# Robust Low-power Digital Circuit Design in nano-CMOS Technologies

**Submitted by**

**Touqeer Azam**

**School of Engineering**

**University of Glasgow**

**In fulfilment of the requirements to award the degree of**

**Doctor of Philosophy in Electronics and Electrical Engineering.**

**Nov 2010**

# Dedicated to

**My parents and teachers for all their support and encouragement …**

# Abstract

Device scaling has resulted in large scale integrated, high performance, low-power, and low cost systems. However the move towards sub-100 nm technology nodes has increased variability in device characteristics due to large process variations. Variability has severe implications on digital circuit design by causing timing uncertainties in combinational circuits, degrading yield and reliability of memory elements, and increasing power density due to slow scaling of supply voltage. Conventional design methods add large pessimistic safety margins to mitigate increased variability, however, they incur large power and performance loss as the combination of worst cases occurs very rarely.

*In-situ* monitoring of timing failures provides an opportunity to dynamically tune safety margins in proportion to on-chip variability that can significantly minimize power and performance losses. We demonstrated by simulations two delay sensor designs to detect timing failures in advance that can be coupled with different compensation techniques such as voltage scaling, body biasing, or frequency scaling to avoid actual timing failures. Our simulation results using 45 nm and 32 nm technology BSIM4 models indicate significant reduction in total power consumption under temperature and statistical variations. Future work involves using dual sensing to avoid useless voltage scaling that incurs a speed loss.

SRAM cache is the first victim of increased process variations that requires handcrafted design to meet area, power, and performance requirements. We have proposed novel 6 transistors (6T), 7 transistors (7T), and 8 transistors (8T)-SRAM cells that enable variability tolerant and low-power SRAM cache designs. Increased sense-amplifier offset voltage due to device mismatch arising from high variability increases delay and power consumption of SRAM design. We have proposed two novel design techniques to reduce offset voltage dependent delays providing a high speed low-power SRAM design. Increasing leakage currents in nano-CMOS technologies pose a major challenge to a low-power reliable design. We have investigated novel segmented supply voltage architecture to reduce leakage power of the SRAM caches since they occupy bulk of the total chip area and power. Future work involves developing leakage reduction methods for the combination logic designs including SRAM peripherals.

# Acknowledgments

Firstly, I would like to thank my supervisor, Prof. David R.S. Cumming, for his support and guidance throughout my PhD studies. I felt very comfortable with his method of supervision. His thoughtful comments and suggestions helped me a lot in improving my technical and writing skills. I would also like to pay thanks to Prof. Asen Asenov and Dr. Scott Roy for their useful feedback and kind support. Dr. Binjie Cheng has given me a lot of technical advice during my research. I am very thankful for his insight comments and suggestions that really helped me improve my designs. I will also like to thank Sonia Paluchowski for her kind advice and help during her stay in the department.

I would like to pay special thanks to James Grant for his time to review my thesis. I am very grateful for his useful comments and suggestions. Also I will like to thank Pete for his comments and feedback. Not to forget is to thank my friends Feng Hong, Mohammed Al-Rawhani, Anis S.M. Zain, and Sandeep Manjunath. Feng has always helped me with his IT support role and we had lots of useful discussions on variability. Thanks to Mohammed for his help in doing the SRAM cell layouts using Cadence. Thanks to Sandeep for his suggestions and advice during my PhD.

Finally I would like to say ‘many thanks’ to my beloved parents for their never ending support and encouragement. They have given great motivation that keeps me driving in every field. Especially I would like to thank my dad for his insight and understanding at times when I needed his support the most.

# Publications

### Journal Papers

1. Azam, T. and D.R.S. Cumming, *Efficient sensor for robust low-power design in nano-CMOS technologies.* Electronics Letters, 2010. 46(11): p. 773-775.
2. Azam, T., et al., *Robust asymmetric 6T-SRAM cell for low-power operation in nano-CMOS technologies.* Electronics Letters, 2010. 46(4): p. 273-274.

### Conference Papers

1. Azam, T., B. Cheng, and D.R.S. Cumming. *Variability resilient low-power 7T-SRAM design for nano-scaled technologies*. in *Quality Electronic Design (ISQED), 2010 11th International Symposium on*. 2010.
2. Azam, T. and D.R.S. Cumming. *Robust low power design in nano-CMOS technologies*. in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*. 2010.

# Table of Content

**Abstract** ..... **i**

**Acknowledgments** ..... **ii**

**Publications** ..... **iii**

**Table of Content** ..... **iv**

**List of Figures** ..... **vii**

**List of Tables** ..... **xii**

***Chapter 1*** ..... ***1***

***1. Introduction*** ..... ***1***

**1.1 Motivation** ..... **1**

**1.2 Aims and objectives** ..... **3**

**1.3 Thesis outline** ..... **3**

***Chapter 2*** ..... ***5***

***2. CMOS variability, challenges and solutions*** ..... ***5***

**2.1 Types of variability** ..... **7**

2.1.1 Systematic variations ..... 7

2.1.2 Non-systematic variations ..... 7

2.1.2.1 Die-to-Die (global or inter-die) variations ..... 8

2.1.2.2 With-in-Die (local or intra die) variations ..... 8

**2.2 Sources of variability** ..... **8**

2.2.1 Static variability ..... 9

2.2.1.1 Random Discrete Dopant (RDD) fluctuations ..... 9

2.2.1.2 Line-edge-roughness (LER) ..... 10

2.2.1.3 Oxide thickness variations ..... 11

2.2.2 Dynamic variability ..... 11

**2.3 Impact of variability on design** ..... **12**

2.3.1 Frequency and leakage variations ..... 12

2.3.2 SRAM reliability ..... 13

2.3.3 Device wear out & degradation ..... 14

2.3.4 Testing and fault modeling ..... 14

2.3.5 Hard logical faults ..... 15

2.3.6 Soft error rate ..... 16

**2.4 Variability tolerant design techniques** ..... **17**

2.4.1 *In-situ* design ..... 17

2.4.1.1 Error detection methods ..... 18

2.4.1.2 Error prediction methods ..... 19

2.4.2 SRAM cell design ..... 20

2.4.2.1 6T SRAM designs ..... 21

2.4.2.2 7T-SRAM design ..... 23

2.4.2.3 8T-SRAM designs ..... 23

2.4.2.4 9T and 10T-SRAM designs ..... 24

2.4.3 Mitigation of the sense amplifier offset voltage ..... 25

2.4.3.1 Conventional transistor sizing ..... 26

2.4.3.2 Digital trimming ..... 27

2.4.3.3 Tunable sense amplifier design ..... 27

2.4.4 SRAM cache leakage reduction techniques ... 28
2.4.4.1 Novel SRAM cell topologies ... 29
2.4.4.2 Back biasing techniques ... 30
2.4.4.3 Power gating methods ... 31
2.4.4.4 Drowsy cache designs ... 32
**2.5 Chapter summary ... 33**
***Chapter 3 ... 35***
***3. In-situ design techniques ... 35***
**3.1 *In-situ* monitoring of timing failures ... 37**
**3.2 A 45 nm delay sensor ... 39**
3.2.1 Proposed 45 nm sensor design ... 39
3.2.2 Soft error correction ... 42
3.2.3 Simulation results ... 43
3.2.3.1 Temperature Variations ... 45
3.2.3.2 Statistical Variations ... 48
3.2.4 Area and power comparison ... 54
**3.3 A 32 nm delay sensor ... 55**
3.3.1 Proposed 32 nm sensor design ... 55
3.3.2 Simulation results ... 60
3.3.2.1 Statistical variations ... 60
3.3.2.2 Temperature variations ... 61
**3.4 Chapter summary ... 64**
***Chapter 4 ... 67***
***4. Variability resilient SRAM designs ... 67***
**4.1 Standard 6T-SRAM design ... 68**
**4.2 SRAM stability metrices ... 70**
4.2.1 Read margin ... 70
4.2.2 Write margin ... 71
4.2.3 Hold margin ... 72
4.2.4 Cell current ... 73
**4.3 An asymmetric 6T-SRAM design ... 73**
4.3.1 Proposed asymmetric 6T-SRAM cell ... 74
4.3.2 Simulation results for the proposed A- 6T SRAM design ... 78
4.3.2.1 Noise margins comparison ... 78
4.3.2.2 Power and delay comparison ... 82
**4.4 An SNM free 7T-SRAM design ... 83**
4.4.1 Proposed 7T-SRAM cell ... 83
4.4.2 Simulation results of a 45 nm 7T-SRAM design ... 88
4.4.2.1 Noise margins comparison ... 88
4.4.2.2 Power and delay comparison ... 92
**4.5 Fully differential 8T-SRAM design ... 94**
4.5.1 Proposed 8T-SRAM cell design ... 95
4.5.2 Simulation results of a 45nm 8T-SRAM design ... 98
4.5.2.1 Noise margins comparison ... 98
4.5.2.2 Read/write delay comparison ... 101
4.5.2.3 Energy comparison ... 103
**4.6 Summary and conclusion ... 104**
***Chapter 5 ... 106***

*5. Sense-amplifier offset voltage mitigation techniques ..... 106*

**5.1 Background to SRAM sense operation ..... 108**

**5.2 Impact of statistical variations on SRAM read delay ..... 111**

**5.3 Proposed discharge assist design ..... 115**

5.3.1 Proposed discharge assist circuit ..... 115

5.3.2 Statistical variability simulation results ..... 120

5.3.3 Energy and area comparisons ..... 123

**5.4 Proposed pre-charge select design ..... 126**

5.4.1 Stability analysis ..... 131

5.4.2 Statistical variability simulations ..... 135

5.4.3 Energy and area comparisons ..... 137

**5.5 Chapter summary ..... 138**

*Chapter 6 ..... 140*

*6. SRAM cache leakage power reduction ..... 140*

**6.1 Types of MOS transistor leakage ..... 141**

6.1.1 Sub-threshold leakage ..... 141

6.1.1.1 Drain induced barrier lowering (DIBL) ..... 142

6.1.2 Gate oxide leakage ..... 142

**6.2 SRAM cell leakage mechanisms ..... 143**

**6.3 Proposed segmented supply voltage method for leakage power reduction ..... 144**

6.3.1 Architecture of the proposed segmented supply voltage design ..... 149

**6.4 Simulation results and discussion ..... 151**

6.4.1 Read noise margins ..... 151

6.4.2 Leakage reductions ..... 153

6.4.3 Impact on discharge delay and power consumption ..... 154

**6.5 Chapter summary ..... 155**

*Chapter 7 ..... 156*

*7. Conclusion and future works ..... 156*

**7.1 Future work ..... 160**

**References ..... 161**

**Appendix 1: Acronyms ..... 170**

# List of Figures

Figure 2.1: Origins and manifestations of variability. ..... 7
Figure 2.2: Sketch of a 20nm MOSFET having less than 50 dopants in the channel [23]. ..... 10
Figure 2.3: Line edge roughness of 6nm of a 30 x30 nm MOSFET [23]. ..... 10
Figure 2.4: Heat flux across in Watts per square centimetre across a die [5]. ..... 11
Figure 2.5: Impact of variations on microprocessor's frequency and leakage power [36]. ..... 13
Figure 2.6: Threshold voltage of CMOS inverter gate lengths 35 nm (mean=510 mV STD=28 mV). ..... 16
Figure 2.7: Inverter threshold voltages for different supply voltages. ..... 16
Figure 2.8: Razor flip-flop design. ..... 18
Figure 2.9: Built In Soft Error Resilience (BISER) flip-flop design. ..... 19
Figure 2.10: Canary flip-flop design. ..... 20
Figure 2.11: Single ended 6T-SRAM cell. ..... 22
Figure 2.12: Single ended sub-threshold 6T-SRAM. ..... 22
Figure 2.13: SNM free 7T-SRAM cell design. ..... 23
Figure 2.14: An 8T-SRAM cell design. ..... 24
Figure 2.15: A 10T-SRAM cell for high SNM and low bit-line leakage. ..... 24
Figure 2.16: Impact of transistor sizing on failure probability for current latch sense amplifier (CLSA) and voltage latch sense amplifier (VLSA). [72] ..... 26
Figure 2.17: Digitally trimmed sense amplifier design. ..... 27
Figure 2.18: Tuneable sense amplifier design. ..... 28
Figure 2.19: Low leakage asymmetric 6T-SRAM cell ..... 29
Figure 2.20: Reverse body biasing of SRAM cell. ..... 30
Figure 2.21: Gated-VDD SRAM cell. ..... 31
Figure 2.22: Drowsy cache design. ..... 32
Figure 3.1: Frequency distribution of a typical design. ..... 37
Figure 3.2: Circuit diagram of the Canary flip-flop. ..... 38
Figure 3.3: Circuit level implementation of the (a) sensor (b) error generation circuit. ..... 40
Figure 3.4: Timing diagram of sensor operation in multiple clock cycles. Data signal makes transition in the guard band in the second clock cycle, and consequently different values are stored in both flip-flops. An error signal is flagged in the third clock cycle. ..... 41

Figure 3.5: Application of the proposed sensor in a pipelined system. ..... 41
Figure 3.6: Sensor design with soft error correction. ..... 43
Figure 3.7: Design flow for gate level simulation of temperature and statistical variability. .. 45
Figure 3.8: Power consumption for two different design methods. ..... 46
Figure 3.9: Error plot at different temperature and voltage conditions. ..... 47
Figure 3.10: Power consumption at different voltage and temperature conditions. ..... 47
Figure 3.11: Relation between error rate and power consumption at different temperature conditions. ..... 48
Figure 3.12: Error rate comparison for two extreme instances of the CSM. ..... 49
Figure 3.13: Error rate plot with decreasing supply voltage for randomized CMS circuits. ... 50
Figure 3.14: Relation between power reduction and error rate for CSM (worst case)..... 51
Figure 3.15: Relation between guard band and the selected supply voltage (a) GB=0.1T Average power per cycle = 19.6uW. (b) Guard band= 0.2T, Average power per cycle = 15.7uW (c) Guard band= 0.3TAverage power per cycle = 13.3uW (d) GB=0.4T. Average power per cycle = 11.5uW. ..... 53
Figure 3.16: Impact of guard band on the average supply voltage of a 30 stage inverter chain. ..... 54
Figure 3.17: Area and power overhead vs. guard band for both test circuits. ..... 55
Figure 3.18: Timing diagram illustrating pre/post-sampling of data. ..... 56
Figure 3.19: Circuit implementation of (a) sensor (b) error generation circuit. ..... 58
Figure 3.20: Timing diagram of sensor operation. ..... 59
Figure 3.21: Application of the sensor in different pipeline stages. ..... 60
Figure 3.22: Inverter chain simulation under statistical variability. ..... 61
Figure 3.23: Plot of actual and pre-detected errors. ..... 62
Figure 3.24: Average power per cycle at different temperatures. ..... 64
Figure 3.25: Relation between power consumed and error rate at different temperatures with decreasing supply voltage. ..... 64
Figure 4.1: Standard 6T-SRAM design (a) 6T-SRAM cell (b) array architecture. ..... 69
Figure 4.2: SNM of standard 6T-SRAM. ..... 71
Figure 4.3: WNM of standard 6T-SRAM. ..... 72
Figure 4.4: Hold noise margin of standard 6T-SRAM. ..... 73
Figure 4. 5: Circuit schematic (a) conventional 6T-SRAM cell (b) proposed 6T-SRAM cell. 76

Figure 4.6: Timing diagram HSPICE simulation (a) conventional 6T-SRAM (b) proposed asymmetric 6T-SRAM. ..... 77
Figure 4.7: Cell area comparison (a) symmetric 6T-SRAM cell (b) proposed asymmetric 6T-SRAM cell. ..... 78
Figure 4.8: SNM comparison (a) Butterfly curves (b) SNM vs. Supply voltage plot. ..... 80
Figure 4.9: Noise margins comparison (a) SNM symmetric 6T-SRAM (b) SNM proposed asymmetric 6T-SRAM (c) WNM symmetric 6T-SRAM (d) WNM proposed asymmetric 6T-SRAM. ..... 81
Figure 4.10: Power and delay comparison (a) write delay (b) power. ..... 83
Figure 4.11: Circuit design of the proposed 7T-SRAM (a) cell schematic (b) row configuration (c) timing diagram. ..... 86
Figure 4.12: Cell area comparison (a) Layout of the conventional 6T-SRAM cell (b) Layout of the proposed 7T-SRAM cell. ..... 87
Figure 4.13: Noise margins comparison (a) SNM (b) WNM. ..... 89
Figure 4.14: Impact of supply voltage scaling on SNM for different cell ratios. ..... 90
Figure 4.15: SNM comparison (a) standard 6T-SRAM, CR=1.5 (b) proposed 7T-SRAM, CR=1 ..... 91
Figure 4.16: Read failure due to high statistical variability. ..... 91
Figure 4.17: WNM comparison under statistical variability (a) standard 6T-SRAM, CR=1.5 (b) proposed 7T-SRAM, CR=1. ..... 92
Figure 4.18: Write delay comparison of standard 6T and proposed 7T SRAM designs ..... 93
Figure 4.19: Power consumption comparison of standard 6T and proposed 7T SRAM design. ..... 94
Figure 4.20: Circuit schematic (a) conventional 8T-SRAM cell (b) proposed 8T-SRAM cell. ..... 96
Figure 4.21: Timing diagram (a) conventional 8T-SRAM (b) proposed differential 8T-SRAM. ..... 97
Figure 4.22: Layout proposed 8T-SRAM cell. ..... 98
Figure 4.23: SNM plot of both SRAM cell designs (6T and 8T). ..... 99
Figure 4.24: Write stability comparison (a) WNM margins without variability (b) WNM conventional 8T-SRAM (c) WNM proposed 8T-SRAM. ..... 101
Figure 4.25: Delay comparison of a 45 nm 64X32 bit SRAM design (a) read operation (b) write operation. ..... 103

Figure 4.26: Energy comparison of conventional and proposed 8T-SRAM designs. ..... 104
Figure 5.1: Circuit schematic of a conventional 6T-SRAM cell. ..... 108
Figure 5.2: Current mode sense amplifier (a) circuit schematic (b) timing diagram (c) HSPICE simulation. ..... 110
Figure 5.3: The impact of variability on read delay (a) Timing diagram and (b) simulation result. ..... 113
Figure 5.4: Discharge delay relation with offset voltage and cell current (a) discharge delay variations (b) percentage contributions of the offset voltage and cell currents to total discharge delay. ..... 114
Figure 5.5: Proposed discharge assist circuit (a) Circuit schematic and (b) timing diagram (c) HSPICE simulation assisted discharge case 1 (d) HSPICE simulation assisted discharge case 2. ..... 118
Figure 5.6: Impact of discharge assist transistors on read delay. ..... 119
Figure 5.7: Discharge current distributions (a) case 2 AT=1 (b) case 2 AT=2 (c) case 2 AT=3 (d) case 1. ..... 121
Figure 5.8: 256x*N* SRAM array setup for statistical variability simulation. ..... 122
Figure 5.9: Error rate comparison at different discharge delays. ..... 123
Figure 5.10: Area comparison (a) conventional sized sense amplifier Area=46.8x11.1 $\mu m^2$ =519.5 $\mu m^2$ (b) sense amplifier for proposed design Area= 21.6x11.1 $\mu m^2$ =240 $\mu m^2$ (c) proposed discharge assist circuit Area=7.4x11.4 $\mu m^2$ = 82 $\mu m^2$. ..... 126
Figure 5.11: Proposed pre-charge select design (a) array structure (b) circuit schematic (c) timing diagram (d) HSPICE simulation. ..... 131
Figure 5.12: Impact of pre-charge voltage on SNM. ..... 132
Figure 5.13: Stability analysis (a) open loop SRAM cell (b) simulation results. ..... 133
Figure 5.14: Inverter threshold plot under statistical variations (a) DC plot (b) PDF of inverter threshold. ..... 134
Figure 5.15: Impact of low pre-charge on offset voltage. ..... 135
Figure 5.16: Offset voltages (a) conventional design (b) proposed design for $n\sigma_{offset}$ = 46 mV, n=1 (c) proposed design $n\sigma_{offset}$ =92 mV, n=2. ..... 137
Figure 5.17: Proposed pre-charge select circuit Area= $(16.4x7.2 - 8.9 - 2)\mu m^2 = 100.28\mu m^2$. ..... 138
Figure 6.1: Leakage paths in an unselected SRAM cell. ..... 143

Figure 6.2: Segmented drowsy cache cell. ..... 145
Figure 6.3: Hierarchal decoding to select 16 segments each of 16 words from a 45 nm 256 words array (a) architecture (b) decoder delay simulation. ..... 147
Figure 6.4: Wakeup latency of raising the virtual ground. ..... 148
Figure 6.5: Proposed virtual supply voltage architecture. ..... 149
Figure 6.6: Detailed implementation of the control circuit for segmented supply architecture. ..... 150
Figure 6.7: SNM analysis in active mode at different supply voltages. ..... 152
Figure 6.8: Dynamic retention voltage (DRV) when subjected to statistical variability. ..... 152
Figure 6.9: Leakage power reduction for a 16x128 bits SRAM cache segment. ..... 153
Figure 6.10: Increase in discharge delay with decrease in pre-charge voltage. ..... 155

# List of Tables

Table 2. 1: Principles of device scaling in nano technologies ..... 6

Table 3.1: Specification of the test circuits. ..... 44

Table 4. 1: Transistor sizing for 45 nm 64x32 bit SRAM ..... 93

Table 5. 1: Energy comparison ..... 124

# Chapter 1

# 1. Introduction

## 1.1 Motivation

The semiconductor industry has benefited from the relentless scaling of metal-oxide-semiconductor (MOS) transistors for four decades, doubling number of transistors per unit area in each new generation that follows the famous Moore's Law [1]. Aggressive scaling has lead to vast adaptation of highly dense, high performance, low power, and low cost systems. Additional improvements in functionality were possible by increasing the die sizes and utilizing the large logic density available on a chip. Although the Moore's law has helped in a phenomenal growth of the semiconductor industry, it now faces serious challenges of intrinsic device variability that exists even under tight process control [2]. MOS transistors show large deviation in their electrical behaviour due to anomalies in manufacturing process and an increase of intrinsic device variability that arises due to discreetness of the charge and matter [3, 4]. Other challenges to the continuous scaling are large dynamic variations, aggressive wear out mechanisms and the increasing soft error rate [5].

The traditional method to cope with the increased variability, for combinational logic circuits, is to add pessimistic safety margins in the form of higher supply voltages [6, 7] or lower clock frequencies [8, 9]. However these methods incur a large power/ performance overhead as worst case conditions happen very rarely and most of the chips meet desired design targets. New adaptive designs are therefore required that minimize the pessimistic margins and allow a fully functional design using unreliable transistors.

SRAM caches represent an important part of modern processors as they have an increasingly large influence on the system speed and power consumption [10]. Standard 6T-SRAM cells are carefully designed to achieve a balance between conflicting read and write

requirements [11]. Increased variations can easily destroy this balance and cause different kinds of failures. New cell topologies are therefore required for a robust SRAM design under variability. A sense amplifier is used to detect a small differential voltage developed at bit-lines during a SRAM read operation and convert it to a full rail output voltage, thereby enhance system speed and power consumption. Variability in the sense amplifier circuit introduces an offset voltage that needs to be overcome by the bit-line differential voltage to enable a reliable sense operation [12], thereby limiting the power/performance of SRAM design. Conventional sizing methods incur high energy and area overheads [13, 14], therefore new design techniques are required to mitigate the SRAM sense amplifier offset voltage.

Leakage power consumption represents an another challenge to the further scaling of SRAM [15]. Supply voltage scaling is required for the reliability concerns as devices are scaled down. This requires a proportionate scaling of the device threshold voltage to achieve the performance gains. However lower threshold voltages lead to high leakage power consumption. Drowsy architectures provide a method to decrease the power consumption in the idle periods [16, 17]. However, they incur a significant latency and energy overhead. A low overhead leakage power reduction method is therefore needed for the low-power circuit operation in nano-CMOS technologies.

This thesis explores the above mentioned areas of research. We proposed novel *in-situ* designs for the combinational circuits [18, 19], novel SRAM cell topologies [20, 21], sense amplifier offset mitigation methods, and a low leakage-power SRAM architecture that provide a foundation for the robust low-power nano-CMOS design. We have used 45 nm BSIM4 models from the University of Glasgow to include statistical variability (random dopant fluctuations, line-edge-roughnesses, and poly-grain variations) [4, 22, 23]. However these models don't include the temperature variations, therefore we used 32 nm PTM models from the Arizona State University [24] to include temperature variations in our delay sensor designs. In addition, 65 nm PTM device and interconnect models were used for the asymmetric 6T-SRAM design to include interconnect capacitances. These interconnect models were scaled to get approximate capacitances for the 45 nm designs. In addition, we have used the 350 nm Austria Micro System (AMS) technology in different layouts for area comparisons.

## 1.2 Aims and objectives

The aim of this work is to develop low-power reliable digital circuit designs in the face of increased variability in nano-CMOS technologies. The following are the key areas addressed in this research,

1. To develop a new design methodology for the combinational logic circuits that can minimize increasingly pessimistic design margins.

2. To develop novel SRAM cell topologies that are more robust to statistical variability as compared to the standard SRAM design.

3. To improve the SRAM discharge delays by minimizing the SRAM sense amplifier effective offset voltage.

4. To develop a new architecture that minimizes the leakage power consumption of the SRAM arrays.

## 1.3 Thesis outline

The rest of the thesis is organized as follows. The second chapter presents some background information to variability, its impact on design, and previously proposed methods to counter variability. Different types of static variations (including random discrete dopant fluctuations, line edge roughness, and oxide thickness variations) and dynamic variations (including temperature and IR drop variations) are discussed. Impact of variability on design including frequency, leakage, SRAM design, soft errors, and hard logic faults is explored in details. Finally we provide an overview of previous research in the areas of *in-situ* design, SRAM, sense amplifiers, and SRAM leakage power reduction.

The third chapter focuses on low-power reliable circuit operation for the combinational logic circuits. It presents an introduction to the *in-situ* monitoring of timing failures to reduce

worst case design margins. A 45 nm and a 32 nm delay sensor are then proposed to detect timing failures in advance. Design, implementation, and simulation results under statistical and temperature variations are described for both sensors.

Chapters 4-6 focus on robust low-power circuit operation for the sequential circuits and discuss novel SRAM cell designs, SRAM sense amplifier offset voltage mitigation methods and SRAM array leakage power reduction techniques. In chapter four, we present an asymmetric 6T-SRAM, SNM free 7T-SRAM, and a fully differential 8T-SRAM design. An efficient application of write and read assist circuits helps achieve higher noise margins for these designs. HSPICE simulations results are presented for noise margin comparisons under statistical variations.

The fifth chapter describes the background to the SRAM sense amplifier and its impact on the SRAM read delays. It also presents two novel digital methods to minimize the sense amplifier offset voltage dependent delays. A pre-charge select and a discharge assist technique are proposed to minimize the effective offset voltage for the sense amplifiers. Design, implementation, and HSPICE simulation results are described in detail. HSPICE simulations are carried out under statistical variations using the 45 nm BSIM4 models from the University of Glasgow.

The sixth chapter focuses on leakage power reduction for the SRAM array. Considering the fact that the SRAM cache takes a major portion of the total chip area, leakage reduction for the SRAM has therefore high impact on the total power reduction. A segmented supply voltage architecture is presented to reduce the leakage power of SRAM arrays during a drowsy mode. HSPICE simulation results are presented for the energy reductions and the performance overheads.

Chapter 7 concludes this work. A summary of the work done in previous chapters is presented and future directions are laid out.

# Chapter 2

# 2. CMOS variability, challenges and solutions

The great success of the semiconductor industry can be attributed to the scaling of devices to lower dimensions in order to achieve large integration, improved performance, and reduced power consumption at a lower cost. Scaling has resulted in a 0.7X reduction of the vertical and lateral dimensions of MOS transistors in each successive generation that has translated to doubling the number of transistors on the same die area [25]. Increasing the die size improved the total transistor count by 3.3X for each process node. Scaling devices increased the clock speed by 1.4X, whereas the use of additional transistor logic further improved it to 1.7X. Table 2. 1 illustrates the scaling trend that has been followed across different process technologies [26]. This has enabled current high performance processors to operate up to 3-4 GHz [10] with 1.72 billion transistors on each chip [27]. However higher clock frequencies lead to an increase of the power consumption by 60% with every 400 MHz rise in speed [28]. While the dynamic power consumption per transistor has decreased with scaling, the total power consumption per chip has increased due to a large die size [29]. The supply voltage scaling and the use of multi-core processors have helped achieve high performance gains due to transistor size scaling without increasing the power dissipation.

As the device scaling moves to sub-100 nm technologies, CMOS devices show considerable spread in their characteristics that result in more dies failing to meet design specifications (power or performance budget). Those dies that don’t meet desired targets, even if they are functional, are either discarded or sold at a lower price which results in a low yield and less revenue [30]. While the threshold voltage has scaled to achieve performance goals, the variations in the threshold voltage have increased that resulted in the percentage variations

to rise with each new technology node [31]. Process variations have a severe impact on the performance, power consumption, reliability, and yield of the VLSI chips. Decreasing the yield and increasing the cost per die lowers the effectiveness of scaling devices to nano dimensions.

**Table 2. 1: Principles of device scaling in nano technologies**

| Scaling parameters | Relationship | Constant field scaling (S>1) |
|---|---|---|
| Width, W<br>Length, L<br>Oxide thickness, $T_{ox}$ | | $1/S$ |
| Supply voltage, VDD,<br>Threshold voltage, $V_{th}$ | | $1/S$ |
| Device area, A | WL | $1/S^2$ |
| Gate oxide capacitance per unit area, $C_{ox}$ | $1/T_{ox}$ | S |
| Gate capacitance, $C_{Gate}$ | $WLC_{ox}$ | $1/S$ |
| Saturation current, $I_{sat}$ | $WVC_{ox}$ | $1/S$ |
| On resistance, $R_{on}$ | $V/I_{sat}$ | 1 |
| Device delay, $\tau$ | $R_{on} C_{gate}$ | $1/S$ |
| Power, P | $I_{sat} V$ | $1/S^2$ |

Variability can be categorized as static (zero time variation) or dynamic that changes device behaviour with time [32]. Static variations (random dopants, line-edge-roughness, oxide thickness, etc.) and dynamic variations (temperature, IR drop) result in variations in the electrical behaviour (gate capacitance, threshold voltage, saturation current, etc.) of the CMOS devices [33]. Variations in the electrical parameters of the devices results in large variations in delay and power consumption of the logic gates that leads to an unreliable system with a low-yield. Figure 2.1 illustrates the origin and manifestations of variability at different levels of abstraction.

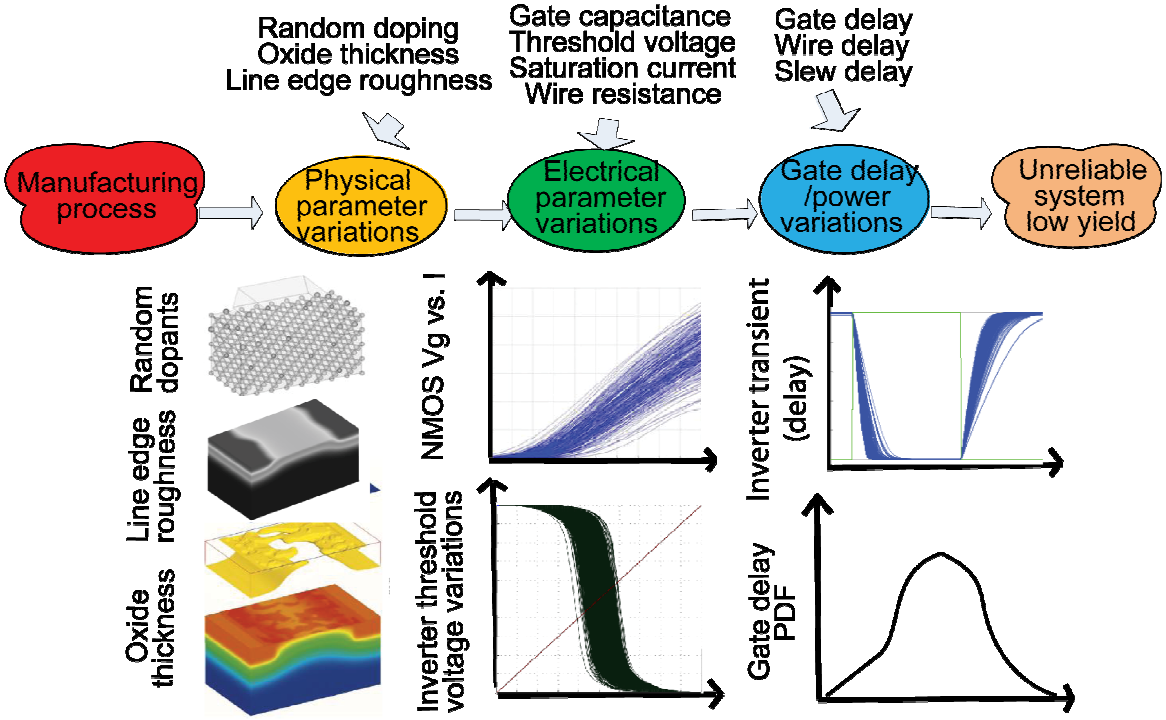



**Figure 2.1: Origins and manifestations of variability.**
***(Figures of parameter variations are taken from [23])**

## 2.1 Types of variability

Variability can also be categorized in two basic types depending on if it is possible to predict it from the layout or not, further classification is made on the spatial reach of variability [33].

### 2.1.1 Systematic variations

Systematic variations are deterministic and can be predicted in advance by analyzing the layout. Even when the transistors may have the same gate length or width, there exists a clear difference in their layout or neighbourhood [32]. These include variations due to the optical proximity effects, chemical mechanical polishing (CMP), and metal fills.

### 2.1.2 Non-systematic variations

Non-systematic variations have a statistical nature and therefore can't be predicted in advance before manufacturing. However they can be represented by random variables to model their statistical characteristic. Further classification of the statistical variability exists

on the basis of its spatial reach to identify the root causes of variations and possible improvement methods [32].

#### 2.1.2.1 Die-to-Die (global or inter-die) variations

"Die-to-die" variations manifest due to the processing shifts that occur from lot to lot, wafer to wafer, and reticle to reticle [33]. However they have a similar impact on all devices on the same die that result in similar electrical parameter variations. All devices may have shorter or longer gate lengths than the mean on a certain die due to the die-to-die variations, but this effect may be different on some other die. Experimental results of a 29 stage ring oscillator frequency indicate that 67% of the total frequency variations arise due to the die-to-die variations in a 90 nm process [34].

#### 2.1.2.2 With-in-Die (local or intra die) variations

"Within die" variations arise from the processing shifts that occur across each die, therefore each device may be affected differently. As an example, the unwanted process shifts can cause different devices to have different oxide thicknesses on the same die. Within die variations can be correlated as devices in neighbourhood suffer a similar process shifts as compared to the far ones. Certain within die variations are totally independent from each other and can cause even neighbourhood transistors to behave quite differently. For independent variations, knowing the characteristic of a transistor doesn't provide any useful information about others. It includes variations due to random discrete dopants (RDD) and line edge roughness (LER).

## 2.2 Sources of variability

Variability can be classified as static or dynamic depending on the sources of variation. Static variability arises from the manufacturing process and occurs during fabrication, whereas the dynamic variability is time and context dependent [5].

### 2.2.1 Static variability

Static variability can be broadly categorized as either intrinsic or extrinsic variability. The intrinsic variability arises from the discreteness of charge and matter which exists even with a tight process control [2] and has become a major limitation to the future scaling [4]. The extrinsic variability arises from the lack of tight process control or from the inability to precisely transfer the mask pattern to a wafer [32]. These include the transistor dimension variations across chip, die-to-die, and wafer-to-wafer. The intrinsic variability includes random discrete dopants (RDD), line-edge-roughness (LER), oxide thickness variations (OTV), polysilicon granularity (PoG), and high-k dielectric morphology [4, 23, 32, 35].

The intrinsic variability adversely affects the reliability of a static random access memory (SRAM), increases the timing violations, and makes the leakage current problem worse. It is believed that RDD fluctuations are the major source of intrinsic variability for channels > 18 nm channel lengths. For channels lengths $\leq$ 18 nm, LER will take over [4]. The impact of poly silicon gate granularity will increase with a further reduction of gate oxide thickness. High-K dielectrics are used in the 45 nm technologies to provide the thicker gate oxide layers in order to reduce the gate leakage currents, however the Si/High-k dielectric interface itself introduces a large variability [35]. Static variations in process parameters can cause a 20X variation in the chip leakage power and a 30% variation in the operating frequency [36].

#### 2.2.1.1 Random Discrete Dopant (RDD) fluctuations

RDD fluctuations arise from the granularity of charge and atomicity of matter [23] and therefore has a significant impact on the threshold voltage variations of the nano-CMOS devices. Channel doping controls the threshold voltage of MOS devices. Due to aggressive scaling, the number of dopants have decreased from 1000's (1 μm technology) to a few dozen (in 45 nm and below), even when the doping concentration increases with the scaling of dimensions. Considering the fact that there are around 100 dopants in the channel for a current generation (45nm) transistor [35], the number and position of dopants can make geometrically identical devices behave quite differently in the future technologies. It was found that RDD contributes 65% of the threshold variations in NMOS at 65 nm and 60% of PMOS at 45 nm [37].

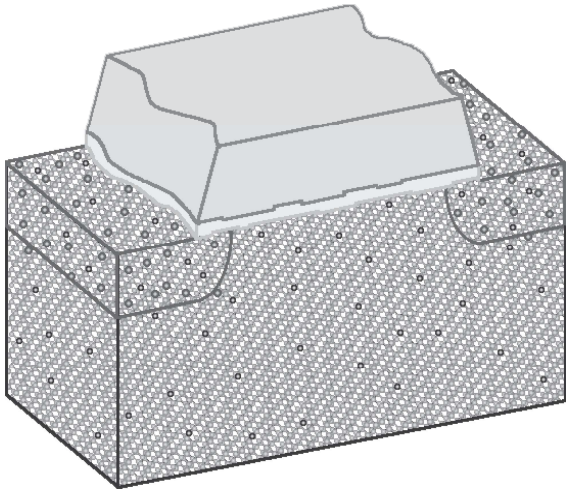


**Figure 2.2: Sketch of a 20nm MOSFET having less than 50 dopants in the channel [23].**

#### 2.2.1.2 Line-edge-roughness (LER)

Using a wavelength of 193 nm for fabricating nano scaled transistors introduces a large variations in the deep sub-micron technologies (130nm till present) and is the primary reason for LER [5]. The impact of LER is expected to supersede RDD at a gate lengths of 18 nm and below [38], until extreme ultra-violet wavelength is used for patterning devices that will minimize the LER and line-width-roughness (LWR). Even a shift from 193 nm to a lower lithography wavelength will not remove all problems, since many variations come from the step-and-repeat process that can cause stepper lens heating, lens focusing, and other aberrations [30]. LER is found to be around 5 nm and doesn't scale with the device scaling [35], therefore, the influence of LER is expected to increase with the further scaling of MOS gate length as predicted in [4, 38].

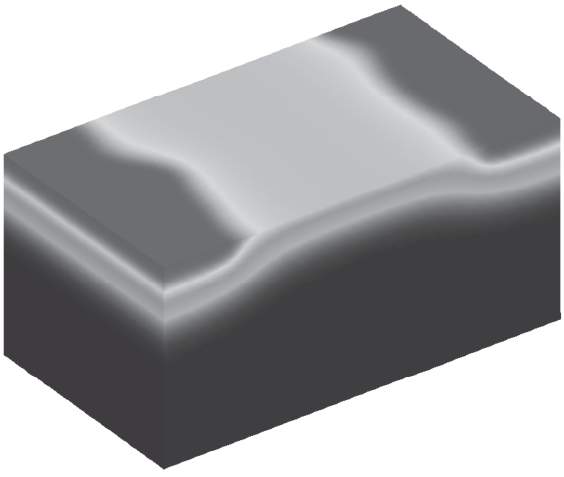


**Figure 2.3: Line edge roughness of 6nm of a 30 x30 nm MOSFET [23].**

#### 2.2.1.3 Oxide thickness variations

Gate oxide thickness is another source of high threshold variations in the deep sub-micron CMOS devices. As the length of gate oxide approaches a few atomic layers with the interface roughnesses of 1 or 2 atomic layers, oxide roughness will lead to more than a 50% variation in the oxide thickness [23, 35]. It is expected that the oxide thickness variations will cause a large threshold variations comparable to RDD for the conventional MOS devices with dimensions 30 nm and below [37].

### 2.2.2 Dynamic variability

Dynamic variability originates from temperature and voltage variations across the die. The heat flux across the chip varies as different blocks have different switching activities and loads. This uneven power dissipation results in uneven temperature variations. Blocks with a higher heat flux put more load on the power distribution network, resulting in a time dependent variations in the supply voltage [5]. This has an adverse effect on the circuit performance and sub-threshold leakage [36].

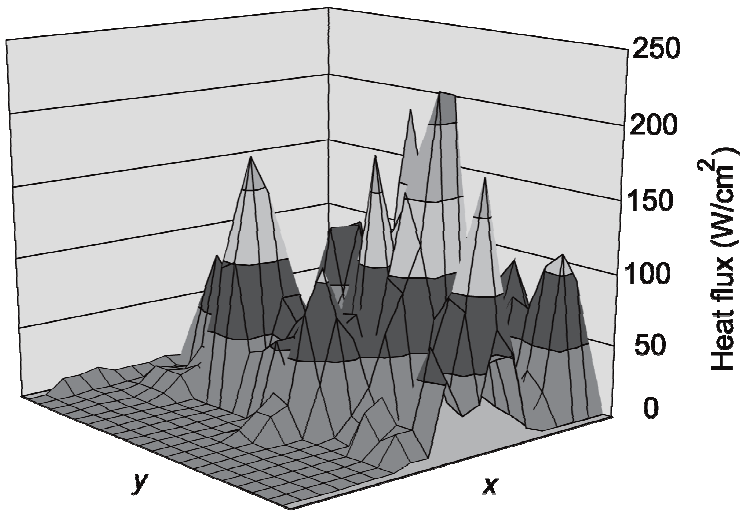



**Figure 2.4: Heat flux across in Watts per square centimetre across a die [5].**

Figure 2.4 shows heat flux across a high performance microprocessor chip [5], indicating large uneven temperature variations. The temperature variations can cause degradation of device and interconnect delays. This can cause performance mismatch between two communication blocks on a chip that can lead to a functional failure. A temperature difference of $4\,^{\circ}-5\,^{\circ}C$ can result in a 20% performance variations in modern processors [39]. The

impact of supply voltage variations will also get worse in future technologies as the small voltage fluctuations will results in large current variations.

## 2.3 Impact of variability on design

Variability in nano scaled devices leads to an increase in the unpredictability of delay and power consumption of VLSI systems. Power and delay have a negative correlation with faster devices contributing more power consumption than the slower devices. Increased variability increases this two sided constraint and therefore results in a low yield [40]. Moreover it has a severe impact on the functionally of the SRAM design, reducing its noise margins and increasing the leakage power [32]. Other aspects of devices and circuits that suffer from the increased variability are device aging [8, 41], soft errors [42, 43], and hard logical faults [44]. This section gives a brief insight of the challenges confronted due to the increased variability.

### 2.3.1 Frequency and leakage variations

Worst case delays and some safety margins are taken to set a processor clock frequency in order to obtain fully functioning chips under worst case conditions [30]. However, as increased variability will lead to the high threshold voltage variations, the delay spread of devices and circuits will rise as well. Therefore even larger design margins will be required for functional designs that will lower the performance gains from scaling. It has already been reported that the worst case margins for a reliable design are increasing due to high variability in scaled technologies [32]. A 30% variation in chip frequency has been observed due to the large process variations for a 180 nm CMOS process [5, 36].

Increased variability has a significant effect on the total power consumption encompassing both static and dynamic components. The device scaling has been accompanied with the supply voltage scaling to lower the total power consumption, and dynamic power in particular. However, an increasing threshold variation that approaches VDD (considering $6\sigma_{Vth}$ from the mean threshold voltage) will limit the VDD scaling to reduce the power consumption. Increasing leakage power due to the lower threshold voltages now accounts for a major portion of the total chip power (50% [41]). The mean off-currents (leakage) are found to have an exponential relation with the off-current variations. Therefore, the high leakage variations will increase the leakage power, and hence the total power

consumption of ICs [32]. Variability can cause 5 to 10 times variations in the leakage power, since the leakage power itself is 30 to 50% of the total power, therefore variability can cause up to a 50% variation in the total power consumption [5]. Figure 2.5 shows the impact of variability on the frequency and leakage power distributions of a microprocessor.

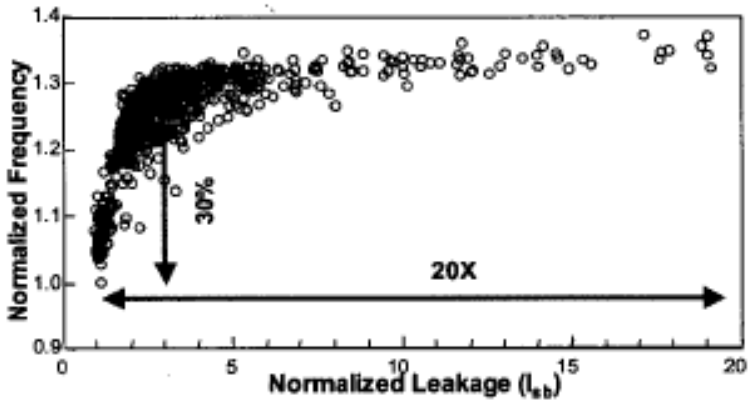



**Figure 2.5: Impact of variations on microprocessor's frequency and leakage power [36].**

### 2.3.2 SRAM reliability

The SRAM caches are an integral part of modern processors taking up to, in some cases, 90 % of the total chip area (Montecito processor) [27]. To achieve a high density level, the SRAM cells are designed using near minimum length devices. However the increased variability makes them susceptible to different kinds of failures including the read, write, and hold failures [32]. Since SRAM cells are quite weak to discharge the large capacitive bit-lines during a read operation, sense amplifiers are used to detect a small voltage differential on the bit-lines and convert it to a full rail output [31]. However the sense amplifier itself suffers from an increased variability as mismatch in its symmetrical transistors induce a large offset voltage variation. A large voltage differential, higher than the offset voltage of the sense amplifier, is required on the bit-lines to allow a reliable sensing [12]. As the required bit-line differential voltage approaches the range of supply voltages for a high reliability ($6\sigma_{\text{offset voltage}}$), the probability of the read failures also increases, incurring more power and performance overhead for the longer discharge periods. Variability also increases the minimum retention voltage needed for a non destructive hold during idle periods which increases the leakage power consumption.

### 2.3.3 Device wear out & degradation

Wear out mechanisms (NBTI, HCI) have a negative impact on performance as devices degrade with time and become slower. Although the amount of degradation varies with temperature, voltage and workload profiles for each chip, a pessimistic performance margin is added to the clock speed to obtain working chips under the worst case conditions [8]. The main degradation mechanisms are negative bias temperature instability (NBTI) and hot carrier injection (HCI) [9]. NBTI occurs due to the generation of the interface traps and positive fixed charges from the electrochemical reaction of Si-H and holes at $Si-SiO_2$ interface [8]. This decreases the driving strength of PMOS FETs and increases the device delay. HCI occurs due to the injection of hot electrons into the gate oxide of NMOS FETs that increases the threshold voltage of the devices making them slower [9].

There is an exponential dependence of different kinds of wear out mechanisms on temperature. Increasing leakage current due to large variations in the scaled technologies can increase the die temperature, and therefore lead to a faster wear out of devices. The time 1% of the processors will have failed can decrease by a 60 % depending upon the heat sink resistance and the total leakage power of the processor [41].

### 2.3.4 Testing and fault modeling

Initial burn-in testing has been a simpler and inexpensive choice to test chips after fabrication. However the reliability of the burn-in testing is threatened by the increased leakage current variations due to a low threshold voltage and its large variations in scaled technologies. Increased latent defects due to the dynamic variations (aging) and the absence of burn-in testing will reduce the effectiveness of the standard present day one time factory testing [5]. IDDQ testing is another inexpensive method of screening faulty chips. IDDQ testing is based on the principal that CMOS devices consume almost zero static current when not switching, quiescent state [27]. A faulty chip due to a metal lines short or gate oxide short consumes a few orders magnitudes higher leakage current than a fault-free chip. Therefore monitoring the power supply current during the IDDQ testing, we can distinguish faulty and fault-free dies. However, IDDQ testing method is less effective in the presence of high

leakage current. Since the failure mechanisms are changing, new testing strategies with an advanced test equipment are necessary that will incur high test overheads [45].

Different kinds of effects like process variations, fabrication defects, and high noise can appear as delay defects. Due to variability these defects have a statistical nature that threatens the effectiveness of corner based models. Therefore, new test and diagnosis methods are necessary in fields of the statistical delay fault simulations, statistical path selection for testing, statistical automatic test pattern generation, and fault diagnosis using statistically generated information [45]. Moreover it may not be possible to achieve the same level of the temperature and switching activity during testing, especially for the structured testing where circuits are broken into pieces to simulate worst case dynamic variations [39]. Therefore pessimistic or optimistic estimates of the temperature variations may affect system performance or reliability of testing, respectively.

### 2.3.5 Hard logical faults

An important challenge for the future generations will be the introduction of hard logical faults. This will happen when a gate (e.g. inverter) will not flip logic state because its threshold voltage will either be very near to the supply or ground voltage, or a small amount of noise will be sufficient to flip its output causing hard logical faults. We performed statistical variability simulations using the 45 nm device models from the University of Glasgow to investigate the inverter threshold variations shown in Figure 2.6. Although a significant amount of variation in the threshold voltage is observed there is still a large noise margin (amount of noise require to flip its output) at VDD=1V. Therefore the probability of stuck-at faults due to extreme variability is very low ($10^{-7}$ to $10^{-9}$) [44]. However, as the supply voltage is scaled to minimize the power consumption, these noise margins are severely degraded as shown in Figure 2.7. At a supply voltage of 300 mV, the $6\sigma_{\text{inverter threshold}}$ approaches VDD (300 mV) or ground (0 mV) resulting in the hard logical faults. These results indicate that variability will limit the amount of voltage scaling in the future generation due to emergence of the hard logical faults.

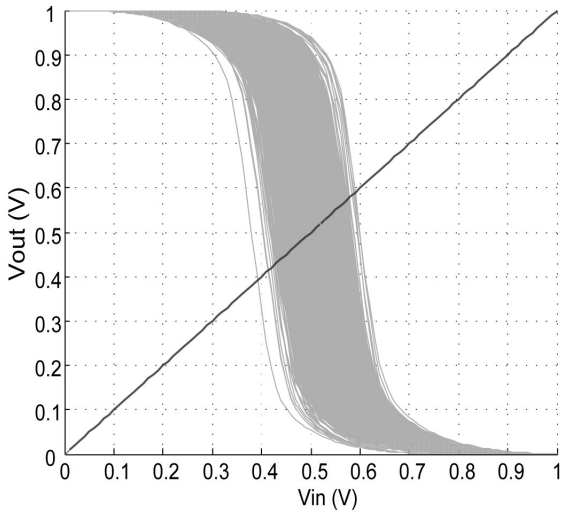



**Figure 2.6: Threshold voltage of CMOS inverter gate lengths 35 nm (mean=510 mV STD=28 mV).**

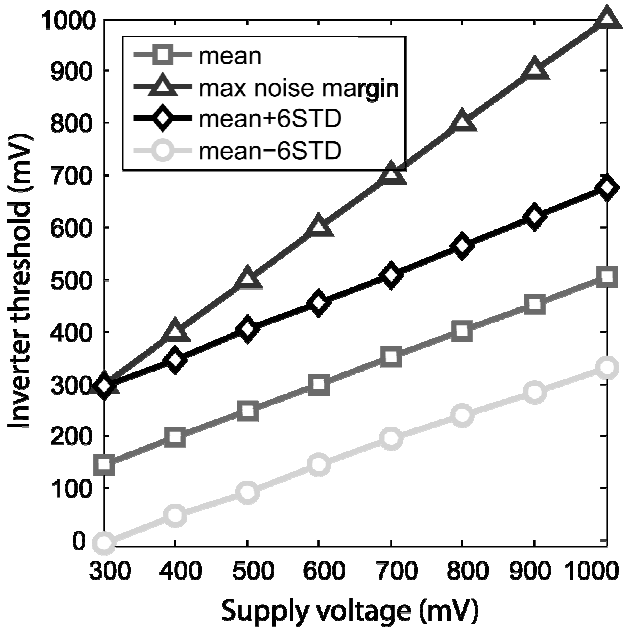



**Figure 2.7: Inverter threshold voltages for different supply voltages.**

### 2.3.6 Soft error rate

Increasing soft error rate is another area of concern for a reliable design in the scaled technologies. These errors occur due to the alpha particles emitted from the packaging materials and cosmic rays (neutrons) from space [9, 46]. In the case of combination circuits, soft errors appear as noise or glitches that can propagate to a latch element and result in a bit error, called a single event transient (SET). For memory circuits they can flip the state (bit) held by a storage element, and is called a single event upset (SEU) [9].

Although transistor scaling results in a lowering the probability of collecting the critical charge that can upset a circuit, the value of the critical charge itself decreases even faster at the lower dimensions resulting in an increasing soft error rate in scaled technologies [47]. It is expected that the soft errors will rise by 8 percent per logic state-bit with each technology generation [5]. Increased variability has a negative impact on soft errors as value of the critical charge changes with parametric variations like gate length, width, threshold voltage, and temperature variations [42]. Previous research shows that value of the critical charge can vary from a -33.5% to 81.7% compared to when no variability is taken into considerations [43]. It shows that the increasing variability will have a significant impact on the soft error rate in future generations.

## 2.4 Variability tolerant design techniques

This section provides a brief overview of the different design techniques previously proposed to provide a robust circuit operation under the increased variability. These include *in-situ* design methods to detect timing failures for the combinational circuits, and robust SRAM cell designs, sense amplifier offset mitigation techniques, and the SRAM cache leakage power reduction techniques for the sequential circuits.

### 2.4.1 *In-situ* design

The era of happy scaling is over as the future scaling of devices confronts challenges like large process and environmental variability, aggressive wear out, and increasing soft error rate, while the user demand for a reliable low-power design is even higher [5]. Variability in the device behavior arising from the process, voltage, and temperature variations results in large circuit delay variations [30]. To achieve functional dies, the worst case delay [8, 9] or the voltage margins [6, 7] are added to typical case values to account for the process and environmental variations that results in a poor performance and power loss, respectively. Since combination of the worst cases happens very rarely, *in-situ* designs provide an opportunity to tune the design margins dynamically, reducing the overhead costs and resulting in an improved yield and higher revenues.

Different sensor designs have been proposed in the past to detect timing errors that can be used with the dynamic compensation techniques (body bias, voltage scaling, and/or frequency scaling [48, 49]) to minimize the delay failures to an acceptable level. Two types of *in-situ* designs are generally presented, error detective and error predictive. Error detection methods allow errors to occur and use data dependency to provide higher energy savings, but require an error recovery mechanism. Error predictive designs detect timing failures in *advance* and therefore don't require an error recovery circuit, however the energy reductions are less. Error detection methods use the Razor flip-flop [6, 7, 50] and error predictive methods use the Canary flip-flop based approach [8, 9, 51]. Fewer sensor designs were presented that could provide soft error corrections [9, 46]. This section provides an overview of these methods.

