robust low voltage SRAM: SRAM analysis method improvement, SRAM bitcell development, SRAM peripheral optimization, and advance device selection.

We first describe a novel yield estimation method for bit-interleaved voltage-scaled 8-T SRAMs. Instead of the traditional trade-off between write and read, the trade-off between write and half select disturb is analyzed. In addition, this analysis proposes a method to find an appropriate Write Word-Line (WWL) pulse width to maximize yield.

Second, low leakage 10-T SRAM with speed compensation scheme is proposed. During sleep mode of a sensor application, SRAM retaining data cannot be shut down so it is important to minimize leakage in SRAM. This work adopts several leakage reduction techniques while compensating performance.

Third, adaptive write architecture for low voltage 8-T SRAMs is proposed. By adaptively modulating WWL width and voltage level, it is possible to achieve low power consumption while maintaining high yield without excessive performance degradation.

Finally, low power circuit design based on heterojunction tunneling transistors (HETTs) is discussed. HETTs have a steep subthreshold swing beneficial for low voltage operation. Device modeling and design of logic and SRAM are proposed.

# CHAPTER I

# Introduction

There has been an ever growing necessity for battery-operated systems. Battery-operated systems include handheld devices as well as sensor applications. Many people have used cell phones and a significant portion of cell phones are now high performance smart phones. In addition to cell phones, battery-operated tablet Personal Computer (PC)s start encroaching into the territory of traditional desktop and laptop PCs. It is highly expected that the market of smart phones and tablet PCs will grow even more. A necessity for sensor applications also has increased. Health-monitoring sensors implanted in a human body and infrastructure monitoring sensor networks are good examples of sensor applications with high demand. To develop more powerful and smaller battery-operated systems, technology scaling and low power design has been two main driving forces.

Technology scaling has acted an important role in the design of high performance System on Chip (SoC) for the past several decades by integrating more devices in a smaller area. First of all, technology scaling makes it possible to build a high performance Central Processing Unit (CPU) which runs at several GHz. Next, it also makes it possible to build a smaller system which has the same or even higher performance than before. For example, an intraocular sensor implanted in a human eye for curing glaucoma [13] needs a small form-factor and technology scaling enables it.

Low power design is indispensible to realize battery-operated systems. It is impossible to use a large battery in a handheld device or a sensor application due to limited sizes of them and a short battery lifetime of a small battery limits the usage of them. Even more, by technology scaling, more devices are integrated in a system and therefore a chip consumes more power. Low power design can extend a battery lifetime by optimizing power consumption of a device.

Analysis and design of robust low voltage SRAM is one of the essential parts for technology scaling and low power design to realize a smaller battery-operated SoC with a long lifetime. In modern SoC, a significant amount of area is used for SRAM and more aggressive scaling is applied to SRAM for denser integration. Hence, maintaining a high yield of SRAM and reducing power consumption of SRAM are necessary. Also, a battery-operated system spends a large portion of its lifetime in standby mode so it is very important to reduce the leakage power of SRAM which cannot be shutdown in standby mode. The most common way for power reduction is the low voltage operation. However, the low voltage operation compromises robustness and performance. Therefore, low voltage SRAM optimized for a given target application is essential to realize more advanced battery-operated systems.

## 1.1 Technology Scaling and Low Power Design

Over the past several decades, the number of transistors in a chip has increased exponentially (Moore's Law [47]). By decreasing the minimum feature size in integrated circuits, more devices are integrated in a small area, performance increases, and power consumption decreases.

However, even though the minimum feature size has decreased, the benefits of technology scaling have diminished today. One of the largest barriers of technology scaling is related to power consumption. The supply voltage has remained almost

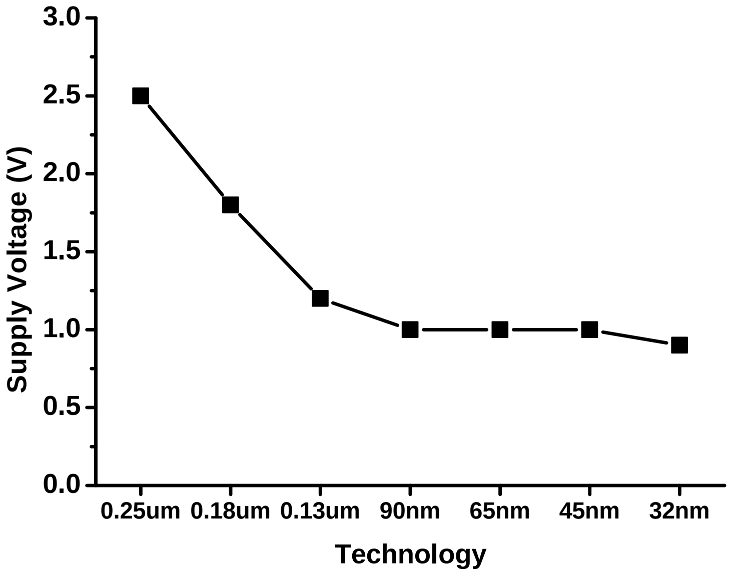


Figure 1.1: Technology scaling trends of supply voltage

constant [19] but leakage current has increased exponentially [48]. Figure 1.1 depicts supply voltage trends with technology scaling. Supply voltage has been stagnant at near 1V after 90nm technology but the total number of transistors in a chip has continuously grown. As a result, it is becoming more difficult to make an energy efficient system.

Variability is another barrier for scaling technology [3]. Random Dopant Fluctuation (RDF) [45] and Line Edge Roughness (LER) [23] induce a significant $V_{th}$ variation and performance variation.

As discussed above, optimizing power consumption is one of the most important issues in advanced technology nodes. Low power design techniques are important over the entire range Near-Threshold Computing (NTC): high performance platforms, personal computing platforms, and sensor-based platforms. If a system is designed for high performance, active power reduction is the most important. However, leakage power reduction techniques are more crucial in sensor-based platforms because an activity ratio in a sensor-based system is extremely low. In cases of personal computing platforms, balancing leakage power reduction techniques and active power reduction techniques is necessary. If voltage scaling is adopted for power reduction, it is also important to mitigate a larger variation at low voltage. Therefore, low power design includes active power reduction techniques, leakage power reduction techniques, and variation mitigating techniques at low voltage.

NTC is one of recently proposed architecture to minimize power consumption and mitigate performance degradation using highly parallelized voltage scaled processors [19]. Finding an optimal trade-off point between power reduction and performance degradation by voltage scaling is the underlying idea of NTC.

## 1.2 Challenges of Designing Low Voltage SRAM

From high performance platforms to sensor-based platforms, the importance of designing robust low power SRAM must be emphasized for several reasons: dominance in area, critical yield issue, and large array sizes.

A significant amount of area is used for SRAM in modern SoC. Figure 1.2 shows the most recent high performance CPU from Intel. Excluding the area of a graphic core and a memory controller integrated in this processor, the L3 cache shown in the die photo and lower level caches integrated in cores spend about 50% of total area. In a case of sensor-based application, the dominance in area of SRAM does not alter. A sensor application developed by University of Michigan also spends about 50% of its area for SRAM (Figure 1.3). Because of the large area dedicated to SRAM, a portion of power consumption by SRAM is significant in modern SoC too.

Maintaining high SRAM yield becomes more challenging because they are particularly vulnerable to process variation. This is because minimum or smaller than minimum sized devices are used in SRAM and the size of array are very large (up to 10s of MB). Voltage scaling for power reduction aggravates robust SRAM operations. Process variation increases as voltage scales and it makes already limited SRAM

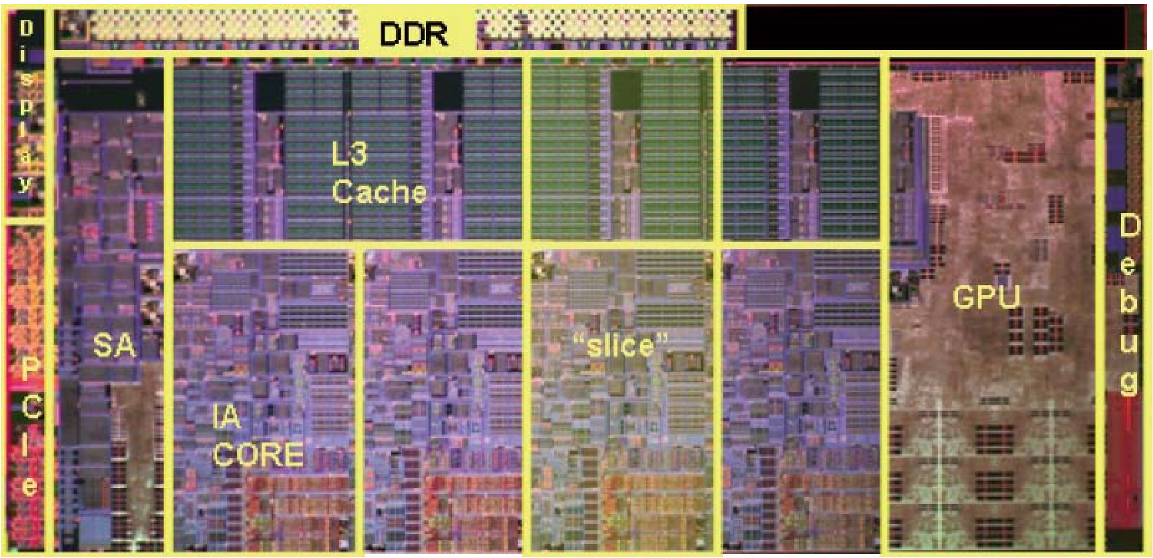


Figure 1.2: Die Photo of Intel High Performance 32nm Processor [71]

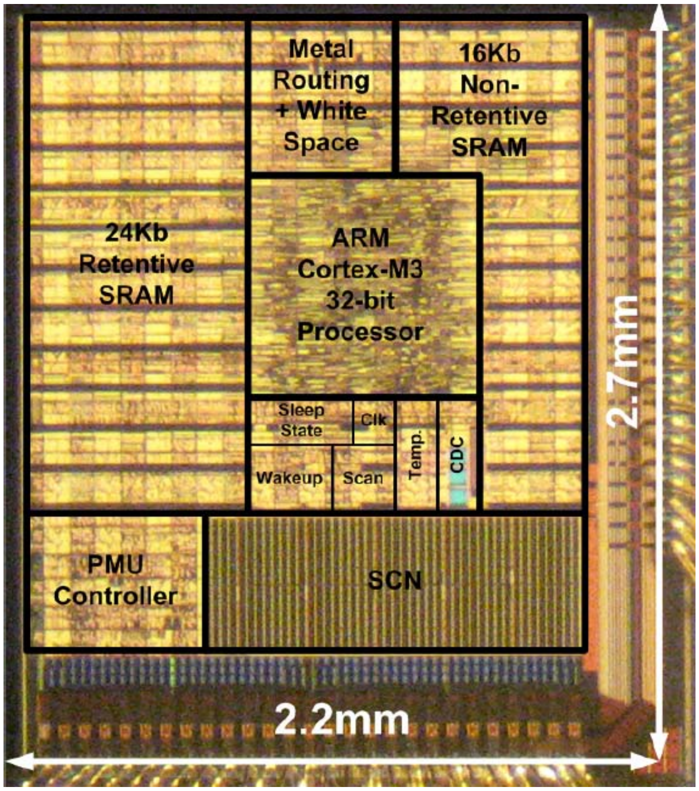


Figure 1.3: Die Photo of University of Michigan 0.18$\mu$m Sensor System [11]

| Solution | Power | Performance | Area | Yield(Robustness) |
|---|---|---|---|---|
| Device Sizing | Worse | Better | Worse | Better |
| Voltage Scaling | Better | Worse | Same or Worse | Worse |
| 8-T SRAM | Vary by Design | Better | Worse | Better |
| Assist Circuits | Vary by Design | Better | Worse | Better |

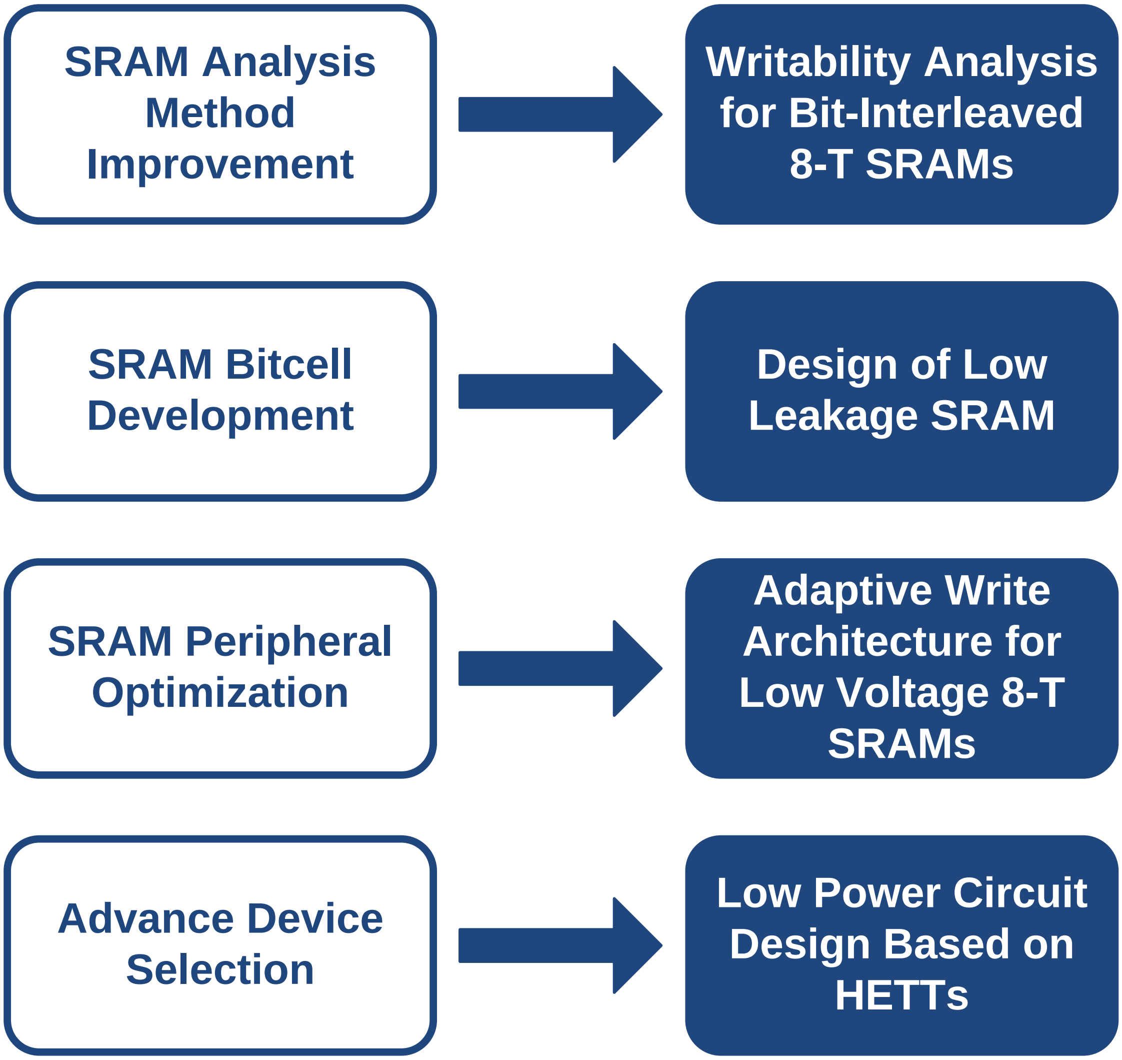
Table 1.1: There are different trade-offs in each solution.

yield worse. Also, low supply voltage impacts the reliability of SRAM [5]. Gate oxide degradation and soft error susceptibility are two important reliability challenges in voltage scaled SRAM design.

Many techniques have been proposed to solve prevailing issues in low voltage SRAM design. In Table 1.1, the traditional solutions for robust low voltage SRAM design are shown and the trade-offs in each solution are compared. The size of devices in SRAM can increase for increased yield and higher performance and but it will result in increased total area and higher power consumption. Voltage scaling can be adopted for power saving but performance will be degraded and yield will be compromised. Using different SRAM bitcells is another approach. With 8-T SRAM bitcell, performance and yield will be improved but the larger area of 8-T SRAM bitcell cannot be avoided. To support a particular SRAM operation, write or read assist circuit can be used. However, there is a chance to spend extra area for the assist circuits. As shown above, there are trade-offs in each solution. Therefore, it is important to choose appropriate techniques and optimize them for a target application.

## 1.3 Contributions of This Work

A main purpose of this work is to analyze and design of low voltage SRAM with high yield and low power consumption. To achieve this goal, there are different approaches (Figure 1.4).


SRAM Analysis Method Improvement

Writability Analysis for Bit-Interleaved 8-T SRAMs

SRAM Bitcell Development

Design of Low Leakage SRAM

SRAM Peripheral Optimization

Adaptive Write Architecture for Low Voltage 8-T SRAMs

Advance Device Selection

Low Power Circuit Design Based on HETTs

Figure 1.4: Approaches for analysis and design of robust low voltage SRAM

The first approach is SRAM analysis method improvement. During design phase or even testing phase, appropriate analyzing methods to measure SRAM yield are necessary for correct and fast yield estimation. Static Noise Margin (SNM) method has acted an important role for robustness estimation of SRAM bitcell for more than 20 years [61]. However, as new SRAM bitcells are developed and complicated yield related issues appear, a dedicated analysis method for a given case is necessary for more accurate and faster estimation of SRAM robustness. In Chapter II [33], a writability analysis method for bit-interleaved voltage-scaled 8-T SRAMs is proposed to maximize writability while minimizing half select disturb. For robust SRAM operations at low voltage, 8-T SRAM bitcell is used. However, it suffers from half select disturb if bitcells are interleaved. This analysis proposes a method to find an appropriate WWL pulse width to maximize yield.

SRAM bitcell development other than 6-T SRAM bitcell is another approach to develop robust low voltage SRAM. A 6-T SRAM bitcell has been a main component of SRAM arrays for the past several decades. However, as technology and voltage scales, it is hard to maintain enough margins for read and write operations with the 6-T SRAM bitcell. To overcome this limitation, an 8-T SRAM bitcell has been introduced [8, 9]. In addition to the 8-T bitcell, other bitcell designs have been proposed for different purposes. A low leakage SRAM with speed compensation scheme is proposed in Chapter III [32]. In this work, a novel 14-T SRAM bitcell is proposed for leakage reduction at the expenses of speed and area. Several leakage reduction techniques are applied to a SRAM array and speed compensation scheme is adopted.

SRAM peripheral optimization can be used at low voltage. In general, process variation is large at low voltage and therefore a large margin is required for successful operations. To minimize the margin and to increase performance and yield at low voltage, adaptive write architecture for low voltage 8-T SRAMs is proposed in

Chapter IV. At low voltage, write operation is a critical operation in 8-T SRAM bitcell. For higher writability, WWL pulse width must be extended and WWL voltage level must increase. However, it will dissipate more power and increase the chance of half select disturb. In this work, performance and yield are improved by adaptively modulating WWL pulse width and WWL voltage level.

The final approach is advance device selection. Complementary Metal-Oxide-Semiconductor (CMOS) has been predominantly used for the past several decades but there have been studies on other future devices. The limited subthreshold swing in Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET) significantly restricts low voltage operation due to a low $I_{on}$ to $I_{off}$ ratio. As an alternative, a tunneling transistor with a steep subthreshold swing has been proposed. In Chapter V low power circuit design based on heterojunction tunneling transistors (HETTs) [31] is discussed. First, device modeling method using Verilog-A is proposed. In addition, benefits and limitations in Heterojunction Tunneling Transistor (HETT) are discussed. At the end of the section, HETT-based SRAM is proposed and leakage reduction using it is discussed.

# CHAPTER II

# Writability Analysis for Bit-Interleaved 8-T SRAMs

As process technology scales, SRAM robustness is compromised. In addition, lowering the supply voltage to reduce power consumption further reduces the read and write margins. To maintain robustness, a new bitcell topology, 8-T bitcell, has been proposed and read and write operation can be separately optimized. However, it can aggravate the half select disturb when write word-line boosting is applied or the bitcell sizing is done to enable robust writability. The half select disturb issue limits the use of a bit-interleaved array configuration required for immunity to soft errors. The opposing characteristic between write operation and half select disturb generates a new constraint which should be carefully considered for robust operation of voltage-scaled bit-interleaved 8-T SRAMs. In this chapter, we propose bit-interleaved writability analysis that captures the double-sided constraints placed on the word-line pulse width and voltage level to ensure writability while avoiding half select disturb issue. Using the proposed analysis, we investigate the effectiveness of word-line boosting and device sizing optimization on improving bitcell robustness in low voltage region. With 57.7% of area overhead and 0.1V of word-line boosting, we can achieve 4.6$\sigma$ of $V_{th}$ mismatch tolerance at 0.6V and the design shows 41% of energy saving.

## 2.1 Introduction

SRAM is an indispensable part of most modern Very-Large-Scale Integration (VLSI) designs and dominates silicon area in many applications. In scaled technologies, maintaining high SRAM yield becomes more challenging since they are particularly vulnerable to process variations due to 1) minimum(close to minimum) sized devices used in SRAM bitcells and 2) the large array sizes (10s of MB). At the same time, low power design is a key focus throughout the semiconductor industry. Since low voltage operation is one of the most effective ways to reduce power consumption due to its quadratic relationship to energy savings, lowering the minimum operating voltage ( $V_{min}$) of SRAM has gained significant interest.

To mitigate variability and reduce the $V_{min}$, it is important to understand SRAM failure modes and quantify immunity to failures. SNM [61] has been widely used as a metric to estimate the immunity to read/write/hold failures. However, it overestimates read failures and underestimates write failures since it assumes infinitely long word-line pulses. For more accurate analysis, dynamic writability has been introduced [18, 66, 69]. In addition, soft error susceptibility of SRAM to particle strikes is a key issue in modern SRAM design [2, 5]. To fix soft errors, bit-interleaved arrays are commonly used; however this leads to the possibility of half select disturb, which degrades robustness. As the supply voltage is lowered, mitigating soft error becomes more important because soft error vulnerability is more critical at low voltage [5]. Hence bit-interleaving array must be adopted and half select disturb issue must be carefully analyzed.

To ensure robust operation at low voltage in nanoscale technologies, an 8-T bitcell has been proposed [8, 9, 57] which can be separately optimized for read and write since bitcells are not interleaved. However, when 8-T bitcells are interleaved for immunity to soft errors, half select disturb issue [29] arises and it limits the freedom to maximize its writability. In terms of dynamic writability, the longer pulse width is favorable to write operation while the shorter pulse width is favorable to immunity to half select disturb. These double-sided constraints placed on word-line pulse width make it difficult to determine appropriate word-line pulse width to maximize SRAM robustness. Also, when write assist method is adopted, half-select disturb is more likely to happen and it will decrease overall yield.

To address these issues, this chapter proposes bit-interleaved writability analysis (both static and dynamic) for a voltage-scaled 8-T bitcell using SRAM worst-case corner simulations. It captures the double-sided constraints to ensure successful write operation and immunity to half select disturb. In addition, we can obtain appropriate word-line pulse width using bit-interleaved dynamic writability analysis.

Compared with the prior works commonly used for SRAM robustness analysis, this work highlights the double-sided constraints on 8-T bitcell write operation which can be mistakenly not considered on the assumption that 8-T bitcell has an unlimited freedom to maximize writability. The common method to avoid the double-sided constraints between read and write at 6-T bitcell is to apply different word-line pulse widths and voltage levels for each operations and this is feasible because read and write at a 6-T bitcell cannot be done simultaneously. However, the new constraints between write and half select disturb in a bit-interleaved 8-T SRAM cannot be solved using those methods because write targeted cell and half selected cell always experience the same word-line pulse width and voltage level. Therefore, special regard is paid to this fact in this chapter. Also, this work uses the dynamic writability analysis and therefore it does not overestimate its failures by assuming an infinitely long word-line pulse.

With this analysis method, we evaluate the effectiveness of two techniques to lower the $V_{min}$: word-line boosting and device size optimization. Poor writability and high soft error susceptibility limit low voltage operation. An SRAM cell in a commercial 45nm low-power CMOS can tolerate only up to $1.7\sigma$ at 0.6V in terms of worst case

$V_{th}$ mismatch which is not acceptable for yield. To achieve iso-robustness($4.5\sigma$) as 1.0V, device sizes need to increase and word-line boosting is needed. We can achieve $4.6\sigma$ tolerance at 0.6V with 57.7% area overhead and 0.1V of word-line boosting. Compared with normal write operation at 1.0V without any technique, it shows 41% of energy saving per operation.

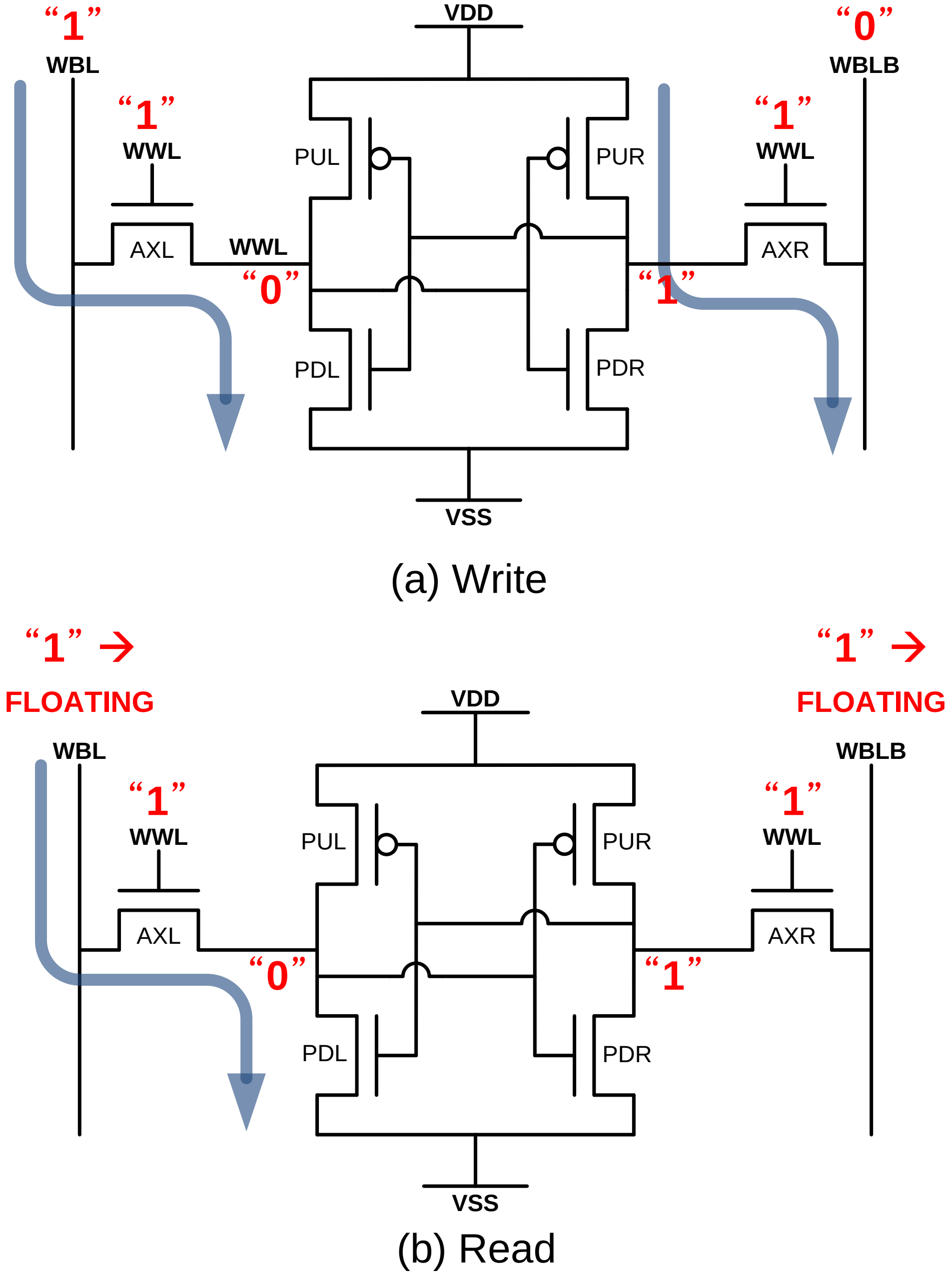
## 2.2 Background and Related Works

Write failure and read disturb are two major SRAM failures. To quantify the probability of these failures, the SNM method [61] has been used for more than twenty years. In addition to write failure and read disturb, soft error [2, 5] and half select disturb [29] have emerged as sources of SRAM failures. This section reviews these SRAM failure modes and related work.

### 2.2.1 Write Failure and Read Disturb

Figure 2.1(a) describes the write operation of a 6-T bitcell. As access transistors (AXL and AXR) are turned on, values on Write Bit-Line (WBL) and Write Bit-Line Bar (WBLB) are driven to internal nodes of a bitcell which attempt to flip both nodes. A write failure occurs when the internal nodes do not flip due to the access transistors being too weak or the pull up P-channel MOS (PMOS) transistors being too strong.

The read operation is depicted in Figure 2.1(b). Both bit-lines are pre-charged to logic “1” before reading. After the word-line pulse is asserted, one of the bit-line (WBL) falls rapidly due to active pull down while another bit-line (WBLB) falls very slowly due to the leakage of other bitcells connected to WBLB. The read operation is completed when a sense amplifier detects a sufficient voltage difference between the two bit-lines. A read disturb occurs when the internal nodes accidentally flip during read operation, caused by a voltage excursion from 0 due to an overly strong access



(a) Write

(b) Read

Figure 2.1: SRAM (a) write operation and (b) read operation

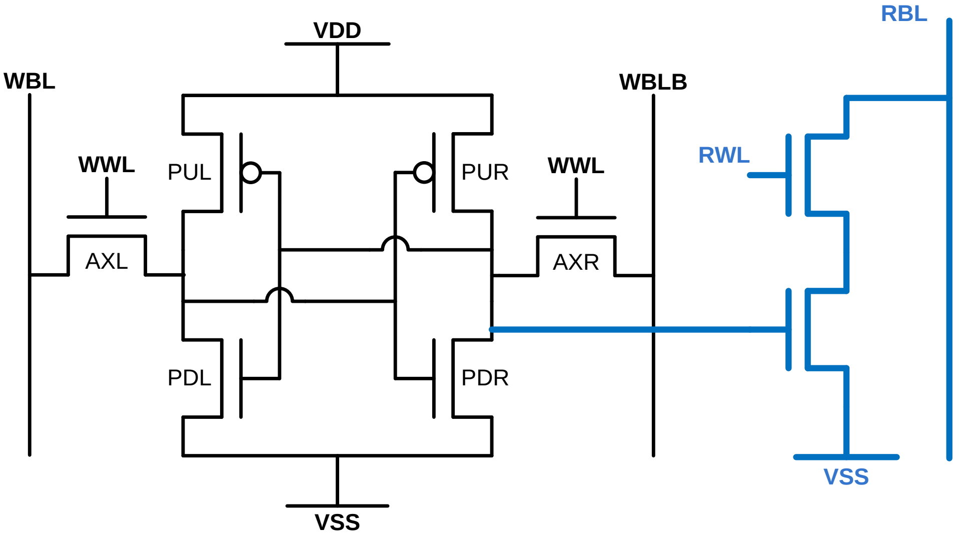


Figure 2.2: 8-T SRAM to decouple write and read

transistor and weak pull-down transistor.

It is difficult to make both the read and write operations highly stable because strong access transistors are preferred for write operations while the opposite is true during read operations. To overcome this limitation, there has been many works using an 8-T bitcell [8, 9, 57] that decouples the read and write paths (Figure 2.2). The write operation is identical to the 6-T but the read operation is executed via a 2-T read path. With 8-T, devices on write and read paths can be optimized separately for each operation.

### 2.2.2 Static Noise Margin Analysis

To estimate SRAM bitcell immunity to the failures mentioned in subsection 2.2.1, the SNM method [61] has been typically used. In addition to read and write failures, it is necessary to understand hold failure before using the SNM method. A hold failure happens when the amount of static noise is large enough to flip the internal state of a SRAM bitcell. The examples of the static noise are offsets and mismatches due to process variation in various operating conditions. To quantitatively analyze

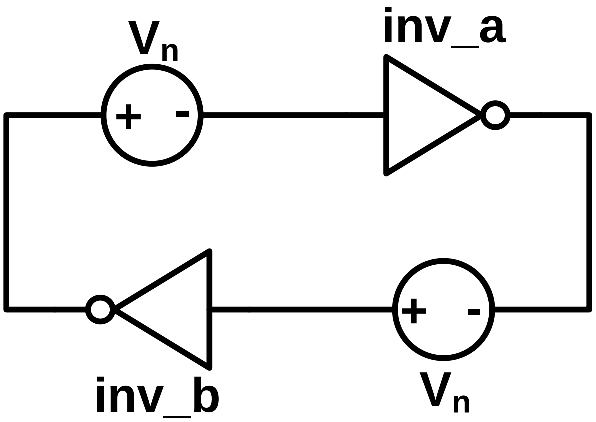


Figure 2.3: Two static noise sources are inserted between two cross-coupled inverters.

hold failure, two imaginary static noise voltage sources are inserted at the two internal nodes (Figure 2.3) and the well-known butterfly curves are drawn assuming zero noise. In the butterfly curves (Figure 2.4), there are two stable points and one metastable point. When two voltage nodes are in the one of stable points, they will remain in the stable point as long as two curves keep the butterfly shape. In other words, the bitcell is stable when a square inside of the butterfly curves can be drawn. If the maximum possible square between the curves is large, it implies that the SRAM bitcell has a large margin against hold failure. Therefore, the SNM is defined as the length of the diagonal line of the largest rectangle which fits in the butterfly curves and it is $\sqrt{2}\times$ larger than the maximum static noise voltage tolerance. Figure 2.4 shows the butterfly curves and the hold SNM at 1.1V and 0.3V respectively. It is clear that the hold SNM decreases as the supply voltage scales and therefore it is more likely to fail at low voltage region.

A schematic for measuring read SNM is shown in Figure 2.5(a). The schematic shows the state when two bit-lines are precharged to $V_{DD}$ and the word-line is turned on. A schematic for measuring write SNM (Figure 2.5(b)) is a little bit different: one bit-line is “1”; another bit-line is “0”; and the word-line is also turned on.

