Power and Area Efficient Sub-threshold 6T SRAM with Horizontal Local Bit-Lines and Bit-Interleaving

by

Sukneet Basuta, B.Eng.

A thesis submitted to the
Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Master of Applied Science in Electrical & Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
Department of Electronics
Carleton University
Ottawa, Ontario
January, 2015

©Copyright
Sukneet Basuta, 2015

Abstract

SRAM in the typical microprocessor consumes a substantial amount of on-chip area and significantly contributes to static power dissipation. Previous studies have shown that sub-threshold operation presents a minimum energy point that is optimal if ultra-low power consumption is desired. However, the standard 6T SRAM cell does not operate at sub-threshold voltages. Instead, designs with higher transistor counts are typically used for sub-threshold operation. These designs generally have low integration density.

This study presents a new SRAM architecture with minimum area that utilizes a modified 6T SRAM cell for sub- and near-threshold operation in ultra-low power applications. This new architecture introduces horizontal bit-lines, mitigates half-select disturb, and supports bit-interleaving. The proposed design's stability was thoroughly tested in the presence of process, temperature, and voltage variations, and compared to the standard 6T and traditional 8T cells. A 32kb SRAM block implementing the proposed architecture was designed, simulated, and contrasted to a traditional 8T SRAM cell block.

The simulated 32kb SRAM block operates at a maximum frequency of 544.8 khz and 6.70 Mhz for the read and write operations, respectively, and consumes 0.586 pJ/bit in the read operation and 0.17 pJ/bit in the write operation. A very similar 32kb SRAM block consisting of the traditional 8T SRAM cell was found to have a maximum frequency of 544.8 kHz and 2.89 Mhz for the read and write operations, respectively, and consumes 0.736 pJ/bit in the read operation and 0.205 pJ/bit in the write operation. The results show that the proposed design has lower power consumption than the 8T SRAM block, comparable read performance, and better write performance. This was all achieved while only having a 10% increase in area per bit over the conventional 6T thin-cell layout.

Acknowledgments

I would like to thank

Professor Maitham Shams, my thesis supervisor, for accepting me as a graduate student, his support, and guidance.

Professor Dimitrios Makrakis for his support at my brief time at the University of Ottawa.

Scott Bruce for helping with all my support problems and putting up with my incessant use of the Linux systems.

My family for their support.

All my friends for being my friends.

Lastly, I want to thank CMC Microsystems and TSMC for providing the design tools and technology kits used in this thesis.

Contents

Abstract	ii
Acknowledgments	iii
Table of Contents	iv
List of Tables	vii
List of Figures	viii
Nomenclature	xvii
1 Introduction	1
1.1 Motivation	1
1.2 Thesis Objectives	3
1.3 Thesis Organization	3
2 SRAM Design & Operation	5
2.1 6T SRAM Cell	5
2.1.1 Read Operation	6
2.1.2 Write Operation	7
2.2 Cell stability	8
2.2.1 Measuring SNM	13
2.2.2 Measuring SNM by Circuit Simulation	15
2.3 Low-power SRAM	18
2.4 Hierarchical bit-lines	21
2.5 Single-ended SRAM	23
2.6 Half-Select disturb	25

2.7	Bit-interleaving	25
2.8	Physical Design	26
3	Literature Review	31
3.1	A Variation-Tolerant Sub-200 mV 6-T Subthreshold SRAM	32
3.2	Differential 6T Sub-threshold SRAM with Low Energy and Variability Resilient Local Assist Circuit	33
3.3	Average-8T Differential-Sensing Subthreshold SRAM With Bit Inter- leaving and 1k Bits Per Bitline	36
3.4	A Single-Ended Disturb-Free 9T Subthreshold SRAM With Cross- Point Data-Aware Write Word-Line Structure, Negative Bit-Line, and Adaptive Read Operation Timing Tracing	40
3.5	Analysis of Past Research	43
4	Proposed SRAM Architecture	45
4.1	Write Operation	47
4.2	Read Operation	53
4.3	Hold Operation	66
4.4	Cell Layout	66
5	Static Noise Margin Evaluations	70
5.1	Process Variations	70
5.1.1	Hold Operation	71
5.1.2	Write Operation	72
5.1.3	Read Operation	76
5.2	Temperature Variations	80
5.2.1	Hold SNM	81
5.2.2	Read SNM	89
5.2.3	Write SNM	97
6	SRAM Block Simulations	106
6.1	SRAM Block	106
6.2	SRAM Block Operation	109
6.3	Comparison to 8T SRAM Block	115
6.3.1	Write Operation	118
6.3.2	Read Operation	122

6.3.3	Idle State	123
6.3.4	Frequency of Operation	126
6.4	Comparison at Various Supply Voltages	127
6.4.1	Frequency of Operation	128
7	Conclusion	131
7.1	Future Work	132
	List of References	133
	Appendix A Static Noise Margin Simulation Results	138
A.1	Hold Operation	138
A.2	Write Operation	144
A.3	Read Operation	151
	Appendix B Copyright Permissions	156
B.1	Intel Copyright Permission	156
B.2	IEEE Copyright Permission	156

List of Tables

6.1	Byte Layout	107
6.2	Summary of comparison between the Proposed SRAM block and the 8T SRAM block.	127
A.1	1-cell Configuration	141
A.2	2-cell Configuration	141
A.3	4-cell Configuration	142
A.4	8-cell Configuration	142
A.5	Traditional 8T	142
A.6	Standard 6T	143
A.7	1-cell Configuration	144
A.8	2-cell Configuration	147
A.9	4-cell Configuration	148
A.10	8-cell Configuration	148
A.11	Traditional 8T	149
A.12	Traditional 8T-SE	149
A.13	Standard 6T	150
A.14	1-cell Configuration	152
A.15	2-cell Configuration	153
A.16	4-cell Configuration	154
A.17	8-cell Configuration	154
A.18	Traditional 8T	154
A.19	Standard 6T	155

List of Figures

1.1	Processor die map of the Intel Sandy Bridge microarchitecture [1]. ©2011 Intel Corporation	2
2.1	Standard 6T SRAM cell.	5
2.2	Standard 6T SRAM cell read operation.	6
2.3	Standard 6T SRAM cell read operation timing diagram.	7
2.4	Standard 6T SRAM cell write operation.	8
2.5	Standard 6T SRAM cell write operation timing diagram.	8
2.6	Butterfly diagram for the hold static noise margin (HSNM) of a stan- dard 6T SRAM cell.	9
2.7	Butterfly diagrams for 6T SRAM.	10
2.8	Test circuit for measuring the hold static noise margin (HSNM). . . .	11
2.9	WSNM butterfly diagrams for 6T SRAM.	12
2.10	Test circuit for measuring the read static noise margin (RSNM). . . .	13
2.11	Butterfly diagram for the read static noise margin (RSNM) for a stan- dard 6T SRAM cell [2].	13
2.12	Test circuit for measuring the write static noise margin (WSNM). . .	14
2.13	Estimation of SNM in a 45° rotated coordinate system [3].	15
2.14	Circuit implementation of (2.2).	16
2.15	Circuit implementation of (2.4).	17
2.16	Butterfly (SNM) curves for the hold operation transposed onto the 45° rotated coordinate system (u,v). The dashed curve is v_1 and the solid curve is v_2 . V_{DD} is 1.0 V in this case, so u was swept from -707 mV to 707 mV.	17
2.17	The difference between the two butterfly (SNM) curves ($v_1 - v_2$) in Figure 2.16.	18
2.18	Read SNM comparision for 6T SRAM cell at different V_{DD}	19
2.19	The traditional 8T SRAM cell.	21

2.20	Classical SRAM array architecture [4].	22
2.21	Hierarchical or divided bit-line architecture [4].	23
2.22	Simplest single-ended SRAM cell architecture [5].	24
2.23	Single-ended SRAM bit-cell with different read and write paths [6]. ©2010 IEEE	25
2.24	Half select disturb problem in a standard 8T SRAM cell array [7]. ©2010 IEEE	26
2.25	SRAM word organization: the top row shows the typical case where all bits in a word are stored adjacently, unlike the bottom row that shows an implementation of bit-interleaving [8]. ©2011 IEEE	27
2.26	Historical layout of 6T SRAM cell [2].	28
2.27	Arrayed historical layout of 6T SRAM cell [2].	29
2.28	Lithographically friendly thin cell 6T cell layout [2].	29
2.29	Thin cell 6T cell layout showing rounding of the notch [9]. ©2009 IEEE	30
2.30	Diffusion-notch-free, or rectangular-diffusion, layout of 6T SRAM cell. Note how all transistors are sized equally [9]. ©2009 IEEE	30
3.1	SRAM Cell design and area [10]. ©2008 IEEE	33
3.2	Cell array structure [10]. ©2008 IEEE	34
3.3	Chiou <i>et al.</i> 's proposed SRAM architecture [11]. ©2013 IEEE	35
3.4	Khayatzadeh and Lian's proposed SRAM cell [12]. ©2013 IEEE	36
3.5	Khayatzadeh and Lian's proposed SRAM cell in the hold state [12]. ©2013 IEEE	37
3.6	Khayatzadeh and Lian's proposed SRAM cell in the read operation [12]. ©2013 IEEE	38
3.7	Khayatzadeh and Lian's proposed SRAM cell when half-selected [12]. ©2013 IEEE	38
3.8	Khayatzadeh and Lian's proposed SRAM cell in the write operation [12]. ©2013 IEEE	39
3.9	SRAM cell proposed in [13]. ©2012 IEEE	40
3.10	Timing diagram of the 9T SRAM cell designed in [13]. ©2012 IEEE	41
3.11	Write operation of the 9T SRAM cell presented in [13]. ©2012 IEEE	41
3.12	Half-selected rows during the write operation of the 9T SRAM cell presented in [13]. ©2012 IEEE	42
4.1	Basic architecture of the proposed cell block in a 2-cell configuration.	45

4.2	Schematic of the proposed cell block in a 1-cell configuration.	46
4.3	Schematic of the proposed cell block in a 2-cell configuration.	46
4.4	Schematic of the proposed cell block in a 4-cell configuration.	47
4.5	Schematic of the proposed cell block in a 8-cell configuration.	47
4.6	Write operation of the proposed SRAM design.	48
4.7	Timing diagram of the write operation when writing a ‘0’.	49
4.8	Timing diagram of the write operation when writing a ‘1’.	50
4.9	Comparison of the write SNM of the proposed design, the differential traditional 8T/standard 6T (they are same in this case), and single-ended 8T write schemes. The V_{DD} used in the simulation is 300 mV.	51
4.10	Half-select disturb in the proposed design.	52
4.11	Read operation of the proposed design when reading a ‘1’.	53
4.12	Read operation of the proposed design when reading a ‘0’.	54
4.13	Bit-line drainage current when a single minimum-size transistor, a single oversized transistor (2 times the minimum width), and a pair of minimum-size parallel stacked transistors (PTS [14]) are used for read-decoupling. Simulation is performed for the FS corner.	55
4.14	Timing diagram of the read operation when reading a ‘1’.	56
4.15	Timing diagram of the read operation when reading a ‘0’.	57
4.16	Drain current vs. width for an NMOS transistor. The inverse narrow width effect can be observed by comparing the drain current of a minimum width transistor with one twice the size (marked on the figure).	58
4.17	Comparison of the read current when a single minimum-size transistor, and a pair of minimum-size parallel stacked transistors (PTS) are used as read decoupling transistors when reading a ‘1’. The read current of the traditional 8T cell is included for comparison purposes. Simulation was performed for the FS corner.	59
4.18	Comparison of the local bit-line voltage when a single minimum-size transistor, and a pair of minimum-size parallel stacked transistors (PTS) are used as read decoupling transistors when reading a ‘1’. The voltage level of the data being read (Node Q) is included for comparison purposes. Simulation was performed for the FS corner.	60

4.19	Comparison of the bit-line leakage for a combination of parallel read-decoupling transistors and stacked read-access transistors, a single read-decoupling transistor and stacked read-access transistors, and parallel read-decoupling transistors and a single read-access transistor. The bit-line leakage for the read path in the 8T SRAM cell is included for comparison. The simulation is performed for the FS corner.	61
4.20	NMOS transistors stacked in series.	62
4.21	Comparison of the sub-threshold current between two minimum-length series stacked NMOS transistors and one NMOS transistor with a length of two times the minimum-length. V_{DD} is 300 mV.	63
4.22	Comparison of the read SNM of the proposed design, the traditional 8T, and standard 6T SRAM cells. The V_{DD} used in the simulation is 300 mV.	65
4.23	Idle operation of the proposed design.	66
4.24	Comparison of the hold SNM of the proposed design, the traditional 8T, and standard 6T SRAM cells for a V_{DD} of 300 mV.	67
4.25	Layout of the proposed bit cell. The layout is based on the layout of the 6T thin-cell.	68
4.26	Layout for the overhead circuitry required for the proposed cell block, including read and write decoupling transistors, and read access transistors.	68
4.27	A section of the proposed SRAM architecture. The layout here shows 2 bit-cells with the overhead circuitry. Any additional cells would be added to the sides.	69
5.1	Simulation results for Hold SNM vs. V_{DD} for the different block configurations proposed, the traditional 8T cell, and the standard 6T cell. The plotted points are $\mu - 5\sigma$. The points in the table are in mV. . .	72

5.2	Estimated probably density function of the hold operation for the 8-cell block configuration and 8T/6T SRAM cells (they are the exact same). V_{DD} is 300mV and the temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. This is similarly true for all sub-threshold operating voltages. Refer to Figure A.1, Figure A.2, and Figure A.3 for the histograms of the 8-cell block configuration, 8T SRAM cell, and 6T SRAM cell, respectively.	73
5.3	Simulation results for Write SNM vs. V_{DD} for the different block configurations proposed, the traditional 8T cell, the traditional 8T cell with single-ended write operation, and the standard 6T cell. The plotted points are $\mu - 5\sigma$. The points in the table are in mV.	74
5.4	Estimated probably density function of the hold operation for the 8-cell block configuration at various supply voltages. The temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. Refer to Figure A.4, Figure A.5, Figure A.6, and Figure A.7 for the histograms of the 8-cell configuration with supply voltages of 200 mV, 250 mV, 300 mV, and 350 mV, respectively.	76
5.5	Estimated probably density function of the write operation for the 8T/6T SRAM cell (they are the same), and 8T single-ended write scheme. V_{DD} is 300mV and the temperature is 27°C. Only the the density function when the V_{DD} is 300mV included because all supply voltages have very similar distributions. Refer to Figure A.8, and Figure A.9 for the histograms of the single-ended 8T, and 8T/6T SRAM cells, respectively.	77
5.6	Simulation results for Read SNM vs. V_{DD} for the different block configurations proposed, the traditional 8T cell, and the standard 6T cell. The plotted points are $\mu - 5\sigma$. The points in the table are in mV. . .	78

5.7	Estimated probably density function of the read operation for the 8-cell block configuration, 8T, and 6T SRAM cells. V_{DD} is 300mV and the temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. This is similarly true for all sub-threshold operating voltages. Refer to Figure A.10, Figure A.11, and Figure A.12 for the histograms of the 8-cell block configuration, 8T SRAM cell, and 6T SRAM cell, respectively.	79
5.8	Hold SNM vs. Temperature (°C) for the 1-cell block.	82
5.9	Hold SNM vs. Temperature (°C) for the 2-cell block.	83
5.10	Hold SNM vs. Temperature (°C) for the 4-cell block.	84
5.11	Hold SNM vs. Temperature (°C) for the 8-cell block.	85
5.12	Hold SNM vs. Temperature (°C) for the traditional 8T SRAM cell. .	86
5.13	Hold SNM vs. Temperature (°C) for the standard 6T SRAM cell. . .	87
5.14	Simulation results for Hold SNM vs. Temperature (°C) for the different block configurations proposed, the traditional 8T cell, and the standard 6T cell. V_{DD} is 300mV. The plotted points are $\mu - 5\sigma$	88
5.15	Read SNM vs. Temperature (°C) for the 1-cell block.	90
5.16	Read SNM vs. Temperature (°C) for the 2-cell block.	91
5.17	Read SNM vs. Temperature (°C) for the 4-cell block.	92
5.18	Read SNM vs. Temperature (°C) for the 8-cell block.	93
5.19	Read SNM vs. Temperature (°C) for the traditional 8T SRAM cell. .	94
5.20	Read SNM vs. Temperature (°C) for the standard 6T SRAM cell. Note that the y-axis has been translated down 20mV to show the standard deviation without changing the scale.	95
5.21	Simulation results for Read SNM vs. Temperature (°C) for the different block configurations proposed, the traditional 8T cell, and the standard 6T cell. V_{DD} is 300mV. The plotted points are $\mu - 5\sigma$	96
5.22	Write SNM vs. Temperature (°C) for the 1-cell block.	98
5.23	Write SNM vs. Temperature (°C) for the 2-cell block.	99
5.24	Write SNM vs. Temperature (°C) for the 4-cell block.	100
5.25	Write SNM vs. Temperature (°C) for the 8-cell block.	101
5.26	Write SNM vs. Temperature (°C) for the traditional 8T SRAM cell. .	102
5.27	Write SNM vs. Temperature (°C) for the single-ended write scheme utilizing the traditional 8T SRAM cell.	103

5.28	Write SNM vs. Temperature ($^{\circ}\text{C}$) for the traditional 8T SRAM cell. .	104
5.29	Simulation results for Write SNM vs. Temperature ($^{\circ}\text{C}$) for the different block configurations proposed, the traditional 8T cell, the single-ended 8T write-scheme, and the standard 6T cell. V_{DD} is 300mV. The plotted points are $\mu - 5\sigma$	105
6.1	Array level block diagram of the 32 kb memory block for the proposed SRAM architecture.	108
6.2	Array level block diagram of the 32kb memory block for the traditional 8T SRAM cell.	109
6.3	Simulated signals involved in the write operation of the 8T block for FS corner at 27°C	110
6.4	The problem with the write operation in the 8T block. Simulation performed for FS corner at 27°C	112
6.5	The signals involved the write operation in the proposed block. Simulation performed for FS corner at 27°C	113
6.6	The signals involved the read operation in the 8T block. Simulation performed for FS corner at 27°C . Note that the reason for the spike of the RWL signal in the beginning of the read operation is due to the slowness of the decoder.	114
6.7	The signals involved the read operation in the proposed block. Simulation performed for FS corner at 27°C . Note that the spikes in the signals in the beginning of the idle state is due to changing the signals at the same time as the clock switches low. It takes time for the decoder to process the low clock state.	116
6.8	Comparison of the power consumption during the start-up, write, idle, and read operations.	117
6.9	Comparison of the power consumption during a write. Close up view of Figure 6.8.	118
6.10	Comparison of the energy consumed during a write.	119
6.11	More accurate comparison of the power consumption during a write. .	120
6.12	More accurate comparison of the energy consumption during a write.	121
6.13	Comparison of the power consumption during a read. Close up view of Figure 6.8.	123
6.14	Comparison of the energy consumed during a read.	124