#### 2.4.1.1 Error detection methods

The Razor flip-flop [6, 7] was designed to reduce the pessimistic voltage margins and use the delay dependence of data to drop the supply voltage well below its critical value while maintaining an acceptable error rate. The critical voltage is selected to ensure robust circuit operation under the worst case process and environmental variations. It consists of a main flip-flop equipped with a shadow latch that holds a valid data and operates at a delayed clock signal as shown in Figure 2.8. In the case of a timing error, the shadow latch keeps a correct data and the error recovery mechanism is used to restore the correct value in the main flip-flop. SPICE-level simulations indicate a substantial energy saving using this technique (up to a 64%).

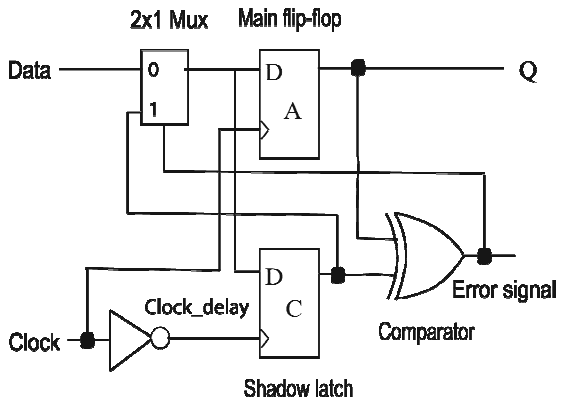



**Figure 2.8: Razor flip-flop design.**

Two major problems with the Razor flip-flop are its susceptibility to the short path delay and the meta-stability. Meta-stability represents a case when the clock signal and the data signal make a simultaneous transition that results in output voltage to hover around VDD/2. Short path delay problem puts a constraint on the minimum path delay in the Razor based designs to be longer than half the clock cycle. In addition it requires an error recovery mechanism since it allows errors to occur resulting in a high performance overhead (about 3%). Razor II [50] is a modified version of the original Razor flip-flop. Based on the architectural replay to execute an erroneous instruction, the Razor II greatly simplifies the error recovery path and reduces the complexity/size of the original Razor flip-flop. However, replaying an erroneous instruction incurs a higher Instruction-Per-Cycle (IPC) overhead.

#### 2.4.1.2 Error prediction methods

A degradation sensor was proposed to allow an early detection of the timing errors by pre-sampling data [8]. It avoids the requirement for an error recovery circuit as the timing errors are detected in advance. However, selection of the appropriate guard band remains a key issue for the aging sensor. A large guard band results in diminishing benefits while too small a guard band makes it difficult to design such circuits due to the large process variations [9]. In addition, there is no soft error protection as in the case of the Razor flip-flop.

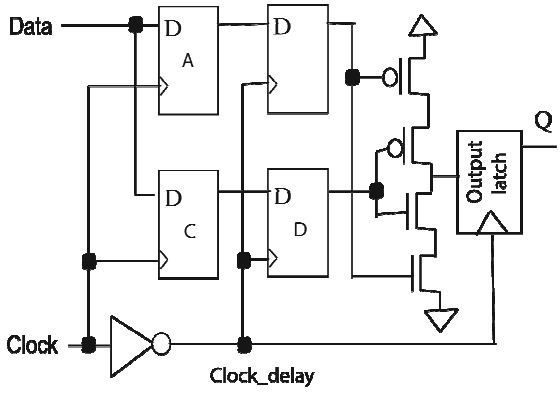



**Figure 2.9: Built In Soft Error Resilience (BISER) flip-flop design.**

Built In Soft Error Resilience (BISER) [46] was proposed to provide soft error correction without any variation detection. Moreover it has a high power (up to 10%) and performance (up to 5%) overhead. Figure 2.9 shows schematic of BISER flip-flop with soft-error correction C element. In the case of a soft error on any of the flip-flops, the outputs of both the

flip-flops differ and the C-element is turned off so that the output latch holds the correct value. Adaptive Variation-and-Error Resilient Agent (AVERA) [9] is a variant of BISER and was proposed to provide variation diagnosis, degradation detection, and soft error protection. However, it can only do one job at a time and the selection of the mode is an important issue that needs to be addressed.

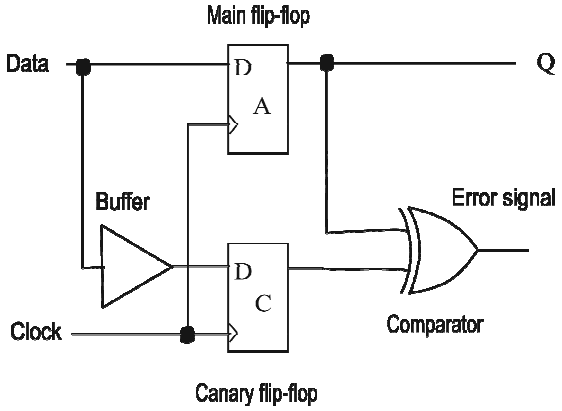



**Figure 2.10: Canary flip-flop design.**

The Canary flip-flop [51] was presented to provide the pre-detection of a timing failure using a delay buffer at the data input. Figure 2.10 shows the circuit diagram of the Canary flip-flop. It consists of a main flip-flop and a Canary flip-flop to store a delayed input using a delay buffer at the data input. The Canary flip-flop doesn't require an error recovery mechanism and eliminates the need of a delay line which simplifies the clock tree design. However it is susceptible to invalidation due to hazards [8] that may happen when a fast switching signal on one branch increases the delay on the slow switching path resulting in a false error signal. In addition the delay buffer can have a significant area and power overhead. A modified version of the Canary flip-flop was presented in [52] to perform dual sensing in order to avoid performance loss due to the voltage oscillations. However, this would worsen its susceptibility to invalidation due to hazards and further increase the area/power overhead.

### 2.4.2 SRAM cell design

SRAM caches play a key role in modern VLSI systems by providing the highest access speeds among the embedded memories. An effective method to improve system speed is to add more SRAM cache. Increasing the cell density has resulted in the SRAM cache to occupy over 70% of the total chip area and a significant portion of the total chip power [10]. Device scaling has resulted in doubling the density of SRAM cache with every new process node. However the small dimensions of the transistors used in SRAM cells make them more

vulnerable to failures under increased process variations. Variability has a significant impact on the reliability of the SRAM read, write, and hold operations. Moreover a conventional 6T-SRAM has constrained read/write requirements as the cell is required to be weak enough to be overwritten easily while also strong enough to preserve its data during the read phase. Conflicting design requirements make it more susceptible to failures and the achievable noise immunity is limited. Although all sources of variability have a significant impact on the yield of SRAM caches, statistical variability poses a major challenge. It can cause symmetric transistors of a SRAM cell, sitting side by side, to behave quite differently and can induce different types of failures [4].

A standard 6T-SRAM cell has a poor read stability represented by the static-noise-margins (SNM). The SNM reflects the maximum noise that can be tolerated at the storage nodes without destroying cell data. Device sizing is normally used to enhance the read stability (SNM) of a SRAM cell. However conventional sizing can be ineffective in nano-scaled technologies due to the large threshold variations [53, 54]. Different SRAM designs have previously been presented that use 6T [55-58], 7T [59], 8T [60-62], 9T[63], and 10T [64-66] (T-transistors) to provide a reliable and/or low power operation. This section provides an overview of these SRAM designs.

#### 2.4.2.1 6T SRAM designs

An SNM free 6T-SRAM cell was proposed for the low-voltage applications in the scaled technologies [56, 67]. Two virtual grounds are provided to achieve expanded read and write margins as shown in Figure 2.11. Write operation is performed by turning on the write access transistor M1, while the ground terminal of the feedback inverter is floating to assist the write operation by weakening the cell storage. Read operation is performed by turning on the read assist transistor MR that allows discharge of the bit-line if the cell stores a one. However this method will increase the write delay because of a single ended operation. Moreover, it can consume large amounts of power during the write operation due to the common ground. For example, two or more cells sharing common ground store ones that will turn on the read access transistor M6 for those cells and their bit-lines are shorted. If a write operation is performed that writes zeros and ones on different bit lines, a short current will flow since the bit-lines are connected through the shared virtual ground. This can result in increased power consumption and may cause write failures.

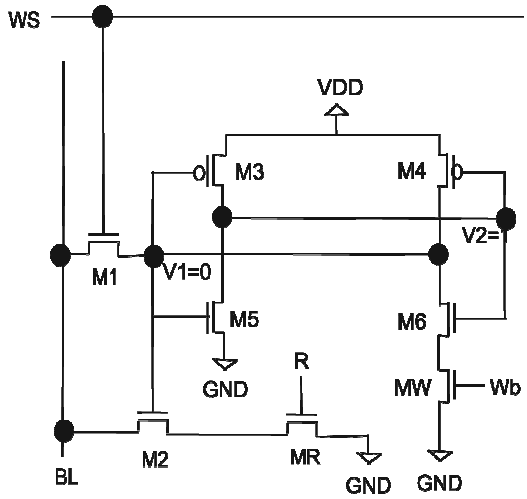



**Figure 2.11: Single ended 6T-SRAM cell.**

To achieve a design with ultra low power operation, a sub-threshold 6T-SRAM cell was proposed [55]. Virtual ground and supply terminals are provided to assist the write operation by collapsing the supply voltages of the feedback inverter as shown in Figure 2.12. However a single ended write operation is slower than a differential write operation. A driving source line 6T-SRAM cell was proposed to increase the bit-line access speed by driving the bit-line negative during the read and left floating during the write [58]. This design provided an improvement by 1/2 in the access delay and reduced the write power consumption by 1/10. However the read margins were not improved since the current ratio of the driver and access transistors remain the same as the conventional 6T-SRAM design. Moreover it requires generation of a negative voltage during the read operation that may degrade the device reliability.

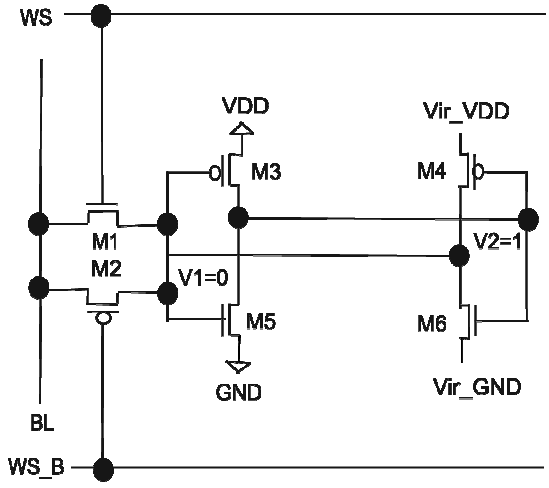



**Figure 2.12: Single ended sub-threshold 6T-SRAM.**

An asymmetric 6T-SRAM cell was proposed to reduce the leakage current for a zero state, however the SNM was degraded and the access delay was higher [68]. Another asymmetric 6T-SRAM was proposed in [57] with an enhanced read and write margins. However the improvements in the noise margins are limited with the conventional sizing due to the constrained requirements for the read and write operations.

#### 2.4.2.2 7T-SRAM design

A 7T-SRAM cell was presented for a low voltage SNM free operation [59] shown in Figure 2.13. A data protection NMOS transistor, M5, is added between node V2 and driver transistor, M7. M5 is turned off during the read access which prevents the node V2 from decreasing even when the disturbance at the node V1 is very high during a read operation. However it suffers from the dynamic retention problem [59] and the cell may loose its data for the longer read delays since the other node (V2) is floating when reading a 1. The proposed 7T-SRAM results in a 13% increase in area overhead as compared to a conventional 6T-SRAM design.

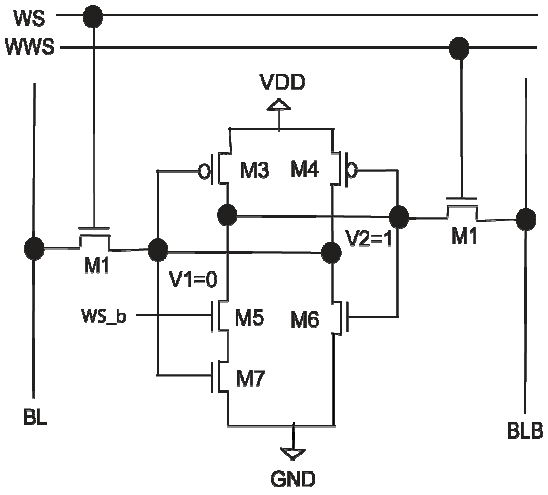



**Figure 2.13: SNM free 7T-SRAM cell design.**

#### 2.4.2.3 8T-SRAM designs

To further improve the read margins without incurring any loss of stability, 8T-SRAM cell designs have been proposed [60-62]. The idea is to use a separate port for the read operation as shown in Figure 2.14. This allows use of the minimum size NMOS driver transistors for a low leakage current and provides an SNM free operation. However the write margins remain

the same and an additional 30% [60] area overhead incurs as compared to a standard 6T-SRAM design.

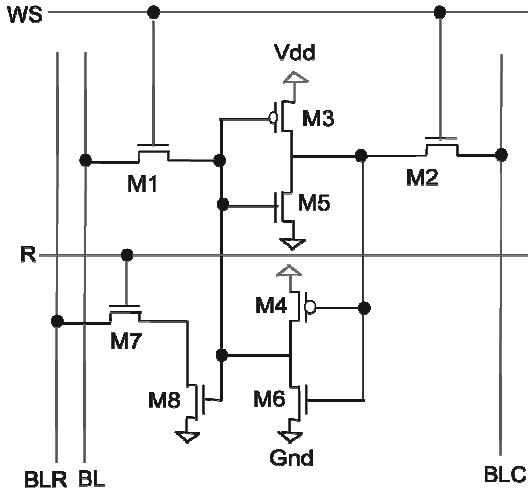



**Figure 2.14: An 8T-SRAM cell design.**

### 2.4.2.4 9T and 10T-SRAM designs

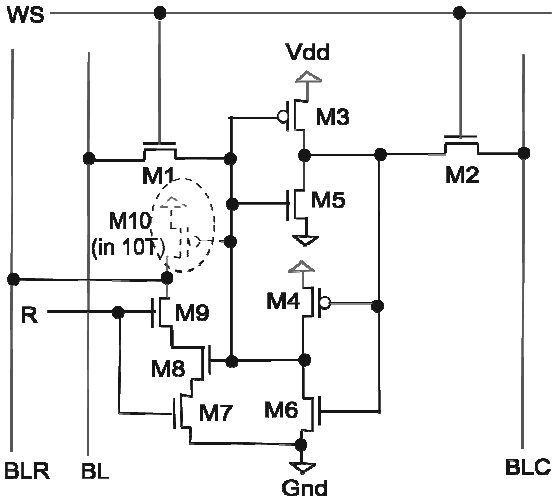



**Figure 2.15: A 10T-SRAM cell for high SNM and low bit-line leakage.**

To enable a sub-threshold ultra-low voltage operation with reduced bit-line leakage current, 10T-SRAM cell designs have been proposed [64-66] as shown in Figure 2.15. Two extra transistors are added to the conventional 8T-SRAM cells creating a strong stacking effect that significantly reduces the bit-line leakage and hence a greater number of cells can be connected to each bit-line. A 9T-SRAM was proposed that eliminates the additional PMOS, M10, in the buffer circuit of the 10T-design to achieve similar leakage reductions [63]. Their

results indicate that the additional PMOS incurs a significant standby leakage current. However both designs increase the area overhead by a 50% or more as compared to the standard 6T-SRAM design.

### 2.4.3 Mitigation of the sense amplifier offset voltage

SRAM cell sizes have reduced by a half every next process generation thereby doubling the on-chip SRAM capacity following the Moore's Law as described earlier. Near minimum length devices are used to achieve a small sized SRAM cell. However weaker transistors in the SRAM cell reduce its driving current. On the other hand, the bit-line capacitance is not scaling in proportion to the scaling of logic circuits, therefore the bit-line discharge times tend to increase [27]. A full discharge of these highly capacitive bit-lines will take a long time and have large power consumption. A sense amplifier is used to detect a small differential voltage developed at the bit-lines and convert it to a full rail output, increasing the speed and reducing the power overhead.

To ensure reliable sensor operation, the minimum differential voltage on the bit-lines must be greater than the mismatch induced offset voltage of the sense amplifier [12]. A high offset voltage will therefore require long discharge delays in order to develop the necessary large voltage differential at the bit-lines. This results in a high power and performance penalty. With the increased process variations, greater the mismatch of the symmetric sense amplifier transistors, higher is the offset voltage. Hence minimizing the required bit-line differential is considered the key to the low power SRAM design which is limited by the offset voltage margins [69]. Maximum SRAM speed is therefore limited by the weakest SRAM cell to discharge a bit-line and by the worst case offset voltage margin of the sense amplifier [70]. Large variations in nano-scaled technologies worsen the offset voltage of a sense amplifier. Due to its significant impact on the total SRAM area, speed, yield, and power, increasing offset of the sense amplifier now requires a special attention. Embedded memories face a clear challenge of the amplifier sense margins in the SRAM design as predicted by ITRS 2009 [71].

Different methods have been proposed in both the analogue and digital domains to mitigate mismatch of the symmetric designs like SRAM, differential amplifiers, comparators, etc. As the SRAM caches take a large portion of the total area and power in modern processor designs, minimizing the mismatch in SRAM design is very important for the high speed and

low power designs. This work focuses on minimizing the mismatch of the sense-amplifier circuit to reduce its offset voltage in order to minimize the energy consumption and enhance system performance. Previously conventional sizing [72, 73], digital trimming [74, 75], and a tuneable sense amplifier [14] have been used to minimize the sense amplifier offset voltage. This section will provide an overview of the previously proposed offset reduction methods.

### 2.4.3.1 Conventional transistor sizing

A conventional method is to employ large sized devices in the sense amplifier design to minimize the delay degradation that arises from a relatively slow scaling of the bit-line capacitance and achieve a near constant offset voltage across different process generations [13]. Recent studies of the sizing techniques to reduce the sense amplifier offset voltage, especially in the presence of statistical sources of variability, can be found in [72, 73]. Figure 2.16 shows the impact of transistor sizing on the failure probability of the sense amplifier. Large sized transistors increase the size of a sense amplifier circuit. Therefore scaling has a low impact on the sense amplifier circuit area as compared to scaling of the SRAM bit cell itself (reduces by half). This increases the area/power overhead of the sense circuit [70] and poses a major challenge to further scaling of SRAM [61]. Large size transistors also increase the energy consumption during sense amplifier switching. A bit-line differential voltage of less than 50 mV ($6\sigma$) is no longer economical due to high energy overhead [14]. The sense amplifier can consume over a 40% of the total energy consumption of the SRAM for a differential voltage of 50 mV ($6\sigma$) for sub-90 nm technologies.

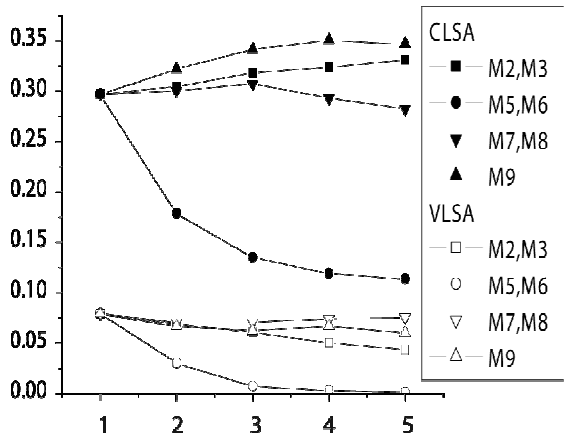



**Figure 2.16: Impact of transistor sizing on failure probability for current latch sense amplifier (CLSA) and voltage latch sense amplifier (VLSA). [72]**

### 2.4.3.2 Digital trimming

Due to the large overhead of the conventional sizing, new techniques have been proposed that perform a dynamic compensation of the sense amplifier offset voltage by analysing the post-silicon data. Digital offset compensation methods add extra transistors to the sense amplifier circuit that are turned on or off by performing a post-silicon calibration [74, 75] The idea is to use these elements (called kicks) to the balance current flow in two identical branches of the sense amplifiers to reduce the offset voltage. However, the use of additional transistors negatively impacts both the speed and power consumption of the sense amplifier circuit. Figure 2.17 shows a schematic of the digitally trimmed sense amplifier circuit. Another method is to use multiple copies (N) of the sense amplifier and select, during the calibration phase, the optimum one i.e. has a minimum offset [61]. Run time selection of the sense amplifier introduces an energy and delay overhead and the offset compensation does not improve substantially with the increasing number of redundant sense amplifiers (N-1) [14].

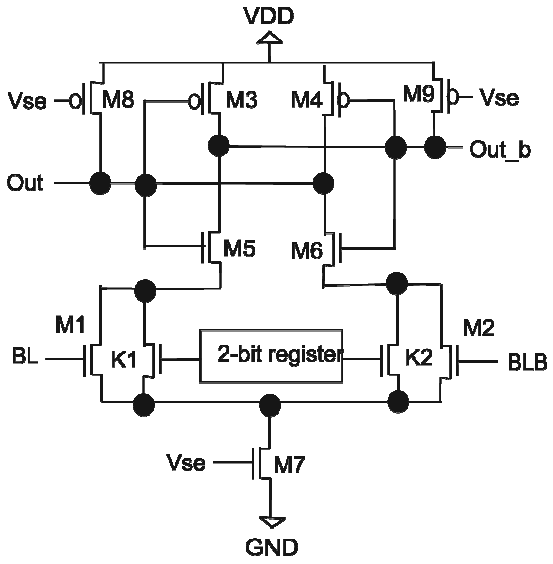



**Figure 2.17: Digitally trimmed sense amplifier design.**

### 2.4.3.3 Tunable sense amplifier design

Another method similar to the digital trimming is to employ multiple reference supply voltages. An appropriate reference voltage ( $V_{ref}$ ) is selected during calibration that minimizes the current difference in the two branches of the sense amplifier to reduce the offset voltage [14]. Figure 2.18 shows a schematic of the tuneable sense amplifier. However, the generation

of a large number of precise supply voltages creates a high overhead considering the fact that near zero offset may not be necessary or optimal for the SRAM sense amplifier [74].

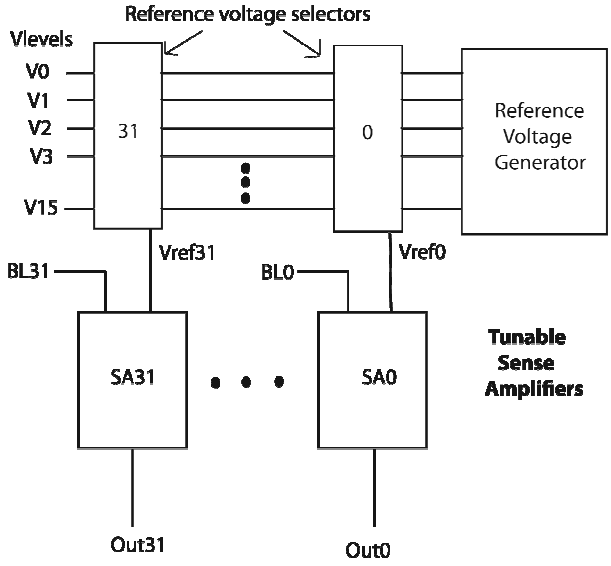



**Figure 2.18: Tuneable sense amplifier design.**

### 2.4.4 SRAM cache leakage reduction techniques

As devices are scaled and the cache density increased, the supply voltage must also scale down. This is required to maintain the device reliability and to decrease the dynamic power consumption of devices. However, lowering the supply voltage will increase the device delay. Therefore the threshold voltage of transistors must also be decreased to achieve the performance gains with scaling [76]. Due to the exponential dependence of the leakage current on the threshold voltage, a reduced threshold voltage results in an increase of the leakage power consumption. The leakage power is proportional to the number of transistors [68]. Since the SRAM cache memories now may occupy over 70% of the total chip area [10, 77], the cache leakage power consumption takes a large portion of the total power. An energy break down of the 8KB instruction cache of multimedia 32-bit RISC (M32R) embedded processor showed that the leakage power now takes over a 50% of the total power consumption at 45 nm technology node [77]. Large threshold variations will result in higher leakage current variations in SRAM arrays. Minimizing the leakage power of SRAM cache is essential in a nano-CMOS low power design due to its substantial impact on the total chip power, cooling system requirements, and reliability.

In the past, many architectural, software, circuit, and device based designs have been proposed to reduce the leakage power of the SRAM caches. This section provides a brief overview of these techniques. Our work mostly focuses on the SRAM array since the bulk of the total SRAM cache is occupied by the SRAM cell arrays that remain in an active state to hold data. Sleep stacking can be effectively used to reduce the leakage power of the SRAM periphery [78]. To reduce the leakage power of the SRAM array itself, several strategies, such as different SRAM cell topologies [65, 66, 68], back biasing techniques [79-81], power gating methods [76, 82-84] and drowsy cache design [16, 17, 85], have been proposed and investigated.

#### 2.4.4.1 Novel SRAM cell topologies

A low leakage 6T-asymmetric SRAM cell was proposed to reduce the leakage current when storing “zeros“ in SRAM cells [68]. However the energy reductions are much smaller when storing “ones”. Furthermore the topology results in an increase of the read delays and requires changes to the peripheral circuitry. The static-noise-margins (SNM) are also degraded due to the asymmetrical nature of the cell. Figure 2.19 shows a circuit schematic of the asymmetric 6T-SRAM cell.

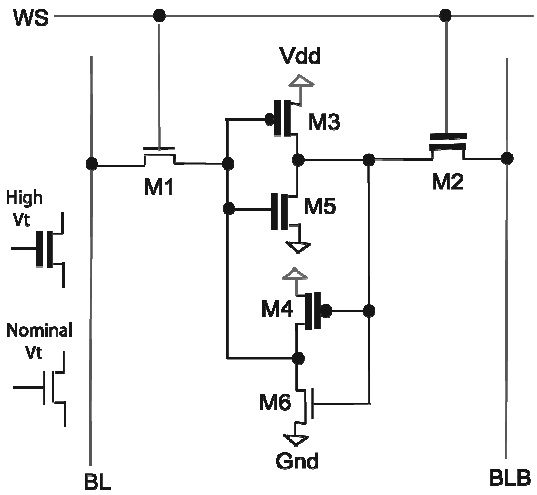



**Figure 2.19: Low leakage asymmetric 6T-SRAM cell.**

A 10T design [65, 66] was proposed to provide a sub-threshold SRAM for an ultra-low voltage operation. These cells were designed to reduce the bit-line leakage to increase the number of cells per bit-line. Each cell had an inverter buffer to provide the SNM free read operation at a reduced leakage. A variant of the 10T SRAM cell consisting of 9 transistors was proposed in [63] to lower the high area overhead. In both cases, however, the area

overhead was 50% or more as compared to a conventional 6T-SRAM design. Figure 2.15 shows schematic of the proposed 10T SRAM cell.

### 2.4.4.2 Back biasing techniques

Reverse body biasing NMOS or PMOS devices can decrease the leakage current exponentially as it increases the threshold voltage exponentially due to the body effect. This method has been used in [79-81] to reduce the leakage current of the SRAM cache. Figure 2.20 shows reverse body biasing of a 6T-SRAM cell. It doesn't impact the access delays during the discharge periods in the active mode as is the case with the gated-VDD designs [82, 83]. However a large delay and energy overhead occurs for a body transition due to the large substrate capacitance and a large $V_{body}$ swing [80]. This method is also less effective in scaled technologies due to the small body coefficient and an increase in the band-to-band tunnelling due to reverse biasing. Band-to-band tunnelling occurs when electrons tunnel through a reverse biased p-n junction under high electric fields, especially when highly doped shallow junctions are used in scaled technologies [86].

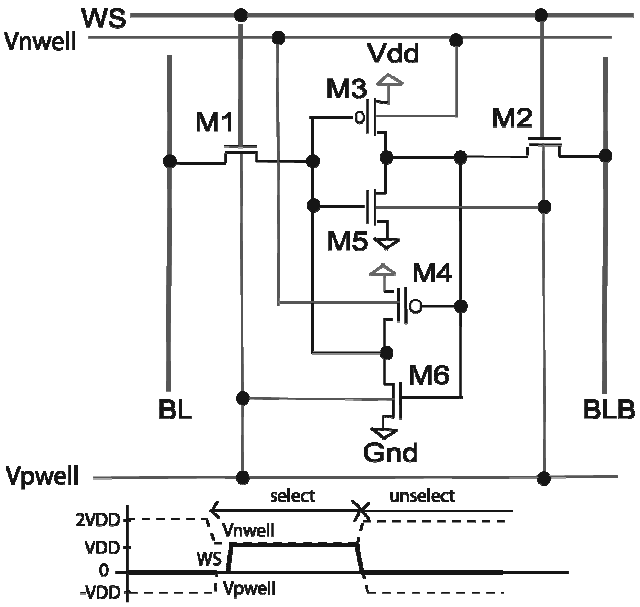



**Figure 2.20: Reverse body biasing of SRAM cell.**

To improve the back biasing, forward biasing with the high threshold voltage ($V_{th}$) transistors has been proposed [80]. The idea is to use super high-$V_{th}$ devices to suppress leakage in unselected parts of the cache and forward biasing the selected section to speed up

the read operation. However this technique also incurs a large energy and delay overhead for switching a large substrate or body capacitance between the active mode and sleep mode.

### 2.4.4.3 Power gating methods

Gated-VDD design was introduced to reduce leakage current of the unused sections of the cache by turning off the ground terminal of the unselected cells [82]. Figure 2.21 shows a gated-VDD SRAM cell. A high-$V_{th}$ transistor (M7) is inserted between the actual ground terminal (GND) and a virtual ground. It is turned on when the cell is accessed and is turned off during the idle periods to reduce the leakage current because of the stacking effect and exponential dependence of leakage on $V_{th}$ [17]. Although this method is very effective to reduce the leakage current (up to 97%), the major drawback comes from the fact that the cell losses its information when put in an idle mode. Therefore a large performance penalty may occur when data in the cache is accessed and conservative cache policies may be required. Moreover putting an extra transistor in the read discharge path increases the access delay.

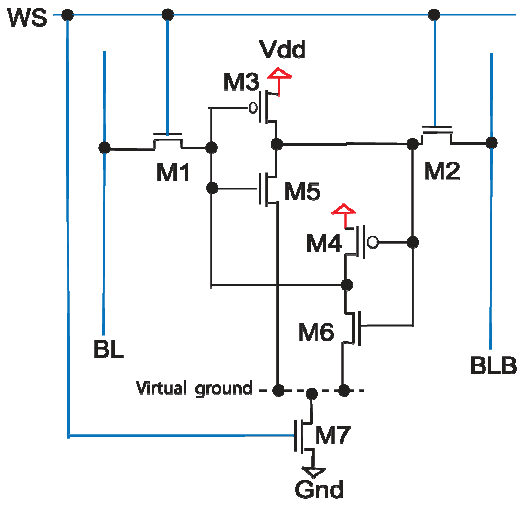



**Figure 2.21: Gated-VDD SRAM cell.**

DRG-cache was proposed [76] to provide a data retention capability to the gated-VDD design by using sophisticated sizing techniques that are sensitive to noise during the sleep periods. This also results in less energy reductions as compared to the gated-VDD caches (97% vs. 47%) with a 5% increase in the execution time. An extension of the DRG-cache is proposed in [84] where the sleep transistor is programmed to achieve the desired level of ground voltage. This reduces the rail to rail voltage on a SRAM cell and significantly reduces the leakage current. However a significant latency occurs for complete discharge of the virtual

ground terminal. A similar design was presented in [83] to provide a virtual ground to a selected segment of the cache. However the access delay was degraded by a 7%. Moreover this method requires three reference voltages and the generation of these voltages will incur a high power overhead.

#### 2.4.4.4 Drowsy cache designs

Drowsy caches [16, 17] lower the supply voltage of the un-accessed cells to reduce the leakage current in the drowsy mode, and a standard supply voltage is provided during the active mode. Drowsy mode refers to idle periods when the SRAM cache segments receive a lower supply voltage to reduce the leakage power without losing cell data. The supply voltage is kept higher than the retention voltage (200 mV-300 mV) to preserve cell data during the sleep mode. Dropping the supply voltage is effective in reducing all kinds of leakage currents, therefore significant energy savings are achieved (over 70%). Figure 2.22 shows a drowsy cache cell with the supply voltage control circuit. Moreover there is no impact on the access delay during the active period as in the case of the gated-VDD technique.

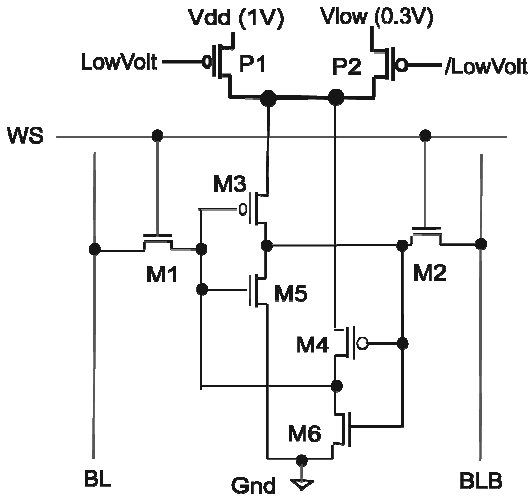



**Figure 2.22: Drowsy cache design.**

However, switching between the sleep and active modes imposes a significant latency (one to two cycles) and energy overhead. The greatest drawback comes from the increased bit-line leakage due to un-accessed cells having low node voltages as compared to the bit-lines. Using high $V_{th}$ access transistors can minimize this leakage. However this would increase the access delay. One method is to leave the bit-lines floating during the sleep period to reduce bit-line leakage current without degrading the access delay. An aggressive drowsy mode cache [85]

was proposed to maximize energy reductions without incurring wake-up latency. However, this may result in a cell data corruption during the read operation as the SRAM cell voltage can be lower than the pre-charged bit-line voltage.

## 2.5 Chapter summary

Device scaling has enabled production of largely integrated, high performance, low power, and low cost VLSI chips. However a move to sub-100 nm technologies has resulted in rise of process variations, aggressive device wear-out, and increased soft errors. This chapter provides a background to the research in the field, variability, its implications for design, and a review of previously established techniques to counter variability. Large variability poses a major challenge to future scaling. Variability can have a systematic or statistical component. While the systematic variability can be compensated to some extent by design, statistical variability (RDD, LER, oxide thickness variations) is harder to cope with. Statistical variability has worsened frequency/leakage power variations, degraded SRAM stability, and threatened the reliability of popular test methods.

*In-situ* methods have been presented to detect timing failures that use different compensation techniques to provide a robust low power operation. Novel SRAM cell topologies are discussed that increase robustness of the read/write operations in SRAM design. Degraded sense margins of the sense amplifiers present a major challenge to the future scaling of SRAM design. Adaptive digital methods to compensate for an increased offset voltage are discussed for a low energy/area overhead SRAM sense circuit. Leakage power reduction is an important area of research as it now takes up to 50% of the total power consumption in scaled technologies. Finally a brief introduction of the previously proposed SRAM leakage reduction techniques is presented.

Chapters 3 - 6 present proposed design techniques to mitigate variability that enable a robust low - power design verified by HSPICE simulations for different test benches. Chapter 3 describes two novel sensor designs to detect timing failures in advance for the combinational circuits. Voltage scaling is employed based on the timing errors information generated by the sensors to provide a robust circuit operation at a low voltage margin that reduces the power overhead. Chapter 4 presents novel 6T, 7T, and 8T SRAM cells that

provide high read/write noise margins even when subjected to high statistical variability. Chapter 5 introduces two novel techniques to reduce the effective offset voltage of the sense amplifier that minimizes the area and energy overhead as compared to a conventional design. Chapter 6 presents a new architecture to reduce leakage power of the SRAM cache array without incurring any wakeup latency. Chapter 7 concludes our work and presents future directions.

# Chapter 3

# 3. *In-situ* design techniques

The semiconductor industry has been scaling MOS transistors for decades to achieve large integration, high performance, and low power consumption. However increasing variability and device degradation in deep submicron technologies, coupled with the quest towards low power applications and stringent reliability demands provide major challenges to future scaling [5]. Increasing soft error rate due to smaller geometries further worsens this problem. Large parametric variations can lead to a excessive timing and power violations that can cause functional failures. One method to cope with the increased variability is to add worst case voltage or frequency margins, however they incur high power/performance loss. Worst case voltage designs select a critical supply voltage that ensures correct circuit operation under the worst case temperature and process variations [6, 7]. Worst case frequency designs consider pessimistic circuit delays and safety margins that ensure a correct circuit operation under large variability [8, 9]. Operating chips at a higher supply voltage significantly increases the power consumption because of its cubic dependence on the supply voltage. This may be unnecessary for most of the cases since the combination of worst case conditions happens very rarely, and bulk of the chips lie near the target frequency bin. Similarly adding large delay margins degrades system performance as the combination of worst case conditions that can violate delay constraints happen rarely. Sensor based design avoids large safety design margins, and allows most of the circuits to operate at a low supply voltage that is selected at run time and corresponds to on-chip variability.

One of the methods to detect the extent of variability or degradation is to detect timing failures of the combinational logic circuits. Circuits with large variability or degradation can cause timing failures that can be detected or predicted in advance. Different compensation techniques like dynamic voltage scaling, body biasing, or frequency scaling [48, 49] can then be used to avoid the actual timing errors. The Razor flip-flop [6, 7] was developed to detect the

minimum supply voltage that maintains an acceptable error rate. Razor is based on error detection, i.e. it allows errors to occur and therefore requires an error recovery circuit that increases the complexity of the design and results in a higher performance overhead [7]. Major problems with the Razor flip flop are its susceptibility to short path delay and meta-stability. Adaptive Variation and Error Resilient Agent (AVERA) was presented to provide the variation diagnosis, degradation detection and soft error correction [9]. But it can do only one job at a time. Therefore, selecting an appropriate mode of operation at a particular time remains a critical issue. Built in soft error resilience (BISER) [46] was proposed to provide soft error correction for the combinational and latch elements by increasing the flip-flop redundancy. However it doesn't provide any variation detection or correction. The Canary flip-flop [51] was proposed to provide the pre-detection of timing errors using a delay buffer at the data input. The Canary flip-flop doesn't require a delayed clock signal hence it simplifies the clock tree design as compared to the Razor flip-flop. However, this type of logic is more prone to invalidation due to hazards [8]. Moreover, the delay buffer can't be shared for different sensors; this will cost an extra power/area overhead. ElastIC [87] was proposed to provide a highly adaptive architecture based on aggressive self diagnosis, adaptation, and self healing that may achieve highest robustness to variation in deeply scaled technologies. However the high amount of redundancy and adaptation poses major obstacles to the practical implementation of this architecture. The details of the previously proposed *in-situ* design techniques can be found in Chapter 2.

In this chapter, we present two novel delay sensor designs (45 nm and 32 nm) that can detect timing failures for the combinational logic circuits in advance and therefore don't need an error recovery mechanism. The proposed 45 nm delay sensor uses the output of the master latch in a conventional master-slave flip-flop to pre-detect timing violations. It avoids the use of an additional delay buffer at the data input as is used in the Canary flip-flop [51]. Therefore it has a negligible impact on the combinational logic delay, reduces the power overhead, and doesn't suffer from invalidation due to hazards. Moreover it does not suffer from the short path delay and meta-stability constraints as in case with the Razor flip-flop. Since the master latch has always a positive delay, the sensor is able to pre-detect timing failures even when there are high temperature and process variations. We have extended the sensor to also provide soft error correction without the requirement of mode selection. The total performance overhead was found to be less than 0.9% for a 32-bit Carry Select Adder (CSA)

and a 16x16 Carry Save Multiplier (CSM). The power overhead was about 5.5% when the sensor was pessimistically applied on 50% of all the critical paths of the CSM. However, this overhead can be minimized by carefully applying sensor only on the most critical paths.

The proposed 32 nm delay sensor uses a main clock signal and an advanced clock signal to create a guard band for sampling data that allows detection of the timing failures in advance. This method is similar to the Razor flip-flop, however using an advanced clock signal removes the need for an error recovery circuit as errors predict timing failures but don't correspond to the actual failures. The proposed design can reduce the power consumption by 1/1.7 as compared to the worst case design. The proposed design also avoids the meta-stability and short path delay constraints and can easily be extended to provide soft error correction.

## 3.1 *In-situ* monitoring of timing failures

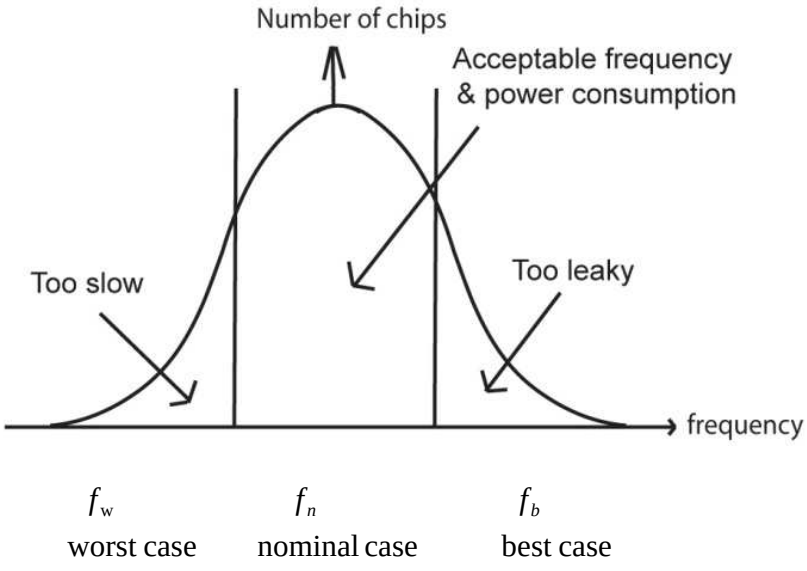



**Figure 3.1: Frequency distribution of a typical design.**

Variability in nano-CMOS devices is very high, and large safety margins are therefore needed that degrade the usefulness of scaling. Consider a sample design A that can operate at a maximum clock speed of $f_n$. Process variations cause a large spread in the circuit delay of design A and a distribution of the operating frequency is obtained as shown in Figure 3.1. Although most of the chips fall in the nominal frequency/ power bin, a significant number of

chips either have excessive leakage or are too slow. One design method is to increase the supply voltage to recover slow chips by increasing their clocking frequency. However this will unnecessarily increase power consumption of those chips which meet the frequency requirements. Another design method is to operate all chips at a lower frequency (than the nominal frequency) to increase the number of functional chips. However this will result in a loss of performance for chips which fall in the nominal and high frequency bins.

*In-situ* monitoring of the timing errors provides a mechanism to dynamically tune chip frequency or supply voltage in proportion to the on chip variations. The idea is to sample data with an early clock edge, called pre-sampling [9, 51] or sample data with a delayed clock edge, called post-sampling [6, 7]. Pre-sampling techniques are simple and have less overhead as they don't allow errors to occur. In contrast, post-sampling techniques allow errors to occur, thus permitting even more down scaling of the supply voltage for higher energy savings, but require an error recovery mechanism that has a higher performance overhead. The Canary flip-flop was presented to provide the pre-sampling of data without the need of a delay line. A buffer is put at the input to the Canary flip–flop to sample data before the main flip-flop as shown in Figure 3.2. Any mismatch between the outputs of the Canary flip-flop and the main flip-flop is flagged as an error signal. We present a 45 nm delay sensor in the next section that extends the idea of Canary flip-flop and incurs very low power/performance overhead.

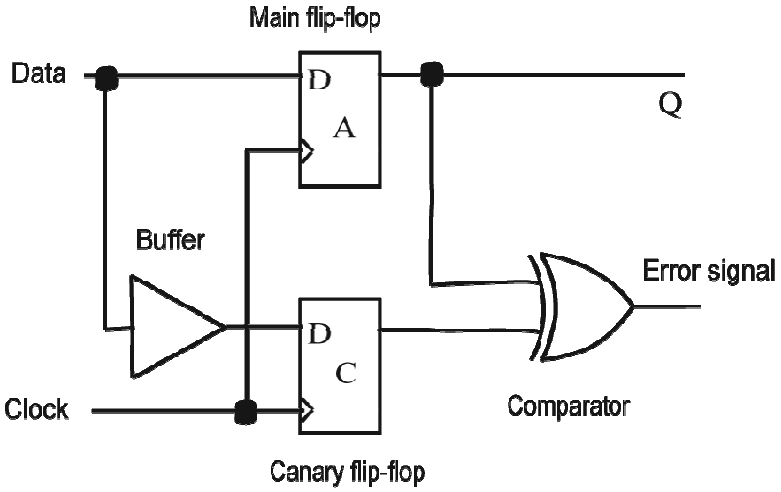



**Figure 3.2: Circuit diagram of the Canary flip-flop.**

## 3.2 A 45 nm delay sensor

We propose a 45 nm delay sensor [19] to predict the timing failures for the combinational logic due to large variations and degradation. The proposed design avoids the need of an error recovery circuit, simplifies clock design, and incurs a very low area/power overhead compared to the previously proposed delay sensor designs. This section provides design, implementation, and simulation results of the proposed 45 nm delay sensor.

### 3.2.1 Proposed 45 nm sensor design

Figure 3.3(a) illustrates a circuit level implementation of the proposed sensor. It consists of a main master-slave flip-flop augmented with an image master-slave flip-flop to pre-sample data. Any difference between the values stored by the two flip-flops is indicated as an error signal. Instead of putting an additional buffer at the input of the image flip-flop, we use the output of the master latch in the main flip-flop as a delayed input to the image flip-flop. Latch A now acts as master to latch B as well as a delay buffer to latch C. This minimizes the performance overhead, an increase in the critical path delay, and makes it more robust to invalidation due to hazards as compared to the Canary flip-flop. Moreover removing the additional delay buffer decreases the area and power overhead of the sensor.

Both latch A and C become transparent on falling edge of the clock signal and a delayed data is latched by latch C, while latch B and D are opaque during this interval and do not pass this data at the output. Since latch A always contributes some positive delay to the input to latch C, a positive guard band in capturing data at latch A and C is ensured even when there are high process and temperature variations. Because of this resilience to variability we can detect variations and degradation before the actual errors start to occur. Any signal transition in this guard band is captured by latch A provided its setup timing requirements are met. However latch C can't capture any transition in the guard band as its setup timing requirement is violated and an error signal is flagged. The comparator is pre-charged when the clock signal is high to avoid generation of a spurious error signal, and it is discharged if the outputs of latches B and D mismatch during the negative clock signal. We have used an error generation circuit [6] that does a logical OR-operation of all the error signals '*i*' generated by each sensor for each pipelined stage as shown in Figure 3.3(b). It sets the 'Error out' signal high when the clock signal is low and at least one error signal '*i*' is high during that period. Since there are

some transients at the start of clock cycle, therefore, it latches data during the negative clock cycle when these transients have settled.

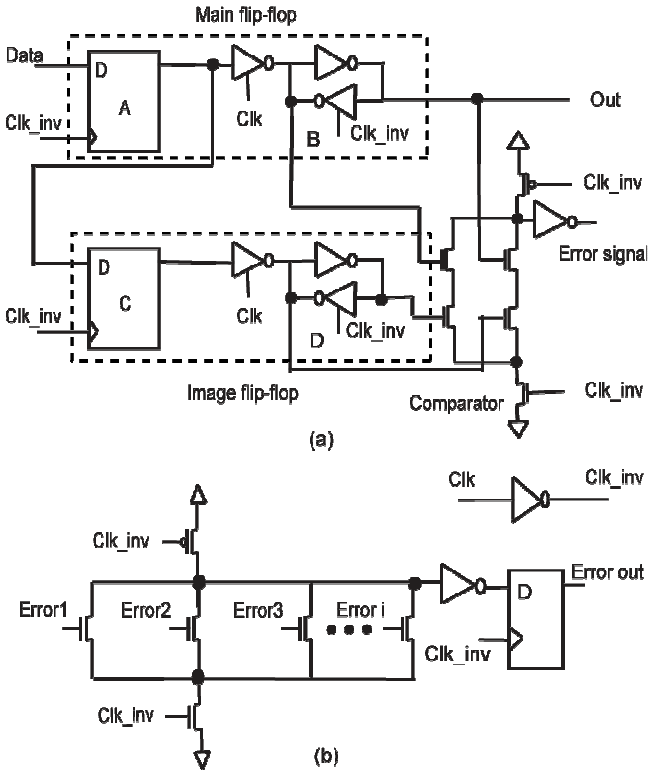



**Figure 3.3: Circuit level implementation of the (a) sensor (b) error generation circuit.**

Figure 3.4 shows the timing diagram of the sensor operation. The image flip-flop receives the Data_delayed signal which is a delayed version of the Data signal. When Clock=0 during the first clock cycle, latches A and C are transparent and pass correct data at their input as their set up time requirement is met. When Clock=1 during the second clock cycle, latches A and C store correct data and output this to latches B and D respectively, which are now transparent. The comparator is pre-charged to allow any transients at the output latches to settle. When Clock=0 during the second clock cycle, latch A becomes transparent again and passes the correct data. However, latch C can't capture correct data as the delayed data misses its set up time requirement. The error signal remains low since both latch B and D store correct data. When Clock=1 in the third clock cycle, latch B receives correct data from latch A and latch D receives wrong data from latch C. An error signal is flagged when Clock=0 in the third clock cycle to indicate a pre-detected timing error. This method avoids data transitions at the main flip-flop when the clock signal makes a transition. This minimizes the chance of meta-stability. Moreover the existence of a short path doesn't invalidate data in the shadow flip-flop which avoids the short path delay constraint.

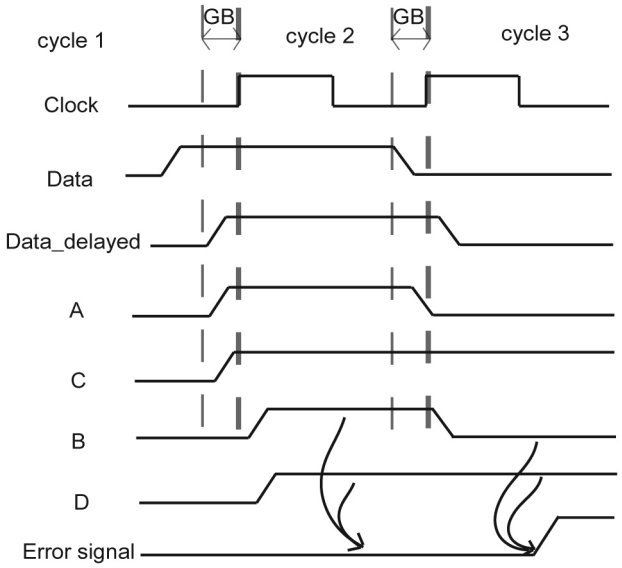



**Figure 3.4: Timing diagram of sensor operation in multiple clock cycles. Data signal makes transition in the guard band in the second clock cycle, and consequently different values are stored in both flip-flops. An error signal is flagged in the third clock cycle.**

Figure 3.5 illustrates application of the proposed sensor at different stages of a pipelined system. At each stage, the sensor can be carefully applied at the most critical paths only to avoid extra power and area overhead. The error outputs of all the sensors at each stage are compressed using the error generation circuitry and a single error signal is outputted. The output signals at each stage can be further compressed to generate a single error out signal for the whole system. Since the error signal represents an early timing failure, the system can avoid the actual timing errors using different compensation techniques [48, 49].

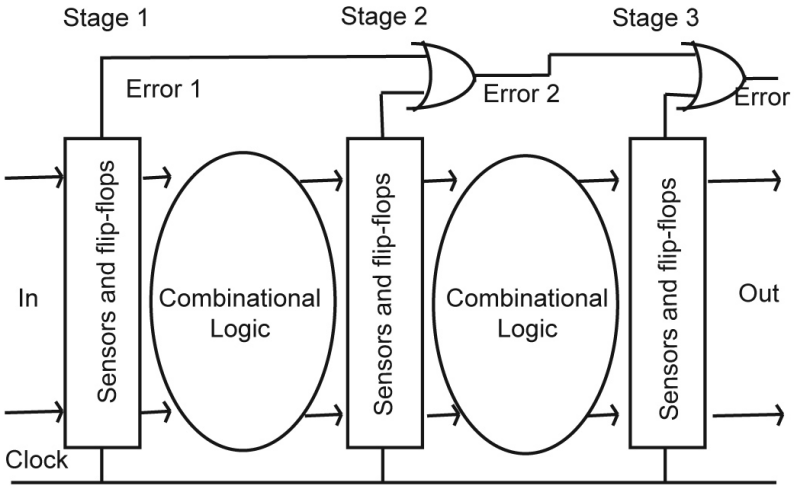



**Figure 3.5: Application of the proposed sensor in a pipelined system.**

The proposed design removes some serious drawbacks associated with the Razor and Canary flip-flops. The main flip-flop in the Razor design can suffer meta-stability due to simultaneous switching of the data and clock signals. That can flag a timing error and trigger a complex error recovery mechanism. In the proposed design the main flip-flop does not suffer from meta-stability since the data transitions occur before the clock signal changes. However it is possible that the input signal at the image flip-flop and the clock signal change simultaneously, resulting in a meta-stability of the image flip-flop. This event may result in a flagging of the Error out signal. Since the data stored in the main flip-flop is not invalidated, an error recovery circuit is not required. In case of the Razor flip-flop, the short path delay constraint requires that the minimum path delay should be larger than the clock delay '$t_{delay}$' plus the hold time of the shadow latch '$t_{hold}$'. Since the path delays can be of any minimum value for the proposed design without invalidating the outputs, our design doesn't suffer from short path problem, i.e. a single phase clock eliminates the short path constraint [51].

The delay buffer at the data input makes the Canary flip-flop prone to invalidation due to hazards as in case of the path delay testing [8]. Delay testing is invalidated when a fast switching path pre-empts another path that shares a common segment with it [88]. Therefore a fast switching branch can increase the delay at the other branch of the Canary flip-flop thus can result in a timing error. Cross talk between the two input segments of the Canary flip-flop can alter the max-path delay that can also result in timing failures. Since both the main flip-flop and the image flip-flop share a single critical path, the proposed design is more robust to hazards and cross talk. Moreover the delay buffer incurs significant area and power overhead, especially when a longer guard band is required to cope with input/process variations. A large number of critical paths might require *in-situ* monitoring due to high variability in future technologies. The proposed sensor avoids the additional delay buffer used in the Canary flip-flop and therefore incurs lower area/power overhead.

