# Low-Power Soft-Error-Robust Embedded SRAM

by

Jaspal Singh Shah

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Jaspal Singh Shah

# Abstract

Soft errors are radiation-induced ionization events (induced by energetic particles like alpha particles, cosmic neutron, etc.) that cause transient errors in integrated circuits. The circuit can always recover from such errors as the underlying semiconductor material is not damaged and hence, they are called soft errors. In nanometer technologies, the reduced node capacitance and supply voltage coupled with high packing density and lack of masking mechanisms are primarily responsible for the increased susceptibility of SRAMs towards soft errors. Coupled with these are the process variations (effective length, width, and threshold voltage), which are prominent in scaled-down technologies. Typically, SRAM constitutes up to 90% of the die in microprocessors and SoCs (System-on-Chip). Hence, the soft errors in SRAMs pose a potential threat to the reliable operation of the system.

In this work, a soft-error-robust eight-transistor SRAM cell (8T) is proposed to establish a balance between low power consumption and soft error robustness. Using metrics like access time, leakage power, and sensitivity to single event transients (SET), the proposed approach is evaluated. For the purpose of analysis and comparisons the results of 8T cell are compared with a standard 6T SRAM cell and the state-of-the-art soft-error-robust SRAM cells. Based on simulation results in a 65-nm commercial CMOS process, the 8T cell demonstrates higher immunity to SETs along with smaller area and comparable leakage power. A 32-kb array of 8T cells was fabricated in silicon. After functional verification of the test chip, a radiation test was conducted to evaluate the soft error robustness.

As SRAM cells are scaled aggressively to increase the overall packing density, the smaller transistors exhibit higher degrees of process variation and mismatch, leading to larger offset voltages. For SRAM sense amplifiers, higher offset voltages lead to an increased likelihood of an incorrect decision. To address this issue, a sense amplifier capable of cancelling the input offset voltage is presented. The simulated and measured results in 180-nm technology show that the sense amplifier is capable of detecting a 4 mV differential input signal under dc and transient conditions. The proposed sense amplifier, when compared with a conventional sense amplifier, has a similar die area and a greatly reduced offset voltage. Additionally, a dual-input sense amplifier architecture is proposed with corroborating silicon results to show that it requires smaller differential input to evaluate correctly.

## Acknowledgements

First of, I would like to express my deepest sense of gratitude to Professor Manoj Sachdev. Professor Sachdev has always been an invaluable source of guidance and encouragement. I am thankful for his persistence and the tremendous amount of faith he put in me which has always motivated me to do better than the best of my abilities. The level of attention given by my co-supervisor Professor David Nairn had a profound impact on the work. Thank you for your feedback and support, which has always aimed at making it better. I would also like to thank Professor Jim Martin, Professor Ajoy Opal, and Professor Karim Karim for serving on my Ph.D committee. Professor Martin, thanks for the effort that you put in understanding the work that is outside your field. Professor Opal, thanks for your insightful questions and comments. Professor Karim, I must thanks you for all the discussions and the guidance on matters are not related to research.

A sincere thank goes to Professor Bruce Cockburn from the University of Alberta, as an external examiner and for contributing your time, sharing your wealth of knowledge, and making a trip to Waterloo. Thanks a lot for all the positive comments and feedback.

I am grateful to Dr. Miachael Trinczek of TRIUMF for his help in irradiating the test chip. Thanks to Paul, Phil, Fernando, and Steve for the computing resource support. Thanks to Phil for responding to my Friday mid-night emails as a first thing on Saturday mornings to resolve Cadence issues which have a knack for showing up before tape-outs. The administrative issues were addressed efficiently, thanks to Annette, Deobrah, Diana, Karen, Lisa, Mary, Sarah, and Wendy Boles. A special thanks goes to the Canadian Microelectronic Corporation for providing access to a great set of tools to do research and for not updating the library kits on time which aided in extending my stay in the degree.

I am privileged to have worked with the past and current members of the CDR group. Your contributions are acknowledged and I am especially grateful for all the technical and non-technical discussions. You all have enriched my experience at Waterloo in multiple ways. Many people have contributed indirectly to this work in a way of being great friends, thoughtful moments shared with Jagtar, Amit, Rohit, Geetu, Jatinder, Karan, and Noman are acknowledged.

A very special thanks to my wife Dr. Inderpreet Kour Shah for her understanding during the past few years. Her encouragement, love and support have made this thesis

possible. She has put in so much sacrifice being at home a lot alone. I want to acknowledge my son Harnamit for being without me on beautiful evenings, missing walks and trips to the park while I was working at the university. Also, I want to thank my brother for providing encouragement whenever it was needed.

I would not be where I am today without the immense support and vision of my parents. They have always been supportive of my endeavours. My father, whose hardwork has always been a tremendous source of inspiration and motivation. I can not thank you enough.

## Dedication

To my parents and my beloved wife.

# Contents

# List of Tables

# List of Figures

# List of Symbols and Abbreviations

| | |
|---|---|
| $\beta$ | Cell Ratio or CR |
| $\Delta$V | Differential Voltage |
| $\gamma$ | Pull-up Ratio or PR |
| $\overline{BL}$ | Complementary Bit-line |
| $BL$ | Bit-line |
| $C_{BL}$ | Bit-line Capacitance |
| $hV_{th}$ | High Threshold Voltage |
| $I_{leak}$ | Leakage Current |
| $I_{read}$ | Read Current |
| $Q_{col}$ | Collected Charge |
| $Q_{crit}$ | Critical Charge |
| $sV_{th}$ | Standard Threshold Voltage |
| $T_{acc}$ | Access Time |
| $V_{BL}$ | Bit-line Voltage |
| $V_{DD}$ | Supply Voltage |
| $V_{gs}$ | Gate-to-Source Voltage |

| | |
|---|---|
| $V_{OS}$ | Offset Voltage |
| $V_{th}$ | Threshold Voltage |
| $V_{WL}$ | Word-line Voltage |
| $Y_{MUX}$ | Data acquisition Phase or Column Multiplexer Enable |
| 10T SRAM cell | Quatro Bit-cell |
| 6T | Six-Transistor SRAM |
| 6T Bit-cell | Six-Transistor SRAM |
| 8T | Eight-Transistor Soft-Error-Robust SRAM Bit-cell |
| 8T Bit-cell | Eight-Transistor Soft-Error-Robust SRAM Bit-cell |
| 8T SER SRAM | Eight-Transistor Soft-Error-Robust SRAM Bit-cell |
| BLB | Complementary Bit-line |
| CF | Calibration Factor |
| CLK | Clock or Input Clock |
| CLSA | Current Latch-type Sense Amplifier |
| CONV | Conventional Sense Amplifier |
| CQFP | Ceramic Quad Flat Package |
| CSA | Classic Sense Amplifier |
| DICE | Dual-Interlocked Storage Cell |
| DICLSA | Dual-Input Current Latch Type Sense Amplifier |
| DICSA | Dual-Input Classic Sense Amplifier |
| DILSA | Dual-Input Voltage Latch Type Sense Amplifier |

| | |
|---|---|
| DRV | Data Retention Voltage |
| DWL | Divided Word-line |
| ECC | Error Correction Circuits/Codes |
| EN | Enable |
| EV | Evaluation Phase |
| FIT | Failures In Time |
| GIDL | Gate-Induced Drain Leakage |
| HWD | Hierarchical Word Decoding |
| JEDEC | Joint Electron Device Engineering Council |
| LET | Linear Energy Transfer |
| LSA | Voltage Latch Type Sense Amplifier |
| MBU | Multi-bit Upset |
| NM | Neutron Monitor |
| PC | Pre-charge Signal |
| PCB | Printed Circuit Board |
| PD | Pre-discharge Phase |
| Quatro | Quad-Node Storage Cell |
| R/W | Read or Write Operation |
| RBB | Reverse Body Bias |
| SA | Sense Amplifier |
| SAE | Sense Amplifier Enable Signal |

| | |
|---|---|
| SAOC | Offset Cancelling Sense Amplifier |
| SE | Soft Error |
| SER | Soft Error Rate |
| SET | Single Event Transient |
| SEU | Single Event Upset |
| SNM | Static Noise Margin |
| SOC | System-On-A-Chip |
| SOI | Silicon-On-Insulator |
| SRAM | Static Random-Access Memory |
| TNF | TRIUMF Neutron Facility |
| TRIUMF | Tri-University Meson Facility |
| VSA | Voltage-Mode Sense Amplifier |
| WLE | Word-line Enable Signal |

# Chapter 1

# Introduction

## 1.1 Problem Statement

As systems-on-a-chip are becoming more and more memory-intensive, new applications require larger memory modules which occupy considerable area of the chip. To reduce the chip area, embedded memories (e.g., static random-access memory (SRAM)) are designed with state-of-the-art technology with minimum feature sizes. Due to nano-scale dimensions and reduced operating voltages, advanced semiconductor technologies have become more sensitive to radiation-induced transients. The sources of radiation are energetic neutrons from cosmic rays and alpha particles from the chip packaging materials [1]. These particles free electron-hole pairs as they pass through a semiconductor device [2]. The high field present at the p/n junctions can efficiently collect the particle-induced charge. The collected charge ($Q_{col}$) leads to a transient which is called single event transient (SET). When such a transient causes sufficient charge to be stored in a memory cell, a latch, a flip flop, or a register a single event upset (SEU) occurs. Since the SET or SEU does not permanently damage the device, it is referred to as a Soft Error (SE).