6.15	Comparison of the power consumption during the idle state, i.e. static power consumption. Close up view of Figure 6.8.	125
6.16	Comparison of the energy consumed during the idle state.	126
6.17	Comparison of the power consumption during the write, read, and idle operations. Simulations performed for FS corner at 27°C.	128
6.18	Comparison of the energy per bit consumed during the write and read operations. Simulations performed for FS corner at 27°C.	129
6.19	Comparison of the maximum write and read speed. Simulations performed for FS corner at 27°C.	129
6.20	Comparison of the maximum write and read speed not including the decoder delay. Simulations performed for FS corner at 27°C.	130
A.1	Histogram for the hold operation of the 8-cell block configuration. V_{DD} is 300mV and the temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. This is similarly true for all sub-threshold operating voltages.	138
A.2	Histogram for the hold operation of the traditional 8T cell. V_{DD} is 300mV and the temperature is 27°C.	139
A.3	Histogram for the hold operation of the standard 6T cell. V_{DD} is 300mV and the temperature is 27°C.	140
A.4	Histogram for the write operation of the 8-cell block configuration. V_{DD} is 200mV and the temperature is 27°C.	144
A.5	Histogram for the write operation of the 8-cell block configuration. V_{DD} is 250mV and the temperature is 27°C.	145
A.6	Histogram for the write operation of the 8-cell block configuration. V_{DD} is 300mV and the temperature is 27°C.	145
A.7	Histogram for the write operation of the 8-cell block configuration. V_{DD} is 350mV and the temperature is 27°C.	146
A.8	Histogram for the single-ended write operation of the traditional 8T cell. V_{DD} is 300mV and the temperature is 27°C.	146
A.9	Histogram for the write operation of the traditional 8T/standard 6T cell. V_{DD} is 300mV and the temperature is 27°C.	147

A.10 Histogram for the read operation of the 8-cell block configuration. V_{DD} is 300 mV and the temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. This is similarly true for all sub-threshold operating voltages.	151
A.11 Histogram for the read operation of the traditional 8T cell. V_{DD} is 300mV and the temperature is 27°C.	152
A.12 Histogram for the read operation of the standard 6T cell. V_{DD} is 300 mV and the temperature is 27°C.	153

Nomenclature

List of Acronyms

Acronym	Definition
BL	Bit-line
$\overline{BL}$ /BLB	Bit-line compliment
CLK	Clock
CMOS	Complementary metal-oxide-semiconductor
FS	Fast NMOS , Slow PMOS
GBL	Global bit-line
GND	Ground
HSNM	Hold static noise margin
HVT	High threshold voltage
LBL	Local bit-line
LP	Low-power transistor variant
MOSFET	Metal-oxide-semiconductor field-effect transistor
NMOS	n-channel MOSFET
PMOS	p-channel MOSFET
PTS	Parallel stacked transistors
PVT	Process-voltage-temperature
RBL	Read bit-line

RSNM	Read static noise margin
RWL	Read word-line
SNM	Static noise margin
SRAM	Static Random Access Memory
TSMC	Taiwan Semiconductor Manufacturing Company Limited
VLSI	Very-large-scale integration
VSS	Source supply
VVSS	Virtual source supply
WBL	Write bit-line
$\overline{WBL}$	Write bit-line compliment
WL	Word-line
WSNM	Write static noise margin
WLA	Word-line A
WLB	Word-line B
WWL	Write word-line

List of Symbols

Symbol	Definition
C	Capacitance
I	Current
I_{OFF}	Drain-source current when MOSFET is off
I_{ON}	Drain-source current when MOSFET is on
P	Power consumption
M_{ac}	Access transistors
M_{rd}	Read decoupling transistors
M_{rac}	Read access transistors
M_{wr}	Write decoupling transistors
M_{pd}	Block mask transistors
N	Number of
R	Resistance
V	Voltage
V_{DD}	Supply voltage
V_{th}	Threshold voltage
V_{tn}	NMOS threshold voltage

Chapter 1

Introduction

Embedded systems such as biomedical implants, and micro-sensor networks require ultra-low power circuits to extend battery life for as long as possible. It is widely regarded that the best way to reduce power in digital circuits is to lower the supply voltage; studies have shown that the minimum energy point is formed below the threshold voltage, in the sub-threshold region [15]. While ultra-low-voltage logic design has been well researched [16], there has been considerably fewer studies on sub-threshold SRAM design. Although there has been a number of papers published on sub-threshold SRAM in the past decade, sub-threshold SRAM design remains a challenge.

Scaling down the voltage permits the grasping of a minimum energy point, but adversely affects speed and leads to large process, voltage, and temperature (PVT) variations. Additionally, the increasingly poor I_{ON}/I_{OFF} ratio in the transistors as the supply voltage decreases only adds another obstacle to the challenge. With an increase in demand for ultra low-power embedded systems due to rapid advancements in CMOS technology, this problem needs to be addressed.

1.1 Motivation

In the past two decades, the 6T SRAM cell has been the standard for SRAM. It is, arguably, the paragon comprise of stability, performance, power consumption, and area utilization in super-threshold operation. When the voltage scales down to sub or near-threshold operation, however, the standard 6T cell has poor stability [17].

To circumvent this problem, manufacturers have been moving to the traditional 8T SRAM cell for low-voltage operation. The 8T SRAM addresses the decreased

stability, but it incurs an area penalty of at least 30%. This increase in area is a substantial issue since the on-die cache can consume a significant amount of on-chip area (see Figure 1.1). In their latest processors, Intel has moved to using the 8T cell for the fast L1 and L2 caches to increase performance at lower voltages while decreasing power consumption. However, the L3 cache is so large that utilizing the 8T cell would result in significant cost and area increase. Thus, to keep using the 6T SRAM cell in the L3 cache, the L3 cache and everything else off-core were given its own voltage plane and operates at a reduced speed. While this is not an example of an ultra-low power embedded system, the same problem applies.

Figure 1.1: Processor die map of the Intel Sandy Bridge microarchitecture [1].
©2011 Intel Corporation

Dynamic RAM (DRAM) may seem like a solution to the area problem, since the basic circuit is just one transistor with one capacitor and it's a volatile memory. Unfortunately, DRAM is generally slower than the all transistor SRAM and has to be refreshed periodically, resulting in a higher power consumption, making it unsuitable for use in caches. Non-volatile memory, such as ROMs and NVRAM, have very low power consumption, but do not have the speed to be suitable for caches and have a high area penalty. Their main purpose is for storing data when power is removed. Thus, there is no suitable replacement for SRAM.

A decent number of studies have been published in the past decade to solve the

ultra-low voltage SRAM problem, but very few focus on minimizing area. Instead, they tend to focus on achieving lower voltage operation, increased performance, and/or increased stability. These are all important issues that need to be solved for certain ultra low-voltage applications, but many other applications would benefit more from a less costly design. Thus, to find a solution to this problem, this study will try to design a sub-threshold SRAM bit cell that minimizes area while having acceptable performance and stability.

1.2 Thesis Objectives

Given the above motivation, the objectives of this research are as follows:

1. Propose a sub-threshold SRAM design that minimizes area.
2. Determine the performance and stability of the SRAM design with respect to voltage-scalability, process-variations, and temperature variations.
3. Compare the proposed design to the standard 6T and traditional 8T SRAM cells to determine its suitability as an alternative prospect at sub-threshold voltages.
4. Develop a SRAM block utilizing the proposed design to ensure proper operation.
5. Evaluate the constructed SRAM block, and compare it to a similar SRAM block consisting of the traditional 8T SRAM bit cell, with emphasis on power consumption.

1.3 Thesis Organization

Following this introductory chapter, Chapter 2 introduces the background knowledge needed to understand this study. The basic SRAM cell operation, cell stability, low-power SRAM, and SRAM design challenges are all tackled in this chapter.

Chapter 3 presents a review of previous studies that directly inspired this research. How each design functions, their advantages, and their problems are all summarized.

Chapter 4 introduces a new SRAM cell design. The reasoning for each aspect of the design is outlined and the operation of the design is discussed in detail.

Chapter 5 shows the simulation results of testing the stability of the proposed SRAM cell. The effects of process, temperature, and voltage variations are all explored to provide good insight on how stable the cell is.

In Chapter 6, an entire SRAM block is simulated, utilizing the proposed cell. Its performance, power, and energy consumption are compared to the traditional 8T SRAM block that is built using a very similar architecture. The metrics are then measured against various supply voltages.

Chapter 7 concludes this thesis and presents ideas for future work.

Chapter 2

SRAM Design & Operation

2.1 6T SRAM Cell

The basic requirements of a SRAM cell is to be able to read and write data, and to hold the data for as long as there is power. To meet these requirements, all memory cells consist of a storage cell and transfer (or access) gates. The storage cell holds the data. The transfer gates allow data to be read from and written to the storage cell. A basic flip-flop meets these demands but does so at a large size.

Figure 2.1: Standard 6T SRAM cell.

The standard 6T SRAM cell, illustrated in Figure 2.1, is considerably smaller than a flip-flop. This area reduction results in a more complex operation when reading and writing. However, given that SRAM dominates the area in modern microprocessors, it is a much welcomed trade-off.

In the standard 6T SRAM, the storage cell is comprised of a pair of weak inverters (M1 and M2, M3 and M4) placed back-to-back. The positive feedback created by cross-coupling the inverters corrects disturbances caused by noise and leakage. The NMOS transistors (M1 and M3) are known as the drive or pull-down transistors, and the PMOS transistors (M2 and M4) are known as the load or pull-up transistors. The transfer gates consist of transistors M5 and M6—the access transistors [18].

When the access transistors are off (i.e. WL is low), the SRAM cell is in the hold state. The back-to-back inverters keep the state of the cell through positive feedback.

2.1.1 Read Operation

Figure 2.2 shows an SRAM cell being read. Before the read operation occurs, the bit-lines (BL and $\overline{\text{BL}}$) are precharged high. When the read operation starts, the bit-lines are left floating high and the Word Line is raised high, enabling the access transistors. As a result, one access transistor will have no potential difference across it, while the other will have a potential difference of V_{DD} causing current to flow through it. In the case of Figure 2.2, this causes $\overline{\text{BL}}$ to be pulled down through the driver and access transistors. While $\overline{\text{BL}}$ is being discharged, it will be read as a logic ‘0’ by the peripheral circuitry. It is important to note that due to the large number of cells on each bit-line, there is a relatively large capacitance that must be discharged through the access and drive transistor. Thus, the bit-line is rarely fully discharged to a logic level ‘0’ [18].

Figure 2.2: Standard 6T SRAM cell read operation.

While $\overline{BL}$ is being pulled down, node $\overline{Q}$ tends to rise due to the current flowing in from M6, but is still held low by M3 (refer to Figure 2.3). Therefore, it is important to ensure the drive and access transistors are sized so that node $\overline{Q}$ remains below the switching threshold of the M1/M2 inverter. In many processes, similarly sized transistors are sufficient to ensure this. However, using equal sized transistors reduces the read margin [2].

Figure 2.3: Standard 6T SRAM cell read operation timing diagram.

2.1.2 Write Operation

Figure 2.4 shows an SRAM cell being written to. In the standard 6T SRAM cell, data is written into the cell by writing a ‘0’ into one node of the storage cell. The internal feedback of the cell sets the value of the other node to ‘1’. Due to the constraint of sizing the pull-down and access transistors so that Q (or $\overline{Q}$) remains below the switching threshold during reads, it is not possible to force a node to ‘1’ without feedback.

Once again, the bit-lines (BL and $\overline{BL}$) are precharged high before the operation begins. When the write operation starts, the bit-line connected to the node where ‘0’ is desired to be written is discharged. In Figure 2.4 it is desired to set Q to ‘0’, and hence $\overline{Q}$ to ‘1’. Thus, BL is discharged, causing Q to be discharged through the access transistor M5 when the Word Line (WL) is enabled (Refer to Figure 2.5). The PMOS transistor M2 opposes this operation by sending current into Q. Therefore, the load transistors (M2 and M4) must be weaker than the access transistors (M5

and M6) so that nodes Q and $\bar{Q}$ can be pulled low enough to change the state of the cell [18].

Figure 2.4: Standard 6T SRAM cell write operation.

Figure 2.5: Standard 6T SRAM cell write operation timing diagram.

2.2 Cell stability

The stability and writability of an SRAM cell is determined by the hold, read, and write margins. These margins are determined by the static noise margins of the cell in each mode of operation. The hold and read static noise margin (SNM) is a quantification of how much noise can be applied to the inputs of the storage cell

before a stable state is changed. The write static noise margin is a quantification of how much noise can be tolerated at the inputs of a storage cell before a second stable state is created.

Figure 2.6: Butterfly diagram for the hold static noise margin (HSNM) of a standard 6T SRAM cell.

Figure 2.6 is a butterfly diagram for a standard 6T SRAM cell during the hold state. This plot is a two-dimensional state space for a bistable circuit, such as a memory cell. Each point along the curve is a possible combination for the voltages on the bistable nodes (Q and $\overline{Q}$). Point A at the top left of the plot (0,1), where the two curves intersect, marks a stable solution for the system. There is a second stable solution at the bottom right (1,0), marked by point C, where the two curves intersect. In the middle (0.4,0.4), point B marks a metastable solution where the two curves intersect. Any amount of disturbance in the cell at this metastable state will quickly swing the cell to one of the stable states [19].

Figure 2.7(a) shows the best-case, or typical, SNM while in the hold state. The SNM is determined by the width of the largest square that can be placed between the curves. In cases where the butterfly diagram is not symmetric, and the high and low static noise margins are not equal, the SNM is the lesser of the two cases. The largest square shows the maximum amount of noise that can be applied before

Figure 2.7: Butterfly diagrams for 6T SRAM.

a stable state is lost, assuming equal but opposite noise sources are applied to both sides of the storage cell (refer to Figure 2.8). In the case of Figure 2.7, the worst-case noise is the maximum noise that can be applied until a stable point (point A) and the metastable point (point B) coincide. Any more noise in the system will yield only one intersection point (point C in this example). Figure 2.7(b) shows noise being applied to the storage cell; Points A and B are closer together, but there are still two stable points and one metastable point. Figure 2.7(c) shows the transfer characteristics when the maximum tolerated noise is applied. Note that the two voltage curves move in opposite directions of equal magnitude. Curve I has had -SNM Volts of noise applied at node V2 and Curve II has had +SNM Volts of noise applied at node V1 [3].

Figure 2.8: Test circuit for measuring the hold static noise margin (HSNM).

The butterfly curve while in the read state is very similar to the hold state. The read SNM determines how much noise can be applied to the storage cell, when reading, before the cell switches its state.

The butterfly curve while in the write state differs from the other two cases, as can be seen in Figure 2.9(a). The write SNM is determined by the width of the smallest square that can be placed between the two curves. When writing, only one stable state is desired to ensure writing the correct value (i.e. Point A in the figure). The static noise margin in this case determines how much noise can be applied to the storage cell during the write operation before a second state is created (Point B in Figure 2.9(b)). When a second state is created, it is not guaranteed which value will be written to the cell.

(a) Butterfly diagram for the best-case write SNM for a standard 6T SRAM cell.

(b) Butterfly diagram for the write mode where noise is applied to the storage cell in a standard 6T SRAM cell. The SNM square from the best-case SNM is included in the diagram for comparison. Curve I has had $+SNM$ V of noise applied at node V_2 and Curve II has had $-SNM$ V of noise applied at node V_1 .

Figure 2.9: WSNM butterfly diagrams for 6T SRAM.

2.2.1 Measuring SNM

Figure 2.8 shows the test circuit for measuring the hold static noise margin (HSNM). Noise is applied to the storage cell when the SRAM cell is in the holding state (i.e. the access transistors are all off). In Figure 2.8, V_n is the amount of noise voltage applied. The butterfly diagram that determines the SNM is the voltage transfer characteristics of the cross-coupled inverters. That is, the butterfly curve is generated by plotting the input voltage vs the output voltage of each inverter. The width of the largest square that can be placed between the two curves is the HSNM (see Figure 2.6).

Figure 2.10: Test circuit for measuring the read static noise margin (RSNM).

Figure 2.11: Butterfly diagram for the read static noise margin (RSNM) for a standard 6T SRAM cell [2].

The test circuit for determining the read static noise margin (RSNM) is depicted in Figure 2.10. Measuring the RSNM of the cell is done the same way as in the hold state, but with the SRAM cell in the read state. That is, by applying noise to the storage cell when the access transistors are on and the bit-lines are high. The access transistors end up pulling the node voltages up, distorting the voltage transfer characteristics and resulting in a lower SNM than in the hold state, as shown in Figure 2.11. The RSNM can be improved by decreasing the strength of the access transistors, and increasing the strength of the drive transistors. Access transistors can be weakened by making them smaller relative to the drive transistors, and/or by reducing the Word Line voltage relative to V_{DD} . The drive transistor can be made stronger by making them larger relative to the access transistors, increasing V_{DD} , and/or increasing V_{tn} .

Figure 2.12: Test circuit for measuring the write static noise margin (WSNM).

Obtaining the write static noise margin (WSNM) is done similarly to the previous two cases. The test circuit can be seen in Figure 2.12. To write to the SRAM cell, both access transistors are enabled while one bit-line is pre-charged high and the other is discharged to ‘0’. The smallest square that can be placed between the curves determines the WSNM. Conversely to improving the RSNM, the WSNM improves by increasing the strength of the access transistors, and by decreasing the strength of the drive transistors. The access transistors are strengthened by making them larger relative to the drive transistors, and/or by increasing the Word Line voltage. The drive transistors can be made weaker by making them smaller relative to the access transistors.

It’s important to note that there is a vast distribution of read, write, and hold static noise margins on each chip due to threshold voltage mismatches caused by random dopant fluctuations in the many cells. This is especially a problem in nanometre

processes [19]. Therefore, to obtain a good estimation of static noise margins, many simulations must be run in the presence of random process variations.

2.2.2 Measuring SNM by Circuit Simulation

Obtaining the SNM by drawing the largest square is a good graphical estimation, but a method to calculate the SNM is needed for running exhaustive simulations (i.e. Monte Carlo Simulations). Seevinck *et al.* created a method to do this [3].

Figure 2.13: Estimation of SNM in a 45° rotated coordinate system [3].

The method involves rotating the axis 45° and plotting the difference between the two curves (see Figure 2.13). This difference is the distance between the two curves. Thus, the maximum absolute difference (multiplied by $\cos 45^\circ$) is the SNM.

By assuming that the voltage transfer characteristic of one inverter is $y=F_1(x)$, and the other is $y=F_2'(x)$, which is the mirrored version of $y=F_2(x)$, the curves can be

transferred to the 45° rotated coordinate system (u,v) by using the following equations

$$\begin{aligned} x &= \frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v, \\ y &= -\frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v. \end{aligned} \quad (2.1)$$

Substituting $y = F_1(x)$ into (2.1) gives

$$v = u + \sqrt{2}F_1 \frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v \quad (2.2)$$

Since $F'_2(x)$ is the mirrored version of $F_2(x)$, to obtain it's transformation, $F_2(x)$ is first mirrored with respect to the v-axis and then transformed to the (u,v) coordinate system. This transformation results in the equations

$$\begin{aligned} x &= -\frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v, \\ y &= \frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v. \end{aligned} \quad (2.3)$$

Substituting $y = F_2(x)$ into (2.3) gives

$$v = -u + \sqrt{2}F_1 - \frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v \quad (2.4)$$

Figure 2.14: Circuit implementation of (2.2).

Equations (2.2) and (2.4) are mathematical representations of the inverters in an SRAM cell. Translating these equations into circuits results in Figure 2.14 and

Figure 2.15: Circuit implementation of (2.4).

Figure 2.15, respectively. By setting the Word-Lines and Bit-Lines to their respective values and sweeping u from $-V_{DD}/\sqrt{2}$ to $V_{DD}/\sqrt{2}$ the SNM of each operation can be obtained.

Figure 2.16: Butterfly (SNM) curves for the hold operation transposed onto the 45° rotated coordinate system (u,v) . The dashed curve is v_1 and the solid curve is v_2 . V_{DD} is 1.0 V in this case, so u was swept from -707 mV to 707 mV.

Figure 2.17: The difference between the two butterfly (SNM) curves ($v_1 - v_2$) in Figure 2.16.