### 3.2.2 Soft error correction

We have also added a soft error correction circuit [9, 46] to extend the sensor's capability to provide robustness against soft errors as shown in Figure 3.6. The proposed design provides a protection against soft errors in the combinational circuit (SET) and in memory elements or flip-flop (SEU). Using the fact that the soft errors in the combinational logic appear as glitches [46], we sample data using the main flip-flop and its delayed version using image

flip-flop. Both flip-flops store different values in case of a soft error in the combinational logic. The soft error correction element latches data from the main flip-flop as it keeps correct output. In case of SEU, we assumed that all five latches (A, B, C, D, and the output latch) can flip their state due to a soft error. When the Clock signal is high, latches B and D pass correct data (assuming latches A and C are not affected by a soft error during this interval). The chances of latches B and D getting a soft error are very low since they are driven by master latches. Similarly the output latch is less susceptible to a soft error when it is driven by latch B. When the clock signal becomes low, then latch A and C becomes transparent. They have very little chance of having a soft error as they are driven by the input signal. However latch B, D and the output latch can be affected by soft errors. A soft error may occur when a particle strike flips the state of latch B or latch D. Since their outputs mismatch, the soft error correction element will not propagate any data at the output and the output latch will hold the correct data in its feedback loop. Similarly if the particle strike flips the state of the output latch, it will recover correct state since it is fed correct data by latch B and D.

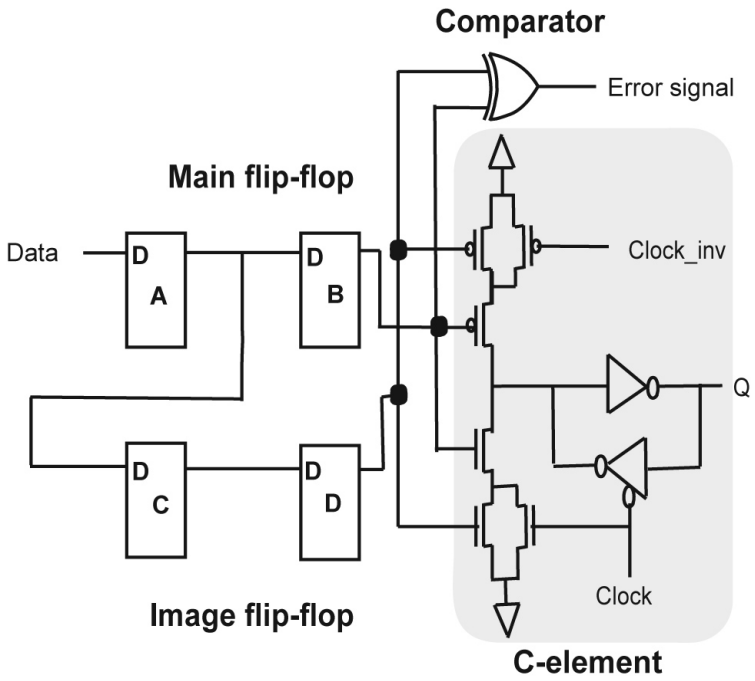



**Figure 3.6: Sensor design with soft error correction.**

### 3.2.3 Simulation results

In order to assess the usefulness of the proposed 45 nm delay sensor for a robust low power circuit operation in nano-CMOS technologies, we designed a 32 bit Carry Select Adder (CSA)

and a 16x16 Carry Save Multiplier (CSM) as combinational logic test benches for the circuit level simulations. Table 3.1 shows the specifications of both test circuits for our simulations. Two different design methods were used to add a pessimistic design margin in terms of delay and power. Static and dynamic variability was then injected separately into different designs to extract possible energy reductions in the case of different variations. We have used the 32 nm Predictive Technology Models (PTM) from Arizona State University [24] and 45 nm High Performance BSIM4 model cards from the University of Glasgow [22, 89] to carry out our simulations. Figure 3.7 illustrates our method for gate level simulation of temperature and statistical variability. Gate level description of the test bench circuits was given in the form of a HSPICE netlist. C/MATLAB scripts were used to process this list to insert statistical and temperature variations. C-scripts were then used to insert random input vectors in HSPICE netlists for circuit simulation at different clock cycles. Finally MATLAB was used for processing HSPICE generated data.

**Table 3.1: Specification of the test circuits.**

| Test circuit | Number of inputs | Number of outputs | Outputs with sensors | Outputs with flip-flop | Number of transistors (with flip-flops only) | Number of transistors (with proposed sensors) | Number of transistors (with Canary flip-flops) |
|---|---|---|---|---|---|---|---|
| 32-bit CSA | 64 | 33 | 13 | 20 | 2760 | 3203 | 3255 |
| 16x16 CSM | 32 | 32 | 16 | 16 | 8896 | 9438 | 9502 |

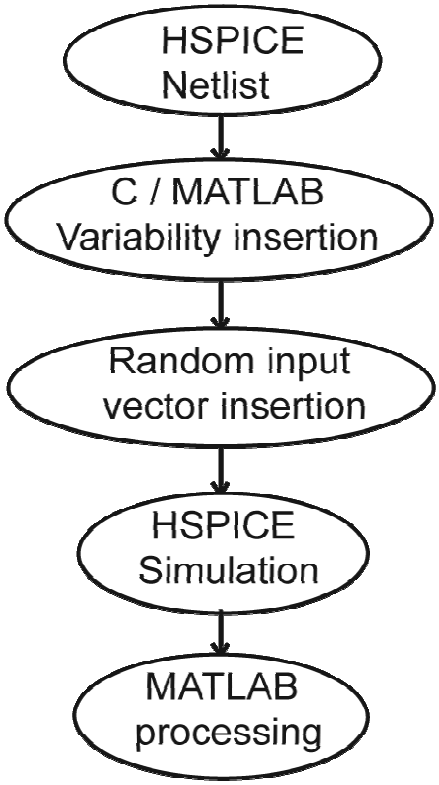



**Figure 3.7: Design flow for gate level simulation of temperature and statistical variability.**

### 3.2.3.1 Temperature Variations

The 32-bit CSA was designed using 32 nm PTM models to evaluate sensor's robustness and to quantify energy savings in the case of temperature variations. We applied 12,000 random input vectors to the CSA to identify the critical paths and the maximum circuit delay. In order to minimize the power overhead, the sensor was applied only on 40% of all the paths that had maximum delay. The critical voltage was found to be 1.02V by running circuit level simulations at 90% of the base line clock period at a temperature of $85^{o}$C to account for process and temperature variations [7]. The same set of random input vectors was applied to the CSA when equipped with and without the sensor at different voltage and temperature conditions. Figure 3.8 illustrates the average power per cycle consumed by the CSA based on the worst case design and the sensor based design. For clarity only those points are plotted for the sensor based design where the circuit operated at a minimum voltage without any pre-detected or actual error. It was found that sensor based design can reduce the power consumption by 1/1.5 ($146\,\mu W$ vs. $97\,\mu W$) as compared to the worst case design. The device delay degrades with the temperature rise and the circuit delay increases. This increases the pre-detected error rate as more outputs fall inside the guard band. The average power for the

worst case design remains nearly the same as it always operates on the same voltage selected at the design time. However for the sensor based design voltage is increased or decreased depending upon the detection of pre-detected errors for a defined number of cycles. The amount of energy saving reduces with increasing temperature since the sensor detects more errors.

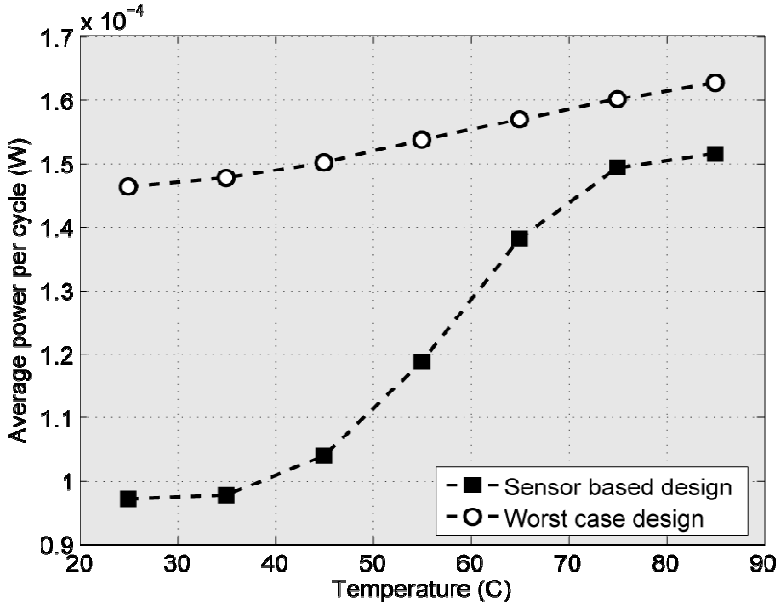



**Figure 3.8: Power consumption for two different design methods.**

Approximately 60,000 HSPICE simulations were carried out that applied random input vectors to the CSA at different voltage and temperature conditions in order to assess the sensor's robustness against temperature variations. The results indicate that the sensor is able to pre-detect timing errors at each temperature condition before they actually happen as illustrated in Figure 3.9. The sensor detects more pre-detected and actual errors as the supply voltage is dropped or the temperature is increased. The circuit delay increases and more output transitions either fall inside the guard band or miss their setup time requirements resulting in pre-detected or actual error respectively. The total number of actual errors exceeds pre-detected errors at very low voltages since the outputs of all the sensors are OR-ed to generate single 'Error out' signal. This provides a method to drop the supply voltage to a very low level that is well below its critical value that ensures correct operation. Moreover there is no need for an error recovery circuit since errors are detected before they actually occur.

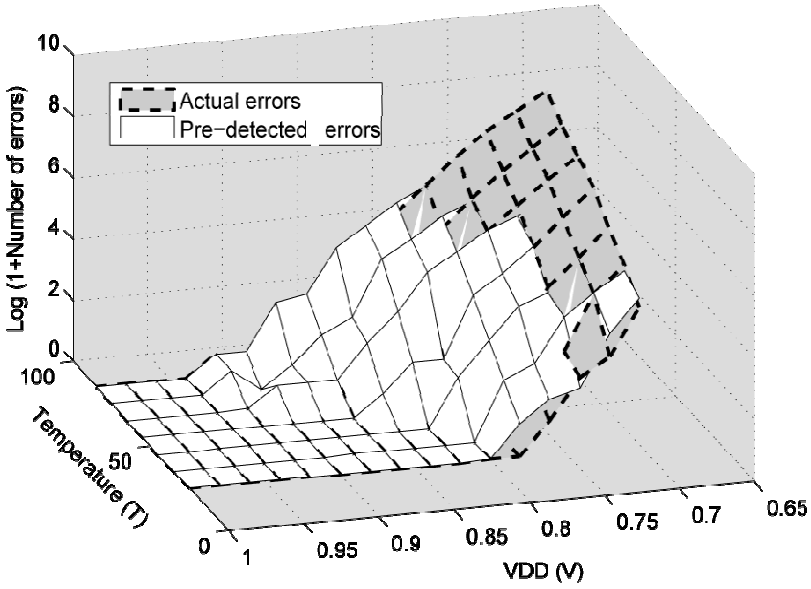



**Figure 3.9: Error plot at different temperature and voltage conditions.**

Figure 3.10 shows the average power consumed per cycle at different voltage and temperature conditions. Since the circuit delay increases with a rise in temperature or a decrease in the supply voltage, the sensor detects more errors in these conditions. The system then responds by increasing the supply voltage to make the circuit run faster and avoid any timing failure. However choosing high supply voltage costs higher power consumption as shown in Figure 3.10. This reduces the power reduction margin at a high temperature. This margin would further decrease in the presence of static variations. However the combination of all worst case conditions occurs very rarely.

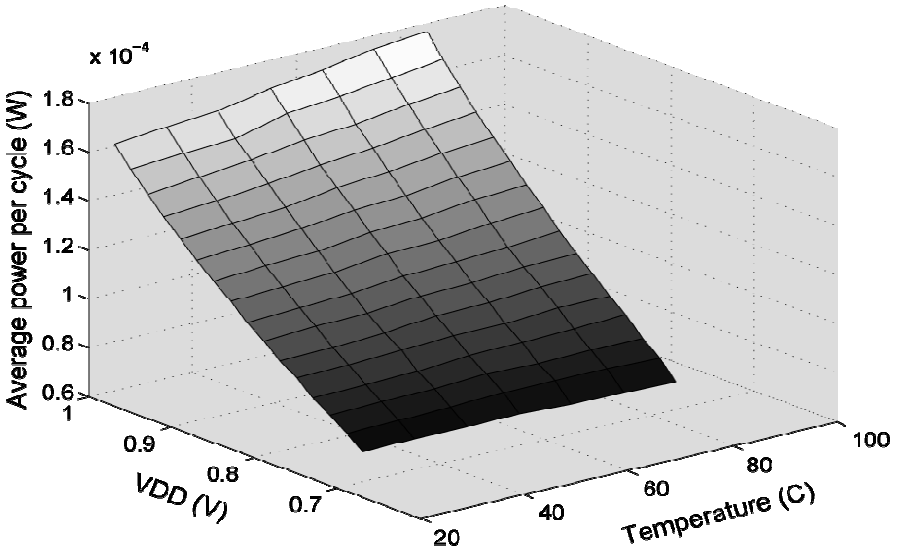



**Figure 3.10: Power consumption at different voltage and temperature conditions.**

Figure 3.11 describes the relationship between the error rate and average power consumption with decreasing the supply voltage at different temperature conditions. A decrease in the supply voltage results in higher energy reductions at the cost of higher error rate. An increase in the temperature results in a large increase in the error rate. However, the average power consumption increases very little. This shows that a supply voltage can be selected either to maintain a given error rate or to meet certain power consumption requirements.

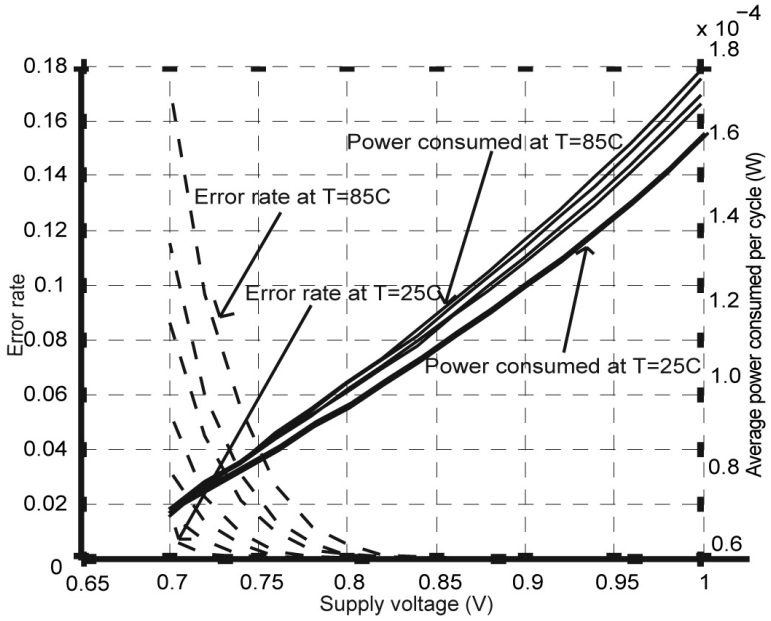



**Figure 3.11: Relation between error rate and power consumption at different temperature conditions.**

### 3.2.3.2 Statistical Variations

A 16x16 Carry Save Multiplier (CSM) was designed using 45 nm BSIM4 models from the University of Glasgow. These models are based on 35 nm gate length devices [89]. Approximately 200 models were extracted based on the variability simulations of different sources of statistical variability including Random Discrete Dopant (RDD), Poly-Si Gate Granularity (PGG), and Line Edge Roughness (LER). A combination of C and MATLAB scripts were used to insert devices randomly from the ensemble of 200 models. Ten randomized versions of CSM (with sensor) were generated to observe robustness of the proposed sensor in case of statistical variability.

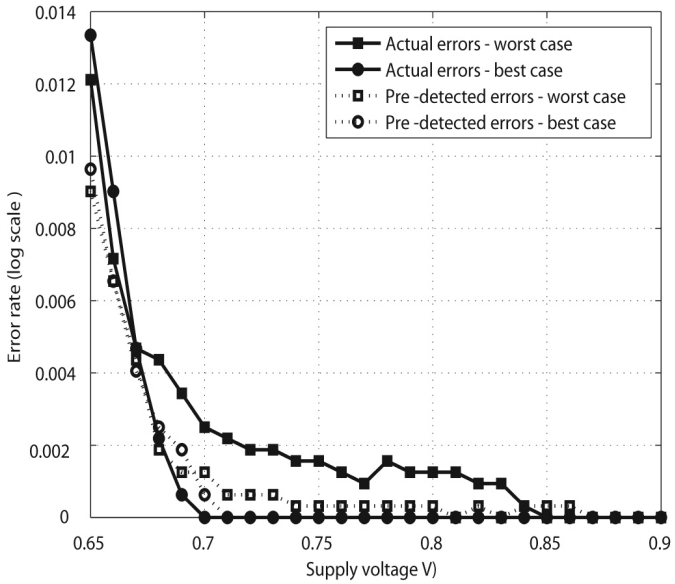



Figure 3.12: Error rate comparison for two extreme instances of the CSM.

Over 12,000 random input vectors were applied to the CSM to find out the maximum path delay and the critical paths in the circuit. Instead of increasing the supply voltage, we added a 20% delay guard band (GB) to the clocking frequency to account for process and temperature variations. The sensor was applied on 50% of all the critical paths to pre-detect timing failures. Since a large computational time is required for the gate level simulation of our design, therefore we selected 100 worst case vectors from the random input vectors. The impact of statistical variability on the design is evident in terms of two extreme cases of the same circuit as illustrated in Figure 3.12. About 5,000 HSPICE simulations were performed that applied worst case vectors to both worst case CSMs at different supply voltages. The sensor was able to detect timing failure at 0.7 V for the best case CSM providing the maximum power reduction. The power reduction is at minimum for the worst case instance of the CSM when the sensor starts pre-detecting timing failure at 0.86 V. The sensor was able to pre-detect timing errors before the actual errors occurred for both the extreme case designs.

Figure 3.13 illustrates the error rate plot for 10 randomized instances of the CSM. We applied the same 100 worst case vectors to all of these CSM circuits at different supply voltages to give a fair comparison. For simplicity, the results are sorted by the error rate to give a clear picture of the outcomes for each circuit. The sensor was able to detect timing failure in all the cases, since the shaded surfaces (representing pre-detected errors) occur

before the transparent surfaces (actual errors). We can see two extreme cases where the statistical variability causes a large difference in the CSM delays. However the error rate remains comparable for most of the circuits. Therefore, using a high voltage or adding a large timing guard band indiscriminately for all the circuits will result in a waste of the useful energy. In contrast the sensor based design selects an appropriate supply voltage or frequency depending on the actual variability or device wear out with time.

To test the efficiency of proposed sensor in the dynamic power regulation, we consider the case where each of these CSM is equipped with a voltage regulator which increases or decreases the supply voltage by observing the error rate for a defined number of clock cycles. The system starts of with pessimistic guard band of 0.2T (T is the clock period) and the voltage regulator decreases the supply voltage until a pre-detected error occurs. The simulation results show that the proposed 45 nm delay sensor reduces the power consumption by 1/1.4 ($328\,\mu W$ vs. $239\,\mu W$) on average as compared to the worst case design. The maximum of 1.7X ($327\,\mu W$ vs.$198\,\mu W$) and a minimum of 1.1X ($328\,\mu W$ vs.$305\,\mu W$) improvements in the average power of the CSM is observed as compared to the worst case design. For high performance applications, this reduction in power can be exploited to boost clocking speed.

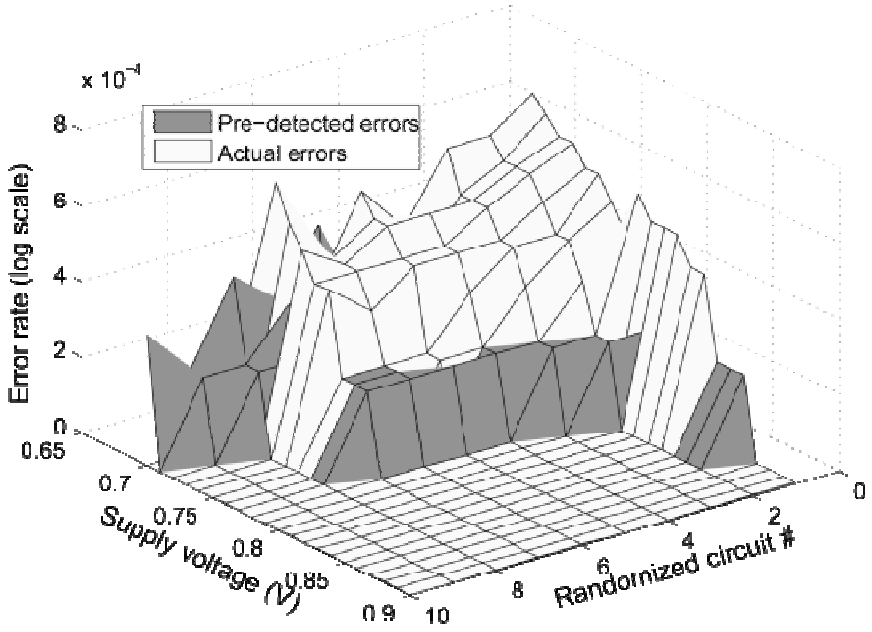



**Figure 3.13: Error rate plot with decreasing supply voltage for randomized CMS circuits.**

Figure 3.14 shows the relationship between error rate and the average power consumption for the worst case instance of the CSM. As the supply voltage is dropped, the average power consumed by the CSM decreases proportionally. However a decrease in the supply voltage increases the circuit delay which results in more output signal transitions in the guard band of the sensor. These are flagged by the sensor as pre-detected errors and the error rate rises exponentially with the decreasing supply voltage. A small decrease in the supply voltage results in a large increase in the error rate, therefore, the voltage supply step size should be kept small to maintain an acceptable error rate. An optimal supply voltage can be selected for each CSM that meets given power requirements or maintains an acceptable error rate.

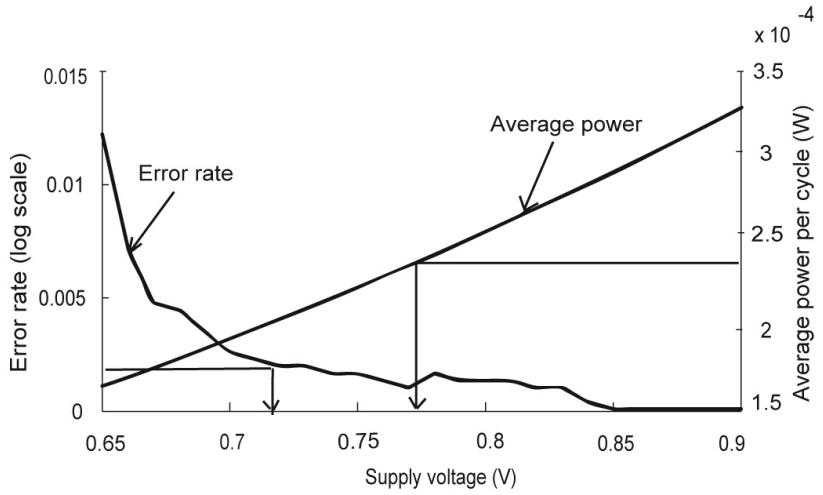



**Figure 3.14: Relation between power reduction and error rate for CSM (worst case).**

An important observation from these simulations is that a guard band of 20% might not be enough to account for both the process and temperature variations. It should be noted that an additional guard band is needed for the degradation as devices wear out with time and become slower. For illustration, we designed a 30 stage inverter chain and the sensor was applied at the output. MATLAB scripts were used to mimic a voltage regulator that would increase or decrease the supply voltage depending on when any pre-detected error is detected, or not, on each clock cycle. Again MATLAB and C scripts were used to create 200 randomized versions of the inverter chain circuit. A guard band of 10%, 20%, 30% and 40% was added to all of these circuits and HSPICE simulations were carried out for 50 clock cycles. The results of these simulations for GB=0.1T and GB=0.4T (where T represent typical case clock period without guard band) are shown in Figure 3.15.

The voltage regulator starts with an initial supply voltage of 0.9V and then decrements this voltage in next clock cycle to reduce design margins provided there is no pre-detected error. When a timing failure is detected then it raises the supply voltage in the next clock cycle, therefore it follows a zig-zag path after few clock cycles. As the length of guard band is increased, the clock speed gets slower. However, more outputs can now meet their timing requirements even under high statistical variations. The voltage regulator had to increase supply voltage to avoid timing failure when the guard band was 0.1T. As the length of guard band increased, more outputs met their timing requirements and the voltage regulator could drop supply voltage to save power. The greatest power reduction is achieved with a guard band of 0.4T when most of the circuits could operate at a very low supply voltage. Figure 3.16 demonstrates the impact of guard band on the average supply voltage of different randomized instances of a 30 stage inverter chain circuit. The average supply voltage decreases with the increase in the guard band. This provides higher power reductions at the cost of a low operating frequency.

These simulations indicate that a guard band of length 0.3T-0.4T (30% - 40%) would be needed to avoid timing failure for these circuits. It can be concluded that future technology nodes will require even large guard bands to avoid functional and timing failures. Moreover, the device degradation is expected to get worse beyond 32 nm technology node [5]. Therefore, a sufficiently large guard band will be needed that will have a significant impact on the performance. This necessitates the use of sensor based designs to avoid large pessimistic guard bands. Another way to use this sensor is to increase the operating frequency to reduce performance overhead for a large guard band, at the cost of lower energy savings. There is a trade off between the maximum clock frequency and power reductions achieved for both cases. However the sensor based design provides an intelligent utilization of the useful performance and power resources.

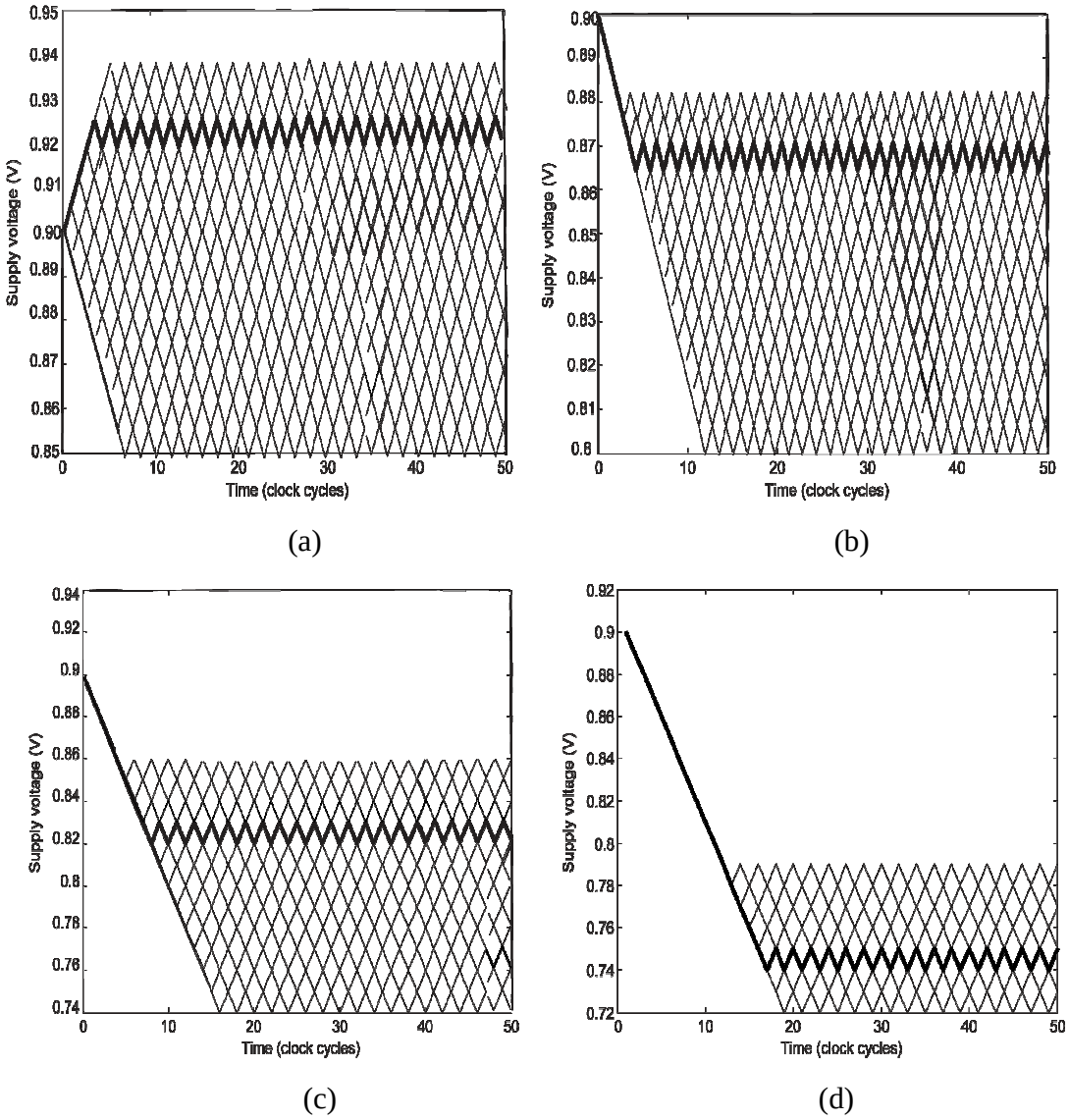



**Figure 3.15: Relation between guard band and the selected supply voltage (a) GB=0.1T Average power per cycle = 19.6uW. (b) Guard band= 0.2T, Average power per cycle = 15.7uW (c) Guard band= 0.3TAverage power per cycle = 13.3uW (d) GB=0.4T. Average power per cycle = 11.5uW.**

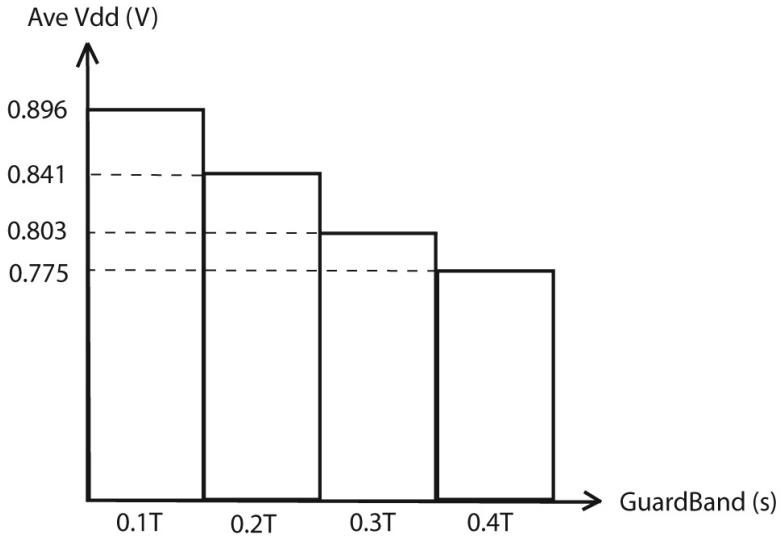



**Figure 3.16: Impact of guard band on the average supply voltage of a 30 stage inverter chain.**

### 3.2.4 Area and power comparison

We simulated both test circuits (CSA and CSM) using a Canary flip-flop and the proposed sensor for a comparative analysis. For a fair performance comparison, a set of 2000 worst case vectors were applied to each of the test circuits. We found that the proposed 45 nm delay sensor had a lower performance overhead (0.9%) as compared to the Canary flip-flop (1.4%). For area and power comparison, we appended the Canary flip-flop and the proposed sensor on 40% and 50% of the critical paths of the CSA and CSM, respectively. Approximately 1500 worst case vectors were applied to both designs for different lengths of the guard bands, the results of these simulations are shown in Figure 3.17. The Canary flip-flop based design incurs a significant power overhead as compared to the proposed sensor design for both CSA (12% vs. 7%) and CSM (11% vs. 3%) for a guard band, GB=35ps. Similar improvements in the area overhead are observed for the proposed design in both test cases, CSA (22% vs. 16%) and CSM (8% vs. 6%). It is expected that the number of critical paths will increase in future generations due to a rise in process variations. Therefore more critical paths will require *in-situ* monitoring. Moreover since the predictive *in-situ* monitoring doesn't have an error recovery mechanism, a longer guard band would be essential to minimize the chances of timing failure due to input variations. The proposed sensor design presents a more area and power efficient alternative to the Canary flip-flop.

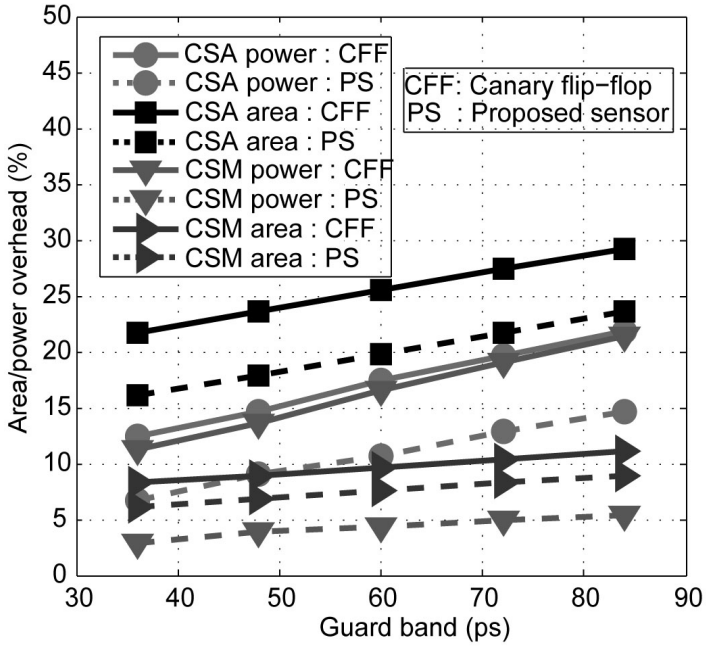



**Figure 3.17: Area and power overhead vs. guard band for both test circuits.**

## 3.3 A 32 nm delay sensor

The previously proposed 45 nm delay sensor provides an extension to the Canary flip-flop design by using the master latch of the main flip-flop as a delay buffer to detect the timing failures. We now present a 32 nm delay sensor [18] that uses an advance clock signal to pre-detect the timing failures. This design works similar to the Razor flip-flop; however, the errors are predicted as opposed to detected in the Razor flip-flop. This section provides design, implementation, and simulation results of the proposed 32 nm delay sensor.

### 3.3.1 Proposed 32 nm sensor design

The structure of the proposed 32 nm delay sensor is very similar to the Razor flip-flop however the difference lies in sampling of the data for both designs. The Razor flip-flop performs post-sampling of the data by using a delayed clock for the shadow latch. It allows the data to violate timing margins of the main flip-flop in order to achieve higher energy savings using the data dependency. However it requires an error recovery circuit that incurs a performance penalty and increases the complexity of the design. The proposed 32 nm delay sensor uses an advanced clock signal to sample data in the shadow latch first and then with a delayed signal (original clock) in the main flip-flop, this method is called pre-sampling. Errors

are predicted in advance before an actual timing error does occur. Therefore it avoids the requirement for an error detection and recovery mechanism. However it doesn't utilize higher data dependency and therefore has lower energy reductions compared to the Razor flip-flop.

Figure 3.18 illustrates the timing operation of the pre-sampling (proposed) and post-sampling (Razor) methods. The gap between the Clock signal and the D_clock forms a guard band to detect timing failures for the proposed design. Any errors after the rising edge of the Clock signal are detected timing failures for the post-sampling design. There is no timing violation during the first clock cycle for both designs; therefore the error signals are low. The Data signal makes a transition inside the guard band for the proposed design in the second clock cycle and an error signal is flagged. The error signal for the post-sampling (Razor) remains low as it doesn't predict timing failures. The Data signal violates setup time in the forth clock cycle, therefore the post sampling method flags a timing failure. However the proposed design can't detect this error because the timing violation (data transition) occurred outside the guard band interval of the proposed design.

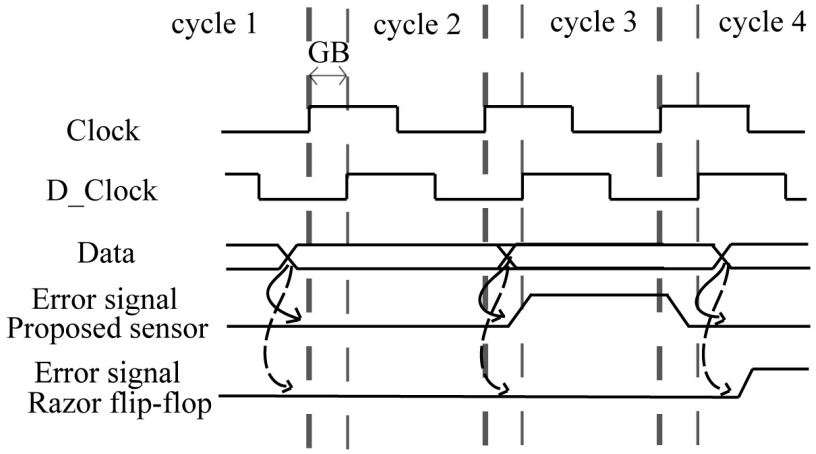
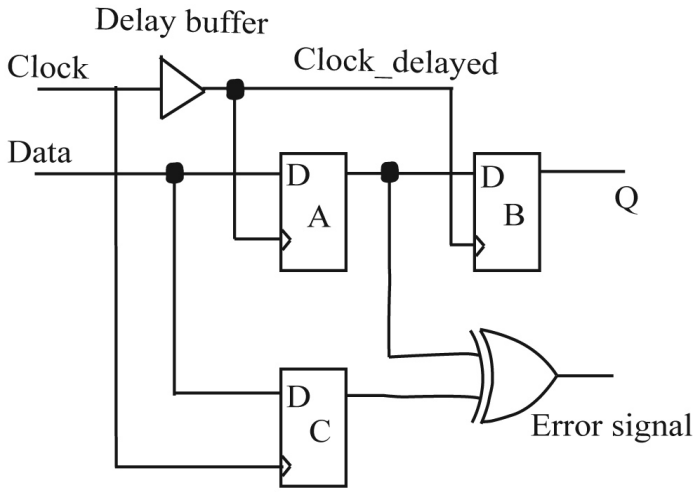



**Figure 3.18: Timing diagram illustrating pre/post-sampling of data.**

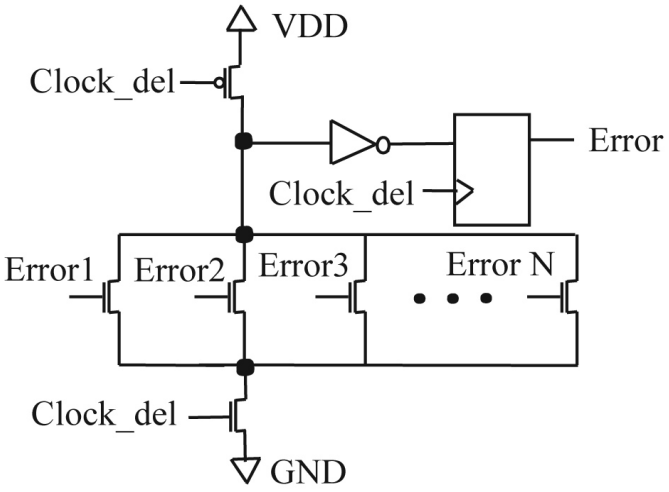
Figure 3.19(a) shows the detailed circuit level implementation of the proposed 32 nm delay sensor. The proposed sensor consists of a conventional master-slave flip-flop augmented with a shadow latch that operates at an advanced clock signal to detect timing failures in advance. An error signal is flagged by the comparator (Exclusive OR) to predict a timing failure when the latches A and C hold different values. Latch A stores data when the delayed clock signal is high and becomes transparent when the delayed clock signal is low. Latch C is transparent

when the clock signal is low and becomes opaque otherwise. The delay between the clock and its delayed version creates a guard band to detect the timing failures. Any signal transition in this interval is detected and flagged as the timing failure due variation, degradation or a too small voltage required in case of dynamic voltage scaled (DVS) processors. The propagation delay of the clock buffer ensures a positive guard band even when the process variations are high that makes it very robust to variability [51]. The shadow latch experiences a timing failure earlier than the main flip-flop and therefore it can detect timing errors in advance. The delay buffer doesn't add any overhead to the clock speed and it serves only to delay the clock signal.

Figure 3.19(b) shows the error generation circuit [6] to perform a logical OR of the error outputs of the individual delay sensors. It is pre-charged to output a zero error when the delayed clock signal is low. An error signal is generated when any sensor flags an error signal 'N' while the delayed clock signal is high as well. Using the delayed clock signal avoids generation of a false error signal that may occur when the latches A and C output different values before the falling edge of the delayed clock signal. It may happen when the latch C, operating with the clock signal (advanced), stores new data while the latch A, operating with a delayed clock signal, still holds an old data. The difference in stored values can raise an error flag by the comparator. Clocking the error generation circuit avoids any unwanted spurious error signal (occurring outside the guard band of a delay sensor) to discharge the circuit and signal false errors.




(a) Sensor circuit




(b) Error generation circuit

**Figure 3.19: Circuit implementation of (a) sensor (b) error generation circuit.**

Figure 3.20 shows the timing operation of the proposed delay sensor for different clock cycles. The data signal makes transition before the start of the guard band, and therefore meets timing requirements of the both latches A and C. The delayed clock signal raises high at the start of second clock cycle. The latch B becomes transparent and passes the data stored by latch A to the output. The error signal remains low since both latches store the same values. The data signal makes a transition inside the guard band in the third cycle. The signal transition satisfies the setup timing requirements of the latch A and a correct data is stored. However latch C misses its set up timing requirements and a wrong data is saved. The comparator generates an error signal in the third clock cycle indicating a timing failure. Compensation methods like voltage scaling or body biasing can then be used to avoid the actual timing failures. Since this method keeps the data transitions to occur before the start of the guard band therefore the chances of the simultaneous data and clock signal transitions are minimized, i.e. it is more robust to meta-stability. In addition any short path doesn't invalidate data in the shadow latch as the main flip-flop stores the correct data; therefore it avoids the short path constraint as well.

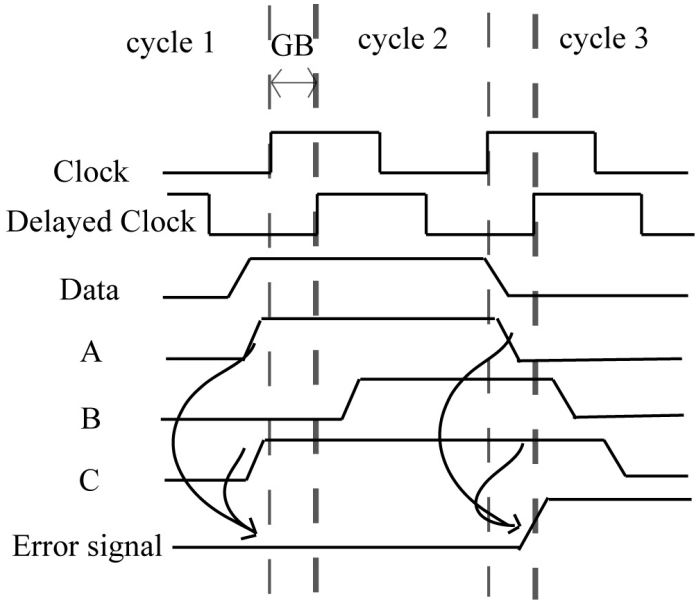



**Figure 3.20: Timing diagram of sensor operation.**

Figure 3.21 shows an application of the proposed delay sensor in different stages of a pipelined processor. Each stage uses a set of delay sensors embedded at the critical paths only to avoid a higher area and power overhead. The flip-flops and the sensors at each stage run on the delayed clock signal while the shadow latch operates at normal clock signal (advanced in this case). The error generator circuit is used to do logical OR of errors at each stage that generates an error signal if any path experiences a delay failure. The error signals at each stage are further OR-ed to generate a global error signal. Different compensation schemes can then be applied to avoid actual timing failures at any of the stages of the pipelined system. For a fine grained system, each stage can have an independent control mechanism that can adjust the supply voltage, frequency or the body bias to avoid actual delay failures of that stage only. This allows maximizing the energy reductions as each stage operates according to its variability. Therefore no worst case matrices (voltage or frequency) are chosen to operate all stages that degrade possible improvements of the *in-situ* design. However this would complicate the design as each stage requires a separate compensation mechanism.

The delayed clock signal can be locally generated for each stage to avoid having an extra clock tree. However this method is more prone to process and environmental variations as each delay buffer (delay element) may give different delay to the original clock signal that can lead to synchronization problems. A single clock delay buffer can avoid an extra overhead of local generation of the delayed clock signal and is more robust to variations. However it

increases the clock tree capacitances and may complicate the clock tree design.

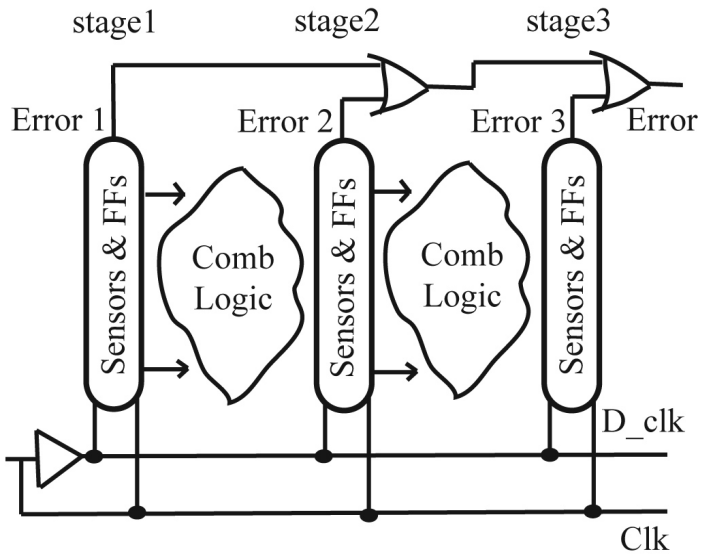



**Figure 3.21: Application of the sensor in different pipeline stages.**

## 3.3.2 Simulation results

We have used the 32 nm PTM models and 45 nm BSIM4 model cards to evaluate effectiveness of the proposed delay sensor under temperature and statistical variations, respectively. This section discusses the results of our simulations for temperature and statistical variations.

### 3.3.2.1 Statistical variations

To illustrate the application of the proposed delay sensor for an *in-situ* design, we designed a simple 30 stages inverter chain circuit as test bed and embedded the proposed sensor at the output. We used 45 nm BSIM4 model cards with the statistical sources of variability to verify functionality of the proposed 32 nm delay sensor for this simple test circuit. A voltage regulator was used to increment or decrement the supply voltage after each cycle depending on if a timing failure occurs or not. The voltage regulator is simulated to decrease the supply voltage until a timing failure is detected. It then raises the supply voltage to avoid any timing failure in future. Since a single critical path exists and circuit delay is dictated by this path, therefore the supply voltage follows a zig-zag pattern after initial cycles that calibrate the circuit according to variability. Figure 3.22 shows the result of statistical variability simulations of the inverter chain for 4 different instances. The voltage-time graphs are plotted

for 50 cycle cycles only for simplicity. The initial supply voltage for the test circuit is set high (1V) to avoid any timing errors (actual). The voltage regulator then keeps decrementing the supply voltage after each cycle as long as no error signal (predicted) is flagged. Once a error signal is flagged by the delay sensor, the regulator starts to increment the supply voltage until the timing error (predicted) is avoided. We observe that a minimum supply voltage can be selected for each instance that will minimize the timing error and provide the highest energy savings. In contrast a conventional design selects the worst case supply voltage (1V) for all instances of the circuit that does incur high energy overhead. Our simulations indicate that the *in-situ* design using timing sensors (proposed) can provide high energy savings compared to a conventional worst case design.

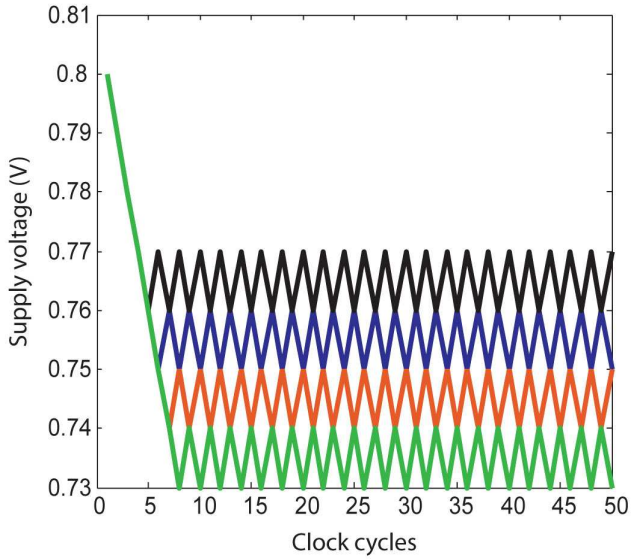



**Figure 3.22: Inverter chain simulation under statistical variability.**

### 3.3.2.2 Temperature variations

A 32 bit Carry Select Adder (CSA) was designed using Predictive Technology Models (PTM) [24] for 32 nm technology node in order to quantify possible power reductions using the proposed sensor design for the combinational logic circuits. C / MATLAB scripts were used to insert random vectors in HSPICE net list, and the simulations were carried out at different temperature and voltage conditions. The Razor flip flop and the Canary flip flop were applied separately to the CSA to have a fair comparison with earlier sensor designs. It was found that the proposed sensor had a much lower performance overhead, less than 0.5% as compared to 1.4% for the Canary logic and 2.2% or the Razor flip-flop. Addition of the

delay buffer causes higher performance overhead in case of the Canary logic. However it is still less than Razor flip flop since it doesn't require an error recovery circuit. For power comparison, we applied the same set of inputs vector to the CSA equipped with different sensors under the same process, temperature, and voltage conditions. The Razor flip-flop had the highest power overhead of 7.3% due to a complex error recovery circuitry. Whereas the proposed 32 nm delay sensor had a 4% power overhead that is lower than the overhead for the Canary flip-flop (6.7%). The delay buffer at the input of the Canary flip flop is non-sharable and therefore incurs a extra power overhead, whereas we use a single clock delay buffer for all the sensors that reduces the power overhead.

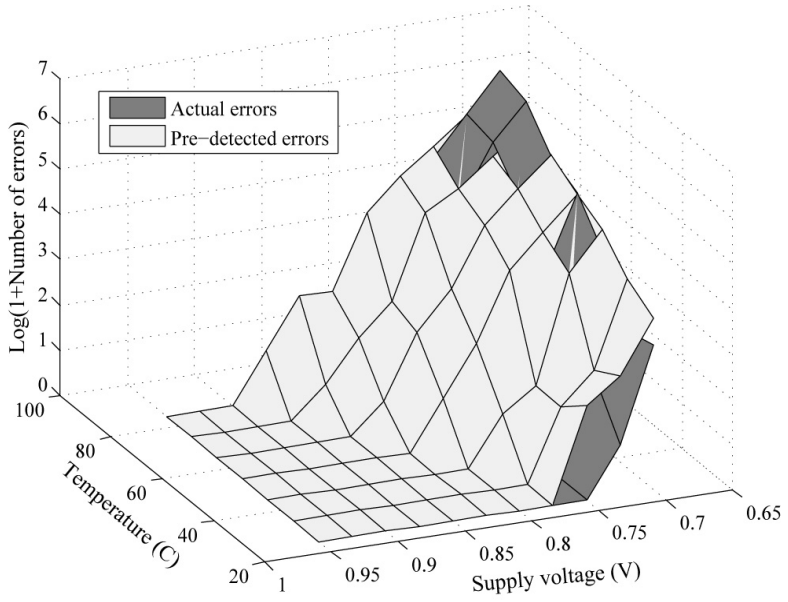



**Figure 3.23: Plot of actual and pre-detected errors.**

To verify the functionality of our sensor circuit we performed 42000 HSPICE simulations that applied random inputs to the CSA at different voltage and temperature conditions. Plot of the actual errors and pre-detected errors at different temperature and supply voltage conditions is shown in Figure 3.23. For each temperature condition the sensor was able to pre-detect timing failure at a higher voltage before the actual errors. The system can then use some compensation schemes like adaptive voltage scaling, adaptive body bias, or adaptive frequency scaling to avoid the actual errors. It is therefore possible to use such a sensor in a DVS processor where the supply voltage can be dropped to a low voltage that ensures the

correct operation without the need for an error correction operation. More errors are detected either at a high temperature or low supply voltage due to a rise in the circuit delay. Because of its resilience to process and temperature variations, the sensor does not fail even at worst case temperature of $85^{o}$ C.

The proposed sensor was applied to only 40% of all the paths which had a maximum delay to minimize the power overhead. The critical voltage was found to be 1.02 volts by running circuit simulations with 90% of the base line clock period at a temperature of $85^{o}$ C to account for process and environmental variations [7]. Figure 3.24 shows a plot of the average power consumed per cycle by the CSA based on the worst case design and with the proposed sensor. For comparison only those points are chosen where the sensor operated at a minimum voltage without any pre-detected error. The device delay increases with temperature and more outputs fall within the guard band. Thus the sensor detects more errors as the temperature is raised. These results in lower power savings as it then needs to be operated at a higher voltage to avoid the actual timing failures, highest power reductions by 1/1.7 ($277\,\mu W$ vs.$162\,\mu W$) are observed at $25^{o}$ C. There is very little increase in the average power with a temperature rise for the CSA worst case design as it always operates at a fixed supply voltage.

Figure 3.25 illustrates the relationship between the power consumption and the error rate at different temperature and voltage conditions. The error rate falls exponentially as the supply voltage increases and the power consumption rises because of its non linear dependence on the supply voltage. A rise in temperature significantly increases the error rate, however, the power consumption increases very little indicating a weak dependence of power on temperature. This plot also shows that a supply voltage can be selected for a sensor based design that either meets power requirements or maintains a given error rate at different temperature conditions.