The soft error does not damage the device, but it is a potential threat to the reliable operation of the system. The first paper to discuss the effect of SEUs was in 1962 by Wallmark, et al. [3]. The effect of technology scaling in terrestrial microelectronics was also reported. The first confirmed anomaly in space electronics due to cosmic rays was reported in 1975 in bipolar J-K flip flops by Binder, et al, [4]. In 1978, phenomenon of soft errors in dynamic random access memories was first reported by May and Woods [5]. The source of SE was identified to be alpha particles emanating from contaminations (uranium) in the packaging materials. The semiconductor manufacturing process and the packaging materials have been purified to a point of diminishing returns. Concrete has been shown to reduce the radiation rate by 1.4x per foot, however, this is not a practical solution [6]. In 2005, Hewlett-Packard acknowledged that a large installation base of 2048-CPU server system in Los Alamos National Laboratory crashed frequently because of cosmic ray strikes to its parity-protected cache array [7]. A similar crash in implantable cardioverter defibrillator, space-borne electronics, aircraft controllers or other mission critical applications can endanger human life. As CMOS technology scales into the sub-100nm regime, less charge is used to store the data owing to smaller node capacitance and lower operating voltage ($Q = CV$). Thus, a particle strike with a relatively small transferable energy can cause an error. As such this problem warrants the need for soft-error-tolerant designs. The SEUs and multi-bit upsets (MBUs) are a major reliability concern in commercial electronics as reported by Texas Instruments [6], Hewlett-Packard [7], Intel [8], Cypress Semiconductors [9], Virage Logic [10], Boeing [11], and IBM [12].

## 1.2 Radiation Effects on Microelectronics

The smaller transistor dimensions and reduced operating voltage has led to an increased sensitivity of integrated circuits to ionizing radiations [1]. The magnitude of the disturbance that an ionizing radiation can cause depends upon the linear energy transfer (LET) of that ion. In other words, the particle loses energy as it traverses through the material and the energy loss of the particle or the energy transfer to the material is a function of the distance it travels through the material and the density of the material. Thus, the LET is reported in energy lost per unit length per unit mass density, i.e., $(MeV/cm)/(mg/cm^3)$ or $MeV - cm^2/mg$. The phenomenon of charge generation and collection can be explained with the help of Fig. 1.1 (adapted from [6]). Fig. 1.1(a) shows that at the onset of the event, electron/hole pairs are generated in high concentrations along the path of the ion. In Fig. 1.1(b), the electrons and holes are collected by the drift mechanism due to the high electric field of the depletion region. In Fig.1.1(c), the charge collection is completed by diffusion. The charge collection through diffusion continues (hundreds of picoseconds to nanoseconds) until all of the excess charge carriers have been collected, recombined or have diffused away from the junction area. The corresponding current pulse is shown in Fig. 1.1(d) [6], [2].

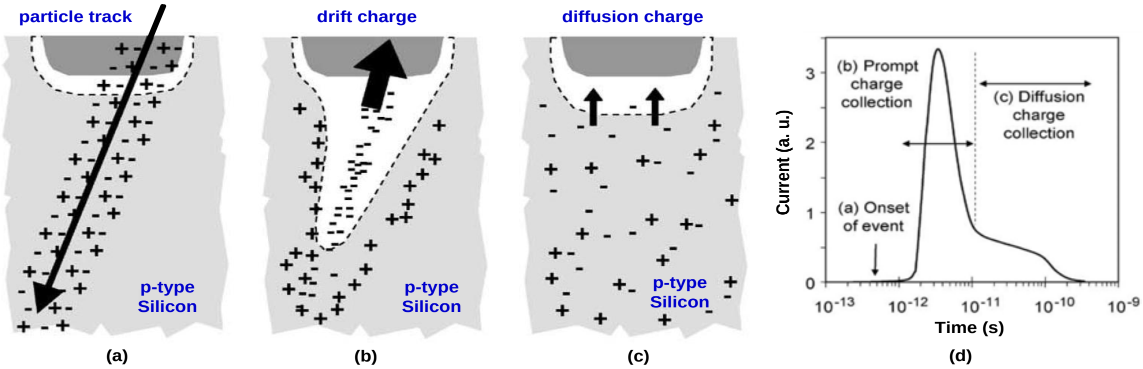


Figure 1.1: Charge generation and collection events at a reverse-biased $p-n$ junction after a particle strike and the resulting current at the collection node.

In general, the farther away from the junction that a particle strikes, the smaller the amount of charge that will be collected and the less likely that the event will cause a soft error. In actual circuits, a node is never isolated but is a part of a complex sea of nodes in close proximity to one another. Thus, charge sharing among nodes and parasitic bipolar action (the formation of an unintentional bipolar transistor between junctions and wells) can greatly influence the amount of charge collected. In fact, the magnitude of collected charge depends on a complex combination of factors including the size of the device, the biasing of various circuit nodes, substrate doping, the type of the particle, its energy, the position of the event, and the state of the device.

The collected charge does not result in a soft error until it exceeds a critical charge ($Q_{crit}$), which is defined as the minimum charge required to trigger a change in the data state [7]. Thus, for an event of particle strike, if $Q_{col} > Q_{crit}$, a soft error will result; otherwise the circuit will survive the event and no soft error will occur. Therefore, the critical charge can be used as a figure of merit to assess the soft error susceptibility. However, the critical charge is not constant since the response of the device to the charge injection is dynamic and dependent on the magnitude as well as the temporal characteristics of the pulse [1]. Consequently, the critical charge becomes a function of the node capacitance, operating voltage, and the strength of the restoring or feedback mechanisms connected to the node, making it difficult to model [8].

The soft error rate (SER) is generally measured in FIT (Failures In Time). One FIT means 1 failure per 1x$10^9$ hours of the device operation. Typically, the failure rates for hard failure mechanisms such as latch up, gate-oxide breakdown etc. add up to 1-500 FIT. The SER can easily exceed 50,000 FITs per chip [13] which is the highest failure rate of all the reliability mechanisms [14].

### 1.2.1 Soft Errors in Logic Circuits

A single event transient in a logic circuit can affect the computation in two ways. One, it can get latched in a memory element. Two, the transient can lead to faulty logic evaluation which can further result in error propagation. The propagated errors can eventually get latched leading to multiple SEUs. In combinational circuits there are a few phenomena which mask the SETs.

#### Logical Masking

An SET at a node in a combinational circuit will not affect the evaluation if it is not a controlling input for the logic gate. As shown in Fig. 1.2 for a NAND gate, the transient affects the input B of the gate while the input A is at logic 0. The transient does not affect the output Y. In the same figure, for a NOR gate, the transient affects the input C of the the gate while the input D is at logic 1. The output Z is not affected by the transient. In this situation, the error is said to be logically masked.

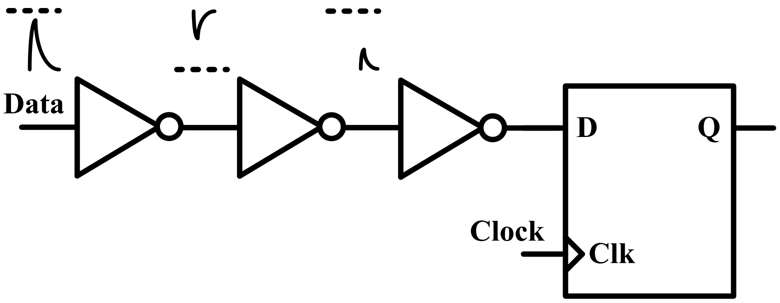
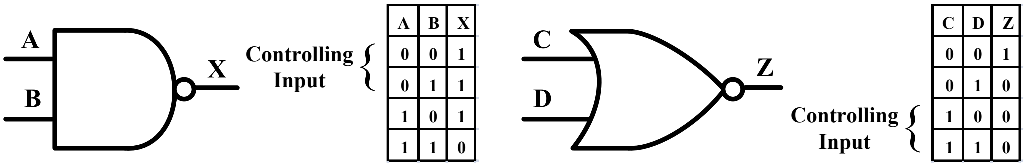


Figure 1.3: Electrical Masking



| A | B | X |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

| C | D | Z |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 0 |

Figure 1.2: Logical Masking

### Electrical Masking

An SET can get attenuated while passing through different stages of the combinational logic due to electrical properties of the gates. In particular, an SET of duration greater than the gate delay will propagate with attenuation. This phenomenon is called electrical masking. Fig. 1.3 shows pulse attenuation produced through a chain of inverters.