Figure 2.16 shows the result of simulating the circuits in Figures 2.14 and 2.15 in the hold operation. The difference between the two solutions ($v_1 - v_2$) can be seen in Figure 2.17. The absolute values of the maximum and minimum are the diagonal lengths of the largest squares. Therefore, dividing the smaller value of the two by $\sqrt{2}$ gives the SNM of the cell [3].

2.3 Low-power SRAM

All SRAM cells have a minimum voltage at which they can reliably operate, since the SNM depends on V_{DD} . Figure 2.18 shows how the read SNM diminishes for a standard 6T SRAM cell as V_{DD} is decreased. Moreover, as the operating voltage decreases, process, voltage, and temperature variations become more apparent; This results in a greater variation in the SNM. When 6T SRAM cells are employed the minimum V_{DD} is typically between 0.7-1.0 V mainly due to the diminishing RSNM.

The minimum voltage of operation is around 0.7 V in a 90 nm process. As such, 6T SRAM cells are not suitable for sub-threshold operation [20].

Figure 2.18: Read SNM comparison for 6T SRAM cell at different V_{DD} .

A few techniques have been proposed to increase the read SNM of the standard 6T SRAM cell to allow for lower voltage operation. Some examples include lowering the word-line voltage in respect to V_{DD} [21], dynamically increasing cell V_{DD} during reads [22], utilizing a negative GND [23], and pulsing the word-line or bit-line (that takes advantage of the larger dynamic noise margin) [24]. However, these methods require additional circuitry and, generally, consume more energy. Lowering the word-line voltage increases the read SNM, but decreases the read speed.

Various techniques have also been proposed to increase the write SNM of SRAM cells, but, as mentioned, the RSNM is predominantly the bottleneck when lowering V_{DD} . Dynamically lowering the cell V_{DD} during writes [22], floating the cell ground

during writes [25], increasing the word-line voltage [12], and keeping a negative voltage on the bit-line [13] all improve the reliability when writing to the cell. These techniques all require additional circuitry, increase power consumption, and, in some cases, decrease the hold SNM of neighbouring cells.

While these assistance methods can help the standard 6T SRAM operate at a lower voltage, it still does not make it suitable for sub-threshold operation. An arguably better solution is to improve the SRAM cell.

The second most common SRAM cell, and the standard choice when operating at voltages below those supported by the standard 6T SRAM cell, is the traditional 8T SRAM cell in Figure 2.19. The 8T SRAM cell addresses the read SNM problem by isolating the read bit-line from the storage cell at the cost of around 30% increase in area [26].

The write operation of the 8T SRAM cell is the same as in the conventional 6T cell, but the bit-lines and word-lines used during a write are only used for the write operation. Exactly like in the 6T SRAM cell, the write bit-lines (WBL and $\overline{\text{WBL}}$) are precharged high before the operation begins. When the write operation starts, the write bit-line connected to the node where ‘0’ is desired to be written is discharged. When the write word line (WWL) is enabled, one side of the storage cell is selectively discharged through the discharged bit-line [26].

The read operation for the 8T SRAM cell has its own read word-line, a separate read bit-line, and, as a result, each cell has an additional access transistor (M8 in Figure 2.19) exclusively for reads. Before the read operation occurs, the read bit-line (RBL) is precharged high. To begin the read operation, the read word line (RWL) is raised high, enabling the read access transistor (M8). Depending on the value stored in the cell, the read decoupling transistor (M7) is selectively enabled. If the read decoupling transistor is enabled, the read bit-line is discharged through the read access and decoupling transistors. Thus, it will be read as a logic ‘0’ by the peripheral circuitry. If the read decoupling transistor is disabled, the read bit-line remains high and will be read as a logic ‘1’ by the peripheral circuitry [26].

Another key issue for sub-threshold SRAM is the relatively poor ratio of I_{ON} to I_{OFF} of the transistors. This poor ratio puts a limit on the number of cells that can be connected to a bit-line. With many cells connected to a bit-line, the total leakage current in a bit-line can exceed the I_{ON} current of a SRAM cell during reads. To overcome this problem without adding more transistors to limit leakage, most

Figure 2.19: The traditional 8T SRAM cell.

architectures implement hierarchical bit-lines.

2.4 Hierarchical bit-lines

The classical SRAM architecture is illustrated in Figure 2.20. The classical structure consists of an array of memory cells with 2^m rows and 2^n columns. Each SRAM cell in the same column share the same bit-lines.

When utilizing conventional 6T SRAM cells, the bit-line capacitance mostly consists of the diffusion junction capacitance of the access transistor connected to the bit-line and metal capacitance of the bit-line. Thus, the bit-line capacitance can be expressed as [27]

$$C_{BL} = C_{JunctionAccessTransistor} \times N_{CellsConnectedToBit-line} + C_{Metal}. \quad (2.5)$$

It is easy to see that reducing the number of cells on the bit-line can have a significant effect on the bit-line capacitance. The bit-line capacitance has a major effect on the RC delay and grows proportionally with each cell added to the bit-line. The bit-line capacitance also affects the power dissipation (P) linearly since

$$P = CV_{DD}V_{sw}f, \quad (2.6)$$

where f is the switching frequency, and V_{sw} is the voltage swing of the bit-line.

Figure 2.20: Classical SRAM array architecture [4].

A modified SRAM architecture is known as hierarchical or divided bit-lines. Hierarchical bit-lines subdivide the single large array into several smaller sub-arrays, reducing the number of SRAM cells attached to a bit-line. Each smaller sub-array is connected to a global bit-line and presents only a small load relative to all the cells in a single row of the classic architecture. This decreases the RC delay and power dissipation. Additionally, it is possible to only enable the required bit-lines in a sub-array, which can further reduce the overall power dissipation. Figure 2.21 shows an example of an hierarchical or divided bit-line architecture. In this architecture, global bit-lines run along the SRAM array columns and several sub-bit-lines connect a small group (4 SRAM cells in this case) of SRAM cells vertically. These sub-bit-lines are connected to the global bit-lines through pass transistors, reducing the number of transistors directly connected to the global bit-line.

Hierarchical bit-lines also help alleviate the problem of bit-line leakage by limiting the number of SRAM cells on each bit-line, without limiting the total number of cells in an SRAM block. There is a small area penalty that is incurred by utilizing

Figure 2.21: Hierarchical or divided bit-line architecture [4].

hierarchical bit-lines, but it can help to allow for sub-threshold operation (by reducing disturbances and the capacitance on a bit-line) and reduces power dissipation [4].

2.5 Single-ended SRAM

Using single-ended SRAM cells is a fairly common approach for reducing leakage current and power consumption. Traditional SRAM architectures contain multiple bit-lines with large capacitances in every column. Due to these large capacitances, charging and discharging the bit-lines results in a relatively significant power consumption. A single-ended SRAM architecture reduces the number of bit-lines to one per column, decreasing the active power consumption by at least half. Moreover, since there is only one bit-line from which leakage current can flow into a cell, there is also a significant reduction in the static power consumption. Figure 2.22 displays the

simplest single-ended SRAM cell architecture. It is basically the standard 6T SRAM cell with one access transistor and bit-line removed.

Figure 2.22: Simplest single-ended SRAM cell architecture [5].

Single-ended SRAM has its follies however. Typically, a single ended approach utilizes the same bit-line and a single access transistor to write a ‘0’ and a ‘1’ into the same node of the storage cell. As discussed previously, writing a ‘0’ into one side of the cross-coupled inverters is not an issue because of the strong NMOS drive (or pull-down) and access transistors. However, writing a ‘1’ requires the ability to overcome the strong NMOS drive transistors. A strong ‘1’ writability requires the use of write assist circuitry, such as weakening one side of the cross-coupled inverters by floating its ground, that comes at the cost of increased area and power consumption. Moreover, write assist circuitry generally has the unintended consequence of lowering the hold static noise margin of half-selected cells [21].

A single-ended cell architecture similar to that in Figure 2.22, where there is a single access transistor connected to a large capacitance bit-line, has a read SNM similar to that in the standard 6T case. As such, it is unsuitable for low-power operation. Many single-ended SRAM bit-cell architectures, however, provide separate read and write paths while utilizing a single bit-line. This allows for optimized paths for each operation to increase the static noise margins for each individual operation. Figure 2.23 shows an example of a single-ended SRAM bit-cell with different read and write paths.

Figure 2.23: Single-ended SRAM bit-cell with different read and write paths [6].
©2010 IEEE

2.6 Half-Select disturb

An issue that becomes more apparent in low-voltage and sub-threshold SRAM architectures is half-select disturb. Half-select disturb occurs in a cell when the word-line for the row is selected, but the column is unselected, leaving the bit-line floating high. This is perhaps better illustrated by Figure 2.24. The disturbed cells in the disturbed columns have their access transistors enabled for the write operation and their bit-lines are left floating high. Therefore, it is possible for noise in the bit-lines to flip the stored data in the disturbed cells. This presents a compromise between write ability and the disturb margin. The current flowing through the access transistors needs to be large enough to flip the stored data in the storage cell, but not too large that the stored data flips when the cell is disturbed [7].

2.7 Bit-interleaving

As SRAM cells enter sub-threshold operation, there is an increase in soft-error rates. Soft errors are incorrect signals or data that result from the radiation of energetic particles, thermal neutrons, and random noise. Soft errors are often undetectable.

Figure 2.24: Half select disturb problem in a standard 8T SRAM cell array [7].
 ©2010 IEEE

Generally, when soft errors occur, adjacent bits are also affected [28]. When only one bit in a word is affected, Error-correcting codes (ECC) can recover the affected bit. However, it is difficult to implement on-chip ECC that can recover multiple-bits.

Typically, all bits in a word are stored adjacent to each other in one contiguous row. This makes multiple bits in a word targets for unrecoverable soft errors. However, bit-interleaving can be implemented to separate the bits in a word, so when soft errors do occur, only one bit in a word is affected. As its name implies, this technique interleaves the bits of different words. Figure 2.25 shows an example of bit-interleaving.

In the figure, each bit in the word is far away from each other. Thus, if for example, the first 4 bits in the row are affected by a soft error, only the first bits of the first 4 words are affected, and can be recovered by ECC [8].

2.8 Physical Design

As discussed in the introduction, SRAM cells often dominate the area of modern chips. Hence, it is desirable to minimize the area of each cell to improve integration density.

The ideal layout of an SRAM cell is a square to balance the capacitive load on

Figure 2.25: SRAM word organization: the top row shows the typical case where all bits in a word are stored adjacently, unlike the bottom row that shows an implementation of bit-interleaving [8]. ©2011 IEEE

row and column lines, and balance the area overhead for the decoder and sense/drive circuits. Up until the 90 nm process generation, a layout similar to that seen in Figure 2.26 was used. This layout is one of the smallest possible layouts for the 6T cell. To minimize the area, cells are designed to be mirrored and overlapped to share V_{DD} , GND, word-lines, and bit-lines with adjacent cells, as shown in Figure 2.27.

Once the feature size gets below 90 nm, the bends in polysilicon and diffusions in the layout are challenging to fabricate. There is also an increase in process variations due to mask misalignments. To get around these issues, processes under 90 nm use the thin cell 6T cell layout shown in Figure 2.28. In this layout, all transistors are aligned in the same direction; all diffusions run in the vertical direction and all polysilicon in the horizontal direction. Since the cell is long and skinny, the bit-line capacitances, which run vertically, are reduced. The word-lines are longer as a result, but their capacitances are less critical.

An overly long or thin SRAM block results in excessive routing area, signal delay, and capacitance. Ideally, the entire SRAM circuit should have a shape close to a square. However, the aspect ratio of the cells, and number of rows and columns can make this difficult. Often, to address the non-square shape of the cells, SRAM

Figure 2.26: Historical layout of 6T SRAM cell [2].

blocks at these feature sizes will consist of more rows than columns to obtain a more square-like shape. This balances the capacitive loads and obtains a more efficient area layout [29].

The width of the drive (or pull-down) transistors shown in Figure 2.28 are larger than that of the load (or pull-up) transistors and access transistors. While this increases the read static noise margin, it also results in a diffusion notch between the access and drive transistors that tends to round out due to lithography limitations. This is illustrated in Figure 2.29. The rounding of the notch may change the effective width of the access transistor and increase variations in the static noise margin. Diffusion-notch-free (or rectangular-diffusion) cells have been proposed in order to mitigate this manufacturing defect. The layout of such cells is illustrated in Figure 2.30. In this layout, all six transistors in the cell have the same minimum widths and lengths. This reduces the read static noise margin along with the process variations in the cell [9].

Figure 2.27: Arrayed historical layout of 6T SRAM cell [2].

Figure 2.28: Lithographically friendly thin cell 6T cell layout [2].

Figure 2.29: Thin cell 6T cell layout showing rounding of the notch [9]. ©2009 IEEE

Figure 2.30: Diffusion-notch-free, or rectangular-diffusion, layout of 6T SRAM cell. Note how all transistors are sized equally [9]. ©2009 IEEE

Chapter 3

Literature Review

Many alternative architectures have been proposed to overcome the problems of the 6T SRAM cell at sub-threshold operation. Many studies have addressed the read stability issue by creating separate read and write paths, similar to the 8T SRAM cell. This allows for the optimization of each operation [6], [30], [31]. However, it practically always comes at the cost of additional transistors in the cell.

Other studies have focused on improving the traditional 8T bit cell by utilizing the reverse short-channel effect [32], or by implementing additional circuitry [25] to further increase stability and allow lower voltage operation. Some have even focused on entirely different cell structures by using dynamic threshold voltage MOSFETs (DTMOS) [33], [34], or Schmidt triggers [35].

Many other papers focus on improving the write ability of the SRAM cell. Most of these investigations focus on making the storage cell weaker or access devices stronger. Reducing/floating the supply, and increasing/floating the ground voltage are common ways to make the cell weaker [25], [22], [36]. Breaking the feedback loop of the decoupled inverters has also been explored [37], [38]. However, both these methods affect the stability of half-selected cells. Making access devices stronger is a more promising approach. Upsizing access transistors ([30], [39]), and boosting the word-line voltage ([25] [12]) are good ways to achieve this, but the effect of upsizing transistors is lessened in sub-threshold operation, and boosting voltage requires additional circuitry.

Papers have also investigated ways to maintain the stability of half-selected cells. Some have prevented word-line sharing among cells in the same row by utilizing sub-rows [25]. Another research group has implemented a cross-point selection so there are no half-selected cells, but this requires two additional transistors for cell

selection [13], [40]. One method designed to overcome this problem is a write-back operation in which the unaccessed cell's data is written back into it at the expense of speed and power [31], [41].

Extensive research has also been done to reduce bit-line leakage. Some studies have utilized replica bit-lines [32]. Others have added additional transistors to the bit cell [20], [31]. Many studies look to virtual grounds to reduce leakage [25], [22]. A few others use negative word-line [38] and bit-line [13] voltages.

The common theme among all these advancements is that they all increase the transistor count and/or power consumption. This results in a substantial increase in cost. However, there have been a small number of studies that attempt to reduce the total transistor count. In the following, some of the important previous work relevant to the topic of this thesis are presented.

3.1 A Variation-Tolerant Sub-200 mV 6-T Sub-threshold SRAM

A few designs have adapted the standard 6T cell for use at sub-threshold voltages. One of the first of these designs was a single-ended 6T bit-cell proposed in [10]. This design, presented in Figure 3.1, employs a full transmission gate in the bit-cell to drive the bit-line from rail to tail. Hence, the need for a sense amplifier is eliminated—reducing the chance of variability problems common in differential designs. Since the architecture is single-ended, noise is isolated to one bit-line, making it more robust to read upsets compared to differential designs. To overcome the reduced write margins associated with single-ended architectures, V_{DD} and GND connected to the feedback inverter in the SRAM cell are drooped during a write. The combination of these design changes effectively decouples the read and write operations.

In order to overcome random dopant fluctuations causing V_{th} mismatch, the authors had to oversize the transistors. This increased the area to two times that of a standard 6T cell. They did note however that device sizes could be reduced if a less stringent supply voltage floor is required.

It was determined by Monte Carlo simulations that at 16 bit-cells per bit-line the read current of a single bit-cell is always greater than the cumulative leakage current from the unaccessed cells. As seen in Figure 3.2, adjustable strength header

Figure 3.1: SRAM Cell design and area [10]. ©2008 IEEE

and footer devices are used. When the *wr_en* signal is asserted, the strong PMOS transistor at the top and the strong NMOS transistor at the bottom are turned off. However, the weak headers and footers are still enabled, resulting in a voltage droop in VirVDD and VirGND that effectively disables the feedback inverter in the cell. This results in an increase in write SNM, but reduces the hold SNM of unaccessed cells that share the same VirVDD and VirGND. Rather than sharing the virtual GND and/or V_{DD} with the entire row, which is typically done, the virtual GND and V_{DD} are only shared by cells on the same local bit-line. While this increases the area, the hold SNM of unaccessed cells is only affected in cells on the same local bit-line rather than the entire row [10].

3.2 Differential 6T Sub-threshold SRAM with Low Energy and Variability Resilient Local Assist Circuit

The design in [10] was arguably improved upon in [11] by employing a differential bit-line. Driving the bit-line from rail to rail was a promising approach to eliminate the read disturb issue, but required large transistors. They determined that a differential

Figure 3.2: Cell array structure [10]. ©2008 IEEE

bit-line would allow them to keep the transistor sizes of the bit-cells as small as possible.

Figure 3.3 shows the proposed SRAM architecture presented in [11]. Minimum sized high threshold voltage (HVT) transistors for each standard 6T bit-cell were utilized to reduce the leakage current. Each local bit-line is comprised of 16 bit-cells and a local sense amplifier. To reduce the bit-line energy consumption, a charge-sharing based pre-charge scheme is used. When a cell in the column is not being accessed, the charges on bit-line (LBL) is shared with it's compliment (LBLB). This results in the bit-lines holding a charge somewhere between V_{DD} and GND. This also has the added advantage of reducing the read disturb voltage, and, thus, improving the read stability.

Typically, a differential bit-line cannot guarantee functionality at ultra-low voltages due to process variations [42]. To combat process variations, a transmission gate is used instead of the typical pass transistor to link between the local bit-line and

Figure 3.3: Chiou *et al.*'s proposed SRAM architecture [11]. ©2013 IEEE

global bit-line. The extra driving current suppresses the intra-die variations. However, this comes at the cost of additional area. With a sense amplifier required for every 16 bit-cells, the required area is vastly increased compared to a standard 6T SRAM architecture.

3.3 Average-8T Differential-Sensing Subthreshold SRAM With Bit Interleaving and 1k Bits Per Bitline

While the previous two studies succeeded in adapting the 6T SRAM cell for sub-threshold operation, they dramatically increased area consumption. The design proposed in [12] presents a sub-threshold differential SRAM architecture with a robust read scheme and minimum area. The SRAM architecture is illustrated in Figure 3.4.

Figure 3.4: Khayatzaheh and Lian's proposed SRAM cell [12]. ©2013 IEEE

The design proposes having cell blocks that vary in transistor count depending on how many bits are stored in each block (up to 16 bits). The fewer the number of bits, the more tolerable it is of lower voltages. This varies the average transistor count from 14 (1 bit per block) to 6.5 (16 bits per block). The study focuses on the average 8 transistor case since that is the closest to the conventional 8T architecture.

As shown in Figure 3.4 that the design utilizes the standard 6T architecture with localized bit-lines. These local bit-lines are decoupled from the global read and write

bit-lines by Mrd_{1-4} and Mwr_{1-2} respectively. In the idle/unselected state (word-lines are disabled for the block), the local bit-lines are left floating. This may turn on Mrd_{1-4} and possibly disturb the read operation of another block. To resolve this issue, block mask transistors (Mpd_{1-2}) are utilized to keep the local bit-lines low and minimize leakage. Figure 3.5 illustrates this unselected/idle state.