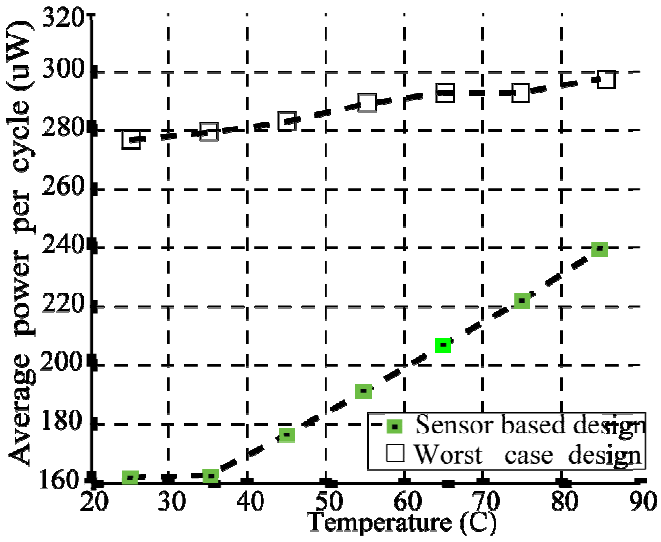



**Figure 3.24: Average power per cycle at different temperatures.**

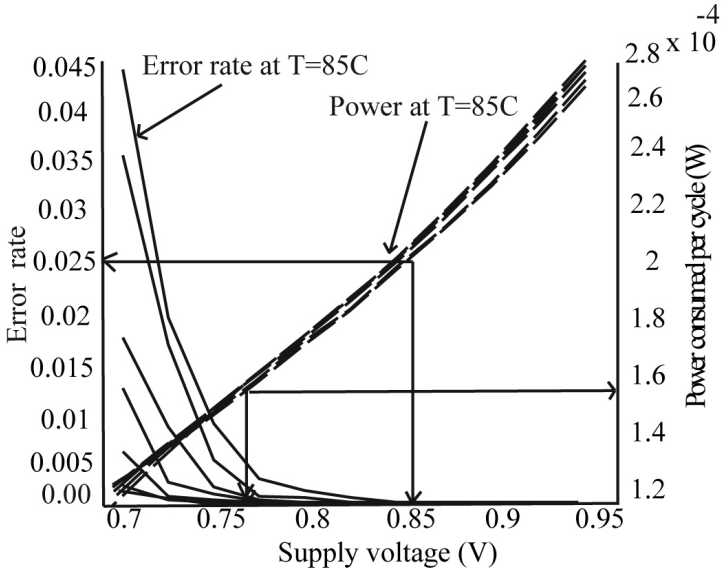



**Figure 3.25: Relation between power consumed and error rate at different temperatures with decreasing supply voltage.**

## 3.4 Chapter summary

Large variations, aggressive degradation, and increasing soft error rate pose serious challenges for a reliable circuit design in future technologies. Conventional design methods add pessimistic voltage or frequency margins to obtain fully functional designs. However these methods incur high power/performance overhead as most of the dies meet desired delay/power specifications. *In-situ* monitoring of the timing failures provides a handle to tune

chip voltage/frequency that corresponds to the on-chip variability, thereby allows an efficient use of power/frequency resources. We have presented two novel delays sensors to detect timing failures in advance. The proposed 45 nm delay sensor uses the delay of a master latch in a conventional master-slave flip-flop to pre-detect timing failures before they cause the actual errors. This provides a handle to either decrease the supply voltage to save energy or increase the clocking frequency for high performance applications while keeping an acceptable error rate. The sensor can avoid large pessimistic design margins for a robust design in future technologies at a minimum performance and power overhead. It has negligible impact on the maximum path delay, while the total performance overhead is also very small (less than 0.9%). The power overhead is about 5.5% when the sensor is applied on 50% of all the critical paths. However, this can be minimized by applying the sensor on fewer but more probable critical paths. The proposed sensor can be extended to provide soft error correction simultaneously at the cost of a small power overhead. HSPICE simulations carried out on a 32bit CSA with temperature variations using 32 nm PTM indicate that the proposed design can reduces the average power consumption by 1/1.5 as compared to the worst case design. We are able to extract a reduction by 1/1.4 in the average power consumption of a 16x16 CSM with statistical variations (RDD, LER, and PGG). Our simulations also indicate that future technologies might require very large guard band for reliable design that would cost high power/performance overhead in the case of conventional worst case designs.

The proposed 32 nm delay sensor uses an early clock edge to detect timing failures in advance. It can be applied to a DVS processor to detect a minimum supply voltage that ensures correct operation without requiring a complex error recovery mechanism which is essential for the Razor based design. By avoiding the delay buffer at data input, it significantly reduces performance overhead as compared to the Canary flip-flop, and the total performance overhead is less than 0.5% which is much smaller compared to earlier designs. Similarly the power overhead is about 4% which is quite small as compared to the Razor flip-flop (7.3%). Simulation results indicate that the proposed 32 nm delay sensor based design can reduce the power consumption by 1/1.7 as compared to the worst case design.

Both the delay sensors can be used to reduce the design margins for lower process technologies as well as their operation remains the same, however the energy reductions may be different. The proposed 32 nm delay sensor has a lower area overhead (less number of

transistors per sensor) as compared to the 45 nm delay sensor, however it may complicate the clock tree design due to the requirement of a delayed clock signal. The energy reductions are higher for the 32 nm delay sensor as it has a lower guard band and therefore can reduce the design margins further. The proposed delay sensors in this section can provide robust circuit operation for the combinational logic. It is important to note that variability has even higher impact on the sequential elements like SRAM cells as compared to the combinational logic due to their symmetrical nature. Increased variations can easily disturb the symmetrical balance achieved for latch elements (e.g. SRAM cells, sense-amplifiers) through careful sizing that can lead to functional failures. The next chapters present our work on SRAM design due to its higher influence on system performance, power, and cost.

# Chapter 4

# 4. Variability resilient SRAM designs

SRAM cache has been the preferred choice for decades to occupy the upper level memory hierarchy, including registers, on chip caches, and memory buffers, because it provides the highest access speed in embedded memories and seamless integration with the logic circuits [10]. Increasing size of the SRAM cache memory has therefore been an effective method to enhance system performance since it allows faster access to most of data/instructions. Current on-chip SRAM caches achieve performances that match state-of-the-art processor core speeds (3-4 GHz) whereas their counterpart, the off-chip DRAM caches, operate just around 600 MHz [10]. However the DRAM caches have capacities in gigabytes as compared to a few kilobytes for the on-chip caches. The emergence of multi-core architectures has driven the use of large SRAM cache memories to support high bandwidth and capacity requirements. This has resulted in SRAM caches to take up to a 90% of the total chip area [27]. Similarly SRAM caches take large portion of the total chip power consumption, especially the leakage power consumption which is proportional to the number of transistors [68].

SRAM cell sizes have reduced nearly by a factor of two with successive generation, doubling the number of on chip transistors to increase cache density that follows the Moore's Law. Aggressive scaling had lead to fabrication of over a 208 Mbit of SRAM caches in a $100\,mm^2$ area, with each cell taking an area of 0.346 $\mu m^2$, in 45 nm technology node for Intel [10]. However scaling the SRAM cells in nano-CMOS technologies faces different challenges including low noise margins and decreased cell currents. Standard 6T-SRAM cells are carefully designed to meet the constrained read/write requirements without increasing the area overhead. However increasing parametric variations and decreased supply voltages have reduced the cell noise margins, thereby decreasing the reliability of read/write and hold operations. Statistical variability in particular can result in each device of a SRAM cell to behave differently, disturbing the symmetrical balance achieved with sizing. Large threshold

voltage variations can cause functional failures, degrading the reliability of SRAM design [4]. Small cell sizes result in low cell currents that take longer to discharge the bit-lines, therefore lead to a higher discharge delay and large power consumption.

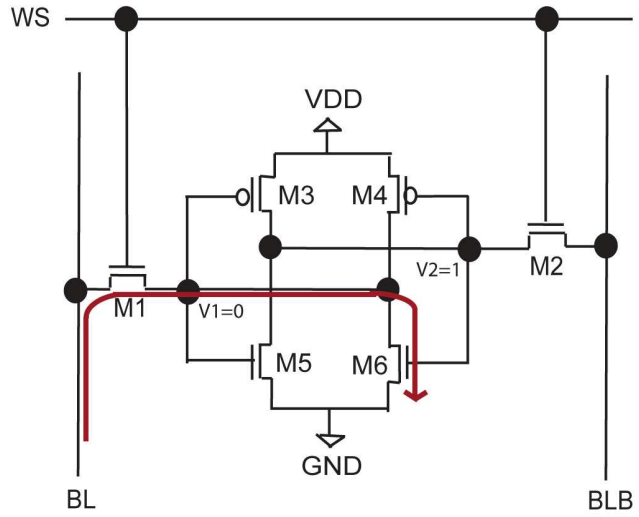
Different SRAM cell topologies have been proposed in the past that increase the robustness of read and write operation. Since SRAM cells have a relatively poor read stability as compared to the write operation, most of the previously proposed designs increase read stability. These include 6T [55, 57, 67], 7T [59], 8T [60-62], 9T [63], and 10T [65, 66] SRAM designs. A detailed discussion of these designs can be found in Chapter 2. This chapter will present operation of a conventional 6T-SRAM design, stability matrices for SRAM design, proposed asymmetric 6T-SRAM design [21], proposed SNM free 7T-SRAM design [20], and proposed fully differential 8T-SRAM design.

## 4.1 Standard 6T-SRAM design

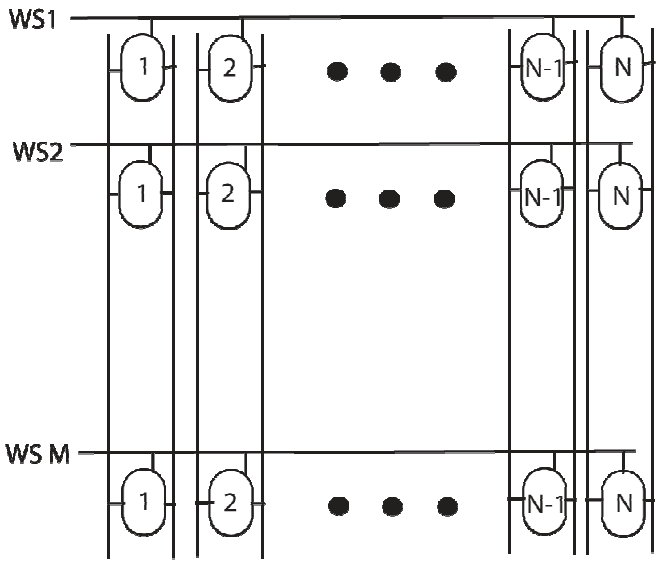
A standard 6T-SRAM cell consists of two cross coupled inverters (M3-M6) to hold the storage data and two access transistors (M1-M2) to provide controlled access for the read/write operation. Figure 4.1(a) shows the circuit schematic of a standard 6T-SRAM cell. The cross coupled arrangement allows a SRAM cell to hold data as long as the power supply is available due to the regenerative mechanism of the inverters. This avoids the need of refresh cycle required in the case of DRAM cells. The access transistors connect the bit-cells with the corresponding bit-lines (BL, BLB) and provide isolation in non-accessed periods. The word select line, WS, is held high to turn on the access transistors during a read/write operation. Figure 4.1(b) shows architecture of a conventional SRAM design. Bit cells are arranged in form of a matrix of size M x N, where M represents the number of rows and N represents the number of columns. A single word select line, 'WSi' is connected to a complete word. SRAM caches are arranged to have multiple bit-cell arrays. A row decoder is used to select a particular row, and a column decoder is then used to select a particular word that enables sharing of the sense amplifier with multiple bit-lines.

The bit-lines (BL, BLB) are pre-charged to VDD and the word select line, WS, is turned high to perform a cell read operation. The bit-line, BL, gets discharged when the internal cell node, V1, holds a 'zero'. The other bit-line, BLB, connected to the node, V2 that holds a 'one'

remains pre-charged. A voltage differential created on the bit-lines is then amplifier by the sense-amplifier to detect a zero or one being read. Voltage division between the access transistor, M1, and the driver transistor, M6, raises the voltage at node V1. This can cause a read failure if the raised voltage at node V1 is higher than the threshold voltage of the inverter (M3-M5) that can flip cell data. In the case of write operation, complementary data is loaded on the bit-lines and the word select line, WS, is held high. Strong full rail voltages on the bit-lines force the bit-cells to overwrite new data. A write failure occurs when a new data can't be loaded in the bit-cells.




(a)




(b)

**Figure 4.1: Standard 6T-SRAM design (a) 6T-SRAM cell (b) array architecture.**

## 4.2 SRAM stability metrices

The robustness of a SRAM cell is measured in terms of its static-noise-margin (SNM), write-noise-margin (WNM), hold-noise-margin (HNM), and cell current ($I_{cell}$). Continuous scaling has lead to a decrease in the noise margins with the increased process variations. Similarly the supply voltages have been scaled down in order to reduce the power consumption. This has resulted in reduced cell currents and consequently degrading the discharge delays. This section explains these measures to quantify the stability improvements with the proposed cell designs that are presented in the next sections.

### 4.2.1 Read margin

A standard 6T-SRAM cell has a very poor read stability and is more prone to failures under increased variations in scaled technologies. The read operation stability margin is expressed in terms of the static-noise-margin (SNM) that is defined as the maximum amount of noise tolerable at either storage nodes of a cell without causing loss or corruption of the stored data [10]. A widely used method to represent the SNM is by the use of butterfly curves to show the cross coupled inverters characteristics. Figure 4.2 shows a graphical illustration of butterfly curves and the corresponding circuit topology. The access transistors are turned on and connected to the supply voltage, VDD, to give the pre-charge voltages. The input voltage is swept from 0 V to 1 V (VDD) and the output node voltage (V1, V2) is plotted. The SNM is represented by an edge of the largest square enclosed by the two curves, taking the length of the minimum of the two squares (s2). Larger is the length of the square edge, the larger is the SNM and hence the read stability of the cell. Increasing the cell ratio is a common method to increase the SNM, where the cell ratio (CR), $\beta$ , [27] represents width ratio of the driver (M5, M6) and access transistors (M1, M2), shown in Figure 4.1(a).

$$\textbf{Cell ratio (CR)} = \beta = \frac{W_{M5}}{W_{M2}} = \frac{W_{M6}}{W_{M1}} \qquad \textbf{Equ 4. 1}$$

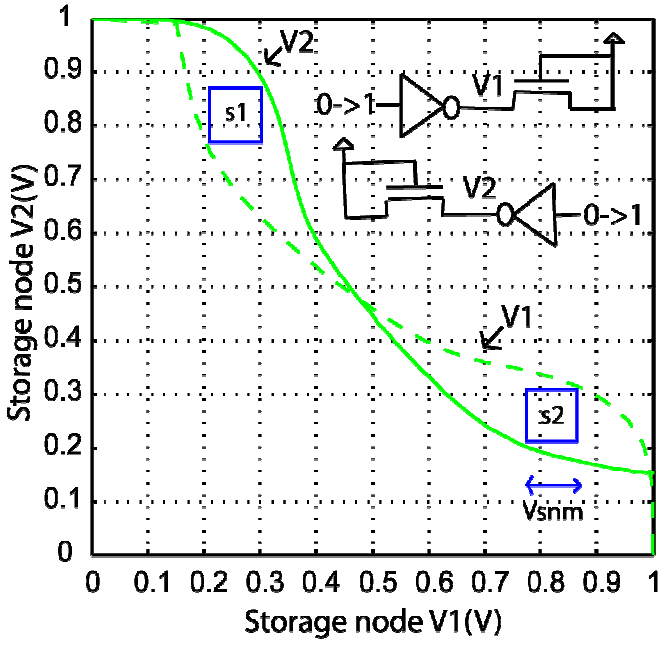



**Figure 4.2: SNM of standard 6T-SRAM.**

### 4.2.2 Write margin

The write stability is described by the WNM of a SRAM cell that represents the minimum bit-line voltage that can flip the state of a bit-cell [27]. Figure 4.3 shows the butterfly curves and the corresponding circuit topology to measure the WNM of a standard 6T-SRAM cell. As we can see the maximum square that can be enclosed in butterfly curve is much larger than the read butterfly curve, therefore, a standard 6T-SRAM cell has a much higher write noise immunity as compared to read. The stability of write operation, called the pull-up ratio (PU), $\gamma$ [27], depends on the width ratio of the pull up transistors (M3, M4) and access transistors (M1, M2) shown in Figure 4.1(a). Higher the $\gamma$ ratio, lowe is the WNM of a standard SRAM cell.

$$\textbf{Pull up ratio} = \gamma = \frac{W_{M3}}{W_{M2}} = \frac{W_{M4}}{W_{M1}} \qquad \textbf{Equ 4. 2}$$

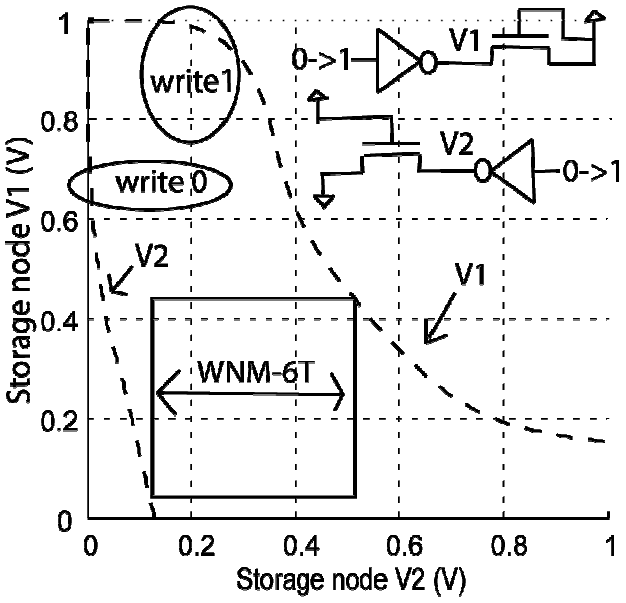



**Figure 4.3: WNM of standard 6T-SRAM.**

### 4.2.3 Hold margin

The hold margin represents the maximum amount of noise tolerable at the storage nodes of a SRAM cell during hold/idle periods without causing any loss of cell information. Figure 4.4 shows the butterfly curves and the corresponding circuit arrangement to determine the hold-noise-margin (HNM). Both the access transistors are tuned off (connected to the GND) as they isolate the bit-cell from the bit-lines during a hold period. Although a SRAM cell has a relatively large hold noise margin as compared to the read operation, however, the bit-cells are put in a low voltage operation in idle periods to reduce the leakage currents that results in a degraded HNM. Increased process variations result in an even further degraded hold stability margins.

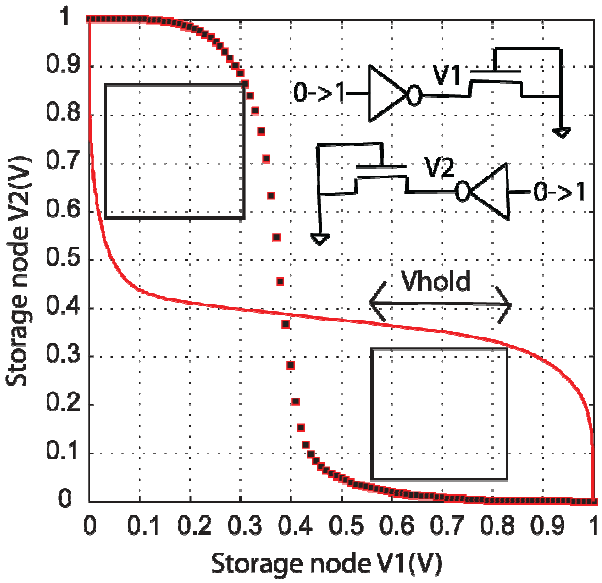



**Figure 4.4: Hold noise margin of standard 6T-SRAM.**

### 4.2.4 Cell current

Once the word line is selected for the read operation, one of the bit lines gets discharged depending if the access transistor connects it to a node holding a 'zero'. The discharge time of the bit-lines depends on the bit-line capacitance, cell current, and the required voltage differential for a reliable sensing [10]. The discharge current depends on the strength of pull down and access transistors that form a series path during the bit-line discharge. Although device scaling has reduced cell size, reducing per cell capacitance, however, the bit-line capacitance is not scaling proportionally that degrades the access delay overhead [13]. Moreover, low supply voltages reduce the cell currents that further degrades read delays.

## 4.3 An asymmetric 6T-SRAM design

Previously proposed single ended 6T-SRAM designs [55, 67] included an assist circuit to improve the read/write margins, however, a single ended write operation degrades the access delay. Another 6T-SRAM design was presented to reduce the write power consumption by a 1/10 and decreased the access delay by a 1/4-1/2 using a virtual ground line [58]. The ground line is floating during the write operation and a negative voltage is applied during the read operation. However this design doesn't provide any improvement in the read noise margins, while the use of negative supply voltage may degrade device reliability. An asymmetric 6T-SRAM design [57] was proposed to provide a differential write operation and a single ended

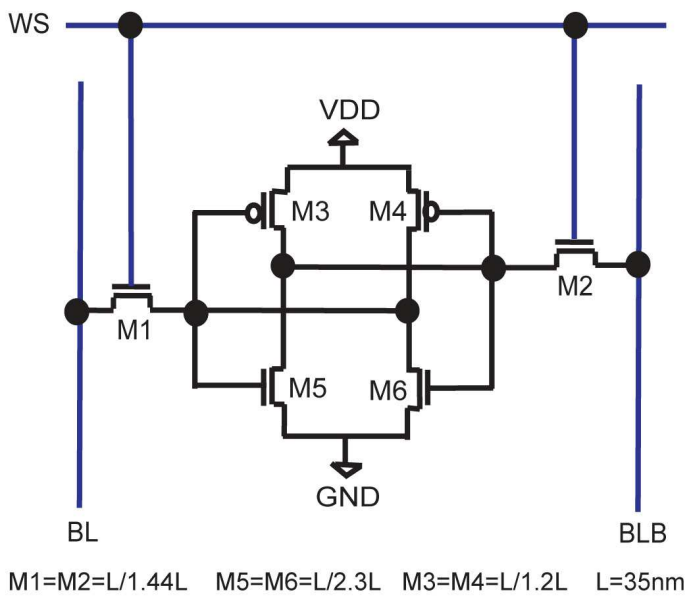
read operation. It thus avoids the write delay degradation, and improves the SNM by strengthening the feedback pull down transistor and/or weakening the forward pull down transistor. However, the constrained nature of this design doesn't allow further improvements in the noise margins. For example, it relies on strengthening the access transistors to increase the WNM that results in the degraded SNM. Thus section describes design, implementation, and simulation results for the proposed low-power asymmetric 6T-SRAM design. The proposed asymmetric 6T-SRAM design achieves a high read stability by strengthening the pull-down transistor of the feedback inverter for a single ended read operation. The access transistors can be kept minimum sized to suppress the bit-line leakage current and increase the SNM, without any degradation in the WNM. The improvements in the WNM come from a low overhead write assist transistor that is turned off during the write operation to weaken cell storage for the low-power write operation.

### 4.3.1 Proposed asymmetric 6T-SRAM cell

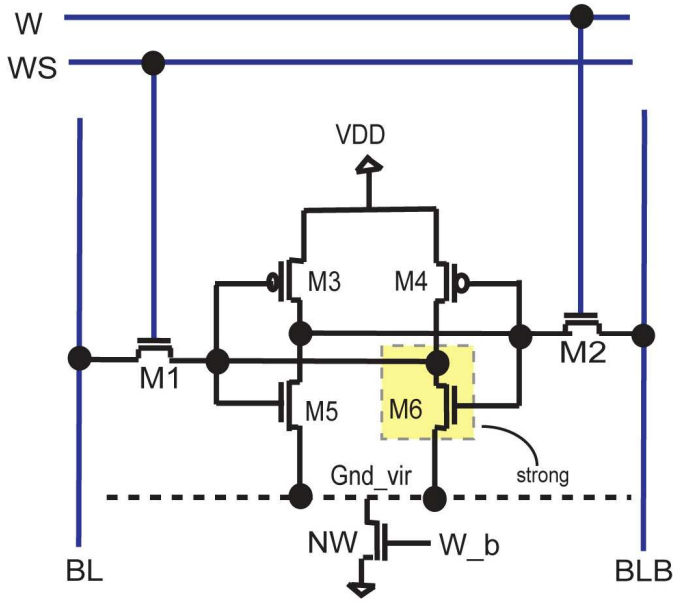
Figure 4. 5 shows the circuit schematic of a conventional symmetric 6T-SRAM cell and the proposed asymmetric 6T-SRAM cell. Both cells use a cross-coupled inverter pair for data storage, however, the proposed design uses an asymmetric inverter pair to enhance the SNM by taking advantage of the single ended read operation. A single ended read operation can result in cell disturbance at only one end of the cell that is connected to the bit-line. Therefore the driver transistor (M6 in (b)) can be made stronger to increase the cell ratio for a higher SNM. Since the other end of the cell is isolated during the read operation, therefore the forward inverter (M3, M5) can be kept minimum sized to reduce cell area. The access transistors are kept minimum size to increase the cell ratio thereby increasing the SNM. However a lower pull-up ratio, $\gamma$- ratio, degrades the WNM. Increasing the size of access transistors can improve WNM at the cost of degraded SNM as in the case of previously proposed asymmetric 6T-SRAM cell [57].

In contrast to the conventional sizing, we use a write assist transistor, NW, to increase the WNM without compromising the SNM as shown in Figure 4. 5(b). The write assist transistor, NW, is turned off during a write operation that eliminates the ground path for the cross coupled inverter pair, providing a virtual ground terminal, Gnd_vir. This stops the regenerative feedback mechanism and weakens cell storage. The cell can quickly change its

state and the write power is reduced due to absence of the true ground terminal (0 V). The write assist transistor is turned on during the read operation or hold period that allows the proposed SRAM cell to retain its data. To allow a differential write and a single ended read operation, the proposed design employs two word select lines, W and WS. The write word select line, W, is turned on only during a write operation that turns on the access transistor, M2, to provide differential write operation. The word select line WS is turned on for both the read and write operations. Figure 4.6 shows the timing diagram from HSPICE simulation for the conventional 6T-SRAM (Figure 4.6(a)) and the proposed asymmetric 6T-SRAM (Figure 4.6(b)). A single write assist transistor, NW, is shared for a complete word to minimize the area overhead, providing a virtual ground, Gnd_vir, to all cells connected to the word select line. A 350 nm process was use to do an area comparison by drawing the cell area layouts for both the conventional and the proposed 6T-SRAM cells. Figure 4.7 shows the area comparison for the layout of both conventional and proposed 6T-SRAM cells. Since the proposed design relies on using the minimum access transistors and minimum sized forward inverter, therefore, a 3% ($141.36\,\mu m^2$ vs. $145.61\,\mu m^2$) area reduction is possible with the proposed design even when the driver transistor, M6, is made larger.



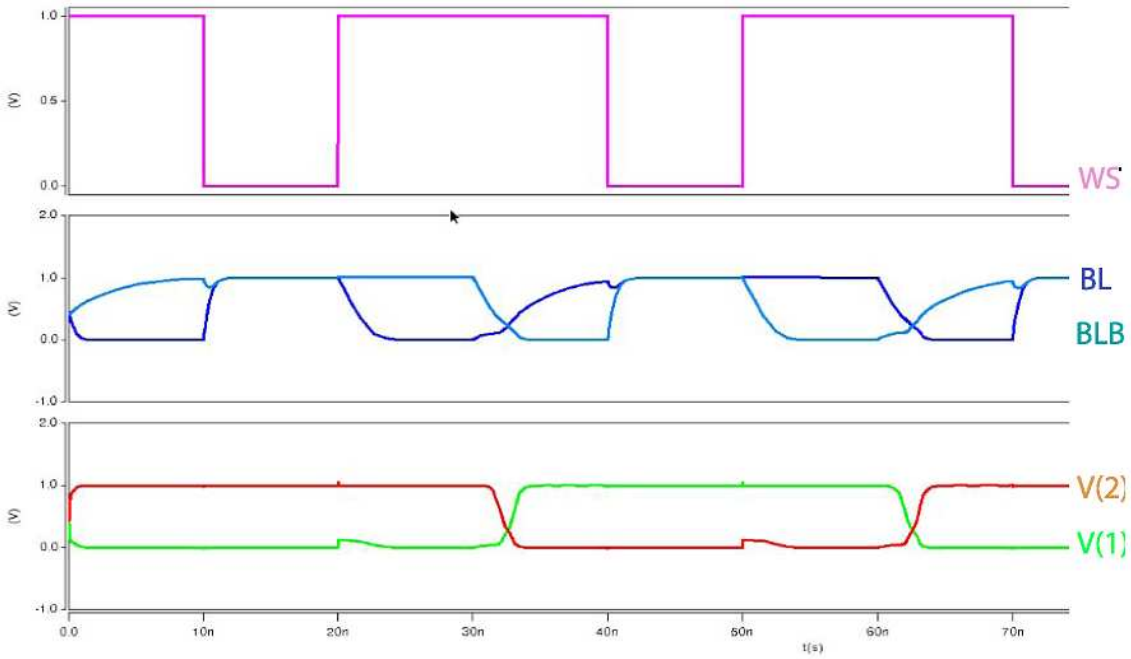

(a) Conventional 6T-SRAM cell



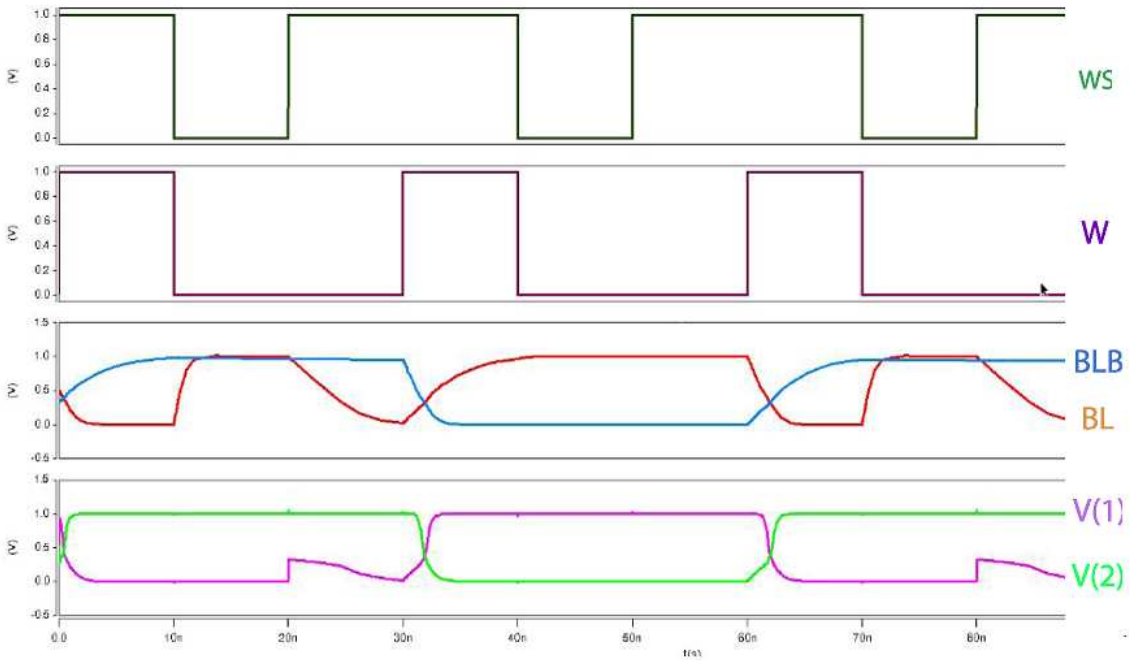

M6=L/2.4L, M1=M2=M3=M4=M5=L/1.2L L=35nm

(b) Proposed 6T-SRAM cell

**Figure 4. 5: Circuit schematic (a) conventional 6T-SRAM cell (b) proposed 6T-SRAM cell.**

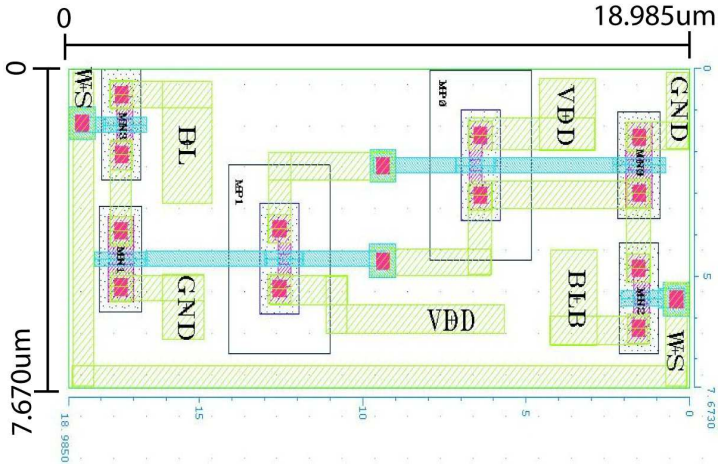



(a) HSPICE timing simulation for conventional 6T-SRAM design.

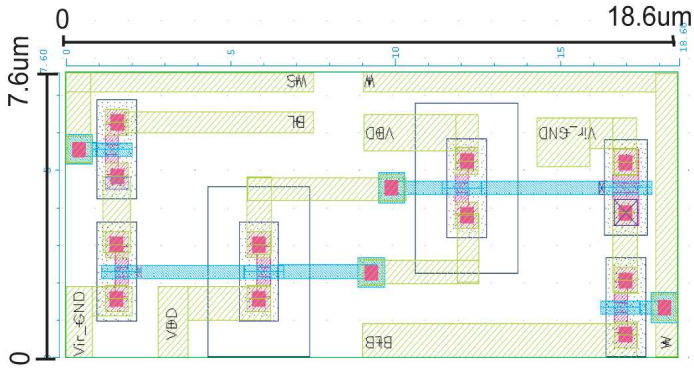



(b) HSPICE timing simulation for the proposed 6T-SRAM design.

**Figure 4.6: Timing diagram HSPICE simulation (a) conventional 6T-SRAM (b) proposed asymmetric 6T-SRAM.**




(a) Layout conventional 6T-SRAM cell




(b) Layout for the proposed 6T-SRAM cell

**Figure 4.7: Cell area comparison (a) symmetric 6T-SRAM cell (b) proposed asymmetric 6T-SRAM cell.**

### 4.3.2 Simulation results for the proposed A- 6T SRAM design

We have used HP-45 nm BSIM4 MOSFETs models [22, 89] with statistical sources of variability including random discrete dopants, line edge roughness, and poly grain granularity variations for comparing the noise immunity of standard and proposed 6T-SRAM designs. A 64x32 bits SRAM array was designed using 65 nm PTM device models [24] for delay and power comparison of both designs.

#### 4.3.2.1 Noise margins comparison

A comprehensive SNM analysis is carried out to quantify the improvements in the SNM for the proposed design. Figure 4.8 shows the simulation results for both 6T-SRAM cells (S/A) using uniform 45 nm device models without any variations. Figure 4.8(a) shows the butterfly curve for the standard 6T, proposed asymmetric 6T, and SNM free 8T SRAM cells. The proposed design provides a 1.6X (168 mV vs. 108 mV) improvement in the SNM over conventional 6T-SRAM design with cell ratios given in Figure 4. 5. It is evident that the proposed design provides better stability with similar cell area as compared to a symmetric 6T-SRAM cell. The butterfly curve (left upper side) for the proposed design resembles the butterfly curve for a SNM free 8T-SRAM design due to single ended read operation. Device sizing can be used, as it is a conventional practice, to increase the SNM for the conventional 6T-SRAM design. However, the advantages of scaling are nullified for low voltage SRAM designs as shown in Figure 4.8(b). By contrast, the proposed design achieves much better noise immunity at low voltages which is only possible because of the novel cell topology. This can be explained with the following equations [26],

*Linear region:*

$$I_d = \mu_n C_{ox} \frac{W}{L}\left((V_{gs} - V_{th})V_{ds} - \frac{V_{ds}^2}{2}\right) \qquad \textbf{Equ 4.3}$$

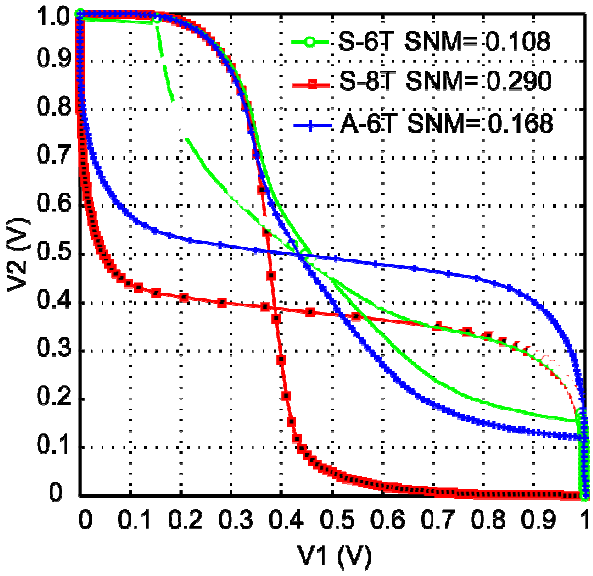
*Sub-threshold region:*

$$I_d = I_s e^{\frac{V_{gs}}{nkT/q}}\left(1 - e^{\frac{V_{ds}}{KT/q}}\right) \qquad \textbf{Equ 4.4}$$

Where,

$\mu_n$ --- Effective carrier mobility.
$C_{ox}$ --- Gate capacitance per unit area.
$V_{gs}$ --- Gate to source voltage.
$V_{ds}$ --- Drain to source voltage.
W, L – Width and length of the device.
$I_s$ and *n* --- Empirical parameters.
K --- Process transconductance.
T --- Absolute temperature.
*q* ---- Electric charge.

Equ. 4.3 shows that the drive current has a linear dependence on device width that results in some improvements in the SNM in the linear region. However the drive current has a quadratic dependence on the supply voltage, $V_{ds}$, that results in aggressive degradation in the SNM when the supply voltage is scaled. In the sub-threshold region, the device on-current has an exponential dependence on the threshold voltage, $V_{th}$, and supply voltage, while there is no dependence on the sizing (Equ. 4.4). Therefore, only increasing the size of the transistors (cell ratio) has a negligible impact on the SNM at low voltages for conventional design and requires new cell topologies [55].




(a)

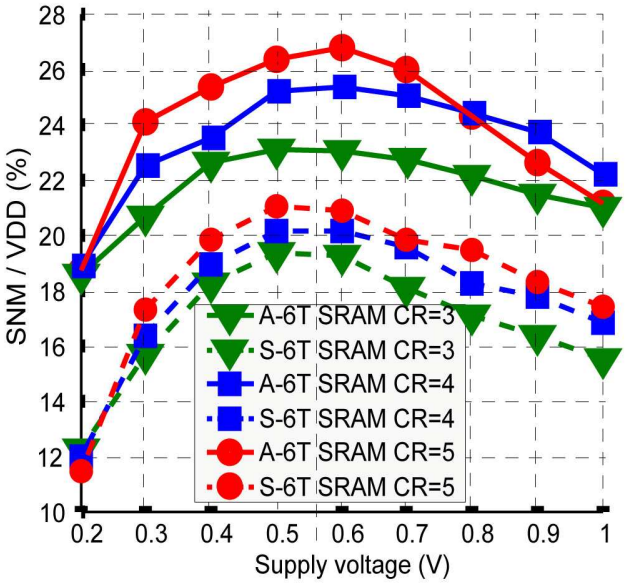



**Figure 4.8: SNM comparison (a) Butterfly curves (b) SNM vs. Supply voltage plot.**

An ensemble of 200 BSIM4 model cards with statistical sources of variability was used to investigate the impact of random variations on SRAM designs. Each transistor from both the cells (S-6T, A-6T) was replaced with a randomly picked model card from the ensemble to simulate statistical variations in SRAM designs. Read and write noise margins were then calculated for each randomized instance of both the cell for the noise margin comparison. Figure 4.9 shows the results of 8000 randomized circuit simulations to calculate the noise margins. Proposed asymmetric 6T-SRAM design provides a 1.9X (175 mV vs. 92 mV) improvement, on average, in the SNM over the conventional design for similar cell areas

when subjected to statistical variability as shown in Figure 4.9(a-b). Large eye opening of the butterfly curves for the proposed design indicate higher noise immunity with improved robustness to variations. The improvements in SNM are higher than found with uniform devices probably because one end of the proposed 6T-SRAM cell remains noise free for the single ended read operation, whereas both ends suffer variation and noise for the conventional 6T-SRAM design. The use of write assist circuit results in significant improvements in the WNM as shown in Figure 4.9(c-d). Virtual floating ground terminal during a write operation weakens cell storage and the cell is easily overwritten, therefore expanded write stability results in a 2.1X (380 mV vs. 789 mV) improvement in the WNM.

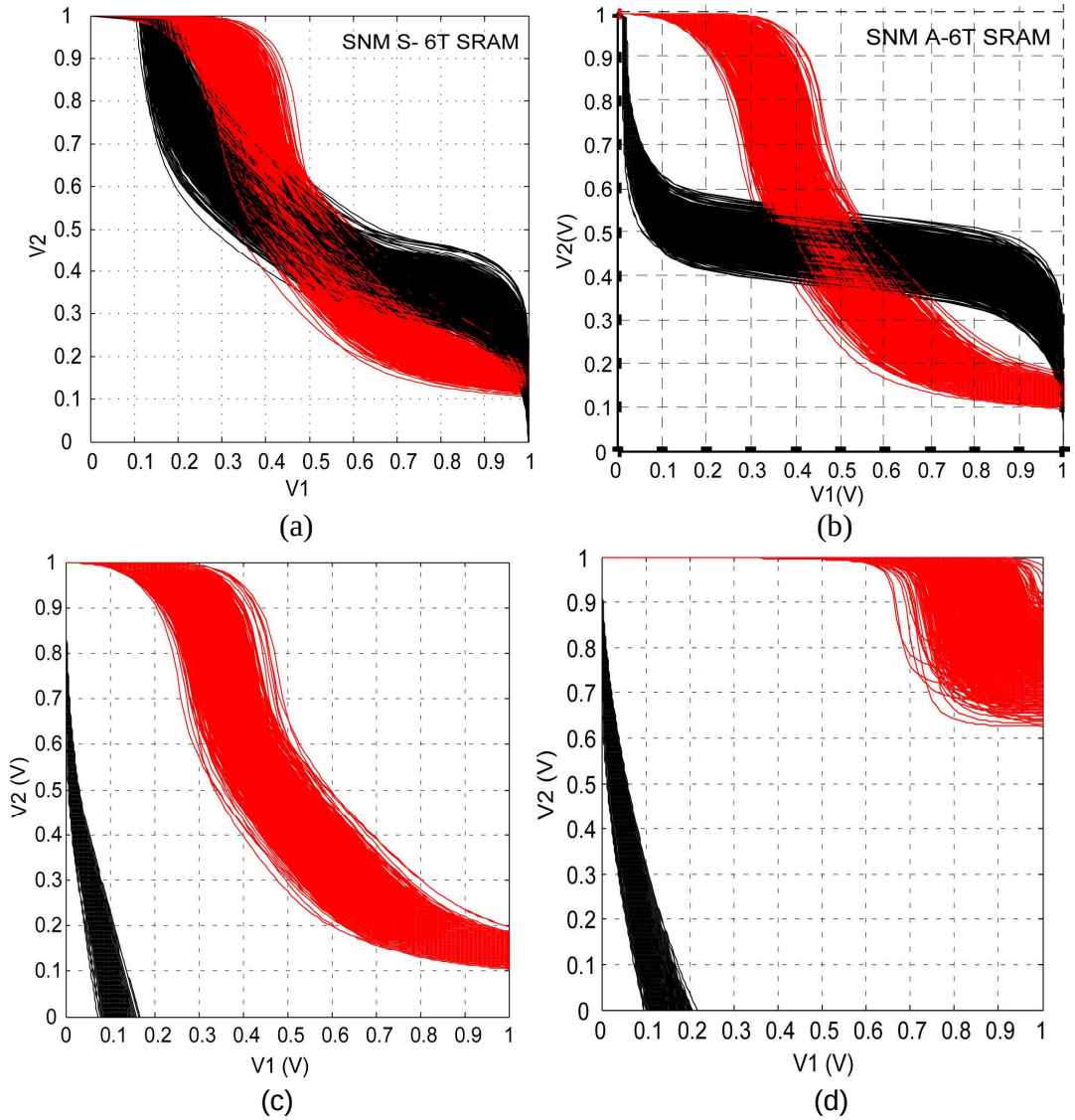
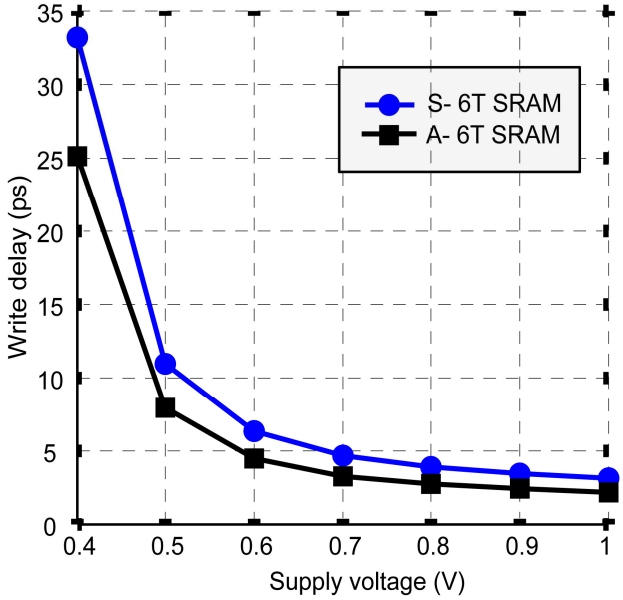


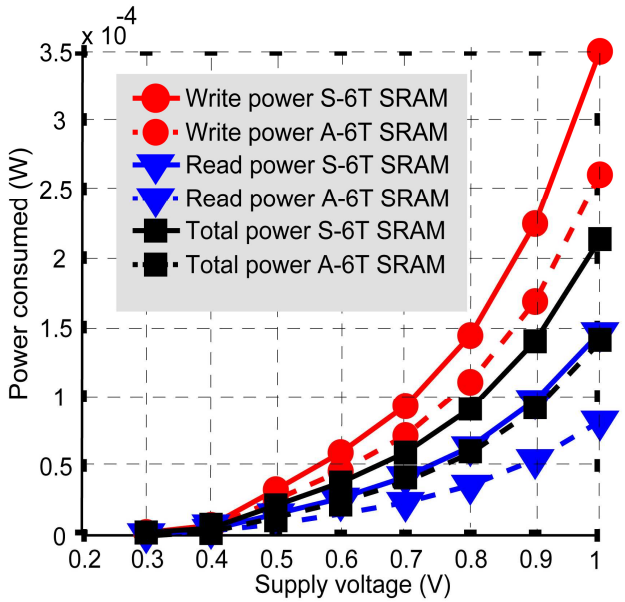

**Figure 4.9: Noise margins comparison (a) SNM symmetric 6T-SRAM (b) SNM proposed asymmetric 6T-SRAM (c) WNM symmetric 6T-SRAM (d) WNM proposed asymmetric 6T-SRAM.**

#### 4.3.2.2 Power and delay comparison

We designed a 64x32 bit SRAM array using 65 nm PTM models for power and delay comparison of both S/A (symmetric/asymmetric) SRAM designs. The 65 nm was chosen because its device and interconnect PTM models are available online. Turning off the write assist transistor during a write operation weakens cell storage that enables a faster write operation. Our simulations results indicate that the write delay reduces by 1/1.5 (3.11 ns vs. 2.13 ns) for a 1 V of supply voltage as shown in Figure 4.10(a). Similar improvements in the write delay are observed at very low supply voltages that makes the proposed design a suitable option for the low voltage applications. Turning off the ground path for the cross coupled inverter pair avoids the flow of short circuit current during switching of the inverter pair. A small short current may flow between the bit-lines (BL-M1-M6-M5-M2-BLB in Figure 4. 5(b)) during the write operation that result in a significant write power reduction. We found that the write power consumption reduces by 1/1.4 ($350\,\mu W$ vs. $260\,\mu W$) for the proposed asymmetric 6T-SRAM design as compared to the conventional symmetric 6T-SRAM design shown in Figure 4.10(b). A single ended read operation means only one bit-line is pre-charged and discharged during the read operation that results in the read power to reduce by 1/1.8 ($145\,\mu W$ vs. $82\ \mu W$). The total power consumption reduces by 1/1.5 ($214\,\mu W$ vs. $141\,\mu W$) for the proposed asymmetric 6T-SRAM design.




(a) write delay comparison




(b) power comparison

**Figure 4.10: Power and delay comparison (a) write delay (b) power.**

## 4.4 An SNM free 7T-SRAM design

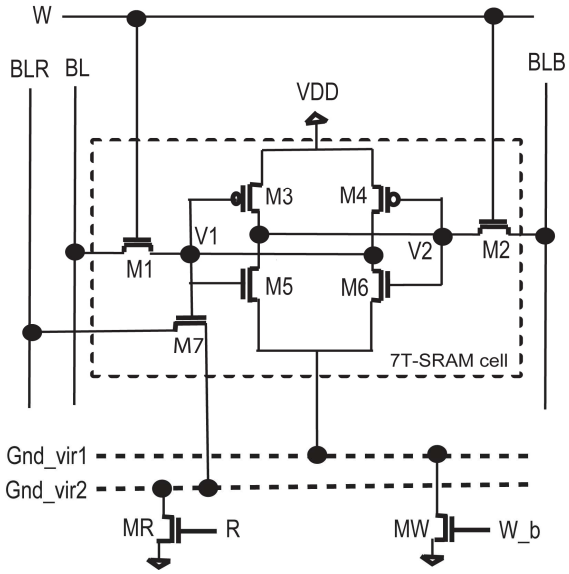
A number of 6T SRAM designs have been presented to improve the SNM as compared to the standard 6T-SRAM design, however either they don't provide a SNM free operation [21, 55, 57, 58] or they had high delay/power overhead [56, 67]. To provide an SNM free read operation without increasing the delay and power, a 7T-SRAM design was presented [59]. However it suffers from dynamic retention when one end storing a 'zero' floats for long periods during the read operation. Moreover the write noise margin is decreased at the low supply voltages and the read operation can destroy cell data. The cell area overhead is about 13% as compared to the standard 6T-SRAM design. 8T-SRAM designs [60-62, 70] do provide an SNM free operation, however they incur a 30% area overhead over standard 6T-SRAM. We propose a 7T-SRAM cell for the SNM free operation without incurring any increase in the write delay and power consumption. Our design is also free from dynamic retention problem found in previously proposed 7T-SRAM design. The cell area overhead is 16% as compared to the standard 6T-SRAM cell design.

### 4.4.1 Proposed 7T-SRAM cell

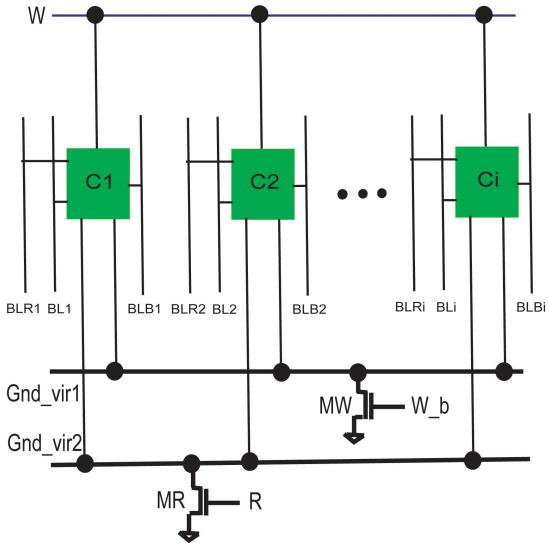
Figure 4.11(a) shows the circuit schematic of the proposed 7T-SRAM cell. It consists of two cross coupled inverters (M3-M5) to provide storage for the cell data as in case of the

standard 6T-SRAM cell. However unlike standard 6T-SRAM cell, we provide a virtual ground terminal Gnd_vir1 to the inverter pair that is floating during a write operation to weaken cell storage. Therefore it is easily overwritten and the write delay is reduced, while the write power consumption is also improved as the cross coupled inverters don't consume high dynamic switching power. The two access transistors (M1-M2) are dedicated for a differential write operation only and not used during the read operation. The word select line, W, is turned high only during the write operation to turn on the write access transistors (M1-M2). An extra transistor, M7, is added in the proposed SRAM cell to provide an SNM free operation. One end of the cell is connected to the read bit-line, RBL, for the read operation and the other end is connected to the virtual ground, Gnd_vir2, that provides a real ground (0 V) only during the read operation. The gate terminal is connected to one of the storage terminals to indicate if a zero or one is being read during the bit-line read operation.

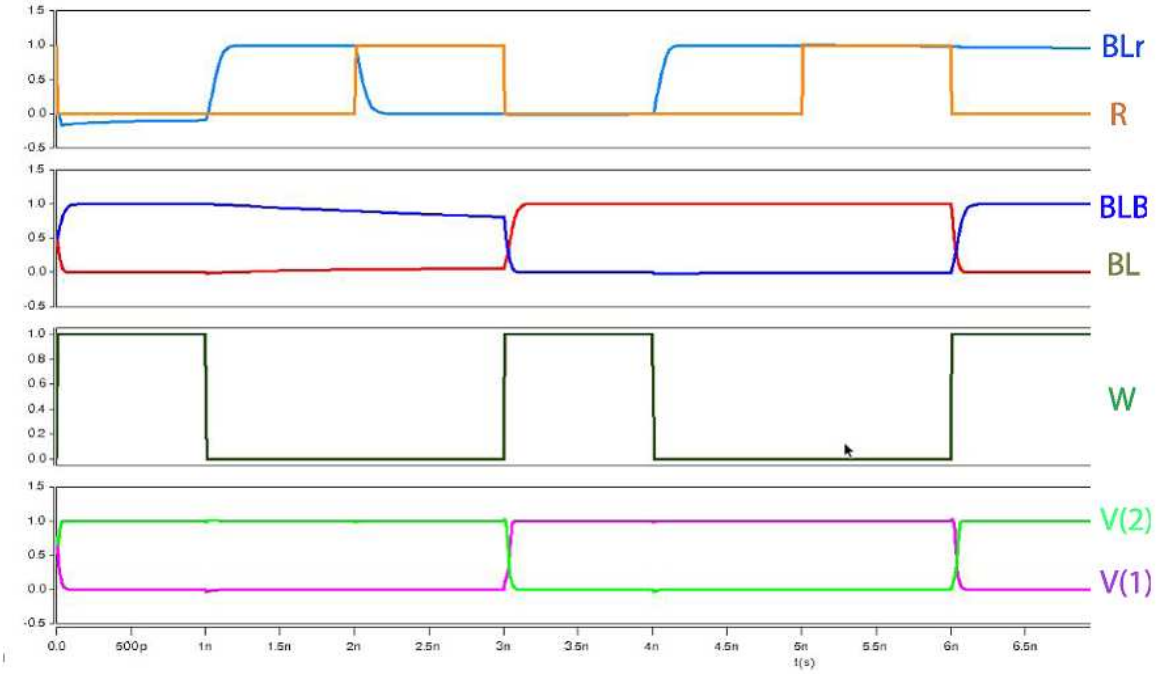
A conventional 6T-SRAM is prone to read failures because it provides a direct access to the storage nodes during the read operation. However the proposed design doesn't allow a direct node access that avoids the chances of cell data being corrupted. This also allows us to optimize the read and write operations independently since read/write operation is controlled by separate access transistors. For example, the driver transistors (M5, M6) in Figure 4.11(a) can be sized minimum with high-$V_{th}$ to reduce the leakage current without degrading stability of the read operation or without increasing the read delay because of the low cell currents. Use of high-$V_{th}$ (low leakage) devices is very important to reduce the total leakage power consumption during hold periods when the cells are not accessed. Similarly the read assist transistor can be made larger to minimize the degradation in read delays. The virtual grounds are shared for a complete word line to minimize the area overhead as shown in Figure 4.11(b). Figure 4.11(c) shows the timing operation of the proposed 7T-SRAM using HSPICE. A 350 nm process was used to do the layout of both (conventional 6T and proposed 7T) cells using 2 metal layers for the cell area comparison as shown in Figure 4.12. The proposed design incurs a 16% area overhead as compared to a standard 6T-SRAM cell when both were designed for minimum dimensions. We don't include the area overhead by assist circuit for cell area comparison as it is common for a complete word line.