### Latching Window Masking

Even if a transient propagates through the logic towards a storage element without significant attenuation, it may not result in an error in a case when the pulse reaches the input of a flip flop outside the latching window. This effect is called latching window masking. The period during which the latch is sensitive to the pulse is called the window of vulnerability [15]. Sometimes, the term timing masking is used to explain a similar phenomenon for the dynamic logic. Dynamic logic operates in two phases during a clock cycle namely, pre-charge and evaluation. During the pre-charge phase, the output node is driven by a power supply. During evaluation the output is driven conditionally to ground if the pull down network evaluates to a logic 1. The output node is less susceptible to SETs during pre-charge phase as output is constantly driven by the supply rails.

The masking effects described above decrease the soft error rate in combinational logic. However, in the nanometric regime, shrinking feature sizes and increased pipeline depths diminish these making effects. Electrical masking is reduced as scaled transistors are faster and cause less attenuation on the pulse. At higher clock rates, the latches cycles more frequently, which can reduce the latching window masking [16].

### 1.2.2 Soft Error Detection Techniques in Logic Circuits

The soft errors in logic circuits can be detected in a few ways such as space and time redundancy. In space redundancy, the same logic operation is performed twice using independent hardware. The output of each stage is latched and compared with a parity circuit to indicate an error. In time redundancy, the output is sampled and latched at two time intervals separated by a certain delay. The delay is chosen such that it is less than the pulse width of an SET [17]. The sampled results are compared with a parity circuit to determine an error. A typical problem with the use of parity circuits is that it will not indicate an even number of errors. Space redundancy leads to an extra hardware and hence, it has a significant area and energy overhead. In time redundancy, the penalty is minimal for area, but higher for the delay, as the system speed has to increase to incorporate the sampling interval. Triple modular redundancy is an error correction technique where the logic operation is triplicated in hardware and a majority voter circuit determines the correct output. A major drawback of this approach is high area and energy footprint. Also, SETs affecting the voter circuit can still lead to an incorrect decision. In all of the detection techniques if the SET affects the input, it may still lead to incorrect logic evaluation.

### 1.2.3 Soft Errors in SRAMs

A typical six-transistor SRAM bit-cell (6T SRAM or 6T bit-cell or 6T) stores data in a cross coupled inverter pair. When an energetic particle strikes a sensitive node (reverse-biased drain junction) in an SRAM cell, as shown in Fig. 1.4 (adapted from [2]), the charge collected by the junction results in a transient current in the struck transistor($N_2$). As this current flows through the struck transistor, the restoring transistor ($P_2$) sources current in an attempt to balance the particle-induced current. The current flow through the restoring transistor therefore induces a voltage drop at its drain. This voltage transient in response to the the single event current transient is the mechanism that causes an upset in the SRAM cell [2].

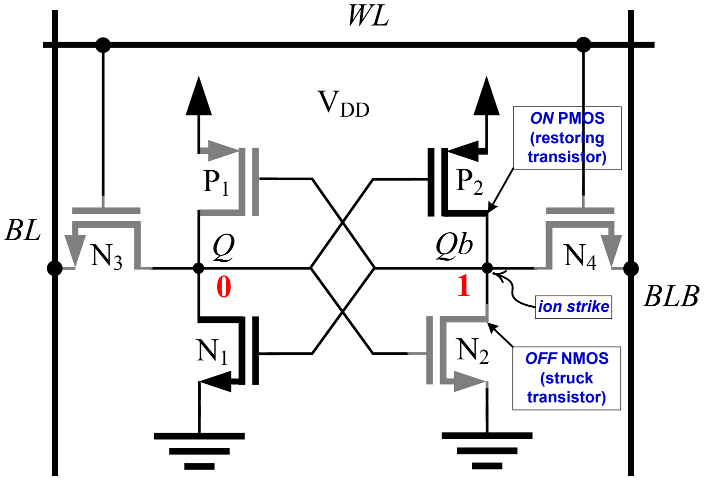


Figure 1.4: SEU in a 6T SRAM bit-cell.

In an SRAM cell, there are four possible sensitive locations i.e., the four transistor drains. The charge collection mechanism is different depending upon whether the junction is located within a well or a substrate. The well-substrate junction provides a potential barrier that prevents any charge deposited within the substrate from diffusing back to the struck drain. For a struck drain which is not in a well, the charge deposited in the substrate can diffuse back to the drain junction. Thus, the reverse-biased junction which is not in a well is the most sensitive part of the circuit.

### 1.2.4 Mitigation of Soft Error in SRAMs

The soft error mitigation techniques in SRAM can be classified into three categories: process, circuit, and architecture.

### Process Techniques

The objective of process techniques is to decrease the sensitivity of the charge collecting nodes. This can be achieved in a variety of ways. Increased doping of the p-well [18] results in a reduced charge collection. The use of a triple well has shown an improvement in the SER for BiCMOS process [19] by limiting the charge collection; however, in CMOS use of a triple well has shown degradation in the SER FIT rate because of the increase in collection volume for holes [20]. A careful placement of n-well and p-well contacts and an increased area has shown SER improvements in the triple well process [20]. Use of the silicon-on-insulator (SOI) process reduces the charge collection in the substrate. A 5x reduction in the SER of SRAM devices is reported by using SOI technology [21]. The use of the modified process offers some benefits at the expense of an increased cost and, sometimes, volume manufacturing is not feasible.

### Circuit Techniques

The SER can be reduced either by slowing down the response of the circuit to the SETs or by increasing the critical charge of the sensitive nodes. This approach involves adding a resistor to the feedback path [22] or a coupling capacitor between the sensitive nodes [23]. SRAM cell with redundant nodes has also been proposed [24], [25] which restores the logic at the node through feedback. These techniques are explained in detail in section 2.8. The use of circuit techniques generally involves area overhead.

### Architecture Techniques

From the architecture perspective, a soft error may not be a problem if the SRAM cell undergoes a write operation before a read operation. Most recent attempts focus on the use of parity circuits to detect single bit errors and then correct them with error correction circuits/codes (ECC) [26]. The ECC being a reactive approach is used once the error has occurred, but it can detect double bit errors with added complexity. Multiword ECC is also reported in the literature which offers reduced energy consumption [25]. Moreover, ECC implementation is prohibitive for multi-bits errors (higher cost of implementation) and cannot be used in areas such as L1 cache (speed constraints) and FPGA configuration memory (distributed nature). With an added design complexity, bit interleaving is another approach which helps in the detection and the correction of errors with the aid of the ECC.

## 1.3 Scaling and Soft Errors

The relationship between the process technology and the soft error rate is illustrated in Fig. 1.5 (adapted from [6]). It shows that with the advent of scaled technologies, soft errors have become a significant problem. The shrinking of device sizes as the manufacturing process advances from one process to another has some mitigating effects. As the die area occupied by a given memory cell decreases with decreasing feature size, as shown in the Fig. 1.6a (adapted from [27]), the probability that a given memory cell being struck by a transient also decreases. However, this is offset by the increase in the density of memory cells. Also, the lower energy needed to upset the cell outweighs the lower probability of an individual element being struck. Moreover, smaller feature sizes increase the probability of MBUs as shown in the Fig. 1.6b (adapted from [27]). The net result is that the probability of SEUs increases in finer geometries and smaller feature sizes [28].