Figure 3.5: Khayatzadeh and Lian's proposed SRAM cell in the hold state [12].
©2013 IEEE

When in the idle/standby state, leakage and power can be further reduced by turning off the block mask transistors. This results in the local bit-lines being left floating. Additionally, the global read bit-lines are set to a high impedance mode by disabling the pull-up networks for the global read bit-lines.

While this design is similar to the standard 6T SRAM design with hierarchical bit-lines, it separates itself by having data-independent leakage, separate differential read and write bit-lines, and decoupled write operation. Data-independent leakage is a result of the block mask transistors keeping the local bit-lines low when unselected. This reduces the leakage during reads. Leakage is further reduced by the stacked read decoupling transistors.

While the storage cell is a standard 6T SRAM cell, the read operation is slightly different, as illustrated in Figure 3.6. Similar to the standard case, the read operation is done by raising the word-line connected to the access transistors of the desired cell. However, since the local bit-lines are not pre-charged high like in the standard case,

Figure 3.6: Khayatzadeh and Lian's proposed SRAM cell in the read operation [12].
©2013 IEEE

the stored data sets the values of the local bit-lines. The high local bit-line enables the read decoupling transistors, discharging one of the pre-charged high global read bit-lines. It's important to note that the value of the high local bit-line will not fully be V_{DD} since the access transistors are NMOS transistors. This can be mitigated by boosting the word-line voltage during reads, but at the cost of a more complex power supply and increased area consumption.

Figure 3.7: Khayatzadeh and Lian's proposed SRAM cell when half-selected [12].
©2013 IEEE

Figure 3.7 depicts the half-selected cells in the same row (access transistors are on). These half-selected cells are relatively stable since all pull-down transistors are off, and the local bit-line is very short. Thus, the disturbance caused by the pre-discharged local bit-line capacitance is relatively small.

The downside of this relatively complex read scheme is the that the timings of the BLK and word-line signals must be carefully designed to not overlap. If the access transistors are enabled before the block mask transistors are turned off, it is possible for the data in the storage cell to be overwritten.

Figure 3.8: Khayatzadeh and Lian's proposed SRAM cell in the write operation [12].

©2013 IEEE

The write operation is shown in Figure 3.8. The basic mechanism of the write operation is the same as in the standard 6T SRAM case. The word-line for the selected row is raised high, and, depending on the data to be written, one of the local bit-lines is discharged. However, since the write operation is single ended, the WSNM suffers compared to the standard 6T case. The half-selected case is the same as in the read operation; there is relatively little disturbance from the local bit-line capacitance.

3.4 A Single-Ended Disturb-Free 9T Subthreshold SRAM With Cross-Point Data-Aware Write Word-Line Structure, Negative Bit-Line, and Adaptive Read Operation Timing Tracing

A very interesting SRAM architecture is proposed in [13], as depicted in Figure 3.9. Rather than decreasing area, they aimed to improve read stability with a read buffer and eliminate half-select disturb during writes while utilizing a single bit-line to improve density and reduce BL power dissipation. Previous studies (e.g. [43], [28], [44], [40], [8]) have shown that write half-select disturb can be eliminated by utilizing a cross-point write structure. In a cross point structure, both row and column based word-lines are used in conjunction to select a cell. The word-line (WL) and virtual VSS (VVSS) are row-based, and the write word-lines (WWLA and WWLB) and bit-line (BL) are column-based. The timing diagram for each operation can be seen in Figure 3.10.

Figure 3.9: SRAM cell proposed in [13]. ©2012 IEEE

Reading is performed somewhat similar to the standard 8T SRAM cell case. The read word-line (WL) is set high and the write word-lines remain disabled, in addition to virtual VSS (VVSS) being forced to ground. Transistors M7-M9 buffer the data during reads to conditionally discharge the BL. Since Q and QB are isolated from the bit-line during reads, the read SNM is very close to the hold SNM. Additionally, unselected cells sharing the same bit-line have greatly reduced leakage. This is due

Figure 3.10: Timing diagram of the 9T SRAM cell designed in [13]. ©2012 IEEE

to the unselected cells in the same column having their VVSS held at V_{DD} and the stacking of the transistors in the read buffer.

Figure 3.11: Write operation of the 9T SRAM cell presented in [13]. ©2012 IEEE

One of the most innovative features of this design is its write operation. Figure 3.11 illustrates how the write operation works. When writing a '1' into the cell, WL and WWLA are brought high while VVSS and BL are forced to ground, and WWLB remains low. This causes QB to discharge by the bit-line (through M6 and M7) and by VVSS (through M8 and M9), setting QB to '0' and Q to '1'. When writing a '0', WL and WWLB are set high while VVSS and BL are forced to ground, and WWLA remains low. This causes Q to discharge by the bit-line (through M5 and M7).

It is important to note that the write ability is degraded by the series-connected

NMOS transistors (M5/M6 and M7). The authors alleviated this by utilizing a negative bit-line during writes. However, this requires additional area and a more complex power supply.

Figure 3.12: Half-selected rows during the write operation of the 9T SRAM cell presented in [13]. ©2012 IEEE

For a write to occur, WL and one of the write word-lines must be enabled to change the data in the cell. Since WL and VVSS are row-based and the write word-lines (WWLA and WWLB) are column-based, it is not possible change the value of a half-selected cell, as depicted in Figure 3.12. In half-selected cells within the active row, the access transistors are disabled. In half-selected cells within the active column, WL is disabled and VVSS is high (effectively eliminating the leakage current through M9 and M8). As a result, nodes Q and QB in half-selected cells in both the active row and column are isolated. Having half-select disturb-free writes allows for

a bit-interleaving architecture.

Most architectures that utilize read buffers require separate bit-lines for the read and write operations. However, this architecture is able to utilize the same bit-line for reads and writes. This allows for power savings as the number of bit-line charges and discharges are reduced.

Perhaps the most interesting aspect of this design is the vertical word-lines (WWLA and WWLB) connected to the gates of the access transistors. These vertical word-lines allow each side of the back-to-back inverters (Q and QB) to be pulled down to 0 using the same bit-line. In the typical case of horizontal word-lines, this is not possible since the bit-lines set the values of the cells and/or both access transistors would be enabled at the same time.

A big downside to this 9T SRAM design is the large area. There is a 97% area overhead over the 6T SRAM cell.

3.5 Analysis of Past Research

Previous studies have shown that a hierarchical bit-line approach utilizing the standard 6T SRAM cell is perhaps the ideal way to minimize area. However, the standard 6T SRAM requires differential bit-lines, which is arguably not ideal for subthreshold SRAM. A single ended approach while keeping the cell transistor count to 6T or lower, requires the ability to strongly write a ‘1’ by overcoming the stronger NMOS pull-down (or load) transistors (such as through the use of large transmission gates in [10]), or a scheme that allows the writing of a ‘0’ to each side of the storage cell.

The SRAM architecture presented in [13] provides a method to do the latter. By utilizing vertical word-lines and connecting the access transistors to the same bit-line, it is possible to discharge the value stored on each side of the storage cell—writing a ‘1’. However, their designed scheme requires a separate read path utilizing more transistors. The Designs in [10], [11], and [12] all feature hierarchical bit-lines with the same read and write paths, saving on area. Maybe combining the two schemes would present an ideal SRAM architecture.

However, utilizing vertical word-lines along with traditional vertical bit-lines presents a problem when selecting cells. Enabling a vertical word-line will select every cell in that column. During a write, the stored value in every selected cell is affected. During a read, the stored value in each selected cell affects the read value.

The design in [13] overcame this problem by having an additional horizontal word-line, so that a cell can only be selected if the horizontal and vertical word-lines are enabled. This requires an additional transistor for each cell, and does not allow for good readability; during the read operation, the data has to pass through two NMOS transistors. Perhaps overcoming this issue may conceivably present a SRAM design that has high integration density.

Chapter 4

Proposed SRAM Architecture

The chapter presents a novel SRAM architecture. Figure 4.1 shows the basic proposed SRAM in a 2-cell configuration. The SRAM cell is similar to the standard 6T SRAM cell, i.e. cross-coupled inverters are utilized for the storage cell along with 2 access transistors to retrieve or alter the data. However, a distinct difference from the standard 6T cell is that the gates of the access transistors are not connected together, and the select word-lines are vertical rather than the traditional horizontal. This completely changes the read and write operations.

Figure 4.1: Basic architecture of the proposed cell block in a 2-cell configuration.

Both of the access transistors are connected to the same horizontal local bit-line (LBL). The local bit-line is decoupled from the vertical global bit-line by the read-decoupling transistors, M_{rd1} and M_{rd2} , and the read access transistors, M_{rac1} and

M_{rac2} . The write operation does not utilize the global bit-line. Instead, the write operation is decoupled by M_{wr} . The word-lines for the the read and write operations are horizontal.

Figure 4.2: Schematic of the proposed cell block in a 1-cell configuration.

Figure 4.3: Schematic of the proposed cell block in a 2-cell configuration.

The number of transistors in each cell block varies depending on how many cells are in each block. Since the goal is to minimize area, it is desirable to maximize the number of bits in each the block. However, the minimum operational voltage

Figure 4.4: Schematic of the proposed cell block in a 4-cell configuration.

Figure 4.5: Schematic of the proposed cell block in a 8-cell configuration.

limits the number of bits that can be on the same bit-line due to bit-line leakage and relatively poor I_{ON} to I_{OFF} ratio. For clarity, this explanation will focus on a configuration of 2-cells in each cell block. The 1-cell configuration can be seen in Figure 4.2, 2-cell in Figure 4.3, 4-cell in Figure 4.4, and 8-cell in Figure 4.5.

4.1 Write Operation

The main rational for employing a horizontal local bit-line is for the write operation. As discussed in Section 2.1.2, data is written into standard SRAM cells by writing a ‘0’ to a node in the storage cell. In traditional architectures, this means separate access transistors for each node connected to different bit-lines. When used in combination with vertical word-lines, horizontal bit-lines allow both access transistors to utilize the same bit-line in its write path. The allows for a decoupled write operation while keeping the transistor count to a minimum. A decoupled write operation isolates the cell node from the long write bit-lines, decreasing the vulnerability of noise.

Due to the vertical bit-lines, the traditional way of selecting a cell cannot be used. The access transistors of the standard 6T cell is still used as a selection switch, but both access transistors cannot be enabled at the same time without contention. The

Figure 4.6: Write operation of the proposed SRAM design.

select word-lines connected to the access transistors (WLA and WLB) can analogously be compared to the write bit-lines, and the read and write word-lines (WWL and RWL) are basically identical to the read and write word-lines in the common 8T SRAM architecture. Since the write operation is completely decoupled, the only bit-line used in the write operation is the local bit-line (LBL).

Cell selection in the write operation is done by enabling the write word-line (WWL) for the row containing the cell, and the respective select word-line, WLA (word-line A) or WLB (word-line B), is enabled for the cell's column, depending on the value being written. Enabling these word lines turns on M_{wr} , and M_{ac1} or M_{ac2} . The selected node is then discharged through the enabled access transistor and M_{wr} .

Determining which select word-line is selected to write a '1' depends on which select word-line is used for reads. When using WLB for reads, selecting WLA during writes will write a '0' and WLB will write a '1'. When using WLA for reads, selecting WLA during writes will write a '1' and WLB will write a '0'. In the read configuration discussed in Section 4.2, WLB is used for reads. Thus, write operation discussed writes a '0' when WLA is selected, and a '1' when WLB is selected.

Figure 4.6 depicts writing a '0' into Q_1 . The access transistor, M_{ac1} , for Q_1 is enabled by $WLA0$, and the write decoupling transistor, M_{wr} , is enabled by WWL. Node Q is then discharged through M_{wr} . This results in node $\bar{Q}$ being forced to '1' through the feedback process by the cross-coupled inverters. The timing diagram for the operation of writing a '0' is illustrated in Figure 4.7. If writing a '1' is desired,

Figure 4.7: Timing diagram of the write operation when writing a '0'.

Figure 4.8: Timing diagram of the write operation when writing a '1'.

WLB0 for the cell should be enabled instead of WLA0. Otherwise, the write process is the exact same with Node $\overline{Q}$ being discharged through M_{wr} , and node Q being forced to '1'. The timing diagram for the operation of writing a '1' is shown in Figure 4.8.

Figure 4.9: Comparison of the write SNM of the proposed design, the differential traditional 8T/standard 6T (they are same in this case), and single-ended 8T write schemes. The V_{DD} used in the simulation is 300 mV.

The write static noise margin (WSNM) is comparable to that of a conventional 8T (or 6T) single-ended write scheme. This is illustrated in Figure 4.9. Compared to a traditional differential write scheme, where the write SNM is 147.1 mV, the write SNM is 27.2 mV worse (119 mV) because only the voltage at one node is being pulled

down. In a differential write scheme, one node is pulled up while the other is pulled down. While pulling a node down is what actually completes the write, pulling the other node up increases the SNM. Rarely is the write SNM the bottleneck of a design however.

The proposed architecture may not seem to support bit-interleaving due to the horizontal bit-lines, but it is actually the traditional word structure, where all bits in a word are adjacent, that is not supported. It is not possible to concurrently write into a cell that is directly adjacent (i.e. any cell that is within the same cell block). While this also makes it impossible to arbitrarily write a bit in any position, it does provide a solution to half-select disturb and, thus, supports a bit-interleaving structure. This bit-interleaving structure provides the ability to concurrently write into any one cell in each block in the same row. It also puts a minimum limit on the number of cell blocks in an SRAM block. There must be at least x cell blocks in a row, where x is the number of bits in a word, assuming the unit of address resolution is a word.

Since the word-lines connected to the access transistors of the cells are vertical, half-select disturb is not an issue for cells in the same row as they remain isolated. Instead, cells in the same column are susceptible to half-select disturb. This is shown in Figure 4.10. In this state, all the pull-down transistors connected to the local bit-line are OFF and a single access transistor is ON. The half-selected bit-cell (Q_1 in Figure 4.10) is exposed to the disturbances in the local bit-line (LBL), resulting in a decrease in stability.

Figure 4.10: Half-select disturb in the proposed design.

Arguably, column half-select disturb is a greater issue than row half-select disturb because, typically, there are more cells in a column than in a row (since the thin-cell layout of the standard 6T SRAM cell is longer horizontally than vertically). However, since the local bit-line is relatively short in contrast to conventional bit-lines, the disturbance caused by other cells on the bit-line is comparatively very small. Additionally, the stability of the half-selected bit cells are the same as the read cell stability (i.e. read SNM), which is somewhat close to the hold cell stability (i.e. hold SNM).

4.2 Read Operation

Cell selection for the read operation is done by enabling the read word-line (RWL) for the row containing the cell and WLB (word-line B) for the column containing the cell. These steps turn on one access transistor for the selected cell and the read access transistors (M_{rac1-2}). The global bit-line is then conditionally discharged depending on the data stored in the selected cell. As discussed in the previous section, it is inconsequential whether WLB or WLA is used for selecting the column as long the same word-line is always selected for the read operation. In this thesis, WLB will be used for reads.

Figure 4.11: Read operation of the proposed design when reading a '1'.

If the data stored in the cell at $\overline{Q}$ is a '1', the read decoupling transistors (M_{rd1}

and M_{rd2}) are turned on. This results in the pre-charged global bit-line to discharge through the read access and decoupling transistors. A full-swing large signal sense amplifier is then used to read the resultant value. Figure 4.11 shows the read operation when the read data is a '1'. The timing diagram when reading a '1' is shown in Figure 4.14. If the data stored in the cell at $\bar{Q}$ is a '0', the read decoupling transistors remain off, and the pre-charged global bit-line remains high. Figure 4.12 shows the read operation when the read data is a '0'. The timing diagram when reading a '0' is shown in Figure 4.15.

When reading a '1', the data stored in the cell at $\bar{Q}$ is a '1'. Thus, when the access transistor for the cell is enabled, the local bit-line obtains the value '1'. However, since NMOS transistors are used for the access transistors (which are required for strong writes), it is not possible for the local bit-line to attain a voltage level of V_{DD} . The local bit-line charges close to V_{DD} , but not fully. As a result, the read decoupling transistors (M_{rd1} and M_{rd2}) are not fully ON.

Figure 4.12: Read operation of the proposed design when reading a '0'.

A single read-decoupling transistor would slow down the read operation compared to the standard 8T cell. However, utilizing 2 minimum sized transistors in parallel further increases the current while keeping all transistors the same size to reduce process variations (as discussed in Section 2.8). Figure 4.13 shows the bit-line drainage current when a single minimum sized transistor, a single oversized transistor (2 times the minimum width), and a pair of minimum-size parallel stacked transistors (PTS

Figure 4.13: Bit-line drainage current when a single minimum-size transistor, a single oversized transistor (2 times the minimum width), and a pair of minimum-size parallel stacked transistors (PTS [14]) are used for read-decoupling. Simulation is performed for the FS corner.

Figure 4.14: Timing diagram of the read operation when reading a '1'.

Figure 4.15: Timing diagram of the read operation when reading a '0'.

[14]) are used for read-decoupling. Using a larger width for the read decoupling transistor actually results in a lower current. This phenomenon is explained by the inverse narrow width effect.

Figure 4.16: Drain current vs. width for an NMOS transistor. The inverse narrow width effect can be observed by comparing the drain current of a minimum width transistor with one twice the size (marked on the figure).

The inverse narrow width effect causes the threshold voltage, and consequently the drain current, to vary as the transistor width changes. This effect is caused by the fringing fields at the sharp corner in the shallow-trench isolation process that is used in the TSMC 65nm process. The fringing field at this corner leads to the formation of inversion charges, making it easier to form a conduction channel at a lower voltage. As the transistor width decreases, the fringing capacitance contributes more to the overall electric field, causing the threshold voltage to lower [45], [46]. This effect can

Figure 4.17: Comparison of the read current when a single minimum-size transistor, and a pair of minimum-size parallel stacked transistors (PTS) are used as read decoupling transistors when reading a ‘1’. The read current of the traditional 8T cell is included for comparison purposes. Simulation was performed for the FS corner.

be observed in Figure 4.16.

While parallel stacked read-decoupling transistors do increase the read current, it is important to remember that the read decoupling transistor is never fully ON. Consequently, the read current is never as strong as in the standard 8T cell, in fact

Figure 4.18: Comparison of the local bit-line voltage when a single minimum-size transistor, and a pair of minimum-size parallel stacked transistors (PTS) are used as read decoupling transistors when reading a ‘1’. The voltage level of the data being read (Node Q) is included for comparison purposes. Simulation was performed for the FS corner.

it is many times worse. As a result, the time it takes to discharge the global bit-line during reads is decently longer. Figure 4.17 compares the read current when a single minimum-size transistor, and a pair of minimum-size parallel stacked transistors are used as read decoupling transistors when reading a ‘1’. Figure 4.18 shows the local

bit-line (LBL) voltage for both cases, as well as the node voltage of the data being read in the SRAM cell.

Figure 4.19: Comparison of the bit-line leakage for a combination of parallel read-decoupling transistors and stacked read-access transistors, a single read-decoupling transistor and stacked read-access transistors, and parallel read-decoupling transistors and a single read-access transistor. The bit-line leakage for the read path in the 8T SRAM cell is included for comparison. The simulation is performed for the FS corner.