(a) Proposed 7T-SRAM cell

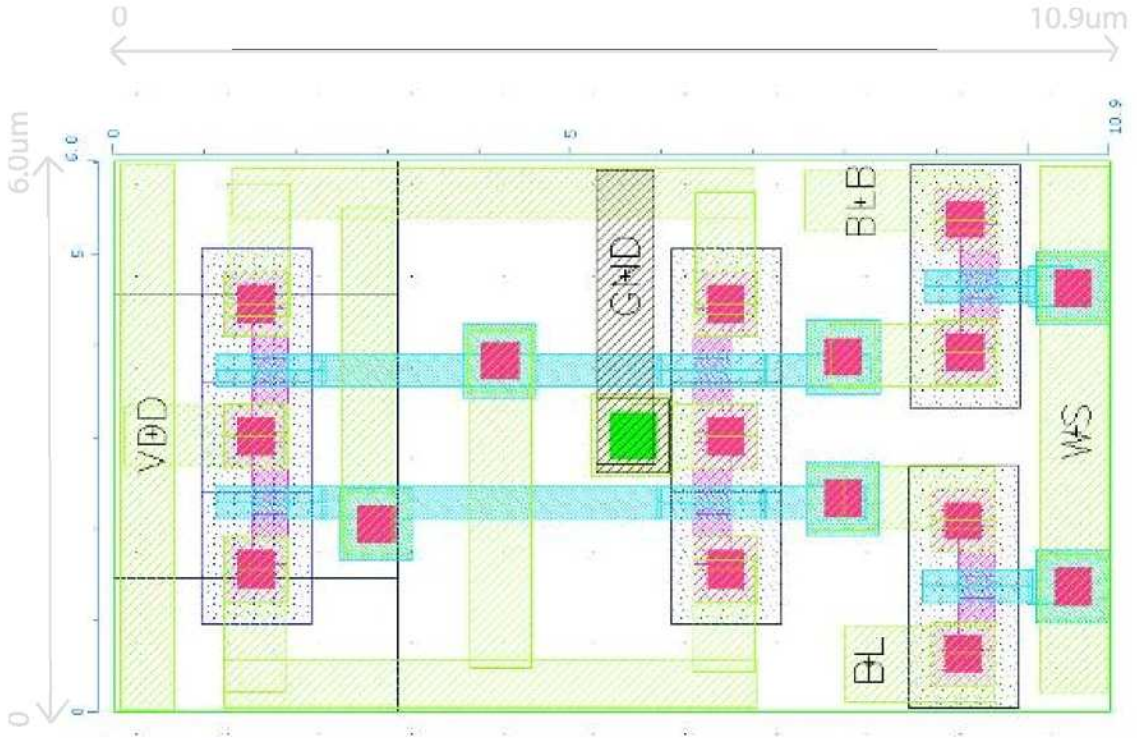



(b) Row configuration proposed 7T-SRAM design

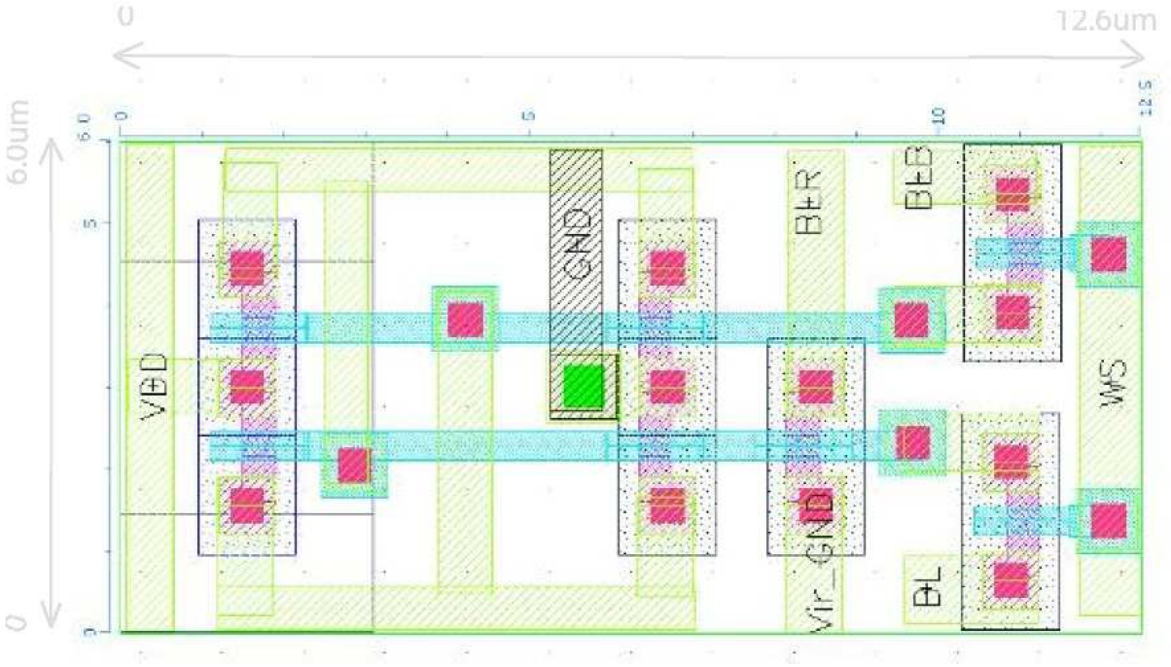



(c) HSPICE timing operation for proposed 7T-SRAM design

**Figure 4.11: Circuit design of the proposed 7T-SRAM (a) cell schematic (b) row configuration (c) timing diagram.**




(a) Conventional 6T-SRAM cell




(b) Proposed 7T-SRAM cell

**Figure 4.12: Cell area comparison (a) Layout of the conventional 6T-SRAM cell (b) Layout of the proposed 7T-SRAM cell.**

We have used a write assist transistor, MW, to provide a virtual ground, Gnd_vir1 shown in Figure 4.11(a). It is turned off during the write operation to weaken cell storage by eliminating the regenerative feedback mechanism that holds cell data. This allows new data to be easily overwritten and improves the write delay. The write margins are improved since a small differential voltage can be loaded on the bit lines that can overwrite old cell data due to sense amplifier behavior of the proposed cell. When the word line, W, is held high for the write operation, its complement, W_b, is turned low that turns off the write assist transistor providing a floating gate to the inverter pair. The word line, W, is held low during other periods (hold and read), therefore its complement, W_b, is high that turns on the write assist transistor, MW, and provides a true ground terminal (0V) to inverter pair connected to the virtual ground, Gnd_vir1. A floating ground terminal during the write operation puts driver transistors (M5-M6) and write access transistor (M1-M2) in series with the bit-lines (BL, BLB) that minimizes the short current that flows during the write operation, yielding significant energy reductions over a conventional design. The write assist transistor can be design for the minimum dimensions to reduce the area overhead as it only serves the purpose of weakening cell storage during the write operation. The other virtual ground terminal for the read operation, Gnd_vir2, is floating during a write operation.

To provide an SNM free read operation, we added a read assist transistor, MR, to provide a virtual ground, Gnd_vir2, to read access transistor, M7, as shown in Figure 4.11(b). A read signal, R, is held high to turn on the read assist transistor that then provides a true ground terminal (0V) to M7. An ON read assist transistor, MR, thus allows bit-line discharge for a read 'zero' if the storage node V1 turns on the read access transistor, M7. If a zero is stored at node V1, then the read access transistor is off and the read bit-line, RBL, remains charged at VDD, indicating a read 'one'. The read assist transistor is turned off by keeping the read control signal, R, low during write or hold operations to provide a floating ground terminal to read access transistor, M7. The read bit-lines, RBL, remain pre-charged since the ground terminal is floating and no major current (short current) flows between the bit-lines. The read assist transistor is shared for a complete word line to minimize the area overhead. However it requires careful sizing to achieve performance goals without incurring high cell area overhead. The write assist transistor, MW, is turned on during a read operation to provide a real ground terminal (0 V) to the cross coupled inverter pairs connected to the virtual ground terminal, Gnd_vir1.

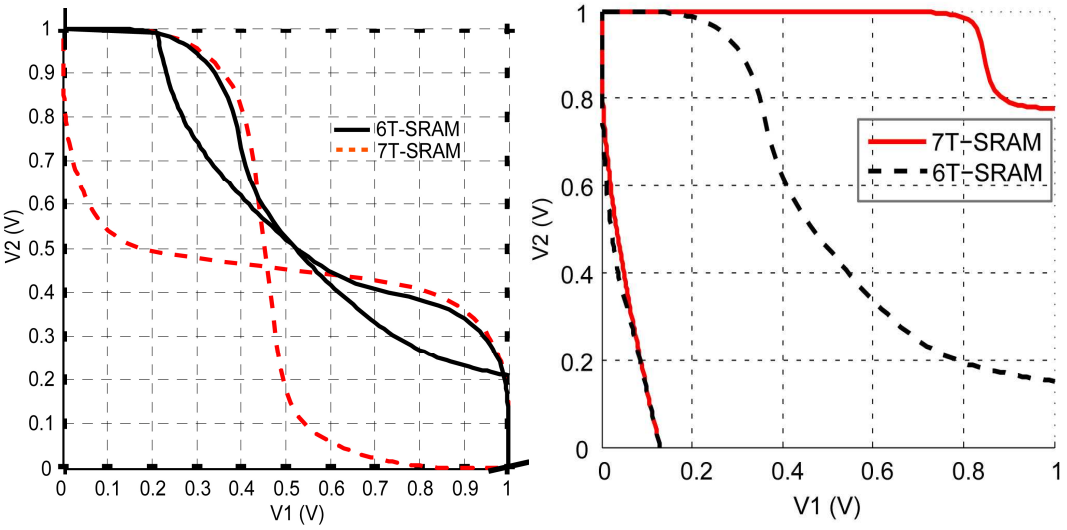
### 4.4.2 Simulation results of a 45 nm 7T-SRAM design

We have used 45 nm BSIM4 model cards for noise margins and energy/delay comparisons of the proposed 7T-SRAM design with conventional 6T-SRAM design. This section provides a discussion on the simulation results.

#### 4.4.2.1 Noise margins comparison

Figure 4.13 shows the read and write stabilities calculated for both designs using 45 nm models without any variations included. The proposed design provides a very high read stability due to the SNM free topology, and a 2.7X (112 mV vs. 299 mV) improvement in the SNM in observed as shown in Figure 4.13(a). The SNM was calculated from the butterfly curves with cell ratio, CR=1.5, for the conventional design and a cell ratio, CR=1, for the proposed design. Although this improvement in the SNM comes at the cost of adding an additional transistor to basic 6T-SRAM cell, however a move to SNM free designs using 7T and 8T SRAM cells would be necessary to provide high robustness in future technologies. Although a standard 6T-SRAM cell provides relatively high write stability compared to its read operation, however the degrading write stability due to high variability may become as

serious problem as is the read stability. Use of low overhead write assist circuit for the proposed design provides very high write noise immunity as shown in Figure 4.13(b). The WNM improved by a 2.1X (406 mV vs. 861 mV) for the proposed design as compared to the conventional 6T-SRAM cell.




(a) (b)

**Figure 4.13: Noise margins comparison (a) SNM (b) WNM.**

A straight forward method to improve the SNM of a conventional 6T-SRAM cell is to perform conventional sizing of the SRAM cell transistors. This is normally accomplished by increasing the cell ratio to increase the SNM or increasing the pull up ratio to increase the WNM. Our simulation results indicate that even conventional sizing may not be sufficient to provide a very large SNM even at the cost of a high cell area. Figure 4.14 shows the impact of supply voltage scaling and increasing the cell ratio, CR, on the SNM/VDD of a conventional 6T-SRAM. There is a reduction in the SNM as the supply voltage is scaled down. Although increasing the cell ratio provides some improvement at high voltages, the advantages of device scaling are negligible at low supply voltages for the conventional design, as explained for the asymmetric 6T-SRAM design previously. A conventional design can't achieve a high SNM even with a large cell ratio, e.g. CR=4, therefore a topological change in circuit design is required for high SNM as provided by the proposed 7T-SRAM design. These results indicate the sizing is less effective to cope with the increase variations and voltage scaling for conventional 6T-SRAM design.

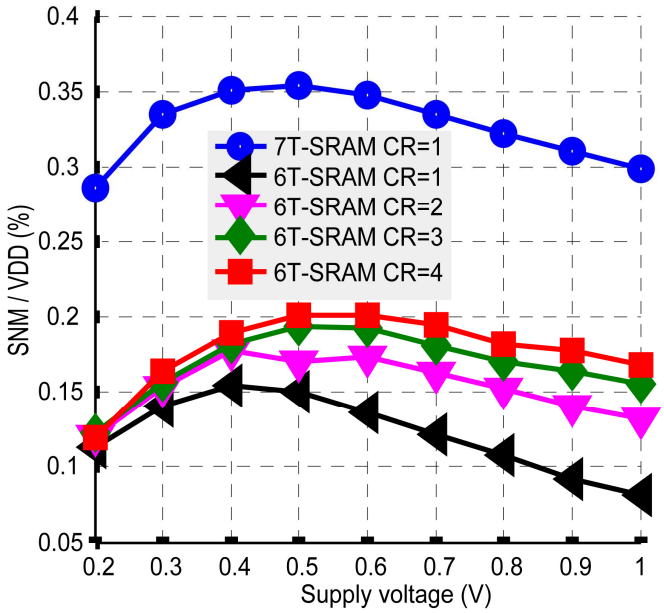



**Figure 4.14: Impact of supply voltage scaling on SNM for different cell ratios.**

Symmetrically designed SRAM cells are prone to random variations as they can cause each device to behave independently of others, and statistical variations pose a major challenge for robust SRAM design. We performed statistical variability simulations to compare the stabilities of both (standard 6T and proposed 7T) designs under high variability. A set of 200 randomized BSIM4 model cards were used to simulate impact of RDD, LER, and PoG variations on SRAM noise margins. Figure 4.15 shows the SNM comparison of both designs when subjected to statistical variability. The proposed design provides an SNM free operation and achieves a 3X (98 mV vs. 294 mV) improvement in the SNM over the standard 6T SRAM design with a higher cell ratio, CR=1.5. Figure 4.16 shows an instance of read failure for the standard 6T-SRAM cell (CR=1) when subjected to extreme statistical variations. The non-overlapping butterfly curves indicate a negative SNM would be required for a reliable read operation.

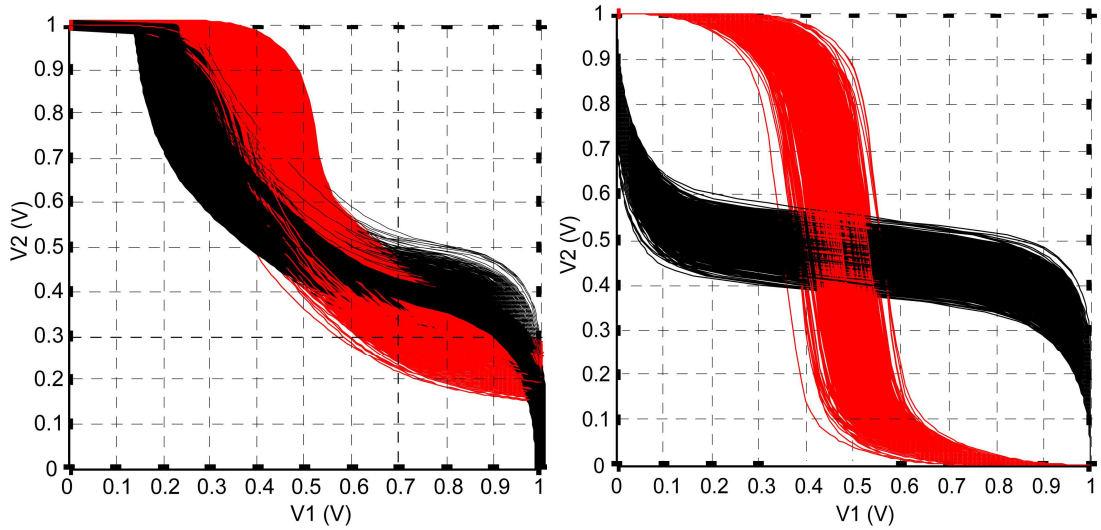



**Figure 4.15: SNM comparison (a) standard 6T-SRAM, CR=1.5 (b) proposed 7T-SRAM, CR=1.**

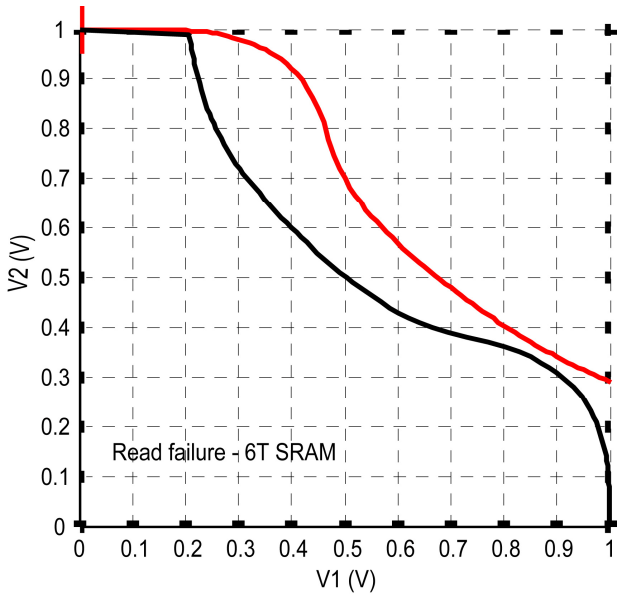
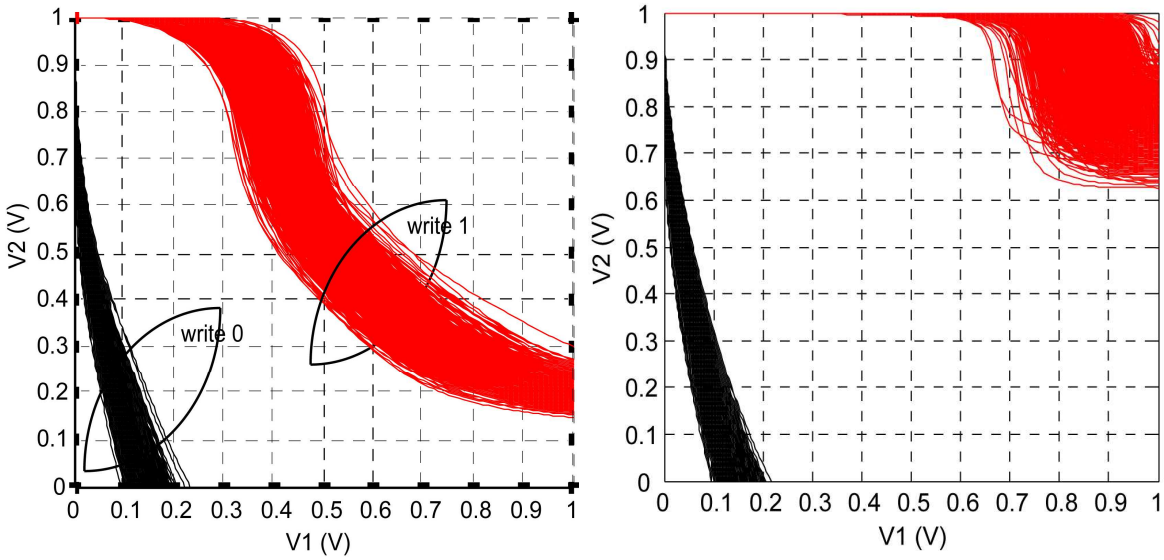



**Figure 4.16: Read failure due to high statistical variability.**

Figure 4.17 shows the results of 4000 statistical variability simulations to compare the WNM of standard 6T and the proposed 7T SRAM designs. The butterfly curves indicate a relatively high WNM for the standard 6T SRAM when compared to its read margin. However the WNM may be a case of concern under high variations in nano technologies considering the $6\sigma$ stability requirements for multi billion bits SRAM chips. With increasing variations, the spread in the WNM butterfly curves increases, and the required noise margin

($\mu_{snm} - 6\sigma_{snm}$) is degraded. A small amount of noise may be sufficient to cause write failures under extreme variations. The proposed design provides higher write stability and achieves a 2.2X (850 mV vs. 380 mV) improvement in the WNM.




(a) (b)

**Figure 4.17: WNM comparison under statistical variability (a) standard 6T-SRAM, CR=1.5 (b) proposed 7T-SRAM, CR=1.**

### 4.4.2.2 Power and delay comparison

We designed a 45 nm 64 x 32 bits SRAM array to perform power consumption and delay comparisons for both SRAM designs (6T vs. 7T). Figure 4.18 shows the write delay plot at different supply voltages for both designs. The proposed design reduces the write delay by a 1/1.3 (55 ps vs. 42 ps) at a 1 V of supply voltage. The weakened cell storage for the proposed design is easily overwritten and a substantial reduction in the write delay is observed. Similar improvements are observed at the low supply voltages as well. The write delay increases with the decrease in the supply voltage. The read discharge delay is higher for the proposed design as a single read assist transistor provides virtual ground to a complete row of SRAM cells. However it depends on sizing of the read assist transistor. A large sized transistor can minimize degradation of the discharge delay but would higher cost area/power overhead. Table 4. 1 shows the sizing arrangement of both the conventional 6T and the proposed 7T-SRAM designs for our power/delay simulations.

**Table 4. 1: Transistor sizing for 45 nm 64x32 bit SRAM**

| Width (L=35nm) | M1,2 | M3,4 | M5,6 | M7 | MW | MR |
|---|---|---|---|---|---|---|
| 6T | L | L | 1.5L | - | - | - |
| 7T | L | L | L | L | L | 16L |

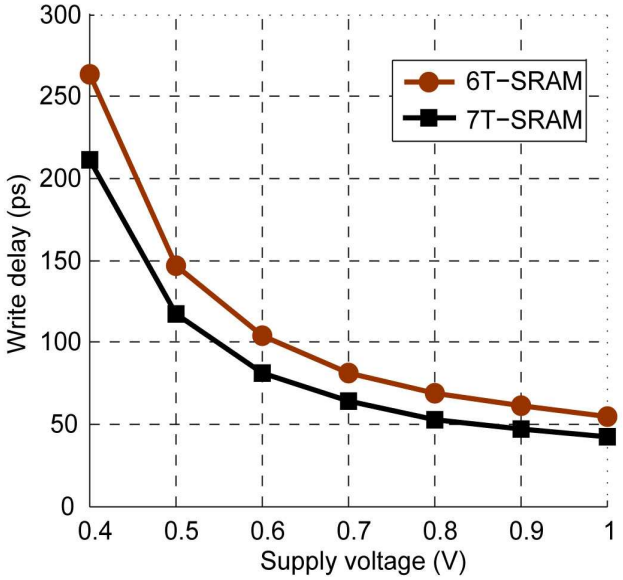



**Figure 4.18: Write delay comparison of standard 6T and proposed 7T SRAM designs.**

Figure 4.19 shows the plot of power consumption for both designs at different supply voltage for power comparison. By eliminating the true ground terminal of the inverter pair during the write operation, we minimize the dynamic power consumption of the inverters during switching. A small amount of short current may flow between the bit-lines as the floating ground puts the driver and access transistors in series with the bit-lines (BL-M1-6-M5-M2-BLB). Therefore the proposed design reduces the write power by 1/1.3 (98$\mu$W vs. 75$\mu$W) as compared to conventional 6T-SRAM design. The use of single ended read operation results in low pre-charge and discharge power consumption compared to the conventional design and the read power decreases by 1/1.6 (46.9$\mu$W vs. 29.3$\mu$W). The total power consumption decreases by a factor of 1/1.4 (63.9$\mu$W vs. 44.4$\mu$W) as compared to the conventional design.

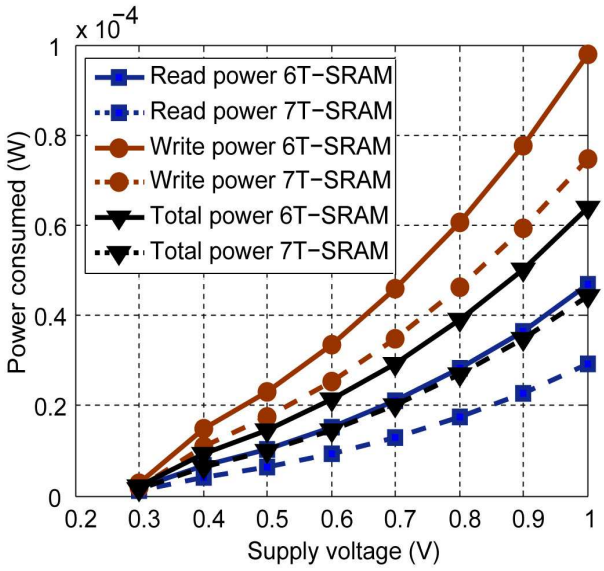



**Figure 4.19: Power consumption comparison of standard 6T and proposed 7T SRAM design.**

## 4.5 Fully differential 8T-SRAM design

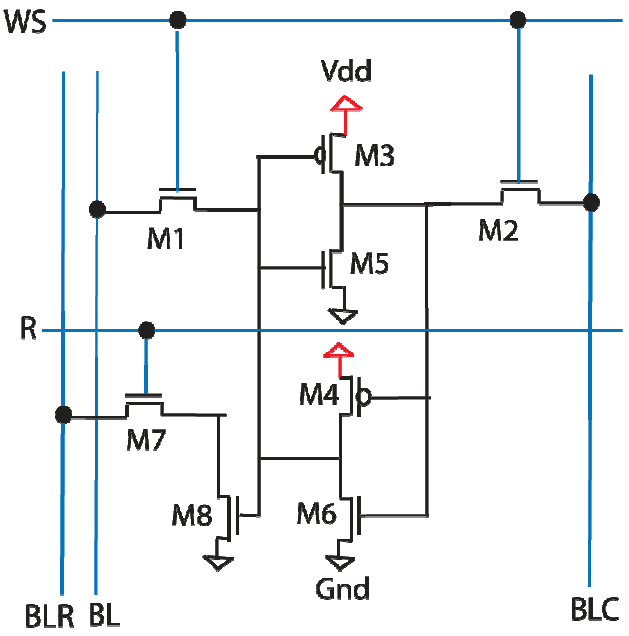
Different 8T- SRAM cells have been proposed in the past to provide differential write and a single ended read operation [60, 62]. Two extra transistors and a separate read bit line is added to a conventional 6T-SRAM cell that isolates read and write operations, providing the SNM free read operation. 8T-SRAM designs provide better stability than either 6T or 7T-SRAM cells. However a single ended read operation has a negative impact on the read speed since a differential sense amplifier is more sensitive to a small differential voltage and has a better common-mode-rejection-ratio (CMRR) as compared to a single ended sense amplifier. A 9T-SRAM cell was proposed for a fully differential read/write operation [90]. However it doesn't improve write margin and has a very high area overhead.

We present an 8T-SRAM cell that provides robust high speed fully differential read and write operations under increased variability. A low overhead write assist transistor is added to avoid the supply voltage to ground path of the cross-coupled inverter pair to weaken cell storage during a write operation as described for the asymmetric 6T and SNM free 7T- SRAM designs. This increases the write stability, increases the write speed, and decreases write power. A separate read assist transistor is added for an entire word line to provide the SNM free differential read operation that provides significant improvements in the read delay as compared to the conventional single ended SRAM designs.

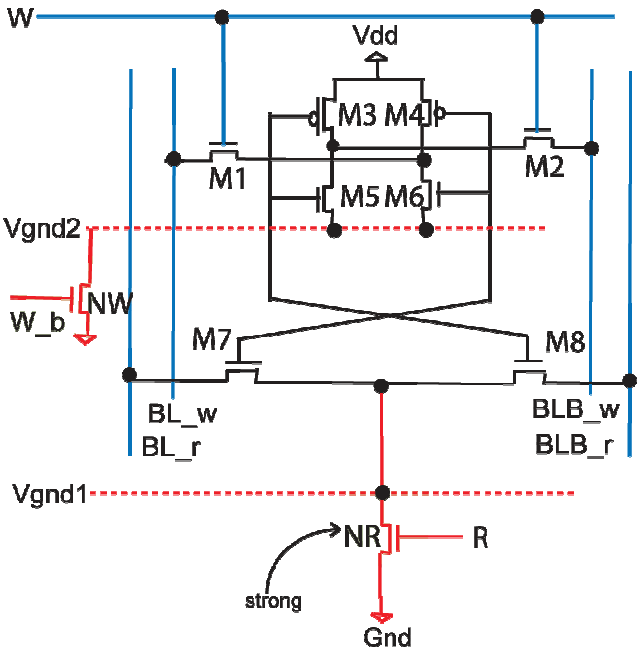
### 4.5.1 Proposed 8T-SRAM cell design

A conventional 8T-SRAM cell provides an SNM free operation by adding two additional transistors (M7-M8) and a separate read bit-line, BLR, to the conventional 6T-SRAM cell structure, shown in Figure 4.20(a). The bit-line discharge occurs when the read word line, R, is held high while the node stores a ‘one’ that turns on M8. Figure 4.20(b) shows the proposed 8T-SRAM cell with its associated read/write assist circuits. It consists of a cross coupled inverter pair as in the case of a conventional 6T-SRAM for storage purpose, two access transistors (M7-M8) used only during the read operation, and another two access transistors (M1-M2) used only during the write operation. We add two additional lines (BL_r, BLB_r) for the read operation only, and the two separate bit-lines (BL_w, BLB_w) for the write operation only. It allows an independent optimization of the SRAM cell design for both read/write operations.

The write access transistors (M1-M2) connect a cell with the write bit-lines (BL_w, BLB_w) when the write signal, W, is turned on. The read access transistors (M7-M8) connect the SRAM cell with the read bit-lines (BL_r, BLB_r) when the read signal, R, is held high. The two virtual grounds (Vgnd1 and Vgnd2) are provided to assist read and write operations. During a write operation, the write assist transistor, NW, provides a floating ground ‘Vgnd2’ to the inverter pair of the cell selected. For a read operation, the read assist transistor, NR, provides virtual ground, Vgnd1, and is connected to the read access transistors (M7-M8). The write assist transistor can be of minimum size as its purpose is to weaken cell storage during a write operation, however the read assist transistor ‘NR’ is carefully selected to minimize the access delay degradation. The driver transistors (M5-M6) can be of minimum size and high threshold to reduce the leakage current without degrading the read stability and speed. Figure 4.21 shows the timing diagrams for the conventional and the proposed 8T-SRAM designs.

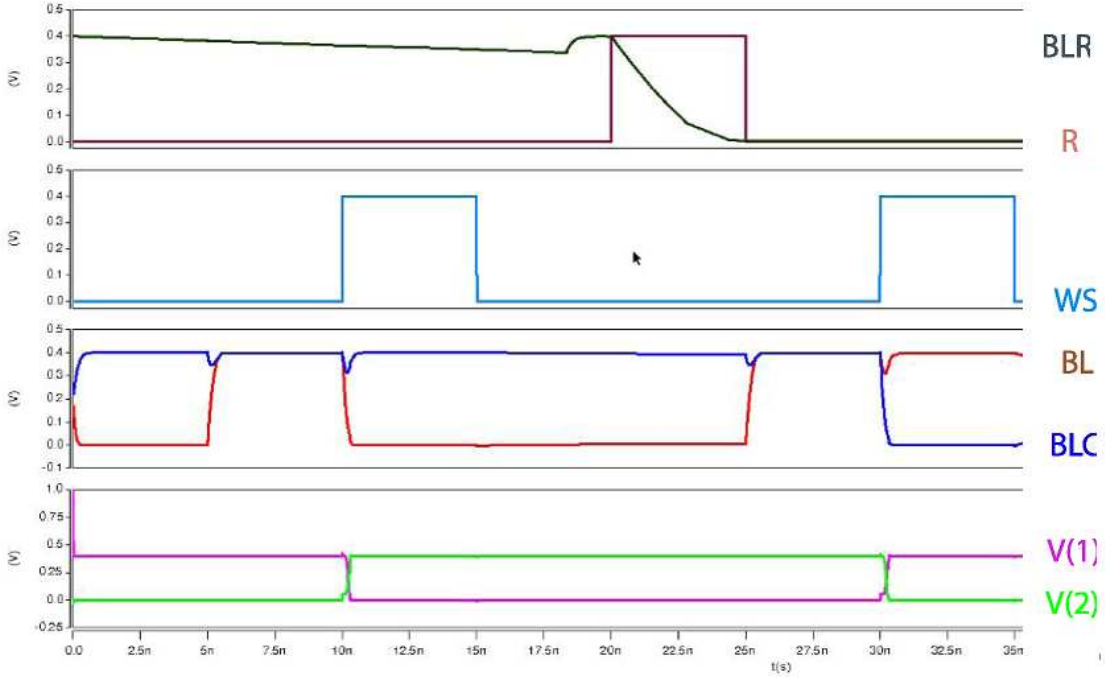



(a)

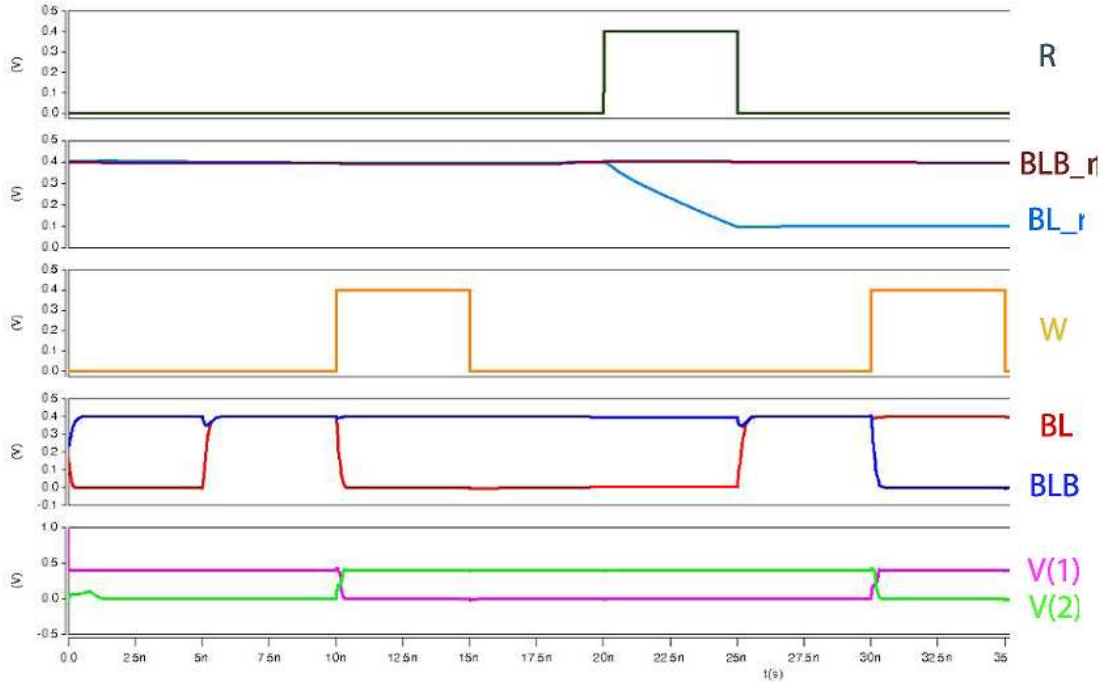



(b)

**Figure 4.20: Circuit schematic (a) conventional 8T-SRAM cell (b) proposed 8T-SRAM cell.**




(a)




(b)

**Figure 4.21: Timing diagram (a) conventional 8T-SRAM (b) proposed differential 8T-SRAM.**

### 4.5.2 Simulation results of a 45nm 8T-SRAM design

In order to investigate the impact of statistical variability on the reliability of SRAM design, we have used 45 nm BSIM4 models for our simulations. An ensemble of 200 model cards that included statistical variability was used in combination with C/MATLAB scripts for Monte Carlo simulations. During randomization our scripts randomly picked model cards from the ensemble and inserted in design. We have used a 350 nm process to perform cell area comparison of the conventional 6T-SRAM and the proposed 8T-SRAM cell. Figure 4.22 shows the layout for the proposed 8T-SRAM cell. The proposed 8T-SRAM cell incurs a 30% cell area overhead as compared to the conventional 6T-SRAM cell (shown in Figure 4.12(a)).

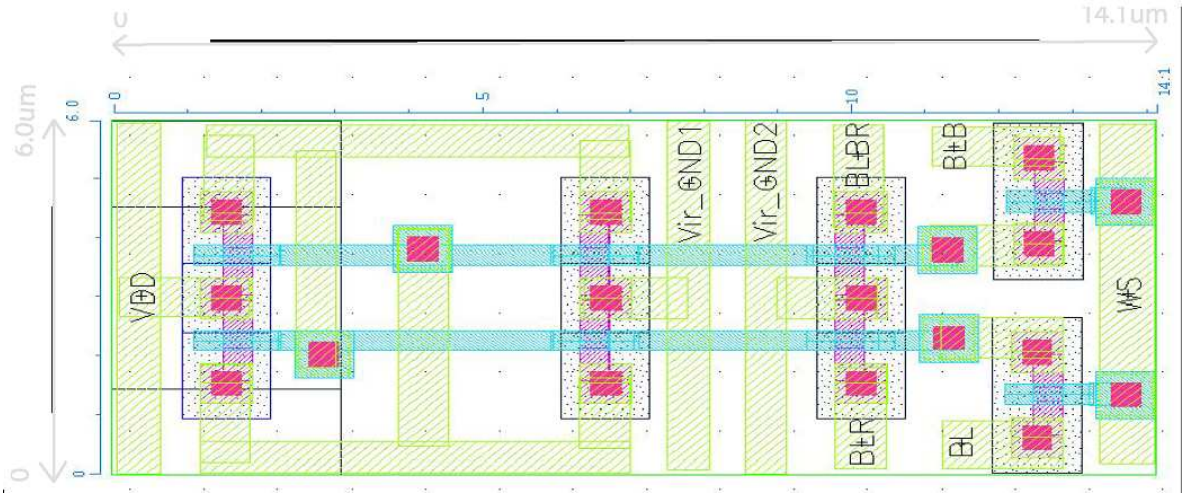



**Figure 4.22: Layout proposed 8T-SRAM cell.**

#### 4.5.2.1 Noise margins comparison

*A. Read operation*

The read bit lines (BL_r, BLB_r) are first pre-charged to VDD and the read signal ‘R’ is turned on during the read operation. The gate terminals of the read access transistors are connected to the outputs of inverter pairs. Assume the storage node connected to the bit-line, BL_r, by the write access transistor, M1, hold a 0. During a read ‘0’ operation, BL_r gets discharged through the read assist transistor NR and M7, while BLB_r stays at pre-charged level. A correct output can then be evaluated by a differential sense amplifier. When reading ‘1’ on bit line BL_r, it stays at pre-charged level and BLB_r gets discharged through the read assist transistor NR and M8. Since the read operation doesn’t disturb the cell content, therefore an SNM free, high performance read operation is performed.

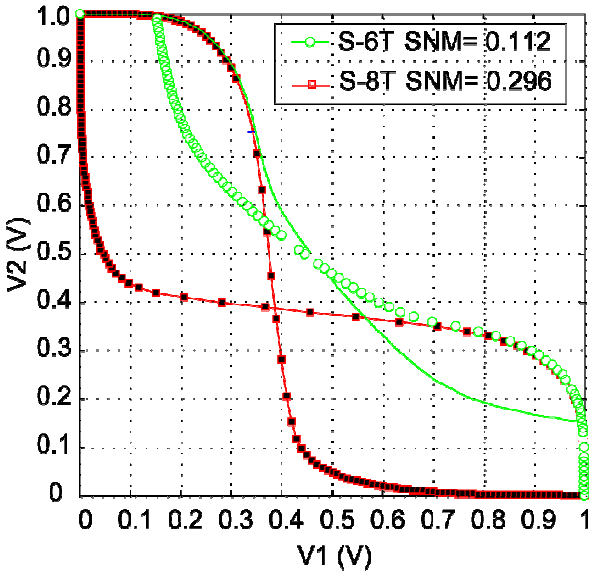



**Figure 4.23: SNM plot of both SRAM cell designs (6T and 8T).**

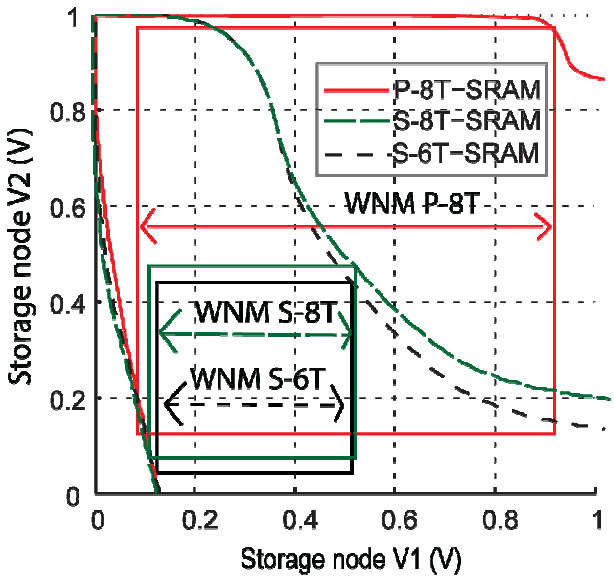
Figure 4.23 shows the SNM plot for a conventional 6T-SRAM cell and the proposed 8T-SRAM cell using 45 nm uniform models (without any variability source). The proposed design provides about 2.6X (296 mV vs. 112 mV) improvement in the SNM for a cell ratio, CR=1, over the conventional 6T-SRAM design with a cell ratio, CR=1.5. This improvement comes from the fact that an 8T-SRAM design doesn't allow a direct access to storage nodes of the cell during a read operation. In order to investigate the impact of statistical variability on the read stability of both SRAM designs (conventional 6T and proposed 8T), we carried out 4000 HSPICE simulations of the randomized instances of both SRAM cells (conventional 6T and proposed 8T). The results achieved show similar improvements in the SNM as for the proposed SNM free 7T-SRAM design, shown in Figure 4.15.

*B. Write operation*

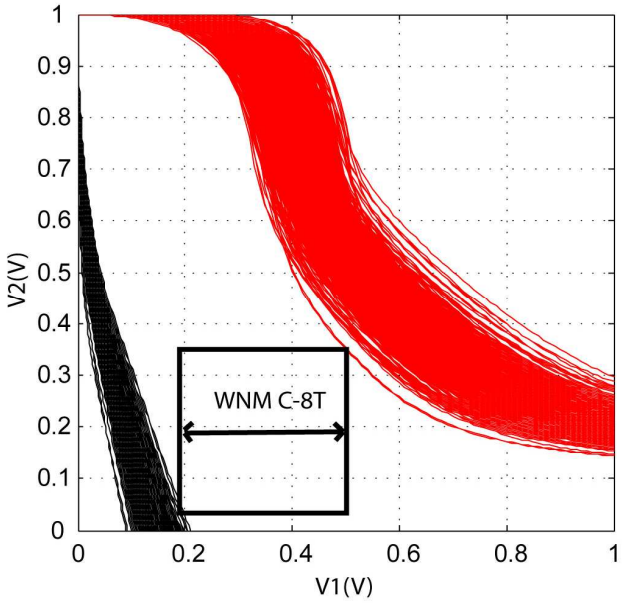
During a write operation the data to be written is loaded on the write bit-lines (BL_w, BLB_w), and the write select line, W, is pulled high. This turns on the access transistors (M1-M2) and the data is written into the cells selected by the write select line. When the word line, W, goes high, W_b goes low and turns off the write assist transistor, NW. This eliminated the supply to ground path for the inverter pair of the cell selected. This breaks the feedback path, and stops regeneration of cell data that weakens cell storage and the cell data is easily over-written as explained for previously proposed 6T and 7T-SRAM designs.

As discussed before for the asymmetric 6T-SRAM and SNM free 7T-SRAM designs, we created 4000 randomized versions of the conventional 8T SRAM cell and the proposed 8T-SRAM cell to analyze the impact of statistical variability on the write stability. Figure 4.24(a) provides write stability comparison of conventional (6T and 8T) SRAM design vs. proposed 8T-SRAM design using uniform 45 nm devices. A conventional design provides little improvement in WNM over 6T design, however proposed design improves it by 2X (861 mV vs. 436 mV) and by a 2.1X (861 mV vs. 406 mV) as compared to conventional 8T and 6T SRAM designs, respectively.

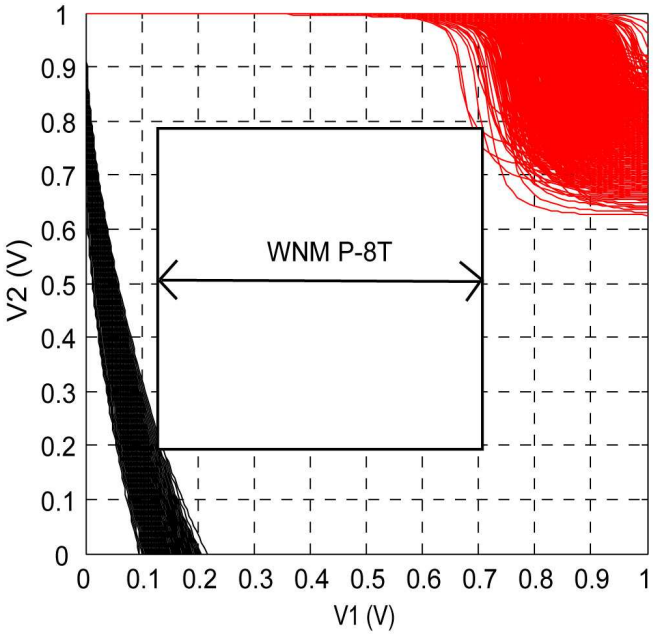
Figure 4.24(b-c) shows impact of statistical variability on the WNM of both 8T-designs. Although a conventional 8T-SRAM cell provides high write noise margins, however extreme statistical variability can significantly reduce this margin, and in worst cases can cause write failure. Figure 4.24(c) illustrates the impact of statistical variability on WNM of the proposed 8T-SRAM cell. By providing a floating ground to the inverter pair during a write operation, we significantly improve the write noise margin. Simulation results indicate on average 2X (430 mV vs. 850 mV) improvement in the write stability for the proposed 8T-SRAM cell over conventional 8T-SRAM design.




(a) Write margins comparison- without variations




(b) Write margins for conventional 8T-SRAM design



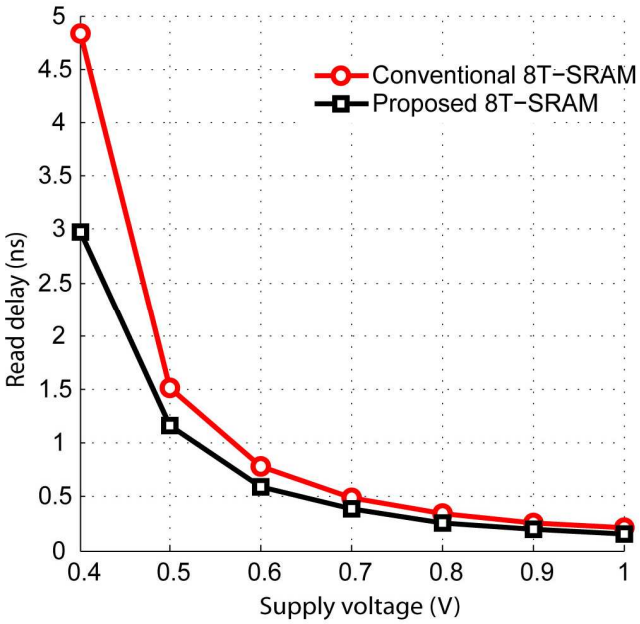

(c) Write margins for proposed 8T-SRAM design

**Figure 4.24: Write stability comparison (a) WNM margins without variability (b) WNM conventional 8T-SRAM (c) WNM proposed 8T-SRAM.**

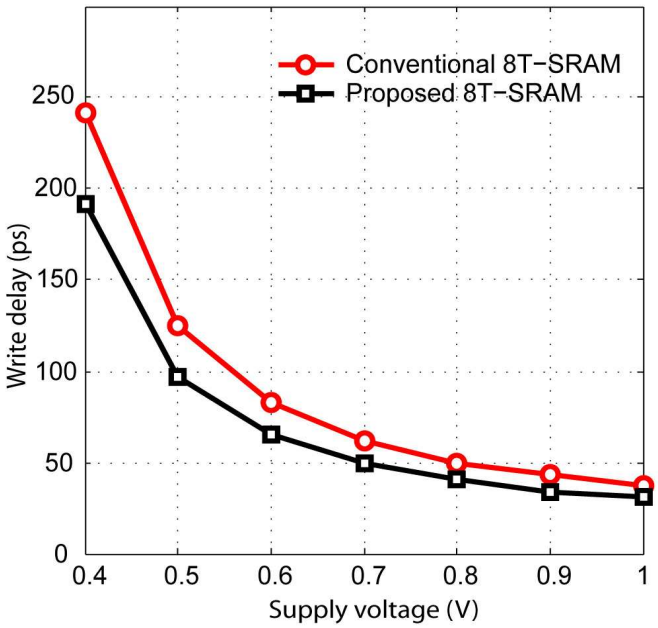
#### 4.5.2.2 Read/write delay comparison

In order to carry out read/write delay analysis for a comparative study of both designs, we designed a 45 nm 64X32 bit SRAM array using both cells (conventional 8T and proposed 8T). Consecutive read after write operation were performed for a 3 bit sequence '010' at different supply voltage conditions. Figure 4.25(a) shows the write delay comparison for both

SRAM designs. The write time is improved by 1/1.2 (37 ps vs. 31 ps) at a supply voltage of 1 V due to weakened cell storage during the write operation. This trend in speed improvement is followed even at very low voltages and the write delay improves by 1/1.3 (242 ps vs. 192 ps) at a supply voltage of 0.4 V. Figure 4.25(b) shows the read delay comparison for both designs when a discharge differential of 400 mV (singe ended 8T-SRAM) and 200 mV (proposed differential 8T-SRAM) is required. A single ended design would require twice as much discharge on a single bit line as compared to the differential discharge [14], therefore it requires more read discharge delay for a reliable sensing. The proposed design provides an improvement by 1/1.3 (208 ps vs. 159 ps) in the read discharge delay at a supply voltage of 1 V. Similar delay reductions are achieved at low supply voltages as well. Although pre-charging both bit-lines results in a higher read power consumption for the proposed design. However the low discharge period for differential sensing offsets this overhead for the proposed design and the total energy consumptions are similar for both designs.




(a) Delay-read operation




(b) Delay-write operation

**Figure 4.25: Delay comparison of a 45 nm 64X32 bit SRAM design (a) read operation (b) write operation.**

#### 4.5.2.3 Energy comparison

Figure 4.26 shows the energy plot for both the conventional 8T and the proposed 8T-SRAM designs. The proposed design has a faster write speed and consumes less power due to a floating ground terminal during the write operation. Therefore the write energy reduces by 1/1.7 (100 fJ vs. 59 fJ) at 1 V of supply voltage. The read delay was calculated for 200 mV of the discharge differential voltage for the proposed differential 8T-SRAM design and 400 mV discharge differential for the conventional single ended 8T-SRAM design. Therefore the read discharge period was lower for the proposed design. However it required a pre-charge of both the bit-lines. The energy comparison shows that both designs consume similar read energies (the overhead is less than 1% at 1 V of supply voltage) as small discharge delay compensates for an increase in the energy consumption due to differential sensing.

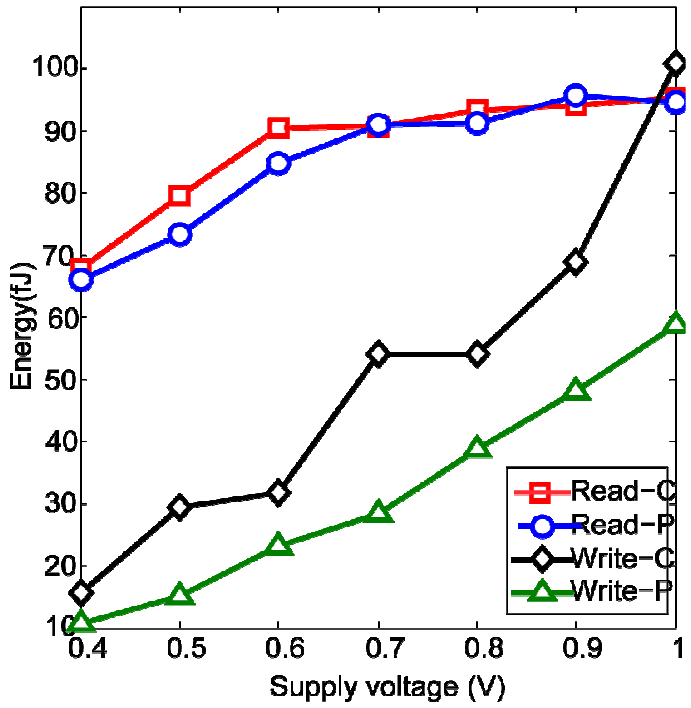



**Figure 4.26: Energy comparison of conventional and proposed 8T-SRAM designs.**

## 4.6 Summary and conclusion

SRAM caches are an important part of modern processor technology and require a handcrafted design to meet constrained stability requirements. Increased process variations in nano-CMOS technologies and the supply voltage scaling have threatened the reliability of conventional 6T-SRAM design. We have presented an asymmetric 6T-SRAM cell that provides a 1.9X improvement in the SNM and a 2.1X improvement in the WNM for similar cell areas. Proposed design use a single ended read operation with a strong driver transistor for the feedback inverter to improve the SNM. A write assist transistor is added to provide a floating ground terminal to the cross coupled inverters during the write operation. It increases write margins, write speed, and decreases write power. The write delay improves by 1/1.5 and the write power improves by 1/1.4.

Although proposed 6T-STAM design provides significant improvement in the SNM over a conventional 6T-SRAM design, however, increased variations may cause stability problems. A 7T-SRAM design is presented for the SNM free read operation. A read access transistor is added to a conventional 6T-SRAM cell structure and two virtual grounds are provided for read and write operations. A floating ground is provided to the latch structure of the cell

during the write operation and a true (0 V) ground terminal is provided during the read operation to read access transistor for the SNM read. Proposed 7T-SRAM design provides a 3X improvement in the SNM and 2.24X improvement in the WNM. The delay decreases by 1/1.3 and the write power decreases by 1/3, while the total power reduction is by 1/1.4. The circuit incurs a 16% area overhead compared to standard 6T-SRAM cell. To further improve on our 7T-SRAM cell in terms of high speed read operation, we propose an 8T fully differential SNM free SRAM design. The proposed design allows differential sensing operation that result in a 1/1.3 improvement in the discharge delays. The write delay improves by 1/1.2 and the write energy decreases by 1/1.7. The proposed design incurs about 30% increase in cell area compared to conventional 6T-SRAM cell.

The read delays can be too long if the bit-lines are required to be fully discharged, this would degrade the system speed and cost high power consumption. A sense amplifier is used to detect a small differential on the bit-lines and convert it to a full rail output, thereby increases the system speed and reduces power consumption. A minimum bit-line differential voltage is required, higher than the offset voltage of the sense amplifier, to enable a reliable read operation. However the offset voltages are getting worse due to large increase in variability. The next chapter presents novel designs to reduce offset voltage dependent read delays for conventional 6T-SRAM design when subjected to large statistical variability.