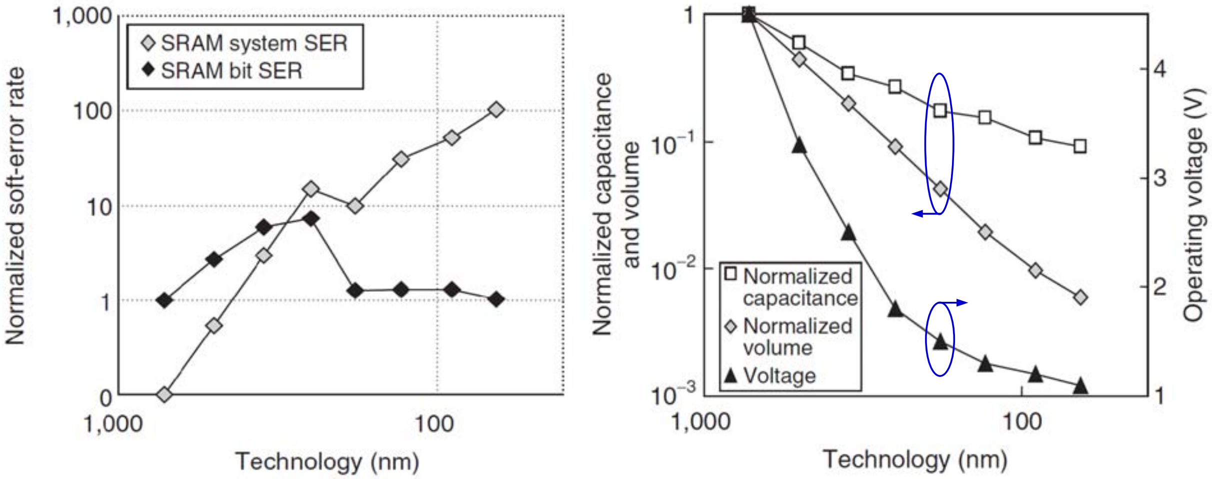
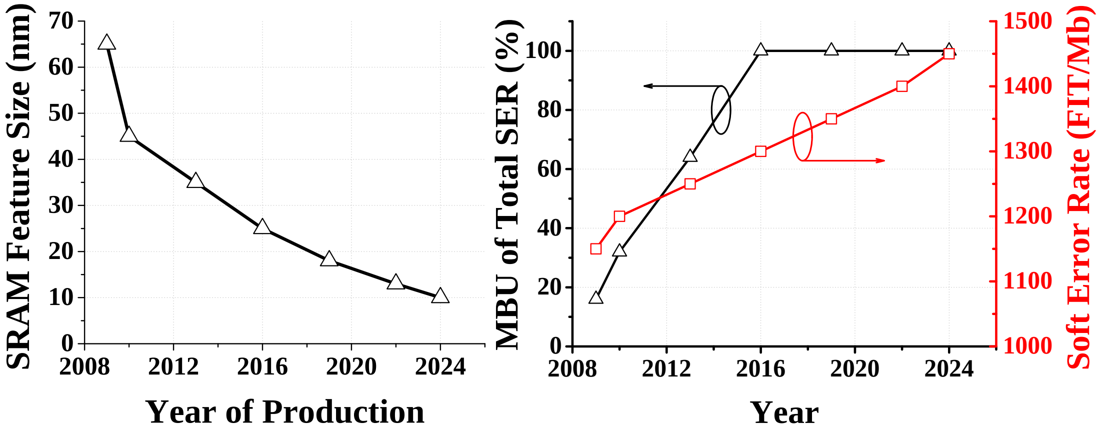


Figure 1.5: SRAM scaling trends: SRAM single-bit and system SER, node capacitance and operating voltage as a function of technology node.



(a) SRAM feature size

(b) SER and MBU

Figure 1.6: SRAM feature size, SER and MBU as a function of time.

Voltage scaling is a technique commonly used to reduce the dynamic power consumption. The technique is also used to reduce the leakage power. In caches, the supply can be reduced while ensuring data stability [29] and is powered up before it is accessed. In a custom implementation [30], a 25% decrease in supply voltage resulted in a 20% reduction in $Q_{crit}$ and a 35% reduction in leakage power. Hence, with a reduced supply voltage there is a loss in immunity to the soft errors.

Higher threshold voltage ($hV_{th}$) devices are often employed to reduce the leakage power. Due to the properties of the $hV_{th}$ transistors, higher energy is required to create electron-hole pairs in the substrate and hence the device is more immune to soft errors [31].

## 1.4 Goal of This Research

In the scaled technologies, memory and logic are more sensitive to SETs due to higher packing density, smaller node capacitance and reduced supply voltage. In particular, SRAM is more vulnerable to soft errors than dynamic RAMs because of lower node capacitance. In modern microprocessors and system-on-a-chips (SoCs), SRAM-based memory elements are a major component of the die. Due to its large size, SRAM consumes a significant portion of the total power budget. Thus, the use of low-power techniques is imperative.

However, techniques such as reduction in data retention voltage [32], use of sleep transistors [33], [34] increase the SER [35]. This motivates the need to develop soft-error-robust memory cells. Typical approaches in the literature include new manufacturing processes such as SOI, increasing the critical charge of the memory element by resistive hardening [22] or capacitive hardening [23], but these ideas are constrained by scaling possibilities. The hardened by design approach, such as dual-interlocked storage cell (DICE) [24], SER register element [36], quad-node ten-transistor cell (Quatro) [25] is attractive because the ideas can be easily scaled between different technology nodes and the implementation is relatively cost-effective. Thus, in comparison a standard 6T, DICE and Quatro cells are considered. In literature, a number of eight transistors cells are proposed, but their objective is primarily read stability or low power, and hence they are not considered in this research. The main focus of this research is to develop a low power soft error robust SRAM cell. Further, voltage scaling techniques are investigated to achieve a balance between low power and soft error robustness.

During a read operation the stored information in a memory cell needs to be determined. Reading the data, which involves sensing a bit from an array, is an important part of embedded memory design. In the scaled technologies, the process-induced variations can cause two neighboring transistors to have different properties causing a sense amplifier (SA) to make an incorrect decision. A novel sensing scheme for SRAMs which can cancel the offset caused by process variations is proposed and implemented. In addition, the inputs of the sense amplifier are sourced by the bit-lines which are highly capacitive in nature. As a consequence, the time required to develop sufficient differential voltage increases as the size of the array increases. Thus, it becomes imperative to use a sense amplifier which demands smaller differential inputs to correctly identify the stored data. A sense amplifier architecture is proposed and implemented which requires a smaller differential input signal to sense correctly. Hence, this thesis presents a low voltage robust design – a soft error robust SRAM and process-aware peripheral circuits.

## 1.5 Outline

The rest of the thesis is organized in the following manner. Chapter 2 discusses SRAM architecture and operation. Additionally, existing soft-error-robust SRAM cells are analyzed. In Chapter 3, a proposed low-power soft-error-robust SRAM cell is presented. Further, simulation and measurement results are provided and analyzed. In Chapter 4, the proposed offset cancellation sense amplifier is presented along with analysis, simulation and measured results. Chapter 4 also includes an architecture of sense amplifiers which requires smaller differential inputs to make a correct decision in the wake of process variations. Finally, Chapter 5 concludes with future research work that lies ahead in the proposed thesis.

# Chapter 2

# SRAM Architecture and Circuits

The static random access memory is capable of storing a large quantity of digital data. The amount of memory required depends upon the type of application. In general, the number of transistors required for data storage is much larger than the transistor count required to implement the logic and other operations. The increasing demand for larger memory capacity has in part led to more compact design rules for manufacturing. The number of stored data bits per unit area determines the memory cost per bit. The access time, which is the time required to store or to read from a memory location, determines the speed of the memory. The static and dynamic power consumption of the bit-cell are additional important factors while designing the memory circuits.

In this chapter, the basic SRAM architecture is described. The peripherals surrounding an SRAM array are explained and various design choices are evaluated.

## 2.1 Architectural Overview

An SRAM array is shown in Fig. 2.1 of size $n$ x $m$, where $n$ is the number of words and $m$ is the number of bits per word. The figure indicates the inputs for a synchronous, single port memory: the input clock (CLK), the address bus provides the memory address for a read or write operation, a control signal specifying the read or write operation (R/W), a

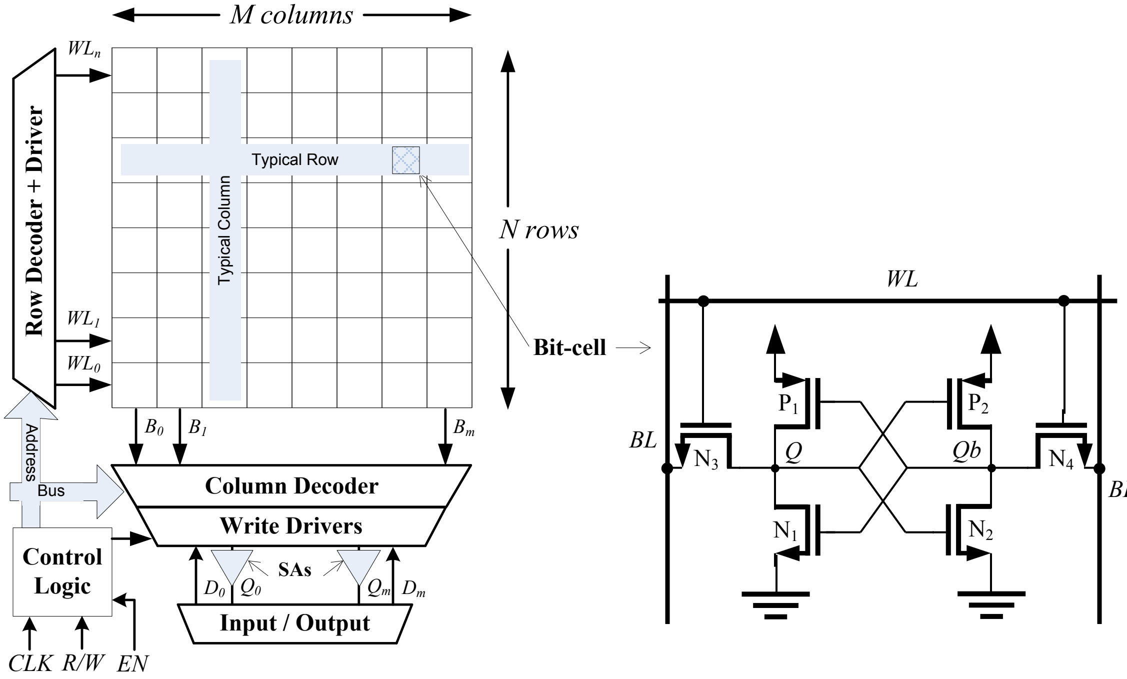


Figure 2.1: A typical 6T SRAM array and bit-cell.

memory enable (EN) to provide access to the memory block for a memory operation at the CLK edge, $WL_i$ and $B_i$ which represents the intermediate word line and bit-line signals, and $D_i$ and $Q_i$ which constitute data input and memory output signals respectively.