It is important to note that the read operation is also susceptible to half-select disturb. The state of the cell block is the same as in Figure 4.10; all the pull-down transistors connected to the local bit-line are OFF and a single access transistor is

ON. Again, the exposed bit-cell (Q_1 in Figure 4.10) has decreased stability and is subject to the disturbances in the local bit-line. However, this has an additional impact on the read operation. Since the value of LBL (the local bit-line) is set by $\overline{Q}$ in Q_1 , this exposure affects the amount of global bit-line leakage. In the worst case, LBL will be set to a logic ‘1’, enabling the read decoupling transistors—greatly increasing the leakage.

Figure 4.20: NMOS transistors stacked in series.

To reduce the global bit-line leakage, the read access transistors are stacked in series. The leakage current through a stack of series-connected transistors is greatly reduced when more than one transistor in the stack is OFF as a result of the stacking effect and is smaller than the leakage in a single transistor. An NMOS transistor stack is illustrated in Figure 4.20. The sub-threshold (and leakage) current can be found by the following equations:

$$\begin{aligned}
 I_{sub} &= I_0 e^{\frac{V_{GS} - V_{TH0} - \eta V_{DS} + \gamma V_{BS}}{nV_T}} \left(1 - e^{\frac{-V_{DS}}{V_T}}\right) \\
 I_0 &= \mu C_{OX} \frac{W}{L} (n-1) V_T^2 \\
 V_T &= \frac{kT}{q}
 \end{aligned} \tag{4.1}$$

where V_{GS} , V_{DS} , and V_{BS} are the gate-to-source, drain-to-source, and bulk-to-source voltages, respectively, V_{TH0} is the zero biased threshold voltage, η is the drain induced barrier lowering coefficient, n is the sub-threshold swing coefficient, V_T is the thermal voltage, γ is the body effect coefficient, μ is the carrier mobility, C_{OX} is the

gate oxide capacitance per unit area, W and L denotes the channel width and length, k is the Boltzmann constant, T is the absolute temperature, and q is the electrical charge of an electrons [47].

Equation 4.1 shows that doubling the channel length reduces the off-current by a factor of 2. However, due to the reverse short channel effect in modern deep sub-micron devices, the threshold voltage tends to decrease for longer channels, making leakage reduction less effective. Transistor stacks are a more promising approach for reducing leakage in these modern devices.

Figure 4.21: Comparison of the sub-threshold current between two minimum-length series stacked NMOS transistors and one NMOS transistor with a length of two times the minimum-length. V_{DD} is 300 mV.

The intermediate voltage (V_{int}) between two OFF stacked transistors is below V_{DD} , but remains positive due to a small drain current. Since the gates of the OFF NMOS transistors are connected to GND and V_{int} is positive, the gate-to-source voltage of the transistor at the top the stack (M_1 in Figure 4.20) will have a negative voltage of $-V_{int}$, reducing drain current. Similarly, the body-to-source voltage of M_1 is also $-V_{int}$, resulting in a larger body effect and, thus, an increase in threshold voltage compared to a single transistor. Furthermore, the drain-to-source voltage of M_1 is lower than the drain-to-source voltage of a single transistor [48]. Substituting the relevant voltages into Equation 4.1 for the stacked transistors and a single transistor

results in

$$I_{sub} = I_0 e^{\frac{-V_{int} - V_{TH0} + \eta(V_{DD} - V_{int}) - \gamma V_{int}}{nV_T}} (1 - e^{\frac{-(V_{DD} - V_{int})}{V_T}}), \text{ and} \quad (4.2)$$

$$I_{sub} = I_0 e^{\frac{-V_{TH0} + \eta V_{DD}}{nV_T}} (1 - e^{\frac{-V_{DD}}{V_T}}),$$

respectively. Since $V_{int} \gg \eta(V_{DD})$, these equations show that the series stacked transistor will have lower sub-threshold current [47]. This can be further observed in Figure 4.21.

Using series stacked transistors for the access transistors allows for the number of cell blocks connected to the global bit-line to be comparable to the number of bit-cells connected to the read bit-line in the standard 8T architecture. In the worst case corner, i.e. FS (Fast NMOS, slow PMOS), the preeminent bit-line leakage is less than the paramount bit-line leakage in the standard 8T architecture, as shown in Figure 4.19. The FS corner is the worst case corner because the read access and decoupling transistors are NMOS and the pull-up transistors are PMOS. Stacked read-access transistors reduces the leakage current by close to 3 times. Utilizing parallel stacked read-decoupling transistors in conjunction with the stacked read-access transistors does increase the leakage slightly when they are OFF, but the increase in ON current is, arguably, worth the tradeoff depending on the design goal.

The horizontal local bit-lines do not only help with the write operation, they also support the read operation. Compared to the long global bit-lines, the short local bit-lines have the advantage of having a lower cumulative bit-line leakage due to a drastic reduction in the number of cells connected to it. This is especially advantageous in sub-threshold operation where the current of hundreds of unselected cells can greatly surpass the on current of a selected cell. This has a positive effect on the read static noise margin.

The read static noise margin (RSNM) of the proposed cell is relatively comparable to that of the traditional 8T SRAM cell, and much better than the standard 6T SRAM cell, in ideal conditions, as shown in Figure 4.22. Introducing PVT variations into the scenario results in a slightly worse read SNM when compared to the standard 8T SRAM cell. Due to the read operation not being isolated from the local bit-line, unlike the 8T SRAM cell, the cell is still subjected to bit-line charge sharing with the other cells connected to the same bit-line. As a result, the stability of the cell suffers

Figure 4.22: Comparison of the read SNM of the proposed design, the traditional 8T, and standard 6T SRAM cells. The V_{DD} used in the simulation is 300 mV.

slightly. Nevertheless, the comparatively short local bit-lines prevent the read SNM from being as abysmally small as in the standard 6T SRAM cell. In terms of cell stability, the read SNM is the weakest part of this design.

It is interesting to note that the proposed architecture is single-ended. However, horizontal bit-lines allows it to go a step further by only having one large capacitance global bit-line per cell block. That is, there is only one global bit-line per x bits, where x is the number of bits in each cell block. This decreases the dynamic power consumption by greatly reducing the number of bit-lines to pre-charge and discharge.

Bit-line leakage is also significantly reduced since there are a limited amount of bit-lines.

4.3 Hold Operation

The hold, or idle, operation is achieved by holding all word-lines low, which disables the access transistors. This leaves the local bit-line (LBL) floating. While global bit-lines may be disturbed, they are decoupled from the storage cells. Thus, the data remains unaffected. Figure 4.24 shows the hold static noise margin (HSNM). Since the storage cell is exactly the same as in the standard 6T and traditional 8T SRAM cells, the hold SNM is roughly the same in all cases.

Figure 4.23: Idle operation of the proposed design.

4.4 Cell Layout

Since the proposed SRAM cell is a modification of the standard 6T SRAM cell, its layout is based on the 6T thin-cell layout. The layout for the SRAM cell is illustrated in Figure 4.25. All transistors in the layout are minimum sized, creating a diffusion-notch-free layout. The thin-cell layout keeps the long global bit-lines, WLA, and WLB as short as possible to reduce their capacitances. This comes at the cost of a longer

Figure 4.24: Comparison of the hold SNM of the proposed design, the traditional 8T, and standard 6T SRAM cells for a V_{DD} of 300 mV.

local bit-line, write word-line, and read word-line, but they are relatively small by comparison. It would be ideal to keep the local bit-line as short as possible to improve the read and half-select static noise margins, but this would result in increased process variations and/or increased delay due to the higher capacitance global bit-lines. The area of the SRAM cell is roughly the same as the thin-cell layout at $2.91 \mu\text{m}^2$.

The layout for the overhead circuitry required for the cell block is depicted in Figure 4.26. The transistors are all aligned in the same direction to reduce process variations. The total area of this overhead section is around $2.14 \mu\text{m}^2$.

Figure 4.25: Layout of the proposed bit cell. The layout is based on the layout of the 6T thin-cell.

Figure 4.26: Layout for the overhead circuitry required for the proposed cell block, including read and write decoupling transistors, and read access transistors.

The combined layout can be seen in Figure 4.27. The overhead block circuitry is placed in the middle of the cell block with the SRAM cells placed to either side. The local bit-line (LBL) stops at the end of the cell block, unlike the write word-lines (WWL), and read word-line (RWL) that run for the entire row. All vertical word-lines and bit-lines are routed in metal 2, and all horizontal word-lines and bit-lines are routed in metal 3. The total area of the 8-cell configuration is $25.42 \mu\text{m}^2$. This results in an average area of $3.18 \mu\text{m}^2$ per bit cell or around 1.1 times the area of the 6T cell. Thus, it can be said that this design has an average area of 6.6T. This is very close to the area of the standard 6T cell, but with much better read SNM.

Figure 4.27: A section of the proposed SRAM architecture. The layout here shows 2 bit-cells with the overhead circuitry. Any additional cells would be added to the sides.

Chapter 5

Static Noise Margin Evaluations

To properly assess the proposed SRAM cell, the stability of the cell must be evaluated under all conditions, including process and temperature variations. Schematic simulations were done using Cadence Spectre (a SPICE class circuit simulator) using the TSMC 65 nm Mixed Signal RF salicide Low-K IMD process configured for 1.2 V supply voltage, which utilizes a BSIM4 (V4.5) model. The minimum length for the process is 60 nm, the minimum width is 120 nm, and the threshold voltage for a minimum sized NMOS transistor is 547.6 mV.

Monte Carlo simulations for the TSMC 65 nm Low Power process show that, at 11 bits in each cell block and a V_{DD} of 300 mV, the read current of a single bit cell is always greater than the cumulative bit-line leakage caused by the unselected devices in the same block. To keep decoding simple, it is ideal for a block to contain a number of bits that is a power of 2. Thus, for the TSMC 65 nm LP process, 8 bits per block is the quintessential choice to minimize area. However, to explore how the number of bits in a cell block affects stability, cell blocks with configurations of 1 bit, 2 bits, 4 bits, and 8 bits will all be tested. The schematics of these configurations are shown in Figure 4.2, Figure 4.3, Figure 4.4, and Figure 4.5.

5.1 Process Variations

The compared points are plotted with a tolerance level of $\mu - 5\sigma$ to study the worst case scenarios, where μ is the arithmetic mean and σ is the standard deviation. For normally distributed data, five standard deviations from the mean ($\mu \pm 5\sigma$) accounts for 99.9999426697% of all values. That is, only 1 out of 1 744 278 manufactured dies will be outside the range of $\mu \pm 5\sigma$ [49]. Only the $\mu - 5\sigma$ values are reported

because the smallest SNM is a concern, not the largest. The histograms the estimated probability density functions are related to, and the values of the simulation results can be found in Appendix A.

5.1.1 Hold Operation

As Figure 5.1 shows, there is very little change in the hold static noise margin as the number of bits in each cell block increases. There is only a 1 to 2 mV difference in hold SNM between the 8-cells and 1-cell configuration. The hold SNM is relatively independent from the number of bits because there are two transistors (the access transistors) in series between every two bits that minimizes how each storage node affects each other. It is consistent, however, that as the number of bits increase, there is a minor increase in hold SNM. While this difference is within a few millivolts, at a V_{DD} of 200 mV there is a 33% improvement in the hold SNM when utilizing the 8-cell configuration over the 1-cell. This small correlation can be attributed to the slight increase in the capacitance of the local bit-line as the number of bits increase.

There is very little difference between the hold SNM for the traditional 8T, standard 6T, and proposed design, but as V_{DD} drops, the difference becomes more apparent. At a V_{DD} of 200 mV, the 8T and 6T cells have a 25% improvement in hold SNM compared to the 8-cells block configuration. The decrease in hold SNM in the proposed design is due to the floating local bit-line. In the 8T and 6T cells, the bit-lines are pre-charged to logic ‘1’, so the vast majority of the leakage will be through the access transistor connected to the node holding ‘0’. In the proposed architecture, the voltage of the local bit-line can vary due to the leakage on the bit-line. Thus, the leakage through the access transistors can be higher than in the 8T and 6T cases.

The density functions of the traditional 8T cell, standard 6T cell, and the proposed architecture for the hold operation are all left-skewed normal distributions with very similar proportions. This can be verified by comparing the density functions in Figure 5.2. Note that since the density functions for all block configurations are very similar, only the 8 block configuration density function is included.

Figure 5.1: Simulation results for Hold SNM vs. V_{DD} for the different block configurations proposed, the traditional 8T cell, and the standard 6T cell. The plotted points are $\mu - 5\sigma$. The points in the table are in mV.

5.1.2 Write Operation

Like the hold operation, there is very little correlation between write static noise margin and the number of bits inside each cell block. There is at most 1.5 mV difference between the 8-cell configuration and 1-cell configuration at a V_{DD} of 200 mV. This relative consistency can be explained by the write decoupling transistor keeping the local bit-line low in all cases.

Nevertheless, there is a slight, but noticeable, decrease in write SNM as more bits

Figure 5.2: Estimated probably density function of the hold operation for the 8-cell block configuration and 8T/6T SRAM cells (they are the exact same). V_{DD} is 300mV and the temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. This is similarly true for all sub-threshold operating voltages. Refer to Figure A.1, Figure A.2, and Figure A.3 for the histograms of the 8-cell block configuration, 8T SRAM cell, and 6T SRAM cell, respectively.

are added to each cell block. This is the result of enabling one access transistor, which exposes the storage cell to the local bit-line. Leakage from non-selected access transistors connect to the same local bit-line negatively affects the write SNM. Adding more cells to the local bit-line only increases the leakage. This, along with the increasingly poor I_{ON}/I_{OFF} ratio as V_{DD} decreases, also explains why there is a bigger change in write SNM among the different block configurations as V_{DD} is lowered.

The differential traditional 8T and standard 6T cells have a much better write static noise margin at higher voltages, but the relative disadvantage diminishes as V_{DD} drops. The differential write operation pulls down one node in the storage cell

Figure 5.3: Simulation results for Write SNM vs. V_{DD} for the different block configurations proposed, the traditional 8T cell, the traditional 8T cell with single-ended write operation, and the standard 6T cell. The plotted points are $\mu - 5\sigma$. The points in the table are in mV.

while pulling up the other. This allows for a higher write SNM compared to a single-ended write operation. Unfortunately, as the voltage drops, and the I_{ON}/I_{OFF} ratio decreases, the weak pull-up is a disadvantage; to pull up the storage cell, the access transistor is enabled, which allows the leakage from other cells on the same bit-line to affect the node. However, at operational voltages for the proposed cell, the differential scheme always provides a better write SNM than the proposed architecture.

Peculiarly, the single-ended write operation for the traditional 8T cell (8T-SE in

Figure 5.3) provides a better write static noise margin than the proposed design. This is intriguing because the write mechanisms are nearly identical. The connected bit-line is brought low and one access transistor is enabled, pulling one side of the storage node down. The key difference is the use of a local bit-line and minimum-size write-decoupling transistor. Since all transistors inside the cell block are of minimum size, there is a slight contention between the load (or pull-up) transistor in the storage cell and the write decoupling transistor. Due to the use of NMOS access transistors and NMOS transistors being stronger than similarly sized PMOS transistors, the write decoupling transistor always wins, but the write operation is slightly weakened. In the traditional 8T cell, the high capacitance global bit-lines connected to the cells require a strong pull-down transistor to discharge the bit-lines and keep the bit-line voltage relatively constant throughout the write operation. Thus, the write operation is stronger than in the proposed architecture.

The density function for the write operation in the proposed design is also dissimilar to the density functions of the 8T and 6T cells. Figure 5.4 shows a left-skewed normal distribution for the 8-cell configuration (again, only 1 configuration is included since they all have similar distributions). However, Figure 5.5 shows right-skewed normal distributions for the single-ended write operation of the traditional 8T SRAM cell and the differential write operation of the traditional 8T/standard 6T cell. This means that in extreme process variations the write operation actually improves in the 8T and 6T cells, unlike the hold and read cases. In the proposed design, the write operation gets worse in the presence of extreme process variations.

Hence, the worst case write noise margin of the 8T and 6T cells is close to the mean, and the noise margin actually improves in most cases when subjected to process variations. This is a very desirable property, which is not present in the proposed architecture. Conversely to the 8T and 6T cases, the best case write noise margin is close to the mean, and the noise margin worsens when subjected to process variations. This contrast is the result of connecting low capacitance local bit-lines to the storage cell instead of high-capacitance global bit-lines. Thus, while the local bit-lines help improve the read operation, it negatively impacts the write operation.

Figure 5.4 show that as V_{DD} increases, the distribution becomes less skewed. The stronger on current, helps negate the negative impact of the local bit-lines, but not enough to have a right-skewed normal distribution like in the 8T and 6T cells.

Figure 5.4: Estimated probably density function of the hold operation for the 8-cell block configuration at various supply voltages. The temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. Refer to Figure A.4, Figure A.5, Figure A.6, and Figure A.7 for the histograms of the 8-cell configuration with supply voltages of 200 mV, 250 mV, 300 mV, and 350 mV, respectively.

5.1.3 Read Operation

The read operation is greatly affected by the number of bits in the cell block. Figure 5.6 displays that there is a large variation in the read static noise margin between each block configuration. At a V_{DD} of 250 mV, the configuration with 8-cells has a read SNM of 11.878 mV, while the 1-cell configuration has a read SNM of 22.226 mV. Thus, as the number of bits in the cell block configuration increases, the read SNM decreases.

This diminishing SNM is the result of subjugating the storage node to the local bit-line. With the access transistor enabled, there is only one transistor separating

Figure 5.5: Estimated probably density function of the write operation for the 8T/6T SRAM cell (they are the same), and 8T single-ended write scheme. V_{DD} is 300mV and the temperature is 27°C. Only the the density function when the V_{DD} is 300mV included because all supply voltages have very similar distributions. Refer to Figure A.8, and Figure A.9 for the histograms of the single-ended 8T, and 8T/6T SRAM cells, respectively.

the storage node from the other storage nodes connected on the same local bit-line. Thus, leakage through the non-selected access transistors increases considerably. As more cells are added to the local bit-line, the total leakage into the local bit-line is further exacerbated.

This leakage is why the standard 6T SRAM cell has abysmal read static noise margins, as shown in Figure 5.6. In most 6T SRAM architectures, there are very large number of bits connected to the same bit-line. While this gives a high bit density, it greatly reduces the read SNM, causing it to fail at sub-threshold supply voltages. The proposed architecture doesn't have this issue since it limits the number of cells connected on the same bit-line. Furthermore, the proposed architecture only enables one access transistor during reads, which limits the disturbance from other

Figure 5.6: Simulation results for Read SNM vs. V_{DD} for the different block configurations proposed, the traditional 8T cell, and the standard 6T cell. The plotted points are $\mu - 5\sigma$. The points in the table are in mV.

storage cells to a single node.

The conventional 8T SRAM cell has a slightly higher read SNM compared to the proposed architecture due to the storage cells being isolated from the bit-lines during reads. The read SNM is exactly the same as the hold SNM because the access transistors are off during the read operation. This also results in the 8T read operation being less affected by a reduction in V_{DD} . Figure 5.6 shows that the proposed architecture has a larger reduction in read SNM than the 8T cell as V_{DD} is lowered. There is a further decrease in read SNM as the number of bits in the cell

block configuration is increased.

Analyzing the density functions in Figure 5.7 shows that the traditional 8T cell and the proposed design both have left-skewed normal distributions and are very similar. The standard 6T cell has a very slightly right-skewed normal distribution. The read SNM for the 6T cell is very poor and can vary wildly due to the lack of stability.

Based on the simulation results, it is obvious that the read static noise margin is the limiting factor in the block configuration. Hence, there is a compromise between integration density and read SNM (and by extension minimum operating voltage). Depending on the yield and noise tolerance required, when using the 65nm LP TSMC process, 250 mV would be the minimum operating voltage for the 8-cell configuration and 230 mV for the 1-cell configuration.