The read SNM (Figure 2.6(a)) is smaller than the hold SNM because of two

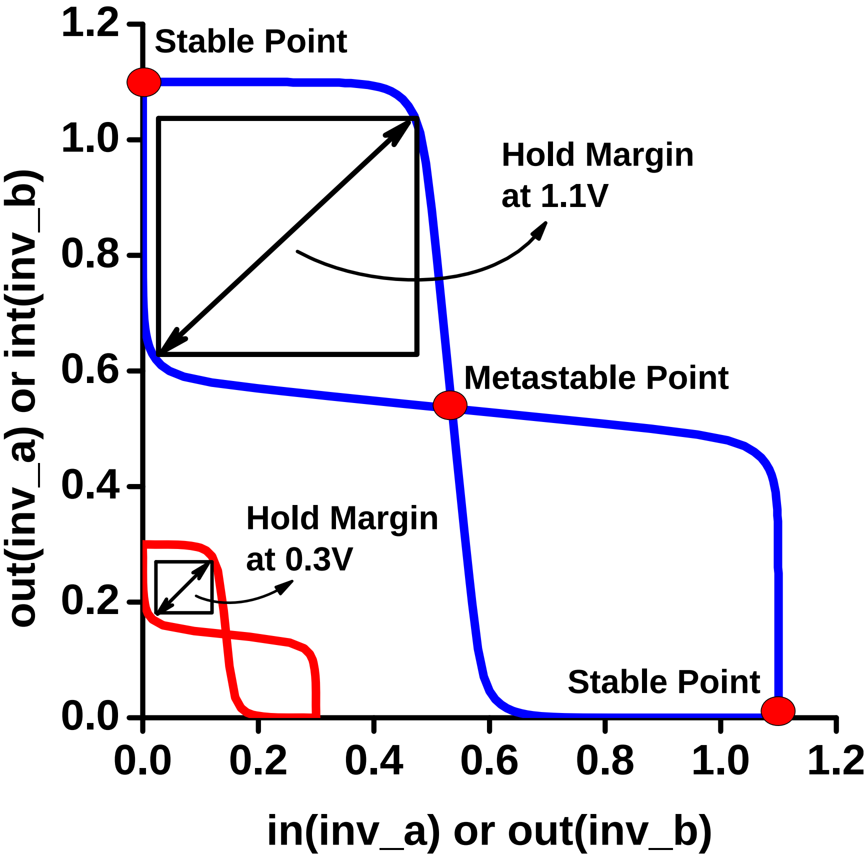
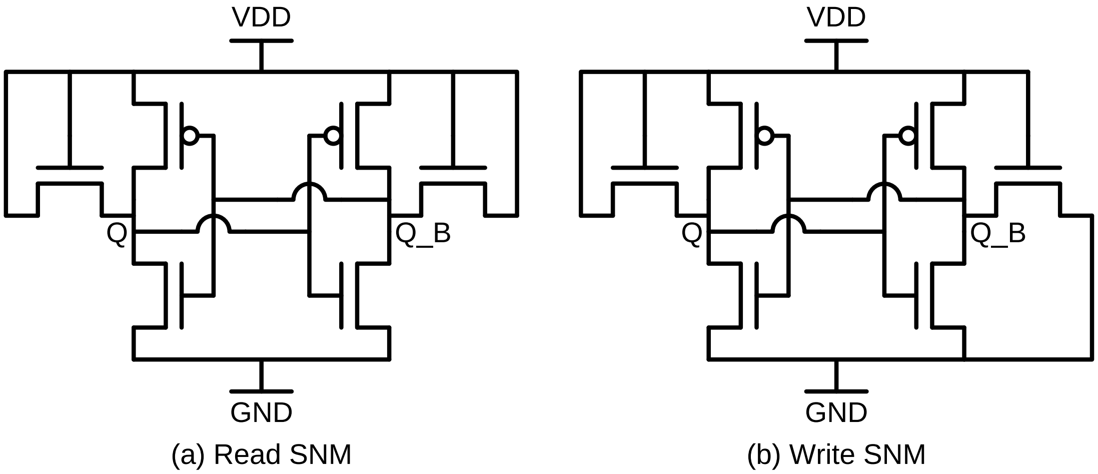


Figure 2.4: The butterfly curves for hold margin are shown.



(a) Read SNM

(b) Write SNM

Figure 2.5: Schematics for measuring (a) read SNM and (b) write SNM

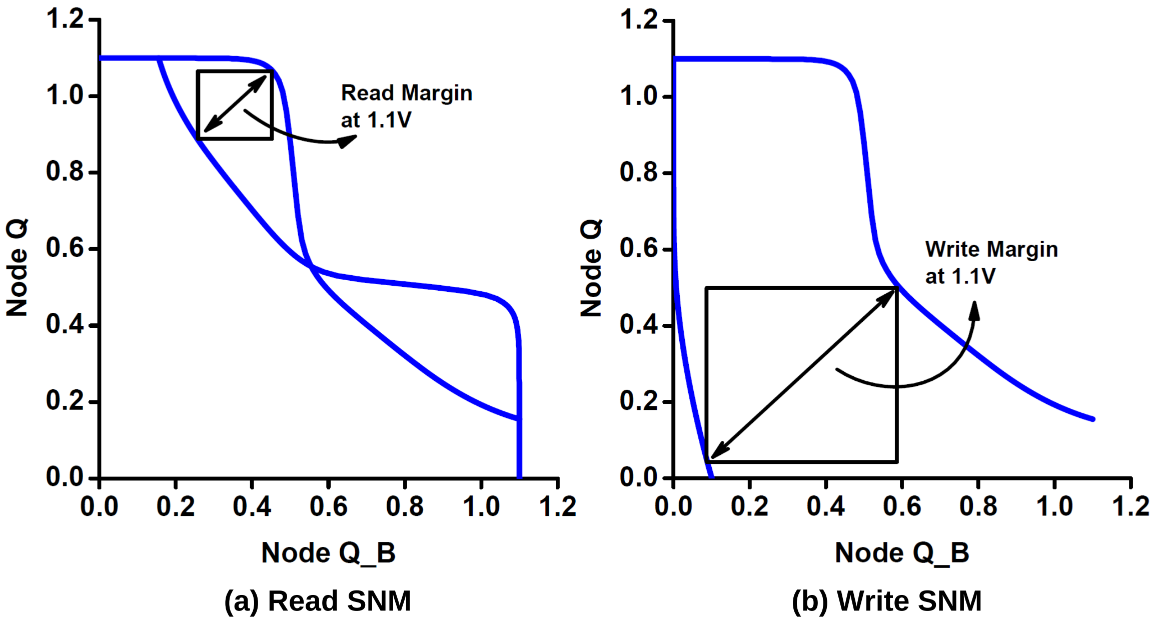


Figure 2.6: The butterfly curves for (a) read and (b) write margins are shown.

direct current paths from $V_{DD}$ to the internal nodes via the access transistors. It is noticeable that the read disturb happens if the read margin is negative. In the case of the write SNM, the definition of SNM is different from the other cases. The hold and read SNM is positive when the internal state does not change while the write SNM is positive when it changes. Therefore, the maximum square is drawn at the outside of the curves (Figure 2.6(b)).

While other works [21, 70] have been recently proposed to supplement the SNM method, SNM remains the standard approach today.

The limitation of the SNM analysis is that all signals are static in the analysis. The infinitely long ("static") word-line pulse width are used in the analysis while there is no such a pulse in reality. In addition, bit-lines are being kept at $V_{DD}$ in the analysis while the voltage levels of them decreases as time goes by. The static word-line pulse causes optimistic and pessimistic estimation for write and read operations, respectively. To overcome this limitation, works related to dynamic stability have been introduced [66, 69].

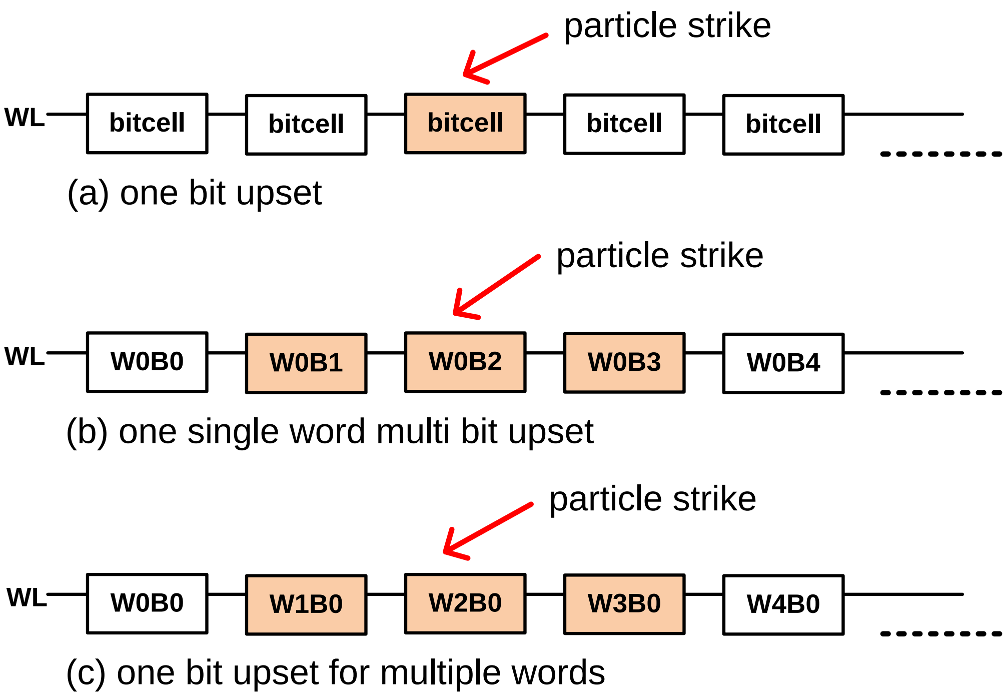


Figure 2.7: Three different scenarios of soft errors

### 2.2.3 Soft Error and Half Select Disturb

Soft errors are faults induced by a particle strike that upsets internal data states while the circuit itself is undamaged [2]. Even though it is unpredictable, soft error susceptibility is a critical reliability challenge for modern SRAM design [5]. Figure 2.7 shows three different scenarios when a soft error occurs. In Figure 2.7(a), a particle hit results in only one upset bit and it can be fixed with Hamming Single Error Correction/Double Error Detection (SECDED) codes [22]. Figure 2.7(b) and Figure 2.7(c) depict single event multi-bit upsets, which become more common in highly scaled technologies with smaller bitcells [51]. In Figure 2.7(b), all bits in a single word are located next to each other and, therefore, a single word has multiple bits corrupted. This type of error cannot be fixed with SECDED codes and requires complicated approaches that incur large area penalties [51]. The easiest way to avoid single word multi-bit upsets is to interleave bits, such that logically adjacent bits are not physically adjacent. Figure 2.7(c) shows such a bit-interleaved array structure

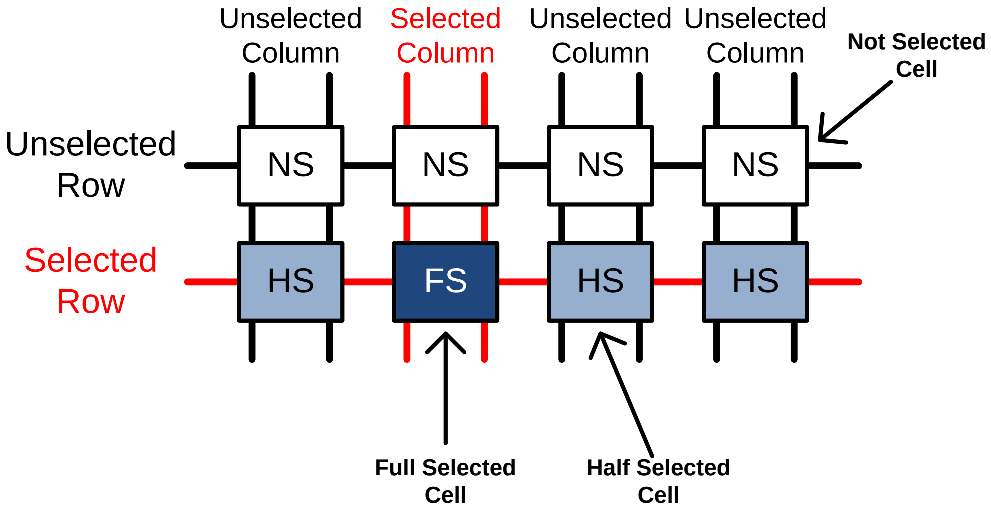


Figure 2.8: Half select disturb in bit-interleaved array

and in this case the multi-bit upset can be easily fixed with SECDED codes since each word contains only a single upset bit.

Although bit-interleaving is effective in avoiding single word multi-bit upset, it induces half select disturb problem. Figure 4.7 illustrates the half select disturb phenomenon in a bit-interleaved array. In a selected row containing 4 words, $\frac{3}{4}$ of columns are unselected, defined as "half selected", and as a result the access transistors in these cells are turned on and the internal data could flip(those half selected columns are in "6-T read-like" mode).

Several works [25, 29] have focused on resolving this issue. Reference [29] proposed a local word-line scheme which does not allow bit-interleaving and reference [25] proposed the electron injection which requires process tweaking.

### 2.2.4 Writability vs. Half Select Disturb in Bit-Interleaved 8-T SRAMs

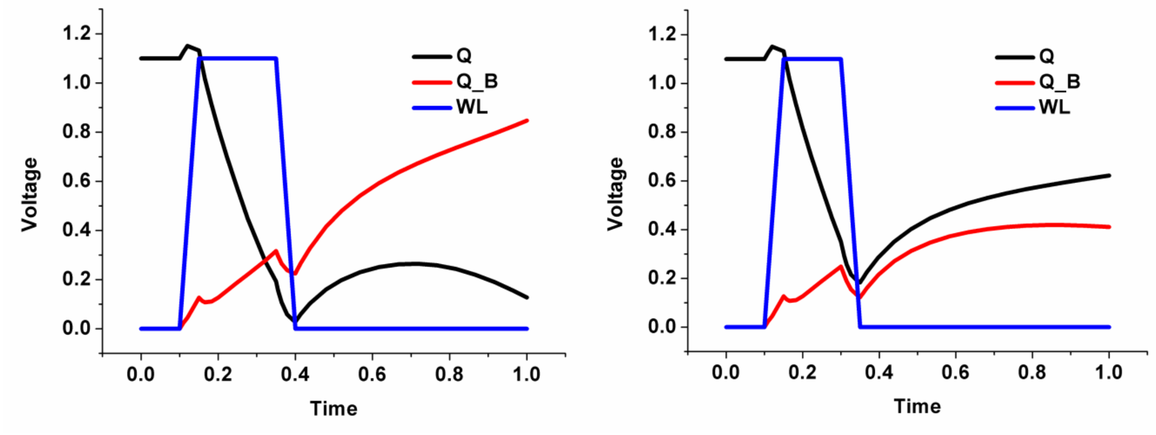
In a 6-T SRAM, half select disturb is nearly identical to read disturb if bit-lines of unselected columns are floated. Therefore, half select disturb is unlikely to happen with appropriately sized 6-T SRAM if write assist method is not applied. However, if the 6-T portion of an 8-T SRAM is optimized for the write operation, half select disturb is more likely to happen and must be carefully considered when 8-T SRAM is designed. In addition, to regain robust writability at low voltage in nanoscale technologies, write assist methods such as word-line boosting are commonly used. Differently from read disturb in 6-T SRAM, half select disturb happens concurrently along with write operation so the write assist methods directly influence half select disturb and therefore the effect on yield must be analyzed before using these methods.

## 2.3 Writability Analysis Method

SNM has long been used to estimate read-stability and writability. However, it assumes an infinitely long word-line pulse, making it optimistic for write and pessimistic for read operations compared with realistic SRAM operation. Recently, several papers [69, 18, 66] have considered dynamic writability to accurately assess SRAM writability. This section analyses SRAM bit-interleaved writability using worst case corner simulation. Because read operation is done through the 2-T read path of the 8-T bitcell, we consider the 6-T part of 8-T bitcell in the rest of the chapter for write operation and the writability does not influence read operation.

### 2.3.1 SRAM Dynamic Writability Metric

SRAM dynamic writability can be defined using the minimum write word-line pulse width required for a successful write operation, $T_{crit}$ [69]. If $T_{wl}$ is longer than $T_{crit}$, the write operation will be successful (Figure 2.9(a)). However, a bitcell cannot be written for $T_{wl}$ shorter than $T_{crit}$ (Figure 2.9(b)) and this is referred to as dynamically limited write failure. If the bitcell cannot be written at all, even with an infinitely long word-line pulse, we refer to this case as statically limited write failure. $T_{crit}$ is infinite in this case, allowing static write failure to be captured with the same metric. In Reference [6], the effectiveness of write assist techniques are compared based on $T_{crit}$.



(a) Write Success

(b) Write Failure

Figure 2.9: There is a minimum word-line pulse width of the successful write.

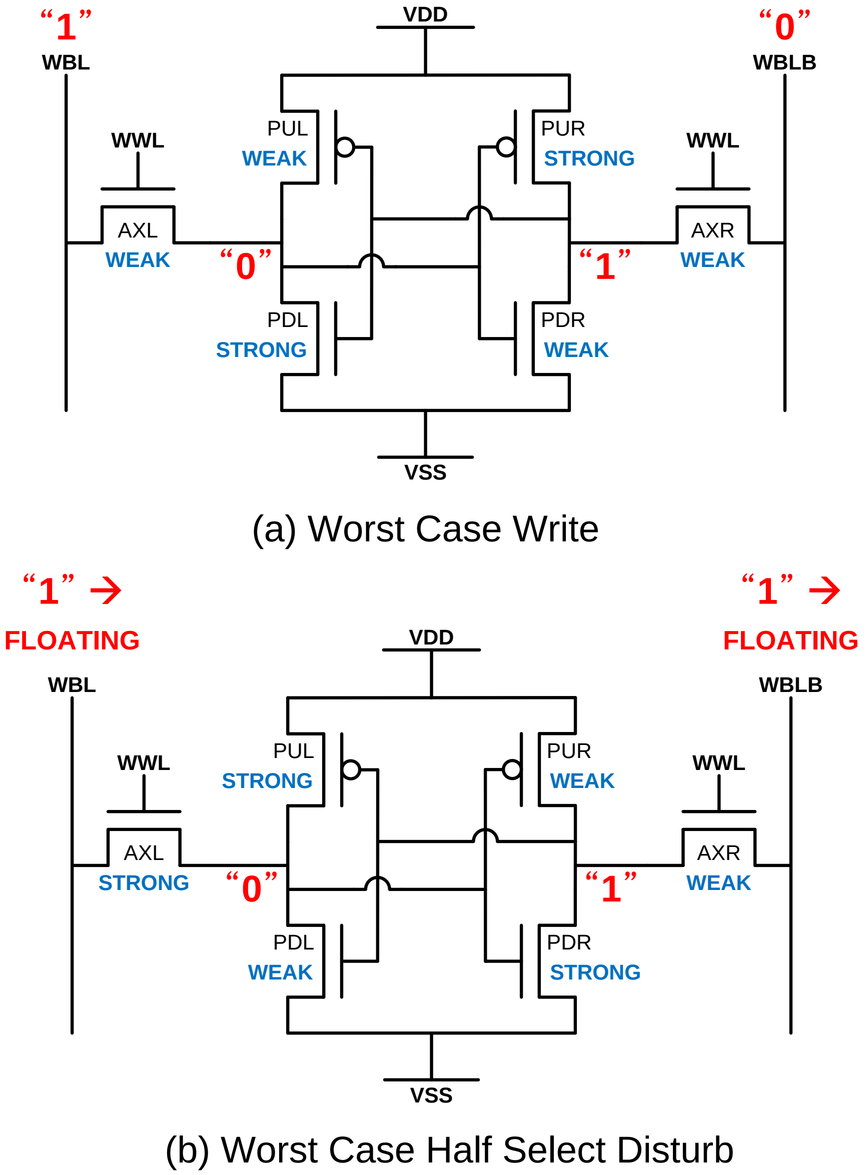
### 2.3.2 SRAM Worst Case Corner Simulation

SRAM worst case corner simulation is used to characterize the SRAM writability. The basic idea of this simulation is to find the maximum $V_{th}$ mismatch allowable before failure occurs, which is then used as the quantitative definition of writability [12].

A device becomes stronger or weaker when its $V_{th}$ decreases or increases, respectively. Initially, there is no $V_{th}$ skew for each device. To worsen the writability of s bitcell, $V_{th}$ of each device is skewed in appropriate directions. Figure 2.10(a) shows the worst case corner directions for each device in a write operation. Weak access transistors, AXL and AXR, worsen writability since it becomes difficult to drive the bit-line values onto the internal nodes through them. The strong left pull-down transistor (PDL) and the weak left pull-down transistor (PUL) tightly hold logic “0” and, therefore, writability is weakened. Similarly, the weak right pull-down transistor (PDR) and right pull-up transistor (PUR) worsen writability. The 6-T structure is symmetric and, therefore, we only consider a single state case (write “1” only) and it will reflect the other state too. Figure 2.10(b) shows the worst case corner directions for half select disturb. While an SRAM cell is designed not to have half select disturb in the absence of variations, transistor mismatch will incur such errors. The worst case corner directions for the half select disturb differ from that of the write operation. To cause a disturbance, AXL should be strong enough to easily drive the logic “1” on WBL to the internal node while AXR is weak such that the logic “1” on WBLB does not help maintain the high state on the right internal node. The skewed directions of the internal four devices are set such that they do not strongly hold the internal nodes, allowing them to more easily flip. Initially, both bit-lines are pre-charged to logic “1”. However, they float when WWL is asserted and the voltage levels of bit-lines are determined by how many bitcells are connected to each bit-line and the values stored in each bitcell. In our simulations, we assume 256 bitcells are connected to each bit-line with all other bitcells storing an opposite value to the target bitcell to ensure the worst case.

Figure 2.11 shows the worst case corner simulation result at 1.1V in a 45nm low-power CMOS process for (a) write operation and (b) half select disturb. All values are normalized to $T_{crit}$ for the write operation at $0\sigma$, i.e., no variation. The write operation (Figure 2.11(a)) can be successfully performed up to $6.3\sigma$, however $T_{crit}$ increases monotonically as $V_{th}$ mismatch increases indicating a steady degradation in write performance with variability. In the real simulation, because we cannot use the infinitely long word-line pulse, we assume that a pulse width of normalized 1000 as a practical limitation before static failure. Putting this in terms of Static Writability (SW) with infinitely long word-line pulse, SW at 1.1V is $6.3\sigma$. On the other hand, Dynamic Writability (DW) depends on $T_{wl}$ (word-line pulse width). For example, if $T_{wl}$ is allowed to be 3× the nominal value, DW at 1.1V is not $6.3\sigma$ but $5.2\sigma$. While SW reveals a theoretical limitation, DW represents a more realistic view of actual writability.

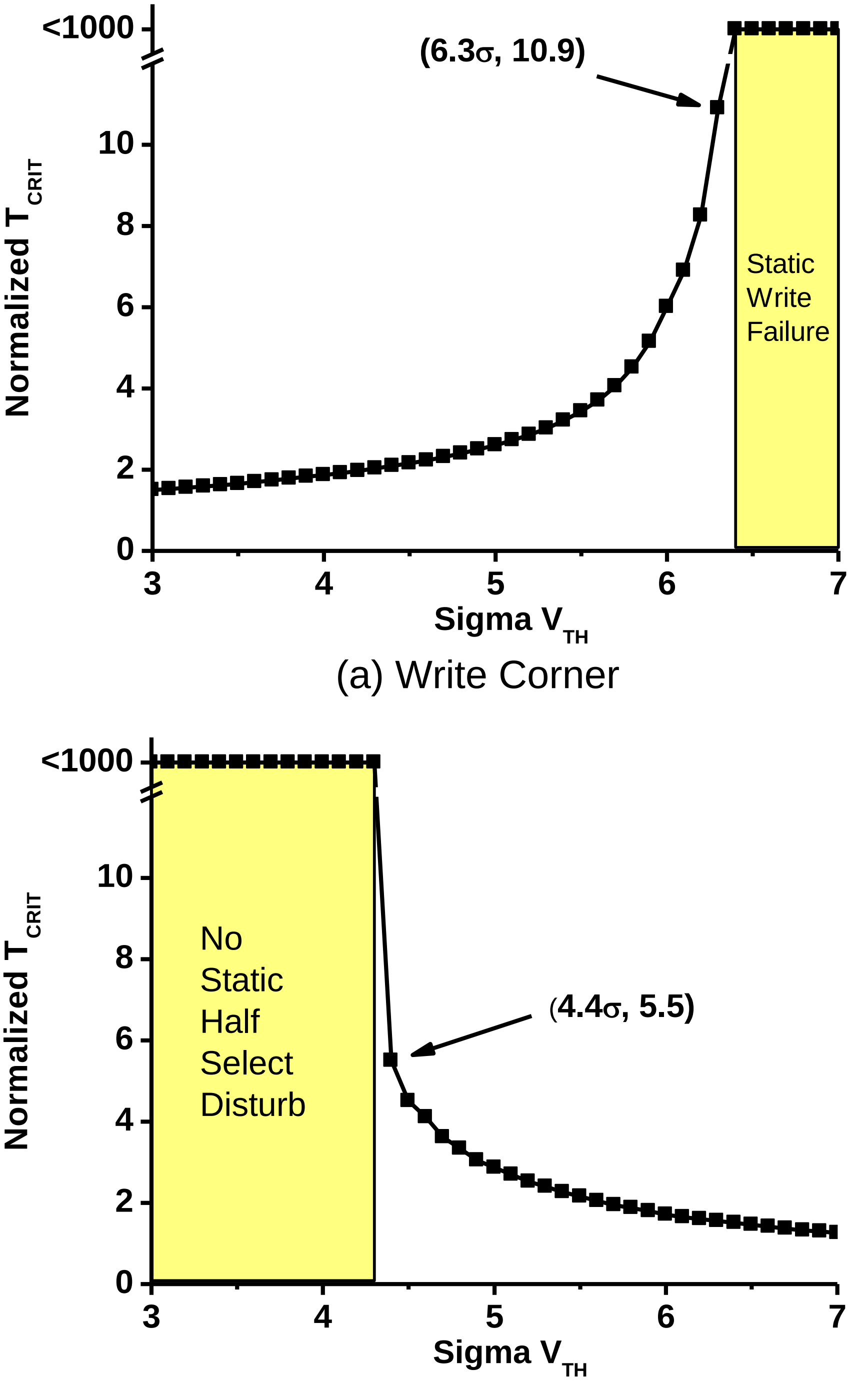
Figure 2.11(b) depicts the half select corner result. Half select disturb does not occur up to $4.3\sigma$ even with an infinitely long word-line pulse hence $4.3\sigma$ is the static



(a) Worst Case Write

(b) Worst Case Half Select Disturb

Figure 2.10: $V_{th}$ mismatch directions of each device (a) for write operation and (b) for half select disturb



(a) Write Corner

(b) Half Select Disturb Corner

Figure 2.11: Corner simulation results at 1.1V

limitation of the half select disturb at 1.1V. As variation increases, the half select disturb likelihood increases, such that at $4.4\sigma$, a $5.5\times$ long word-line pulse is required to cause half select disturb. With more variation, the necessary word-line pulse width decreases, indicating that the cell becomes more vulnerable to half select.

### 2.3.3 Bit-Interleaved Writability Analysis

The previous subsection investigated static and dynamic writability and half select disturb. The simulation results show that a longer word-line pulse is simultaneously favorable for write and unfavorable for half select disturb. This is problematic when an SRAM array uses bit-interleaving for soft-error immunity. Without bit-interleaving, the dynamic writability can be improved at the expense of operation speed (e.g., by allowing for longer word-line pulses when variability is large). However, with bit-interleaving, such longer word-line pulses will generate half select disturbs, limiting overall array robustness. To analyze this trade-off, the worst corner simulation results for the write operation and half select disturb are overlaid in Figure 2.12. With an infinitely long word-line pulse, the write operation tolerates up to $6.3\sigma$ variability while half select disturb starts to occur beyond $4.3\sigma$. The Bit-interleaved Static Writability (BSW) can be defined as the maximum $V_{th}$ mismatch until the write failure OR the half select disturb occurs, assuming an infinitely long word-line pulse. Therefore, BSW at 1.1V is $4.3\sigma$. Up to $4.3\sigma$, the write operation is successful if $T_{wl}$ is larger than $T_{crit,write}$ ($T_{crit}$ of the write operation). However, BSW is pessimistic because the infinitely long word-line pulse is unrealistic. To overcome this, the Bit-interleaved Dynamic Writability (BDW) can be defined as the maximum $V_{th}$ mismatch until write failure and half select disturb occurs at a given word-line pulse width. At $4.4\sigma$, if $T_{wl} > T_{crit,write}$ yet smaller than $T_{crit,half}$ ($T_{crit}$ of the half select disturb), the write operation can be successfully performed without incurring half select disturb. This leads to a BDW of $5\sigma$ at 1.1V. In this way, BDW best captures the trade-off

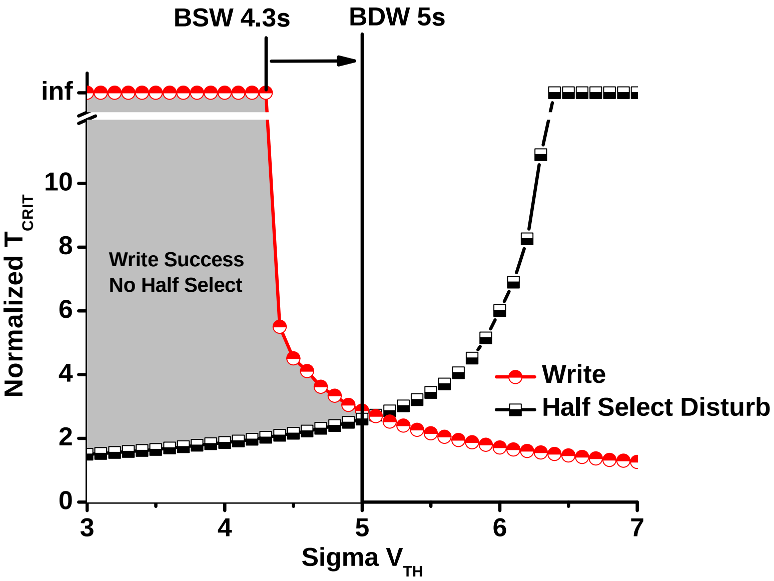


Figure 2.12: Bit-interleaved writability analysis at 1.1V

between writability and half select while capturing the negative correlation between these parameters.

## 2.4 Writability Analysis at Near-Threshold

Figure 2.13 depicts the bit-interleaved writability with supply voltage scaling. Figure 2.13 clearly shows that the writability is very limited at near-threshold region. When BDW and BSW are overlapped, it has very poor writability and therefore the writability is statically limited before half-select happens. In this section, we investigate how to increase the writability using two writability enhancement techniques: word-line boosting and device sizing optimization.

### 2.4.1 Word-line Boosting

The first approach to enhance the writability is word-line boosting. This is a commonly used technique for SRAM operation in low voltage regime [57, 6]. In Figure 2.14, the two access transistors are over driven by the boosted word-line. By

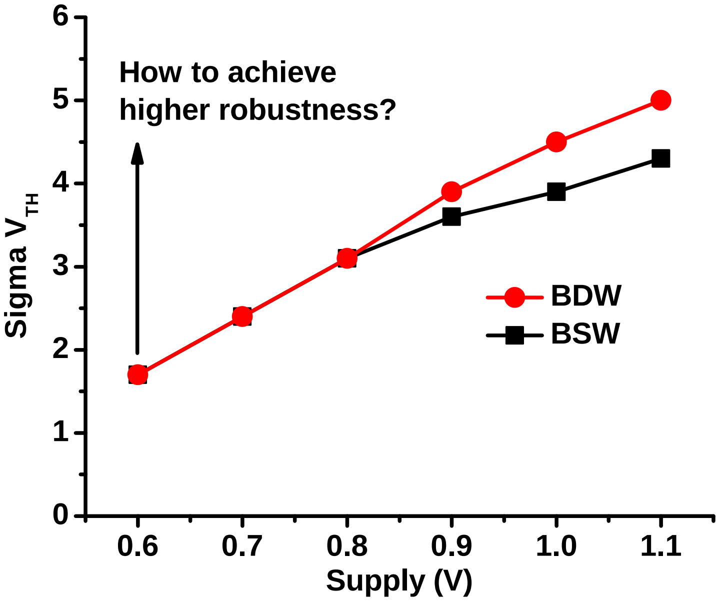


Figure 2.13: Bit-interleaved writability as voltage scales

doing this, the current driving abilities of both transistors are enhanced; therefore, the SRAM cell becomes more favorable to the write operation. At the same time, the SRAM cell is more likely to experience read disturb and half select disturb with the word-line boosting. Figure 2.15 depicts the writability as the word-line boosting voltage increases when the supply voltage is 0.6V. Without boosting, both BDW and BSW are 1.7$\sigma$. After the boosted word-line is used, the writability is enhanced. However, BSW gets worse beyond 0.7V of the boosted supply because it makes the bitcells more prone to half select disturb. On the other hand, BDW monotonically increases as the boosted supply increases but saturates sooner. In conclusion, the word-line boosting is effective up to 0.75V and 3.5$\sigma$ of BDW is achieved.

### 2.4.2 Device Sizing Optimization

The second approach is device sizing optimization. Reference [10] shows that sizing optimization can achieve an iso-robustness condition while lowering the supply

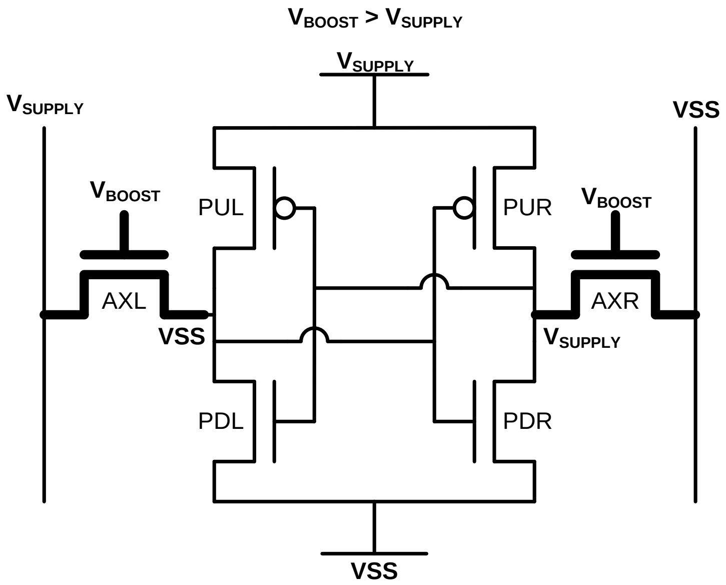


Figure 2.14: Word-line boosting

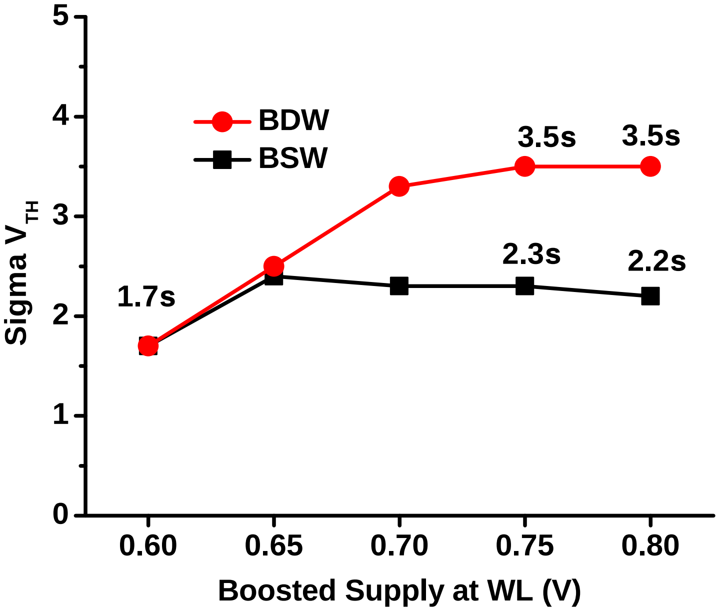


Figure 2.15: Bit-interleaved writability with word-line boosting at 0.6V

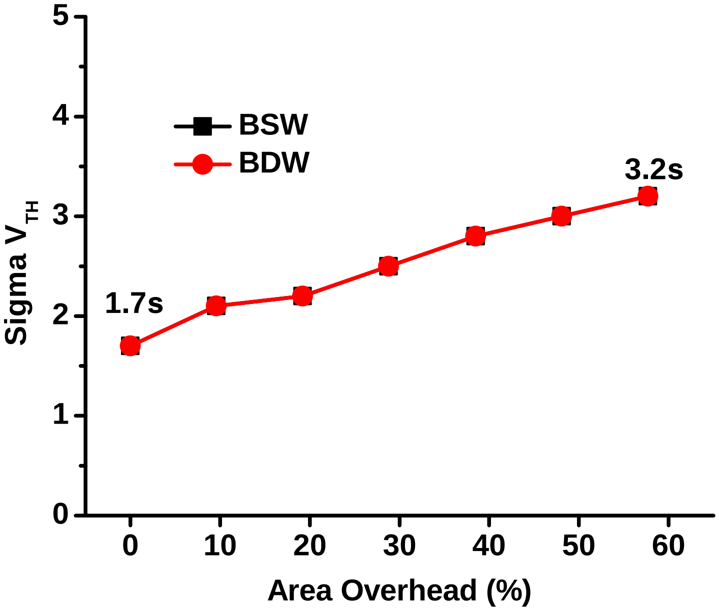


Figure 2.16: Bit-interleaved writability with device sizing at 0.6V

voltage, at the cost of density. Referring back to Figure 2.10 (a), write operation is mainly driven by AXR and PUR since AXR can drive logic "0" into the internal node and PUR keeps logic "1" in the internal node. The writability can be enhanced by increasing the width of access transistors or increasing the length of pull-up devices. On the other hand, strong pull-down devices are favorable to avoid half select disturb. Here we increase the width of access transistor and pull-down transistor simultaneously. Increasing the length of pull-down devices is not used because it makes a notch in poly which is not favorable for design for manufacturability. Area overhead is calculated based on layout. With device sizing, BDW and BSW monotonically increase at the same time (Figure 2.16). Because the way of device sizing in this subsection is favorable to both the writability and the half select disturb immunity, BSW also increases monotonically. With 57.7% of area overhead, BDW and BSW are extended to 3.2$\sigma$ from 1.7$\sigma$.