# Chapter 5

# 5. Sense-amplifier offset voltage mitigation techniques

We examined novel SRAM cell designs in the preceding chapter to provide noise tolerant SRAM read/write operations. In this chapter, we now investigate the impact of statistical variations on the SRAM sense amplifier and possible measures to counter its offset voltage. As described earlier, SRAM cache is probably one of the most vulnerable and valuable resources on a VLSI chip that requires handcrafted design so that it is very robust against device variations. SRAM now accounts over 70% of the total chip area [10], and has a substantial impact on the system speed and total power consumption. Sense amplifier is a critical component in SRAM design that is used to amplify a small differential signal developed between the bit-lines during a read operation. A good sense amplifier will improve system speed and reduce the power consumption during the bit-line discharge. For reliable data sensing, the sense amplifier is triggered only after the value of differential voltage developed at the bit-lines has exceeded its offset voltage [12]. Minimizing the voltage swing on highly capacitive bit-lines is considered as the key to lower the power dissipation of SRAM read operation. However the minimum voltage swing is limited by sense amplifier offset voltage [69]. Similarly the maximum SRAM speed is limited by a weakest bit-cell and the input offset voltage of an worst case sense amplifier [70] since the delay margins are added considering the worst cases.

Offset voltage arises from the mismatch between otherwise identical transistors in a sense amplifier. Devices show deviation in their nominal behaviour due to geometrical or statistical variations that makes a sense amplifier asymmetric [12]. Systematic components of variability can be minimized through a careful layout design [74]. Statistical variability arising from the discreetness of the charge and matter is a major limitation to device scaling and has adverse effects on SRAM design [4]. Different sources of statistical variability include random

discrete dopant fluctuation, line edge roughness, interface roughness, oxide thickness variations, and high k-dielectric morphology, and these sources can cause neighbouring transistors in a sense amplifier to behave quite differently even if they have the same geometry and dimensions in design, resulting in an ever increased offset voltage. Due to its significant impact on the total SRAM area, speed, yield, and power, increasing offset of the sense amplifier now requires special attention. According to ITRS 2007, embedded memories face a clear challenge of the amplifier sense margins in SRAM design [71].

In this chapter, two novel digital methods are presented to reduce the offset voltage dependent SRAM read delay. First proposed method uses a discharge assist circuit for a faster development of the required differential voltage to speed up the read operation. Depending on the asymmetry of the sense amplifier, discharge assist circuit creates an additional discharge path on a bit-line to reinforce the bit-line discharge. There is no performance overhead since this method doesn't add any corrective elements to the sense amplifier structure. Moreover, the energy overhead due to simultaneous discharge (by standard 6T-SRAM cell read and assist circuit) is compensated by lower discharge delays. The proposed discharge assist design results in a 20% reduction in the read energy and a 38% reduction in the sense area over conventionally sized sense amplifier design.

The second method is to add a pre-charge select circuit that chooses an appropriate supply voltage for the bit-line pre-charge that minimizes the discharge differential required for a reliable sense operation. Monte Carlo simulations indicate a 37% reduction in the effective offset voltage using a $1\sigma_{offset}$ calibration for the proposed design when subjected to statistical variability. The kick size can be made of $3\sigma_{offset}$ that can reduce the effective offset voltage by $3\sigma_{offset}$ for the worst case sense amplifiers whose offset voltages lie in the range $3\sigma_{offset}$ - $6\sigma_{offset}$. The proposed design results in a 42% reduction in the read energy consumption and a 15% reduction in the sense area as compared to a conventional sized sense amplifier design.

The chapter presents some background to SRAM sense amplifier, impact of variability on the SRAM read delays, the proposed offset mitigation methods, and a discussion on the simulation results. The details of previously proposed techniques to mitigate SRAM sense amplifier are given in Chapter 2.

## 5.1 Background to SRAM sense operation

Figure 5.1 shows the circuit schematic of a conventional 6T-SRAM cell. It consists of a cross-coupled inverter pair for data storage and two access transistors to control cell read and write operations. Data is loaded on the bit-lines (BL, BLB) during a write operation and the world select line, WS, is held high to write new data in the bit cell. For a read operation, the bit-lines are first pre-charged to supply voltage, VDD. The word select line, WS, is turned high to allow the bit-line discharge. Considering node V1=0, the bit-line BL gets discharged through driver transistor M6. Due to low cell current and large capacitive bit-lines, the discharge time can be very long, and it would degrade SRAM speed and cost high energy consumption. A sense amplifier is used to detect small differential signal developed at the bit-lines during the read operation and convert it to full rail output. This result in a high speed and low power read operation.

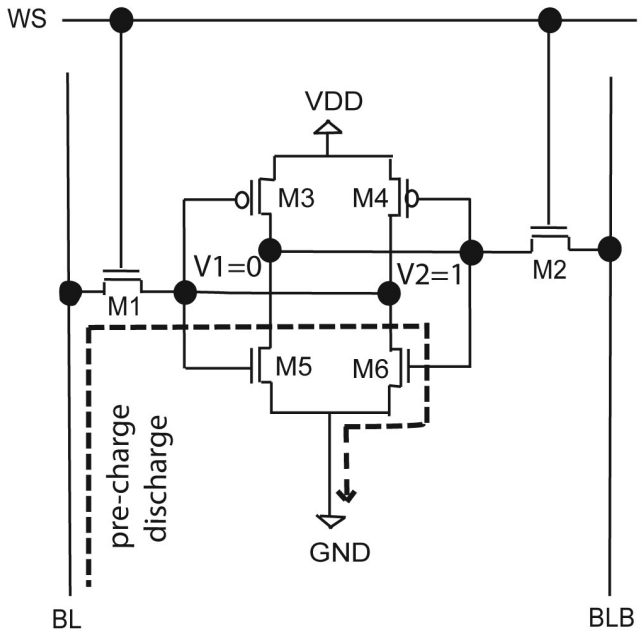
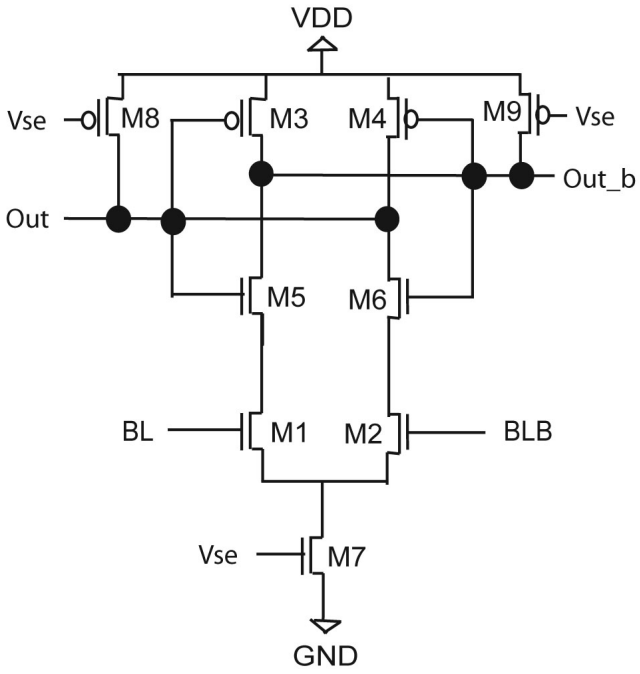


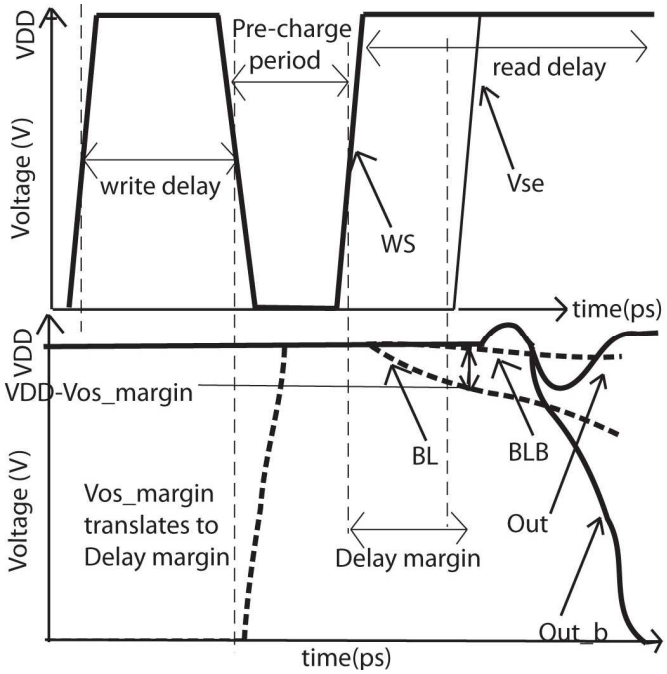

**Figure 5.1: Circuit schematic of a conventional 6T-SRAM cell.**

Figure 5.2(a) shows circuit schematic of a current mode sense amplifier [91]. It consists of two differential input transistors (M1, M2) serially connected to a latch circuit (M3-M6), a clocking transistor (M7), and two pre-charging transistors (M8-M9). A current difference is created between the input differential transistors (M1, M2) due to a differential input voltage. This difference is converted to a large voltage difference by the latch circuit (M3-M6) when the clocking transistor (M7) is turned on [12]. Short current that flows during switching of the cross-coupled inverter pair automatically stops when the sense amplifier outputs settle.

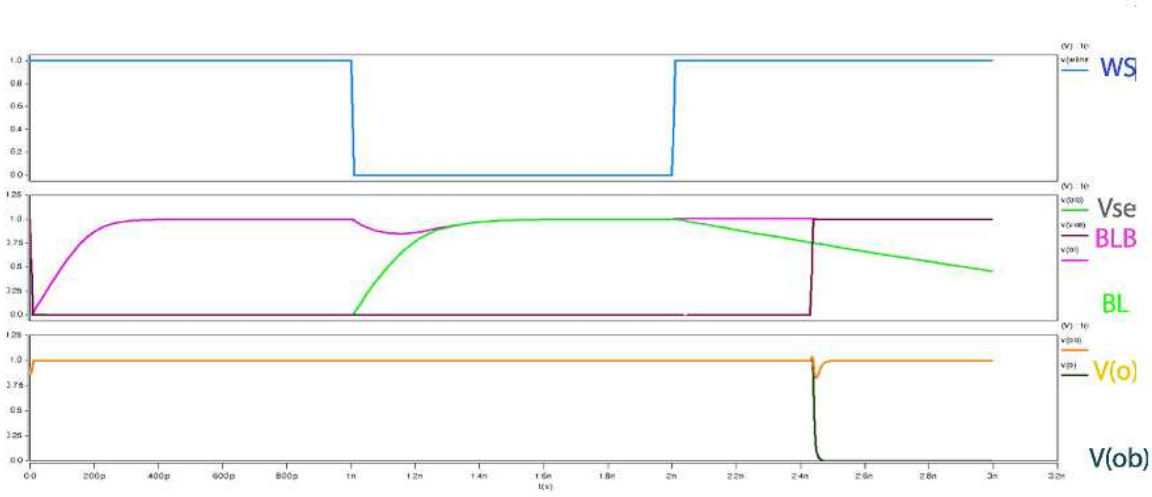
Therefore the circuit doesn't dissipate static power during the read operation [91]. Moreover the current flow itself is small because the latch circuit doesn't drive highly capacitive bit-lines (BL, BLB) directly.




(a) Circuit diagram of conventional current mode sense amplifier




(b) Timing diagram of conventional current mode sense amplifier operation




(c) HSPICE timing simulation of a current mode sense amplifier operation.

**Figure 5.2: Current mode sense amplifier (a) circuit schematic (b) timing diagram (c) HSPICE simulation.**

Figure 5.2(b) shows the timing diagram simulating the transient behaviour of the sense amplifier. The bit-lines are pre-charged and equalized to VDD before the read operation commences. The word select line, WS, is held high to allow the bit-line discharge during the read operation. Assuming the bit-line, BL, is connected to node V1 of the read bit cell that holds '0'. Therefore the bit-line, BL, gets discharged through a path (BL-M1-M6) terminating at SRAM bit cell transistor M6 as shown in Figure 5.1. Sense amplifier is in a sleeping state since the clocking transistor M7 is off and the outputs (Out, Out_b) are held at VDD by the sensor amplifier pre-charge transistors (M8-M9). No major current flows during this period except for the leakage currents. A delay margin is set between the bit-line discharge and the start of sensing operation for a reliable SRAM read operation. This delay depends upon the bit cell discharge current, bit-line BL capacitance, required bit-line discharge level, and the sense amplifier timing margin [10]. We refer the required bit-line discharge level as offset margin Vos_margin since it depends on the offset voltage. A higher offset margin ($V_{os\_margin} \geq n\sigma_{offset}$, where *n* is a multiple of standard deviation of the offset voltage) would be selected for high reliability that would result in longer discharge delay and higher power consumption.

The sensing operation starts when the control signal $V_{se}$ is set high that turns on the clocking transistor M7. A small current starts to flow through the two branches (M7-M1-M5-

M3 and M7-M2-M6-M4) of the sense amplifier that begin to discharge the output nodes (Out, Out_b). The branch current is determined by the discharge voltage drop developed at its respective bit-line connected to the input transistor. Since one of the bit-line gets discharged (in this example, BL), a differential voltage drop at the input transistors results in an imbalanced current flow in two branches of the sense amplifier. This difference is then amplified by the latch circuit (M3-M6) and converted to a full rail output voltage. Ideally a branch (M7-M2-M6-M4) connected to a higher voltage bit-line (BLB in this example) current would discharge the corresponding output node more quickly as shown in Figure 5.2(b). However the two branches can have imbalanced current flow due to variability even when the bit-lines (BL, BLB) have the same voltage. This mismatch in matched devices of the sense amplifier results in its offset voltage. Therefore an input differential voltage higher than the offset voltage of sense amplifier is required for reliable sensing. Figure 5.2(c) shows an actual timing diagram of a 45 nm sense amplifier operation using HSPICE.

## 5.2 Impact of statistical variations on SRAM read delay

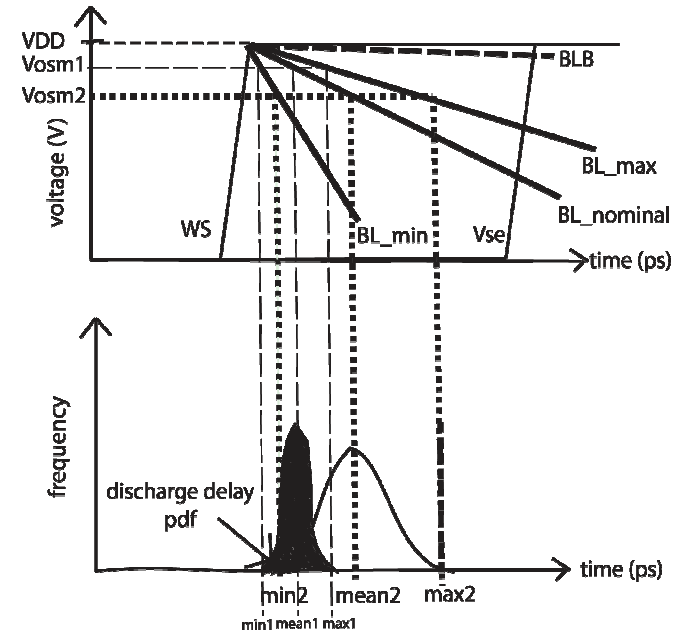
Figure 5.3 illustrates the combined impact of cell current variation and the offset voltage variation of the sense amplifier due to variability on the read delay of the conventional 6T-SRAM design. Figure 5.3(a) shows the timing diagram of the read operation and corresponding discharge delay probability distribution function (PDF). When the word select line, WS, is held high to start the read operation, the bit-line BL gets discharged to a given offset voltage margin $V_{osm}$ depending upon the discharge current of the bit cell, and the bit-line capacitance. Both the cell current and the offset voltage variations degrade the read delays, therefore, impact of variability on both conventional 6T-SRAM cell and sense amplifier is taken together to estimate the worst case delays.

The variability in matched devices of the sense amplifier require a certain amount of offset voltage margin, $V_{osm}$, to be met during the read operation. Meanwhile, variability in the conventional 6T-SRAM cell itself will result in large cell current variations. A cell with high cell current will quickly discharge the bit-lines (e.g. BL_min in Figure 5.3(a)), whereas a

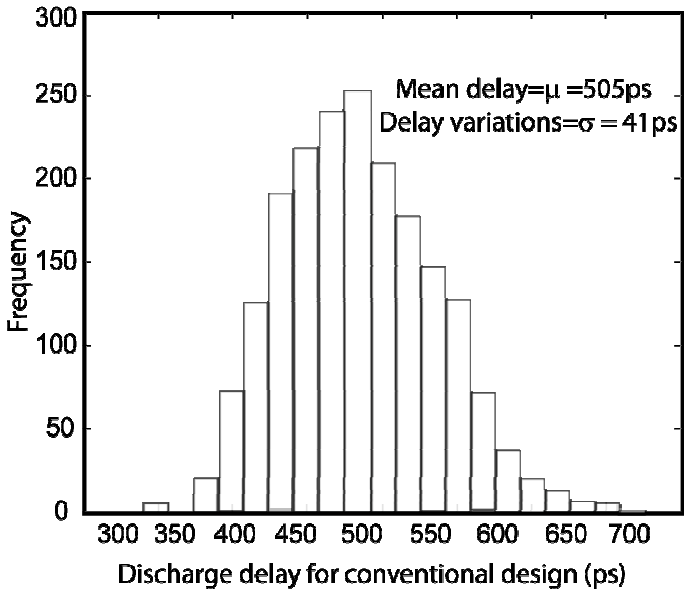
weak cell will take longer to establish desired the differential voltage $V_{osm}$. This will result in a discharge delay PDF as illustrated in Figure 5.3(a), a sense amplifier with small offset margin ($V_{osm1}$) results in a small mean discharge delay with small variations as compared to a sense amplifier with large offset margin ($V_{osm2}$).

Sensing is delayed to cover worst-case discharge current for the weakest SRAM cell (corresponds to discharge of the bit-line BL_max). The increased variability in nano-scaled technologies will result in high offset voltage variations and large cell current variations. Large delay margins will, therefore, be necessary for the reliable sense operation that would incur a high power and performance overhead. Figure 5.3(b) shows discharge delay PDF on a bit-line ($C_{BL,BLB} = 48$ fF) connected to a 45 nm technology generation conventional 256-SRAM cells column simulated for an offset voltage margin $V_{osm}$=300 mV ($n\sigma_{offset}$), when statistical variability including RDD, LER, and PGG was inserted in SRAM cells ($W_{M1,2,3,4} = L, W_{M5,6} = 1.5L$).

Figure 5.4 shows the relationship of the discharge delays with offset voltage (margin) and SRAM cell current, which subjected to statistical variability. The mean delays correspond to the discharge delays for a given offset margin of the sense amplifiers, and the $6\sigma$ of the delay is taken to consider extreme case cell current variations. Figure 5.4(a) shows an increase in the mean delays with the increase of required differential voltage margin. The conventional 6T-SRAM cell current variations increase when large offset margin is required as seen by increasing $\sigma$ of the discharge delays. However the sense amplifier offset margin has a higher impact on the total discharge delay as shown in Figure 5.4(b). Over 60% of the discharge delay is attributed to offset voltage margin and less than 40% is due to SRAM cell current variations, considering the $6\sigma$ delay variations. The higher the sense amplifier offset voltage, higher is the impact on the discharge delays by the sense amplifier.

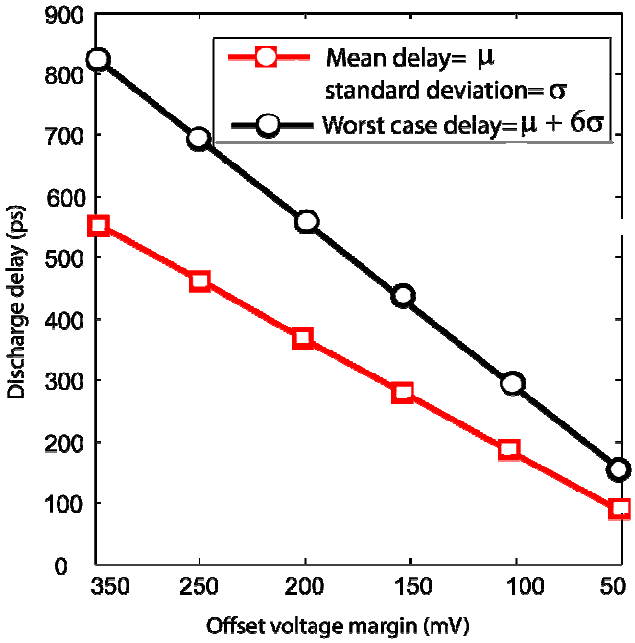



(a)

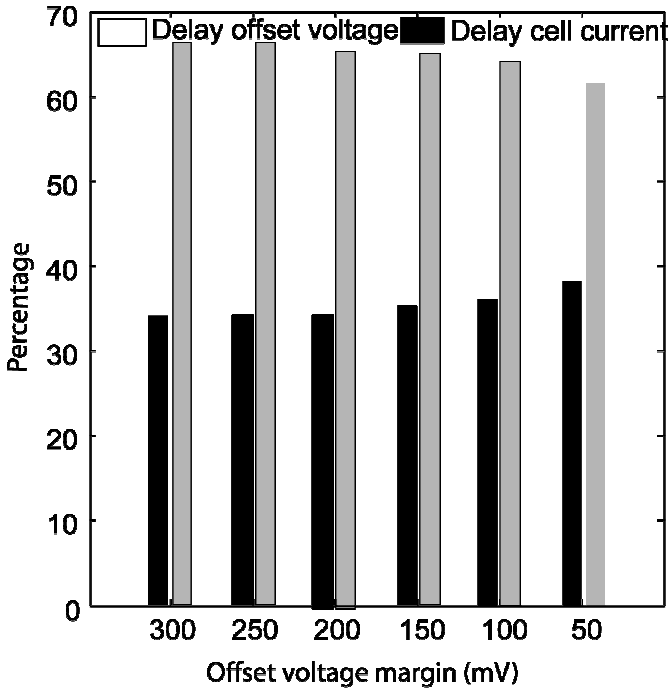



(b)

**Figure 5.3: The impact of variability on read delay (a) Timing diagram and (b) simulation result.**




(a)




(b)

**Figure 5.4: Discharge delay relation with offset voltage and cell current (a) discharge delay variations (b) percentage contributions of the offset voltage and cell currents to total discharge delay.**

## 5.3 Proposed discharge assist design

A straight forward method to achieve a relatively constant offset voltage, across different generations, is the use of traditional sizing of the amplifier transistors. It avoids the delay degradations that arises due to device scaling in the lower technologies [13]. A number of sizing based techniques have been presented in the past to mitigate the offset voltage [72, 73]. However the size of the sense circuit doesn't scale with technology as rapidly as it does for SRAM cells, that increases the sense circuit overhead [70]. It represents a major trade off between the size of the sense amplifier and an acceptable offset voltage [61]. Large sized sense amplifiers consume a large dynamic energy that makes a significant portion of the total energy consumption. One method is to add corrective elements to a conventional sense amplifier and use digital trimming after fabrication to recover worst case amplifiers [74, 75]. However the addition of corrective elements to the basic structure of the sense amplifier increases its delay and power consumption. Sense amplifier redundancy can be used to select best case amplifier during run time, however it its increases run time cost [61]. A number of reference voltages, $V_{ref}$, can be generated and a particular voltage can be selected that minimizes the offset voltage [14]. However the overhead is the generation of multiple precise voltages and a number of storage devices to save configuration settings.

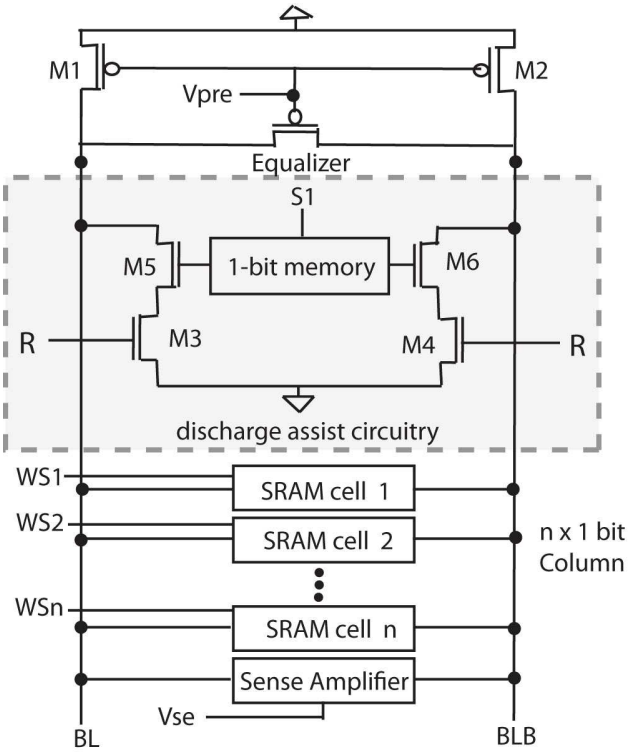
### 5.3.1 Proposed discharge assist circuit

The proposed design uses asymmetry information of the sense amplifier, generated during the post-silicon calibration [74], to assist the bit-line discharge process during the SRAM read operation. Figure 5.5(a) shows the circuit schematic of the proposed design. A 1-bit storage node S1 (flip-flop or latch) keeps configuration settings for each sense amplifier to allow intelligent assisted bit-line discharge. The discharge assist transistors ((M3 and M4 in Figure 5.5(a))) are turned on when the read operation starts with the read signal, R, held high. Depending upon the values in the storage nodes that correspond to asymmetry of the sense amplifier, a discharge control transistor (M5 or M6 in Figure 5.5(a)) is turned on to enable assisted discharge on the bit-line connected to a faster branch of the sense amplifier. Proposed assisted discharge method therefore improves the discharge process to quickly overcome a voltage or current imbalance (offset) in the sense amplifier to allow a reliable sense operation at a reduced discharge delay.

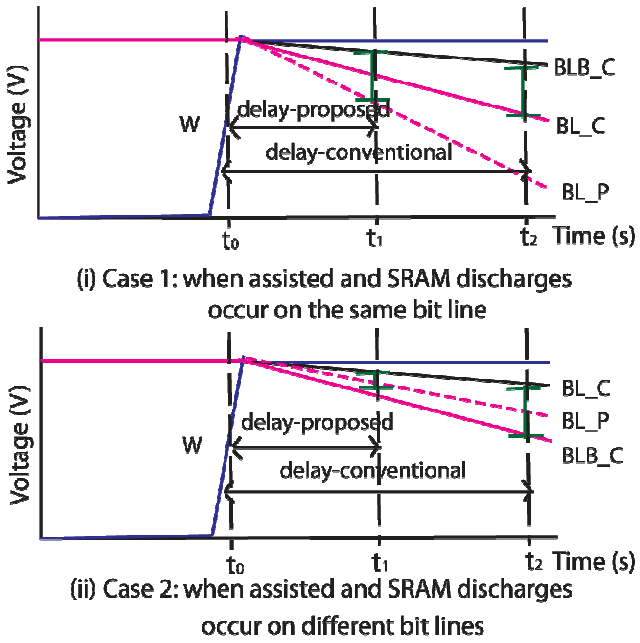
Figure 5.5(b) illustrates the timing operation for the proposed discharge assist method. Assuming the faster branch of the sense amplifier is connected to the bit-line BL and the slower branch is connected to the bit-line, BLB. Appropriate values are loaded in the storage devices during calibration phase that configure the proposed circuit to turn on the discharge control transistor M5 (Figure 5.5(a)) to speedup the discharge of the bit-line BL. For a differential read operation, the bit-line discharge can occur on both the bit-lines (BL, BLB) depending upon the stored value in the SRAM cell being read. There are two possible cases for the discharge, one when the SRAM cell and the assisted discharge occur on the same bit-line (case 1), or when the cell and assisted discharges occur on different bit-lines (case 2) as shown in Figure 5.5(b). When the assisted discharge and the SRAM cell discharge (read 0) occur on the same bit-line (e.g. BL), in such case the bit-line for the proposed design, BL_P, establishes required bit-line differential more quickly at time $t_1$ as compared to a conventional design that takes time $t_2$ (where $t_1 < t_2$) to develop same differential voltage as shown in Figure 5.5(b-i).

The other case can be when assisted bit-line discharge occurs on a bit-line (e.g. BL) and the discharge by the SRAM cell (read 1) occurs on a different bit-line (e.g. BLB) as shown in Figure 5.5(b-ii). In such a case, the proposed design requires a significantly longer discharge delay, compared to the conventional sensing, to develop the same bit-line differential since both the bit-lines (BL_P, BLB_P) are simultaneously discharged by the SRAM cell read and assisted discharge circuit. However, we don't need to wait for the same differential voltage to be developed on the bit-lines, as in the case of a conventional design, since the sense amplifier is skewed (faster) on the side connected to the bit-line BL_P. Therefore we can trigger the sense amplifier at $t_1$ (as shown in Figure 5.5(b-i) and still sense correct output as long as assisted discharge on the bit-line BL_P is lower than the SRAM cell discharge (read 1). Discharge speed is limited by the required voltage drop on the bit-line BL for correct sense operation on read 0, since the sense amplifier can be fired as soon as the bit-line discharge starts in case 2. Figure 5.5(c) shows the result of HSPICE simulation of a 45 nm randomized sense amplifier for the case 1 with the standard 45nm 256x1 6T-SRAM cell array. Note the bit-line differential developed in case 1 is larger than the bit-line differential voltage developed by a conventional design simulated in Figure 5.2(c), considering the same discharge period. Figure 5.5(d) shows HSPICE simulation of a 45 nm randomized instance of the sense amplifier for the case 2. The bit-line voltage differential is lower than the

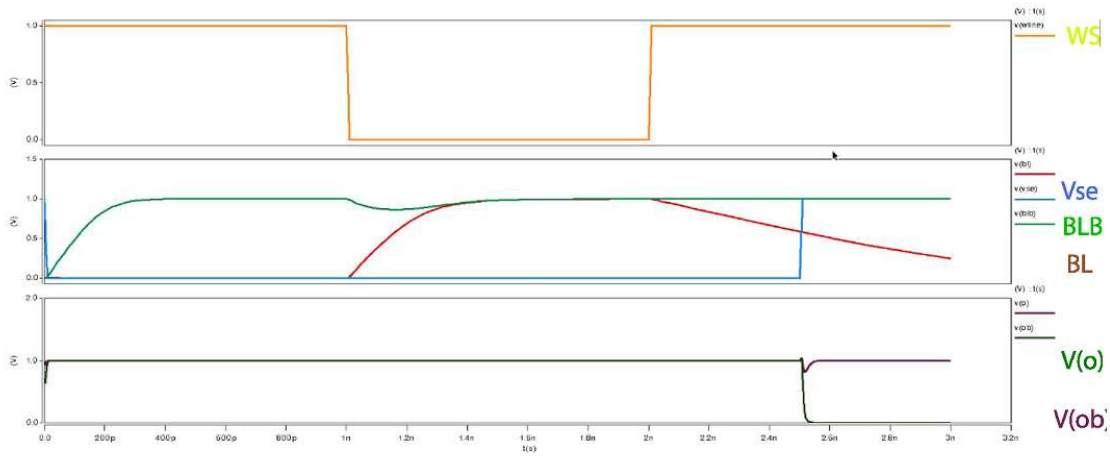
conventional design, for the same discharge period, however it still yields a correct output due to asymmetry of the sense amplifier arising from statistical variations.



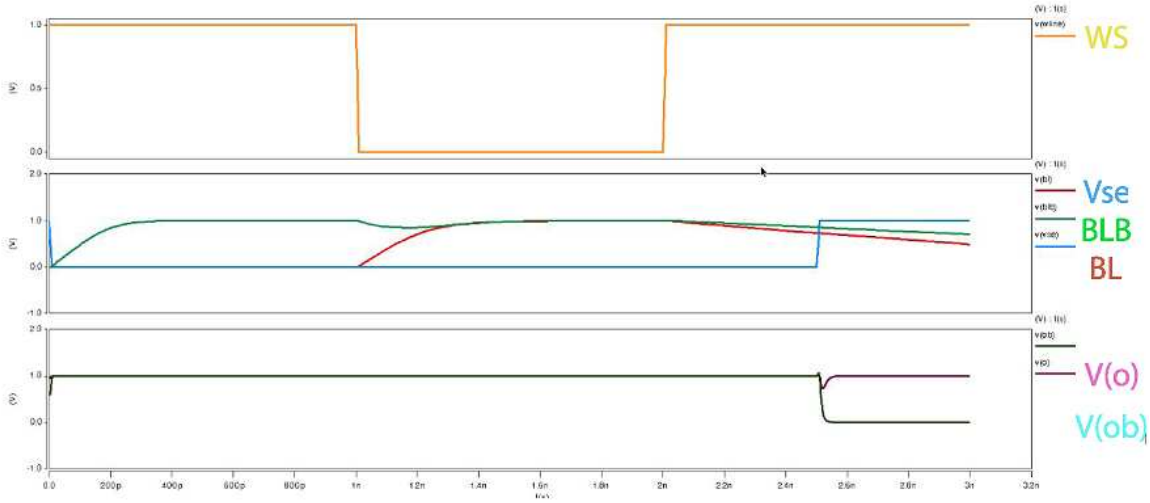

(a) Proposed dis-charge assist circuit diagram




(b) Timing operation for the proposed discharge assist design




(c) HSPICE simulation of the proposed discharge assist design - case 1




(d) HSPICE simulation of the proposed discharge assist design – case 2

**Figure 5.5: Proposed discharge assist circuit (a) Circuit schematic and (b) timing diagram (c) HSPICE simulation assisted discharge case 1 (d) HSPICE simulation assisted discharge case 2.**

Excessive assisted discharge on a bit-line can cause read failures when it is higher than the SRAM cell discharge. Therefore we employ minimum sized high threshold (weak) discharge assist and control transistors to avoid unwanted assisted discharge, as the transistor pairs M3, M5 or M4, M6 (Figure 5.5 (a)) form an additional discharge path. An actual 6T-SRAM cell is designed to achieve high speed discharge with an acceptable read margins by keeping a high cell ratio (CR), β, where the cell ratio β represents the width ratio of the driver and pass

transistors. Since minimum sized transistors are used for the assist/control transistors, weak discharge is ensured by the proposed discharged assist circuit. For an ideal SRAM cell without variation, the maximum delay improvement that can be achieved by the assist discharge approach would be 50% when cell and assist currents are the same.

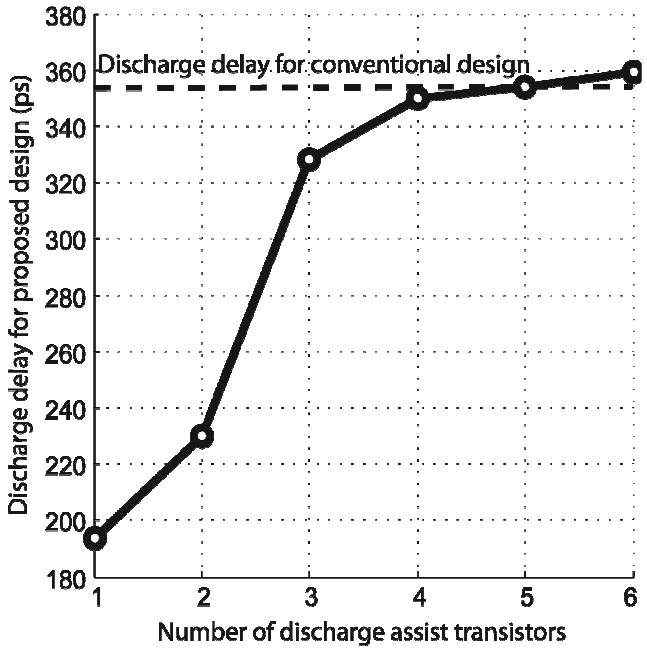



**Figure 5.6: Impact of discharge assist transistors on read delay.**

High process variations, especially due to statistical sources of variability, cause large threshold variations in scaled technologies. When assisted and SRAM cell discharge occur on different bit-lines, large variability can weaken cell drive currents during the read operation to be lower than the assisted discharge current that may lead to read failures. One method is to use long channel devices to weaken assist current that increases read stability. An alternate method is to put multiple assist transistors in series to minimize the unwanted discharge. We employ multiple discharge assist transistors in this work due to availability of the minimum length device models for our simulations. Figure 5.6 shows the impact of multiple assist transistors (1-6) on the discharge delay when required a 200 mV differential on the bit-lines (BL, BLB). Using 1-4 assist transistors improves the read delay by boosting the discharge process. Maximum delay improvement of a 45% (354 ps vs. 194 ps) is observed for one

control/assist transistors, however it may increase read failures due to large threshold variations. Having 5 or more assist transistors nullifies the proposed design, as assist discharge current is too small to overcome the delay increases due to the additional bit-line loading. Using 2-4 assist transistors (in series) provides a trade-off between the maximum delay improvement and error rate (stability).

### 5.3.2 Statistical variability simulation results

Figure 5.7 shows discharge current distributions obtained from 14,000 simulations of a 45 nm 256x1 conventional 6T-SRAM column array. Statistical variability was inserted in both (conventional and the proposed) designs, and the discharge currents were calculated after some fixed discharge period. Figure 5.7(a),(b),(c) show discharge current for the proposed design in case 2 (SRAM cell and assist discharge occur on opposite bit-lines) when using one (AT=1), two (AT=2), three (AT=3) assist transistors, respectively. A current overlap with conventional design in Figure 5.7(a) indicates probable read failures at the tails of these distributions when assist discharge may become higher than conventional SRAM cell discharge. However increasing the number of assist transistor to two (AT=2) in Figure 5.7(b) and three (AT=3) in Figure 5.7(c) removes this overlap even at the tails of the distributions. This removal of the tails indicates that although increasing the number of assist transistors decreases assist discharge current, it can increase the reliability of correct sense operation. However this decreases the total discharge current in case 1 (assist and SRAM cell discharge occur on the same bit-line). The decreasing gap between the current distributions, of the proposed and conventional designs, with the increasing number of assist transistors is shown in Figure 5.7(d). Therefore, longer discharge delays will be required to achieve the required bit-line differential voltage due to reduced total discharge currents when using large number of assist transistors.

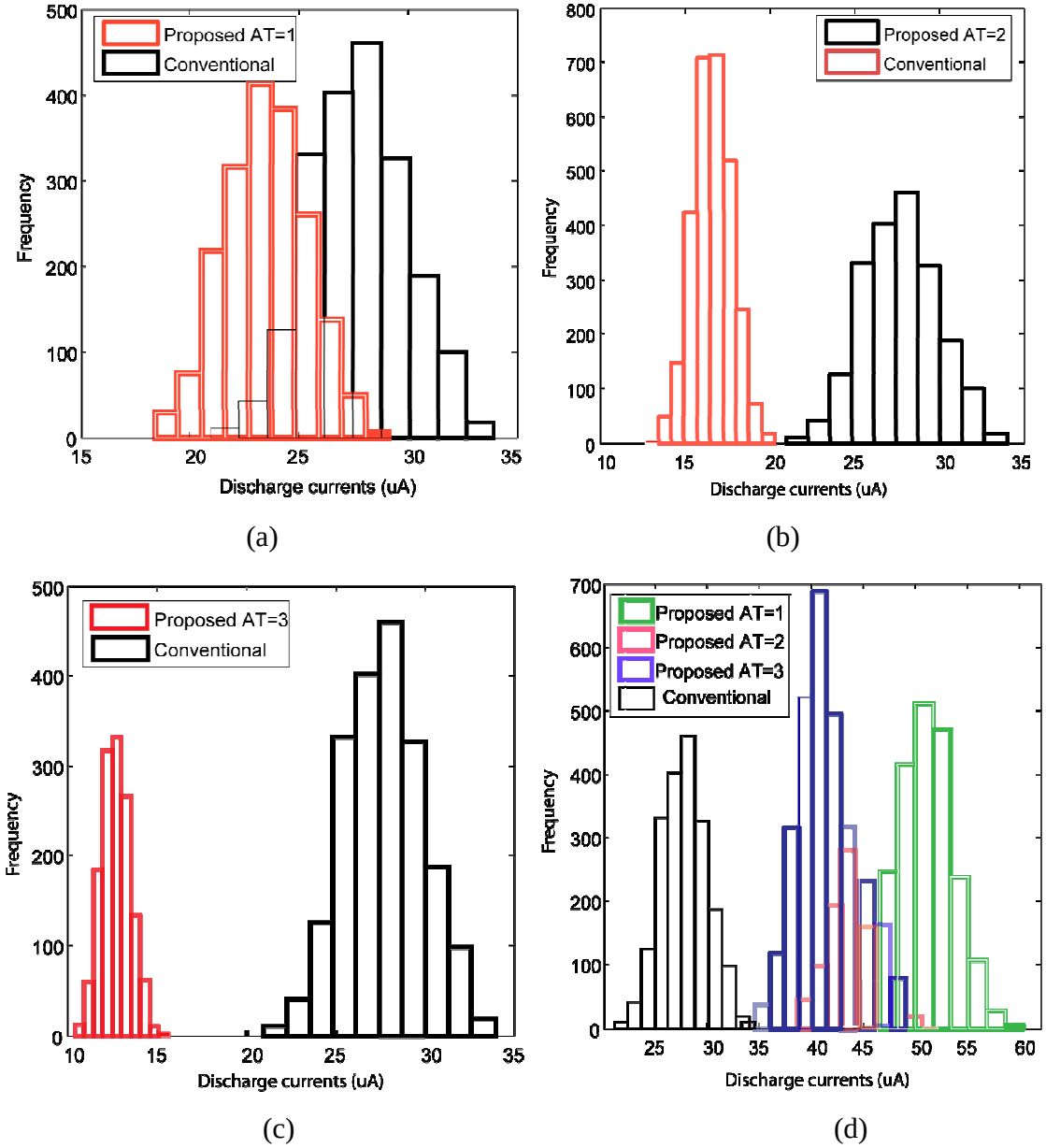



**Figure 5.7: Discharge current distributions (a) case 2 AT=1 (b) case 2 AT=2 (c) case 2 AT=3 (d) case 1.**

We designed a 45 nm 256x1 bit conventional 6T-SRAM array (column, $C_{BL,BLB} = 48\,fF$) in order to quantify effectiveness of the proposed design in reducing read failures by assisting discharge process during the SRAM read operation, shown in Figure 5.8. Over 100,000 Monte Carlo simulations were performed by inserting statistical variability (RDD, LER, and PGG) in both the conventional 6T-SRAM cells and sense amplifier circuit to analyse error rate reductions at different discharge delays using multiple number of assist transistors. This allows a more comprehensive SRAM read delay analysis under variability when taking into account both discharge current variation of the SRAM cells and offset voltage variations of

the sense amplifier. For 100,000 Monte Carlo simulations, an error rate of 0% is required for reliable SRAM read sense operation. Simulations were performed using one control transistor in series with one, two, and three assist transistors (AT) in our discharge assist circuit for comparative study. The results of these simulations are shown in the error rate plot in Figure 5.9 for both the proposed and conventional design. The proposed design provides significant reduction in the error rate for low discharge delays.

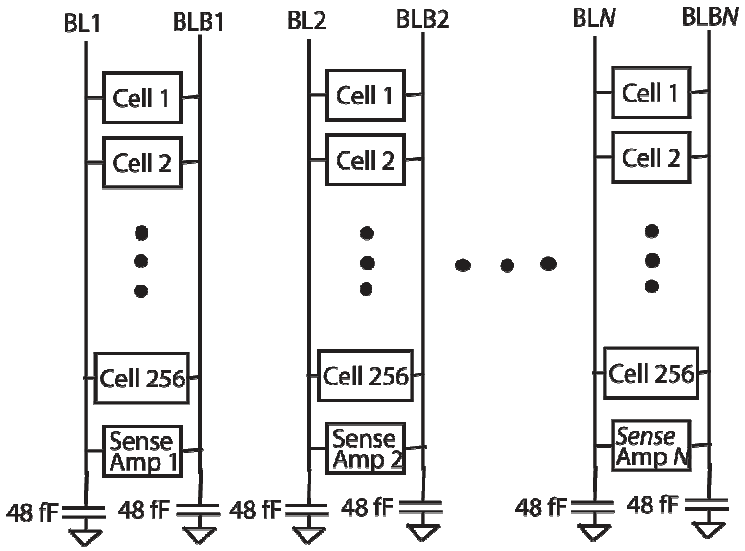



**Figure 5.8: 256x*N* SRAM array setup for statistical variability simulation.**

Increasing the number of assist transistors degrades error rate performance for low discharge delays since the assisted discharge current reduces that offsets effectiveness of the proposed design. However, for a reliable SRAM sense operation, it's important to look at the lowest delay time required for a 0% error rate. For the conventional design, a 325 ps delay time is required in order to guarantee the successful read operation. For the assist discharge approach with only one assist transistor (AT=1), it can achieve the highest error rate reduction in low discharge delays compared with the conventional counterpart. However, it can not achieve error rate of 0% due to the huge conventional 6T-SRAM cell discharge current variation and the relatively large assist discharge current. Using two assist transistors (AT=2) provides both the high robustness and speed improvements, as evident by the low error rate at both high and low discharge delays (shown in Figure 5.9) for AT=2. It can reduce the discharge delay time to 200 ps for a read error rate of 0%, which represents a 38% improvement as compared to a conventional design. This improvement reduces to 23% (325

ps vs. 250 ps) when using three assist transistors (AT=3) due to weak assisted discharge currents.

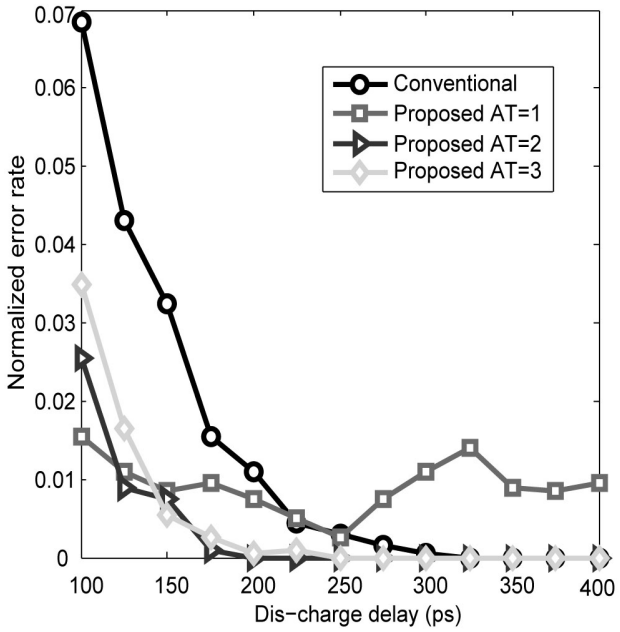



**Figure 5.9: Error rate comparison at different discharge delays.**

### 5.3.3 Energy and area comparisons

For a comparative study with the conventional sizing technique, we carried out an energy and area analysis, on a 45nm 256x1 bit conventional 6T-SRAM SRAM column array for the same performance requirement ($\tau_{sense-amp} + \tau_{discharge}$ =124 ps). For the conventional design, traditional sense amplifier sizing technique [72] was applied to achieve a differential swing of $6\sigma_{offset} = 53\,\text{mV}$ [14] to meet the given delay requirement. The proposed circuit was designed using one assist and one control transistors. The sense amplifier was sized smaller for the proposed design to achieve a differential of $6\sigma_{offset} = 102\,\text{mV}$, in order to meet given performance metric (124ps). Note the fact that the differential required for the given discharge delay is halved when cell and assist discharge currents are equal,

$$\tau = \frac{VC}{I_{cell} + I_{assist}} \qquad \text{Equ. 5. 1}$$

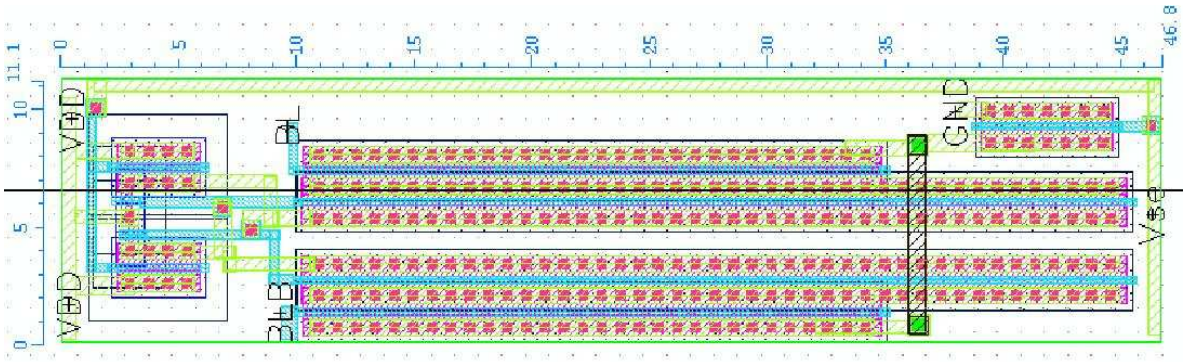
Large sizing of the sense amplifier NMOS transistors was performed for a low offset margin that resulted in a high energy and area overhead for the conventional design. For the proposed discharge assist design, small size transistors have been employed in the sense amplifier circuit since it can accommodate relatively large offset voltage (102 mV vs. 53 mV), which increases the energy dissipation by 77% (3.3 fJ vs. 5.8 3fJ) during the bit-line discharge process, however, there was a 62% (6.65 fJ vs. 2.53 fJ) reduction in energy consumption during the sensing period (20 ps) and a total of 16% (9.95 fJ vs. 8.35 fJ*)* reduction in the total energy consumption over a traditional design. Sense operation took 67% of the total energy consumption for a traditional design as compared to 30% for the proposed design. The total energy reductions improved by 20% (9.95 fJ vs. 7.96 fJ) for a two assist and one control transistor configuration due to reduced bit-line discharge current. However the sense amplifier was sized larger to achieve low offset margin ($6\sigma_{offset} = 88\,mV$) that met the delay requirements (124 ps), however this in turn led to an increased area overhead.

**Table 5. 1: Energy comparison**

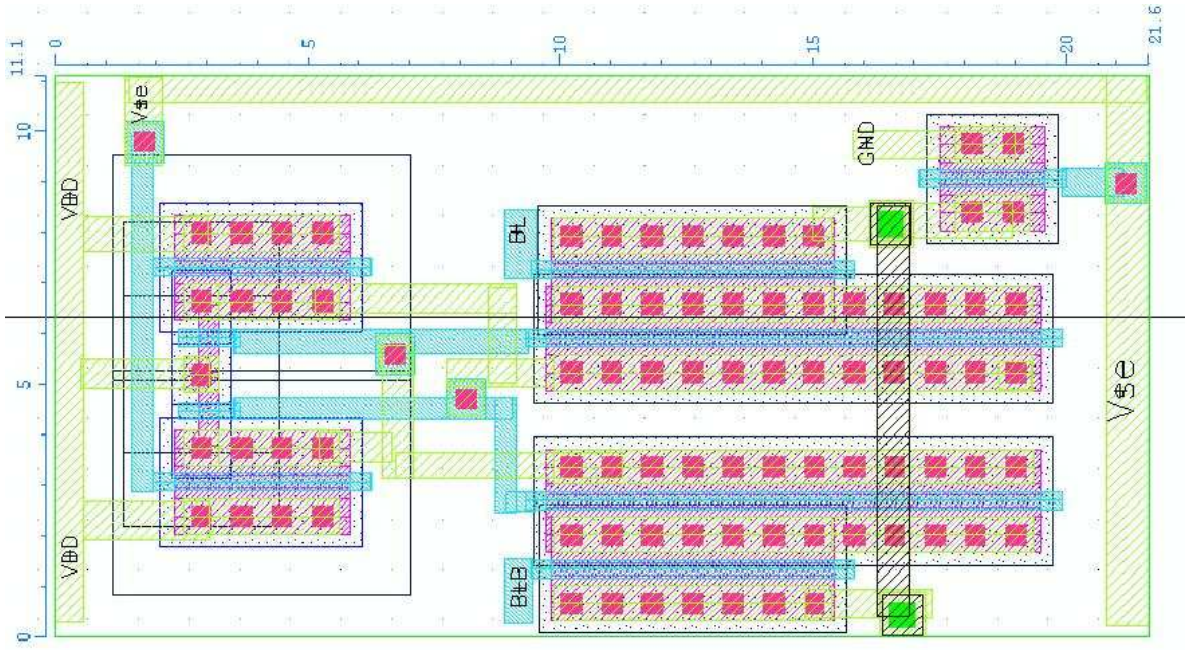
| | Discharge energy (fJ) | Sense energy (fJ) | Total energy (fJ) |
|---|---|---|---|
| Conventional sense amplifier ($6\sigma_{offset} = 53\,mV$) | 3.3 | 6.65 | 9.95 |
| Proposed sense amplifier ($6\sigma_{offset} = 102\,mV$) | 5.83 (77%↑) | 2.53 (62%↓) | 8.35 (16%↓) |

A layout study of the sense amplifier for the conventional design and the proposed design (one assist/control circuit and sense amplifier) has been carried out. It was found that the proposed design requires a 38% ($519\,\mu m^2 vs. 322\,\mu m^2$) less sense area as compared to the conventional design. Figure 5.10 shows the layout for both sense amplifier designs. The area savings reduce to a 27% ($519\,\mu m^2 vs. 378\,\mu m^2$) for one control and two assist transistors configuration due to large sized sense amplifier and additional assist transistors. The proposed method can also be used in addition to the conventional sizing to reduce read delays. The area

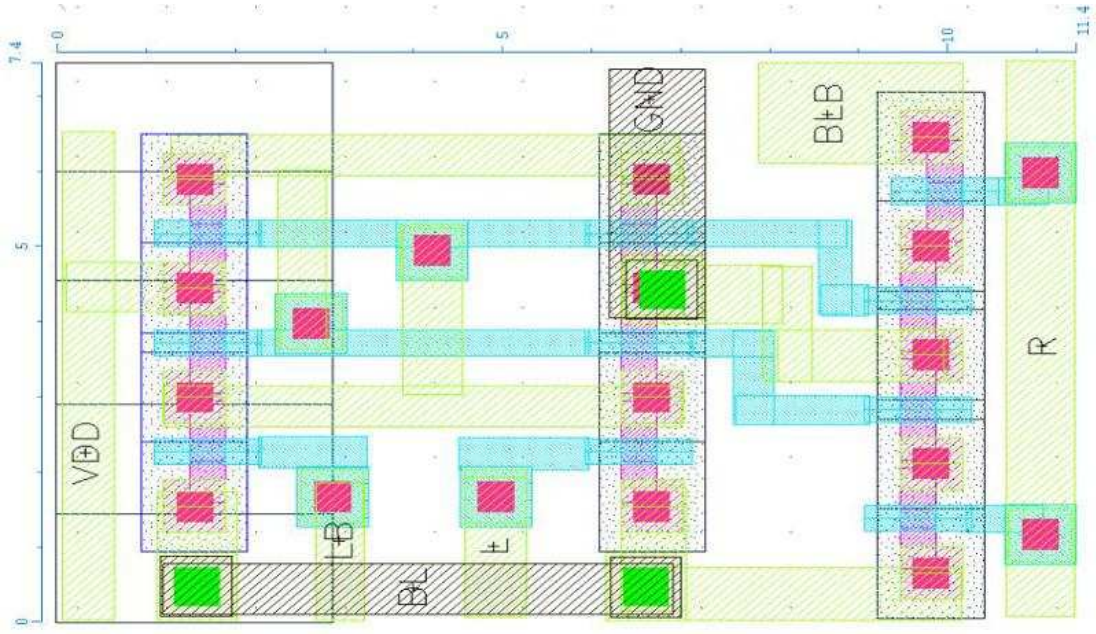
overhead in that case is less than 2% ($322\,\mu m^2$ vs.$16742\,\mu m^2$) for a 256 bit SRAM column array (cell area x word length = $10.9x6x256\,\mu m^2 = 16742\,\mu m^2$).