Figure 5.7: Estimated probably density function of the read operation for the 8-cell block configuration, 8T, and 6T SRAM cells. V_{DD} is 300mV and the temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. This is similarly true for all sub-threshold operating voltages. Refer to Figure A.10, Figure A.11, and Figure A.12 for the histograms of the 8-cell block configuration, 8T SRAM cell, and 6T SRAM cell, respectively.

5.2 Temperature Variations

The equations in 4.1 show that the sub-threshold current depends on several temperature-dependent parameters and, thus, varies with temperature. The temperature dependence on the threshold voltage, V_T , can also be expressed by

$$V_{TH} = V_{TH0} - \kappa T, \quad (5.1)$$

where V_{TH0} is the threshold voltage at 0 K, T is the absolute temperature, and κ is the temperature coefficient of V_{TH} . From this equation, it is clear that the threshold voltage decreases as the temperature increases [50]. Furthermore, the mobility, μ , of the MOSFET carriers can be expressed by

$$\mu(T) = \mu(T_0)(T/T_0)^{-m}, \quad (5.2)$$

where $\mu(T_0)$ is the carrier mobility at room temperature T_0 , and m is the mobility temperature exponent. This equation shows that the mobility of the carriers decreases as the temperature increases [50].

By utilizing equations 4.1, 5.1, and 5.2, the temperature coefficient of the sub-threshold current of a MOSFET with a fixed gate-to-source voltage and $V_{DS} > 0.1$ V (making the current relatively independent of V_{DS}) can be found. The temperature coefficient is as follows:

$$\begin{aligned} T.C_{Sub} &= \frac{1}{I_{sub}} \frac{dI_{sub}}{dT} \\ &= \frac{1}{\mu} \frac{d\mu}{dT} + \frac{1}{V_T^2} \frac{dV_T^2}{dT} + \frac{1}{e^{\frac{(V_{GS}-V_{TH})}{nV_T}}} \frac{d}{dT} e^{\frac{V_{GS}-V_{TH}}{nV_T}} \\ &= \frac{2-m}{T} - \frac{1}{nV_T} \frac{(V_{GS}-V_{TH0})}{T} \end{aligned} \quad (5.3)$$

Thus, the sub-threshold current has a positive temperature coefficient and increases with temperature [50].

Given this temperature dependence, there can be a lot of variance in the SNM at differing temperatures. As such, it is important to evaluate the stability of the proposed SRAM design at multiple temperatures. The following test simulations show that the TSMC 65 nm LP process has increased variations at low temperatures. Thus,

to more accurately show temperature trends, box plots are used to show temperature alterations. The central mark in the box plots is the median, and the edges of the box are the 25th and 75th percentiles. The whiskers are defined as larger than $q3 + 1.5(q3 - q1)$ or smaller than $q1 - 1.5(q3 - q1)$, where $q1$ and $q3$ are the 25th and 75th percentiles. This corresponds to approximately $\pm 2.7\sigma$ (or 99.3%) coverage when the data is normally distributed [51], [49].

5.2.1 Hold SNM

As shown in Figure 5.8, Figure 5.9, Figure 5.10, and Figure 5.11, the temperature trends are the same for all block configurations. There isn't much change in the hold SNM as we increase the number of bits. The hold SNM remains the same in all block configurations at the same temperatures. As the temperature increases, there is a notable decrease in SNM due to the increase in leakage current. Unexpectedly, below 0°C the SNM decreases as the temperature drops. The variability of the SNM decreases as the temperature increases.

Examining Figure 5.12 and Figure 5.13 show that the increase in variability in the hold SNM experienced in the proposed SRAM cell as the temperature drops similarly occurs in the traditional 8T and standard 6T SRAM cells. It can be seen that the standard deviation in the proposed SRAM cell is roughly equal to the 8T SRAM cell at temperatures greater than or equal to 0°C , but at below 0°C the standard deviation is larger. The difference in standard deviation further increases as the temperature drops. This is a result of there being additional transistors on the local bit-line that affect leakage. The affect of temperature on the hold SNM of the SRAM cells is summarized in Figure 5.14.

Figure 5.8: Hold SNM vs. Temperature (°C) for the 1-cell block.

Figure 5.9: Hold SNM vs. Temperature (°C) for the 2-cell block.

Figure 5.10: Hold SNM vs. Temperature (°C) for the 4-cell block.

Figure 5.11: Hold SNM vs. Temperature (°C) for the 8-cell block.

Figure 5.12: Hold SNM vs. Temperature (°C) for the traditional 8T SRAM cell.

Figure 5.13: Hold SNM vs. Temperature (°C) for the standard 6T SRAM cell.

Figure 5.14: Simulation results for Hold SNM vs. Temperature (°C) for the different block configurations proposed, the traditional 8T cell, and the standard 6T cell. V_{DD} is 300mV. The plotted points are $\mu - 5\sigma$.

5.2.2 Read SNM

The temperature has a greater affect on the read SNM than on the hold SNM. Figure 5.15, Figure 5.16, Figure 5.17, and Figure 5.18 show that the read SNM of the proposed SRAM cell decreases as the temperature drifts above or below 0°C. Once again, the variability decreases as the temperature increases. The standard deviation drastically increases and the average SNM decreases exponentially as the temperature drops below 0°C.

The temperature has a greater affect on the read SNM as the number of bits increase inside the block. When the temperature drops below 0°C, there is a significant decrease in the average SNM and a much greater variability as the number of bits in the block increases. However, when increasing the temperature above 0°C, there is not a significant change in the average SNM and standard deviation as the number of bits increases. There is a minor drop in the average SNM and a small increase in the standard deviations when additional bits are added to the block. These changes can be attributed to the storage cell being exposed to the local bit-line on reads. Additional cells in the cell block increases the number of transistors that are affected by temperature variations.

Comparing Figure 5.18 and Figure 5.19 shows that temperature has a greater impact on the read SNM of the proposed design than on the traditional 8T SRAM cell. Compared to the 8T SRAM cell, there is increased variability at all temperatures. The increase in standard deviation is especially pronounced below 0°C. There is a larger decrease in average SNM in the proposed design than in the 8T cell as the temperature deviates from 0°C. This decrease is more prominent with each additional bit added to the block configuration. Again, this can be attributed to the storage cell being exposed to the local bit-line, whereas, in the 8T SRAM cell, the storage cell is isolated during reads.

Temperature has relatively little affect on the read SNM of the standard 6T SRAM cell at sub-threshold voltages, as shown in Figure 5.20. The change in the average SNM as the temperature changes is decently larger in the proposed SRAM architecture than in the 6T SRAM cell. The proposed SRAM architecture also has lesser

Figure 5.15: Read SNM vs. Temperature (°C) for the 1-cell block.

Figure 5.16: Read SNM vs. Temperature (°C) for the 2-cell block.

Figure 5.17: Read SNM vs. Temperature (°C) for the 4-cell block.

Figure 5.18: Read SNM vs. Temperature (°C) for the 8-cell block.

Figure 5.19: Read SNM vs. Temperature (°C) for the traditional 8T SRAM cell.

Figure 5.20: Read SNM vs. Temperature (°C) for the standard 6T SRAM cell. Note that the y-axis has been translated down 20mV to show the standard deviation without changing the scale.

variability at all temperatures, except at -50°C , compared to the 6T SRAM cell. The variability in the 6T SRAM cell remains more or less the same at all temperatures, with a slight decrease as the temperature moves away from 0°C . This is in stark contrast with the proposed design which has a noticeable decrease in average SNM as the temperature diverges from 0°C , and a decrease in a standard deviation as the temperature increases. This shows how the 6T SRAM cell is less stable at sub-threshold voltages compared to the proposed design and the 8T SRAM cell. This difference is due to the use of the relatively short local bit-line. The long global bit-lines utilized in the 6T SRAM cell subjects the storage cell to all the PVT variations in all the cells connected to the bit-line.

Curiously, the 6T SRAM cell has an even spread above and below the median, whereas the proposed SRAM cell (and 8T SRAM cell) have an increased spread below the median. This means in the average case, the read SNM is close to the upper limit, and only in extreme process variations does the read SNM decrease. The affect of temperature on the read SNM of the SRAM cells is summarized in Figure 5.21.

Figure 5.21: Simulation results for Read SNM vs. Temperature ($^{\circ}\text{C}$) for the different block configurations proposed, the traditional 8T cell, and the standard 6T cell. V_{DD} is 300mV. The plotted points are $\mu - 5\sigma$.

5.2.3 Write SNM

The write operation is more resilient to temperature variations than the read and hold operations. Figure 5.22, Figure 5.23, Figure 5.24, and Figure 5.25 show that as the temperature changes there is relatively little change in the average write SNM of the proposed Design. Additionally, there is only a minor increase in variability as the temperature drops.

Since the write SNM is not dependent on the number of bits in each block, the temperature has an equivalent affect on all block configurations. As is consistent with the read and write static noise margins, there is a decrease in the average write SNM as the temperature deviates from 0°C. However, the decrease is not as significant. Interestingly, the rate of decrease of the average SNM below 0°C is not as substantial as in the read and write cases. This is true for the increase of variability as well. This is evident by comparing Figure 5.11, Figure 5.18, and Figure 5.25.

Temperature variations have a smaller affect on the write operation of the proposed design than they do on the write operation in the traditional 8T and standard 6T SRAM cells. Figure 5.25 shows that the average SNM and variability of the proposed design slightly decreases as the temperature diverges from 0°C. However, Figure 5.26 and Figure 5.28 show that the median SNM and standard deviation of the 8T and 6T cells have a greater reduction than the proposed design as the temperature decreases.

Comparing Figure 5.25 and Figure 5.27 shows that the write operation of the proposed SRAM cell is not as tolerant of temperature variations as the single-ended write operation in traditional 8T SRAM cells. Compared to the single-ended write scheme, there is a larger decrease in average SNM and increased variability in the proposed design as the temperature drifts from 0°C. In contrast to 8T SRAM cells utilizing a single-ended write scheme, as the temperature drops below 0°C there is a decrease in the write SNM of the proposed SRAM cell. Thus, it can be stated that the write operation for the proposed cell follows the same trend as the 8T read and hold operations instead of following the single-ended write operation. The single-ended

Figure 5.22: Write SNM vs. Temperature (°C) for the 1-cell block.

Figure 5.23: Write SNM vs. Temperature (°C) for the 2-cell block.

Figure 5.24: Write SNM vs. Temperature (°C) for the 4-cell block.

Figure 5.25: Write SNM vs. Temperature (°C) for the 8-cell block.

Figure 5.26: Write SNM vs. Temperature (°C) for the traditional 8T SRAM cell.

Figure 5.27: Write SNM vs. Temperature (°C) for the single-ended write scheme utilizing the traditional 8T SRAM cell.

Figure 5.28: Write SNM vs. Temperature (°C) for the traditional 8T SRAM cell.

write operation is the same in both cells, so it seems that the cause for this difference is the use of the write-decoupling transistor inside the cell block.

The proposed design has an almost even spread above and below the median, while the differential and single-ended write operation for the 8T SRAM cells have an increased spread above the median. This shows that the write operation in the SRAM cell blocks are more susceptible to process variations than the traditional 8T and standard 6T SRAM cells. The affect of temperature on the SRAM cells is summarized in Figure 5.29.

Figure 5.29: Simulation results for Write SNM vs. Temperature (°C) for the different block configurations proposed, the traditional 8T cell, the single-ended 8T write-scheme, and the standard 6T cell. V_{DD} is 300mV. The plotted points are $\mu - 5\sigma$.

Given all these results, the static noise margins of the proposed design is very comparable to the 8T SRAM cell. In many cases it is slightly inferior, but it is definitely an acceptable trade off for the area savings. It is undoubtedly more suitable for sub-threshold operation than the 6T SRAM cell, which has abysmal read SNM. Now that the stability of the cell is proven to be adequate, its performance and power requirements must be examined.

Chapter 6

SRAM Block Simulations

To assess the proposed SRAM architecture more accurately, a complete SRAM array block was implemented, simulated, and compared to a traditional 8T SRAM block. Since the purpose of this study is to design a minimum area SRAM cell, the 8-cell configuration of the proposed design is used. All simulations were done using the worst-case corner, i.e. FS (Fast NMOS, Slow PMOS). The FS corner is the worst-case scenario in terms of SRAM stability, i.e. SNM, because the minimum-size PMOS transistors are inherently weaker than the minimum-size NMOS transistors in the storage cell. It is not necessarily the worst-case for speed or power consumption.

6.1 SRAM Block

Figure 6.1 shows the architecture of the implemented 32 kilobit memory block. The array is composed of 16 columns and 256 rows of the 8-cell block configuration, which translates to an array of 128 x 256 bits. To be homologous with modern computers, the block was constructed to be byte-addressable. Thus, there is an address space of 2^{12} , or 4096, 8-bit words.

The structure of the block was designed to support a bit-interleaving structure; the architecture writes concurrently into one cell in multiple cell blocks contained in the same row. The bits are interleaved between the blocks. The following shows the composition of a row, where A is the first byte, and I is the second byte:

$$| A_1 | I_1 | A_2 | I_2 | A_3 | I_3 | A_4 | I_4 | A_5 | I_5 | A_6 | I_6 | A_7 | I_7 | A_8 | I_8 | \quad (6.1)$$

As such, 8 out of 16 columns are read-from or written-to at the same time. The

Table 6.1: Byte Layout

Col 1	Col 2	Col 3	...	Col 16	Col 17	...	Col 128
A_1	B_1	C_1	...	A_2	B_2	...	P_8

composition of the data inside a cell block would be:

$$| A_x | B_x | C_x | D_x | E_x | F_x | G_x | H_x |, \quad (6.2)$$

where A, B, C, D, E, F, G, and H are all separate bytes, and x is an arbitrary position in the byte. The value of x is the same in all the bits stored in the block. For example, the block storing A_1 in 6.1, would have a composition of:

$$| A_1 | B_1 | C_1 | D_1 | E_1 | F_1 | G_1 | H_1 |, \quad (6.3)$$

and the block storing I_1 would have a composition of:

$$| I_1 | J_1 | K_1 | L_1 | M_1 | N_1 | O_1 | P_1 |. \quad (6.4)$$

In this example, A is the first byte in the row, B the second, etc, with P being the sixteenth.

The address decoder used is a simple clock-based decoder with a predecoding stage designed for maximum integration density. It is not optimized for speed or minimum logical effort. The data inputs for the block are connected to the word-line demultiplexers. Each bit input is shared between two demultiplexers (i.e. word-line demultiplexers 1 and 2 are connected to ‘Data In [1]’). The read bit-line multiplexers selects between two global read bit-lines. The sense amplifiers utilized are standard high-skewed inverters, which are arguably ideal for sub-threshold large-single sensing single-ended designs.

The conventional 8T SRAM array block is shown in Figure 6.2. The array is composed of 128 columns and 256 rows of the 8T cell, which translates to an array of 128 x 256 bits. Again, there is an address space of 2^{12} , or 4096, 8-bit words.

Bit-interleaving was implemented in the same manner as the proposed architecture for the 8T array block. One bit in a byte is written every 16 bits in a row. This is

Figure 6.1: Array level block diagram of the 32 kb memory block for the proposed SRAM architecture.

perhaps best explained by Table 6.1, which shows the layout of bytes in a row. In Table 6.1, A is the first byte in the row, B the second, etc, with P being the sixteenth.

With the exception of the composition of the array, the architecture is very similar to the proposed design block. The main difference is the use of a write driver for the write bit-lines and pull-ups needed for all bit-lines. In the proposed design, the word-lines WLA and WLB are brought high when selected, but are low when not. In the 8T cell, the write bit-lines are pre-charged high, but then are selectively discharged depending on the value being written. Despite these differences in operation, the schematics of the write driver and word-line multiplexer are very similar, with the biggest difference being the additional circuitry in the word-line multiplexer to handle read selection as well.

Since all bit-lines are required to pre-charged, strong pull-ups are needed for the write bit-lines in addition to the read bit-lines. Due to the differences in architectures, there are 384 bit-lines in the 8T block, while only 16 global bit-lines in the proposed block. Another, albeit relatively minor, difference is that the read bit-line multiplexer in the 8T array block is connected to 16 bit-lines instead of 2 like in the proposed architecture.

Figure 6.2: Array level block diagram of the 32kb memory block for the traditional 8T SRAM cell.

6.2 SRAM Block Operation

While the architecture of the SRAM blocks are very similar, the write operations of each block is fairly different. This is verified by comparing the signals in Figure 6.5,

for the proposed design, and Figure 6.3, for the 8T block. In the 8T SRAM block, the write operation starts when the WR signal is enabled and CLK (the clock) is high. As soon as CLK is high, the pull-ups for the bit-lines are disabled. This causes the write bit-lines to steadily drop due to bit-line leakage. However, once the decoder processes the address and write/read signals, the write driver further pulls down the appropriate write bit-line for the desired written data ($\overline{WBL}$ in Figure 6.3), while leaving the other write bit-line floating.

Figure 6.3: Simulated signals involved in the write operation of the 8T block for FS corner at 27°C.

Slightly before the write driver is activated, the decoder enables the WWL (write

word-line) signal for the selected row, which allows for the modification of the data stored in the selected cell. The write bit-line that has been discharged sets the value of the connected cell node to logic ‘0’. Feedback through the back-to-back inverters then raises the voltage of the other node. In Figure 6.3, $\overline{WBL}$ discharges Cell Node 2 that, subsequently, raises the voltage of Cell Node 1.

It is evident that the voltage of Cell Node 1 never becomes logic level ‘1’ due to it being connected to the high capacitance write bit-line. However, once the write operation is completed and the SRAM block is in the idle state, feedback by the cross-coupled inverters brings the nodes to their appropriate level of V_{DD} and GND. Figure 6.3 also shows that it takes some time after the clock switches low for all relevant signals to be disabled, allowing the SRAM block to go into the idle state. This is due to the decoder taking time to process the switching clock signal. This delay is present in all operations.

In the 8T SRAM block architecture used for comparison in this study, the worst-case leakage in the write bit-lines can severely affect the write ability of a cell. The worst-case leakage in the write bit-line occurs when all cells in a column are storing the same value. The write bit-line connected to the node storing ‘0’ will suffer from a relatively substantial amount of leakage. If it is desired to change the value of one cell in this column, the write bit-line connected to the node storing a ‘1’ must be discharged to change the node’s value. This presents a race condition on which can discharge the bit-line faster, the write driver or the cumulative leakage.

Figure 6.4 presents this problem when running a simulation in the worst-case FS corner. It can be seen that there is competition between WBL and $\overline{WBL}$. Once the pull-ups are disabled, $\overline{WBL}$ steadily decreases due to leakage. When the write driver is enabled, it quickly pulls down WBL and is able to discharge the desired bit-line slightly faster than the leakage on its counterpart, $\overline{WBL}$. However, this results in a marginal difference between the two nodes and bit-lines. Since a cell is written to by discharging one node, in Figure 6.4 the difference is enough to recover the desired data values after the end of the write operation. This issue can be mitigated by implementing a faster decoder, a stronger write driver, or by having the write driver keep the other node high. Although, all of these would result in a greater area consumption and/or increase in power consumption.