An SRAM block consists of several peripheral units such as row and column decoders, row and column drivers, sense amplifiers, input output storage units and the control logic. The row decoder decodes the binary encoded input address to a physical location within the array. Usually, a group of cells are selected in a given R/W operation. The number of cells selected is determined by the word length of the design which is typically <128 bits. The address of the selected word in a block is decoded by the column decoder. For example, a row with $m$ columns will have $m/32$ words of 32 bits each. Thus, the column decoder needs $log_2(m/32)$ address bits. The control logic generates the timing signals necessary to initiate communication with the memory block such as block select, address decode, word line activation, and read or write operation. In the following sub-sections the constituents of an SRAM block are discussed.

## 2.2 SRAM Cell

The memory bit-cell consists of two inverters connected back to back, and two complementary access transistors. As long as the power supply is available, the cell preserves one of the two possible states of the data. The design of a 6T CMOS SRAM cell involves balancing a number of design criteria. The most significant requirement of a typical memory cell is that a read operation should not destroy the stored information and during a write operation the stored information can be modified.

### 2.2.1 Read Operation

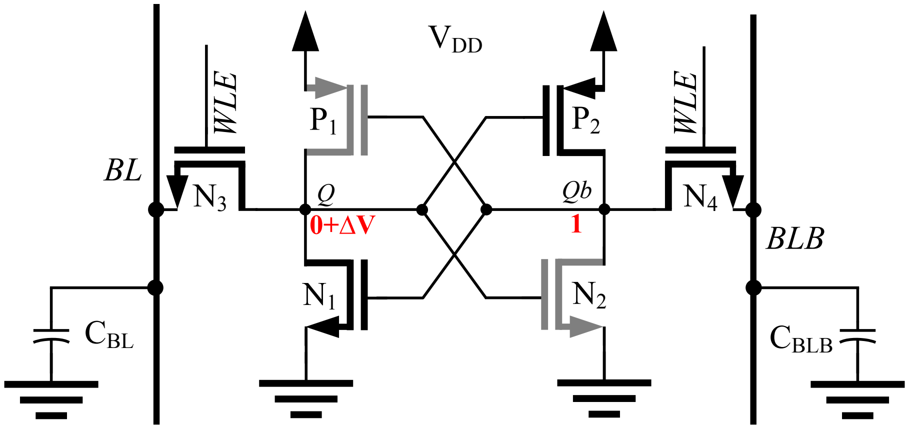


Figure 2.2: Voltage levels in the 6T SRAM cell at the beginning of the read operation.

The Fig.2.2 shows the read operation of the cell. In this figure, the cell is storing a logic 0 at node $Q$ and logic 1 at node $Qb$. Thus, the gray transistors $N_2$ and $P_1$ are off, while the transistors $N_1$ and $P_2$ operate in linear mode. At the beginning of the read operation the bit-lines are pre-charged to logic 1. The access transistors $N_3$ and $N_4$ are turned on by the word-line enable signal (WLE) which belongs to row selection circuitry. The bit-line voltage

$V_{BL}$ is discharged through the transistors $N_1$ and $N_3$ connected in series. The transistors $N_1$ and $N_3$ form a voltage divider whose output (node $Q$) is no longer at zero volts. In other words, $N_3$ and $N_1$ conduct a nonzero current discharging the bit-line capacitance (also called the column capacitance). While discharging the column capacitance the voltage at node $Q$ increase from its initial value of 0 V to 0+ΔV. This voltage drives the input of inverter $N_2$-$P_2$. The key design issue during a read operation is to ensure that the raised voltage at node $Q$ does not exceed the threshold voltage ($V_{th}$) of $N_2$. This is determined by the cell ratio (CR or $\beta$). The $\beta$ is given by the aspect ratio of the driver transistor ($N_1$) to the access transistor ($N_3$).

$$\beta = \frac{W_{N_1}/L_{N_1}}{W_{N_3}/L_{N_3}} \tag{2.1}$$

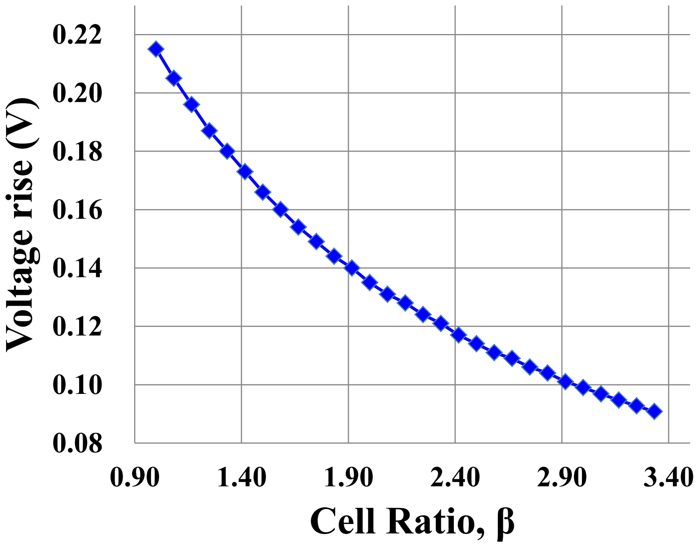


Figure 2.3: Voltage rise in the cell at the node holding a 0 during read versus cell ratio.

The dependence of differential voltage (ΔV) developed between bit-lines on $\beta$ is shown in Fig. 2.3. To ensure a non destructive read, the $\beta$ is usually kept greater than 1 and can be varied depending upon the target application. A larger $\beta$ provide a higher read current and hence a higher speed at the expense of a larger cell area. A typical sizing approach is to keep the access transistors of minimum size and of slightly larger than the minimum length and width of the driver transistors [37]. Once a sufficient bit-line ΔV is developed, the sense amplifier circuitry can amplify it to a full scale output signal.

A similar argument dictates the aspect ratios of $N_2$ and $N_4$. During a read operation, the WLE is activated for a limited duration as determined by the read access time. The read operation is successful if a pre-charged bit-line is discharged by a value ΔV large enough to trigger the sense amplifier within the WLE duration.

### 2.2.2 Write Operation

The write operation involves writing a logic 0 at node $Qb$ which is storing a 1. The node $Q$ of the cell cannot be pulled high enough to write a 1 because of constraints imposed by read stability that ensures that a 0 node does not exceed the switching threshold of the inverter. The pull-up transistor ($P_2$) helps to maintain the high level on the node $Qb$ and prevents its discharge during data retention. Thus, to accomplish a successful write operation, logic 0 is written at node $Qb$ by pulling the node below the switching threshold of $N_2$-$P_2$. The condition for the successful write can be derived by writing out the dc current equation which involves the pull-up ratio (PR or $\gamma$).

The pull-up ratio is given by:

$$\gamma = \frac{W_{P_2}/L_{P_2}}{W_{N_4}/L_{N_4}} \tag{2.2}$$

The dependence of the voltage at node $Qb$, $V_{Qb}$ on PR is shown in Fig. 2.4. The lower PR leads to lower $V_{Qb}$, in order to pull the node below the $V_{th}$ of NMOS, the PR has to be below 1.9. Typically, a stronger write capability is achieved by making the pull-up device weaker that the access transistor. However, a stronger pull-up PMOS improves read stability. The read stability and write ability are thus two conflicting design requirements.