(a) Conventionally sized sense amplifier layout




(b) Sense amplifier layout for the proposed design




(c) Proposed discharge assist circuit layout

**Figure 5.10: Area comparison (a) conventional sized sense amplifier Area=46.8x11.1 $\mu m^2$ =519.5 $\mu m^2$ (b) sense amplifier for proposed design Area= 21.6x11.1 $\mu m^2$ =240 $\mu m^2$ (c) proposed discharge assist circuit Area=7.4x11.4 $\mu m^2$ = 82 $\mu m^2$.**

## 5.4 Proposed pre-charge select design

Device mismatch in a sense amplifier results in an unbalanced current flow in the two branches of a sense amplifier even when the same input voltage is applied that appears as the sense amplifier offset voltage. The input offset voltage of the sense amplifier refers to the differential voltage that will force the latch circuit (M3-M6 in Figure 5.2(a)) to enter meta-stability, V(Out) = V(Out_b) [12]. We use this fact to provide a lower bit-line pre-charge voltage at an input transistor of a faster branch of the sense amplifier. The pre-charge voltage is selected during the calibration phase that will minimize the differential (offset voltage) required to achieve meta-stable outputs. Since this method doesn't change the total differential voltage required to achieve meta-stable outputs of the sense amplifier, therefore, the intrinsic offset voltage of the sense amplifier remains the same. However the bit-line discharge differential voltage required for meta-stable outputs changes after calibration that we refer as the effective offset voltage.

Figure 5.11(a) shows a SRAM array structure for the proposed design. Each column of the array is provided with a pre-charge select circuit that is calibrated to minimize the required discharge differential for reliable sensing of the corresponding sense amplifier. Only two DC supply voltages (1V, V_pre), where V_pre = VDD - $n\sigma_{offset}$ and n is an integer multiple, are provided for selection of the pre-charge levels in a 2 cycle calibration process. These voltages are selected depending upon the intrinsic offset voltage of the sense amplifier that will minimize the effective offset voltage. A single pre-charge select circuit can be used for multiple columns to minimize area overhead, when a single sense amplifier is shared by N columns (N>1).

During calibration, which is performed at initial system power-on phase, each sense amplifier is calibrated to identify the pre-charge voltages that minimize its effective offset voltage. The calibration starts by applying the same VDD voltage on both the bit-lines and then sensing the output of the sense amplifier. Depending on if the output is zero or one, one branch of the sense amplifier is identified as fast or slow. In the next cycle, the storage nodes are loaded with an appropriate value to apply a low pre-charge voltage, V_pre=VDD - $3\sigma_{offset}$ (n=3), on a faster branch to minimize the current difference between the two branches. If the outputs are flipped this time, it shows the offset lies in the range 0 - $3\sigma_{offset}$ and no correction is therefore needed. The storage nodes are loaded with the default values that select VDD as the pre-charge voltage for both the bit-lines. In the other case, when the outputs don't flip, it indicates a worst instance of the sense amplifier whose offset lies in the range $3\sigma_{offset}$ - $6\sigma_{offset}$ . Therefore the lower pre-charge voltage (VDD - $3\sigma_{offset}$ ) is selected to kick it back in the range 0 - $3\sigma_{offset}$ , effectively recovering nearly all instances of the worst case sense amplifiers.

Figure 5.11(b) shows a detailed implementation of the proposed pre-charge select design. Two supply voltages (1 V, V_pre) are shown here as an example of a 2-step bit line pre-charge voltage calibration for the sense amplifier. A pre-charge select circuit is added to each bit-line pair (BL, BLB) that selects an appropriate voltage during the calibration phase for pre-charging. A 2-bit storage register is provided to store the configuration setting that is derived from the asymmetry information of the sense amplifier. The storage element is pre-set at start

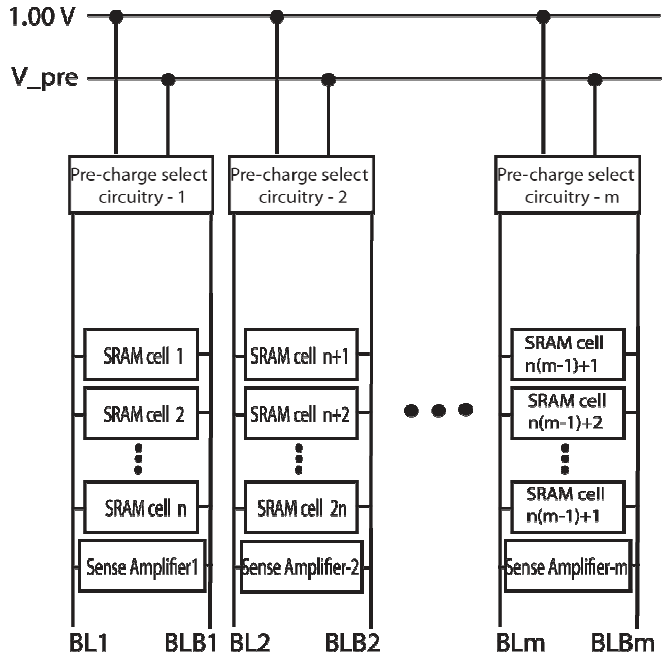
to select VDD for both bit-lines (BL, BLB). At the end of the first calibration cycle, the system can identify a faster branch of the sense amplifier connected to the bit-lines (BL, BLB), and a lower pre-charge voltage is selected for the corresponding bit-line attached to the input transistor of the sense amplifier. Outputs Out and Out_b of the sense amplifier (Figure 5.2(a)) indicate which bit-line should receive a lower pre-charge voltage to minimize the effective offset voltage. Assuming the branch connected to the bit-line BLB (M9-M2-M6) is slower than the other branch connected to the bit-line BL (M9-M1-M5) that causes the output node, of the sense amplifier, Out, to discharge to zero. Therefore a lower pre-charge voltage is selected for the bit-line BL during calibration phase that reduces the discharge differential needed for reliable sensing, that in turn minimizes the effective offset voltage.

Figure 5.11(c) shows timing diagram of the sense amplifier operation for the proposed design. We assume that one branch (M7-M1-M5-M3) of the sense amplifier connected to the bit-line BL is faster than the other branch (M7-M2-M6-M4) that is connected to the bit-line BLB. Therefore a lower pre-charge voltage, $V_{psel}$, is selected for the bit-line BL during calibration phase that will minimize the effective offset voltage. During the pre-charge phase, the bit-line BL is charged to $V_{psel}$ and BLB to VDD=1 V. Sense amplifier is triggered at $t_1$ for the proposed design and at $t_2$ for the conventional design that allows same bit-line voltage differential but different effective offset voltages. Note $t_1 < t_2$ because the proposed design has a lower effective offset voltage (Voffset-P) than conventional design (Voffset-C). Therefore a lower delay margin for the proposed design will result in a high speed and low energy consumption. Figure 5.11(d) shows the timing operation simulated using HSPICE for a randomized 45 nm sense amplifier. A lower pre-charge select voltage results in early start to the sense operation for the proposed design compared to the conventional design, simulated in Figure 5.2(c). The relation between effective and intrinsic offset is given as,

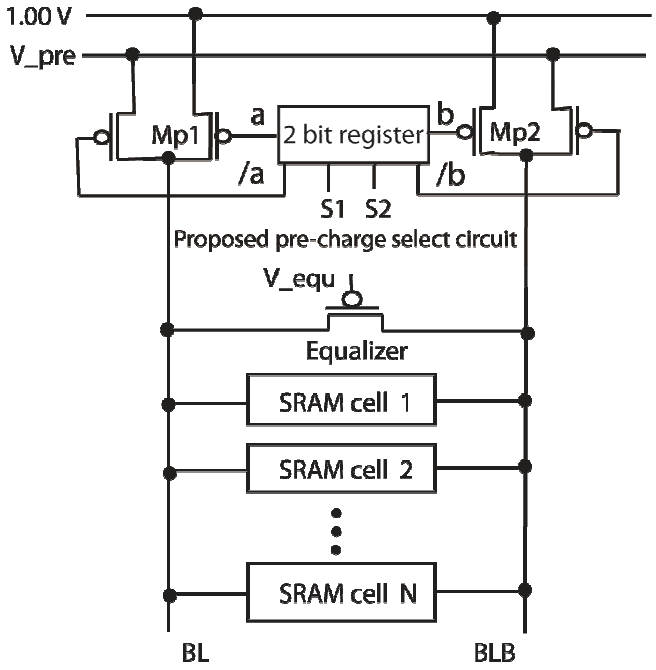
$$V_{offset-p} = (V_{offset-c} + V_{psel}) - VDD \qquad \textbf{Equ. 5. 2}$$

A positive effective offset indicates that the faster branch of the sense amplifier is still faster after compensation, but a lower discharge differential is required for correct operation. For a negative effective offset voltage the opposite will be true, however, the absolute value of the

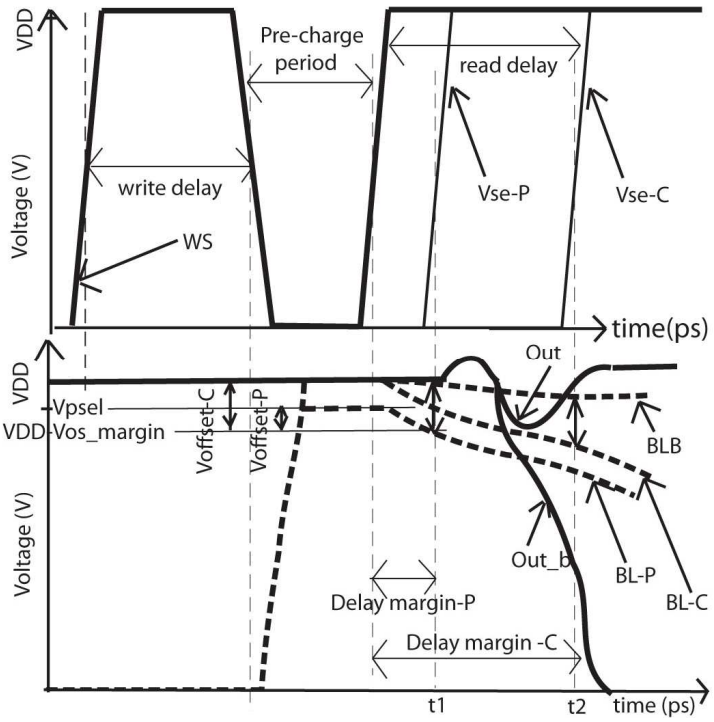
effective offset voltage will be always smaller than the voltage step, in this case, $3\sigma_{offset}$. Since two different pre-charge voltages are applied at the bit-line pair (BL, BLB) in this circuit, the bit-lines may not require an equalization circuit. However an equalization circuit is useful to speed up the pre-charge phase. If we keep a conventional equalization scheme in the design, we avoid any speed and power penalty (worst bit line bias conditions) that occurs when the bit-lines have no equalization. However this reduces the bit-line pre-charge differential voltage due to a voltage division of the pre-charge differential between pre-charge and equalization transistors. We avoid this problem by increasing the step size (decreasing the resolution) to achieve the desired differential voltage at the bit-lines. Lower pre-charge voltages result in a low pre-charge power consumption. Moreover, the static noise margins (SNM) improve since a lower pre-charge voltage produces a small disturbance on the storage node (holding 0) during the SRAM cell read.



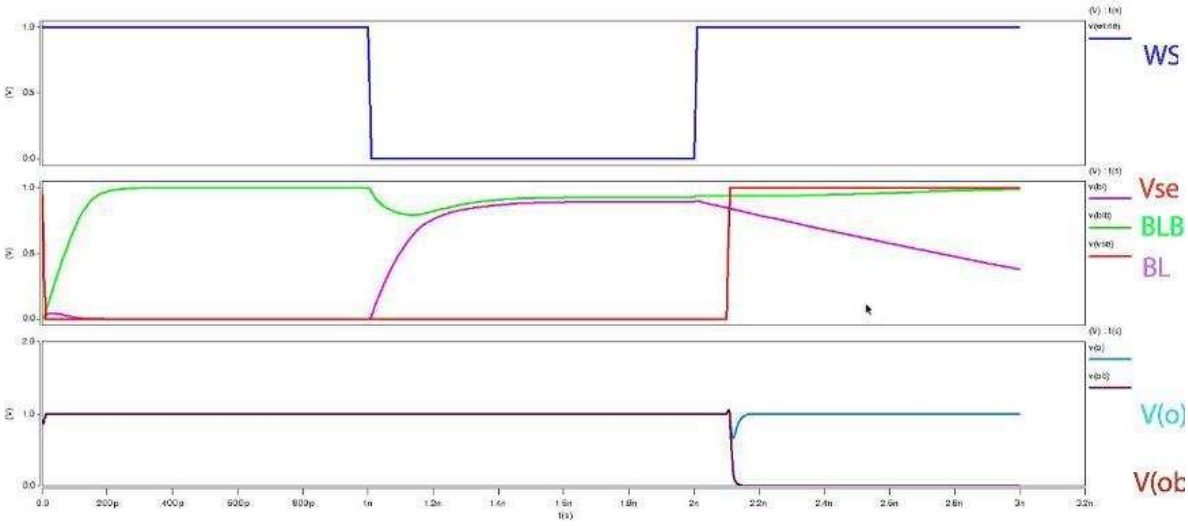

(a) Array structure for proposed design.




(b) Pre-charge select circuit




(c) Timing diagram for the proposed design




(d) HSPICE simulation of the proposed pre-select design

**Figure 5.11: Proposed pre-charge select design (a) array structure (b) circuit schematic (c) timing diagram (d) HSPICE simulation.**

### 5.4.1 Stability analysis

Pre-charging the bit-lines (BL, BLB) to a voltage below VDD results in faster and low power pre-charge operation due to the reduced voltage swing required at the bit-lines [92]. Decreasing the pre-charge voltage reduces the voltage rise at the storage node '0', thereby increases the SNM [93]. However when the pre-charge voltage falls below a certain value then it may degrade read speed and SRAM cell stability. It may not be a problem for the proposed design as kick size of $3\sigma_{offset}$ wouldn't be very large considering the fact that the required voltage differential is normally very small (<100 mV [31]). Figure 5.12 shows the SNM plot for different pre-charge voltages. Read stability increases till 0.65 V of pre-charge voltage, reaching a maximum at 0.7 V. Below 0.65 V, the SNM starts to fall below the SNM value at 1 V pre-charge voltage.

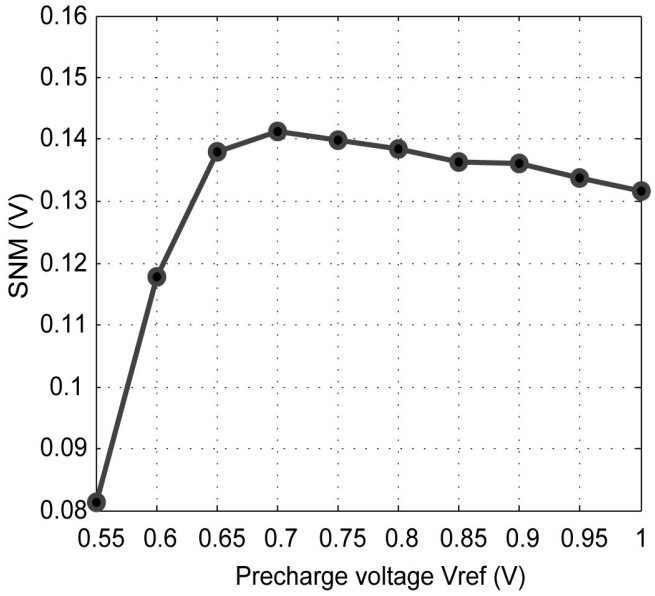
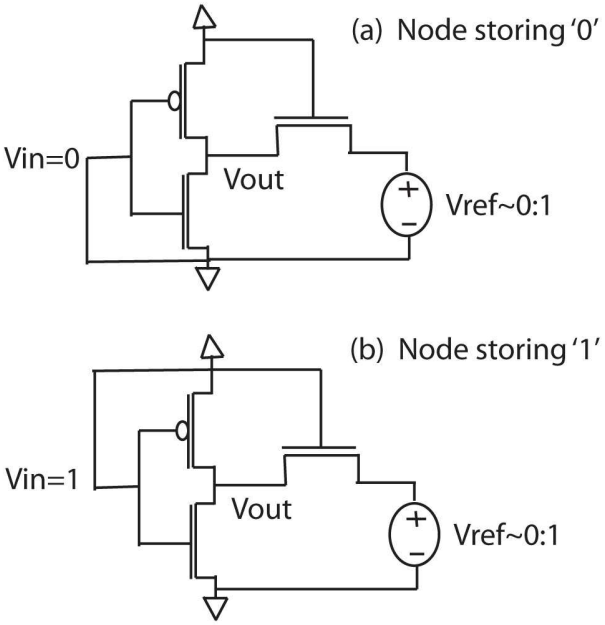
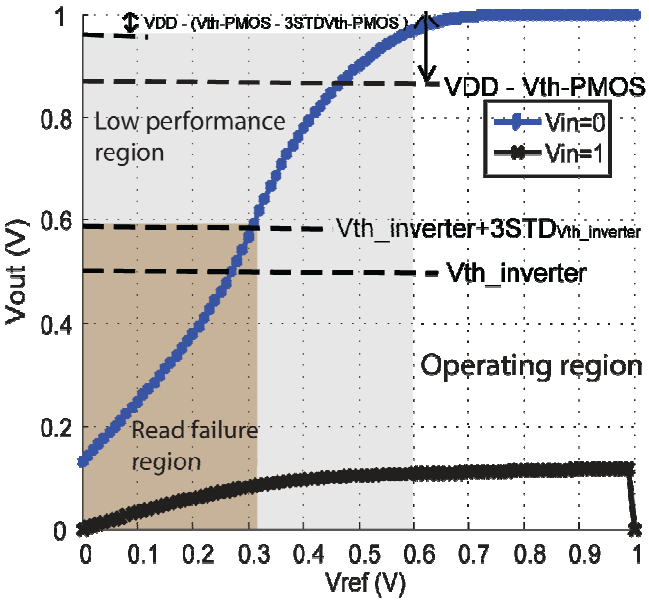



**Figure 5.12: Impact of pre-charge voltage on SNM.**

Figure 5.13(a) shows the set-up for stability analysis using open loop inverters with access transistors. A reference voltage $V_{ref}$ representing the bit-line pre-charge voltage is applied to observe the behaviour of the storage node, $V_{out}$. Simulation results of the given setup to observe cell stability are plotted in Figure 5.13(b). The storage node holding a '1' (Vin =0) keeps holding a strong '1' as long as $V_{ref}$ is higher than 0.7 V for 1 V of VDD, below which the storage node gets weakened. However this will not affect discharge delay as long as VDD - $V_{out}$ is less than the threshold voltage of PMOS connected to the storage node holding a '1' in a SRAM cell. We set a margin to account for the threshold variations due to variability, it now requires that we use $V_{ref}$ values for which VDD - $V_{out}$ $<< V_{th\_PMOS} - 3\sigma_{Vth\_PMOS}$ to account for worst case conditions. We call this as operating region to avoid any read speed penalty. When VDD - $V_{out}$ falls below the threshold of the PMOS connected to $V_{out}$ in close loop configuration, it (PMOS) is turned on that degrades the discharge speed. We call it a low performance region since the discharge delay increases under this condition. However the cell storage remains intact until $V_{out}$ falls below the inverter threshold voltage.




(a)

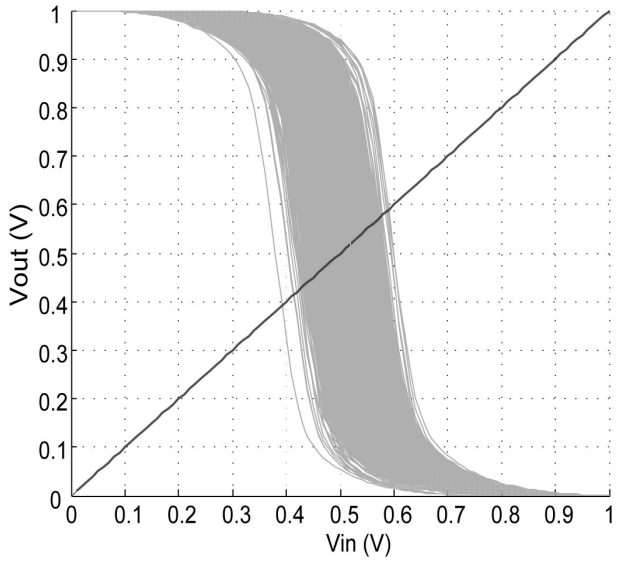



(b)

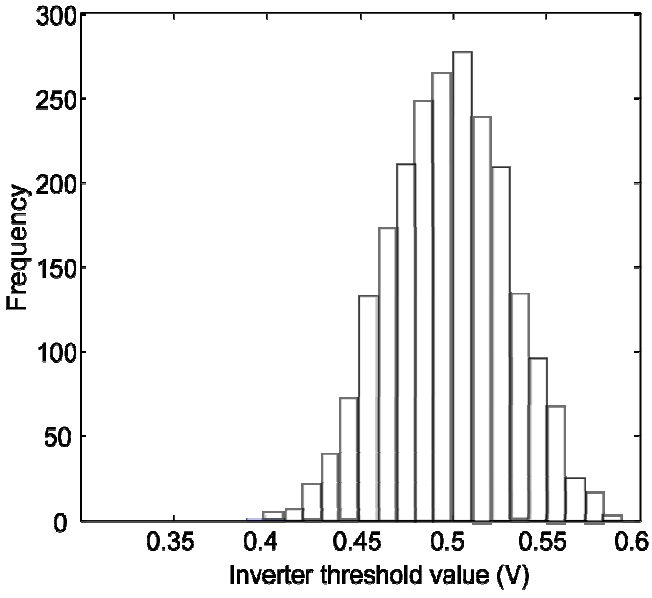
**Figure 5.13: Stability analysis (a) open loop SRAM cell (b) simulation results.**

To account for process related inverter threshold voltage variations, it requires that $V_{ref}$ should be selected such that, $V_{out} \geq V_{th\_inv} + 3\sigma_{Vth\_inv}$, any values below that are referred as failure region. Figure 5.14 shows result of inverter threshold variations under statistical variability. Mean inverter threshold lies on 502 mV with 28 mV of standard deviation (STD).

Figure 5.13(b) shows that $V_{ref} \geq 320$ mV ($V_{th\_inv} + 3\sigma_{Vth\_inv}$) would be sufficient to avoid a destructive read due to the low pre-charge voltage. However a pre-charge level lower than 0.6 VDD degrades the sense amplifier delay [12], therefore, we set 0.6 V as the minimum pre-charge voltage for the proposed pre-charge select design.




(a)




(b)

**Figure 5.14: Inverter threshold plot under statistical variations (a) DC plot (b) PDF of inverter threshold.**

The proposed pre-charge select design chooses a low supply voltage for the bit-line pre-charge that may impact offset voltage of the sense amplifier. A lower discharge voltage is developed at input transistors of the sense amplifier when the initial pre-charge levels are low.

This results in a lower drain current flow in the two branches of the sense amplifier that causes a large initial voltage difference in the latch circuit of the sense amplifier [12]. A high initial voltage difference results in a better sensing ability of the sense amplifier that corresponds to a lower offset voltage. Figure 5.15 shows the impact of low pre-charge levels on the standard deviation of the offset voltage of a current mode sense amplifier. Variations in the offset voltage decrease as the pre-charge voltage are lowered. There is no speed penalty of the sense amplifier as long as pre-charge voltage is 60% of VDD [12], below which operational current decreases due to a low drain to source voltage of the clocking transistor (M9 in Figure 5.2(a)). Therefore lowering the pre-charge levels for the proposed design reduces the offset variations without incurring any performance overhead.

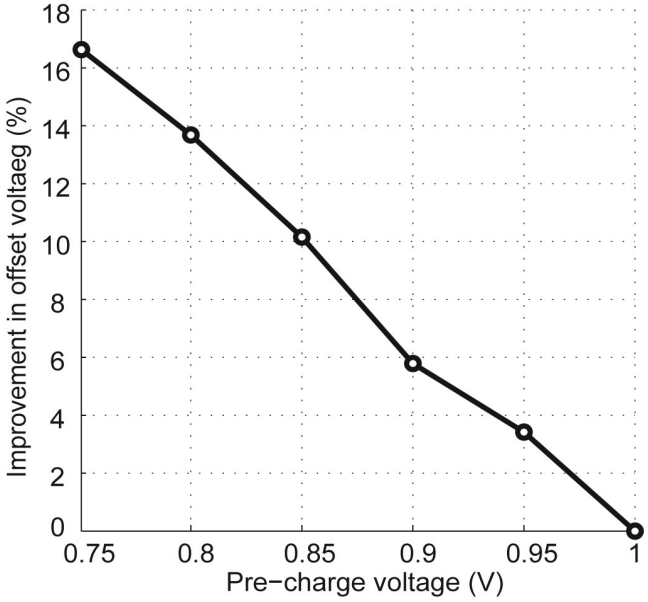
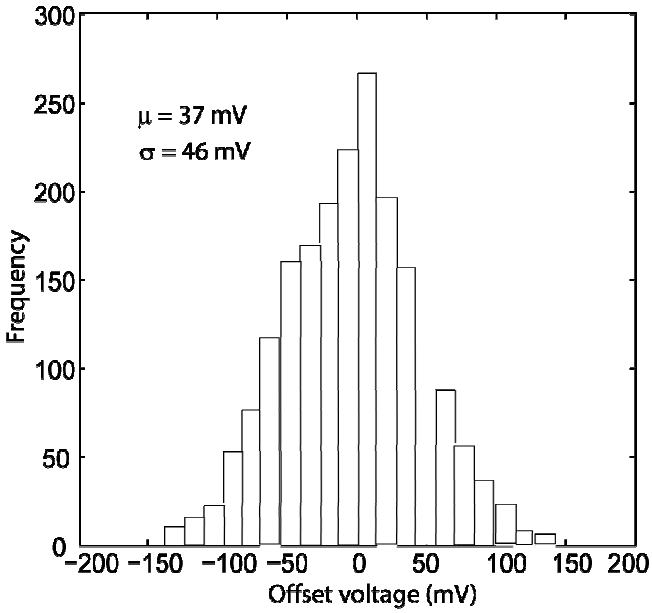



**Figure 5.15: Impact of low pre-charge on offset voltage.**

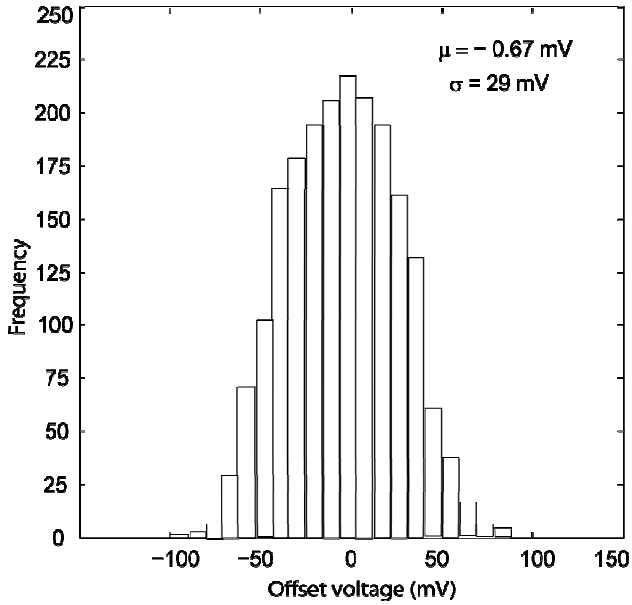
## 5.4.2 Statistical variability simulations

In order to investigate offset reductions using the proposed design, we implemented a 45nm 256x1 bit conventional 6T-SRAM column array and appended a sense amplifier with optimized sizing given in [72] to minimize offset voltage. An ensemble of 45 nm 200 model cards with random dopant fluctuations, line edge roughness, and poly-granularity [4] were used to insert statistical variability in design. Figure 5.16 shows result of 6,000 statistical variability simulations to calculate offset voltage for a comparative analysis. Conventional design has a 46 mV STD (standard deviation) of the offset voltage. Proposed design reduces it to 29 mV using a calibration of kick size $1\sigma_{offset}$ =46 mV, resulting in a 37% reduction in the

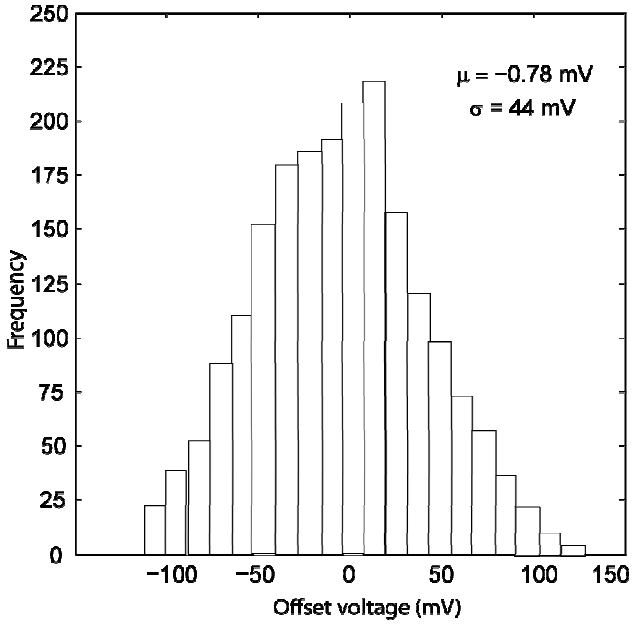
STD of the offset voltage. Increasing the kick size to $2\sigma_{offset}$ =92 mV reduces the effective offset variations to 44 mV that results in a 4% improvement. Although the improvement is less, however it can squeeze the worst case sense amplifiers in the range $2\sigma_{offset}$ - $4\sigma_{offset}$ to a range 0 - $2\sigma_{offset}$ . Large numbers of simulations are required to show improvements in the offset reduction with higher kick sizes.




(a) Conventional design




(b) Proposed design, $n\sigma_{offset}$ = 46 mV, n=1




(c) Proposed design, $n\sigma_{offset}$ =92 mV, n=2

**Figure 5.16: Offset voltages (a) conventional design (b) proposed design for $n\sigma_{offset}$ = 46 mV, n=1 (c) proposed design $n\sigma_{offset}$ =92 mV, n=2.**

### 5.4.3 Energy and area comparisons

For a comparative analysis with the conventional sizing technique, we designed a 45 nm 256x1 bits conventional 6T-SRAM array and measured energy consumption of both designs (conventional and proposed) for similar performance requirements ($\tau_{sense-amp} + \tau_{discharge}$ =124 ps). Conventional design requires a minimum differential voltage of 53 mV ($6\sigma_{offset}$) to achieve a given performance target (126 ps), while the proposed design requires a 106 mV ($6\sigma_{offset}$ using $3\sigma_{offset}$ kick) voltage differential to achieve the required discharge delay targets. Since large sized devices are used for the conventional design to achieve a low offset margin, therefore it costs a high dynamic power overhead. Proposed design provides a 42% (9.95 fJ vs. 5.78 fJ) reduction in the read energy as compared to the conventional design. Note the fact we didn't consider the case when both the bit-lines have different pre-charge voltages for the proposed design. Since a short current may flow during equalization when both the bit-lines have different voltages and are connected by an equalizer. However, the probability of such a case (worst case), when the low pre-charge voltages are selected for the bit-line pre-charge, is very small, Pr (offset> $3\sigma_{offset}$ ) < 1%.

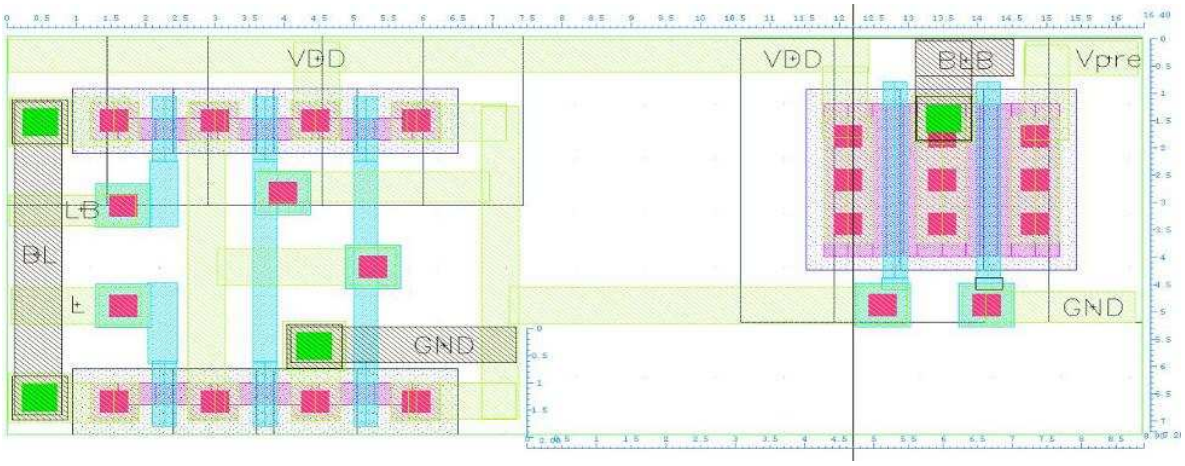



**Figure 5.17: Proposed pre-charge select circuit Area=** $(16.4\text{x}7.2-8.9-2)\,\mu m^2 = 100.28\,\mu m^2$.

A 350 nm process was used for the sense amplifiers and the proposed pre-charge select circuit layouts in order to carryout the area comparison of conventional and the proposed designs. Although the sense amplifier for the proposed design takes nearly half the area required by a conventional sense amplifier ($519\,\mu m^2 vs. 240\,\mu m^2$), however large sized pre-charge select transistors (W=16L), of area $200.5\,\mu m^2$, were used to allow a faster pre-charge of the bit-lines, therefore the total area reductions were reduced to a 15% ($519\,\mu m^2 vs. 440.5\,\mu m^2$). Figure 5.17 shows the layout of the proposed pre-charge select circuit with pre-charge select transistors. The layout of the sense amplifier for the proposed and conventional designs is given in Figure 5.10.

## 5.5 Chapter summary

Large parametric variations in the scaled technologies increase the offset variations of a typical SRAM sense amplifier design. To overcome the effect of variability the sense operation has to be delayed longer for a reliable SRAM read operation. However this increases the power consumption and decreases the read speed. In this chapter, two novel digital methods are presented to mitigate the offset voltage dependent discharge delays in order to minimize energy consumption and boost performance. The proposed discharge assist design method adds a discharge assist circuit to improve the bit-line discharge based on the asymmetry information of the sense amplifier. Monte Carlo statistical variability simulations indicate a 38% improvement in the discharge delay; however, energy reductions (16%) are not very substantial as compared to the performance improvements since this method results

in simultaneous discharge by the assist circuit and SRAM cell. The proposed design requires a 38% less area and consumes 16% less energy for the same speed measure when compared to a traditionally sized sense amplifier.

The other method adds a pre-charge select circuit to select an appropriate supply voltage for the bit-line pre-charge that reduces the differential voltage required for a reliable sensing. The pre-charge voltage on a bit-line, which is connected to a faster branch of the sense amplifier, is dropped to minimize the current difference in the two branches of the sense amplifier. Statistical variability simulations show a 37% reduction in the offset voltage using a $1\sigma_{\text{offset}}$ kick to recover the worst case sense amplifiers. The sense area reduction is 15% and the read energy reduction is 42% for the proposed design over a conventionally sized sense amplifier. The proposed pre-charge select method is more energy efficient as compared to the proposed discharge assist design since a single discharge path exists. However the proposed discharge method is more area efficient due to a low overhead assist circuit. The proposed design methods provide a means to low power robust SRAM design using *in-situ* digital offset compensation. The next chapter considers leakage reduction of the SRAM caches in idle periods.

# Chapter 6

# 6. SRAM cache leakage power reduction

CMOS technology has been the preferred choice of the semiconductor industry as CMOS devices consumed power only when switching. However, device and threshold voltage scaling has resulted in a high rise in the static power consumption of these devices, degrading the advantages of CMOS logic. Device miniaturization has resulted in a scaling of the lateral and vertical dimensions of CMOS transistors. Supply voltage has been scaled to maintain device reliability and low power consumption. The threshold voltage has scaled proportionally to the voltage in order to maintain the performance gains of device scaling. However, narrow oxide thicknesses and low threshold voltages result in a huge rise in gate leakage and sub-threshold leakage currents, respectively. Leakage power now takes a major portion of the total chip power and may exceed the dynamic power in future generations if left unchecked. Increased leakage power also degrades the reliability of popular test methods such as IDDQ and burn-in tests, tightens the requirements of cooling systems, and degrades system reliability.

SRAM caches are an important part of microprocessors as they typically take over 70% of the total chip area [10]. Since the total leakage power is proportional to the number of transistors [68], a reduction of the SRAM cache leakage is therefore critical for low power design. The cache memory can stay in long idle periods when not accessed, especially L2 cache. In those circumstances the leakage power for SRAM caches can exceed the dynamic power as seen for the 8KB instruction cache of the M32R processor at 45 nm technology [77]. Coupled with high leakage, SRAM caches face other challenges, such as small signal voltages and a large device mismatch of symmetric MOS transistors in SRAM cells and sense amplifiers. Increasing the device size can reduce this mismatch. However it increases the area overhead and will result in a rise in sub-threshold leakage current which is proportional to the device size.

This chapter will provide an introduction to device leakage, in the case of SRAM cache, and a proposed leakage minimization technique. This work focuses on SRAM cache arrays since they take the largest portion of the total SRAM cache area. Previously proposed leakage reduction techniques include power gating methods [76, 82-84], drowsy caches [16, 17, 85], and body biasing [79-81]. A brief overview of the previous research for cache leakage power reduction can be found in Chapter 2.

## 6.1 Types of MOS transistor leakage

Device scaling has resulted in improvements of device delay by approximately 30% every two years. However, to keep the power consumption under control, supply voltage scaling was also necessary. A low supply voltage has a negative impact on the device and circuit delay, therefore, the threshold voltage was scaled in proportion with the voltage scaling to keep leaps in performance gains. Threshold voltage scaling has resulted in a huge rise in the sub-threshold leakage current for sub-100 nm technologies. Scaling of the gate oxide thickness was also necessary to achieve a constant electric field scaling and minimize short channel effects. The short channel effect is the decrease in threshold voltage with a decrease in device gate length [86]. However, very small oxide thicknesses of a few atomic layers in nano-CMOS technologies have lead to a high gate leakage current. The classical model of infinite gate input impedance of MOS transistors is therefore no longer valid for deeply scaled devices due to high gate leakage currents. Introduction of the High-K devices at 45 nm technologies reflects a move to reduce the high gate leakage current by many folds, however, it may confront with the same challenges of scaling as with the current silicon-dioxide (SiO2) dielectrics. While the sub-threshold leakage current can be minimized better in design, the gate leakage has to be controlled through process technology [15]. Other kinds of leakage mechanisms are band to band tunnelling (BTBT), drain induced barrier lowering (DIBL), and body effect. The total leakage current depends on the supply voltage, threshold voltage, oxide thickness, drain/source junction depths, and device dimensions [86].

### 6.1.1 Sub-threshold leakage

Sub-threshold current refers to the drain current that flows from the source to the drain of a MOS device when the gate voltage is below the threshold voltage, $V_{th}$ [86]. Technology scaling has lead to the scaling of supply voltage for power reduction that requires threshold

voltage scaling to achieve a 30% delay reduction every next generation. For an ON transistor with the gate source voltage ( $V_{gs}$ ) higher than the threshold voltage, drift current is the major current from source to drain. Drift current is proportional to $(V_{ds} - V_{th})^{\alpha}$, where $1 \leq \alpha \leq 2$ and $V_{ds}$ is drain to source voltage [15]. Therefore to achieve reductions in the device delay while exploiting the supply voltage scaling, $V_{th}$ is also required to be scaled. The MOS transistor is OFF when the gate source voltage is zero $V_{gs}$ =0, diffusion becomes the major source of drain to source (threshold) leakage current. Low threshold voltages have therefore lead to large sub-threshold currents in scaled technologies.

#### 6.1.1.1 Drain induced barrier lowering (DIBL)

DIBL refers to a decrease in the threshold voltage at high drain voltages in the short channel devices. High drain voltages have a little impact on the sub-threshold current for long channel devices. However, a significant increase in the drain to source current occurs in short channel devices due to DIBL. A high drain voltage lowers the barrier potential causing the source terminal to inject more carriers into the channel, independent of the gate voltage [86]. This can be mitigated to some extent by increasing the channel doping concentration near the source and body junctions to reduce barrier lowering, called halo doping [15]. However the source to body and the drain to body junctions have finite lengths, limiting the minimum channel length, below which they are shorted to cause direct tunnelling current.

### 6.1.2 Gate oxide leakage

As scaling moves to nano dimensions, short channel effects (SCE) pose a major challenge to device reliability. SCE lead to a low gate control of the transistors in order to completely turn them on-off and a threshold voltage dependence on the device gate length. With smaller gate lengths, a MOSFET doesn't behave as a planar capacitor. To achieve ideal MOS behaviour, proportionate scaling of the lateral and vertical dimensions of devices is required to have good aspect ratio [15]. Aspect ratio represents the ratio of the vertical and horizontal dimensions of a MOS transistor. Gate oxide thickness is reduced to mitigate short channel effects [86] by providing an electrostatic field that resembles a planar capacitor. However, small oxide thicknesses and high electric fields lead to a very high gate tunnelling leakage from gate to substrate through the oxide, and vice versa. With very low oxide thicknesses, the

gate leakage currents may approach off-state sub-threshold leakage current level when the oxide thickness approaches 1 nm, limiting further scaling of the gate oxide thickness. However High-K devices provide a means to decrease leakage current and allow further scaling of the oxide thickness.

## 6.2 SRAM cell leakage mechanisms

SRAM caches can cause significant leakage current when put in an idle state because a minimum supply voltage is required all the time to hold data. Moreover, the SRAM cells are designed to be high speed to meet processor frequency requirements. High speed devices however contribute more to leakage currents due to low threshold voltages as explained in section 6.1.1. Figure 6.1 shows different leakage paths of an unselected SRAM cell. The leakage path L1, passing through the access transistor, M1, and driver transistor, M6, contributes a high amount of leakage since the bit-lines are pre-charged high normally. The other leakage path L3, passing though the pull up transistor, M3, and driver transistor, M5, can contribute high leakage as the node voltage is high, 1V. A negligible amount of current flows on the leakage path L2 since the pull up transistors (M3, M4) are normally kept high $V_{th}$ to minimize leakage and improve write stability. Similarly, the leakage current on path L4 through access transistor, M2, and driver transistor, M5, is negligible since both of them are OFF.

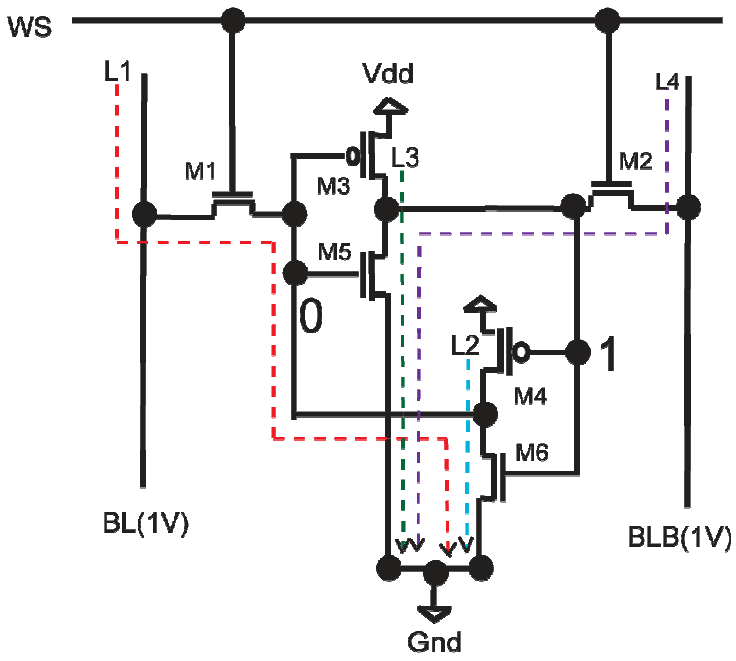



**Figure 6.1: Leakage paths in an unselected SRAM cell.**

## 6.3 Proposed segmented supply voltage method for leakage power reduction

The idea of decreasing the supply voltage of SRAM arrays during idle periods, in order to reduce the leakage current, has been previously investigated for drowsy caches [16, 17]. Leakage power reductions are quite high as decreasing the supply voltage decreases all kinds of leakage currents. However this method incurs a high wakeup latency and energy overheads. An aggressive drowsy cache was proposed to eliminate the wakeup latency [85]. But the access delay was degraded and read failures may occur when the bit-lines are pre-charged higher than the cell node voltages. Moreover the access energy overhead can be substantial if the cells are accessed often and put into idle mode after each access.

We propose a segmented supply voltage design to reduce the leakage power consumption of SRAM cache. A standard supply voltage is provided to an entire segment of the SRAM cache during an active mode period to achieve robust read and write operations with a minimum power/delay overhead. The supply voltage of the un-accessed segments is lowered to reduce the leakage power consumption in idle periods. Once accessed, each segment is left in an active state for a definite number of cycles since future accesses are expected to take place in that segment. There is no wakeup latency overhead for the transition from the drowsy mode to an active mode and the energy overhead is very small as it is amortized over a large number of access cycles. The chances of read failures are minimized by activating the high voltage (VDD) on the virtual supply line of the selected segment before the word line selects a particular word in that segment for a read operation. This allows the virtual supply voltage to achieve a voltage level that can enable a reliable read operation with a minimum delay degradation. Weak node voltages during a write operation are easy to be overwritten by full rail bit-lines voltage, therefore the write operation suffers no delay degradation and has a relatively low energy overhead.

Figure 6.2 shows an SRAM cell when used in a segmented drowsy cache. A virtual supply voltage, Vvdd, is provided to a complete cache segment. The voltage control transistors (MH, ML) allow switching between active and idle modes depending upon the activation signals (HighVolt, LowVolt). We avoid the initial wakeup latency by using the fact that address

decoding takes place in a hierarchal order. A small cache segment can therefore be selected earlier to wakeup before the word line is activated. Read failures may occur if the bit-lines are pre-charged to VDD and the read operation commences while the cell node voltages are not strong, a possible case for the aggressive drowsy cache. We observed that the bit-lines can be pre-charged to a lower voltage, less than VDD, to increase the robustness against read failures with a negligible increase in access delay. The increase in the discharge delay is very small as compared to a high discharge delay overhead incurred in the case of the segmented virtual ground architecture [83]. Moreover, the delay degradation is graceful, and is proportional to the pre-charge voltages. Low pre-charge voltages increase robustness against read failures at the cost of a small increase in the discharge delay. However the pre-charge voltages may not be lower than 0.6VDD, below which the sense amplifier delay tends to rise [12]. The accessed segment can be put to remain in an active mode for a definite number of clock cycles to reduce the wakeup energy overhead. The energy overhead may be quite high if the word lines are accessed often and put into idle mode soon after each access. The bit-lines are left floating during the idle period to minimize the bit-line leakage currents. It also removes the need of high threshold access transistors to minimize the high bit-line leakage when bit-lines are kept pre-charged during the idle periods as well.

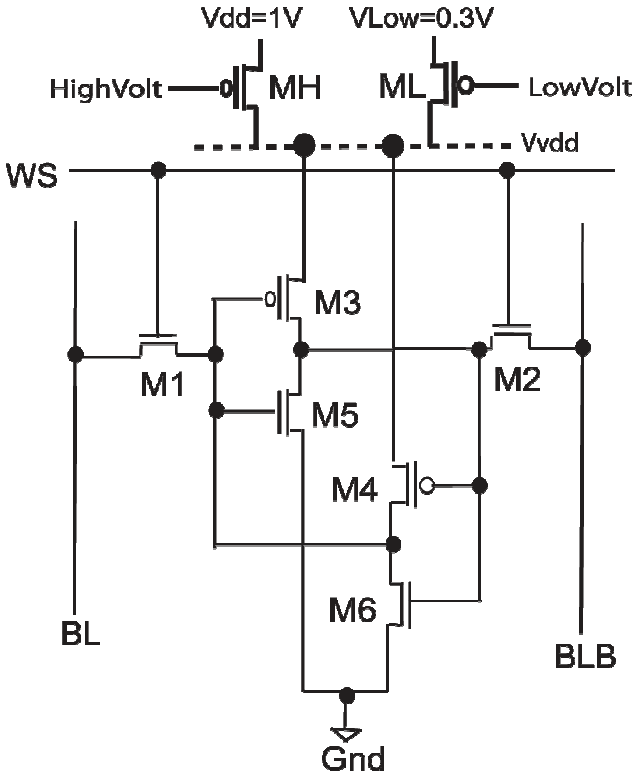
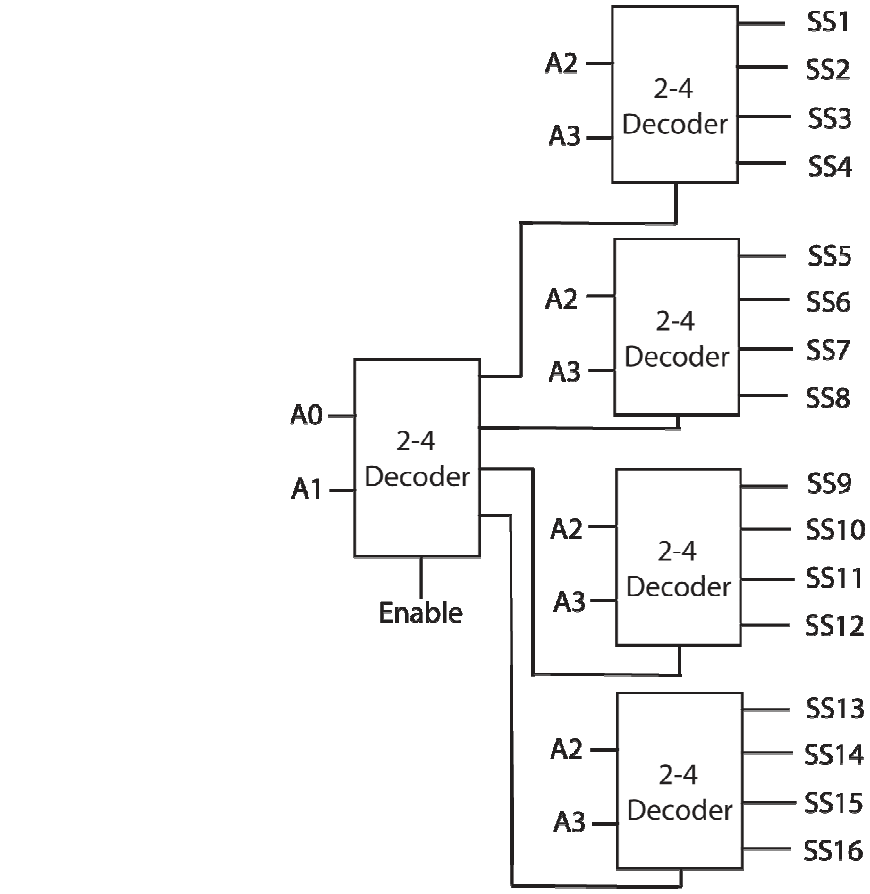


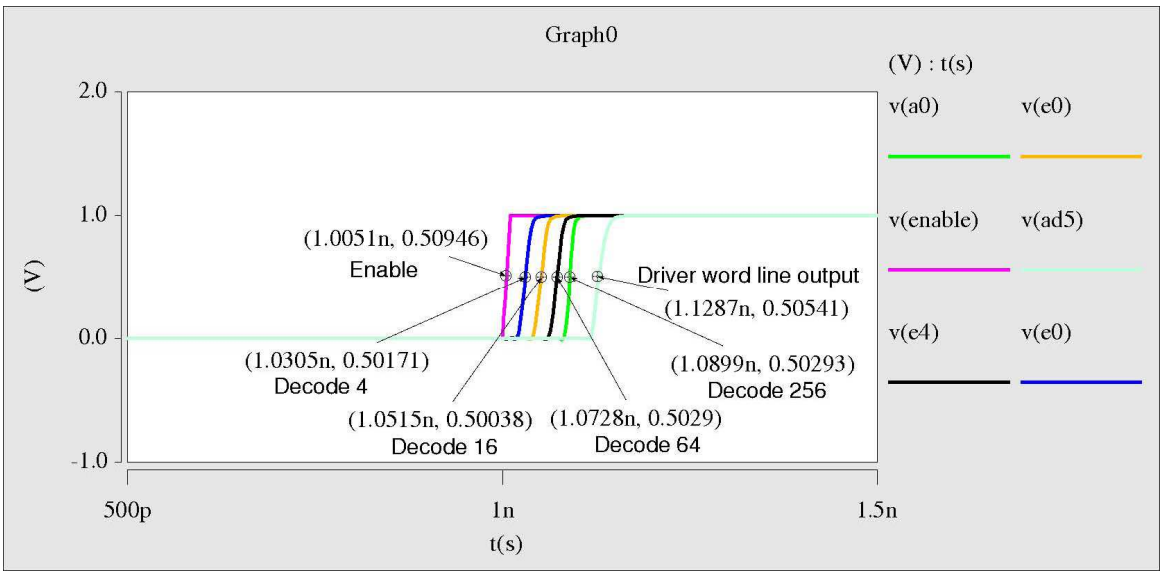

**Figure 6.2: Segmented drowsy cache cell.**

Figure 6.3(a) shows a hierarchal design of an 8x256 bits decoder. For clarity only two stages of decoding are displayed that select 16 segments each of 16 words from a 256 rows cache array. A 2x4 decoder can be used as a basic building block of this decoding process, consisting of 4 NAND gates. Once the enable signal is set high and a data/instruction address is placed at the input (A0A1A2A3…A*N*), the decoder selects first an accessed segment using the first four bits (A0A1A2A3). The supply voltage of the selected segment is then turned high upon selection. The time taken by the rest of the decoding stages and the time taken by the word line driver to raise a word line high is sufficient to enable a robust read operation with a negligible delay degradation. Figure 6.3(b) shows HSPICE-simulation results of a 45 nm 8x256 bits decoder implemented with NAND gates. The word line capacitance was approximated to be 24 fF. A 6 stages word line driver was used to drive the output of the decoder on a high capacitive word line, WS. The decoder takes 46 ps to select a 16 words segment from the 256x128 bits SRAM array, each word has 128 bits. A total of 77 ps is taken for the rest of the decoding, from the selection of 16 segments to a particular word line selection, including the delay of the word line driver.