The write operation of the proposed SRAM block was described in Chapter 4. The decoder processes the input, setting WWL high and enabling the proper select

Figure 6.4: The problem with the write operation in the 8T block. Simulation performed for FS corner at 27°C.

world-line (WLA or WLB). Setting WWL high brings the local bit-line (LBL) in the selected block to GND. Once the proper word-line is selected, a node in the selected cell is exposed to the local bit-line, setting its value to ‘0’. Feedback by the cross-coupled inverters then sets the opposite node to ‘1’. This is illustrated in Figure 6.5. In this example, WLA is selected, forcing Cell Node 2 to ‘0’ that, in turn, sets Cell Node 1 to ‘1’.

Figure 6.5 shows that the node being set to ‘1’ is actually set to a logic ‘1’ during the write operation, unlike in the 8T SRAM block. This is the result of the node remaining isolated during the operation. However, the value of the node doesn’t reach V_{DD} until after the write operation is over. This is due to the leakage through the access transistor as a result of the local bit-line is connected to GND during the write operation. After the write operation, this node does initially decrease in voltage as a result of RWL going low before WLA, but, after WLA is low and the access transistors

are OFF, feedback forces the node to V_{DD} . The initial decrease can be avoided by ensuring WLA goes low before WWL. After the write operation, the local bit-line is left floating, reducing the access transistor leakage. There isn't a race condition in the write bit-lines like in the 8T SRAM block since there aren't any write bit-lines. Nor are any bit-lines left floating during the operation, so they cannot be affected by leakage. There is still delay caused by the decoder once the write operation is over, however.

Figure 6.5: The signals involved the write operation in the proposed block. Simulation performed for FS corner at 27°C.

Conversely to the write operation, the read operation is fairly similar in both SRAM blocks. In the 8T SRAM block, the read operation occurs when the decoder enables the RWL (read word-line) signal for the selected row. This enables the read access transistor in all 8T cells in the selected row. When the stored value in a cell

is '1', the read decoupling transistor in the cell is also enabled. This results in the drainage of the read bit-line (RBL). The multiplexer then selects the desired bit for read, sending it to the sense amplifier, which reads the value. This operation is shown in Figure 6.6. Note that the reason for the rise of the RWL signal at the beginning of the read operation is due to the slowness of the decoder.

Figure 6.6: The signals involved the read operation in the 8T block. Simulation performed for FS corner at 27°C. Note that the reason for the spike of the RWL signal in the beginning of the read operation is due to the slowness of the decoder.

The read operation of the proposed block similarly occurs when the decoder enabled the RWL signal for the selected row, which enables the read access transistors in all cell blocks in that row. At the same time, the word-line demultiplexer selects

the desired column by enabling the WLB signal. The WLB signal enables one access transistor in every cell in the selected row. When reading a ‘1’, the enabled access transistor connects the selected storage cell to LBL, setting LBL to ‘1’ and enabling the read decoupling transistors for that block, draining RBL. The read bit-line multiplexer then sends the selected column to the sense amplifier, which reads the value. This operation is illustrated in Figure 6.7. Again, the signals begin to go high at the beginning of the idle state as a result of changing the signals at the same time as the clock switches low. It takes time for the decoder to process the low clock state.

If the read bit-lines between both blocks are compared, it is evident that RBL in the 8T block drains faster once RWL is asserted due to the read decoupling transistor in the selected cell being fully ON with a value of V_{DD} being applied to the gate. As RBL drains more and more, the drain current slowly decreases due to V_{DS} decreasing in the read access transistor. However, by this point, the sense amplifier has already started to go high, making the decreasing drain current trivial. RBL in the proposed block drains much slower due to the read decoupling transistors not being fully ON. The drain current also slowly decreases as RBL is drained, but the change is not as drastic as the drain current is already impaired. RBL does go to GND faster in the proposed design, but this is inconsequential, as the high-skewed sense amplifier has already read the value by then.

The Idle state in both blocks is achieved by disabling all control signals and pre-charging all bit-lines. In the 8T SRAM block this means that all write bit-lines (WBL and $\overline{WBL}$) and read bit-lines (RBL) are pre-charged and kept high in the idle state. Only RBL is pre-charged and kept high in the proposed block. It is necessary to put the blocks in the idle state between operation so the bit-lines can pre-charge. Since all control signals are disabled, the local bit-lines in all the blocks are left floating in the idle state. This causes the local bit-line to raise slightly in voltage as it reaches an equilibrium in regards to leakage, as seen in Figure 6.7.

6.3 Comparison to 8T SRAM Block

In order to compare the proposed SRAM block to the 8T SRAM block, data was written to the same byte in both blocks and then read out. The basic test done was to write 0101 0101 (i.e. 85 in decimal) to register 0001 0000 (16), and then see if the data was read back successfully. Both blocks completed the test successfully with

Figure 6.7: The signals involved the read operation in the proposed block. Simulation performed for FS corner at 27°C. Note that the spikes in the signals in the beginning of the idle state is due to changing the signals at the same time as the clock switches low. It takes time for the decoder to process the low clock state.

a V_{DD} of 300 mV, with a time period of 4 μs for the clock (to account for the slow read operation). The results of this simulation are shown in Figure 6.3, Figure 6.5, Figure 6.6, and Figure 6.7. The first period of operation in the write diagrams is when the block has just started up, so the values stored in the cells are of random arbitrary value.

Since this test encompasses all operation cases, it was also used to measure the power and energy consumption. Figure 6.8 shows the power consumption of both blocks throughout all states. The power consumption in the start cycle is inconsequential since none of the bits are initialized and many nodes in the storage cells are

Figure 6.8: Comparison of the power consumption during the start-up, write, idle, and read operations.

in an unstable state (resulting in increased leakage). Note that there is a sudden drop in the power consumption in the 8T block near the end of the start cycle. At this point, before any signals have been enabled or changed, the unstable states inside the storage cells have reached a state of stability. The storage cells in the proposed block do not reach a point of stability before the end of the start cycle. This is because both sides of the SRAM cells are connected to the same local bit-line, whereas each side of the 8T SRAM cell is connected to a different bit-line. The unequal amount of leakage into each node in the 8T cell forces the nodes to a state of stability through feedback.

6.3.1 Write Operation

By analyzing the power consumption of the SRAM blocks in Figure 6.9, the power consumption of writing one byte into each SRAM block can be found. Looking at the 8T block, it is apparent that the power consumption drops as the pull-ups for the bit-lines are turned off and slowly discharge. Then, as WWL is enabled, which turns on the access transistors in the selected row, there is a spike in the power consumption followed by a steady increase as the values in the storage cell are written to. Once the values are written to the cells, the power consumption reaches a steady state. The average power consumption over the 4.0 μs period is 1.601 μW . Figure 6.10 shows the total energy consumed during the write operation, which is 6.39 pJ.

Figure 6.9: Comparison of the power consumption during a write. Close up view of Figure 6.8.

Figure 6.10: Comparison of the energy consumed during a write.

The power consumption of the proposed design remains relatively constant at the start. Once the decoder has processed the high clock signal, there is an increase in power consumption as a result of enabling WWL and WLA/WLB, which turn on the write decoupling transistors in the selected row and the access transistors in the selected column, respectively. Once the values are written to the bit cells, the power consumption begins to drop until the storage cells reach a steady state. The power consumption is quite large in this case because there are numerous cells that are not in a steady state until the access transistors have been turned on, forcing them into a steady state through feedback. As such, this test is not an accurate measure of the power consumption when writing a single byte in the block. Nevertheless, the average power consumption over the $4.0\ \mu\text{s}$ period is $1.717\ \mu\text{W}$. Figure 6.10 shows the total energy consumed during this write operation is $6.87\ \text{pJ}$.

For a more accurate portrayal of the power consumption during the write operation, another test was done. In this test 0101 0101 (85 in decimal) was written to register 0001 0000 (16), then in the next operation, 1010 1010 (170 in decimal) was written to the same register. Therefore, this test allows for the measurement of the amount of power required to switch all the bits in a byte, where half the bits are 1s. The power consumption for the second operation in this test is shown in Figure 6.11.

Figure 6.11: More accurate comparison of the power consumption during a write.

The power consumption of the 8T write operation is very similar to the power consumption in the original test because, in both cases, all the bits in the block have reached a steady state before the operation began. Conversely, the proposed block has a lower power consumption than in the original test. There is a steady increase in

Figure 6.12: More accurate comparison of the energy consumption during a write.

power consumption once the decoder has enabled the WWL and WLA/WLB signals. Once the write operation has completed writing the new data, the power consumption again reaches a steady state. The steady state power consumption is slightly, but noticeably, higher than in the original. The increase in power consumption is because the opposite word-line (i.e. WLB instead of WLA) is enabled to write the opposite value. The value of the local bit-line in unselected blocks in the same column thus differs, and is the opposite of what it was in the first write. In this write operation, the values of the local bit-lines in unselected blocks are mostly ‘1’, which results in higher read bit-line leakage.

In comparison to the original test, the average power consumption of the 8T block during the write operation is $1.617 \mu\text{W}$, and $1.677 \mu\text{W}$ for the proposed block. The total energy consumption during the $4.0 \mu\text{s}$ period is 6.47 pJ for the 8T block and

6.71 pJ for the proposed block. If the power consumption during the actual write operation is only considered (from 12 μ s to 13 μ s), the average power consumption of the 8T block is 1.635 μ W, or a total energy consumption of 1.64 pJ, and 1.59 μ W for the proposed block, or a total energy consumption of 1.36 pJ. Overall, the write operations are very comparable with a write energy consumption of 0.205 pJ/bit for the 8T block and 0.17 pJ/bit for the proposed block. If the write period only consisted of the actual write operation, the proposed design definitely seems superior in terms of power consumption.

6.3.2 Read Operation

The power consumption during the read operation is shown in Figure 6.13. During the read operation an entire byte is read out by the block. Looking at the power consumption of the 8T block, there is an immediate drop in the power consumption as the read operation begins. This drop is due to the pull-ups being disabled, leaving the bit-lines floating. When the RWL signal is enabled by the decoder, the power consumption shoots up as all the read access transistors in the selected row are enabled, selectively draining the read bit-lines. As the bit-lines slowly discharge, the power consumption drops due the decrease in leakage. A steady state is not achieved since the bit-lines are still draining at the end of the cycle. The average power consumption throughout this 4 μ s period is 1.947 μ W, for a total energy consumption of 7.788 pJ (as found in Figure 6.14).

In the proposed block, the power consumption raises once the RWL and WLB signals are raised, enabling the read access transistors for all blocks in the selected row, and access transistors for all cells in the selected column. When the global read bit-lines begin to discharge, the power begins to slowly drop. This power decrease is not as drastic as in the 8T SRAM block because 384 bit-lines (128 being read bit-lines) are discharging in the 8T block, while only 16 are discharging in the proposed. This results in an average power consumption of 1.579 μ W, and a total energy consumption of 6.317 pJ.

If only the power consumption of the actual read operation is looked at (the period between 12 μ s and 15 μ s), the average power consumption of the 8T block is 1.961 μ W, and the average for the proposed design is 1.563 μ W. The total energy consumption

Figure 6.13: Comparison of the power consumption during a read. Close up view of Figure 6.8.

during this period is 5.884 pJ for the 8T block, and 4.688 pJ for the proposed block. This translates to a read energy consumption of 0.7355 pJ/bit for the 8T block and 0.586 pJ/bit for the proposed block.

6.3.3 Idle State

The power consumption during the idle state is relatively constant since no signals are enabled and no bit-cell is changing states. In fact, this is the static power consumption of the circuit. There is a significant decrease in power consumption at the beginning of Figure 6.15. This initial decrease is due to the decoder being slow to process the

Figure 6.14: Comparison of the energy consumed during a read.

low clock state. Therefore, the idle (static) power consumption is more accurately portrayed by the period between 9 – 12 μs . In the this period, there is a slight increase in power consumption in the proposed block due to the local bit-lines reaching a steady-state voltage as a result of leakage in the bit-line.

Taking into account the entire period of the cycle, the average power consumption is 2.053 μW for the 8T block, and 1.251 μW for the proposed block. This results in a total energy consumption of 8.21 pJ and 5.01 pJ for the 8T and proposed blocks, respectively. If the average power consumption is measured from the period between 9 – 12 μs , the static power consumption is 1.982 μW for the 8T block, and 1.211 μW

Figure 6.15: Comparison of the power consumption during the idle state, i.e. static power consumption. Close up view of Figure 6.8.

for the proposed block. In this period, the total static energy consumption is 5.934 pJ for the 8T block, and 3.616 pJ for the proposed block.

The reason for the large discrepancy in static power consumption between the blocks is due to the 8T SRAM block having 384 bit-lines that have to be pre-charged and kept high while in the idle state. The proposed block only has 16 bit-lines that need to be pre-charged and kept high. This results in significant power savings for the proposed SRAM block.

Figure 6.16: Comparison of the energy consumed during the idle state.

6.3.4 Frequency of Operation

By measuring the delay of the read and write operations in Figure 6.3, Figure 6.5, Figure 6.6 and Figure 6.7, the maximum frequency for each operation can be calculated. The frequency of operation for the 8T SRAM block are 406.2 kHz and 975.8 kHz for the read and write operations respectively. This includes the decoder delay. Since the decoder is far from optimized for speed, the operation speed was also measured after the decoder was finished processing the high clock signal. This occurs when RWL and WWL for the selected row are enabled. Thus, the frequency of operation, not including the decoder delay, for the 8T SRAM block are 548.8 kHz and 2.887 MHz for the read and write operations respectively. Conversely, in the proposed design, the frequency of operation for the read and write operations are 399.8 kHz and

Table 6.2: Summary of comparison between the Proposed SRAM block and the 8T SRAM block.

$V_{DD} = 300 \text{ mV}$	Proposed	8T	% Improvement
Max Write speed	6.70 Mhz	2.887 MHz	132%
Write Power Consumption	1.59 W	1.635 μW	2.75%
Write Energy/bit	0.17 pJ/bit	0.205 pJ/bit	17.1%
Max Read Speed	544.8 kHz	548.8 kHz	-0.73%
Read Power Consumption	1.563 W	1.961 μW	20.3%
Read Energy/bit	0.586 pJ/bit	0.736 pJ/bit	20.4%
Idle Power Consumption	1.211 W	1.982 μW	38.9%
Area increase over 6T	10%	$\sim 30\%$	17.5%

1.304 MHz, respectively. When not including the decoder delay, the frequencies of operation are 544.8 kHz and 6.70 Mhz for the respective operations.

The results of this section is summarized in Table 6.2. The table shows that the proposed SRAM block has similar or better performance, lower operational and idle (static) power consumption, and higher integration density than 8T SRAM block.

6.4 Comparison at Various Supply Voltages

Since this design is intended to be a generic design, the power consumption, energy consumption per operation, and frequency of operation were measured at various supply voltages to better test and analyze it. The minimum operation voltage for the read operation was determined to be 300 mV through testing. Below 300 mV the worst-case leakage overwhelms the ON current of the read-decoupling transistors. When the number of rows in the array of blocks is decreased, a lower read operation voltage is obtainable. This smaller SRAM block was not thoroughly tested, however.

With this minimum operating voltage in mind, the supply voltages explored range from 300 mV – 500 mV. The average power consumption and total energy for each operation extends for the entire 4 μs period, not just until the operation is successfully completed.

Figure 6.17: Comparison of the power consumption during the write, read, and idle operations. Simulations performed for FS corner at 27°C.

Figure 6.17 and Figure 6.18 show that the power consumption of the write operation dramatically increases above a V_{DD} of 400 mV and that the minimum power consumption is achieved at 300 mV. It also shows that the power consumption of the read operation increases cubically as V_{DD} increases. The power consumption in the idle state increases linearly as the supply voltages increases. It is, thus, advantageous to run the SRAM block at a V_{DD} of 300mV.

6.4.1 Frequency of Operation

Figure 6.19 and Figure 6.20 show that both the read and write speeds increase exponentially as the supply voltage increases.

With these results, it is clear that the proposed design is suitable for sub-threshold and near-threshold operation. When compared to the 8T SRAM block, it matches or beats its operation speed and power consumption at a lower transistor count. The bit array in the 8T block contains 262144 transistors, while there is only 217088 in the proposed design. This count doesn't include any of the ancillary circuitry,

Figure 6.18: Comparison of the energy per bit consumed during the write and read operations. Simulations performed for FS corner at 27°C.

Figure 6.19: Comparison of the maximum write and read speed. Simulations performed for FS corner at 27°C.

Figure 6.20: Comparison of the maximum write and read speed not including the decoder delay. Simulations performed for FS corner at 27°C.

but the total transistor count is pretty similar in both cases. The proposed design's performance, low power consumption, and low area requirements make it an ideal substitute for the 8T SRAM in many applications.

Chapter 7

Conclusion

This study presented a new sub- and near-threshold SRAM architecture. The proposed design's stability was thoroughly tested at the schematic level in the presence of process, temperature, and voltage variations and compared to the standard 6T and traditional 8T cells. Then, a 32kb SRAM block implementing the proposed architecture was designed, simulated at the schematic level, and compared to a traditional 8T SRAM cell block—showing favourable results.

Thus, this study successfully designed and implemented a new SRAM cell design that is comparable in performance to the 8T SRAM cell, but has lower operational and idle power consumption, in addition to higher integration density. The proposed SRAM architecture does all this while also mitigating half-select disturb, unlike the 8T SRAM cell, and supporting bit-interleaving. The SRAM cell is not as stable as the 8T SRAM cell, nor does its operating supply voltage go as low. However, considering that its average area per bit is very close to the standard 6T SRAM cell while still working at sub-threshold voltages, these are reasonable trade offs. Perhaps this provides a solution to the problem of a low-cost SRAM design for sub-threshold operation.

The ultra-low power embedded market is a relative niche currently. In the current market, there is perhaps a higher demand for SRAM cells that work at lower voltages, or cells that have better performance, or, in particularly the biomedical segment, cells that have increased stability, than there is for an SRAM cell that minimizes area. However, the demand is growing. With little improvement in battery technology over the past decade but rapid advancements in CMOS technologies, ultra-low power systems may soon become more common. If that is the case, there will definitely be demand for low-cost, low-area SRAM blocks.

7.1 Future Work

A number of possible areas can be tackled as an addendum to this work. It would be advantageous to improve the read SNM and reduce bit-line leakage in order to decrease the minimum read operation voltage. Numerous methods can be explored such as increasing the word-line voltage during reads, or a different bit cell structure. Perhaps the best path would be exploring separate write and read paths while retaining the horizontal bit-lines for reads may yield positive results. Alternatively, reducing the number of transistors in the read path will be advantageous to further reducing area. Using HVT transistors in only the read path may help reduce the overall transistor count, but it's debatable if this will yield lower area consumption.

It would also be interesting to explore implementing horizontal bit-lines into other SRAM architectures to see if it is universally advantageous or if this design is just an enigma. Horizontal bit-lines is certainly a fascinating idea for sub-threshold write operation. While the current proposed architecture doesn't improve write stability, it does improve power consumption and performance.

There is still a decent amount work that can be done on the proposed architecture. A vertex of lowest power consumption was not found, if there is indeed one. If possible, making the block operate at lower voltages and finding the power supply voltage that facilitates the lowest power consumption would be interesting to find. Different size SRAM blocks should be examined to see how the number of rows and columns affects the power consumption, speed, and voltage operation for the proposed design. A chip could be manufactured to properly assess the design. Even just simulating the entire layout with parasitic capacitances and resistances would be advantages to examine. Seeing how well it scales to smaller processes would be an interesting experiment as well.