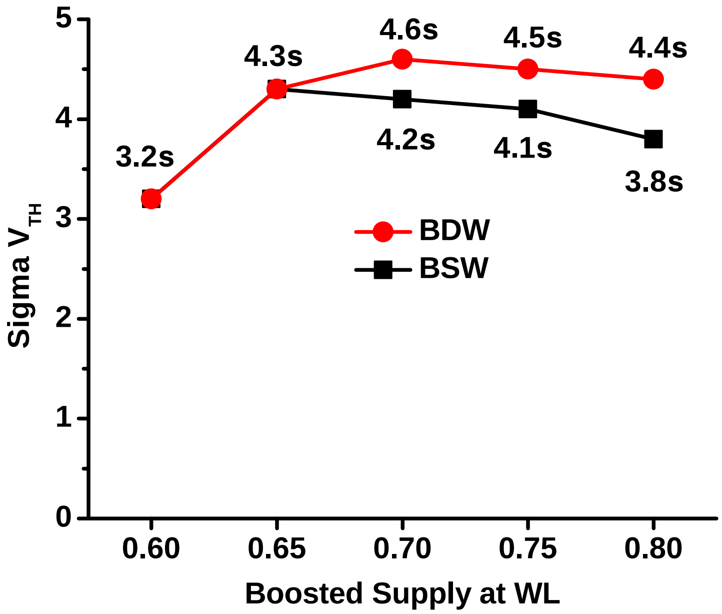


Figure 2.17: Bit-interleaved writability with word-line boosting and 57.7% area overhead

### 2.4.3 Dual Writability Enhancement

In the previous subsections, two writability enhancement techniques are used. However, both techniques are practically limited to below $4\sigma$. To achieve higher robustness, both techniques are applied simultaneously.

Figure 2.17 depicts how BSW and BDW change as the boosted word-line supply increases with a SRAM cell sized 57.7% larger than the nominal when the supply voltage is 0.6V. Since the two techniques are applied at the same time, BDW is extended to $4.6\sigma$ at 0.7V of boosted supply. Beyond 0.7V, the half select disturb overwhelms the writability so BDW decreases. This implies that higher word-line boosting does not guarantee better bit-interleaved writability. Also, we can clearly observe that BSW is too pessimistic and BDW reflects SRAM writability appropriately. In terms of energy consumption, the design shows 41% saving over normal 1.0V operation.

## 2.5 Conclusion

In this section, we discuss the writability and the half select disturb immunity of bit-interleaved 8-T SRAM arrays. The bit-interleaved static and dynamic writability analysis is proposed using worst case corner simulation to estimate the writability more precisely. At the end, two SRAM writability enhancement techniques are compared using the newly proposed analysis. The results show that device sizing and word-line boosting need to be used simultaneously to achieve higher robustness. To obtain the same robustness as 1.0V while lowering the supply voltage down to 0.6V, 0.1V of word-line boosting and 57.7% larger area are required. With these two techniques, we can successfully save the energy consumption per operation by 41%. In addition, the result confirms that higher word-line boosting does not guarantee better robustness because it lowers the half select immunity.

# CHAPTER III

# Design of Low Leakage SRAM

A low leakage memory is an indispensable part of any sensor application that spends significant time in standby (sleep) mode. Although using High Threshold Voltage (HVT) devices is the most straightforward way to reduce leakage, it also limits operation speed during active mode. In this chapter, a low leakage 10-T SRAM cell, which compensates for operation speed using a readily available secondary supply, is proposed in a 0.18$\mu$m CMOS process. It achieves the lowest-to-date leakage power consumption and achieves robust operation at low voltage without sacrificing operation speed. The 10-T SRAM has a bitcell area of 17.48$\mu$m$^2$ and is measured to consume 1.85fW per bit at 0.35V.

## 3.1 Introduction

Sensors with long lifetime are becoming increasingly popular in areas such as medical, infrastructure, and environmental monitoring [11, 13, 24]. In sensor applications, reducing the standby power consumption is as important as reducing the active power consumption since the sensors spend significant time in standby mode. To minimize the standby power consumption, designing low leakage memory is indispensible [36, 37, 42, 68]. Often, the leakage power consumption from memories dominates the total standby power consumption, since data stored in memory must

| Modules | Power Consumption during Sleep Mode at 400mV |
|---|---|
| Retention SRAM | 80.53% |
| Timer | 19.46% |
| CPU | <0.01% |
| Non-Retention SRAM | <0.01% |

Table 3.1: Sleep Power Breakdown of a Sensor Application [11]

be retained while most other blocks such as CPU, radios, and sensors can be fully power gated. Table 3.1 shows retention SRAM which is for storing data consumes more than 80% of its sleep power.

A low leakage 14-T SRAM cell with stacked HVT devices [24] has been previously proposed; however, its area is 9.1× larger than the traditional 6-T cell [17] and the HVT devices degrades write performance by more than 10× compared to the read speed. To overcome these limitations and reduce leakage further, this work proposes a new ultra low leakage SRAM, referred to as the low leakage 10-T SRAM, that exploits a boosted supply. We show how the boosted supply can increase operation speed and reduce leakage power simultaneously. Sensor applications typically operate using batteries, such as thin film batteries which tend to have high supply voltages. To obtain the subthreshold operating voltages, a common method for Direct Current (DC)-DC conversion is to use a Switched-Capacitor Networks (SCN) followed by a Low-Dropout Regulator (LDO). In this case, boosted supply can be obtained with minimal overhead since it can be directly obtained from the input of the LDO or from a higher voltage output from the ladder SCN [11]. Also, several circuit techniques, including a floating bit-line scheme, word-line keeper, and read buffer, are introduced to reduce leakage further and guarantee robust read and write operation.

A prototype chip, which has 24kb of the low leakage 10-T SRAM, shows that a bitcell consumes 1.85fW of standby power at 0.35V with 0.5V of boosted supply. To our knowledge, this marks the lowest-to-date SRAM leakage power. The bitcell area (Figure 3.1) is 17.48$\mu$m$^2$, 3.97× larger than a traditional 6-T cell [17] but 2.3×

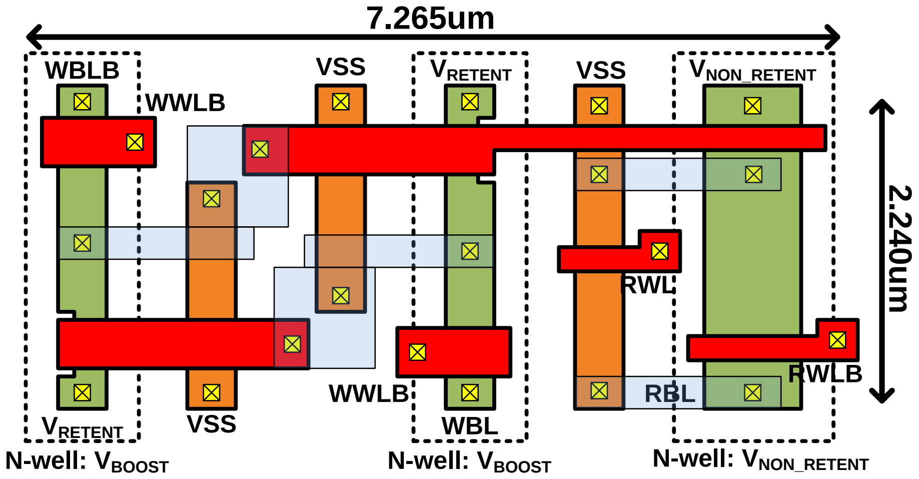


Figure 3.1: A low leakage 10-T SRAM layout is shown. Logic design rules are used for the 10-T SRAM layout

smaller than the previous low leakage 14-T SRAM [24]. If pushed SRAM design rules are used, the area overhead due to logic design rules can be mitigated. This SRAM is successfully demonstrated as a part of an integrated sensor system with a CPU, power management unit, solar cells, and battery [11, 13].

## 3.2 Background and Related Works

As technology scales, leakage becomes no longer negligible and must be minimized for low power operations. In this chapter, basic leakage mechanisms in Metal-Oxide-Semiconductor (MOS) devices are explained. In addition, common leakage reduction techniques are introduced and the effects of these techniques on SRAM leakage reduction are explained. At the end of this section, technology selection for leakage reduction is discussed.

### 3.2.1 Leakage Mechanisms

In figure 3.2, three types of basic leakage source are shown: subthreshold leakage, gate leakage, and junction leakage [59]. The subthreshold leakage and the gate leakage are expected to increase with the technology scaling while the junction leakage is a dominant factor in long-channel devices [59].

The first type of leakage is subthreshold leakage. When a MOS device is in the cut-off mode, there is still current between drain and source even though the amount of current is relatively very small compared with $I_{on}$. To understand the subthreshold leakage in analytic way, the well-known subthreshold leakage equation is following [64]:

$$I_{ds} = \mu C_{ox} \frac{W}{L} (m-1) \left( \frac{kT}{q} \right)^2 e^{\frac{V_{gs}-V_{th}}{mkT/q}} \left( 1 - e^{\frac{-V_{ds}}{kT/q}} \right) \tag{3.1}$$

where

$$m = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{3t_{ox}}{W_{dm}} \tag{3.2}$$

where $\mu$ is the mobility, $C_{ox}$ is the gate oxide capacitance, $\frac{kT}{q}$ is the thermal voltage, $C_{dm}$ is the bulk depletion capacitance, $t_{ox}$ is the gate oxide thickness, and $W_{dm}$ is the maximum depletion layer width. The equation itself is not simple but it is easy to notice that the subthreshold leakage is exponentially proportional to $(V_{gs} - V_{th})$. When $V_{gs} = 0$, the subthreshold leakage is $I_{off}$. As technology scales, $V_{th}$ scaling results in higher subthreshold leakage [15] and therefore it becomes a significant source of power dissipation.

The second type of leakage current is gate leakage. Gate leakage is due to thin gate oxide and the amount of gate leakage is significant in sub-100nm technology [4, 46]. The thin gate oxide results in higher electric field across the oxide and it causes the tunneling of electrons. To reduce the gate leakage, high-$\kappa$ gate dielectric material was introduced with 45nm technology [44].

The third type of leakage current is pn junction reverse-bias leakage. There are two main components of the junction leakage: minority carrier diffusion/drift and electron-hole pair generation [59].

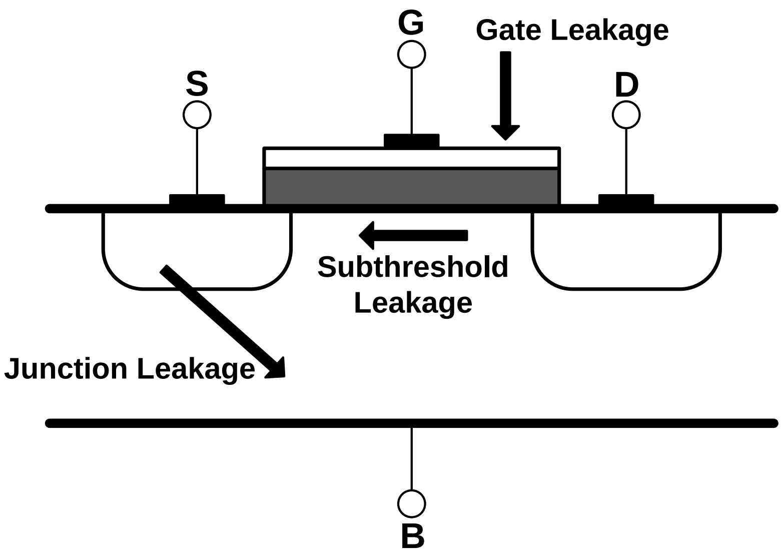


Figure 3.2: There are three types of leakage source in MOS devices [59].

### 3.2.2 Leakage Reduction Techniques

The most straightforward way to reduce leakage is to use a device with low $I_{off}$. In other words, using a HVT device or a device with longer channel is beneficial in terms of leakage reduction. However, a device with low $I_{off}$ usually has low $I_{on}$ (subsection 5.3.1) and it will limit the operating speed of a device during active mode. Therefore, appropriate circuit techniques to reduce leakage while minimizing performance degradation are indispensable for low power operations.

In this subchapter, four circuit techniques to minimize leakage are introduced: transistor stacking, power gating, body biasing, and supply voltage ramping. These techniques are used to reduce mainly the subthreshold leakage current. The transistor stacking is for both active and sleep modes while the other techniques are for sleep mode only.

#### 3.2.2.1 Transistor Stacking

It has been observed that the leakage current through multiple stacked off devices are significantly smaller than the leakage current through one off device [50].

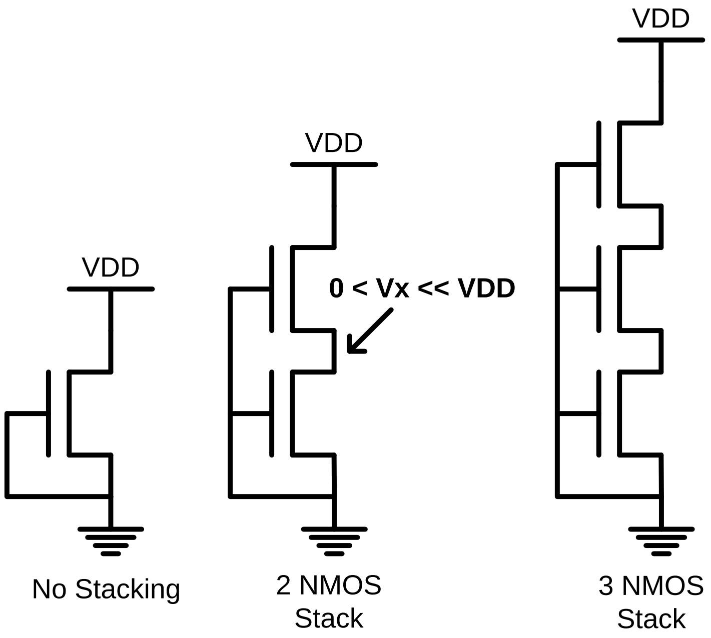


Figure 3.3: Transistor stacking schematic

The transistor stacking concept is shown in Figure 3.3. No stacking, 2 N-channel MOS (NMOS) stack, and 3 NMOS stack are parts of inverter, 2-input Negated AND (NAND) gate, and 3-input NAND gate, respectively. In 2 NMOS stack, the intermediate voltage between two NMOS devices ($V_x$) is much lower than $V_{DD}$ but slightly higher than ground. Because $V_x$ is higher than ground, $V_{gs}$ of upper NMOS is negative and therefore leakage through the NMOS stack is much smaller than without stacking. The same explanation can be applied to 3 NMOS stack and it shows even lower leakage.

As the number of devices in a stack increases, both $I_{off}$ and $I_{on}$ decrease. With 2 NMOS stack, 80% of leakage current reduction is shown in Figure 3.4. However, the leakage reduction in >2 devices stacking is not as effective as 2 devices stacking.

Without changing any circuits, leakage can be minimized for a certain input pattern during standby mode and the input pattern can be found using the transistor stacking concept. The more stacks with off transistors, the less leakage power. Another application of the transistor stacking is stack-forcing: redundant devices which

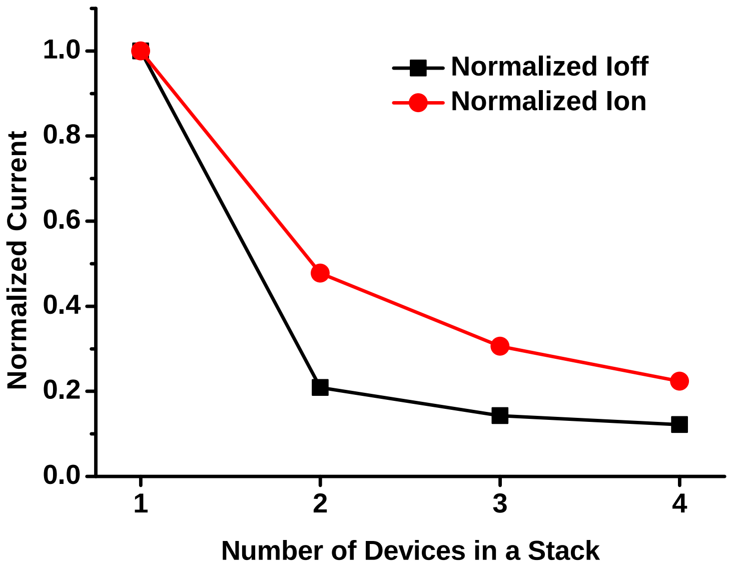


Figure 3.4: Normalized $I_{off}$ and $I_{on}$ current as the number of devices in a stack increases. This is a simulated result with 45nm CMOS technology.

do not change the logic function of the gate are inserted only for leakage reduction purpose. At the expense of area, the stack-forcing can be used to reduce leakage during active mode without overall performance degradation when it is used in a non-critical path.

The transistor stacking is also used in SRAM for leakage reduction. Reference [24] adopts a 14-T SRAM with stack-forced cross-coupled inverters to minimize leakage at the expense of area and write operation speed.

#### 3.2.2.2 Power Gating

Using Low Threshold Voltage (LVT) devices is essential in performance-driven circuit design but it has non-negligible leakage current. To minimize leakage current in standby mode while keeping high performance in active mode, power gating has been proposed [49, 67]. The power gating is also called Multithreshold-Voltage CMOS (MTCMOS) since it requires two different threshold devices: HVT devices for power

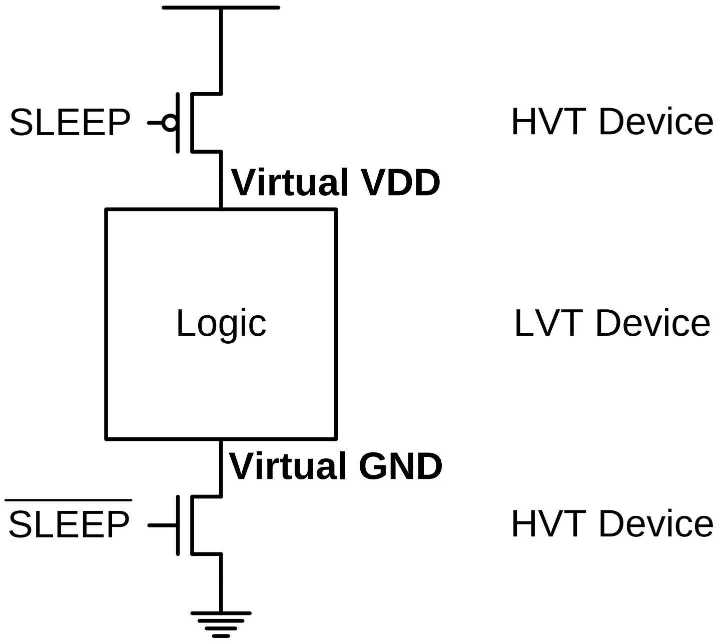


Figure 3.5: There are power gating switches using HVT devices. The logic is connected to virtual $V_{DD}$ and virtual ground.

gating switches (sleep transistors) and LVT devices for logic. When the power gating technique is used, logic with LVT devices is connected to virtual $V_{DD}$ and virtual ground instead of being connected directly to $V_{DD}$ and ground (Figure 3.5). The power gating switches between virtual power supplies and real power supplies control the voltage level of virtual power supplies and the amount of leakage current. In active mode, the power gating switches are completed on and the voltage levels of virtual power supplies are very close to the real power supplies. However, in standby mode, virtual supplies are collapsed together and the leakage will be limited by the leakage through HVT power gating switches. Even though there are both header (PMOS power gating switch) and footer (NMOS power gating switch) in Figure 3.5, it is also possible to use header only or footer only.

There is still a trade-off between leakage and performance by sleep transistor. If a sleep transistor is designed very small for further leakage reduction, it will also limit the operation speed of logic. To overcome this limitation and to make power gating more effective, a boosted sleep transistor was proposed [28]. By using boosted supply at the gate of power gating switches, the trade-off between leakage and performance is mitigated.

When the power gating is adopted, it is not possible to preserve the state during sleep mode due to the nature of collapsing supplies. Therefore, the power gating cannot be used in SRAM bitcells for data retention. In reference [11, 13], only the read path of the bitcell not used in sleep mode is power gated for leakage reduction.

#### 3.2.2.3 Body Biasing

If the voltage levels of source and body in a MOS device is different, the voltage difference induces a change in threshold voltage. This is called the body effect and the threshold voltage is determined by

$$V_{th} = V_{th0} + \gamma(\sqrt{|(-2)\phi_F + V_{sb}|} - \sqrt{|2\phi_F|}) \tag{3.3}$$

where $\gamma$ is the body-effect coefficient and $\phi_F$ is the *Fermi Potential* [55]. In table 3.2, body biasing operation modes are summarized. When a device is reverse biased, $V_{th}$ increases and therefore both $I_{on}$ and $I_{off}$ decrease. To avoid performance degradation during active mode, the reverse body biasing can be applied only in sleep mode to reduce sleep power consumption.

Even though the body biasing is very effective in adaptive post-silicon tuning, there are some drawbacks. When NMOS body biasing is used with P-substrate silicon, it requires the triple-well technology which has an enormous area penalty. Also, there is an energy cost of charging and discharging the substrate/well capacitance. In addition, a dedicated circuit such as a charge pump is necessary for generating voltage levels other than $V_{dd}$ and $V_{ss}$.

In Reference [35], a deep sleep mode is proposed to minimize leakage in SRAM bitcells. Instead of modulating the voltage of the body, the voltage of the source

| Device | No Body Biasing | Forward Body Biasing | Reverse Body Biasing |
|---|---|---|---|
| NMOS | $V_b = V_{ss}$ | $V_b > V_{ss}$ | $V_b < V_{ss}$ |
| PMOS | $V_b = V_{dd}$ | $V_b < V_{dd}$ | $V_b > V_{dd}$ |

Table 3.2: Body Biasing Operation

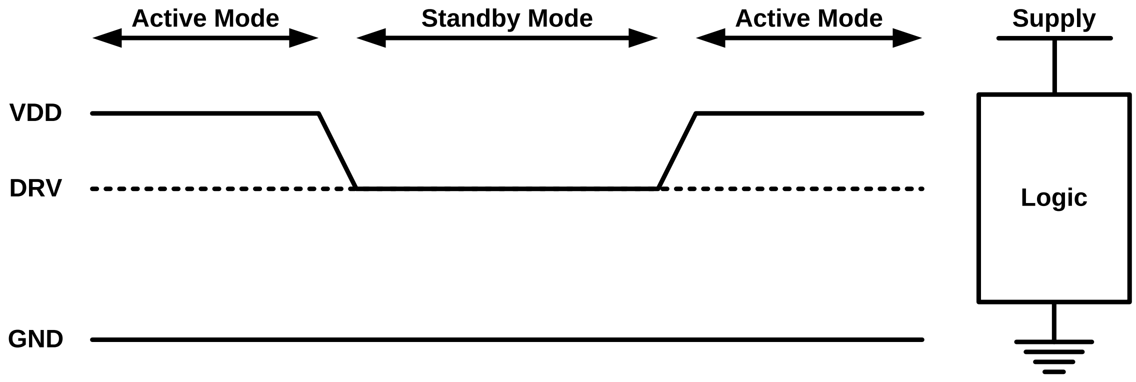


Figure 3.6: The supply voltage decreases to data retention voltage (DRV) in standby mode.

of NMOS devices increases and $V_{gs}$ and $V_{bs}$ become negative. Negative $V_{bs}$ leads leakage reduction by the reverse body biasing.

#### 3.2.2.4 Supply Voltage Ramping

The leakage power consumption is calculated by the following equation:

$$P_{leakage} = V_{DD} \times I_{leakage} \tag{3.4}$$

where $V_{DD}$ is the supply voltage and $I_{leakage}$ is the leakage current. If the supply voltage decreases in sleep mode, both $V_{DD}$ and $I_{leakage}$ decrease and therefore huge leakage power saving is expected. If the state in logic must be preserved, the supply voltage scales down to the data retention voltage (DRV) in standby mode (Figure 3.6) and it is restored to the normal voltage level ( $V_{DD}$) in active mode.

According to equation 3.1, the leakage current decreases as $V_{ds}$ $(=V_{DD})$ decreases

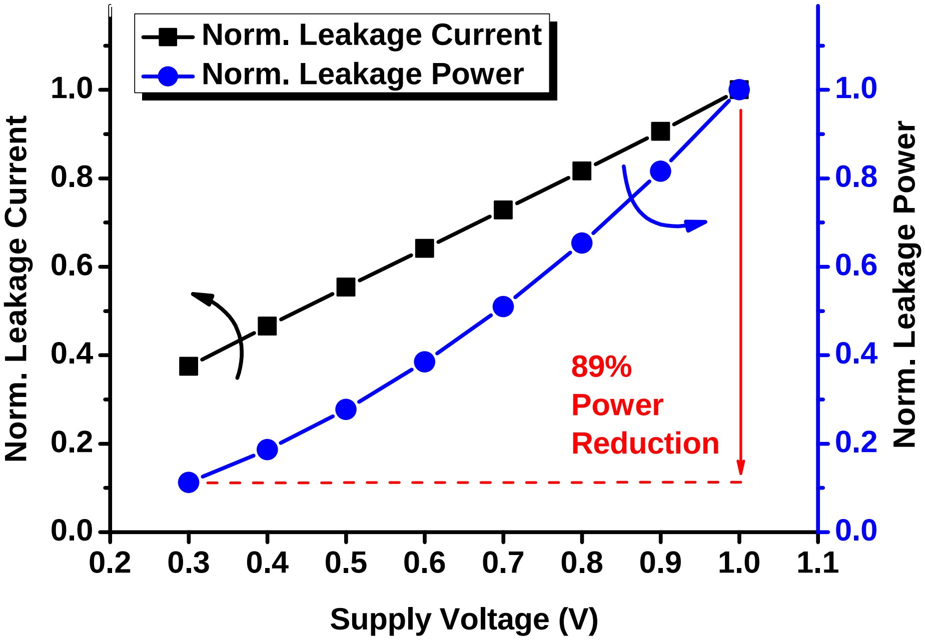


Figure 3.7: 89% of leakage power saving is observed in 45nm CMOS technology if the supply decreases from 1.0V to 0.3V.

and $V_{th}$ increases. In addition to $V_{ds}$ decreasing by the supply voltage ramping, $V_{th}$ also increases due to drain-induced barrier lowering (DIBL) [55]. Figure 3.7 depicts the leakage power saving by the supply voltage ramping. If the DRV is 0.3V, 89% of leakage power saving is observed.

However, the supply voltage ramping needs a second voltage source (for DRV) or a voltage controllable voltage regulator. Also, it usually requires long re-activation time from standby mode to active mode.

The supply voltage ramping is used in drowsy cache [34]. When a cache bank is not used, the supply voltage of the cache bank scales down to the data retention voltage of SRAM and it shows huge leakage power reduction while data in SRAM are preserved.

### 3.2.3 Technology Selection for Leakage Reduction

Scaling in technology leads higher $I_{on}$ but not lower $I_{off}$. This implies that we need to carefully select technology based on the target application to achieve the lowest energy consumption. Reference [62] explains that the old technology with higher $V_{th}$ can be beneficial in terms of energy consumption in a sensor application which spends most of its lifetime in sleep mode.

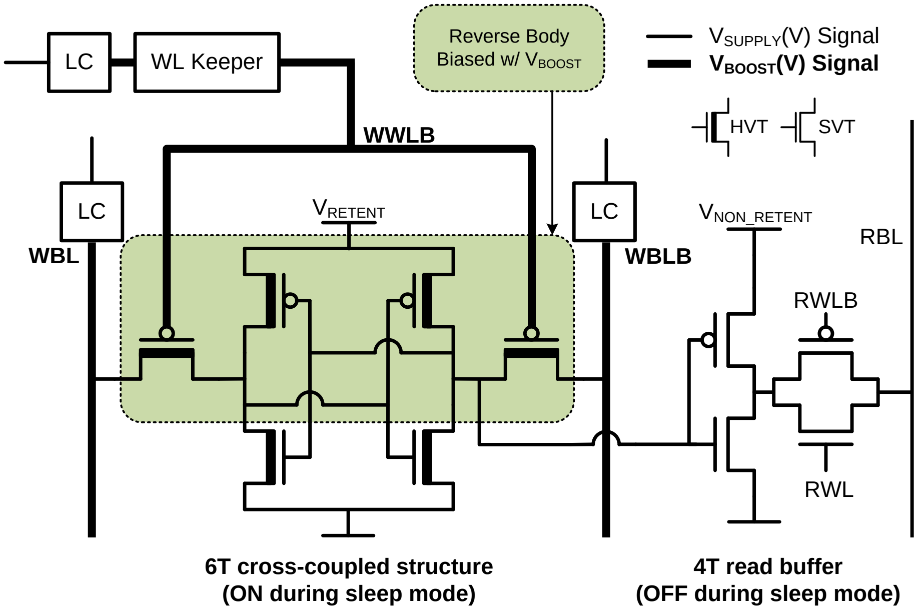
## 3.3 Design of Low Leakage SRAM

Low leakage SRAM is designed for sensor applications using carefully selected 0.18$\mu$m CMOS technology. This section includes SRAM bitcell design as well as SRAM peripheral design for low leakage operation.

### 3.3.1 SRAM Bitcell and Operation Modes

Figure 3.8 shows the proposed 10-T SRAM schematic to minimize leakage current without sacrificing operation speed. It consists of a 6-T cross-coupled structure and a 4-T read buffer. The read buffer can be power gated while the cross-coupled structure must remain on to retain data. Thus, Standard Threshold Voltage (SVT) devices are used in the read buffer for fast read operation and HVT devices are used in the cross-coupled structure for minimizing leakage. The layout is shown in Figure 3.1 and logic design rules are used.

This SRAM operates with three different power supplies: $V_{RETENT}$, $V_{NON\_RETENT}$, and $V_{BOOST}$. $V_{RETENT}$ and $V_{NON\_RETENT}$ have the same voltage level but are connected to different power gating switches. $V_{BOOST}$ has a higher voltage level than the other two supplies and is used for boosting bit-lines and reverse body biases the four HVT PMOS devices. Boosting bit-lines enhances write operation speed, and reverse body biasing allows further leakage reduction. The cell is still functional if $V_{BOOST}$



LC
WL Keeper
Reverse Body Biased w/ V_BOOST
V_SUPPLY(V) Signal
V_BOOST(V) Signal
HVT
SVT
WWLB
LC
WBL
V_RETENT
LC
WBLB
V_NON_RETENT
RBL
RWLB
RWL
6T cross-coupled structure (ON during sleep mode)
4T read buffer (OFF during sleep mode)

Figure 3.8: A low leakage 10-T SRAM schematic is shown. Three signals (WBL, WBLB, WWLB) are boosted to $V_{BOOST}$ using level converters to enhance write operation. Four PMOS devices in 6-T cross-coupled structure are reverse body biased with $V_{BOOST}$ for further leakage reduction.

| Mode | Active | Standby | Shutdown |
|---|---|---|---|
| $V_{RETENT}$ | $V_{SUPPLY}$ | $V_{SUPPLY}$ | 0 |
| $V_{NON\_RETENT}$ | $V_{SUPPLY}$ | 0 | 0 |
| $V_{BOOST}$ | $\geq V_{SUPPLY}$ | $\geq V_{SUPPLY}$ | 0 |

Table 3.3: Low Leakage 10-T SRAM Operation Modes

is the same as $V_{RETENT}$ and $V_{NON\_RETENT}$.

There are three operation modes (See Table 3.3). During active mode, $V_{RETENT}$ and $V_{NON\_RETENT}$ are at $V_{SUPPLY}$ while $V_{BOOST}$ is higher than the two others. When the power gating switch connected to $V_{NON\_RETENT}$ is turned off, the system moves to standby mode. To retain data, $V_{RETENT}$ still remains at $V_{SUPPLY}$. $V_{BOOST}$ must also be kept on to turn off the access transistors and bias the n-well. If no data retention is required, all supplies can be turned off.

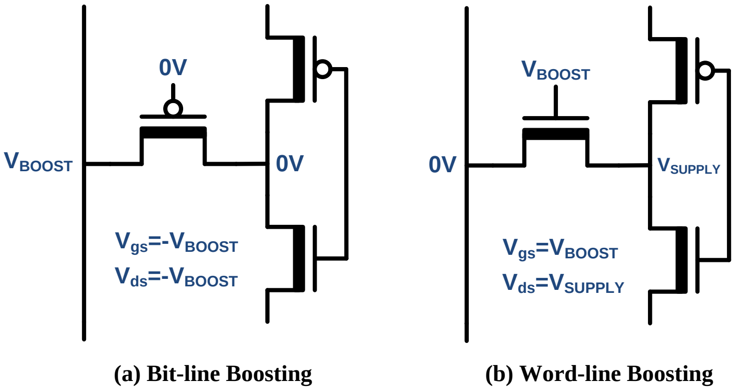


Figure 3.9: (a) Bit-line boosting with PMOS access transistor. (b) Word-line boosting with NMOS access transistor.

### 3.3.2 Bit-line Boosting for Fast Write Operation

The read operation already has an acceptable access time of below 20 SVT Fan-Out of 4 (FO4) delays including cascaded read buffer delays because SVT devices are used in the 4-T read buffer. However, without further modification, the write operation limits the performance of this SRAM cell at more than 1000 SVT FO4 delays because of slow HVT devices in the 6-T cross-coupled structure.

Write speed can be improved by increasing $I_{on}$ of the access transistor. First, PMOS access transistors are used instead of the traditional NMOS access transistors since, at low voltage in this technology, HVT PMOS devices have larger $I_{on}$ than HVT NMOS devices. Second, bit-line boosting is adopted (Figure 3.9(a)). With NMOS access transistors, writing "0" is dominant and a boosted word-line can increase $I_{on}$ by raising $V_{gs}$ of the NMOS (Figure 3.9(b)). On the other hand, with PMOS access transistors, writing "1" is dominant and bit-line boosting is applied. With bit-line boosting, both $V_{gs}$ and $V_{ds}$ of the PMOS are boosted and therefore it results in better performance improvement. Since bit-line boosting is more effective than word-line boosting and since a negative power supply is not readily available, PMOS devices were selected.