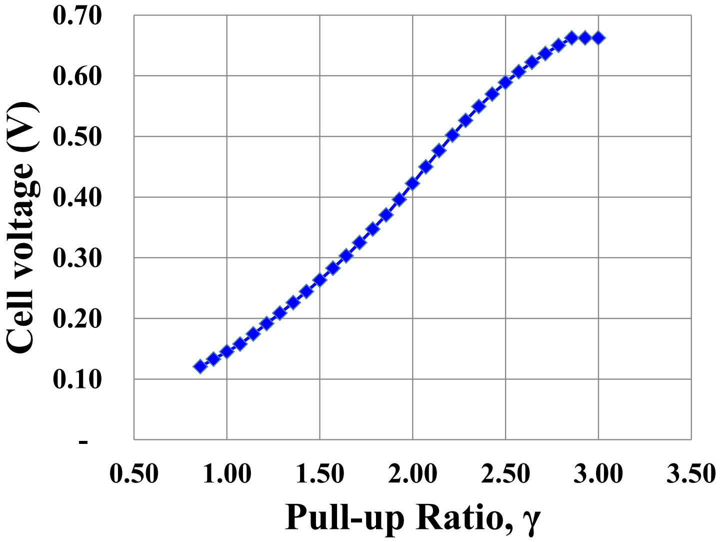
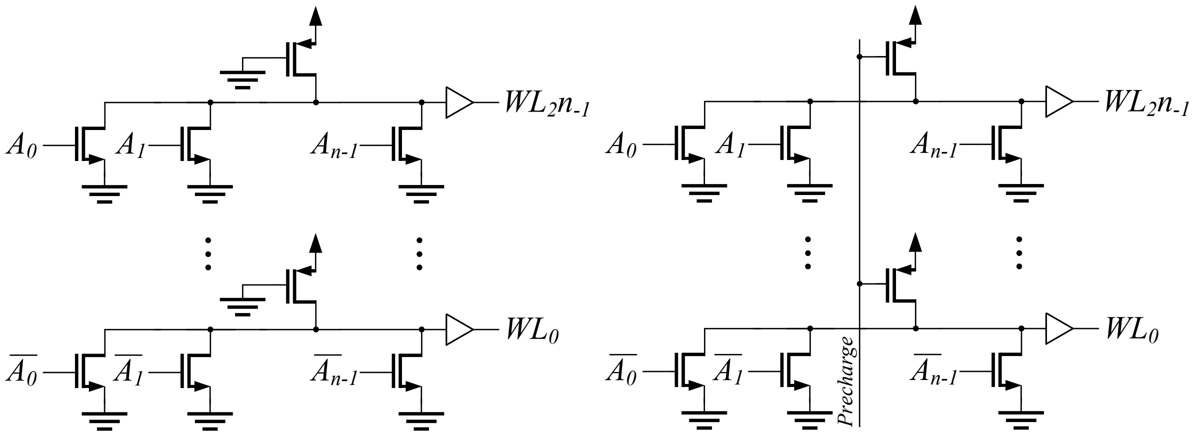


Figure 2.4: Voltage written into the cell versus pull-up ratio.

## 2.3 Row Decoder

The memory address space is defined as the total number of address bits required to access a particular bit or word. The address space depends upon the requirements of the implementation. For example, a 1-Mb SRAM if implemented in a bit-oriented fashion needs a 20-bit address space (1Mb = $2^{20}$); however, a word-oriented implementation with 32-bit word length ($2^5$) needs 15-bit address ($2^{20}/2^5 = 2^{15}$). Alternatively, a 64-bit word length ($2^6$) requires a 14-bit address. On the other hand, a 32-bit word length realization of 1-Mb can be executed as an organization of 32 blocks ($2^5$) where each block has 256 rows ($2^8$) and 128 columns or 4 words ($2^2$). The address space in this case will be 15-bits. The SRAM row decoder can be realized based on a single or multi-stage architecture. In a single stage decoder, all of the decoding is realized in a single block, such as a wide-NOR gate. The fan-in for the NOR gate equals the number of the address bits. To simplify the circuit and reduce the layout area, such decoders are designed using static PMOS transistor loads (Fig. 2.5a). Alternatively, the PMOS load can be clocked leading to a dynamic implementation (Fig. 2.5b). The wide NOR implementation has several challenges [37]. First, the layout of the wide NOR must fit within the word-line pitch. Second, the large fan-in impacts the propagation delay, thus the read/write access time is increased. Third, the gate has to drive the large load of the word-line while not overloading the input addresses. Fourth, the power dissipation has to be limited. Thus, the multi-stage decoder is a viable alternative.



(a) Static row decoder

(b) Dynamic row decoder

Figure 2.5: Single-stage wide-row decoders.

The multi-stage decoder architecture can have multiple flavours. One implementation is the Divided Word-line (DWL) structure shown in Fig. 2.6. The SRAM array is divided into blocks and a local or block-level word-line is asserted when both global word-line and block select are enabled. Since only one block is activated, the DWL structure reduces both word-line delay and power consumption [38]. By dividing the word-line into three or more levels, hierarchical word decoding (HWD) can be implemented as shown in Fig. 2.7. For example, a word-line can have hierarchical structure such as global word-line, sub-global word-line, and a local word-line. With HWD, the load capacitance is efficiently distributed resulting in reduction of both the delay time and the power consumption [39]. However, the HWD is better than DWL only for larger memories (>256 K) [39].

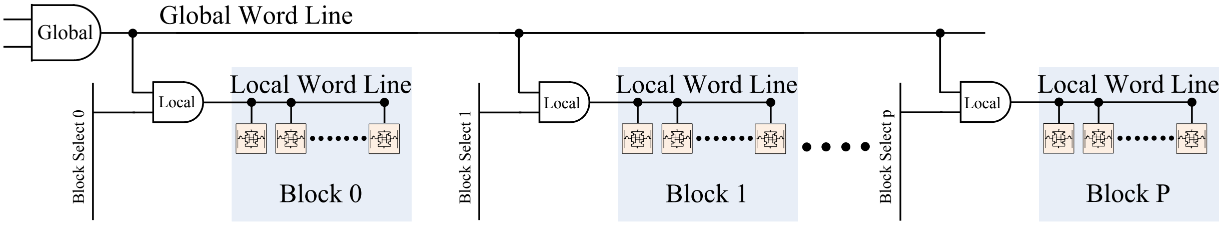


Figure 2.6: Divided word-line architecture.

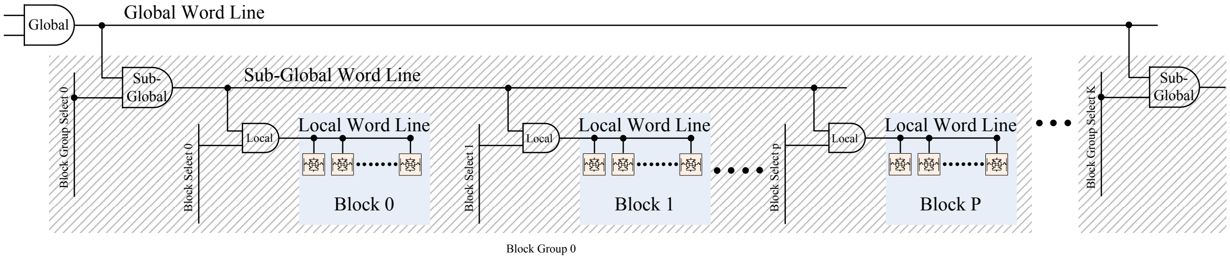


Figure 2.7: Hierarchical word-line architecture.

A conventional row decoder based on two stages, where the address bits are grouped to form two pre-decoders and a post decoder, is shown in Fig. 2.8. In this example, $A_3$, $A_2$ and $A_1$, $A_0$ represents two 2-to-4 decoders and the AND gate generates signal based upon the pre-decoded outputs. To generalize, in the pre-decoder, the first logic state is decoded and the post-decoder generates the final WL signal. The shortest delay in this case is realized when the address field is divided equally between two pre-decoders [40]. The post-decoder consists of a plurality of AND gates. Each AND gates has an input driven by each of the pre-decoder outputs, thus, at any given time only one of the post-decoder's output is high. An important constraint is that the pitch of the AND gate in the post-decoder must match the pitch of the row-driver as well as the height of the SRAM cell in the array.

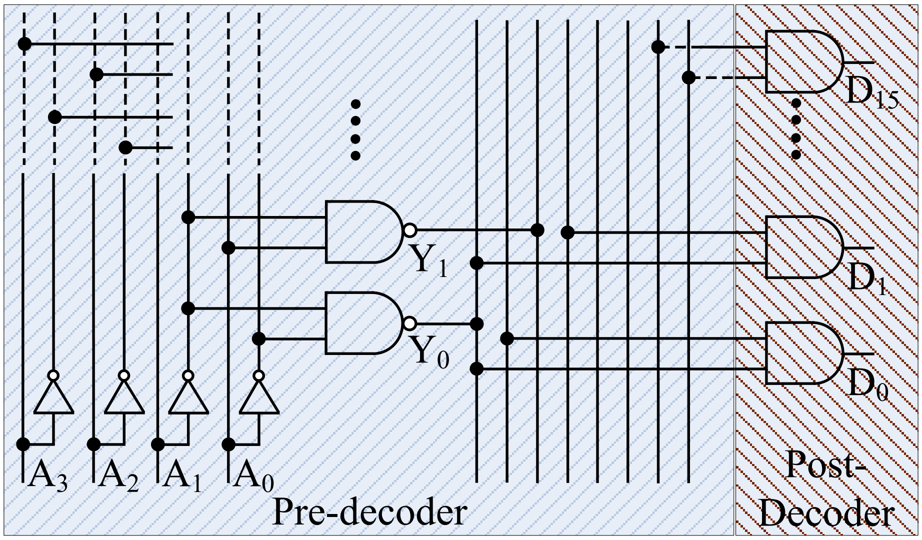


Figure 2.8: Two-stage 4-to-16 AND decoder.