List of References

- [1] Intel Newsroom, “40th anniversary of the microprocessor,” 2011, Sept. [Online]. Available: <http://newsroom.intel.com/docs/DOC-2383> [Accessed: Jan. 15, 2015].
- [2] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*. USA: Addison-Wesley Publishing Company, 4th ed., 2010.
- [3] E. Seevinck, F. J. List, and J. Lohstroh, “Static-noise margin analysis of mos sram cells,” *Solid-State Circuits, IEEE Journal of*, vol. 22, no. 5, pp. 748–754, 1987.
- [4] S. V. Avadhanam, “Design of a flexible high temperature sram with reduced design time,” Master’s thesis, Oklahoma State University, 2007.
- [5] S. Nalam and B. Calhoun, “Asymmetric sizing in a 45nm 5t sram to improve read stability over 6t,” in *Custom Integrated Circuits Conference, 2009. CICC ’09. IEEE*, pp. 709–712, Sept 2009.
- [6] M.-H. Tu, J.-Y. Lin, M.-C. Tsai, S.-J. Jou, and C.-T. Chuang, “Single-ended subthreshold sram with asymmetrical write/read-assist,” *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 57, no. 12, pp. 3039–3047, 2010.
- [7] K. Honda, K. Miyaji, S. Tanakamaru, S. Miyano, and K. Takeuchi, “Elimination of half select disturb in 8t-sram by local injected electron asymmetric pass gate transistor,” in *Custom Integrated Circuits Conference (CICC), 2010 IEEE*, pp. 1–4, 2010.
- [8] D. Anh-Tuan, J. Y. S. Low, J. Y. L. Low, Z.-H. Kong, X. Tan, and K.-S. Yeo, “An 8t differential sram with improved noise margin for bit-interleaving in 65 nm cmos,” *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 58, no. 6, pp. 1252–1263, 2011.
- [9] M. Khellah, N.-S. Kim, Y. Ye, D. Somasekhar, T. Karnik, N. Borkar, G. Pandya, F. Hamzaoglu, T. Coan, Y. Wang, K. Zhang, C. Webb, and V. De, “Process, temperature, and supply-noise tolerant 45 nm dense cache arrays with diffusion-notch-free (dnf) 6t sram cells and dynamic multi-vcc circuits,” *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 4, pp. 1199–1208, 2009.
- [10] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, “A variation-tolerant sub-200 mv 6-t subthreshold sram,” *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 10, pp. 2338–2348, 2008.

- [11] L.-Y. Chiou, C.-R. Huang, C.-C. Cheng, and Y.-L. Tsai, "A 300mv sub-1pj differential 6t sub-threshold sram with low energy and variability resilient local assist circuit," in *Next-Generation Electronics (ISNE), 2013 IEEE International Symposium on*, pp. 337–340, 2013.
- [12] M. Khayatzadeh and Y. Lian, "Average-8t differential-sensing subthreshold sram with bit interleaving and 1k bits per bitline," 2013.
- [13] M.-H. Tu, J.-Y. Lin, M.-C. Tsai, C.-Y. Lu, Y.-J. Lin, M.-H. Wang, H.-S. Huang, K.-D. Lee, W.-C. Shih, S.-J. Jou, and C.-T. Chuang, "A single-ended disturb-free 9t subthreshold sram with cross-point data-aware write word-line structure, negative bit-line, and adaptive read operation timing tracing," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 6, pp. 1469–1482, 2012.
- [14] M. Nabavi and M. Shams, "A gate sizing and transistor fingering strategy for subthreshold cmos circuits," *IEICE Electronics Express*, vol. 9, no. 19, pp. 1550–1555, 2012.
- [15] D. Markovic, C. C. Wang, L. P. Alarcon, T.-T. Liu, and J. M. Rabaey, "Ultralow-power design in near-threshold region," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 237–252, 2010.
- [16] A. Wang, B. H. Calhoun, and A. P. Chandrakasan, *Sub-threshold Design for Ultra Low-Power Systems (Series on Integrated Circuits and Systems)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc, 2006.
- [17] A. Teman, A. Mordakhay, J. Mezhibovsky, and A. Fish, "A 40-nm sub-threshold 5t sram bit cell with improved read and write stability," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 59, no. 12, pp. 873–877, 2012.
- [18] K. Ishibashi and K. Osada, *Low Power and Reliable SRAM Memory Cell and Array Design*. Springer, 2011.
- [19] A. Pavlov and M. Sachdev, *CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies: Process-Aware SRAM Design and Test*. Springer, 2008. 2008924192.
- [20] B. H. Calhoun and A. Chandrakasan, "A 256-kb 65-nm sub-threshold sram design for ultra-low-voltage operation," *Solid-State Circuits, IEEE Journal of*, vol. 42, no. 3, pp. 680–688, 2007.
- [21] F. Moradi, G. Panagopoulos, G. Karakonstantis, D. Wisland, H. Mahmoodi, J. K. Madsen, and K. Roy, "Multi-level wordline driver for low power srams in nano-scale cmos technology," in *Proceedings of the 2011 IEEE 29th International Conference on Computer Design, ICCD '11*, (Washington, DC, USA), pp. 326–331, IEEE Computer Society, 2011.
- [22] M. E. Sinangil, N. Verma, and A. Chandrakasan, "A reconfigurable 8t ultra-dynamic voltage scalable (u-dvs) sram in 65 nm cmos," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 11, pp. 3163–3173, 2009.

- [23] B. Zimmer, S. O. Toh, H. Vo, Y. Lee, O. Thomas, K. Asanovic, and B. Nikolic, "Sram assist techniques for operation in a wide voltage range in 28-nm cmos," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 59, no. 12, pp. 853–857, 2012.
- [24] M. Khellah, Y. Ye, N. S. Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb, and V. De, "Wordline & bitline pulsing schemes for improving sram cell stability in low-vcc 65nm cmos designs," in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, pp. 9–10, 2006.
- [25] N. Verma and A. Chandrakasan, "A 256 kb 65 nm 8t subthreshold sram employing sense-amplifier redundancy," *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 1, pp. 141–149, 2008.
- [26] F. Bauer, G. Georgakos, and D. Schmitt-Landsiedel, *A Design Space Comparison of 6T and 8T SRAM Core-Cells*, vol. 5349 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009.
- [27] A. K. Singh, *Digital Vlsi Design*. PHI Learning Pvt. Ltd, 2010.
- [28] M.-H. Chang, Y.-T. Chiu, S.-L. Lai, and W. Hwang, "A 1kb 9t subthreshold sram with bit-interleaving scheme in 65nm cmos," in *Low Power Electronics and Design (ISLPED) 2011 International Symposium on*, pp. 291–296, 2011.
- [29] M. Jagasivamani, "Development of a low-power sram compiler," Master's thesis, Virginia Polytechnic Institute and State University, 2000.
- [30] J. P. Kulkarni, A. Goel, P. Ndai, and K. Roy, "A read-disturb-free, differential sensing 1r/1w port, 8t bitcell array," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 9, pp. 1727–1730, 2011.
- [31] T.-H. Kim, J. Liu, J. Keane, and C. H. Kim, "A 0.2 v, 480 kb subthreshold sram with 1 k cells per bitline for ultra-low-voltage computing," *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 2, pp. 518–529, 2008.
- [32] T.-H. Kim, J. Liu, and C. H. Kim, "A voltage scalable 0.26 v, 64 kb 8t sram with v lowering techniques and deep sleep mode," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 6, pp. 1785–1795, 2009.
- [33] M.-E. Hwang and K. Roy, "A 135mv 0.13w process tolerant 6t subthreshold dtmos sram in 90nm technology," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, pp. 419–422, 2008.
- [34] S. Soleimani-Amiri, A. Afzali-Kusha, and A. Sammak, "Design and analysis of a new sub-threshold dtmos sram cell structure," in *Computer Architecture and Digital Systems (CADS), 2012 16th CSI International Symposium on*, pp. 50–53, 2012.
- [35] J. P. Kulkarni and K. Roy, "Ultralow-voltage process-variation-tolerant schmitt-trigger-based sram design," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 20, no. 2, pp. 319–332, 2012.

- [36] G. Pasandi and S. M. Fakhraie, "A new sub-threshold 7t sram cell design with capability of bit-interleaving in 90 nm cmos," in *Electrical Engineering (ICEE), 2013 21st Iranian Conference on*, pp. 1–6, 2013.
- [37] N. BAI, X. WU, J. YANG, and L. SHI, "A robust high density 7t sram bitcell for subthreshold applications," *Chinese Journal of Electronics*, vol. 20, no. 2, pp. 243–246, 2011.
- [38] M.-F. Chang, S.-W. Chang, P.-W. Chou, and W.-C. Wu, "A 130 mv sram with expanded write and read margins for subthreshold applications," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 2, pp. 520–529, 2011.
- [39] Z. Liu and V. Kursun, "Characterization of a novel nine-transistor sram cell," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 16, no. 4, pp. 488–492, 2008.
- [40] R. V. Joshi, R. Kanj, and V. Ramadurai, "A novel column-decoupled 8t cell for low-power differential and domino-based sram design," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 5, pp. 869–882, 2011.
- [41] J.-J. Wu, Y.-H. Chen, M.-F. Chang, P.-W. Chou, C.-Y. Chen, H.-J. Liao, M.-B. Chen, Y.-H. Chu, W.-C. Wu, and H. Yamauchi, "A large v/v_{dd} tolerant zigzag 8t sram with area-efficient decoupled differential sensing and fast write-back scheme," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 4, pp. 815–827, 2011.
- [42] V. Sharma, S. Cosemans, M. Ashouei, J. Huiskens, F. Catthoor, and W. Dehaene, "Ultra low power litho friendly local assist circuitry for variability resilient 8t sram," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012*, pp. 1042–1047, 2012.
- [43] I.-J. Chang, J.-J. Kim, S. P. Park, and K. Roy, "A 32 kb 10t sub-threshold sram array with bit-interleaving and differential read scheme in 90 nm cmos," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 2, pp. 650–658, 2009.
- [44] M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Ohbayashi, Y. Nakase, and H. Shinohara, "A 45nm 0.6v cross-point 8t sram with negative biased read/write assist," in *VLSI Circuits, 2009 Symposium on*, pp. 158–159, 2009.
- [45] A. Bhattacharyya, *Compact MOSFET Models for VLSI Design*. Wiley, 2009.
- [46] N. Shigyo and T. Hiraoka, "A review of narrow-channel effects for {STI} mosfet's: A difference between surface- and buried-channel cases," *Solid-State Electronics*, vol. 43, no. 11, pp. 2061 – 2066, 1999.
- [47] A. Wang, B. Calhoun, and A. Chandrakasan, *Sub-threshold Design for Ultra Low-Power Systems*. Integrated Circuits and Systems, Springer, 2006.
- [48] K. Roy, S. Mukhopadhyay, and 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.

- [49] T. P. Ryan, *Modern Engineering Statistics*. Wiley, 2007.
- [50] K. Iniewski, *Advanced Circuits for Emerging Technologies*. Wiley, 2012.
- [51] MATLAB, *version 8.3.0.532 (R2014a)*. Natick, Massachusetts: The MathWorks Inc., 2014.

Appendix A

Static Noise Margin Simulation Results

A.1 Hold Operation

Figure A.1: Histogram for the hold operation of the 8-cell block configuration. V_{DD} is 300mV and the temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. This is similarly true for all sub-threshold operating voltages.

Figure A.2: Histogram for the hold operation of the traditional 8T cell. V_{DD} is 300mV and the temperature is 27°C.

Figure A.3: Histogram for the hold operation of the standard 6T cell. V_{DD} is 300mV and the temperature is 27°C.

Table A.1: 1-cell Configuration

Hold SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	50.9937	9.16324	14.34074	5.1775
250 mV	74.6007	8.94069	38.83794	29.89725
275 mV	86.3046	8.87616	50.79996	41.9238
300 mV	97.9107	8.81488	62.65118	53.8363
350 mV	120.97	8.80096	85.76616	76.9652

Table A.2: 2-cell Configuration

Hold SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	51.2282	9.03835	15.0748	6.03645
250 mV	74.7571	8.83082	39.43382	30.603
275 mV	84.947	9.47921	47.03016	37.55095
300 mV	97.9994	8.74372	63.02452	54.2808
350 mV	121.015	8.76163	85.96848	77.20685

Table A.3: 4-cell Configuration

Hold SNM				
V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	51.4155	8.93937	15.65802	6.71865
250 mV	74.8944	8.73782	39.94312	31.2053
275 mV	84.9524	9.47568	47.04968	37.574
300 mV	98.0853	8.67716	63.37666	54.6995
350 mV	121.064	8.72073	86.18108	77.46035

Table A.4: 8-cell Configuration

Hold SNM				
V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	51.5421	8.86813	16.06958	7.20145
250 mV	74.9925	8.67106	40.30826	31.6372
275 mV	86.6172	8.64551	52.03516	43.38965
300 mV	98.1519	8.62609	63.64754	55.02145
350 mV	121.105	8.68699	86.35704	77.67005

Table A.5: Traditional 8T

Hold SNM				
V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	51.8192	8.51484	17.75984	9.245
250 mV	75.085	8.53884	40.92964	32.3908
275 mV	86.6686	8.57489	52.36904	43.79415
300 mV	98.1805	8.58752	63.83042	55.2429
350 mV	121.116	8.67459	86.41764	77.74305

Table A.6: Standard 6T

Hold SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	51.8192	8.51484	17.75984	9.245
250 mV	75.085	8.53884	40.92964	32.3908
275 mV	86.6456	8.59021	52.28476	43.69455
300 mV	98.1805	8.58752	63.83042	55.2429
350 mV	121.116	8.67459	86.41764	77.74305

A.2 Write Operation

Figure A.4: Histogram for the write operation of the 8-cell block configuration. V_{DD} is 200mV and the temperature is 27°C.

Table A.7: 1-cell Configuration

Write SNM				
V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	65.6669	11.7205	18.7849	7.0644
250 mV	92.4534	13.5829	38.1218	24.5389
275 mV	105.513	14.259	48.477	34.218
300 mV	118.309	14.8241	59.0126	44.1885
350 mV	143.12	15.6947	80.3412	64.6465

Figure A.5: Histogram for the write operation of the 8-cell block configuration. V_{DD} is 250mV and the temperature is 27°C.

Figure A.6: Histogram for the write operation of the 8-cell block configuration. V_{DD} is 300mV and the temperature is 27°C.

Figure A.7: Histogram for the write operation of the 8-cell block configuration. V_{DD} is 350mV and the temperature is 27°C.

Figure A.8: Histogram for the single-ended write operation of the traditional 8T cell. V_{DD} is 300mV and the temperature is 27°C.

Figure A.9: Histogram for the write operation of the traditional 8T/standard 6T cell. V_{DD} is 300mV and the temperature is 27°C.

Table A.8: 2-cell Configuration

Write SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	65.4857	11.7298	18.5665	6.8367
250 mV	92.3955	13.5898	38.0363	24.4465
275 mV	105.481	14.2639	48.4254	34.1615
300 mV	118.291	14.8275	58.981	44.1535
350 mV	143.114	15.6964	80.3284	64.632

Table A.9: 4-cell Configuration

Write SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	65.1295	11.7474	18.1399	6.3925
250 mV	92.2802	13.6036	37.8658	24.2622
275 mV	105.417	14.2737	48.3222	34.0485
300 mV	118.254	14.8343	58.9168	44.0825
350 mV	143.102	15.6998	80.3028	64.603

Table A.10: 8-cell Configuration

Write SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	64.4401	11.7793	17.3229	5.5436
250 mV	92.0519	13.6304	37.5303	23.8999
275 mV	105.288	14.2931	48.1156	33.8225
300 mV	118.254	14.8343	58.9168	44.0825
350 mV	143.077	15.7065	80.251	64.5445

Table A.11: Traditional 8T

Write SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	104.893	17.1443	36.3158	19.1715
250 mV	129.194	15.7165	66.328	50.6115
300 mV	153.462	14.8217	94.1752	79.3535
350 mV	177.011	14.314	119.755	105.441

Table A.12: Traditional 8T-SE

Single Ended Write SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	82.4992	9.7812	43.3744	33.5932
250 mV	106.214	11.171	61.53	50.359
275 mV	117.755	11.4396	71.9966	60.557
300 mV	129.233	11.7245	82.335	70.6105
350 mV	152.115	12.412	102.467	90.055

Table A.13: Standard 6T

Write SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	104.893	17.1443	36.3158	19.1715
250 mV	129.194	15.7165	66.328	50.6115
300 mV	153.462	14.8217	94.1752	79.3535
350 mV	177.011	14.314	119.755	105.441

A.3 Read Operation

Figure A.10: Histogram for the read operation of the 8-cell block configuration. V_{DD} is 300 mV and the temperature is 27°C. Only the 8-cell configuration is included because all configurations have very similar distributions. This is similarly true for all sub-threshold operating voltages.

Figure A.11: Histogram for the read operation of the traditional 8T cell. V_{DD} is 300mV and the temperature is 27°C.

Table A.14: 1-cell Configuration

Read SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	48.1531	10.7043	5.3359	-5.3684
250 mV	72.4415	10.0431	32.2691	22.226
275 mV	84.3788	9.85534	44.95744	35.1021
300 mV	96.2311	9.72389	57.33554	47.61165
350 mV	119.704	9.53813	81.55148	72.01335

Figure A.12: Histogram for the read operation of the standard 6T cell. V_{DD} is 300 mV and the temperature is 27°C.

Table A.15: 2-cell Configuration

Read SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	45.7816	12.0834	-2.552	-14.6354
250 mV	71.7763	10.3138	30.5211	20.2073
275 mV	83.8431	10.2638	42.7879	32.5241
300 mV	96.6968	9.37254	59.20664	49.8341
350 mV	119.488	9.64976	80.88896	71.2392

Table A.16: 4-cell Configuration

Read SNM				
V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	45.5673	11.5122	-0.4815	-11.9937
250 mV	70.5921	10.7534	27.5785	16.8251
275 mV	82.9585	10.5468	40.7713	30.2245
300 mV	95.1573	10.2375	54.2073	43.9698
350 mV	119.074	9.85372	79.65912	69.8054

Table A.17: 8-cell Configuration

Read SNM				
V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	42.955	12.2036	-5.8594	-18.063
250 mV	68.641	11.3526	23.2306	11.878
275 mV	81.3352	10.9978	37.344	26.3462
300 mV	95.1245	10.2317	54.1977	43.966
350 mV	118.313	10.191	77.549	67.358

Table A.18: Traditional 8T

Read SNM				
V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	51.8192	8.51484	17.75984	9.245
250 mV	75.085	8.53884	40.92964	32.3908
275 mV	86.6686	8.57489	52.36904	43.79415
300 mV	98.1805	8.58752	63.83042	55.2429
350 mV	121.116	8.67459	86.41764	77.74305

Table A.19: Standard 6T

Read SNM

V_{DD}	μ	σ	$\mu - 4\sigma$	$\mu - 5\sigma$
200 mV	16.0453	23.4658	-77.8179	-101.284
250 mV	26.2094	23.139	-66.3466	-89.4856
275 mV	31.6987	22.8435	-59.6753	-82.5188
300 mV	37.3789	22.4316	-52.3475	-74.7791
350 mV	49.3672	21.0806	-34.9552	-56.0358

Appendix B

Copyright Permissions

B.1 Intel Copyright Permission

Automatic permission to Sukneet Basuta:

Please consider this message a grant of permission but only in the manner you describe in your request.

This message provides a limited copyright license, the scope of which is implied by your original request. No other intellectual property licenses (trademark, patent or otherwise) are granted. All materials are provided by Intel ‘AS IS,’ with no warranties whatsoever. Any disputes over the license granted herein will be governed by California law.

Thank you for visiting our site. Intel Corporation.

Regards,

Intel Copyright Permission Department.

B.2 IEEE Copyright Permission

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line ©2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line ©[Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior authors approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: ©[year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.