Simulated behavior in Figure 3.10 depicts HVT SRAM write speed improvement

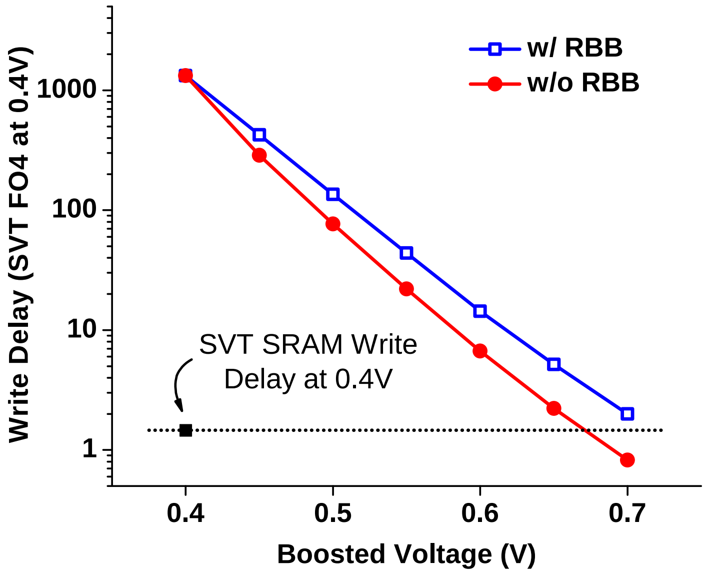


Figure 3.10: Bit-line boosting significantly improves write speed (simulated results).

as boosted supply increases. The write speed in this plot does not include peripherals to directly compare the effect of bit-line boosting. The effect of reverse body biasing will be discussed in the following subsection. Without boosting, HVT SRAM needs more than 1000 SVT FO4 delays and, therefore, a processor must run many cycles for a single write operation. As boosted supply increases, the write speed is dramatically improved and the write operation can be executed in a single or a few cycles.

### 3.3.3 Body Biasing with Boosted Supply

Reverse body biasing of four HVT PMOS devices in the 6-T structure is adopted for leakage minimization without increasing bitcell area. Figure 3.11 compares leakage current of an HVT PMOS device with stack forcing and reverse body biasing (RBB) at 0.4V in simulation. It shows that stack forcing is not as effective as reverse body biasing at low voltage. With more than 50mV of reverse body biasing, leakage reduction is better than stacking two devices. In addition, stack forcing needs more devices and, therefore, increases bitcell area. Because of the optimized layout of the

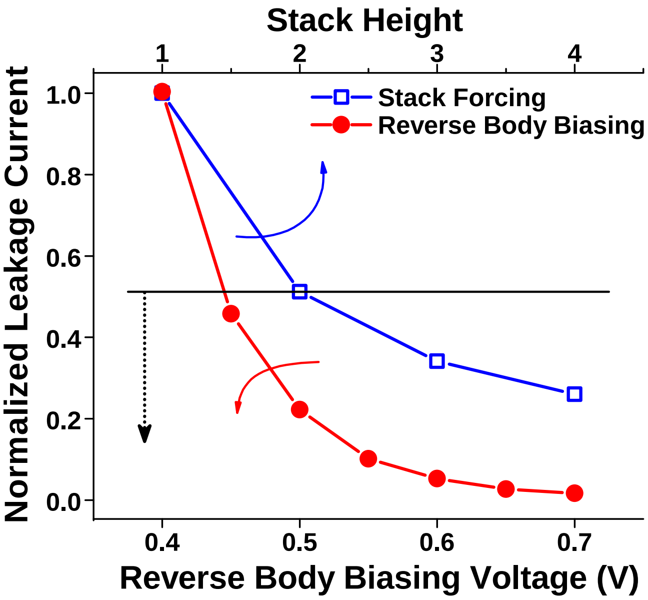


Figure 3.11: Leakage current of a PMOS device is shown. Stack height means the number of devices in a stack. Reverse body biasing is more effective than stack forcing (simulated results).

6-T structure, adding stacked devices causes a more than 2× area increase.

Reverse body biasing decreases both $I_{on}$ and $I_{off}$, so it can degrade write operation significantly. However, the access transistor in the dominant writing "1" path does not experience reverse body biasing during write since the bit-line boosting scheme increases voltage level of source while the body biasing increases by the same amount. Vbs is still 0V during write operation, so reverse body biasing for all four HVT PMOS devices in the 6-T structure does not weaken write operation. In Figure 3.10, HVT SRAM can achieve sufficient speed improvement even with reverse body biasing.

### 3.3.4 Leakage Reduction during Standby Mode

There are four different subthreshold leakage paths in an SRAM cell during standby mode (Figure 3.12). The 0.18$\mu$m CMOS technology has a thick gate oxide so the gate leakage is ignored. The 4-T read path is power gated so it is not considered as a leakage path. $V_{BIT}$ and $V_{BIT\_B}$ affect $I_{AXL}$ and $I_{AXR}$, but do not

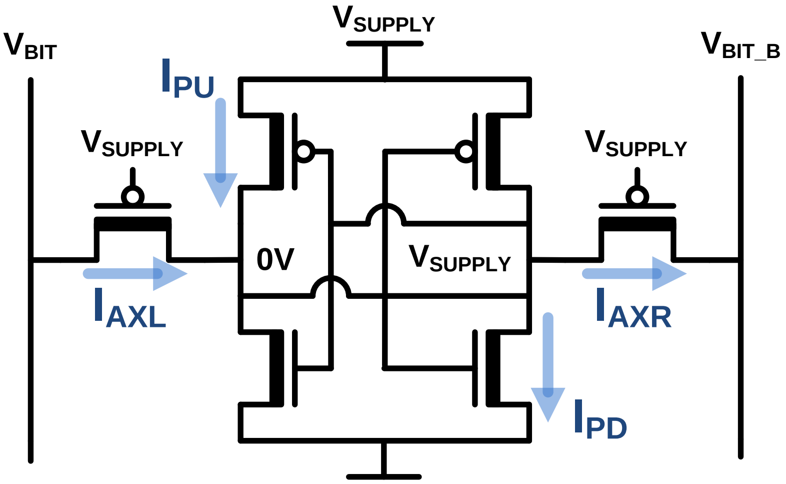


Figure 3.12: There are four leakage paths during standby mode.

impact $I_{PU}$ and $I_{PD}$. If $V_{BIT}$ and $V_{BIT\_B}$ keep either $V_{SUPPLY}$ or 0V, leakage current exists in only one path between $I_{AXL}$ and $I_{AXR}$ and the amount of leakage through each path is the same. This implies that the total leakage current does not change as long as bit-lines are driven to $V_{SUPPLY}$ or 0V. We propose using bit-lines that are floating. In this case, the voltage levels of bit-lines are determined by data stored in the cells connected to the same bit-line. With all the same data in a bit-line, there is no leakage through access transistors. Otherwise, $V_{BIT}$ and $V_{BIT\_B}$ are in an intermediate voltage between $V_{SUPPLY}$ and 0V and therefore an access transistor whose internal node is “0” is super cut-off. In simulation, this mechanism allows at least an addition 18% leakage reduction, and could decrease leakage further depending on the data stored in the cells (Figure 3.13).

Simulated results show that PMOS reverse body biasing is also effective to minimize leakage during standby mode (Figure 3.14). $I_{AXL}$ and $I_{AXR}$ can be minimized with BL floating and RBB, while $I_{PU}$ can be minimized with RBB only. If two techniques are applied, the only remaining leakage path is $I_{PD}$. NMOS reverse body biasing to reduce $I_{PD}$ is not practical since a triple well process is required, increasing bitcell area tremendously. Additionally, it is relatively difficult to obtain negative

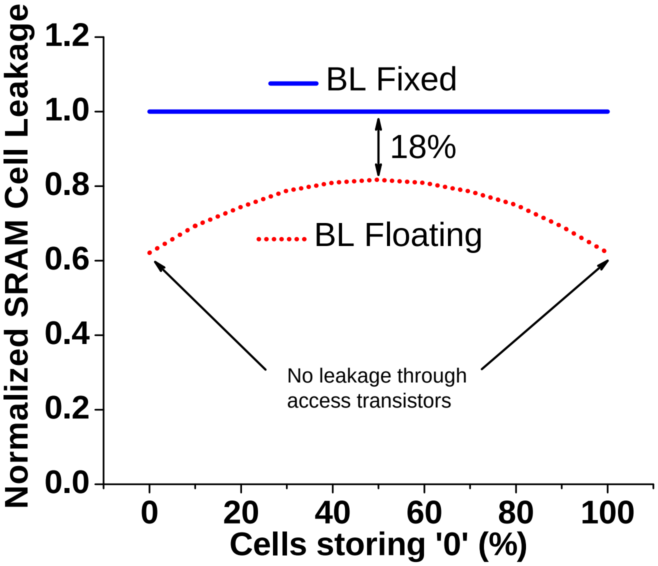


Figure 3.13: Bit-line floating shows at least 18% leakage reduction (simulated results).

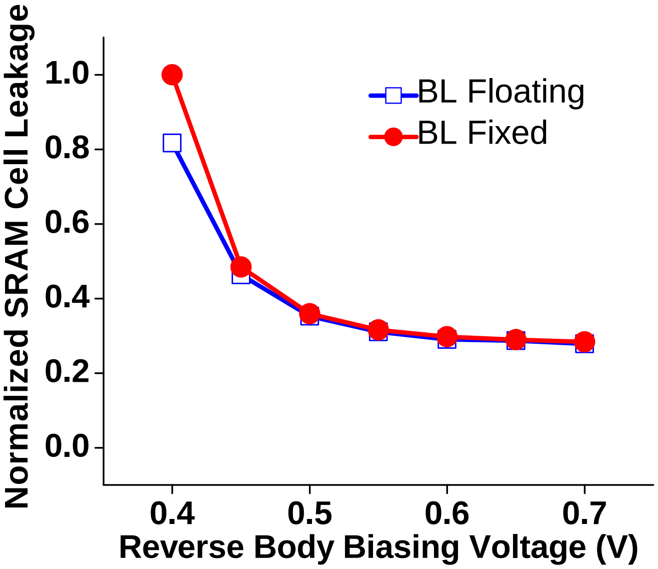


Figure 3.14: SRAM leakage and PMOS reverse body biasing (simulated result).

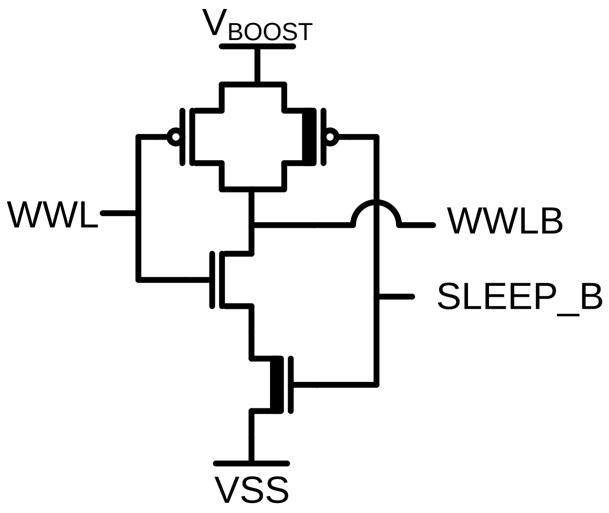


Figure 3.15: Word-line keeper for high performance operation during active mode and low leakage operation during standby mode

power supply compared to boosted power supply since the boosted supply can be easily obtained from the higher voltage output from the ladder switched-capacitor networks in a DC-DC converter.

A special purpose word-line keeper (Figure 3.15) is designed to obtain two goals: no speed degradation and no SVT leakage path. The word-line keeper operates just like a normal word-line driver during active mode, but its output must be kept high in standby mode to fully turn off PMOS access transistors and prevent data corruption. The voltage level of SLEEP_B is higher than 3V (output voltage of small form-factor battery such as a Li battery) since this control signal is generated to control power gating switches. The use of the battery voltage level signal is justified since a small form-factor battery is used in most sensor applications.

### 3.3.5 Read Buffer Design

An improved 4-T read buffer is designed for robust and fast read at low voltage. A static logic circuit (4-T read buffer) can prevent erroneous bit-line discharge, which may occur in a dynamic logic circuit (2-T read buffer [8]), due to its relatively small

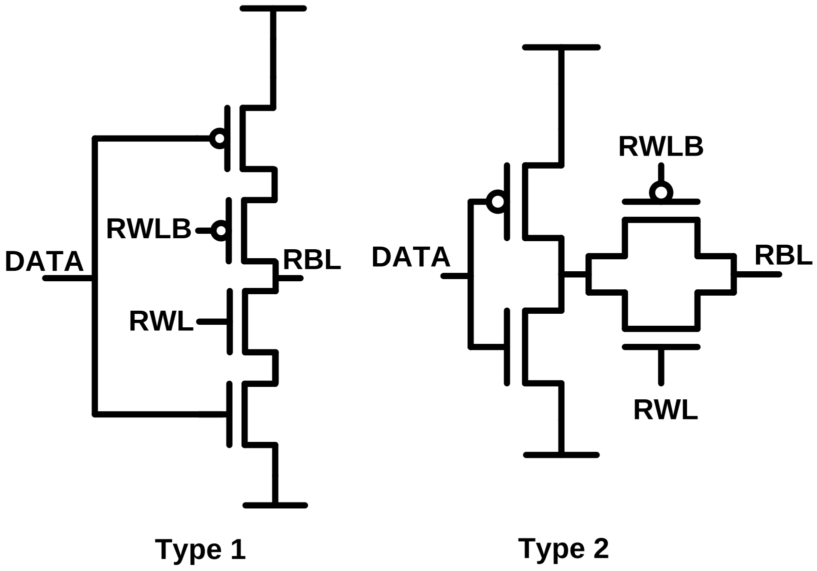


Figure 3.16: Two different 4-T read buffers

ON-OFF current ratio at low voltage. A clocked-gate type 4-T read buffer was used in [24] (Figure 3.16 Type 1) while a tri-state buffer type 4T read buffer ((Figure 3.16 Type 2) is used in this design. Type 2 is faster than Type 1 since both NMOS and PMOS in Type 2 can drive RBL when only one device can drive RBL in Type 1. With this new 4-T read buffer, RBLs are cascaded instead of directly connecting all bitcells to one global RBL. In RBL cascading, eight bitcells are connected to a local RBL and then local RBLs are cascaded to a global RBL. In the worst case, data in the all unselected cells are different from data in the targeted cell and therefore RBL leakage disturbs read operation. Since RBL leakage can be minimized with cascading, it can improve read speed. With new 4-T read buffer with cascading, read speed can be improved by 72% in simulation (Figure 3.17).

## 3.4 Measurement Results

A 24kb low leakage 10-T SRAM was fabricated in a 0.18$\mu$m CMOS process with nominal voltages of 1.8V and 3.3V for SVT and HVT respectively.

The SRAM array has 768 words and each word has 32-bit data. In table 3.4, the

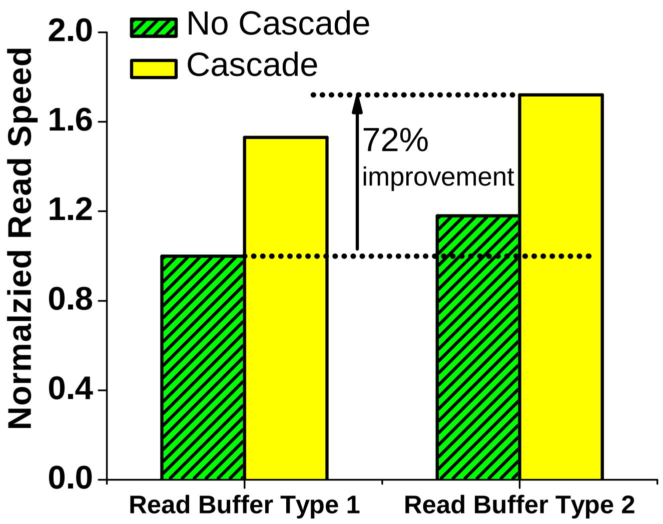


Figure 3.17: Read buffer type 2 with cascading increases read speed 72% (simulated result).

| Supply Voltage (V) | 1kHz | 2kHz | 4kHz | 8kHz | 16kHz |
|---|---|---|---|---|---|
| 0.300 | 0 | 23 | 32 | 32 | 32 |
| 0.325 | 0 | 0 | 31 | 32 | 32 |
| 0.350 | 0 | 0 | 1 | 32 | 32 |
| 0.375 | 0 | 0 | 0 | 3 | 32 |
| 0.400 | 0 | 0 | 0 | 0 | 12 |

Table 3.4: The number of failure words

first 32 words are tested to measure how many words fail as supply and operating frequency are swept. This table shows that there are more words that fail at low voltage and high frequency.

Figure 3.18 depicts speed improvements as boosted supply increases. At 0.35V, the whole SRAM array (768 words) with peripherals operates at 3.5kHz without read and write fail. If the boosted supply is applied, the write speed is enhanced and therefore the system can operate substantially faster. With 0.5V of boosted supply, the operating frequency can reach 52.5kHz, which is ~185 SVT FO4 delays at 0.35V, and this is 15× speed improvement. However, the speed improvement is saturated since a read buffer, peripherals, and CPU still run at 0.35V and the boosted supply

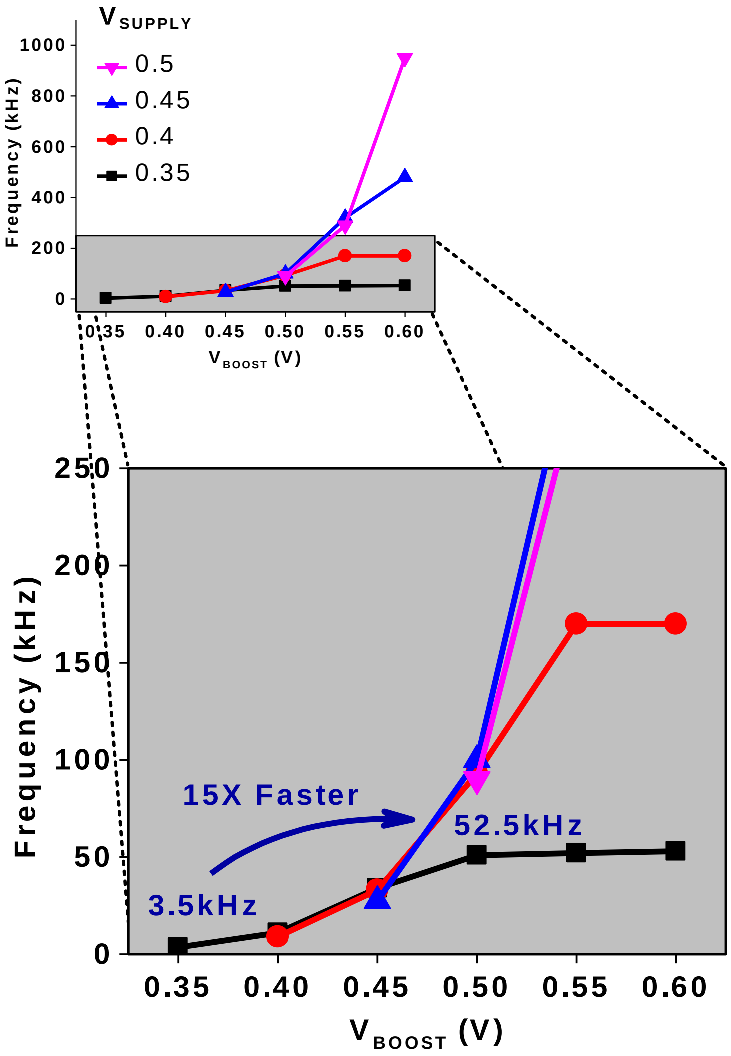


Figure 3.18: Operation speed is improved 15× with boosted supply (measured result).

does not change their operation speed.

Figure 3.19(a) shows the measured leakage power per bitcell as supply and boosted supply are swept in two different dies. Without boosted supply, the total leakage power monotonically increases as supply increases. If boosted supply is applied, the leakage power through $V_{RETENT}$ significantly decreases as it reverse body biases PMOS devices. However, the leakage power through $V_{BOOST}$ increases since it includes leakage through word-line keeper as well as body leakage. Because of these two different power trends, the optimal power minimum ($P_{min}$) point can be found.

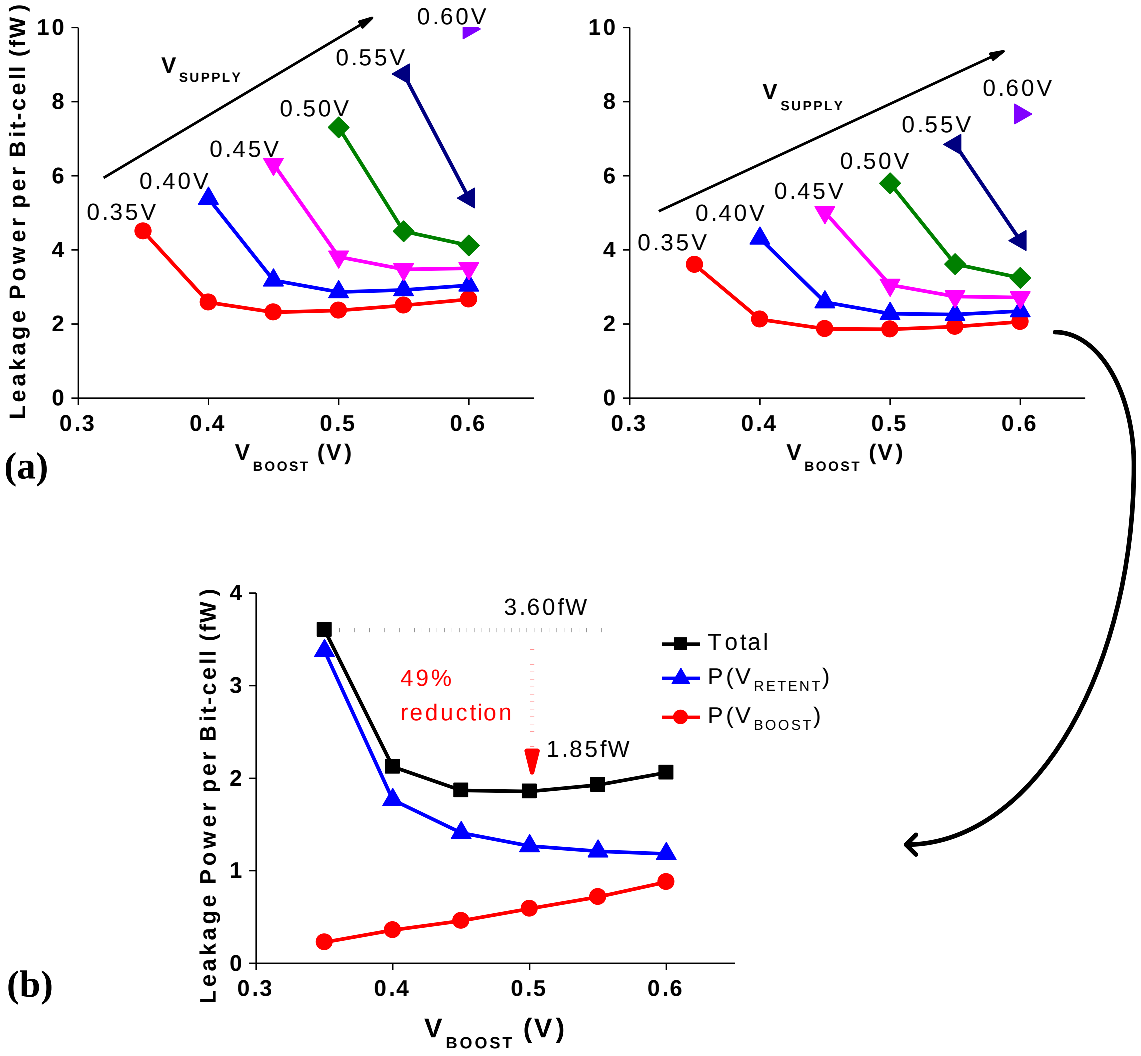


Figure 3.19: (a) Total leakage power is shown as $V_{SUPPLY}$ and $V_{BOOST}$ are swept in two different dies (measured result). (b) Leakage power can be minimized with boosted supply (measured result).

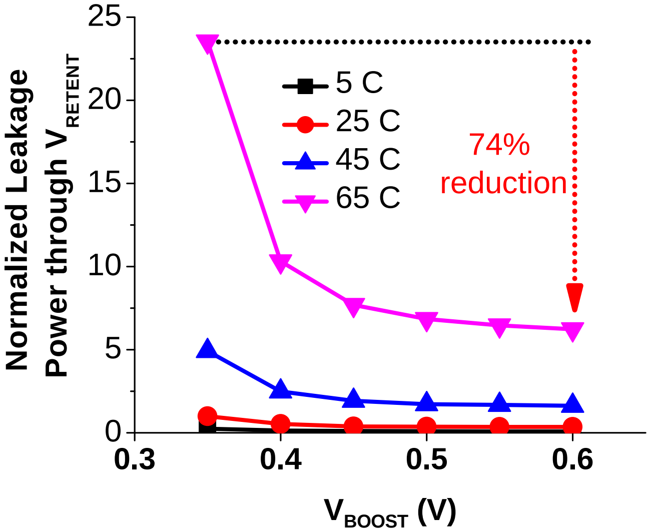


Figure 3.20: Temperature variation of normalized leakage through $V_{RETENT}$ (measured result).

Figure 3.19(b) shows the power breakdown. Two dies in Figure 3.19(a) show almost equivalent leakage reduction characteristics. At 0.35V (Figure 3.19(b)), the initial leakage power is 3.6fW without boosting, but it can be minimized down to 1.85fW with 0.5V of boosted supply. This is 49% leakage power reduction. In Figure 3.20, leakage powers through $V_{RETENT}$ are normalized in the different temperatures at 0.35V. In the different temperature, the boosted supply still allows large leakage reduction. Figure 3.21 shows the chip micrograph and dimension.

## 3.5 Conclusion

A low leakage 10-T SRAM which consumes femtowatt-scale leakage power is proposed for long lifetime sensor applications. A boosted supply is exploited to compensate slow write operation caused by HVT devices and minimize leakage further. A prototype chip fabricated in a 0.18$\mu$m CMOS process shows that 1.85fW of leakage power at 0.35V and operates at 52.5kHz with boosted supply. The boosted supply allows 49% of leakage power reduction and 15× speed improvement at 0.35V.

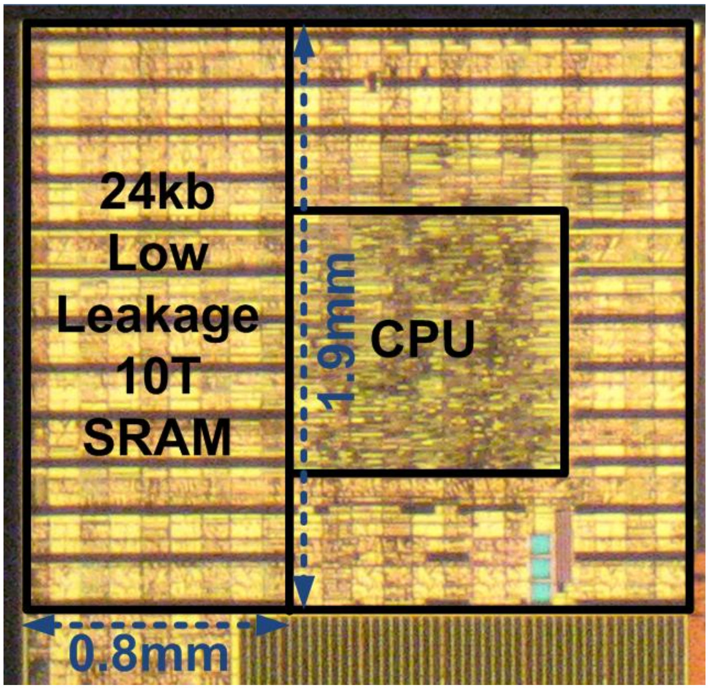


Figure 3.21: Chip micrograph and dimension in 0.18$\mu$m CMOS process

## CHAPTER IV

# Adaptive Write Architecture for Low Voltage 8-T SRAMs

To maintain enough robustness of SRAM at a scaled technology, an 8-T bitcell has been introduced. However, as we discussed in Chapter II, there is a limitation on optimizing write operation at a bit-interleaved array due to the trade-off between write and half select disturb. As the supply voltage is lowered to reduce power consumption, the margin between write and half select disturb becomes smaller and it limits lowering the minimum operation voltage ($V_{min}$). However, it is noticeable that the margin is calculated based on the worst case bitcell (random variation) and the worst case condition (systematic variation). Because write operation is done for one word at a time not for the whole array, most write operations do not need the worst case margining. In this chapter, we propose an adaptive WWL width and voltage modulation technique for bit-interleaved 8-T SRAMs to maximize yield and lower the $V_{min}$ by monitoring write completion. A prototype chip fabricated in 65nm CMOS process shows 3.96× leakage power reduction and 4.24× active power reduction with the lowered $V_{min}$.

## 4.1 Introduction

Low voltage operation is one of the most effective ways to reduce power consumption due to its quadratic power saving. However, power saving can be achieved at the expense of performance degradation. The first reason for performance degradation at low voltage is decreased $I_{on}$. In addition to this intrinsic performance degradation, the speed of a system at low voltage is further limited by larger variation. Therefore, mitigating variation is important to extend the usage of low voltage operation [19] without excessive performance variation.

Developing an SRAM array at low voltage is another challenge of low voltage operation. With a 6-T traditional SRAM bitcell, robust operation at low voltage is not feasible because it is not possible to maintain enough write and read margins at the same time. At the expense of two more transistors, an 8-T SRAM bitcell has been proposed to mitigating variation [8, 9]. By decoupling read and write, an 8-T bitcell enables lowering voltage further. However, it is observed that write operation is a critical operation at low voltage and it requires a tremendous margin for successful write operation. Even more, static write failures happen at low voltage and it is a critical factor to limit lowering the supply voltage.

In this work, we propose an adaptive WWL pulse width and voltage modulation scheme to lower the supply voltage while maintaining yield and mitigating performance degradation. By adaptively modulating WWL pulse width, excessive margin is minimized. In addition to this, adaptive WWL pulse voltage level modulation fixes static write failures while preventing half select disturb. The adaptive modulation is possible by monitoring write completion using the decoupled read path of the 8-T bitcell. WWL pulse is on until to-be-written data and read-out date become the same. If write operation is not completed in the first cycle, the voltage level of WWL increases to fasten write operation and to fix static write failures. However, increased WWL pulse voltage level can generate half select disturb. To prevent half select disturb, internal data in half selected bitcells are read during the first cycle and WBL and WBLB are set to certain values according to read-out data in the first cycle. By doing this, the chance of experiencing half select disturb will be minimized. Overall, the adaptive WWL pulse width and voltage modulation scheme lower the $V_{min}$ to lower power consumption while minimizing performance degradation and maximizing yield. In this work, 65nm CMOS technology is used.

## 4.2 Background and Related Works

### 4.2.1 Variation at Low Voltage

Variation increases as the supply voltage decreases. Figure 4.1 depicts variation as supply scales. The distribution of FO4 delay is measured using 100,000 Monte Carlo simulations. At low voltage, performance degrades by larger variation as well as smaller $I_{on}$. The performance degradation by larger variation limits lowering the supply voltage so variation compensation techniques are required for low voltage operation.

### 4.2.2 8-T SRAM Operations at Low Voltage

As already discussed in Chapter II, the 8-T SRAM bitcell is a good candidate as a SRAM bitcell at low voltage since write operation and read operation can be separately optimized. Between read and write, write operation is a critical operation at low voltage because it has more variation. Figure 4.2 shows 40,000 Monte Carlo simulation results and it clearly depicts that write operation is more vulnerable to variation. Also, there are five write failures out of 40,000 at 0.5V while there is no read failure. Because of the five write failure at 0.5V, it is not possible to lower the supply voltage down to 0.5V.

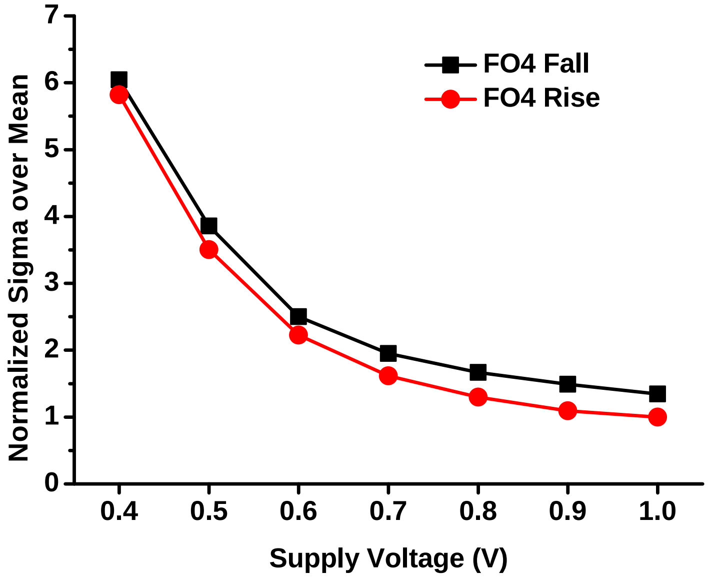


Figure 4.1: More variation exist at lower voltage

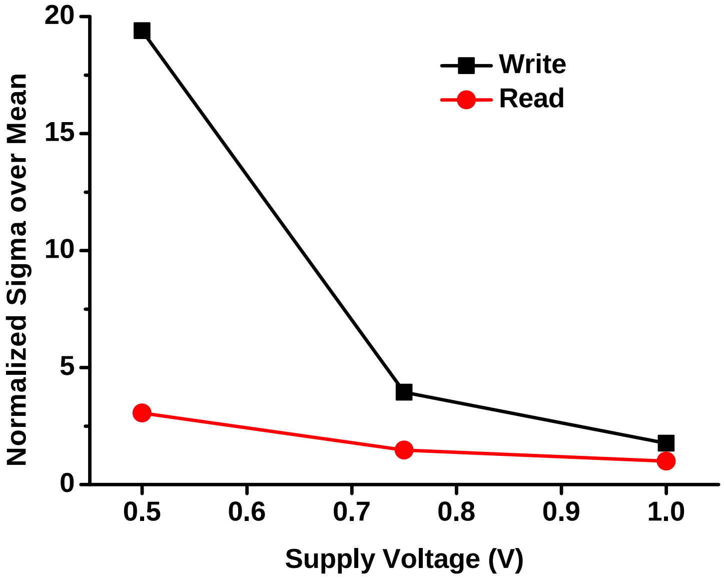


Figure 4.2: Write operation is a critical operation at low voltage because it is more vulnerable to variation.

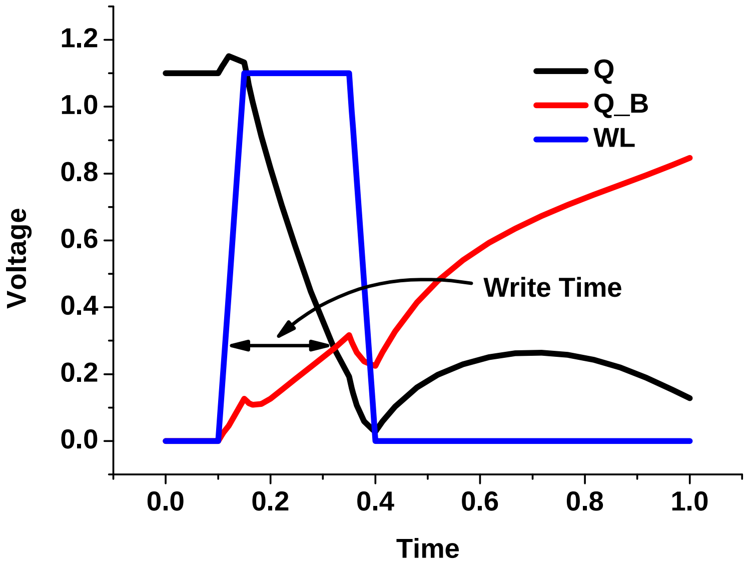


Figure 4.3: SRAM write time is a time difference from WWL on to two internal nodes crossing.