Figure 6.4 shows the wakeup latency of activating the virtual supply voltage from an idle (drowsy) state (0.3 V) to an active state (1 V). The virtual voltage supply line, Vvdd, was connected to 16 words in each particular segment. Selection of too small (fine) a segment results in a small delay margin between segment selection and the word line activation. Therefore the cells may not achieve very high node voltages when the word line is selected, that can lead to a corrupt cell data. An alternate method is to use low pre-charge voltages that don’t exceed node cell voltage to avoid corruption of the cell data. However this will result in an increase in the access delay, albeit it is small. Too large a segment selection can avoid any access delay degradation, however it lowers the effectiveness of the drowsy schemes, as large segments may stay in the active state for a longer period of time.




(a)




(b)

**Figure 6.3: Hierarchal decoding to select 16 segments each of 16 words from a 45 nm 256 words array (a) architecture (b) decoder delay simulation.**

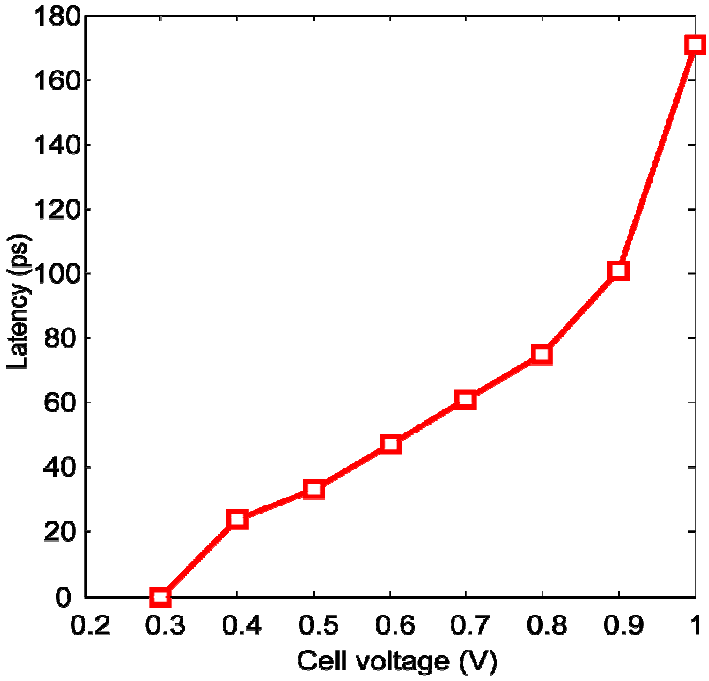



**Figure 6.4: Wakeup latency of raising the virtual ground.**

The virtual supply voltage, Vvdd, was approximated to have a 24x16 fF=384 fF total line capacitance. In Figure 6.2, the high voltage control transistor, MH, was sized larger 32Lx16=512 L, where L is the minimum channel length, to wake up the highly capacitive virtual supply line. The wakeup latency is not large when the high voltage is less than VDD=1V, e.g., the voltage controller takes about 75 ps to raise the supply voltage level from 0.3 V to 0.8 V, however it takes 171 ps to raise the virtual supply to 0.99 V. The decode process from 16-64-256 with word line driver takes 78 ps, therefore, the virtual supply can be raised up to 0.8 V without incurring any wakeup delays. We found that the degradation in the discharge delay is quite graceful if the virtual supply is lower than VDD. Therefore the access delay overhead is small even if the virtual supply voltage reaches 0.7 V. The results of these simulations indicate that the virtual supply voltage can be raised to a high voltage this enables reliable SRAM read operation without incurring any wakeup latency, while a negligible increase in read delay occurs.

### 6.3.1 Architecture of the proposed segmented supply voltage design

Figure 6.5 shows an implementation of the proposed segmented supply voltage architecture. A 45 nm 16x128 SRAM array segment was designed to demonstrate the proposed architecture. Each word line had 128 cells with 24 fF line capacitance. The bit-lines were approximated to have 19.2 fF bit-line capacitances. A virtual supply voltage was provided to the entire segment with a line capacitance of 16x24 fF=384 fF. The voltage levels of the virtual supply are controlled through voltage control transistors, MH and ML. The voltage control signal HighVolt is held low to wakeup the drowsy segment when a segment is to be selected for a read or write operation. A standard supply voltage (1 V) is then provided to the entire segment. The control signal LowVolt is turned low and the HighVolt is held high to put the entire segment in the drowsy state. The size of the MH transistor is kept large to enable quick recovery of the standard supply voltage from a drowsy mode. ML transistor can be kept minimum sized since a fast transition to drowsy state isn't necessary. We used a 512 L wide MH transistor and a 10 L wide ML transistor for our simulations, where L denotes minimum gate length, 35 nm.

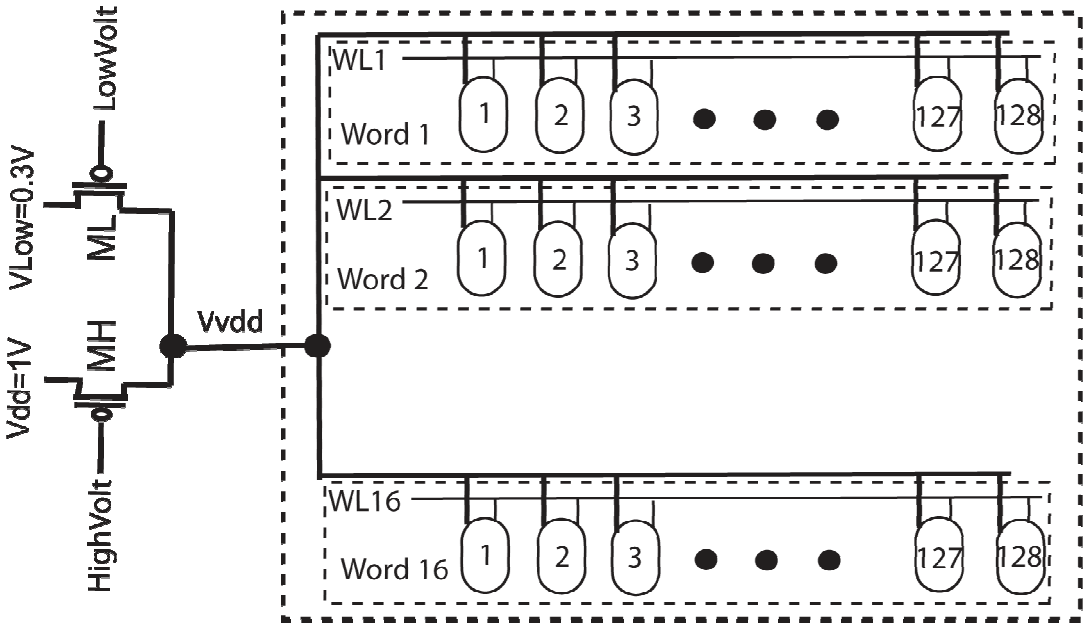



**Figure 6.5: Proposed virtual supply voltage architecture.**

Figure 6.6 shows a detailed implementation of the control circuit and gating mechanisms for the proposed segmented supply voltage design. A simple latch is used to hold the drowsy bit for each segment. When it holds a 0, M4 transistor is turned on to keep the segment in drowsy mode. To put a segment in drowsy mode, the MOS transistor MS is turned on by holding the Set signal high. A drowsy signal is generated for all segments and is AND with each Segment select signal (SS) to turn the Set signal high. The set transistor MS can be minimum sized as its output is driving a weaker voltage control transistor, M4. To reset the drowsy bit of a segment, a Reset signal is held high to turn on the reset transistor, M3. The reset signal is an AND of the /Drowsy signal and the Segment select signal (SS). Since the standard supply control transistor, M3, is very large (512L, L is 35 nm), we sized the reset transistor MR to be large enough (200 L) to quickly reset the drowsy bit and enable a fast turn ON of M3. The world select line, WSL, is gated with the /Drowsy signal to select a word line, WL, only when the segment is to be activated from the drowsy state. The inverted drowsy signal, /Drowsy, is set low and the Segment select signal (SS) is held high to put a segment in the drowsy mode without activating the word line.

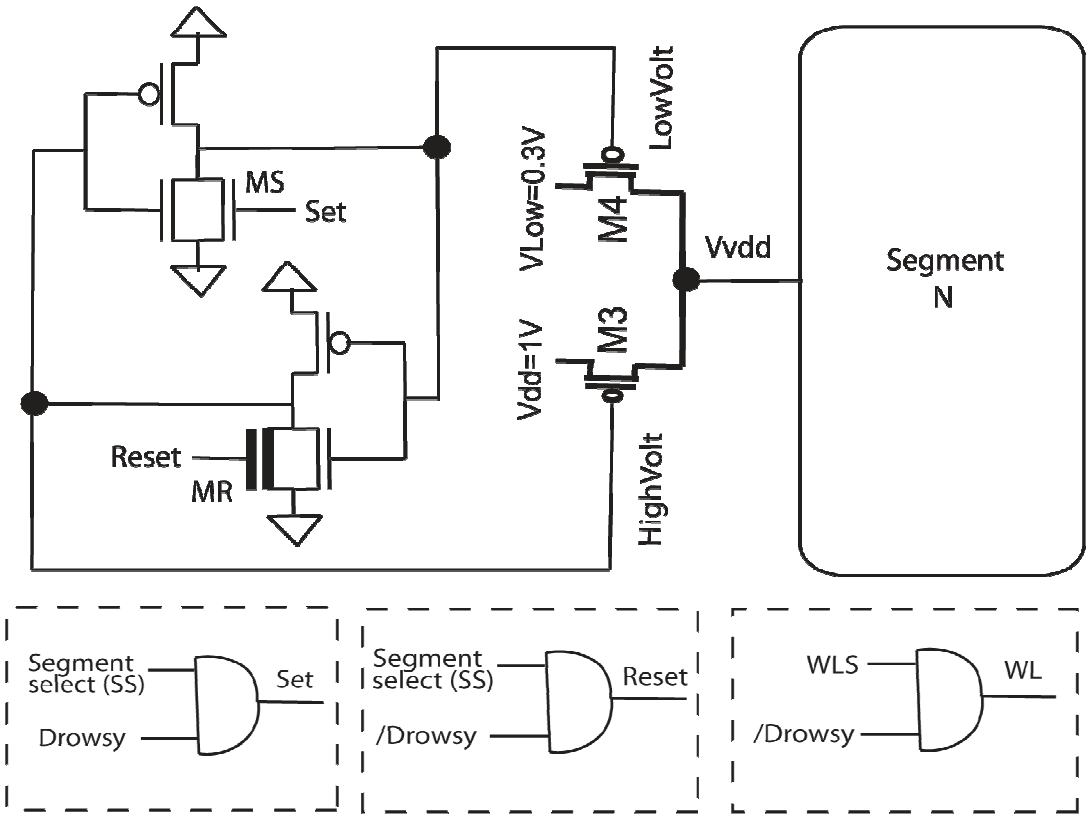



**Figure 6.6: Detailed implementation of the control circuit for segmented supply architecture.**

## 6.4 Simulation results and discussion

We have used 45 nm BSIM4 model cards with statistical variability [4, 22] to evaluate the effectiveness of the proposed segmented supply voltage architecture. This section provides simulation results of leakage power reductions, SNM analysis, and a comparison of power/delay overheads to conventional designs.

### 6.4.1 Read noise margins

A decrease in the supply voltage results in a lowering of the read margins that degrade the stability of the SRAM read operation. SNM analysis is carried out for the proposed design because it uses the fact that bit-lines can be pre-charged low (<VDD) and the supply voltage may not be VDD during the read operation. Figure 6.7 shows SNM ($\mu_{snm} - 3\sigma_{snm}$) of a 6T-SRAM cell for 8000 randomized simulations under statistical variability at different supply voltages. As evident in the plot, a supply voltage of less than 0.4 V may not be sufficient even for a $3\sigma$ design as some of the cells may have negative SNM. This plot also shows that the aggressive drowsy cache may be more prone to read failures as the SNM is not sufficient at low supply voltages. The proposed design raises the supply voltage of a selected segment in advance before a word line is selected. This provides sufficient margin for the cell voltages to rise higher than noise margins to enable reliable read operation. In the case of write operation, weak cell voltages are easier to be overwritten by higher bit-lines voltage.

Another important consideration for the leakage power reduction is the selection of minimum retention voltage. The dynamic retention voltage (DRV) should be chosen to minimize leakage power without destroying the cell data during hold. A uniform device may behave robustly up to 200 mV of the supply voltage in the hold period. However it will not be sufficient considering the impact of high variability with $6\sigma$ design, for a large number of SRAM cells. Figure 6.8 shows simulations results for the hold margins at different retention voltages. We observe that at a 200 mV retention voltage, many instances of the SRAM cells suffer from extreme variability and may lose cell data. However the cells are more stable at a 300 mV retention voltage even for a $6\sigma$ design. We have chosen 300 mV as the low voltage for the idle state to provide maximum energy savings with acceptable reliability.

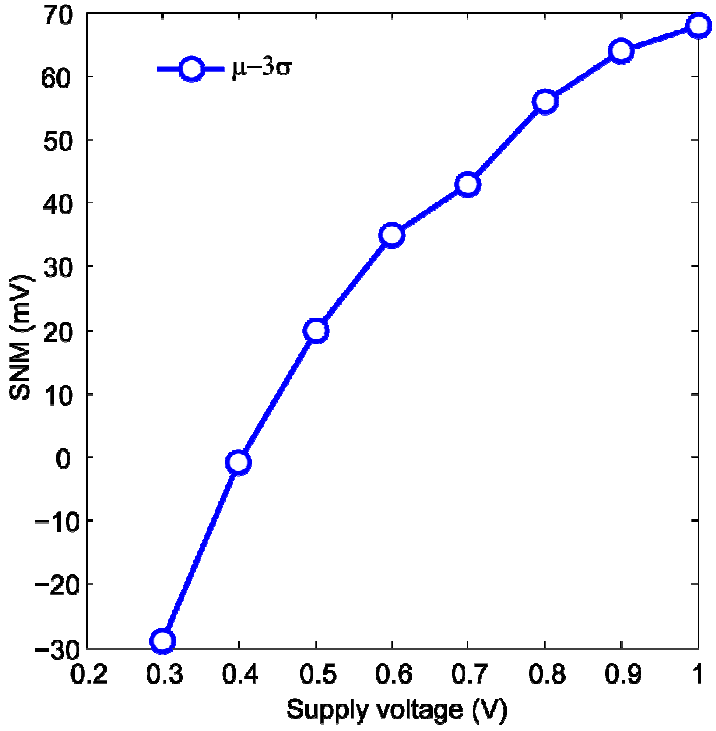



**Figure 6.7: SNM analysis in active mode at different supply voltages.**

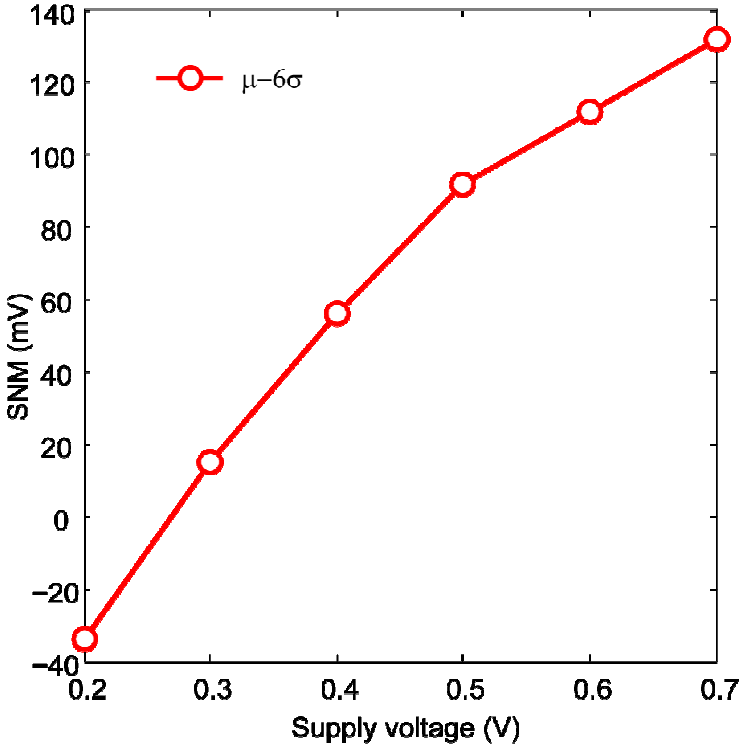



**Figure 6.8: Dynamic retention voltage (DRV) when subjected to statistical variability.**

### 6.4.2 Leakage reductions

A 45 nm 16x128 bits SRAM segment was designed with the voltage control circuitry as described in section 6.3. Figure 6.9 shows the leakage power reductions achieved for different retention voltages. As expected, the greatest power reductions are achieved when a minimum retention voltage of 300 mV is adopted. Maximum power reductions of 69% ($7\,\mu W$ vs. $22\,\mu W$) are achieved at 300 mV retention voltage. These savings decrease to 42% ($13\,\mu W$ vs. $22.6\,\mu W$) for 800 mV retention voltage. Increasing the retention voltage exponentially increases the total leakage power as leakage currents have an exponential dependence on the supply voltage. It should also be noted that maximum power reductions are achieved at a cost of low hold noise margins. Low noise margins therefore degrade the stability of the SRAM cells during hold periods.

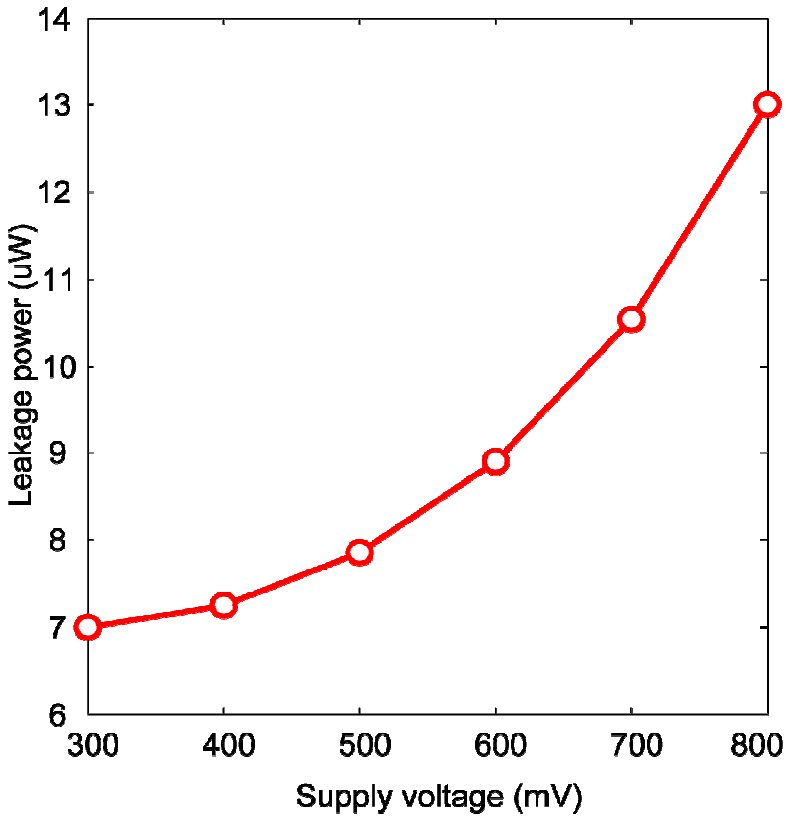



**Figure 6.9: Leakage power reduction for a 16x128 bits SRAM cache segment.**

### 6.4.3 Impact on discharge delay and power consumption

The transition from the drowsy state to an active state requires some wake up delay and energy overhead. We avoid wakeup delay by restoring the standard supply voltage of the selected segment in advance. However, when the virtual ground is not at VDD, some loss in performance occurs due to a small increase in the discharge delay. The wake up latency was calculated for a 16x128 bit segment. The supply line was approximated to have a 384 fF of capacitance. When the pre-charge voltages are set to VDD for a read operation, the performance loss is only 2.2% (140 ps vs. 137 ps). The delay was calculated as the time taken for the development of a 200 mV discharge differential voltage on the bit-lines. Although this results in a negligible impact on total read delay as the bit-line discharge is a small fraction of the total read access delay which includes address buffer delay, decoder delay, bit-line, sense-amplifier delay, data bus, and output buffer delay [83]. However, the bit-lines may be pre-charged to a voltage, slightly less than VDD, for more stability which results in lower discharge delays, as shown in Figure 6.10. When the bit-lines are pre-charged to 0.8 V, an increase of 5.8% (145 ps vs. 137 ps) occurs in the discharge delay. However, it has a very small impact on the total read access delay. There was no degradation in the write delay as the weak cell node voltages during the wakeup period are easily overwritten by full rail bit-lines.

The increase in the read energy was 50% (775 fJ vs. 517 fJ) when the segment makes a transition from the drowsy state to an active state. However, this is amortized over a large number of access cycles when the segment is left in active mode after being accessed for a read/write operation. There is no power overhead during the active mode, however a small energy (power x time) overhead of 2% (525 fJ vs. 514 fJ) is incurred as the discharge delays take longer to account for the worst case delays during wakeup. The write energy increases by 18% (1.73 pJ vs. 1.47 pJ) during the wakeup period, however, it remains the same as for the conventional design in the active mode, i.e. no write energy overhead occurs during write operation in the active mode.

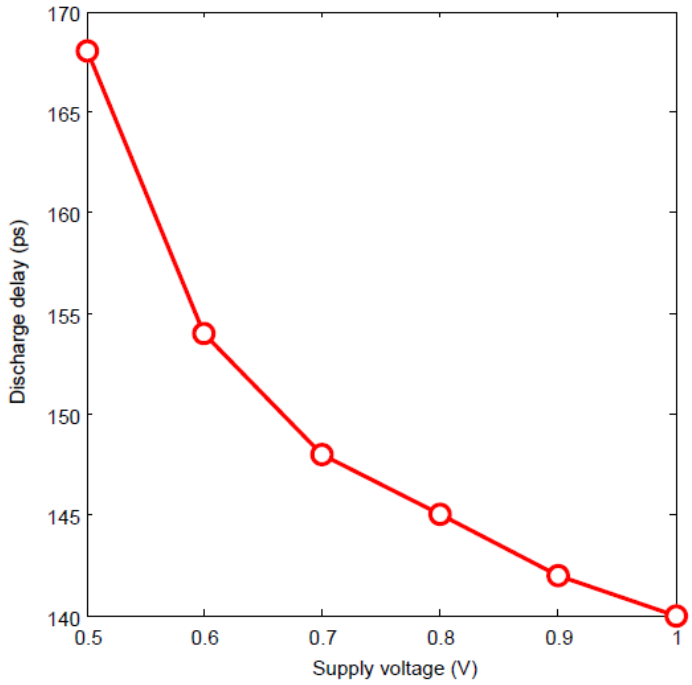



**Figure 6.10: Increase in discharge delay with decrease in pre-charge voltage.**

## 6.5 Chapter summary

SRAM caches occupy the bulk of the total chip area and take a major share of the total chip power since the leakage power is proportional to the number of transistors. An effective method to reduce the leakage power is to put the SRAM caches in a low voltage drowsy mode during idle periods since it reduces all kinds of leakages. Previously proposed drowsy mode designs either have a high performance overhead or degrade the reliability of the read operations. We propose a segmented supply voltage architecture that provides up to 69% reduction in the total leakage power without incurring any wakeup latency while the discharge delay increase is very small (2.2%). The fact that the address decoding takes place in a hierarchal fashion, it means an array segment can be selected before a word line is selected. The delay between the selection of a segment and the word line gives sufficient time to wakeup the supply line this can minimize the chances of read failures during the read operation. One other finding is that the use of low pre-charge voltages results in graceful degradation of the discharge delay. Therefore the bit-lines can be pre-charged to low voltages for robust read operation without incurring high wakeup latency and with very small impact on the total read delay.

# Chapter 7

# 7. Conclusion and future works

The aim of this research was to develop new designs and methodologies that enable the low-power robust circuit operation in nano-CMOS technologies. Increased variability arising from the manufacturing process and environmental conditions pose major challenge to reliable circuit design. Manufacturing processes result in identically designed devices to behave differently from each other due to the inaccuracies in fabricating nano-scaled geometries. Large variations in device behaviour still arise even under tight process control due to the discrete nature of the charge and matter. RDD, LER, and PoG are a few of those sources of statistical variations that may limit future scaling of transistors. Due to the random nature of these variations they can cause each transistor to behave differently from the others in its neighbourhood and can result in timing/power violations and even functional failures. The other type of variability, environmental variability, includes temperature and IR drop variations that arise from the varied load and switching activities in different blocks.

Static variations, especially the intrinsic variability can lead to high frequency and leakage power variations that may require large margins for functional design, lowering the power and performance gains of scaling. Variability has serve impact on the reliability of the sense amplifier and SRAM designs that lead to degraded yield and lower revenue. The supply voltage has been scaled slower due to large variations that lead to high power consumption. High power density and switching activities result in generation of temperature hot-spots and large supply voltage variations. Dynamic variations can cause timing failures for different functional blocks and high temperature variations may speed up the degradation of the devices with time. In addition, the soft error rate rises with scaling and large variations in soft error rate are observed due to variability that further worsens the prospects of a robust low-power design in nano-scaled technologies.

Variability results in large timing variations in the combination elements, therefore observing the timing failures provides an opportunity to quantify the extent of variability and counter measures can be taken to minimize its impact. Different pre-sampling and post-sampling techniques can be used to pre-detect or detect timing failures, respectively. We have presented two delay sensors (32 nm and 45 nm) in this work that are based on timing error prediction. Simulations results indicate high robustness to detect timing failures in advance under different variations and significant reductions in the energy consumption are achieved as compared to the conventional worst case design. The impact of variability can be even worst for the sequential elements as compared to the combinational circuits since different kinds of failures (read, write, hold, access, etc) can occur for the conventional 6T-SRAM design, due to high variability. We have presented different SRAM cell designs to meet high robustness, improved performance, and low-power requirements. Two novel digital offset mitigation methods are presented to decrease the read delays that result in a reduced energy and area overhead of the sense circuits as compared to the conventional sizing methods. Last part of the work focused on minimizing the leakage power of the SRAM arrays at a reduced wakeup overhead. The proposed in-situ designs for the combination circuits (e.g. in the pipeline stages) can be combined with the proposed variability resilient sequential circuits (SRAM and sense amplifiers) to enable a robust low-power digital circuit design in scaled technologies. We didn't focus on system level implementation (that includes combinational elements working with the sequential elements) in this work due to large computational constraints, and this remains a part of future works.

Chapter 2 presented a background to the sources of variability, their impact on design, and previously proposed techniques to counter variability in design. The sources of variability can be static that occurs during fabrication or they can be dynamic that originate at run time. Statistical variability represents a major obstacle to future scaling since it can cause each transistor to behave differently from others even in its neighbourhood. It can lead to large timing/leakage violations for the combinational logic and degraded stability in the case of SRAM design. Previously proposed methodologies and circuit designs for a low-power and/or robust circuit design were described in details with their constraints laid out to build a foundation to present the proposed designs in the later chapters.

Chapter 3 presented the proposed 32 nm and 45 nm delay sensors that enable low-power robust circuit operation for the combinational circuits. The proposed 45 nm delay sensor uses the delay offered by the master latch in a conventional master-slave flip-flop to create a guard band to detect timing failures before they cause an actual timing error. The delayed data and the original data, stored in the main flip-flop, are compared to detect any signal transition in the guard band that flags a timing error signal. The proposed 32 nm delay sensor uses an advanced clock signal for the shadow latch as opposed to a delayed clock signal in the Razor flip-flop to capture timing violations. Any mismatch in the data stored by the main flip-flop and by the shadow latch indicates a timing failure. The errors flagged by both the sensors predict possibility of an actual timing error if counter measures are not taken. Since an actual error doesn't occur, therefore an error recovery mechanism isn't necessary and different compensation techniques (body bias, voltage scaling, or frequency scaling) can be used to avoid actual timing errors in future. The proposed 32 nm delay sensor may complicate clock tree design due to generation of a delay clock signal. However the energy reductions are higher than the 45 nm delay sensor. Both designs can be extended for lower technologies, however further work is required to quantify the energy reductions.

The sequential elements (SRAM cache) represent another area of the digital design that requires careful attention under high process variations in scaled technologies. Conventional 6T-SRAM design provides very low read stability due to constraint requirements for the read and write operations. Chapter 4 presented 6T-asymmetric, SNM free 7T, and fully differential 8T SRAM cell designs that enable low-power and highly noise tolerant SRAM read/write operations. The proposed asymmetric 6T-SRAM cell strengthens the driver transistor of the feedback inverter in a 6T-SRAM cell, taking advantage of the single ended read operation to increase the SNM. A write assist transistor is used to provide virtual ground to the cross coupled inverter pairs of the cell connected to one word line. The virtual ground is left floating during the write operation weakening cell storage and thereby increasing write speed, enhancing write margins and lowering write power consumption. Although the asymmetric 6T-SRAM cell provides significant improvement in the SNM over a conventional 6T-SRAM design, however the read operation is still prone to failures under large variations. A single ended 7T-SRAM design is therefore presented to provide SNM free read operation and a highly robust low power write operation. We improved the 7T-SRAM design further to improve read delays and presented a fully differential 8T-SRAM, as the differential sense

operation is faster than the single ended read operation. The asymmetric 6T-SRAM design provides a better option when no area overheads are tolerated with some improvements in write margins. The 7T-SRAM design provides SNM free operation at the cost of 16% cell area overhead, and is useful when robust operation is required with moderate increase in the cell area. Whereas the fully differential 8T-SRAM design presents an option of highly robust and high speed design at the cost of large area overhead (30%).

Large offset voltage variations of the sense amplifiers pose serious challenge for robust SRAM design, as they result in a high power and performance loss. Chapter 5 presented two novel digital techniques to mitigate SRAM sense amplifier offset. The proposed pre-charge select design selects a low pre-charge voltage on a bit-line which is connected to a faster branch of the sense amplifier. This minimizes the current difference that is responsible for the large offset voltage of the sense amplifier in the two branches, and allows a low-power reliable read sense operation. The proposed design results in a 15% reduction in sense area and a 42% reduction in the energy consumption over a conventionally sized sense amplifier for similar performance metrics. The second method (discharge assist design) is based on the idea of minimizing the offset voltage dependent delay by assisting the bit-line discharge on a bit-line connected to a faster branch of the sense amplifier. The assisted discharge method results in a faster development of the required differential voltage, improving performance and saving energy. It results in a 27% reduction in the sense area and a 20% reduction in the total energy consumption during the read sense operation. The discharge assist method is a better choice when area overhead is of major consideration, while the pre-charge select design may be used when large energy reductions are required with lower area savings.

Chapter 3-5 presented different design methodologies and designs to mitigate the impact of variability on digital design. The last part of this work was focused on reducing the leakage power consumption for digital designs. Chapter 6 presented the proposed segmented supply voltage architecture to reduce the leakage power of SRAM arrays. SRAM caches are put in the drowsy mode during idle periods to save leakage power; however they incur a significant latency and energy overhead during wakeup. The proposed segmented supply voltage architecture selects a larger segment to wake up before the word line selects a particular word. This avoids the wakeup latency incurred in the previous drowsy cache designs. Using the fact that address decoding takes place in a hierarchal order, we can select a larger segment to wake

up before an actual word line is selected. We also found that the pre-charge voltages can be kept lower to enable more robust read operation in the case of drowsy caches and it incurs very small delay overhead. The proposed leakage reduction design can be combined with in-situ timing error monitoring and novel SRAM designs (proposed SRAM cell designs and offset voltage mitigation methods) to enable variability tolerant low-power digital design for future technologies.

## 7.1 Future work

There are a number of designs presented in this thesis which can be further investigated. For the combinational logic circuits, the delay sensors can be extended to provide dual sensing that avoids performance loss due to useless voltage scaling that can introduce voltage oscillation. Another area of improvement for the error predictive *in-situ* designs is to avoid an actual timing failure. Since there is no error recovery mechanism present in these designs, an actual timing error due to high variability and data dependency may occur. For the sequential circuits, it would be interesting to implement complete SRAM design including SRAM arrays, decoders, sense-amplifiers, output buffers, etc for a more detailed performance/power analysis. However it would require a system level design that may involve use of RTL or C languages for simulation. The proposed SRAM sense amplifier offset mitigation methods are implemented for differential sense amplifiers. It would be a useful investigation to implement them for single ended SRAM read designs to determine any energy/delay improvements. The final area of research that needs more investigation is the leakage power reduction for peripheral components of the SRAM or combinational circuits. We have investigated leakage reductions for SRAM arrays only in this work as they take bulk of the SRAM cache area and require a constant supply voltage to retain data. However, the peripheral components such as row decoders and word line drivers consume a significant portion of the leakage power consumption. An interesting setup would be to use the proposed supply voltage architecture along with peripheral leakage reduction methods together to achieve maximum leakage power savings.

# References


[1] Moore, G.E., *Cramming more components onto integrated circuits.* Electronics, 1965. **38**: p. 114-117.

[2] Asenov, A., *Random dopant induced threshold voltage lowering and fluctuations in sub 50 nm MOSFETs: a statistical 3D `atomistic' simulation study* Nanotechnology, 1999. **10**(2): p. 153–158.

[3] Roy, G., Adamu-Lema, F., Brown, A. R., Roy, S., and Asenov, A. *Simulation of combined sources of intrinsic parameter fluctuations in a 'real' 35 nm MOSFET*. in *Solid-State Device Research Conference, 2005. ESSDERC 2005. Proceedings of 35th European*. 2005.

[4] Brown, A.R., Roy, G., and Asenov, A., *Poly-Si-Gate-Related Variability in Decananometer MOSFETs With Conventional Architecture.* Electron Devices, IEEE Transactions on, 2007. **54**(11): p. 3056-3063.

[5] Borkar, S., *Designing reliable systems from unreliable components: the challenges of transistor variability and degradation.* IEEE Micro, 2005. **25**(6): p. 10-16.

[6] Das, S., Pant, S., Roberts, D., Lee, S., Blaauw, D., Austin, T., Mudge, T., and Flautner, K. *A self-tuning DVS processor using delay-error detection and correction*. in *VLSI Circuits, 2005. Digest of Technical Papers. 2005 Symposium on*. 2005.

[7] Ernst, D., Kim, N.S., Das, S., Pant, S., Rao, R., Pham, T., Ziesler, C., Blaauw, D., Austin, T., Flautner, K., and Mudge, T. *Razor: a low-power pipeline based on circuit-level timing speculation*. in *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*. 2003.

[8] Agarwal, M., Paul, B. C., Ming, Zhang, and Mitra, S. *Circuit Failure Prediction and Its Application to Transistor Aging*. in *VLSI Test Symposium, 2007. 25th IEEE*. 2007.

[9] Zhang, M., Mak, T., Tschaz, J., Kim, K.S., Seifert, N., and Lu, D. *Design for Resilience to Soft Errors and Variations*. in *On-Line Testing Symposium, 2007. IOLTS 07. 13th IEEE International*. 2007.

[10] Kevin, Z., *Embedded Memories for Nano-Scale VLSIs*. 2009: Springer Publishing Company, Incorporated. 400.

[11] Agarwal, K., and Nassif, S., *The Impact of Random Device Variation on SRAM Cell Stability in Sub-90-nm CMOS Technologies.* Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, 2008. **16**(1): p. 86-97.

[12] Wicht, B., Nirschl, T., and Schmitt-Landsiedel, D., *Yield and speed optimization of a latch-type voltage sense amplifier.* Solid-State Circuits, IEEE Journal of, 2004. **39**(7): p. 1148-1158.

[13] Zhang, K., Hose, K., De, V., and Senyk, B. *The scaling of data sensing schemes for high speed cache design in sub-0.18 μm technologies*. in *VLSI Circuits, 2000. Digest of Technical Papers. 2000 Symposium on*. 2000.

[14] Cosemans, S., Dehaene, W., and Catthoor, F. *A 3.6pJ/access 480MHz, 128Kbit on-Chip SRAM with 850MHz boost mode in 90nm CMOS with tunable sense amplifiers to cope with variability*. in *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*. 2008.

[15] Narendra, S.G., and Chandrakasan, A.P. , *Leakage in Nanometer CMOS Technologies*. 2006: Sringer.

[16] Flautner, K., Nam Sung, K, Steve, M., David, .B, and Trevor, M., *Drowsy caches: simple techniques for reducing leakage power.* SIGARCH Comput. Archit. News, 2002. **30**(2): p. 148-157.

[17] Nam Sung, K., Krisztian, F., David, B., and Trevor, M., *Circuit and microarchitectural techniques for reducing cache leakage power.* IEEE Trans. Very Large Scale Integr. Syst., 2004. **12**(2): p. 167-184.

[18] Azam, T., and Cumming, D.R.S. *Robust low power design in nano-CMOS technologies*. in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*. 2010.

[19] Azam, T., and Cumming, D. R. S., *Efficient sensor for robust low-power design in nano-CMOS technologies.* Electronics Letters, 2010. **46**(11): p. 773-775.

[20] Azam, T., Cheng, B., and Cumming, D. R. S. *Variability resilient low-power 7T-SRAM design for nano-scaled technologies*. in *Quality Electronic Design (ISQED), 2010 11th International Symposium on*. 2010.

[21] Azam, T., Cheng, B., Roy, S., and Cumming, D. R. S., *Robust asymmetric 6T-SRAM cell for low-power operation in nano-CMOS technologies.* Electronics Letters, 2010. **46**(4): p. 273-274.

[22] Cheng, B., Moezi, N., Dideban, D., Roy, G. ,Roy, S., and Asenov, A. *Benchmarking the Accuracy of PCA Generated Statistical Compact Model Parameters Against Physical Device Simulation and Directly Extracted Statistical Parameters*. in *Simulation of Semiconductor Processes and Devices, 2009. SISPAD '09. International Conference on*. 2009.

[23] Asenov, A., Brown, A. R. ,Davies, J. H. , Kaya, S., and Slavcheva, G., *Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs.* Electron Devices, IEEE Transactions on, 2003. **50**(9): p. 1837-1852.

[24] *Predictive Technology Models (PTM) are available online at : http://www.eas.asu.edu/~ptm*

[25] Siva, G.N., *Challenges and design choices in nanoscale CMOS.* J. Emerg. Technol. Comput. Syst., 2005. **1**(1): p. 7-49.

[26] Rabaey, J.M., *Digital Integrated Circuits : A Design Perspective* 1ed. 1995: Prentice Hall. 702.

[27] Andrie, S., and Manjov, S., *CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies: Process-Aware SRAM Design and Test* 2008, Springer.

[28] Knight, W., *Two heads are better than one [dual-core processors]*. IEE Review, 2005. **51**(9): p. 32-35.

[29] Hofstee, H.P., *Future microprocessors and off-chip SOP interconnect.* Advanced Packaging, IEEE Transactions on, 2004. **27**(2): p. 301-303.

[30] Osman, S.U., Tschanz, J.S., Bowman, K., De, V., Vera, X., Gonzalez, A., and Ergin, O., *Impact of Parameter Variations on Circuits and Microarchitecture.* IEEE Micro, 2006. **26**(6): p. 30-39.

[31] Heald, R., and Wang, P. *Variability in sub-100nm SRAM designs*. in *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*. 2004.

[32] Saxena, S., Hess, C., Karbasi, H., Rossoni, A., Tonello, S., McNamara, P., Lucherini, S., Minehane, S., Dolainsky, C., and Quarantelli, M., *Variation in Transistor Performance and Leakage in Nanometer-Scale Technologies.* Electron Devices, IEEE Transactions on, 2008. **55**(1): p. 131-144.

[33] Blaauw, D., Chopra, K. ,Srivastava, A., and Scheffer, L., *Statistical Timing Analysis: From Basic Principles to State of the Art.* Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, 2008. **27**(4): p. 589-607.

[34] Onodera, H. *Variability modeling and impact on design*. in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*. 2008.

[35] Saha, S., *Modeling Process Variability in Scaled CMOS Technology.* Design & Test of Computers, IEEE, 2010. **PP**(99): p. 1-1.

[36] Borkar, S., Karnik, T., Narendra, S., Tschanz, J., Keshavarzi, A., and De, V. *Parameter variations and impact on circuits and microarchitecture*. in *Design Automation Conference, 2003. Proceedings*. 2003.

[37] Kuhn, k., Kenyon, C., Kornfeld, A., Liu, M., Maheshwari, A., Shih, W., Sivakumar, S., Taylor, G., VanDerVoorn, P., and Zawadzki, K., *Managing Process Variation in Intel's 45nm CMOS Technology*. Intel Technology Journal, 2008. **12**(2): p. 92-110.

[38] Roy, G., Adamu-Lema, F. , Brown , A.R., Roy, S., and Asenov, A. , *Intrinsic parameter fluctuations in conventional MOSFETs until the end of the ITRS: A statistical simulation study.* Journal of Physics: Conference Series, 2006. **38**(1).

[39] Mak, T.M., Krstic, A., Cheng, K. T., and Wang, Li C., *New challenges in delay testing of nanometer, multigigahertz designs.* Design & Test of Computers, IEEE, 2004. **21**(3): p. 241-248.

[40] Sylvester, D., Agarwal, K., and Saumil, S., *Variability in nanometer CMOS: Impact, analysis, and minimization.* Integration, the VLSI Journal, 2008. **41**(3): p. 319-339.

[41] Greskamp, B., S.R. Sarangi, and J. Torrellas. *Threshold Voltage Variation Effects on Aging-Related Hard Failure Rates*. in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*. 2007.

[42] Cannon, E.H., KleinOsowski, A., Kanj, R., Reinhardt, D. D., and Joshi, R. V., *The Impact of Aging Effects and Manufacturing Variation on SRAM Soft-Error Rate.* Device and Materials Reliability, IEEE Transactions on, 2008. **8**(1): p. 145-152.

[43] Qian, D., Rong, L., and Yuan, X. *Impact of process variation on soft error vulnerability for nanometer VLSI circuits*. in *ASIC, 2005. ASICON 2005. 6th International Conference On*. 2005.

[44] Taylor, E., and Fortes, J., *Device variability impact on logic gate failure rates*, in *Government Microcircuit Applications and Critical Technology Conference (GOMAC)*. 2007.

[45] Roy, K., Mak, T. M., and Kwang-Ting, Cheng. *Test consideration for nanometer scale CMOS circuits*. in *VLSI Test Symposium, 2003. Proceedings. 21st*. 2003.

[46] Mitra, S., Zhang, M., Seifert, N., Mak, T. M., and Kim, K.S. *Built-In Soft Error Resilience for Robust System Design*. in *Integrated Circuit Design and Technology, 2007. ICICDT '07. IEEE International Conference on*. 2007.

[47] Constantinescu, C., *Trends and challenges in VLSI circuit reliability*. Micro, IEEE, 2003. **23**(4): p. 14-19.

[48] Tschanz, J.W., Narendra, S., Nair, R., and De, V., *Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low power and high performance microprocessors.* Solid-State Circuits, IEEE Journal of, 2003. **38**(5): p. 826-829.

[49] Chen, T., and Naffziger, S., *Comparison of adaptive body bias (ABB) and adaptive supply voltage (ASV) for improving delay and leakage under the presence of process variation.* IEEE Trans. Very Large Scale Integr. Syst., 2003. **11**(5): p. 888-899.

[50] Das, S., Tokunaga, C., Pant, S., Ma, W. H., Kalaiselvan, S., Lai, K., Bull, D. M., and Blaauw, D. T., *RazorII: In Situ Error Detection and Correction for PVT and SER Tolerance.* Solid-State Circuits, IEEE Journal of, 2009. **44**(1): p. 32-48.

[51] Toshinori, S., and Kunitake, Y., *A Simple Flip-Flop Circuit for Typical-Case Designs for DFM*, in *Proceedings of the 8th International Symposium on Quality Electronic Design*. 2007, IEEE Computer Society.

[52] Kunitake, Y., Toshinori, S., Yasuura, H., *Mitigating Performance Loss in Aggressive DVS Using Dual-Sensing Flip-Flops*, in *VLSI-SoC*. 2008. p. 543-546.

[53] Cheng, B., Roy, S., Roy, G., Brown, A., and Asenov, A. *Impact of Random Dopant Fluctuation on Bulk CMOS 6-T SRAM Scaling*. in *Solid-State Device Research Conference, 2006. ESSDERC 2006. Proceeding of the 36th European*. 2006.

[54] Bo, Z., Blaauw, D., Sylvester, D., and Hanson, S. *A Sub-200mV 6T SRAM in 0.13um CMOS*. in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*. 2007.

[55] Bo, Z., Hanson, S., Blaauw, D., and Sylvester, D., *A Variation-Tolerant Sub-200 mV 6-T Subthreshold SRAM.* Solid-State Circuits, IEEE Journal of, 2008. **43**(10): p. 2338-2348.

[56] Singh, J., Pradhan, D.K., Hollis, S., and Saraju P. M., *A single ended 6T SRAM cell design for ultra-low-voltage applications.* IEICE Electronics Express, 2008. **5**(18): p. 750-755.

[57] Keunwoo, K., Jae-Joon, K., and Ching-Te, C. *Asymmetrical SRAM Cells with Enhanced Read and Write Margins*. in *VLSI Technology, Systems and Applications, 2007. VLSI-TSA 2007. International Symposium on*. 2007.

[58] Mizuno, H., and Nagano, T., *Driving source-line cell architecture for sub-1-V high-speed low-power applications.* Solid-State Circuits, IEEE Journal of, 1996. **31**(4): p. 552-557.

[59] Takeda, K., Hagihara, Y., Aimoto, Y., Nomura, M., Nakazawa, Y., Ishii, T., and Kobatake, H., *A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications.* Solid-State Circuits, IEEE Journal of, 2006. **41**(1): p. 113-121.

[60] Chang, L., Montoye, R. K., Nakamura, Y., Batson, K. A., Eickemeyer, R. J., Dennard, R. H., Haensch, W., and Jamsek, D., *An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches.* Solid-State Circuits, IEEE Journal of, 2008. **43**(4): p. 956-963.

[61] Verma, N., and Chandrakasan, A. P., *A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy.* Solid-State Circuits, IEEE Journal of, 2008. **43**(1): p. 141-149.

[62] Tae-Hyoung, K., Liu, J., and Kim,C. H. *An 8T Subthreshold SRAM Cell Utilizing Reverse Short Channel Effect for Write Margin and Read Performance Improvement*. in *Custom Integrated Circuits Conference, 2007. CICC '07. IEEE*. 2007.

[63] Sheng, L., Yong-Bin, Kim, Fabrizio, and Lombardi, *A low leakage 9t sram cell for ultra-low power operation*, in *Proceedings of the 18th ACM Great Lakes symposium on VLSI*. 2008, ACM: Orlando, Florida, USA.

[64] Noguchi, H., Okumura, S., Iguchi, Y., Fujiwara, H., Morita, Y., Nii, K., Kawaguchi, H., and Yoshimoto, M. *Which is the best dual-port SRAM in 45-nm process technology? - 8T, 10T single end, and 10T differential-*. in *Integrated Circuit Design and Technology and Tutorial, 2008. ICICDT 2008. IEEE International Conference on*. 2008.

[65] Calhoun, B.H., and Chandrakasan, A. P., *A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation.* Solid-State Circuits, IEEE Journal of, 2007. **42**(3): p. 680-688.

[66] Tae-Hyoung, K., Liu, J., Keane, J., and Kim, C. H. *A High-Density Subthreshold SRAM with Data-Independent Bitline Leakage and Virtual Ground Replica Scheme*. in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*. 2007.

[67] Singh, J., Pradhan, D.K., Hollis, S., and Mohanty, S.P., *Single Ended 6T SRAM with Isolated Read-Port for Low-Power Embedded Systems*, in *Design Automation and Test in Europe (DATE), 2009, 12th IEEE International Conference* 2009.

[68] Navid, A., Farid, N. N., and Andreas, M., *Low-leakage asymmetric-cell SRAM*. IEEE Trans. Very Large Scale Integr. Syst., 2003. **11**(4): p. 701-715.

[69] Singh, R., and Bhat, N., *An offset compensation technique for latch type sense amplifiers in high-speed low-power SRAMs.* Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, 2004. **12**(6): p. 652-657.

[70] Sinangil, M.E., Verma, N.,and Chandrakasan, A.P. *A 45nm 0.5V 8T column-interleaved SRAM with on-chip reference selection loop for sense-amplifier*. in *Solid-State Circuits Conference, 2009. A-SSCC 2009. IEEE Asian*. 2009.

[71] *ITRS Roadmap can be found at : http://www.itrs.net/.*

[72] Yeung, J., and Mahmoodi, H. *Robust Sense Amplifier Design under Random Dopant Fluctuations in Nano-Scale CMOS Technologies*. in *SOC Conference, 2006 IEEE International*. 2006.

[73] Saibal, M., Rajiv, V. Joshi, Keunwoo, Kim, and Ching-Te, Chuang, *Variability Analysis for sub-100nm PD/SOI Sense-Amplifier*, in *Proceedings of the 9th international symposium on Quality Electronic Design*. 2008, IEEE Computer Society.

[74] Bhargava, M., McCartney, M. P., Hoefler, A., and Mai, K. *Low-overhead, digital offset compensated, SRAM sense amplifiers*. in *Custom Integrated Circuits Conference, 2009. CICC '09. IEEE*. 2009.

[75] Lee, M.J.E., W. Dally, and P. Chiang. *A 90 mW 4 Gb/s equalized I/O circuit with input offset cancellation*. in *Solid-State Circuits Conference, 2000. Digest of Technical Papers. ISSCC. 2000 IEEE International*. 2000.

[76] Agarwal, A., L. Hai, and K. Roy. *DRG-cache: a data retention gated-ground cache for low power*. in *Design Automation Conference, 2002. Proceedings. 39th*. 2002.

[77] Maziar, G., Tohru, I., and Hamid, N., *Variation-aware software techniques for cache leakage reduction using value-dependence of SRAM leakage due to within-die process variation*, in *Proceedings of the 3rd international conference on High performance embedded architectures and compilers*. 2008, Springer-Verlag: Goteborg, Sweden.

[78] Homayoun, H., Makhzan, M., and Veidenbaum, A. *ZZ-HVS: Zig-zag horizontal and vertical sleep transistor sharing to reduce leakage power in on-chip SRAM peripheral*

*circuits*. in *Computer Design, 2008. ICCD 2008. IEEE International Conference on*. 2008.

[79] Kawaguchi, H., Itaka, Y., and Sakurai, T. *Dynamic leakage cut-off scheme for low-voltage SRAM's*. in *VLSI Circuits, 1998. Digest of Technical Papers. 1998 Symposium on*. 1998.

[80] Kim, C.H., Jae-Joon, K, Mukhopadhyay, S., and Roy, K. *A forward body-biased-low-leakage SRAM cache: device and architecture considerations*. in *Low Power Electronics and Design, 2003. ISLPED '03. Proceedings of the 2003 International Symposium on*. 2003.

[81] Islam, R., A. Brand, and D. Lippincott. *Low power SRAM techniques for handheld products*. in *Low Power Electronics and Design, 2005. ISLPED '05. Proceedings of the 2005 International Symposium on*. 2005.

[82] Michael, P., Se-Hyun, Y, Babak, F., Kaushik, R., and Vijaykumar, T. N., *Gated-VDD: a circuit technique to reduce leakage in deep-submicron cache memories*, in *Proceedings of the 2000 international symposium on Low power electronics and design*. 2000, ACM: Rapallo, Italy.

[83] Sharifkhani, M., and Sachdev, M., *Segmented Virtual Ground Architecture for Low-Power Embedded SRAM*. Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, 2007. **15**(2): p. 196-205.

[84] Zhang, K., Bhattacharya, U., Zhanping, C., Hamzaoglu, F., Murray, D., Vallepalli, N., Yih, W., Zheng, B., and Bohr, M., *SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction.* Solid-State Circuits, IEEE Journal of, 2005. **40**(4): p. 895-901.

[85] El-Dib, D.A., Abid, Z., and Shawkey, H. A. *Investigating an aggressive mode for drowsy cache cells*. in *Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on*. 2008.

[86] Mukhopadhyay, S., Mahmoodi-Meimand, H., Neau, C., and Roy, K. *Leakage in nanometer scale CMOS circuits*. in *VLSI Technology, Systems, and Applications, 2003 International Symposium on*. 2003.

[87] Sylvester, D., Blaauw, D., and Karl, E., *ElastIC: An Adaptive Self-Healing Architecture for Unpredictable Silicon.* Design & Test of Computers, IEEE, 2006. **23**(6): p. 484-490.

[88] Pramanick, A.K., and Kundu, S. *Design of scan-based path delay testable sequential circuits*. in *Test Conference, 1993. Proceedings., International*. 1993.

[89] Wang, X., Roy, S., and Asenov, A., *Impact of Strain on the Performance of high-k/metal replacement gate MOSFETs,* in *Proc. 10th Ultimate Integration on Silicon (ULIS 2009)*. 2009.

[90] Zhiyu, L., and Kursun, V., *Characterization of a novel nine-transistor SRAM cell.* IEEE Trans. Very Large Scale Integr. Syst., 2008. **16**(4): p. 488-492.

[91] Kobayashi, T., Nogami, K. ,Shirotori, T., and Fujimoto, Y., *A current-controlled latch sense amplifier and a static power-saving input buffer for low-power architecture.* Solid-State Circuits, IEEE Journal of, 1993. **28**(4): p. 523-527.

[92] T. Nirschl, B.W., and D. Schmitt-Landsiedel. *High speed, low power design rules for SRAM precharge and self-timing under technology variations*. in *Proc. 11th Int. Workshop Power and Timing Modeling, Optimization and Simulation*. 2001. Yverdon-les-Bains, Switzerland.

[93] Halupka, D., Sheikholeslami, A. *Cross-coupled bit-line biasing for 22-nm SRAM*. in *Research in Microelectronics and Electronics, 2009. PRIME 2009. Ph.D.* 2009.

# Appendix 1: Acronyms

SNM Static Noise Margin

WNM Write Noise Margin

HNM Hold Noise Margin

BTBT Band to Band Tunnelling

RDD Random Discrete Dopants

LER Line Edge Roughness

PoG Polly Granularity

SCE Short Channel Effects

DIBL Drain Induced Barrier Lowering

SiO2 Silicon Di-oxide

DRV Dynamic Retention Voltage

STD Standard Deviation

PDF Probability Distribution Function

SRAM Static Random Access Memory

DRAM Dynamic Random Access Memory

MOS Metal Oxide Semiconductor

PTM Predictive Technology Model

CMP Chemical Mechanical Planarization

OPC Optical Proximity Correction

LWR Line Width Roughness

NBTI Negative Bias Temperature Instability

HCI Hot Carrier Effect

SET Single Event Transient

SEU Single Event Upset