## 2.4 Column Decoder

A column decoder or multiplexer, commonly known as $Y_{MUX}$, allows multiple columns to be connected to a single SA. Typically, $Y_{MUX}$ allows insertion of multiple words in a row which aids in balancing the aspect ratio of an SRAM block and helps to reduce the number of I/O circuits in the memory bank. Additionally, it reduces the bit-line capacitance at the expense of increased word-line capacitance. In a given memory access, a word is selected to perform a read/write operation.

A typical implementation of the $Y_{MUX}$ involves the use of pass transistors. It can also be implemented with a pre-decoder or a tree-based design. In a pre-decoder based implementation, decoded signals enables one column using pass transistors. Fig. 2.9 shows an example of 2-4 pre-decoder with PMOS pass-transistors. The $BL_i$, $BLB_i$ represents a typical column and one out of the four columns is connected to the read/write circuitry. The main advantage of this approach is the speed because only a single transistor is inserted in the signal path. The disadvantage of the structure is large transistor count e.g., $2^K$ input decoder needs $(K+1)2^K+2^K$ transistors. A tree decoder offers an efficient implementation based on the binary reduction such that $2^K$ input decoder needs $2(2^K-1)$ transistors. The advantage is that in the absence of pre-decoder, fewer devices are required. However, the propagation delay increases quadratically because of K-series connected transistors. It is interesting to note that in order to share the multiplexer between read and write operations, a complementary transmission gate should be used to allow a full logic swing.

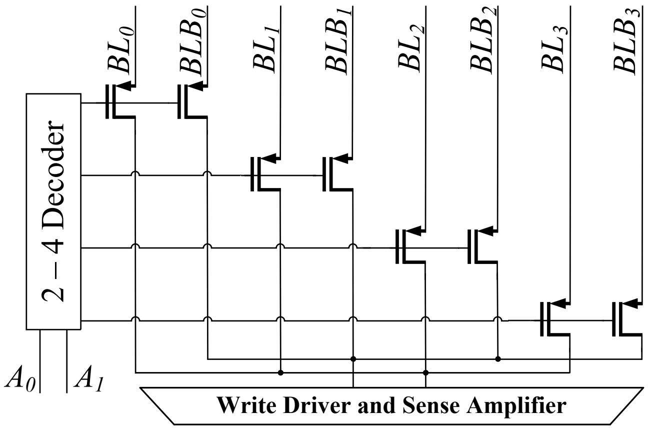
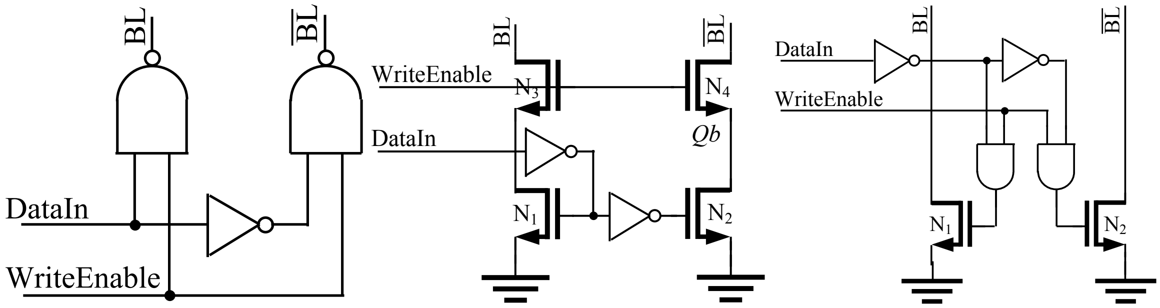


Figure 2.9: Column decoder with a 2-4 predecoder and PMOS pass transistors.

## 2.5 Write Driver

The write driver aids in writing into the SRAM cell. It pulls down one of the bit-lines from the pre-charge level to below the write margin based on the input data to be written. The write driver can be implemented in a few different ways. In Fig. 2.10a, at the onset of write enable bit-line $BL$ or $BLB$ is connected to ground based upon the input data. In Fig. 2.10b, NMOS transistors $N_1$ and $N_3$, and $N_2$ and $N_4$ are stacked for a pass transistor based AND gate. When WriteEnable is enabled, depending upon DataIn, bit-line $BL$ or $BLB$ is connected to ground. In Fig. 2.10c, when WriteEnable is asserted, one of the AND gates activates transistor $N_1$ or $N_2$, thus discharging the corresponding bit-line to ground.

Since only one driver is required to write in a column, it can be upsized if necessary with minimal impact on the overall area.



(a) NAND gate based driver (b) Stacked NMOS based driver (c) AND gate based driver

Figure 2.10: Different write driver circuits.

## 2.6 Timing and Control Circuits

In order to communicate with the memory, a timing block is required which generates and controls the signals such as pre-charge (PC), word-line enable, sense amplifier enable (SAE), and Read/Write. Accurate timing control is of paramount importance to keep the memory block in read or write or retention mode. Accurate timing generation is always a challenge as technology continues to scale down. Threshold voltage variation, process variation, and reduced over-drive impacts the cell as well as the peripheral circuits. The key timing hazards which should be avoided are:

- During a read operation, if WLE is assertion precedes the pre-charge deactivation, the selected SRAM cell will see both the bit-lines high and may flip state.
- A change in address state before the completion of the read operation can result in more than one SRAM discharging the bit-lines which may lead to incorrect decision by the sense amplifier.
- During a write operation, if an SA is enabled, the data being written would appear at the output resulting in a write through.
- In order to successfully read or write, the timing block should provide sufficient timing margins to account for process variations in the target yield.

Some of the timing methods which can be incorporated in the design to address timing issues are:

- Delay-Line-based Timing: A delay line based timing make use of FSM and delay paths to generate necessary timing signals. The delay paths are inverters connected in series and delay of which can be manipulated by either using non-minimal length or using current starved inverters. One of the drawbacks of this scheme is its inability to track delay variations caused by process variations [41].
- Self-Timed Replica Loop: A dummy row and column containing the same number of cells as the main array are used to generate the reference delay signals. Once the dummy bit-line discharges to the dummy SA switching threshold, it resets the FSM and generates the SAE. The key design element is that the dummy bit-line discharge time should be for the statistically worst-case SRAM cell to develop sufficient differential voltage on the active bit-lines. The replica signal thus provides realistic delays as it mimics the capacitive loads and provides precise timing for WLE and SAE signals and, in addition, it tracks the process variations well. The overhead associated with the switching of a dummy column is inversely proportional to the number of simultaneously accessed columns [42].
- Other timing schemes, such as a pipelined timing control signal, has a data output latency of one clock cycle. In the direct clocking method, the WLE and SAE signals are realised from direct clock, limiting the speed as large timing margins are required for reliable operation [43] .

## 2.7 Low-Power Techniques and Figures of Merit

### 2.7.1 Bit-cell Stability

The vital property of an SRAM array after meeting power and performance bounds is the density and yield. In order to guarantee yield at the highest possible density, sufficient design margins need to be maintained and as such read stability and writeability of the

SRAM needs to be understood. There are a number of ways to characterize the read stability and writeability, some of which will be discussed in the following paragraphs.

The stability of the SRAM can be used to characterize the cell's ability to retain data. It can be used to determine the sensitivity to process variations and operating conditions. The Static Noise Margin (SNM) is defined as the maximum static spurious noise that the bit cell can tolerate while still maintaining a reliable operation [44]. SNM is implied if the noise is DC in nature such as variation in transistor sizes due to process spread, $V_{th}$ mismatch due to random dopant fluctuations and mask misalignments. The SNM is given by the side of the largest square that fits into the eye of the voltage transfer characteristic of the SRAM cell (Fig. 2.11). During retention mode, the size of the square that fits is larger than during the read access mode when the driver transistor and access transistor forms a voltage divider and degrades the 0 level of the SRAM cell. The reduced size of the square, and hence the smaller SNM is shown in Fig. 2.11. The SNM of the read accessed cell represents the worst-case SNM. Typically, the 0 value degradation is chosen at design time by the cell ratio of the SRAM cell. An idle cell can hold the data quite well as compared to in the access mode.