### 4.2.3 8-T SRAM Write Time

To analyze the write operation of 8-T SRAM at low voltage, write time is simulated using Monte Carlo simulation. Figure 4.3 describes the definition of write time used in this work. WWL is turned on to start write opearation. After some time, two internal nodes in an SRAM bitcell are crossed each other. Write time is defined as a time between WWL on and two internal nodes crossing. For successful write operation, WWL pulse width must be larger than this write time.

The Monte Carlo simulation results with 100,000 iterations of write time as supply scales are shown in Figure 4.4. At 1.0V, the worst case write time is ∼2.2× larger than typical. If the WWL pulse width is 2.2× larger than typical write time, high yield is expected. However, the required margin at low voltage is much larger than nominal voltage. At 0.65V, at least 58× margin is required for the successful write of all 100,000 iterations. Below 0.65V, the worse case is static write failure: write operation cannot be done even with infinitely long WWL pulse. Based on this simulation, the

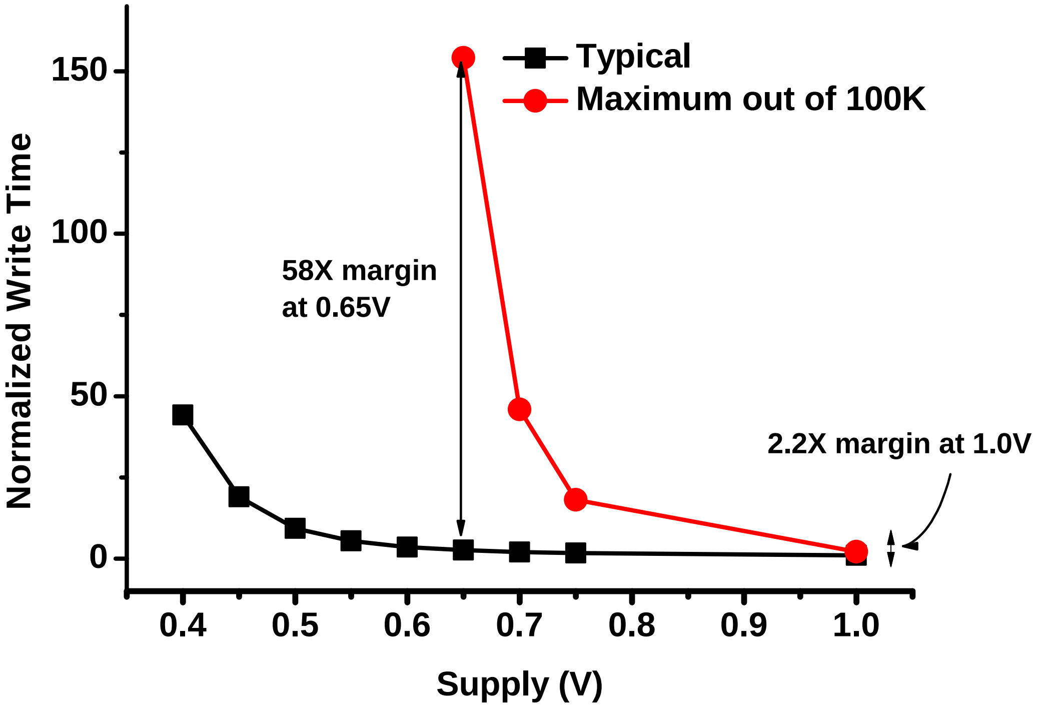


Figure 4.4: Write time degrades as supply scales. However, the degradation of the worst case write time is much worse that typical cases. Below 0.65V, write failure happens (static write failure).

$V_{min}$ of this SRAM bitcell is 0.65V because of static write failures under 0.65V.

In Table 4.1, the number of static failures and the estimated yield of a 16Kb array are summarized as the supply voltage decreases. As already shown in Figure 4.4, the first static failure happen at 0.6V and the number of static failures increases as supply scales. The estimated yield is calculated based on the equation below:

$$\text{Yield} = \left(1 - \frac{\text{the number of static failures}}{\text{the total number of iterations}}\right)^{(\text{the total number of bitcells})} \tag{4.1}$$

when WWL pulse width is longer than the worst case write time.

In terms of dynamic write failure (Chapter II), the yield is compromised as WWL pulse width decreases. Figure 4.5 shows the write time distributions of two supply

| Supply Voltage (V) | Static Failures out of 100K | Estimated Yield of 16Kb Array |
|---|---|---|
| 1.0 | 0 | >84.9% |
| 0.75 | 0 | >84.9% |
| 0.7 | 0 | >84.9% |
| 0.65 | 0 | >84.9% |
| 0.6 | 1 | 84.9% |
| 0.55 | 4 | 51.9% |
| 0.5 | 12 | 14.0% |
| 0.45 | 32 | 0.5% |
| 0.4 | 51 | <0.01% |

Table 4.1: Static write failures happen at below 0.65V and estimated yield compromised

| WWL Pulse Width | Dynamic Failures out of 100K | Estimated Yield of 16Kb Array |
|---|---|---|
| >58× larger than typical | 0 | >84.9% |
| >20× larger than typical | 1 | 84.9% |
| >15× larger than typical | 2 | 72.1% |
| >10× larger than typical | 10 | 19.4% |
| >5× larger than typical | 101 | <0.01% |

Table 4.2: Dynamic write failures increases as WWL pulse width decreases at 0.65V.

voltages: 1.0V and 0.65V. The distribution at 1.0V is much narrower than 0.65V because the worst case write time is just 2.2× larger than typical. At 1.0V, if WWL pulse width is 2.2× larger than typical, more than 84.9% of yield in 16Kb array is expected. To achieve the same yield as 1.0V, at least 58× larger WWL pulse width than typical is required at 0.65V. Table 4.2 summarizes the number of dynamic write failures and the estimated yield of 16Kb array at 0.65V as WWL pulse width decreases for higher performance. At low voltage, the larger distribution of write time requires much longer WWL pulse to maintain the decent yield of an SRAM array and it results in lower performance.

However, not a whole array (all bits) but a word (8, 16, or 32 bits) is written during one write operation. This implies that the worst case WWL pulse width of "the whole array" is not necessary for each write operation but the worst case WWL

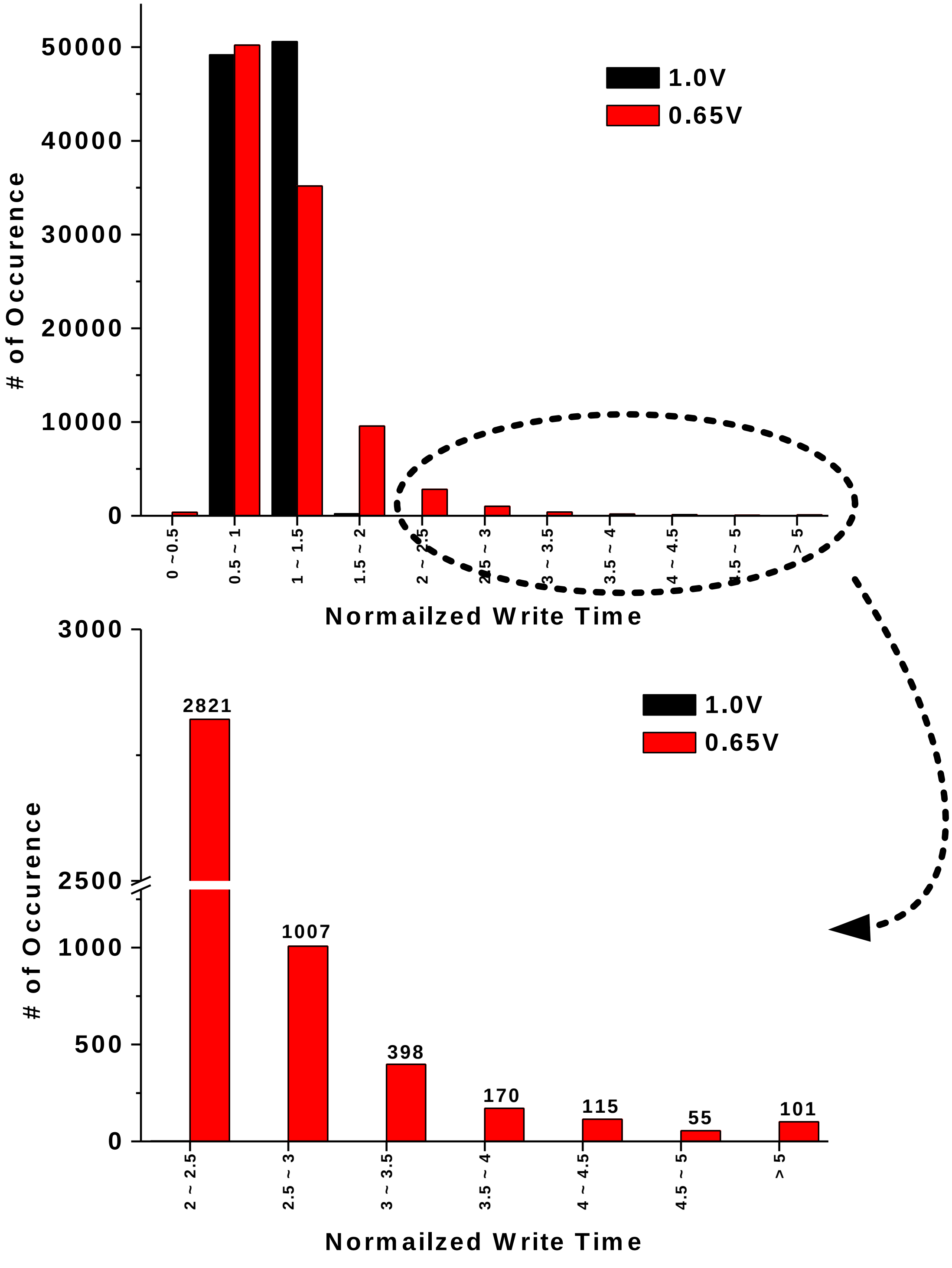


Figure 4.5: Write time distribution at 0.65V and 1.0V are shown. At 0.65V, the distribution is much wider than 1.0V.

pulse width of “a word” is only required for successful write at a time. Moreover, write time distribution is a long tail distribution at low voltage (Figure 4.5) and 99.9% of write time is smaller than 5× typical. This also implies that the worst case margining is only for the small number of cases.

### 4.2.4 8-T SRAM Static Failures and Half Select Disturb

If the supply voltage decreases down to 0.4V, there are 51 static write failures out of 100,000 Monte Carlo simulations (Table 4.1). To fix these static write failures, WWL boosting is applied; WWL pulse voltage level increases to 0.45V. Figure 4.6 shows the effectiveness of boosting. All static write failures are fixed and the distribution becomes narrower than without boosting.

However, it is more likely that half select disturb happens with WWL boosting (Chapter II). Half select distrub is simulated assuming that WBL and WBLB are tied to $V_{dd}$. Figure 4.7 shows the results of 100,000 Monte Carlo simulations. Figure 4.7 clearly shows that half select disturbs happen when WWL pulse width is modulated for successful write operations. Therefore, a way to mitigate half select disturb must be introduced when WWL boost is used.

### 4.2.5 Previous Works

This proposed work detects the worst case WWL pulse width of a word during write operation and makes in-situ WWL pulse modulation available. There have been works [24, 53, 57] which adaptively control word line pulse to enhance yield. Reference [24] proposed a 14-T SRAM bitcell with write completion detection scheme. However, the purpose of write completion detection scheme is not because large variation but rather because extremely slow write operation due to HVT devices. Reference [53, 57] are for compensating systematic (global) variation only with WWL voltage level while this proposed work also compensates random(local) variation as well with WWL pulse

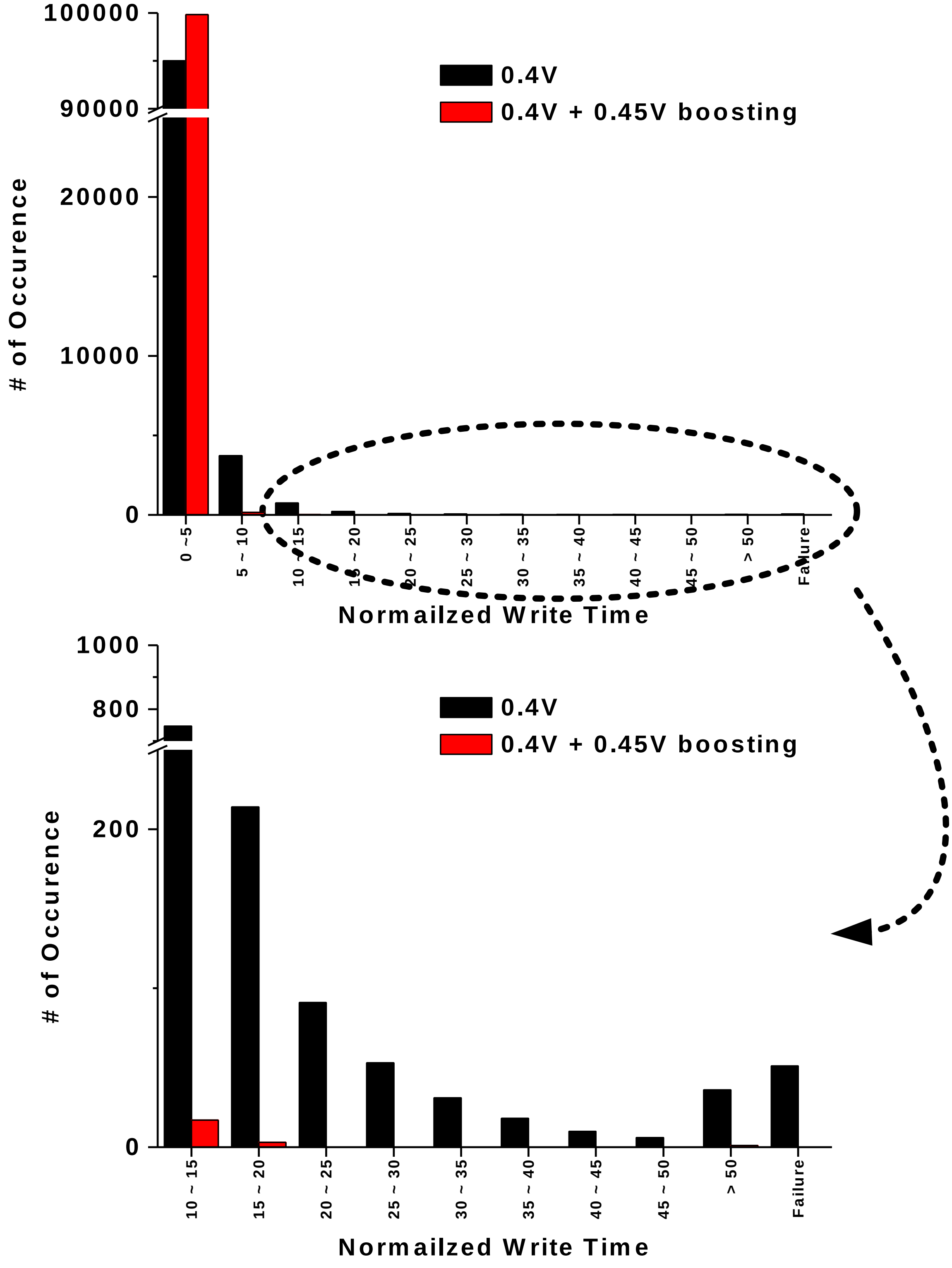


Figure 4.6: Write time distribution at 0.4V with and without boosting are shown. 0.45V of WWL boosting removes all static write failures and makes the distribution narrower than without boosting.

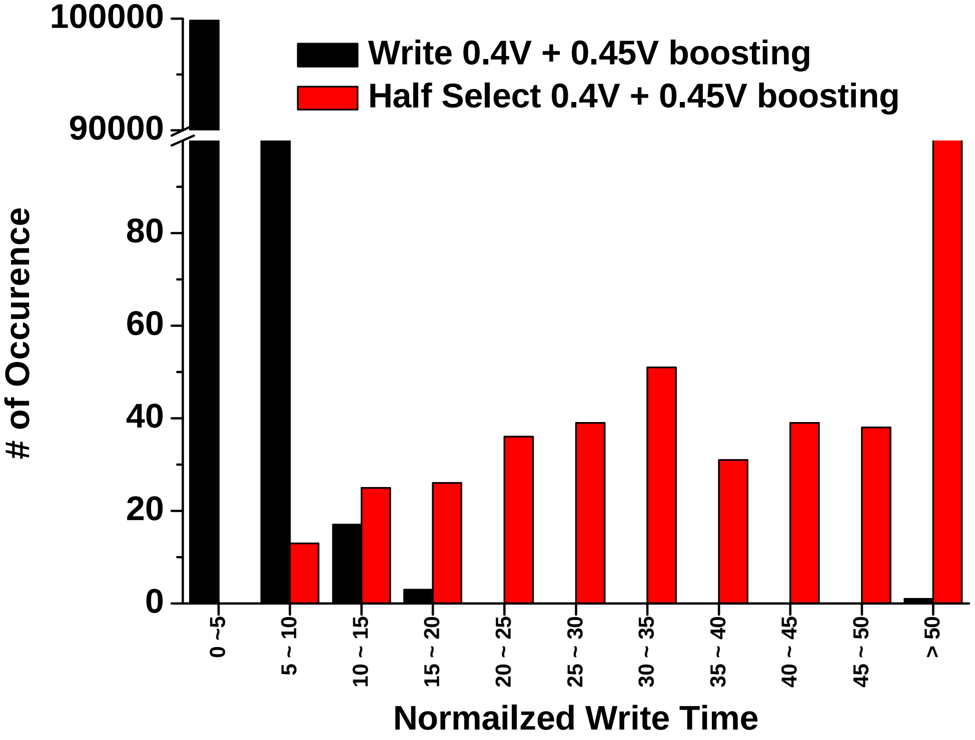


Figure 4.7: Half select disturbs happen with WWL boosting.

width and voltage level.

## 4.3 Design of Adaptive Write Architecture

### 4.3.1 Write Word-Line Pulse Width Modulation

Instead of a long single cycle write operation with excessive margin, a short multiple cycles write operation is proposed. The long single cycle write operation has a fixed WWL pulse width and, therefore, the pulse width must be long enough to perform successful write operation for the worst case bitcell. In contrast, the short multiple cycle write operation adaptively modulates its pulse width until write operation is completed by increasing the number of cycles. In this case, it is important to determine when the write operation is completed. Because the 8-T bitcell has separate read and write paths, it is possible to read the data while the write operation is on-going. By comparing read out data at Read Bit-Line (RBL) and to-be-written data at WBL and WBLB, it is possible to determine whether write operation is completed or not. Similar write completion scheme is used in [24] but they do not use 8-T bitcells.

Figure 4.8 depicts timing diagram for write completion when write operation requires two cycles for successful write. In each cycle, there are two phases: read and evaluation. Read operation is performed during read phase while read out data and to-be-written data are bit-wisely compared during evaluation phase. Write operation is concurrently running during read and evaluation phase. If write operation is not completed yet after evaluation phase, WWL pulse is extended and another write cycle is executed. If write completion is detected, WRITE_DONE signal is turned on and in turn WWL is turned off. The schematic for write completion is shown in Figure 4.9. Because this scheme allows in-situ WWL pulse width modulation to compensate random variation as well as Process, Voltage, and Temperature (PVT)

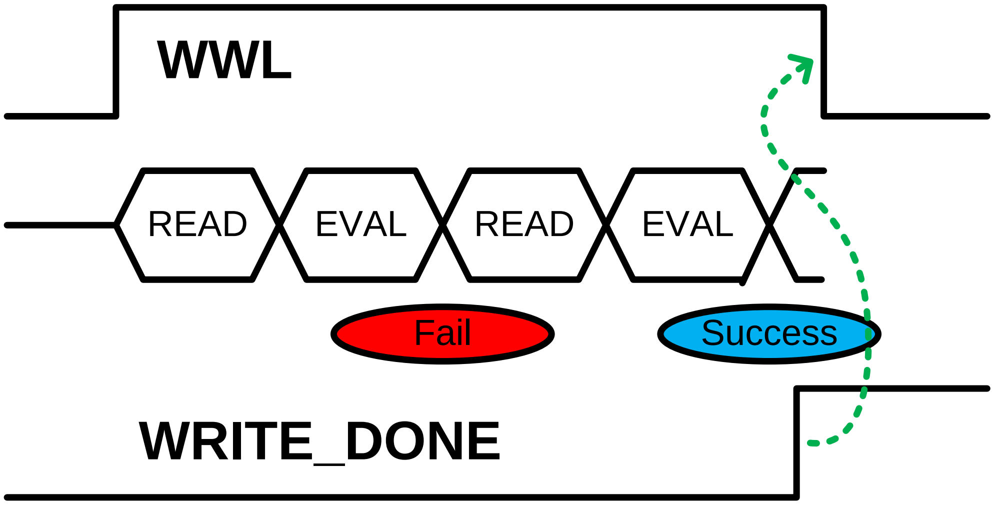


Figure 4.8: High level timing diagram of adaptive write architecture

variation, we call it “Write Assurance”.

### 4.3.2 Write Word-Line Voltage Level Modulation

The write assurance concept is effective to modulate WWL pulse width exactly required for write operation but it cannot solve static write failures. One of the most common ways to mitigate static write failures is boosting WWL voltage. However, as we discussed in the previous section, WWL boosting makes half select disturb immunity worse. An adaptive WWL voltage level modulation is proposed to solve static write failures while mitigating half select disturb.

Table 4.3 summarizes write and read operations for full selected bitcell and half selected bitcell when the adaptive WWL voltage modulation is adopted. In the first cycle, WWL voltage is the same as the supply voltage and, therefore, normal write and normal read are executed for the full selected bitcell. Meanwhile, half selected write is executed for the half selected bitcell but the possibility of half select disturb within the first cycle is very low because the WWL pulse width is not long enough yet to generate half select disturb. If write operation is completed within the first cycle, WWL is turned off and there will be no second cycle. Otherwise, there are

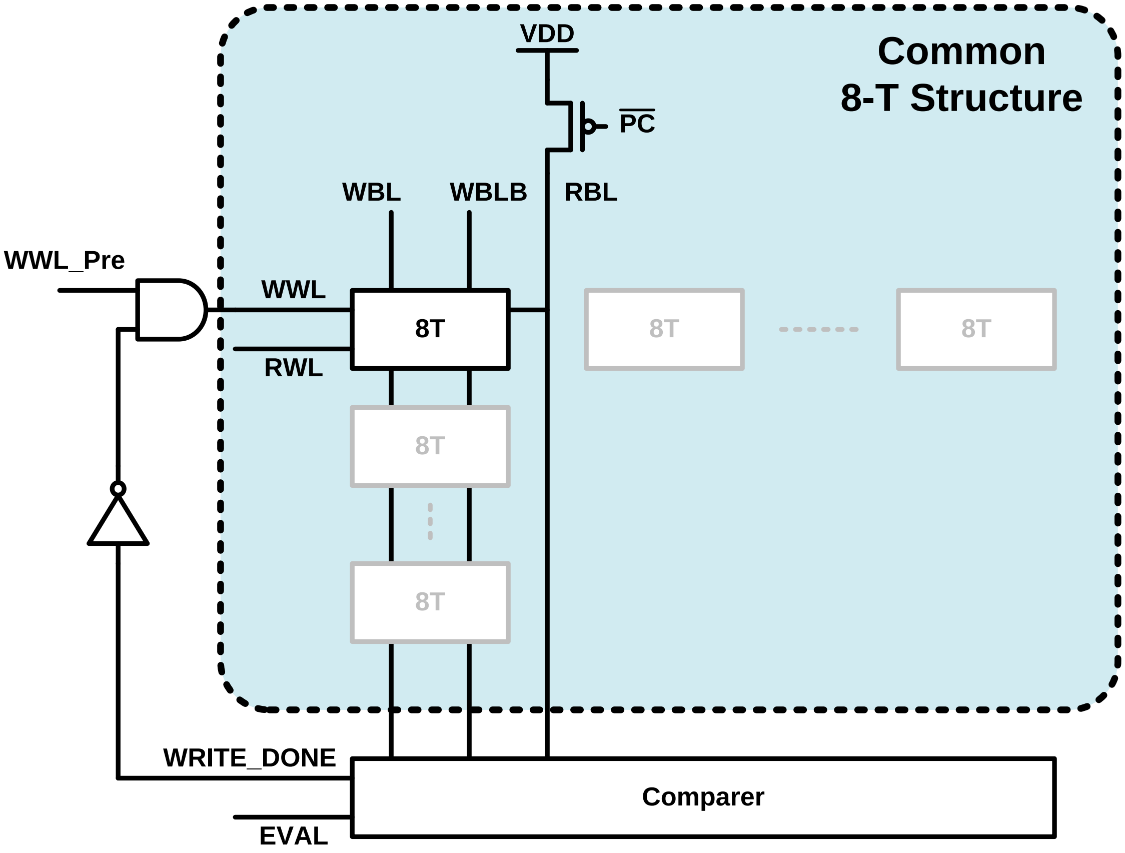


Figure 4.9: High level block diagram of adaptive write architecture

| Cycle | WWL Voltage | Target Bitcells | Half Selected Bitcells |
|---|---|---|---|
| First Cycle | Normal | Normal Write<br>Normal Read | Half Selected Write<br>Normal Read |
| After the First Cycle | Boosted | Boosted Write<br>Normal Read | No Write<br>Normal Redundant Read |

Table 4.3: Adaptive Voltage Level Modulation

dynamic write failures or static write failures and, therefore, more cycles are necessary for write completion. After the first cycle, WWL voltage is boosted to solve dynamic failures and static failures at the same time. However, WWL boosting makes half select disturb immunity worse. To prevent half select disturb after the first cycle, WBL and WBLB of half selected bitcell are driven by read-out data obtained during the first cycle. By applying this technique, half select disturb is prevented even with boosted WWL.

## 4.4 Measurement Results

### 4.4.1 Prototype Implementation

A prototype chip was fabricated in a 65nm CMOS process. Figure 4.10 describes a block diagram of the prototype chip. It basically consists of two types of sub-modules: control modules and a Design Under Test (DUT) module. There are three sub-modules to control and test the DUT module. A Built In Self Test (BIST) module is basically a processor and it executes SRAM write and read operations. An Instruction Memory (IMEM) stores instructions run by the BIST and it is a baseline 8-T SRAM bitcell array which does not have special features such as WWL pulse width modulation. A scan module is a scan chain that communicates to the outside of the chip. These control modules run at the nominal voltage (1.0V). A Data Memory (DMEM) is the DUT module with WWL pulse width and voltage level modulation scheme. To execute voltage level modulation discussed in subsection 4.3.2,

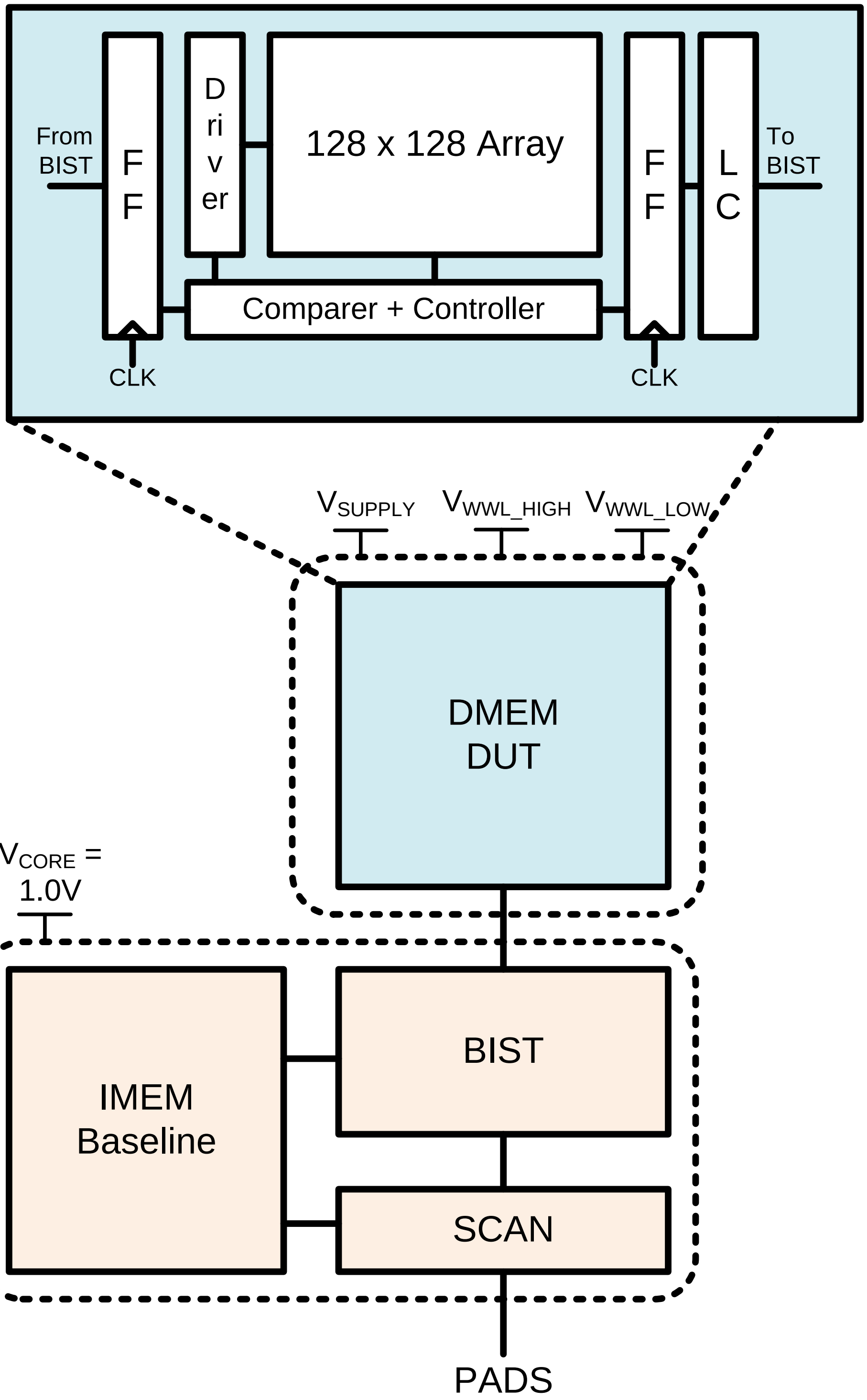


Figure 4.10: A block diagram of prototype implementation. BIST, IMEM, and SCAN are control modules and DMEM is the DUT module.

there are two dedicated power supplies for WWL: $V_{WWL\_HIGH}$ and $V_{WWL\_LOW}$. Also, a dedicated $V_{SUPPLY}$ is connected to DMEM to find the $V_{min}$ of DMEM. Because the voltage levels of DMEM and BIST can be different from each other, level converters are used as output buffers in DMEM.

An 128×128 bitcell array in DMEM has a cascaded bit-line structure (Figure 4.11) and bit-interleaved bitcells (Figure 4.12). For a better performance and functionality, a bit-line has the cascaded structure with several local blocks. Eight bitcells, a precharger, a keeper, and a tri-state buffer construct a local block. To build an 128-bit tall global bit-line, 16 local blocks are connected. The 128×128 bitcell array is bit-interleaved with four 32-bit words.

Figure 4.13 is a micrograph of the prototype chip with dimension.

### 4.4.2 Measurement Data

A shmoo plot to find the $V_{min}$ of an SRAM array is shown in Figure 4.14 when $V_{SUPPLY}$, $V_{WWL\_HIGH}$, and $V_{WWL\_LOW}$ are the same. We can observe the double-sided constraints on frequency by write/read and half select disturb. The $V_{min}$ is a cross point of write/read and half select disturb and it is 0.775V. Below 0.775V, half select disturb is critical; write and read are still functional. The $V_{min}$ of read and write operation is 0.5V.

By lowering $V_{WWL\_HIGH}$ and $V_{WWL\_LOW}$ voltage levels at the same time (WWL underdrive), half select disturb can be mitigated (Figure 4.15). Down to 0.6V, half select disturb is fixed with WWL underdrive while write operation is still functional. In the 8-T bitcell used in this work, the sizes of access transistor and pull-down transistor are the same so write operation is relatively strong compared with a traditional 6-T bitcell. However, below 0.6V, write failures happen when WWL is low enough to fix half select disturb. In other words, we cannot find an appropriate WWL voltage level with no write failure and no half select disturb.

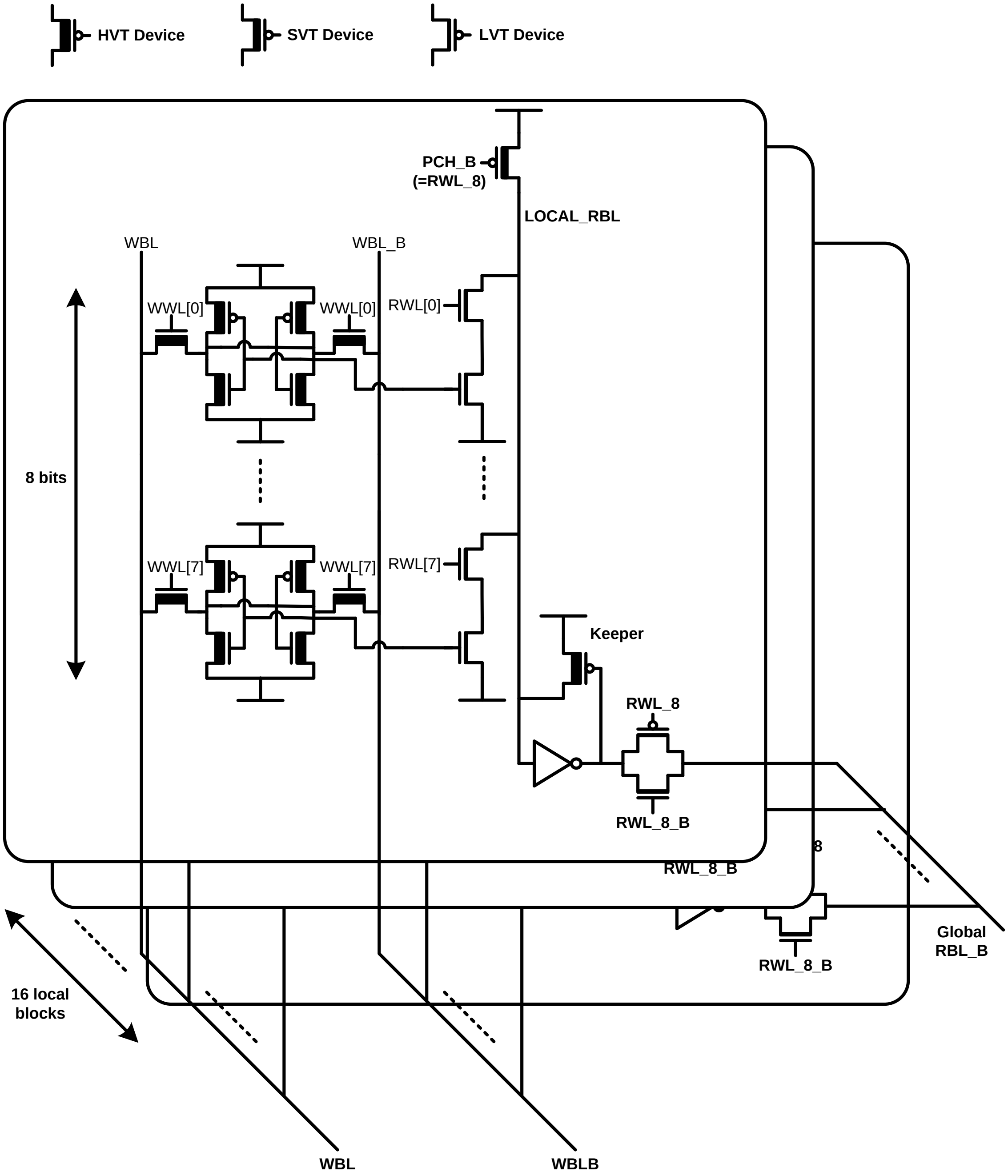


Figure 4.11: A cascaded bit-line structure for DMEM is shown. Eight bitcells, a precharger, a keeper, and a tri-state buffer construct a local block. 16 local blocks are connected; 128 bits tall structure is built.

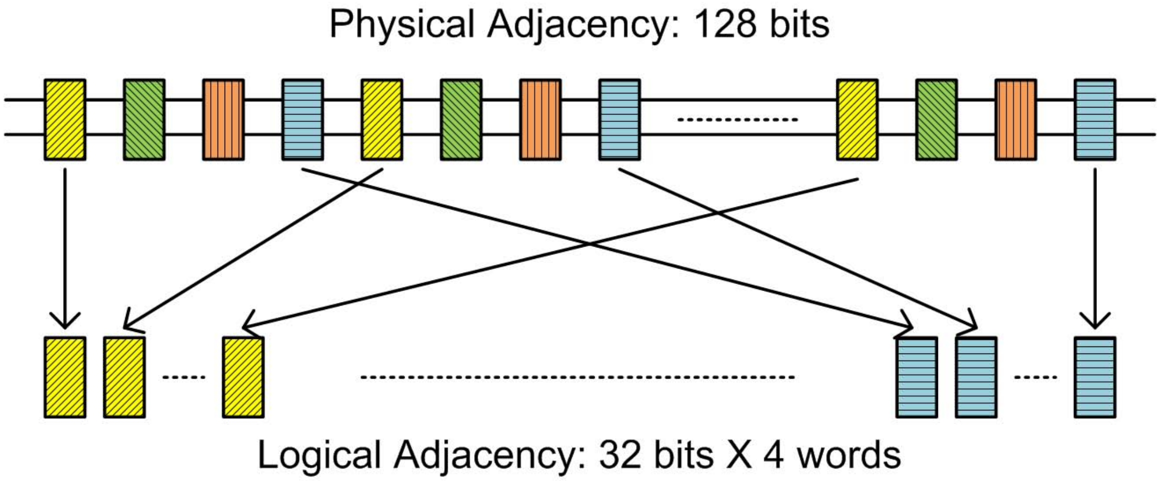

Physical Adjacency: 128 bits
Logical Adjacency: 32 bits X 4 words

Figure 4.12: Bit-interleaved bitcells are shown. There are four words and each word has 32 bits.

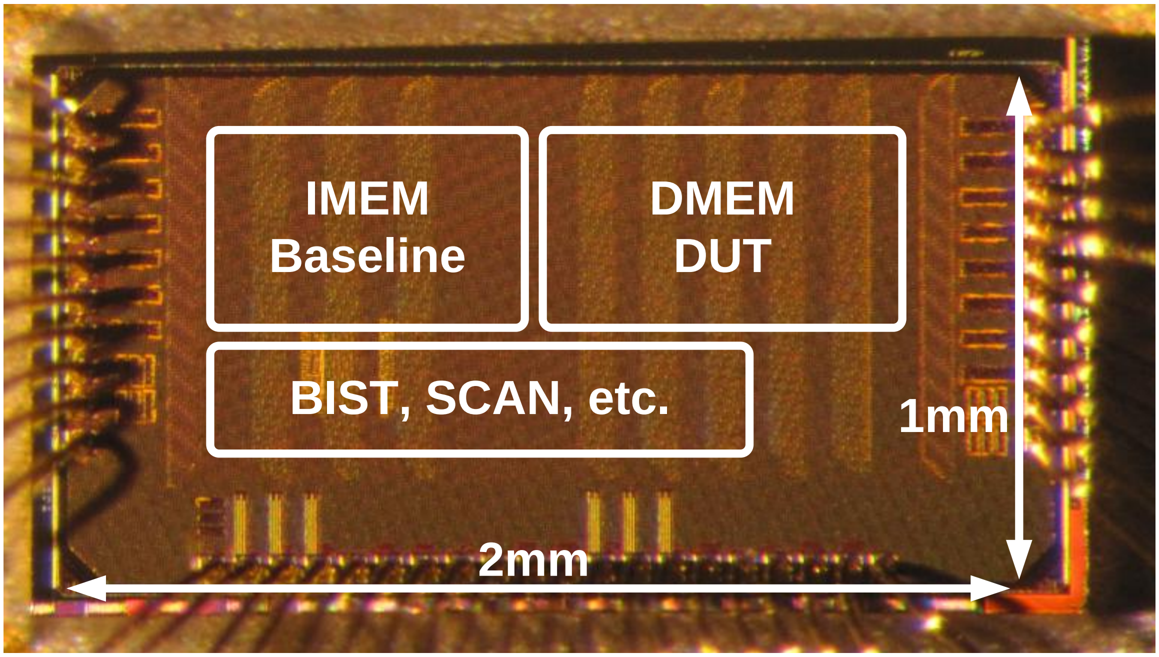


Figure 4.13: A chip micrograph is shown with dimension.

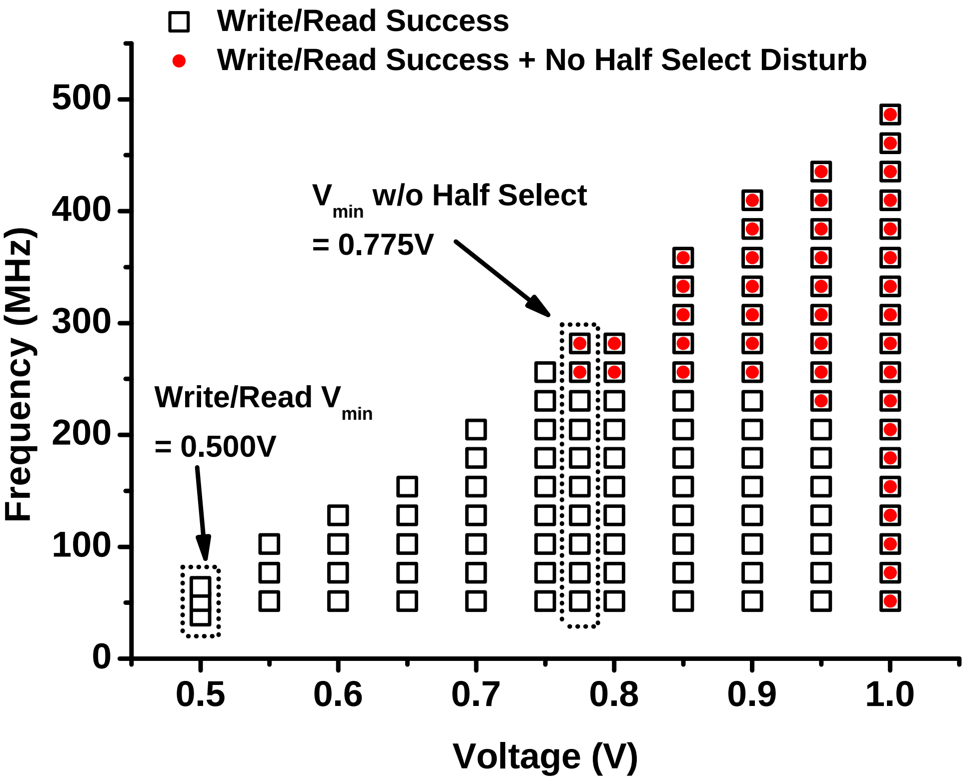


Figure 4.14: A shmoo plot to find the $V_{min}$ with normal write/read operation is shown.

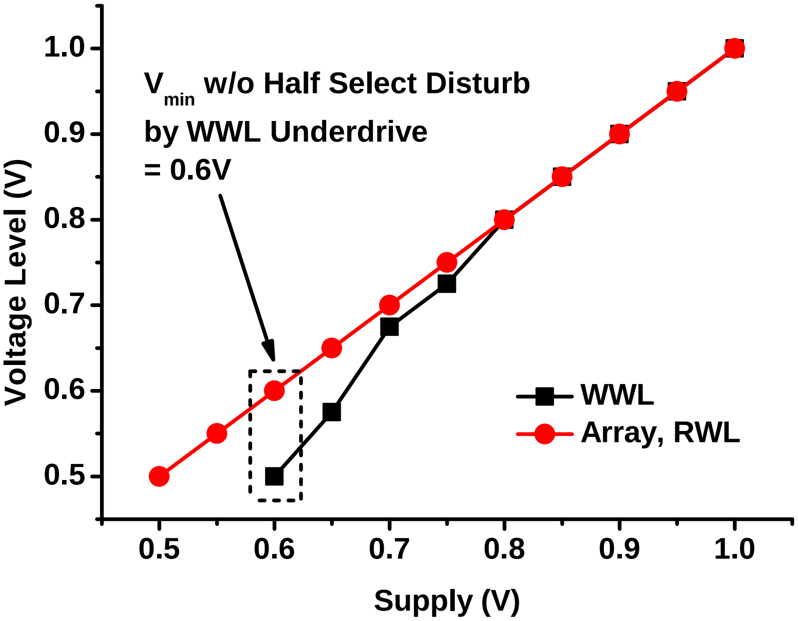


Figure 4.15: The voltage level of WWL decreases to fix half select disturb.

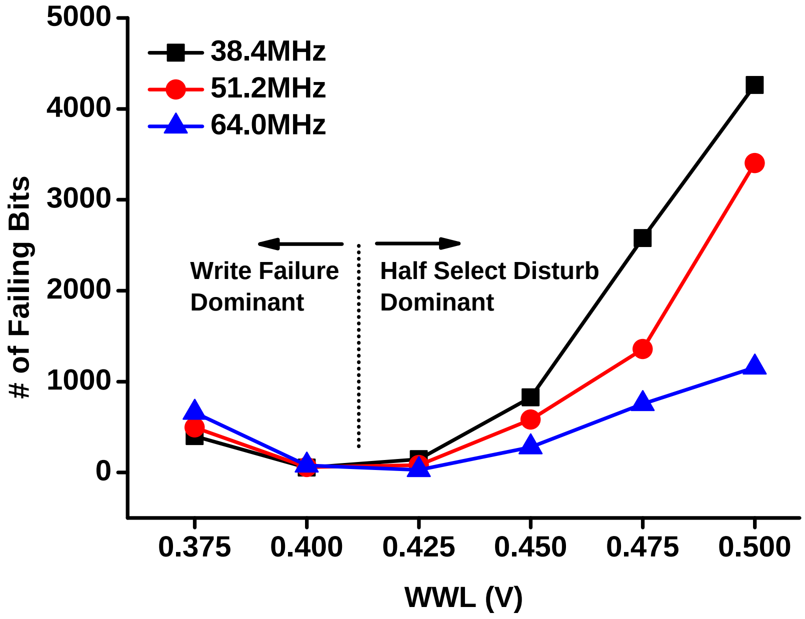


Figure 4.16: The number of failing bits at 500mV with WWL underdrive is shown. As the voltage level of WWL decreases, immunity to half select disturb is enhanced while writability becomes worse.

Figure 4.16 analyze failures at 0.5V ( $V_{SUPPLY}$ = 0.5V) as the voltage level of WWL (both $V_{WWL\_HIGH}$ and $V_{WWL\_LOW}$) decreases. Initially, half select disturb is dominant. Hence, at higher frequency, there are fewer half select when WWL is 0.5V. As the voltage level of WWL decreases, half select disturb is mitigated but write failures becomes worse. If WWL is lower than 0.425V, write failure start dominating failures. We can observe that there are more failures with higher frequency when WWL is 0.375V. These results show the opposite characteristic between write failure and half select disturb on WWL pulse width.

To fix half select disturb and read/write failures at the same time, adaptive voltage level modulation scheme is applied (Figure 4.17) at 0.5V ( $V_{SUPPLY}$ = 0.5V). If $V_{WWL\_LOW}$) is not low enough, half select can be happen even before read out the data of half selected bitcells at the first cycle of adaptive voltage level modulation. No half select disturb is observed when $V_{WWL\_LOW}$ is lower than or equal to 390mV. Also, $V_{WWL\_HIGH}$ must be high enough to fix half select disturb and to write data to

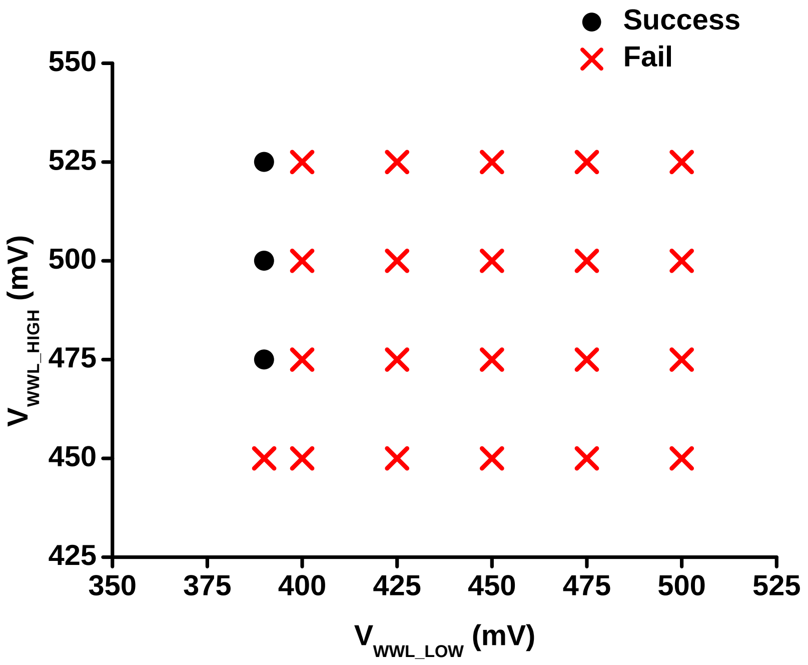


Figure 4.17: $V_{WWL\_HIGH}$ and $V_{WWL\_LOW}$ are separately optimized when $V_{WWL\_HIGH}$ is 0.5V.

| Cases | $V_{SUPPLY}$ | $V_{WWL\_HIGH}$ | $V_{WWL\_LOW}$ |
|---|---|---|---|
| Normal Voltage Scaling | 775mV | 775mV | 775mV |
| WWL Underdrive | 600mV | 500mV | 500mV |
| WWL Voltage Level Modulation | 500mV | 500mV | 390mV |

Table 4.4: Supplies for each $V_{min}$ cases.

not-yet-written bitcells at the second cycle. When $V_{WWL\_HIGH}$ is 475mV, all failures are fixed. Based on the measured results from the prototype implementation, we can simplify the system by only underdriving $V_{WWL\_LOW}$ while keeping $V_{SUPPLY}$ and $V_{WWL\_HIGH}$ the same to maximize writability and half select disturb immunity.

The $V_{min}$ cases shown above are summarized in Table 4.4. The $V_{min}$ with normal voltage scaling is 775mV and it can be lowered down to 600mV using WWL underdrive. Using voltage level modulation scheme, the $V_{min}$ is lowered down to 500mV with 390mV of $V_{WWL\_LOW}$ underdrive.

The measured power consumption results are summarized in Table 4.5. WWL pulse width and voltage modulation scheme allows lower the $V_{min}$ and 3.96× leakage

| Cases | Frequency | $P_{LEAKAGE}$ | $P_{ACTIVE}$ |
|---|---|---|---|
| Normal Voltage Scaling | 256MHz | 317.8$\mu$W | 705.3$\mu$W |
| WWL Underdrive | 128MHz | 129.3$\mu$W | 272.3$\mu$W |
| WWL Voltage Level Modulation | 64MHz | 80.2$\mu$W | 166.5$\mu$W |

Table 4.5: Power measurement results are shown. Each frequency is selected for lower power consumption while avoiding any failures.

power reduction and 4.24× active power reduction are achieved.

## 4.5 Conclusion

The adaptive WWL pulse width and voltage level modulation scheme is proposed. SRAM is more vulnerable to variation at low voltage so write time distribution has a long tail and there can be write failures. By monitoring write completion, write failures are fixed using WWL voltage level modulation while half select disturb is mitigated. A prototype chip fabricated in 65nm CMOS process shows 3.96× leakage power reduction and 4.24× active power reduction with the lowered $V_{min}$.

# CHAPTER V

# Low Power Circuit Design Based on Heterojunction Tunneling Transistors

The theoretical lower limit of subthreshold swing in MOSFET (60 mV/decade) significantly restricts low voltage operation since it results in a low $I_{on}$ to $I_{off}$ ratio at low supply voltages. This chapter investigates extremely-low power circuits based on a new Si/SiGe HETT that has subthreshold swing $<$ 60 mV/decade. Device characteristics as determined through Technology Computer Aided Design (T-CAD) tools are used to develop a Verilog-A device model to simulate and evaluate a range of HETT-based circuits. We show that a HETT-based Ring Oscillator (RO) shows a 9-19× reduction in dynamic power compared to a CMOS RO. We also explore two key differences between HETTs and traditional MOSFETs, namely asymmetric current flow and increased Miller capacitance, analyzing their effect on circuit behavior and proposing methods to address them. Finally, HETT characteristics have the most dramatic impact on SRAM operation and hence we propose a novel 7-transistor HETT-based SRAM cell topology to overcome, and take advantage of, the asymmetric current flow. This new HETT SRAM design achieves 7-37× reduction in leakage power compared to CMOS.

## 5.1 Introduction

Low voltage operation is one of the most effective low power design techniques due to its quadratic dynamic energy savings. Recently, a number of works [7, 24, 30, 54] have shown aggressive supply voltage reduction to near or below the $V_{th}$ of MOSFET devices with considerable reduction in power consumption. However, this power improvement has come at the cost of operation speed (typically $<$ 10 MHz). At such low supply voltages, $I_{on}$ drops dramatically due to lack of gate overdrive resulting in large signal transition delays. To regain this performance loss it is possible to reduce the threshold voltage. However, this exponentially increases $I_{off}$, which is particularly problematic in applications that spend significant time in standby mode [72]. For instance, lowering the supply voltage from 500mV to 250mV while enforcing iso-performance by reducing the $V_{th}$ increases leakage power by 275× in a commercial bulk-CMOS 45nm technology, which is unacceptable.

To address this dilemma, there has been recent interest in new devices with significantly steeper subthreshold slopes than traditional MOSFETs [14, 16, 26, 38, 56, 65]. A steep subthreshold slope enables operation with a much lower threshold voltage while maintaining low leakage. In turn, a low $V_{th}$ enables low voltage operation while maintaining performance. Hence, steep subthreshold slopes can provide power efficient operation without loss of performance.

In this paper, we investigate circuit design using the recently proposed Si/SiGe HETT [52]. The Si/SiGe heterostructure uses gate-controlled modulation of band-to-band tunneling to obtain subthreshold swings of less than 30 mV/decade with a large $I_{on}$ of 0.42mA/$\mu$m at $V_{ds}$ = 0.5V. Furthermore, Si/SiGe heterostructures are fully compatible with current MOSFET fabrication and can leverage the extensive prior investment in CMOS fabrication technology. Currently, several industry and university teams are actively developing Si/SiGe HETT type transistor structures, and initial devices have been experimentally demonstrated [40, 43].

We explore the key differences between HETTs and traditional MOSFETs that must be considered in the design of circuits using these new devices. Most significantly, HETTs display asymmetric conductance. In MOSFETs, the source and drain are interchangeable, with the distinction only determined by the voltages during operation. However, in HETTs, the source and drain are determined at the time of fabrication, and the current flow for $V_{ds} < 0$ is substantially less than for $V_{ds} > 0$ (in an N-channel HETT (NHETT)). Hence, HETTs can be thought to operate "uni-directionally", passing logic values only in one direction, which has significant implications on logic and especially SRAM design. Our analysis shows that another effect is a large increase in gate-to-drain capacitance (i.e., Miller capacitance) in HETTs compared to MOSFETs. This excess Miller capacitance can cause undesirable artifacts in the switching behavior of HETTs that is not present in MOSFETs. These differences in device operation and characteristics require careful study to understand their circuit design implications. In this paper, we show that HETT-based logic circuits are capable of improving energy efficiency by 19× compared to CMOS when operated at a supply voltage of 0.23V. We particularly study SRAM design which is most impacted by the novel characteristics of HETTs. We show that the unidirectional characteristic of HETTs can actually be exploited in SRAM design to enable a novel 7-T robust SRAM cell.

My main contributions in this work are HETT device modeling and HETT-based circuit analysis. I also partially contribute to HETT-based SRAM design.

## 5.2 Background and Related Works

### 5.2.1 Necessity for Steep Subthreshold Swing

To understand the motivation of developing devices with high subthreshold swing, it is necessary to understand two power consumption equations:

$$P_{dynamic} = \alpha \times C \times {V_{DD}}^2 \times f \tag{5.1}$$

where $\alpha$ is the activity factor, $C$ is the switched capacitance, $V_{DD}$ is the supply voltage, and $f$ is the frequency.

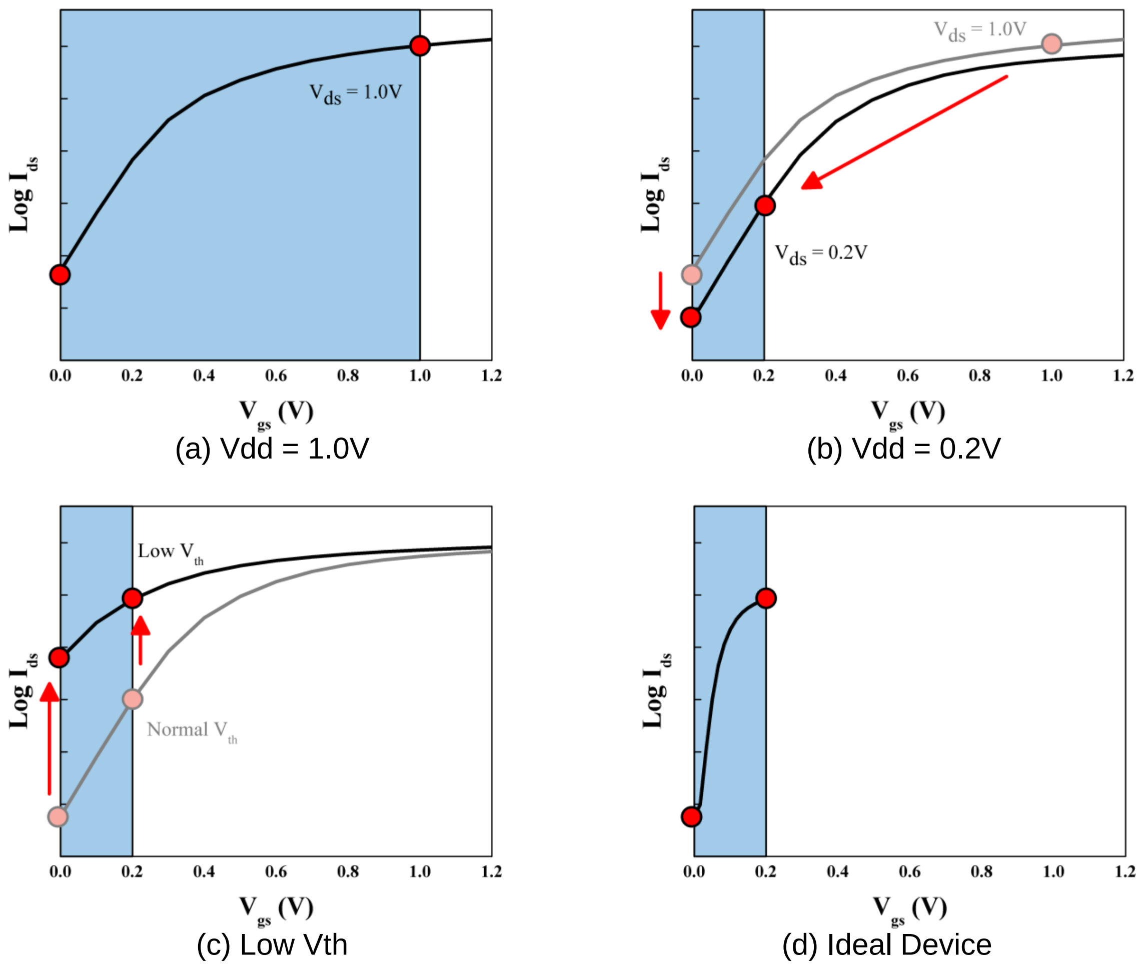
$$P_{leakage} = V_{DD} \times I_{leakage} \tag{5.2}$$

where $V_{DD}$ is the supply voltage and $I_{leakage}$ is the leakage current. For low power operations, it is important to reduce both dynamic power consumption ($P_{dynamic}$) and leakage power consumption ($P_{leakage}$).

The most effective way for power reduction is voltage scaling due to its quadratic dynamic power reduction (Equation 5.1). In Figure 5.1(a), the lower red dot and the upper red dot represent $I_{off}$ and $I_{on}$ respectively when $V_{dd}$ is 1.0V. If the supply voltage scales down to 0.2V (Figure 5.1(b)), both $V_{dd}$ and $I_{off}$ decreases and there is a huge power saving. However, performance degradation is unavoidable because of smaller $I_{on}$. In Figure 5.1(c), the threshold voltage increases for larger $I_{on}$ while keeping the same low $V_{dd}$. However, larger $I_{off}$ with low $V_{th}$ is inevitable due to the limited subthreshold swing in MOSFET and it will increase $P_{leakage}$. The subthreshold swing $S$ is following [55]:

$$S = n\left(\frac{kT}{q}\right)ln(10) \tag{5.3}$$

where $S$ is expressed in mV/decade, $n$ is an empirical parameter, with n $\geq$ 1, and $(kT/q)$ln(10) is 60mV/decade at room temperature. This equation 5.3 means the



(a) Vdd = 1.0V

(b) Vdd = 0.2V

(c) Low Vth

(d) Ideal Device

Figure 5.1: This figure explains the motivation on developing devices with steep subthreshold swing.

subthreshold leakage decreases by a factor of 10 with $V_{gs}$ drop of $S$. This implies that it is not possible obtain the characteristic of the ideal device (Figure 5.1(d)) with MOSFET.

In conclusion, for the ultimate power reduction, it is important to develop a device with a steep subthreshold swing to achieve larger $I_{on}$ at the relatively low voltage while having smaller $I_{off}$.

### 5.2.2 Devices with Steep Subthreshold Swing

There have been several works to achieve steep subthreshold swing. Si-based tunneling transistors are developed [14] but it suffers from the use of the homogeneous Si Structure which limits tunneling. As a different approach, carbon-nanotube tunnel transistors are also presented [1, 56]. However, the fabrication of carbon-nanotube is not mature enough. While these two approaches are based on tunneling effect, impact-ionization can be a different option [65]. The issue in impact-ionization transistor is that it requires quite large voltage for impact-ionization.

## 5.3 HETT Device Physics and Modeling

### 5.3.1 HETT Device Physics

The 60 mV/decade subthreshold slope limitation of conventional MOSFETs arises due to the thermionic nature of the turn-on mechanism. Tunneling transistors do not suffer from this fundamental limitation, since the turn-on in these devices is not governed by thermionic emission over a barrier.

Figure 5.2 illustrates the basic concept of tunneling transistor operation. In an n-type tunneling transistor, the source is doped p-type, the channel is undoped or lightly doped, and the drain is n-type. As shown in Figure 5.2, when the gate is biased positively the device is turned on because electrons in the valence band of the p-type source can tunnel into the conduction band of the channel. If the Fermi level in the source is less than a few thermal voltages ($kT$) below the valence band edge, the bandgap acts as an “energy filter”, precluding tunneling from the exponential portion of the Fermi-Dirac distribution. If the gate bias is reduced sufficiently so that the bottom of the conduction band in the channel rises above the top of the valence band in the source, the tunneling abruptly shuts off. Due to this filtering of the Fermi-Dirac distribution function by the bandgap, the subthreshold slopes can

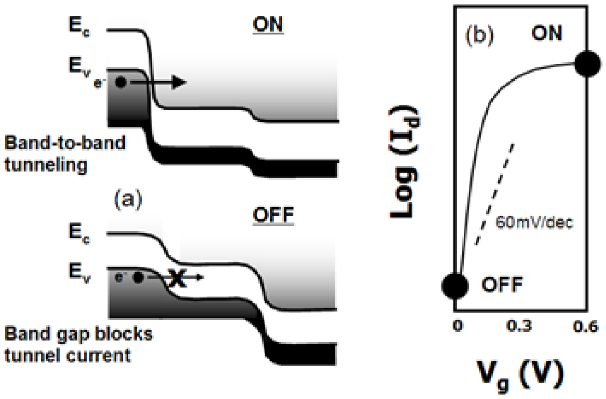


Figure 5.2: Tunneling FET device concept as depicted by a) band diagrams in the source-to-drain direction, and b) qualitative current-voltage characteristics

be significantly less than 60 mV/decade.

A potential problem with tunneling transistors is that a very narrow bandgap semiconductor must be used to obtain sufficiently high $I_{off}$. However, narrow bandgap materials also lead to higher $I_{off}$, and are often incompatible with standard CMOS processing. To avoid this problem, a type-II HETT can instead be employed. In such a case, the source-to-body contact has a staggered band lineup that creates an effective tunneling band gap, $E_{geff}$, which is smaller than that of the constituent materials. Such a band structure can also be realized in the Si/SiGe heterostructure material system, and complementary N- and P-HETTs can be fabricated, making this technology fully CMOS compatible. Figure 5.3 shows a schematic diagram of a complementary Si/SiGe HETT technology.

For the circuit simulations in this work, an optimized device structure was used. The simulated HETT devices have a gate length of 40 nm, and a high-k gate dielectric with effective gate oxide thickness of 1.2 nm. For NHETT, the source consists of pure Ge, with 3% biaxial compressive strain, and Si channel with 1% biaxial tensile strain. The complementary P-channel HETT (PHETT) design includes a strained Si source and pure Ge channel. Using band offsets from [58], the effective bandgap for this

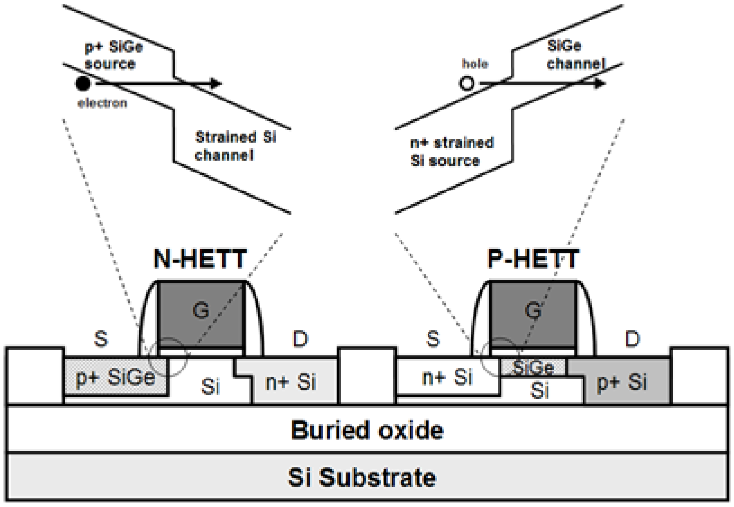


Figure 5.3: CMOS-compatible implementation of complementary tunneling FETs with type-II source-to-body hetero-junctions to improve device drive current

structure is 0.22 eV. For the transport calculations, a non-local tunneling model [27] with a 2-band dispersion relationship within the gap was used. Effective masses are $0.17m_0$ near the conduction band and $0.105m_0$ near the valence band in the silicon channel, and $0.10m_0$ near the conduction band and $0.055m_0$ near the valence band in the pure Ge source [20]. The device has a 2nm gate overlap of the source and an abrupt source doping profile. A gate work function of ∼4.4eV is used to set the $I_{off}$ to <1pA/$\mu$m.

### 5.3.2 HETT Device Modeling

Since accurate analytical models for HETTs are not available, we first built a look-up table based model using Verilog-A to enable circuit simulations. This technique is a simple and accurate way of compact modeling for emerging devices [41] where analytical expressions for the I-V characteristics are not well established.

A look-up table model is built for I-V and C-V characteristics using T-CAD simulation data based on the device parameters described in the above section. The HETT is modeled as a three-terminal device (source, gate, and drain) and current is assumed to flow only between source and drain since gate leakage is negligible with high-$\kappa$ gate dielectrics. Two parasitic capacitors are modeled; $C_{gd}$ and $C_{gs}$, which include inner fringing capacitance and overlap capacitance between gate and drain and between gate and source, respectively. Channel capacitance is negligible because the device has a fully-depleted channel and junction capacitance is also negligible due to its SOI-type substrate. As a result, we build three two-dimensional tables that are functions of two input voltages, $V_{gs}$ and $V_{ds}$, for modeling HETTs: $I_{ds}$ ($V_{gs}$, $V_{ds}$), $C_{gd}$ ($V_{gs}$, $V_{ds}$), and $C_{gs}$ ($V_{gs}$, $V_{ds}$). $V_{gs}$ and $V_{ds}$ are swept in 50mV steps in general, however in the slightly reverse biased region (-0.2V $< V_{ds} <$ 0V) where $I_{ds}$ transition is rapid $V_{ds}$ steps are 10mV for the $I_{ds}$ tables.

Based on the three tables stored at Comma-Separated Values (CSV) files, NHETT and PHETT are modeled using Verilog-A. The source code below shows a sample Verilog-A code for NHETT. This sample Verilog-A code is reasonably self-explanatory. $table_model function needs three types of inputs: variables; a data file which has a data table; and control signals for interpolation and extrapolation. In this sample, the degree of the splines used for the interpolation process is 1. To evaluate a point beyond the interpolation area, the S (spline) extrapolation method is used. In this model, the length of a device is fixed and the width is modulated by parameter “width”.

Sample Verilog-A Code for NHETT

```
module NHETT (s, g, d);
inout s, g, d;
electrical s, g, d;
real cap_gd_value, cap_gs_value;
parameter real width = 1;
analog begin
        cap_gd_value = $table_model( V(g) - V(s), V(d) - V(s), \
        "./cgd_table.csv", "1S, 1S");
        cap_gs_value = $table_model( V(g) - V(s), V(d) - V(s), \
        "./cgs_table.csv", "1S, 1S");
        I(g, d) <+ cap_gd_value * ddt( V(g) - V(d) ) * width;
        I(g, s) <+ cap_gs_value * ddt( V(g) - V(s) ) * width;
        I(d, s) <+ $table_model( V(g) - V(s), V(d) - V(s), \
        "./ids_table.csv", "1S,1S") * width;
```

```
end
endmodule
```

In Figure 5.4, new symbols for NHETT and PHETT are presented. An arrow inside the conventional MOSFET symbol denotes the direction of forward biased current, which is from drain to source for NHETT and vice versa for PHETT.

To verify that Verilog-A device modeling is accurate enough to generate reasonable simulation results, conventional Berkeley Short-channel IGFET Model (BSIM) modeling and Verilog-A modeling are compared in Figure 5.5. The tables of capacitance and current are extracted from BSIM model and Verilog-A model is built based on extracted tables. Figure 5.5 shows the normalized delay and the normalized dynamic power consumption of an 11-stage ring oscillator from two models: BSIM and Verilog-A. The difference between two simulation results is acceptable for investigating the basic characteristics of a futuristic device.

## 5.4 HETT-Based Circuit Analysis

The steep subthreshold swing and larger $I_{on}$ of HETTs compared to MOSFETs allow aggressive voltage scaling at iso-performance, enabling dynamic power reductions. To quantify this power reduction, ring oscillators are simulated with HETTs and compared with a commercial bulk CMOS 45nm technology. In addition, the circuit design impact of HETT limitations is also addressed in this section.

### 5.4.1 Dynamic Power Reduction

A 31-stage ring oscillator with minimum sized inverters is used to evaluate dynamic power consumption. Leakage power is subtracted from total power to focus only on dynamic power in this section since the leakage power contribution was less than 10%. In addition, minimum sized inverters are used since minimizing size results the least power for a given switching period.

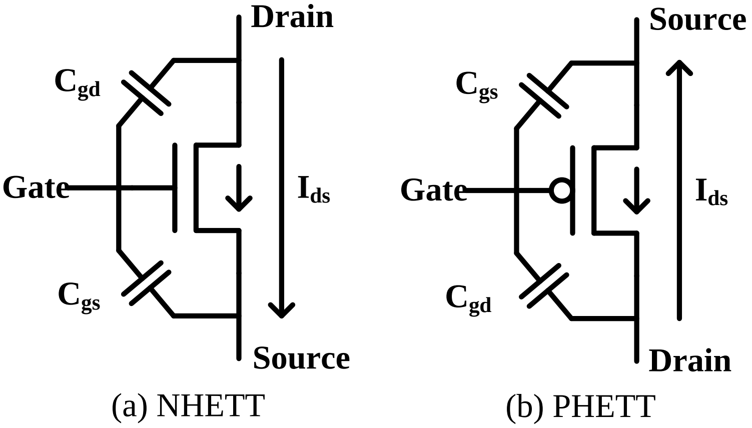


Figure 5.4: Device symbols for (a) NHETT (b) PHETT

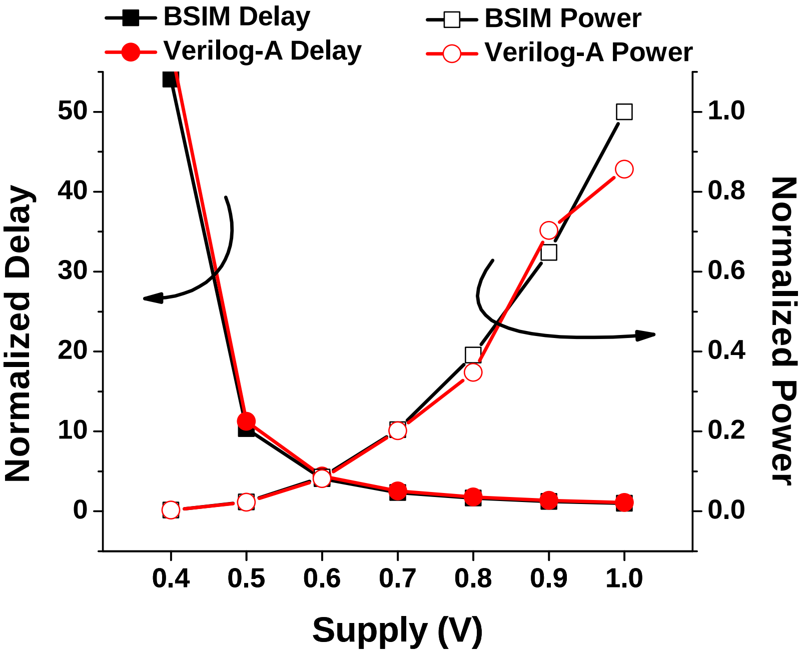


Figure 5.5: BSIM vs. Verilog-A

Figure 5.6 shows the dynamic power reduction of the 31-stage ring oscillator with HETT devices compared to the commercial bulk CMOS 45nm technology. The CMOS technology has two types of logic devices: LP and GP. The LP devices are designed for low power operation and exhibit lower leakage than GP devices. Iso-speed dynamic power consumption of LP devices is expected to be worse because $I_{on}$ in LP is smaller than in GP. With identical device sizes in both CMOS and HETT technology, supply voltage is lowered from 1.0V to 0.3V in CMOS and from 1.0V to 0.15V in HETT with 0.05V steps. At 1.0V, the GP-based ring oscillator has a period of 450ps and 53.9$\mu$W dynamic power consumption. To maintain the same period, the ring oscillator with HETT consumes only 5.74$\mu$W at 0.355V, achieving a 9.4× dynamic power reduction. For 45nm LP, more dynamic power reduction is observed. At 1.0V, the LP ring oscillator period is 980ps and consumes 19.98$\mu$W while the HETT-based ring oscillator consumes 19× less power (1.05$\mu$W) at 0.226V with the same period.

### 5.4.2 Limitations of HETT-Based Circuits

#### 5.4.2.1 Asymmetric Current Flow

HETT source and drain are determined at fabrication time and current flow between the two nodes is not symmetric. Figure 5.7 demonstrates this asymmetric current flow in an NHETT. We assume that the nominal voltage of HETTs will be <0.5V as HETTs target ultra-low voltage applications and are well suited for this voltage regime. Figure 5.7(a) shows forward bias current with $V_{gs}$ swept from 0V to 0.5V. The drain current curves look similar to CMOS devices. However, reverse bias current, where the voltage across the drain and source is negative, differs from CMOS devices as shown in Figure 5.7(b). Note that Ids is negative in Figure 5.7(b). For most regions of $V_{ds}$, drain current is several orders of magnitude smaller than forward current. However, there are two cases where the reverse bias current becomes non-negligible. First is when $V_{ds}$ is approximately -0.5V, at which point drain

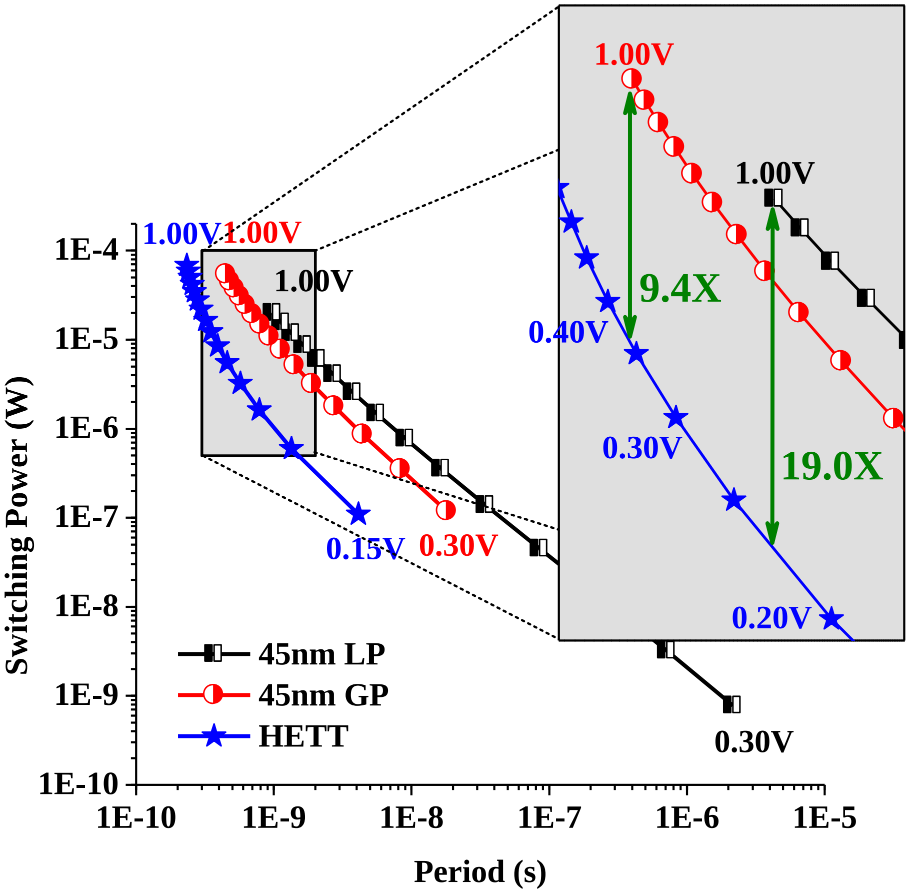


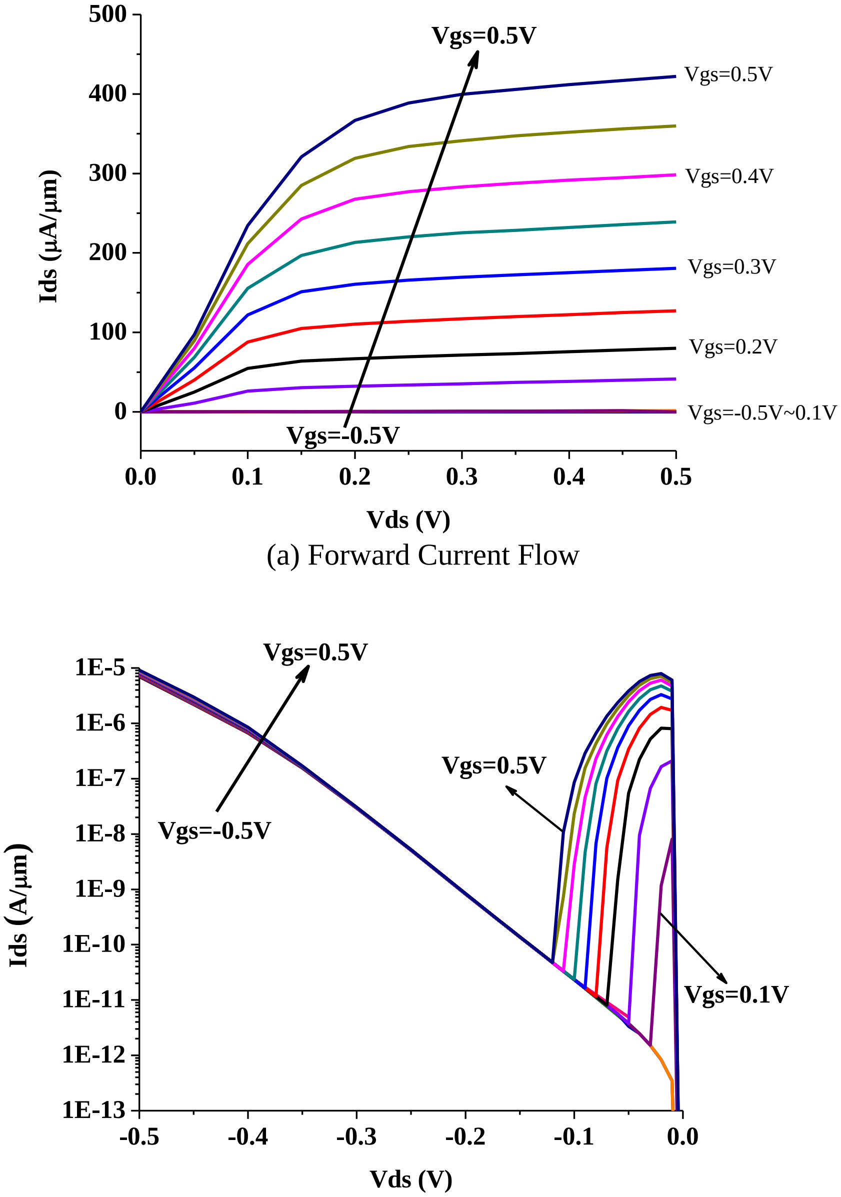
Figure 5.6: Comparison of dynamic power consumption with commercial bulk-CMOS 45nm LP, 45nm GP, and HETT devices

current become non-negligible regardless of $V_{gs}$. The second case occurs for positive $V_{gs}$ combined with a small negative $V_{ds}$. PHETTs exhibit similar asymmetry in their current flow.

The asymmetric current flow does not restrict the use of traditional static CMOS logic circuits with Pull-Up Network (PUN) and Pull-Down Network (PDN) because the current flow of each device in the PUN and PDN is uni-directional. However, pass-transistor and transmission-gate operation is limited since they require current flow in both directions. Figure 5.8 details the limitation of HETT-based pass-transistor circuits. Because the drain and source of the device are fixed, there are two ways to implement a pass-gate in a circuit: oriented left and right. In both cases, the current flow characteristics are classified again by two cases: passing logic "1" and passing logic "0".

A pass-gate propagating logic "1" is shown in Figure 5.8(a), where left and right configurations are both illustrated. Before the input at the gate of pass-gate is switched at 2ns, the output of the rightward pass-gate stays near 0V while the output of leftward pass-gate is pulled up to ∼150mV. This is due to the fact that reverse $I_{off}$ can be larger than forward $I_{off}$. When the input switches at 2ns, the output of the rightward pass-gate immediately switches to $\sim V_{DD}$ while the output of the leftward pass-gate remains near 200mV and increases very slowly. This clearly shows that forward $I_{on}$ can strongly drive the output but reverse $I_{on}$ cannot. For pass-gate passing logic "0" (Figure 5.8(b)), similar trends can be observed and only the leftward pass-gate functions well. This directional current driving capability renders pass-gate logic useless for HETT-based circuits. The asymmetric current flow also limits the use of the standard 6-T SRAM cell and static latches/registers, which exploit pass-gates and transmission-gates as key components. Latches and registers can be implemented without pass-gates and transmission-gates by using clocked CMOS logic.

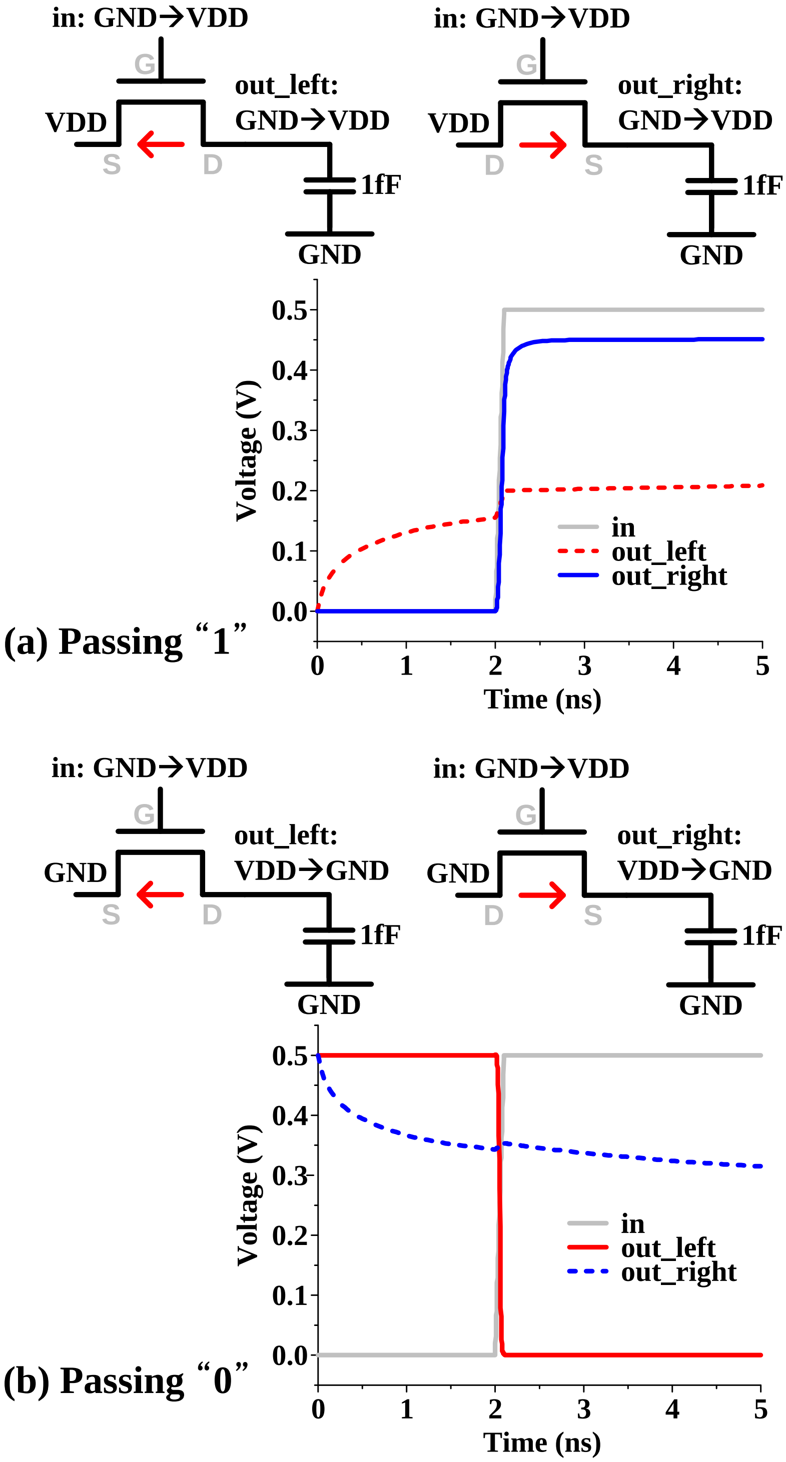
Differently from pass-transistor, the unfavorable effect of uni-directional current



(a) Forward Current Flow

(b) Reverse Current Flow

Figure 5.7: (a) Forward bias and (b) reverse bias drain current of HETT device with L=40nm



(a) Passing "1"

(b) Passing "0"

Figure 5.8: Two orientations (left and right) for implementing NHETT-based pass-gates passing a "1" (a) and passing "0" (b)

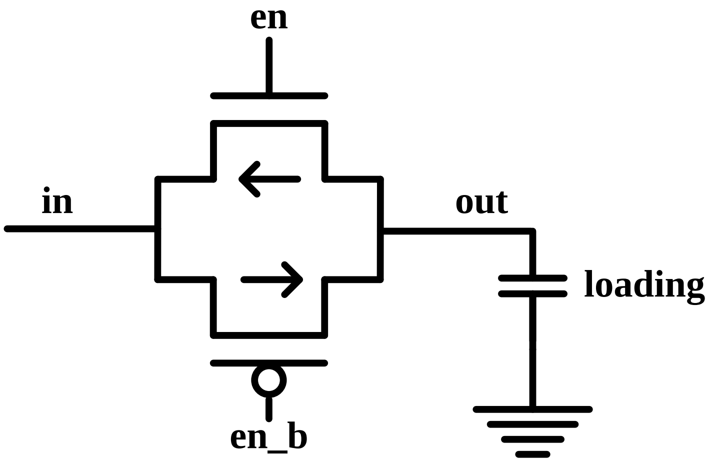


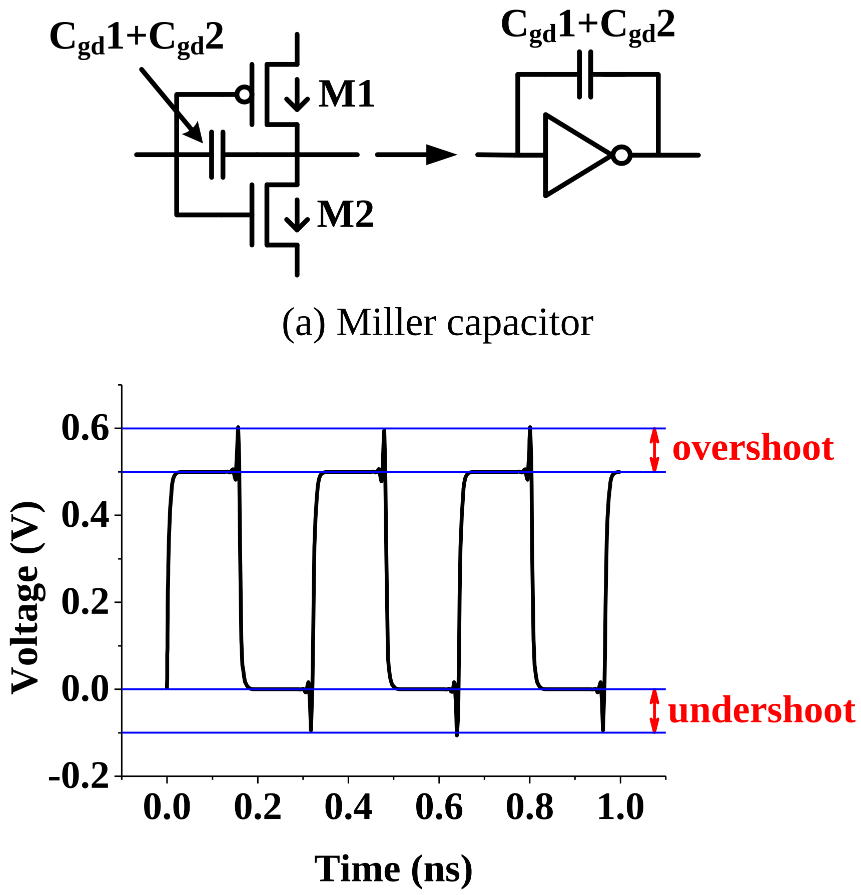
Figure 5.9: HETT-Based Transmission-Gate

flow on transmission-gate operation is mitigated if the source and drain directions of NHETT and PHETT are appropriately chosen. As already shown in Figure 5.8, NHETT is better at passing logic “0” than logic “1” due to $V_{th}$ drop at the output of a gate. In the same reason, PHETT is good at passing logic “1”. Therefore, the current direction of NHETT must be from output to input while the current direction of PHETT must be opposite (Figure 5.9).

### 5.4.3 Increased Miller Capacitance

The capacitance between gate and drain is often referred to as the Miller capacitance as it is impacted as the Miller effect [60]. During a voltage transition, the two terminals of the Miller capacitor are moving in opposite directions such that the voltage change across the capacitor is twice the absolute voltage change (Figure 5.10(a)), hence this capacitance significantly impacts loading. In addition, it causes overshoots and undershoots during transitions due to capacitive coupling between input and output of the gate (Figure 5.10(b)), which results in additional capacitive loading, and performance overhead.

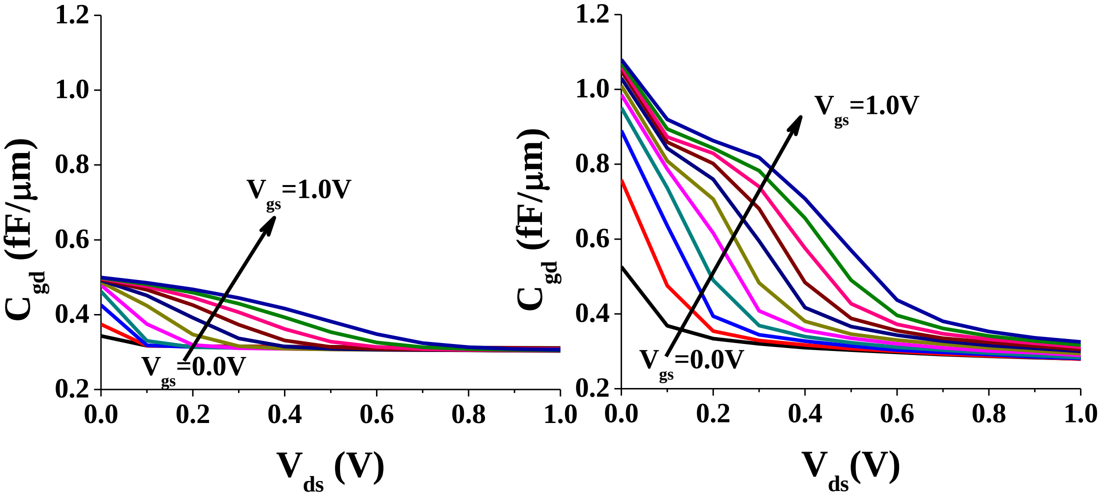
The Miller capacitance in HETTs is larger than the Miller capacitance in MOSFETs.



(a) Miller capacitor

(b) overshoot/undershoot

Figure 5.10: (a) Miller capacitor acting as 2× larger capacitive loading and (b) overshoot and undershoot caused by capacitive coupling



(a) CMOS

(b) HETT

Figure 5.11: $C_{gd}$ comparison of (a) CMOS (NMOS) and (b) HETT (NHETT)

This arises from the linking of the inversion layer in HETTs to the drain rather than the source, as is the case in MOSFETs. In HETTs with large gate bias, what can be viewed as a parasitic inversion layer forms with carriers drawn from the drain side – this inversion layer is not the primary form of current conduction in the device, hence the term parasitic. Under this bias condition, $C_{gd}$ becomes essentially equivalent to the entire channel capacitance due to the parasitic inversion layer. This principle is the same as that described in detail in [39] for carbon nanotube-based tunneling FETs.

In Figure 5.11, we find that the extracted $C_{gd}$ of an NHETT is $\sim 2\times$ larger than $C_{gd}$ of NMOS in a commercial bulk CMOS 45nm technology. To evaluate the impact of this larger Miller capacitance in HETTs, average overshoot and undershoot (as a percentage of the 0.5V supply) is evaluated and shown in Figure 5.12. If the electrical effort (from logical effort [63]) is larger than four, overshoot effects in HETTs are comparable to that in commercial 45nm CMOS technologies. Hence we conclude that for typical loads, the increased $C_{gd}$ will not have significant impact on circuit performance, although it should be considered for very lightly loaded gates.

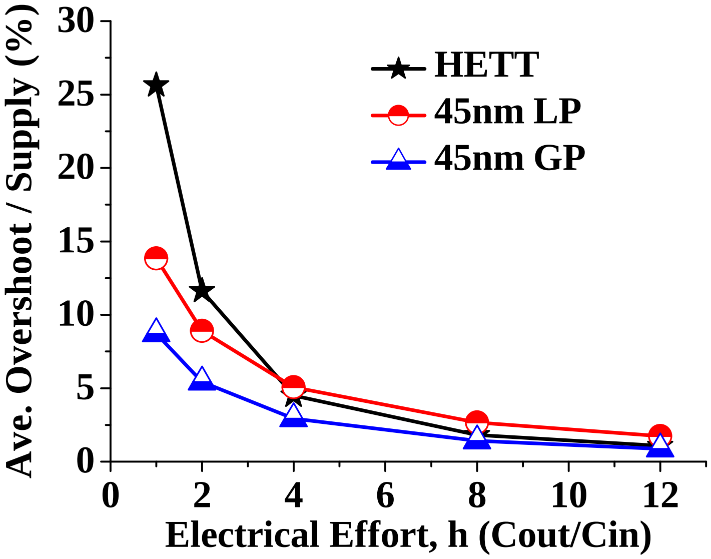


Figure 5.12: Overshoot effects in HETT are not significant with electrical effort of 4 or larger despite the larger $C_{gd}$

## 5.5 HETT-Based SRAM Design

The asymmetric current flow of HETT places restrictions on the use of pass-gate and transmission-gate. While this limitation is not severe for logic circuits, it poses a significant problem for the standard 6-T SRAM, which uses pass gates for access transistors. In this section, we first analyze the implications of asymmetric current flow on SRAM operation and go on to propose an alternative 7-T HETT-based SRAM cell topology. We then compare 7-T performance and robustness to that of a CMOS-based 6-T SRAM design.

### 5.5.1 Limitations in Standard 6-T SRAM

#### 5.5.1.1 CMOS Standard 6-T SRAM

To understand the difference between HETT-based 6-T SRAM and CMOS-based 6-T SRAM, we trace current flow paths in read and write operations. Figure 5.13 shows a CMOS 6-T SRAM cell storing “0”. To read the stored value, bit lines (BIT, BIT_B) are pre-charged to $V_{DD}$ and as Word-Line (WL) is driven high, NPDL pulls

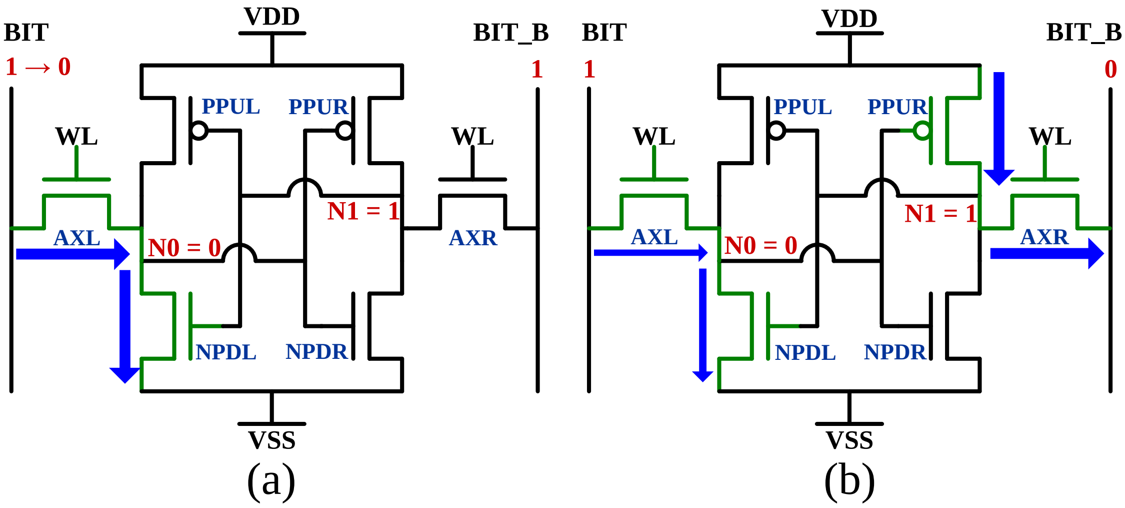


Figure 5.13: Current flow paths in (a) read and (b) write operations in CMOS 6-T SRAM

down the voltage at BIT as shown in Figure 5.13(a). This pull down current or voltage can be sensed by a sense amplifier to determine the stored value. For writing a value "1", as shown in Figure 5.13(b), AXL pulls up internal node N0 while AXR pulls down internal node N1. However, since both access transistors are NMOS, which are better at pulling low, AXR plays the major role in write 1 operation. AXL aids in writing a 1 by pulling up N0 to a certain extent and making the bit flip more easily.

For this type of SRAM, read stability can be improved by increasing the sizing ratio of NPDL to AXL (or NPD to AX), which is commonly referred to as the cell $\beta$-ratio. As cell $\beta$-ratio increases, NPDL in Figure 5.13(a) holds the voltage at node N0 to ground more strongly during read, making it more stable. At the same time, this worsens writability of the cell by making it more difficult to change the voltage at node N0. However as shown in Figure 5.13(b), since the pull down current path (AXR) plays the major role in writing, the size ratio of AXR to PPUR, or AX to PPU, is the critical one for writability and can be improved by increasing this ratio. This implies that, up to a point, readability and writability in CMOS 6-T SRAM can be improved individually at the cost of larger area.

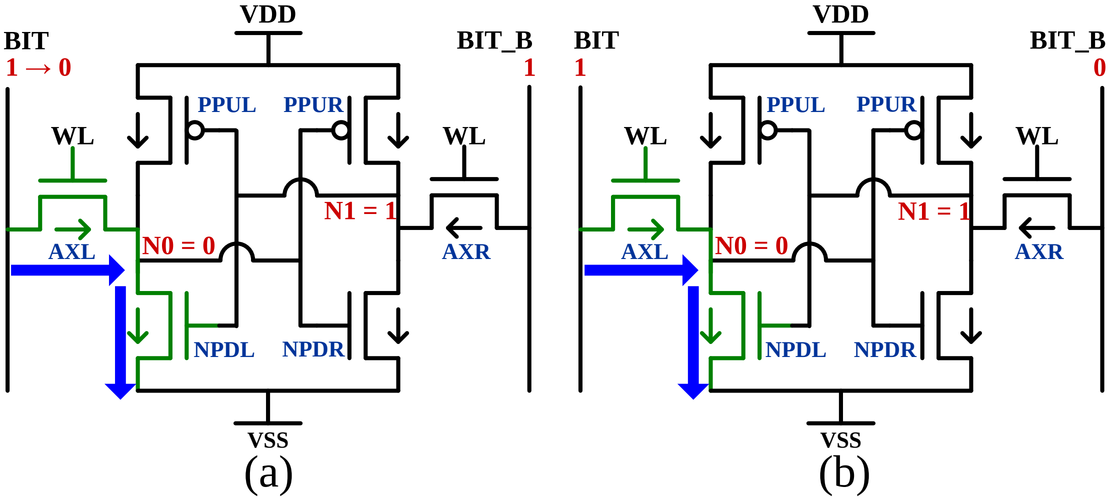


Figure 5.14: Current flow paths in (a) read and (b) write operations in HETT 6-T SRAM with inward direction access transistors

#### 5.5.1.2 HETT Standard 6-T SRAM with Inward Access Transistors

Due to its uni-directional nature, access transistors in HETT 6-T SRAM can drive current either inward or outward only. Figure 5.14 shows a HETT 6-T SRAM structure with inward current flow configuration and storing “0”. Read operation for this SRAM is similar to a CMOS 6-T SRAM. Bit-lines are precharged and current flows through AXL and NPDL. Therefore, similar to CMOS 6-T SRAM, higher cell $\beta$-ratio is preferred for preventing read upset.

However, to write “1” to this cell, AXR cannot pull down the voltage at N1 since it can only conduct current inward, implying that AXL must pull up the voltage at N0 without differential aid, as shown in Figure 5.14(b). Therefore, the write operation is performed only by one side and the stronger current path is removed in HETT 6-T SRAM. Since we are relying on an N-type transistor to drive the internal node voltage high, writability of this cell is substantially worse than a CMOS 6-T SRAM. To overcome poor writability, AXL should be strengthened compared to NPDL, i.e., the cell $\beta$-ratio should be decreased. However, decreasing the cell $\beta$-ratio negatively affects the read margin.

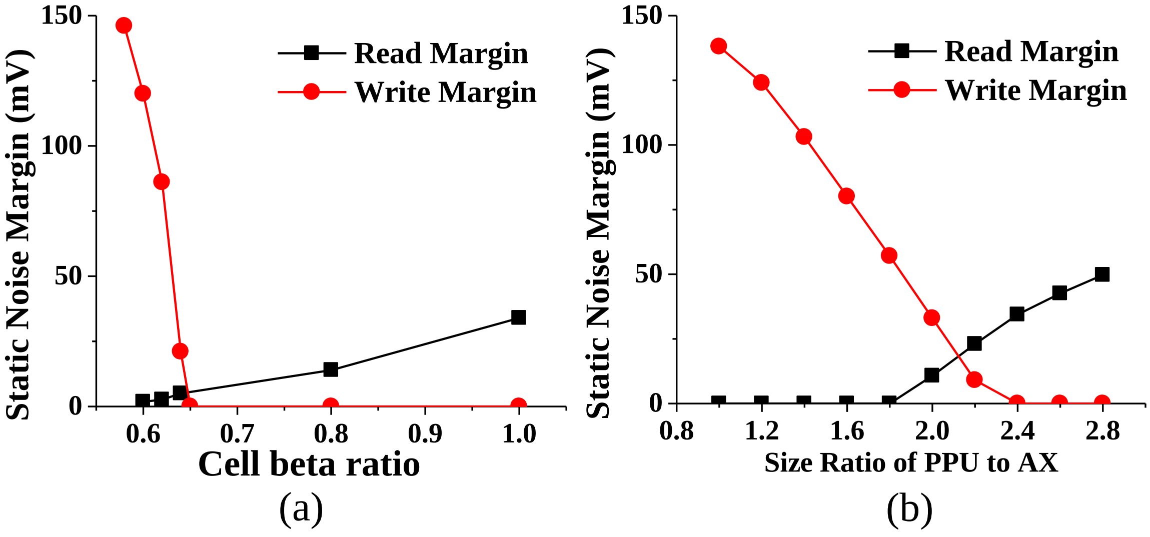


Figure 5.15: Static noise margins of HETT 6-T SRAM with (a) inward and (b) outward access transistor with $V_{DD}$=0.5V

This trade-off between readability and writability can be clearly seen if we plot SNM of read and write operation versus cell $\beta$-ratio, as shown in Figure 5.15(a). SNM is the maximum DC voltage of the noise that can be tolerated by the SRAM and it is widely used for modeling stability of SRAM cells [61]. SNM can be defined for three different operations – read, write, and standby (hold) – but only read and write margins are compared here since they limit SRAM stability. In SNM analysis for HETT-based SRAMs, all simulations use $V_{DD}$ = 0.5V since HETTs are aimed at this voltage regime. For HETT 6-T SRAM with inward access transistors with cell $\beta$-ratio of 1, read margin is 34mV but write margin is 0V, meaning that write operation is impossible. As we decrease the cell $\beta$-ratio to improve writability, write margin becomes positive at a cell $\beta$-ratio of 0.64, however read margin at this point has degraded to <3 mV, indicating that the cell is highly vulnerable to read upset at this design point. From this we conclude that HETT 6-T SRAM with inward access transistors is not feasible.

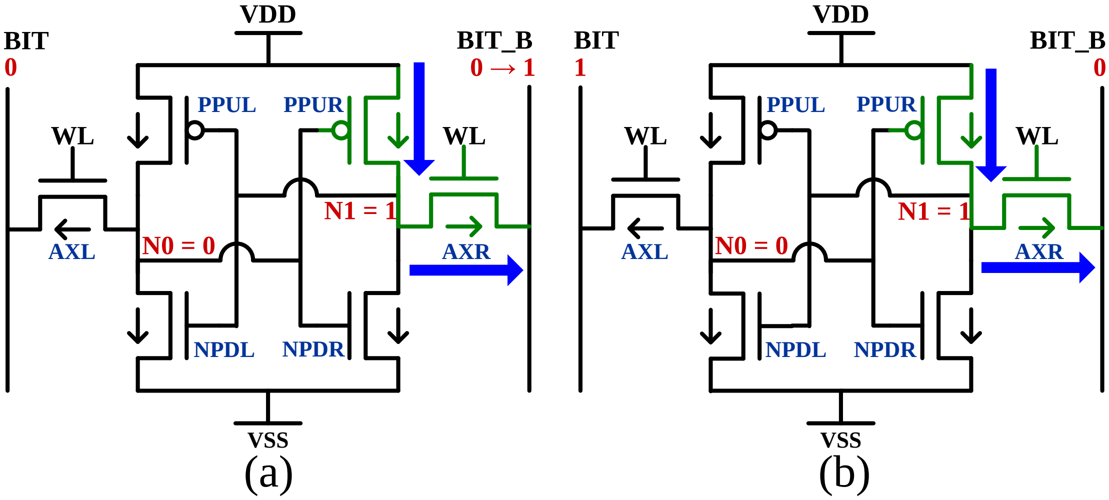


Figure 5.16: Current flow paths in (a) read and (b) write operations in HETT 6-T SRAM with outward direction access transistors

#### 5.5.1.3 HETT Standard 6-T SRAM with Outward Access Transistors

HETT 6-T SRAM with outward access transistors has a similar limitation. Figure 5.16(a) shows a read operation, where bit lines (BIT, BIT_B) are pre-discharged and BIT_B is charged through AXR and must be sensed. For writing, AXR must drive internal node N1 to ground and flip the stored value without differential assistance from AXL. Since both of these operations involve PPUR and AXR, adjusting the ratio of PPUR to AXR strengths will improve one operation and worsen the other. This trade-off can be clearly seen in Figure 5.16(b). The read operation requires PPUR to AXR ratio higher than 1.8, while the write operation malfunctions when the ratio is higher than 2.4. In the remaining design space the SNM for read/write operations is limited to <50 mV, which is insufficient. Therefore, an alternative SRAM topology is needed to achieve robust low leakage SRAM with HETTs.

### 5.5.2 HETT-Based 7-T SRAM

Figure 5.17 shows the proposed 7-T SRAM structure that overcomes the trade-off between read and write in HETT-based 6-T SRAM. The basic structure is based

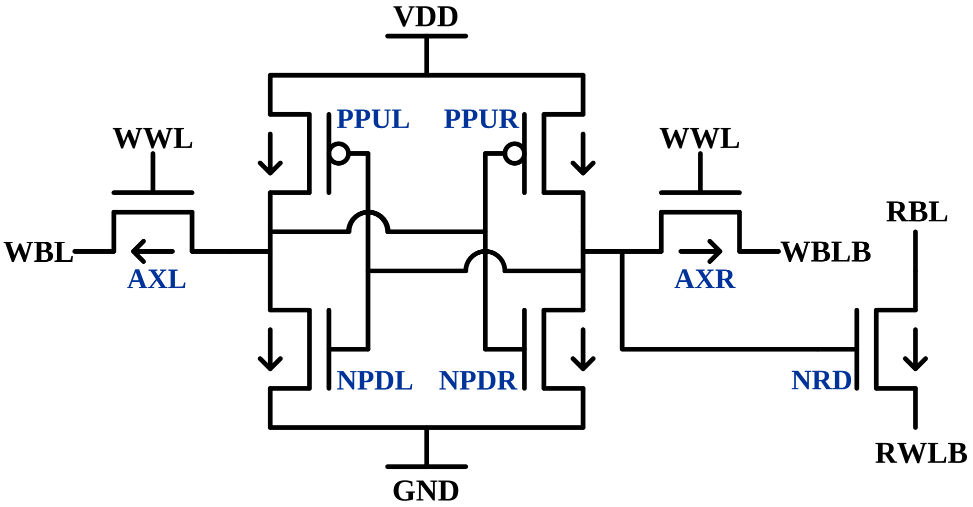


Figure 5.17: Proposed HETT 7-T SRAM structure

on HETT 6-T SRAM with outward access transistors, but includes an additional NHETT labeled "NRD" that is used to read out the cell contents.

Unlike the other 6-T SRAM structures, read operation in 7-T SRAM is conducted solely through NRD. Figure 5.18 illustrates how NRD in each cell is connected in the array structure. The source of NRD is connected to that of other cells in the same word (RWLB), while the drain is connected to that of other cells in same column (RBL). To read values in word[0] (top row of Figure 5.18), bit-lines (RBL[0], RBL[1]) are precharged and RWLB[0] is asserted (driven to ground) while all other RWLBs are set to $V_{DD}$. Since the source of the NRDs in word[0] are set to ground, cells that store value "1" can discharge the bit line, as depicted with the thick arrow in Figure 15. With CMOS transistors, this read scheme does not work because, as RBL[0] is discharged, other cells storing "1" on the same bit line can start charging up RBL[0] as in the case of the bottom-left cell in Figure 5.18. However, by leveraging the asymmetric nature of HETTs, this unwanted reverse-direction charging current is eliminated without the cost of an additional transistor, as in the well-known 8-T structures [8]. The HETT 7-T SRAM is estimated to have <15% area overhead over a standard 6-T while 8-T SRAM exhibits 29% cell area overhead [9]. Figure 5.19 shows

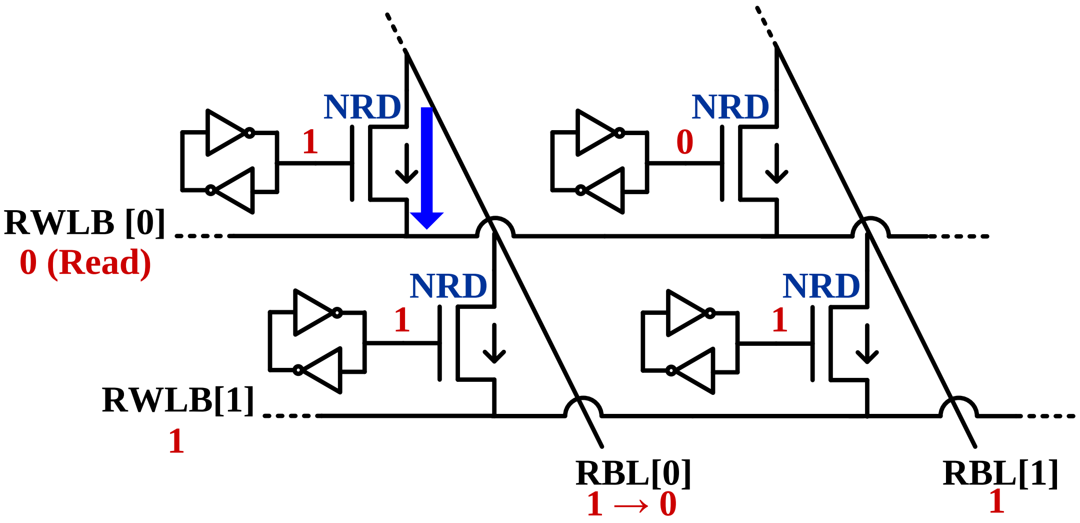
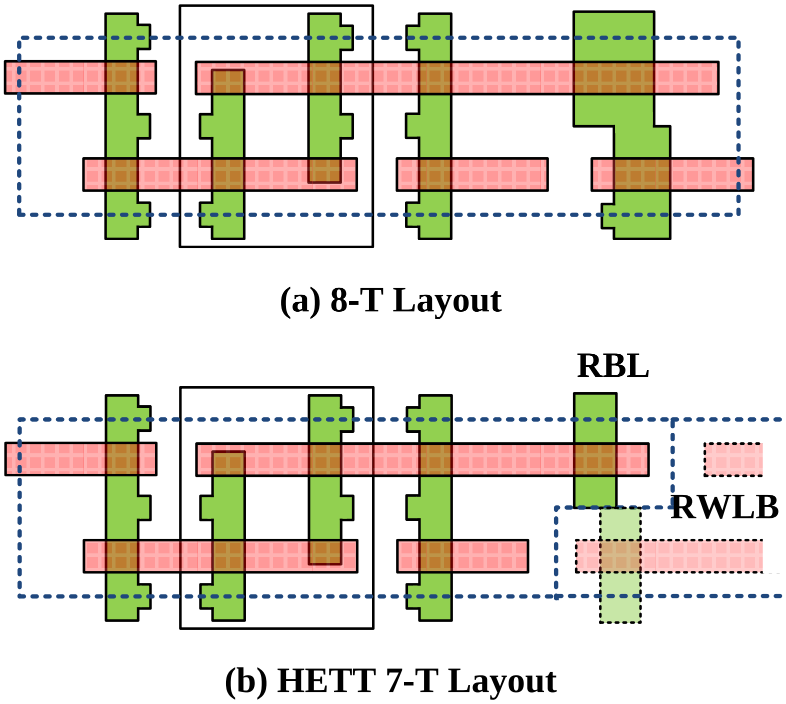


Figure 5.18: Read operation in 7-T SRAM array

that two read transistors (NRD in Figure 25) from adjacent cells can be abutted in 7-T SRAM, making the overhead for two 7-T cells equal to that of one 8-T cell. Moreover, as will be shown below the 7-T cell with all transistors at minimum size shows improved robustness over 6-T at low voltage, hence if an upsized 6-T were used to achieve iso-robustness the area penalty would be much smaller than 15%.

A write operation in this 7-T structure is equivalent to the HETT 6-T SRAM with outward access transistors. However, since the read/write operations are performed by separate current paths, device sizes for all transistors other than NRD can be chosen to favor writability. The outward access transistor scheme is used for its superior writability over the inward configuration when all transistors are of near-minimum width to improve density.

As the additional benefit of unidirectional current flow, the half select disturb in a bit-interleaved array can be mitigated with HETT-based 7-T SRAM. The half select disturb accidently flips internal data in half selected bitcells which share the same write word line with targeted bitcells for write operation (Section II [33]). In HETT-based 7-T SRAM, if the bit lines of half selected bitcells are being kept at $V_{DD}$, the amount of current flow via access transistors is limited to the leakage current


(a) 8-T Layout

(b) HETT 7-T Layout

Figure 5.19: (a) 8-T layout [9] and (b) corresponding HETT 7-T layout

level. Therefore, HETT-based 7-T SRAMs have improved immunity during half select accesses.

We compare SNM of HETT-based 7-T SRAM to a 45nm commercial bulk CMOS 6-T SRAM cell provided by a foundry. All HETT devices are set to equal (minimum) width for maximum density. Read and write margins of both types of SRAMs across a range of supply voltages are plotted in Figure 5.20. SNM for HETT is analyzed with supply voltages up to 0.9V only since HETT is designed for low voltage (∼ 0.5V) operation. Write margins of HETT 7-T SRAM are more than 30% higher than CMOS 6-T SRAM for supply voltages of >0.4V as shown in Figure 5.20.

Since the read operation uses an additional read transistor in the HETT 7-T SRAM and all other transistors are in standby (hold) state during read operation, hold margin is equivalent to read margin in HETT 7-T SRAM. Given this, HETT 7-T read margin is 232 mV at $V_{DD}$=0.9V and 129 mV at 0.5V, which is 41% and 37%

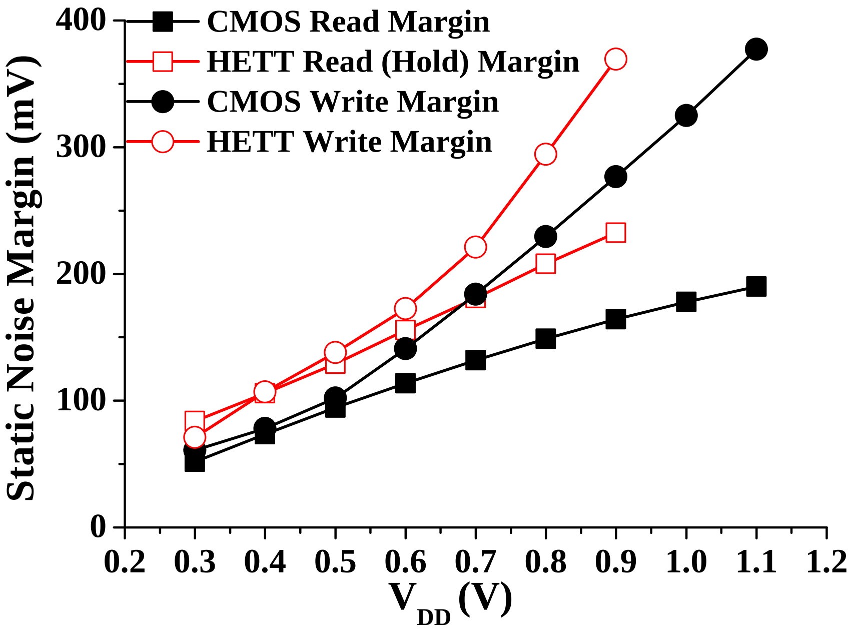


Figure 5.20: Read/Write margin of 45nm commercial bulk CMOS 6-T SRAM and HETT 7-T SRAM

higher than commercial bulk CMOS 6-T SRAM, respectively. Such improvements in read/write margin can be observed for $V_{DD}$ down to 0.3V, suggesting that improved read/write robustness can be achieved with HETT 7-T SRAM over traditional CMOS at low voltage.

Finally, HETT-based SRAM standby power is significantly reduced compared to CMOS 6-T SRAM, as seen in Figure 5.21. At a supply voltage of 0.9V, standby power is reduced by 36.8× and at 0.5V, by 7.4×. This clearly shows the promising low-leakage properties of HETT devices for future memory-dominated low-power applications.

## 5.6 Conclusion

A circuit perspective of a new promising tunneling transistor, HETT, with steep subthreshold swing for extremely low power applications was presented in this paper. We observed 9-19× dynamic power reduction with HETT-based circuits due to their

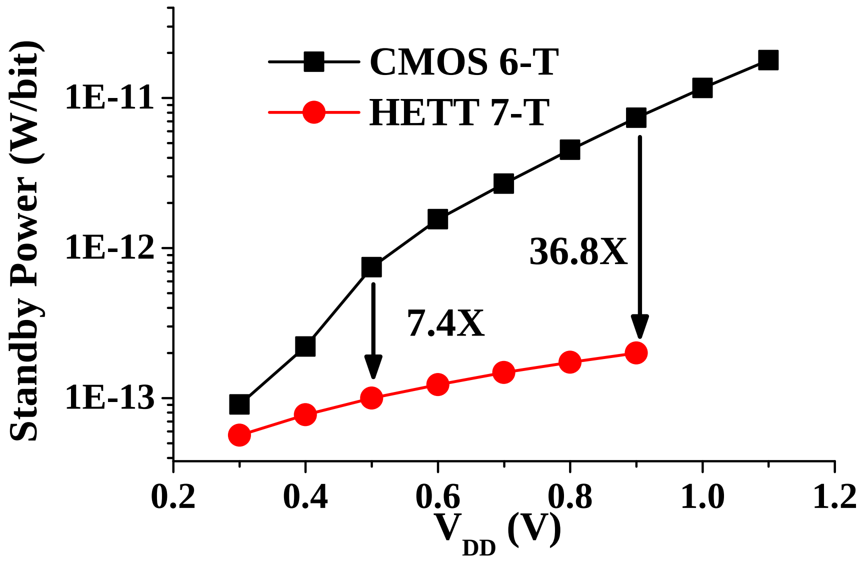


Figure 5.21: Standby power of CMOS 6-T and HETT 7-T SRAM

improved voltage scalability. We examined the limitations of HETTs as they relate to circuit operation. To overcome and exploit the inherent device asymmetry, a new HETT-based SRAM cell topology was presented with 7-37× leakage power reduction.

# CHAPTER VI

# Conclusion

Designing robust low voltage SRAM is one of the key challenges in modern VLSI design. Technology scaling and voltage scaling are two driving factors to enable more efficient battery operated systems. SRAM design is important since SRAM yield is critical as technology scales and voltage scales. This thesis defines challenges in design and analysis of robust low voltage SRAM from different angles and proposes solutions based on the target applications.

A new SRAM yield estimation method is proposed in Chapter II. Bit-interleaved 8-T SRAM is a good candidate at low voltage in scaled technology because read and write can be separately optimize and soft errors can be easily fixed with SECDED. When the 8-T SRAM bitcell is used, people usually assume that write operation can be optimized without any limit; however, the trade-off between write and half select disturb is discussed in this chapter. Because of the double-sided constraints on 8-T bitcell write operation, optimal WWL pulse width can be found where maximizing writability and minimizing half select disturb. With this analysis, we can find appropriate device sizing and optimal write assist techniques during SRAM design phase.

The underlying idea in Chapter II is extended to the adaptive WWL pulse width and voltage level modulation scheme in Chapter IV. At low voltage, the distribution of write time is considerably wide and has a long tail. Instead of giving excessive margin, WWL pulse width is adaptively modulated based on the required pulse width of target bitcells. This can be enabled by concurrent read via decoupled read path in the 8-T bitcell. To solve static write failures, adaptive WWL voltage level modulation is introduced. However, the boosted WWL degrades half select disturb immunity. Therefore, bit-line regeneration scheme is used for half selected bitcells to mitigate half select disturb. With this approach, the minimum operating voltage ($V_{min}$) is lowered and overall performance and yield are improved.

A low leakage SRAM for sensor applications is proposed in Chapter III. Since sensor applications usually spend their most of lifetime in standby mode, it is important to reduce leakage in SRAM which is not turned off even during standby mode. To develop a low leakage bitcell, HVT devices are used in 6-T part which keeps the data while SVT devices are adopted for read path since it can be shut down. To reduce leakage further, reverse body biasing and floating bit-lines are used. For compensating low performance due to HVT devices, bit-line boosting is adopted. This low leakage SRAM is successfully demonstrated with various sensor applications [11, 13].

Finally, low power circuit design based on HETT is discussed in Chapter V. MOSFET has a limited subthreshold swing of 60mV/dec and it limits lowering voltage. By using HETT with steep subthreshold swing, both active power and leakage power can decrease. Based on HETT's specific characteristic, its implications on circuit design and SRAM design are discussed.

This thesis provides issues in state-of-the-art SRAM design from different angles and proposes several ideas. However, there are still a significant amount of challenges in designing low voltage robust SRAM. Mitigating variation, maintaining high yield, providing decent performance at low voltage with newly developed devices will be driving factors for future SRAM development.

# BIBLIOGRAPHY

# BIBLIOGRAPHY

[1] J. Appenzeller, Y.-M. Lin, J. Knoch, and P. Avouris, "Band-to-Band Tunneling in Carbon Nanotube Field-Effect Transistors," Vol. 93, pp. 196805-1–196805-4, Nov. 2004.

[2] R.C. Baumann, "Soft Errors in Advanced Semiconductor Devices-Part I: the Three Radiation Sources," *Transaction on Device and Materials Reliability*, Vol. 1, pp. 17–22, Mar. 2001.

[3] K. Bernstein *et al.*, "High-performance CMOS Variability in the 65-nm Regime and Beyond," *IBM Journal of Research and Development*, Vol. 50, pp. 433–449, July/Sept. 2006.

[4] K. M. Cao *et al.*, "BSIM4 Gate Leakage Model Including Source Drain Partition," *International Electron Device Meeting*, pp. 815–818, 2000.

[5] V. Chandra, R. Aitken, "Impact of Voltage Scaling on Nanoscale SRAM Reliability," *Design, Automation & Test in Europe*, pp. 387–392, 2009.

[6] V. Chandra, C. Pietrzyk, R. Aitken, "On the Efficacy of Write-Assist Techniques in Low Voltage Nanoscale SRAMs," *Design, Automation & Test in Europe*, pp. 345–350, 2010.

[7] I.J. Chang, J-J. Kim, S.P. Park, K. Roy, "A 32kb 10T Subthreshold SRAM Array with Bit-Interleaving and Differential Read Scheme in 90nm CMOS," *International Solid-State Circuits Conference*, pp. 388–389, 2008.

[8] L. Chang *et al.*, "Stable SRAM Cell Design for the 32nm Node and Beyond," *Symposium on VLSI Circuits*, pp. 128–129, 2005.

[9] L. Chang *et al.*, "A 5.3GHz 8T-SRAM with Operation Down to 0.41V in 65nm CMOS," *Symposium on VLSI Circuits*, pp. 252–253, 2007.

[10] G. Chen *et al.*, "Yield-driven Near-Threshold SRAM Design," *International Conference on Computer-Aided Design*, pp. 660–666, 2007.

[11] G. Chen *et al.*, "Millimeter-Scale Nearly Perpetual Sensor System with Stacked Battery and Solar Cells," *International Solid-State Circuits Conference*, pp. 288–289, 2010.

[12] G. Chen *et al.*, "Crosshairs SRAM – An Adaptive Memory for Mitigating Parametric Failures," *European Solid-States Circuits Conference*, pp. 366–369, 2010.

[13] G. Chen *et al.*, "A 1 Cubic Millimeter Energy-Autonomous Wireless Intraocular Pressure Monitor," *International Solid-State Circuits Conference*, pp. 310–312, 2011.

[14] W.Y. Choi *et al.*, "70-nm Impact-Ionization Metal-oxide-semiconductor (I-MOS) Devices Integrated with Tunneling Field-Effect Transistors (TFETs)," *International Electron Device Meeting*, pp. 955–958, 2005.

[15] V. De, S. Borkar, "Technology and Design Challenges for Low Power and High Performance," *International Symposium on Low Power Electronics and Design*, pp. 163–168, 1999.

[16] G. Dewey *et al.*, "Fabrication, Characterization, and Physics of III-V Heterojunction Tunneling Field Effect Transistors (H-TFET) for Steep Sub-Threshold Swing," *International Electron Device Meeting*, to appear, 2011.

[17] C. H. Diaz *et al.*, “A 0.18 $\mu$m CMOS Logic Technology with Dual Gate Oxide and Low-k Interconnect for High-Performance and Low-Power Applications,” *Symposium on VLSI Circuits*, pp. 11–12, 1999.

[18] W. Dong, P. Li, G.M. Huang, “SRAM Dynamic Stability: Theory, Variability and Analysis,” *International Conference on Computer-Aided Design*, pp. 378–385, 2008.

[19] R. G. Dreslinski *et al.*, “Near-Threshold Computing: Reclaiming Moore’s Law Through Energy Efficient Integrated Circuits,” *Proceedings of The IEEE* Vol. 98, pp. 253–266, Feb. 2010.

[20] M. V. Fischetti, S. E. Laux, “Band structure, deformation potentials, and carrier mobility in strained Si, Ge, and SiGe alloys,” *Journal of Applied Physics*, Vol. 80, pp. 2234–2252, Aug. 1996.

[21] E. Grossar *et al.*, “Read Stability and Writability Analysis of SRAM Cells for Nanometer Technologies,” *Journal of Solid State Circuits*, Vol. 41, pp. 2577–2588, Nov. 2006.

[22] R.W. Hamming, “Error Detecting and Error Correcting Codes,” *Bell System Technical Journal*, Vol. 29, pp. 147–160, Apr. 1950.

[23] M. Hane, T. Ikezawa, T. Ezaki, “Atomistic 3D Process/Device Simulation Considering Gate Line-Edge Roughness and Poly-Si Random Crystal Orientation Effects,” *International Electron Device Meeting*, pp. 9.5.1–9.5.4, 2003.

[24] S. Hanson *et al.*, “A Low-Voltage Processor for Sensing Applications with Picowatt Standby Mode,” *Journal of Solid State Circuits*, Vol. 44, pp. 1145–1155, Apr. 2009.

[25] K. Honda *et al.*, "Elimination of Half Select Disturb in 8T-SRAM by Local Injected Electron Asymmetric Pass Gate Transistor," *Custom Integrated Circuits Conference*, pp. 1–4, 2010.

[26] Q. Huang *et al.*, "Self-Depleted T-gate Schottky Barrier Tunneling FET with Low Average Subthreshold Slope and High $I_{ON}/I_{OFF}$ by Gate Configuration and Barrier Modulation," *International Electron Device Meeting*, to appear, 2011.

[27] M. Ieong *et al.*, "Comparison of raised and Schottky source/drain MOSFETs using a novel tunneling contact model," *International Electron Device Meeting*, pp. 733–736, 1998.

[28] T. Inukai *et al.*, "Boosted Gate MOS (BGMOS): Device/Circuit Cooperation Scheme to Achieve Leakage-Free Giga-Scale Integration," *Custom Integrated Circuits Conference*, pp. 409–412, 2000.

[29] R. Joshi *et al.*, "6.6+ GHz Low Vmin, Read and Half Select Disturb-free 1.2 Mb SRAM," *Symposium on VLSI Circuits*, pp. 250–250, 2007.

[30] H. Kaul *et al.*, "A 300mV 494GOPS/W Reconfigurable Dual-Supply 4-Way SIMD Vector Processing Accelerator in 45nm CMOS," *International Solid-State Circuits Conference*, pp. 260–261, 2009.

[31] D. Kim *et al.*, "Low Power Circuit Design Based on Heterojunction Tunneling Transistors (HETTs)," *International Symposium on Low Power Electronics and Design*, pp. 219–224, 2009.

[32] D. Kim *et al.*, "A 1.85fW/bit Ultra Low Leakage 10T SRAM with Speed Compensation Scheme," *International Symposium on Circuits and systems*, pp. 69–72, 2011.

[33] D. Kim *et al.*, "Variation-Aware Static and Dynamic Writability Analysis for Voltage-Scaled Bit-Interleaved 8-T SRAMs," *International Symposium on Low Power Electronics and Design*, pp. 145–150, 2011.

[34] N. Kim *et al.*, "Drowsy Instruction Caches. Leakage Power Reduction using Dynamic Voltage Scaling and Cache Sub-Bank Prediction," *International Symposium on Microarchitecture*, pp. 219–230, 2002.

[35] T. Kim *et al.*, "A Voltage Scalable 0.26V, 64kb 8T SRAM with Vmin Lowering Techniques and Deep Sleep Mode," *Custom Integrated Circuits Conference*, pp. 407–410, 2000.

[36] T. Kim, J. Liu, J. Keane, C. Kim, "A 0.2 V, 480 kb Subthreshold SRAM With 1 k Cells Per Bitline for Ultra-Low-Voltage Computing," *Journal of Solid State Circuits*, Vol. 43, pp. 518–529, Feb. 2008.

[37] T. Kim, J. Liu, C. Kim, "A Voltage Scalable 0.26V, 64kb 8T SRAM with Vmin Lowering Techniques and Deep Sleep Mode," *Journal of Solid State Circuits*, Vol. 44, pp. 1785–1795, June 2009.

[38] J. Knoch, S. Mantl, J. Appenzeller, "Impact of the dimensionality on the performance of tunneling FETs: Bulk versus one-dimensional devices," *Solid-State Electronics*, Vol. 51, pp. 572–578, Apr. 2007.

[39] S.O. Koswatta, M.S. Lundstrom, D.E. Nikonov, "Comparison between p-i-n Tunneling Transistors and Conventional MOSFETs," *Transactions on Electron Devices*, Vol. 56, pp. 456–465, Mar. 2009.

[40] T. Krishnamohan, D. Kim, S. Raghunathan, K. Saraswat, "Double-Gate Strained-Ge Heterostructure Tunneling FET (TFET) With Record High Drive Currents and <60mV/dec Subthreshold Slope," *International Electron Device Meeting*, pp. 947–949, 2008.

[41] J. Lin *et al.*, “Compact HSPICE model for IMOS device,” *Electronics Letters*, Vol. 44, pp. 91–92, Jan. 2008.

[42] H. Mair *et al.*, “A 65-nm Mobile Multimedia Applications Processor with an Adaptive Power Management Scheme to Compensate for Variations,” *Symposium on VLSI Circuits*, pp. 224–225, 2007.

[43] F. Mayer *et al.*, “Impact of SOI, Si1-xGexOI and GeOI substrates on CMOS compatible Tunnel FET performance,” *International Electron Device Meeting*, pp. 163–166, 2008.

[44] K. Mistry, *et al.*, “A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging,” *International Electron Device Meeting*, pp. 247–250, 2007.

[45] T. Mizuno, J. Okamura, A. Toriumi, “Experimental Study of Threshold Voltage Fluctuation Due to Statistical Variation of Channel Dopant Number in MOSFET’s,” *Transactions on Electron Devices*, Vol. 41, pp. 2216–2221, Aug. 1994.

[46] H. S. Momose, *et al.*, “1.5nm direct-tunneling gate oxide Si MOSFET,” *Transactions on Electron Devices*, Vol. 43, pp. 1233–1242, Aug. 1996.

[47] G. E. Moore, “Cramming more Components onto Integrated Circuits,” *Electronics*, Vol. 38, pp. 114–117, Apr. 1965.

[48] G. E. Moore, “No Exponential is Forever: but “Forever” Can Be Delayed!,” *International Solid-State Circuits Conference*, pp. 20–23, 2003.

[49] S. Mutoh *et al.*, “1 -V Power Supply High-speed Digital Circuit Technology

with Multithreshold-Voltage CMOS," *Journal of Solid State Circuits*, Vol. 30, pp. 847–854, Aug. 1995.

[50] S. Narendra, *et al.*, "Scaling of Stack Effect and its Application for Leakage Reduction," *International Symposium on Low Power Electronics and Design*, pp. 195–200, 2001.

[51] R. Naseer, J. Draper, "Parallel Double Error Correcting Code Design to Mitigate Multi-Bit Upsets in SRAMs," *European Solid-States Circuits Conference*, pp. 222–225, 2008.

[52] O.M. Nayfeh *et al.*, "Design of Tunneling Field-Effect Transistors Using Strained-Silicon/Strained-Germanium Type-II Staggered Heterojunctions," *Electron Device Letters*, Vol. 29, pp. 1074–1077, Sep. 2008.

[53] H. Nho *et al.*, "A 32nm High-$\kappa$ Metal-Gate SRAM with Adaptive Dynamic-Stability Enhancement for Low-Voltage Operation," *International Solid-State Circuits Conference*, pp. 346–347, 2010.

[54] Y. Pu, J.P. de Gyvez, H. Corporaal, Y. Ha, "An Ultra-Low-Energy/Frame Multi-Standard JPEG Co-Processor in 65nm CMOS with Sub/Near-Threshold Power Supply," *International Solid-State Circuits Conference*, pp. 146–147, 2008.

[55] J. M. Rabaey, A. Chandrakasa, B. Nikolic, "Digital Integrated Circuits: A Design Perspective," *Second Edition, Prentice Hall*, 2009.

[56] A. Raychowdhury, X. Fong, Q. Chen, K. Roy, "Analysis of Super Cut-off Transistors for Ultralow Power Digital Logic Circuits," *International Symposium on Low Power Electronics and Design*, pp. 2–7, 2006.

[57] A. Raychowdhury *et al.*, "PVT-and-aging Adaptive Wordline Boosting for 8T SRAM Power Reduction," *International Solid-State Circuits Conference*, pp. 352–353, 2010.

[58] M. M. Rieger, P. Vogl, "Electronic-band parameters in strained Si1-xGex alloys on Si1-yGey substrates," *Physical Review B*, Vol. 48, pp. 14276–14287, Nov. 1993

[59] K. Roy, S, Mukhopadhyay, H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of The IEEE*, Vol. 91, pp. 305–327, Feb. 2003.

[60] A. Sedra, K. Smith, "Microelectronic Circuits," *Fourth Edition, Oxford University Press*, 1998.

[61] E. Seevinck, F.J. List, J. Lohstoh, "Static-Noise Margin Analysis of MOS SRAM Cells," *Journal of Solid State Circuits*, Vol. 22, pp. 748–754, Oct. 1987.

[62] M. Seok, D. Sylvester, D. Blaauw, "Optimal Technology Selection for Minimizing Energy and Variability in Low Voltage Applications," *International Symposium on Low Power Electronics and Design*, pp. 9–14, 2008.

[63] I. Sutherland, R. Sproull, D. Harris, "Logical Effort: Designing Fast CMOS Circuits," *Morgan Kaufmann*, 1999.

[64] Y. Taur, T. H. Ning, "Fundamentals of Modern VLSI Devices," *Second Edition, Cambridge University Press*, 2009.

[65] E.H. Toh *et al.*, "I-MOS Transistor With an Elevated Silicon-Germanium Impact-Ionization Region for Bandgap Engineering," *Electron Device Letters*, Vol. 27, pp. 975–977, Dec. 2006.

[66] S.O. Toh, Z. Guo, B. Nikolic, “Dynamic SRAM Stability Characterization in 45nm CMOS,” *Symposium on VLSI Circuits*, pp. 35–36, 2010.

[67] J. W. Tschanz *et al.*, “Dynamic Sleep Transistor and Body Bias for Active Leakage Power Control of Microprocessors,” *Journal of Solid State Circuits*, Vol. 38, pp. 1838–1845, Nov. 2003.

[68] N. Verma, A. P. Chandrakasan, “A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy,” *Journal of Solid State Circuits*, Vol. 43, pp. 141–149, Jan. 2008.

[69] J. Wang, S. Nalam, B. H. Calhoun, “Analyzing Static and Dynamic Write Margin for Nanometer SRAMs,” *International Symposium on Low Power Electronics and Design*, pp. 129–134, 2008.

[70] M. Wieckowski *et al.*, “A Black Box Method for Stability Analysis of Arbitrary SRAM Cell Structures,” *Design, Automation & Test in Europe*, pp. 795–800, 2010.

[71] M. Yuffe *et al.*, “A Fully Integrated Multi-CPU, GPU and Memory Controller 32nm Processor,” *International Solid-State Circuits Conference*, pp. 264–266, 2011.

[72] B. Zhai, D. Blaauw, D. Sylvester, K. Flautner, “Theoretical and Practical Limits of Dynamic Voltage Scaling,” *Design Automation Conference*, pp. 868–873, 2004.