A successful write operation depends upon the pull-up ratio of the SRAM cell. It is possible that a write operation may fail in the presence of process variations, such as a variation that would strengthen the load PMOS as compared to the NMOS access transistor. The SNM is a useful metric to measure the robustness of a cell during read as well as retention mode. With $V_{DD}$ scaling the leakage power can be reduced at the expense of decreased SNM. Drawbacks of SNM: Inability to measure with inline testers and inability to generate statistical information on SRAM fails.

### 2.7.2 Data Retention Voltage

During retention mode, the cell is not accessed and the main function of the cell is to retain data until the next operation. Thus, minimizing the leakage current while holding a stable state is important. Reducing the $V_{DD}$ of the SRAM array during the retention mode to limit its static power consumption is one of the viable methods. The minimum supply voltage at which the cell can reliably retain the data is called the Data Retention Voltage

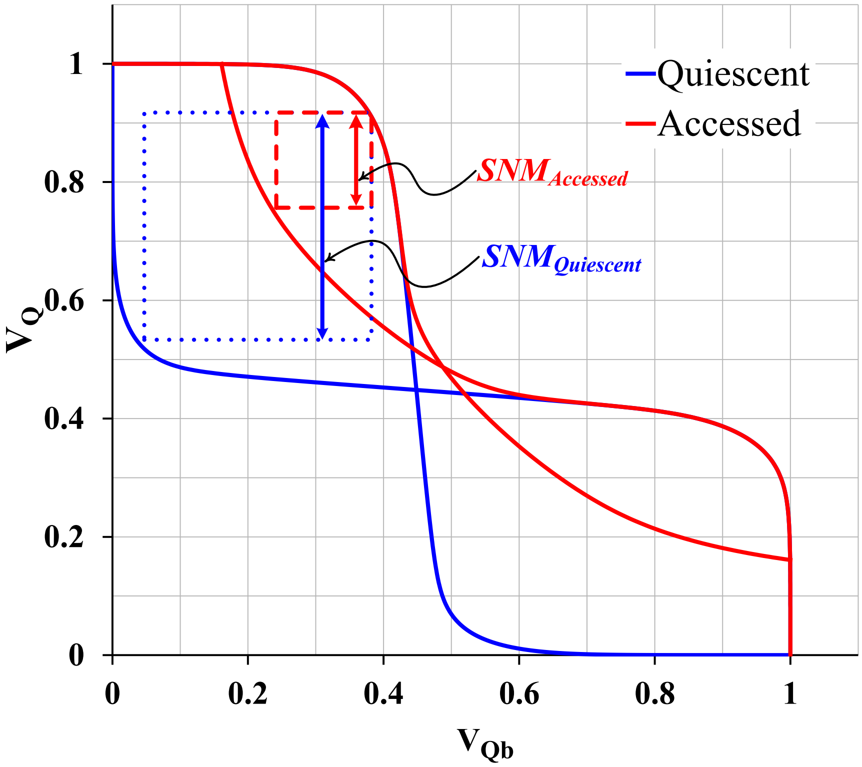


Figure 2.11: SRAM VTC in read-accessed and quiescent mode.

(DRV). However, noise on the supply rails and SETs make this scheme less interesting. Typically, a guard band of 100 $mV$ above DRV for standby $V_{DD}$ gives 60 $mV$ in SNM. Moreover, at reduced $V_{DD}$, the $Q_{crit}$ of the cell will be reduced mandating the use of error correction techniques or low voltage radiation-hardened SRAM cell. Typically, up to 90% reduction in leakage by lowering the supply to within 100 mV of the DRV has been reported [32]. An important consideration for this implementation is the energy cost in lowering the $V_{DD}$ during hold mode and switching it back to nominal value during an active operation should be carefully evaluated in addition to soft error robustness.

### 2.7.3 Virtual Ground and Reverse Body Bias

An alternative to lowering the $V_{DD}$ is to raise the ground node ($V_{SS}$) of the bit-cells. It lowers the $V_{DS}$ of the cell transistors leading to reduced sub-threshold conduction because of drain induced barrier lowering. It also reduces the gate leakage and gate-induced drain leakage. High $V_{SS}$ also results in negative $V_{BD}$ which increases the $V_{th}$, and thus lowers the sub-threshold current. Another alternative is applying reversed body bias (RBB) to the transistors in retention mode. RBB results in increased $V_{th}$ leading to a decrease in the sub-threshold leakage. The RBB can be applied to both the PMOS and NMOS transistors. The choice of implementation is process-dependent. For example, it is easier to apply RBB to PMOS because of control over N-well; however, in a triple-well process NMOS transistors can be placed in their own wells. Triple-well process implementation will also incur some area penalty.

The techniques described above typically play with the $V_{th}$ of the transistors. An alternative can be the use of high-$V_{th}$ or low-$V_{th}$ transistors for some transistors in the cell. Low-$V_{th}$ transistors for access transistors provide improved drive current during read and is a good tradeoff between power and delay [45]. To selectively reduce the leakage during retention, the use of high-$V_{th}$ transistors has also been proposed. Additionally, it increases the $Q_{crit}$ of the cell because of the weaker pull-up of the high-$V_{th}$ PMOS transistor. On the negative side, the high-$V_{th}$ decreases the drive current of the bit-cells and thus limits the speed of the cell [31]. The use of a leakage reduction technique is governed by the application, area, and power budget.

### 2.7.4 Power Consumption in SRAM

The embedded SRAM is the work horse of on-chip data storage owing to its speed, robustness and low power consumption as compared to other options. A large proportion of the memory cells are not accessed at a given time, but they are sort of ready to be accessed. It contributes to the power budget in two ways; one, the active power when the SRAM cell is accessed, and two, leakage power when the cell is in the retention mode. Typically, the on-chip SRAM constitutes 50-90% of the total transistor count. In order to retain the data, the SRAM must remain powered. The large number of transistors constantly draw leakage power. In low-power applications, the leakage power can dominate the standby power and active power. The total power consumption of an SRAM unit is given by

$$P_{Total} = P_{Leakage} + P_{Active} \tag{2.3}$$

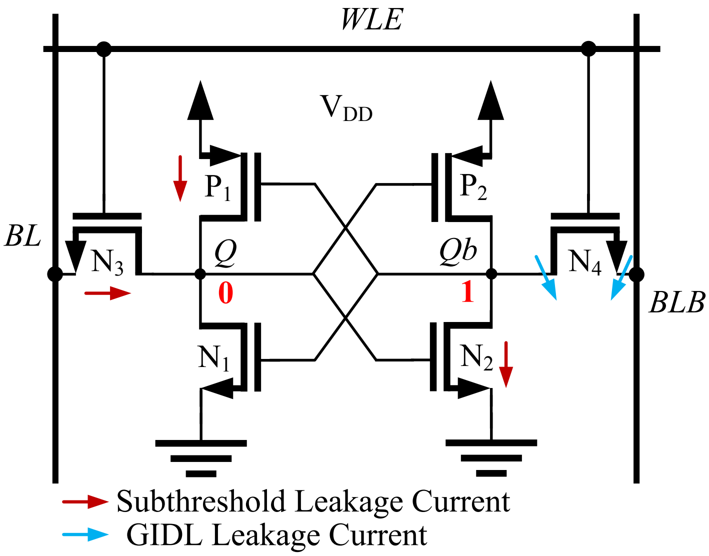


Figure 2.12: Leakage currents in a non-accessed cell

Leakage power is a major contributing factor in large memories while active power is important when the speed of operation is high.

### Leakage Power

The leakage power, also called the static power, is the power consumed by the bit-cell to retain data i.e., when it is not accessed. The amount of power required by an SRAM cell to keep its data is small, but when implemented in an array of bit-cell columns and segments, the total leakage power can become significant. Also, in low-frequency SRAMs and in scaled technologies, leakage can be a significant source of power consumption. If $I_{leakage}$ is the leakage current and $V_{DD}$ is the supply voltage, the leakage power is given by:

$$P_{Leakage} = I_{Leakage} \times V_{DD} \tag{2.4}$$

The sources of leakage current are the subthreshold leakage currents of the off transistors and the gate-induced drain leakage (GIDL) [46]. Fig. 2.12 shows the leakage paths in a

6T bit-cell. The leakage current associated with the off transistors is given by

$$I_s = I_0.e^{(V_{gs}-V_{th}/nV_{th})}(1-e^{-V_{ds}/V_T}) \tag{2.5}$$

where $V_T = kT/q$ and $I_0 = \mu_0 C_{ox}(W_{eff}/L_{eff})V_T^2 e^{1.8}$

**Active Power**

The active power or dynamic power consumption in SRAM constitutes charging and discharging of various capacitances during read and write operations. Typically, the long interconnects of word-line, data-in, data-out, and address decoders dominates the active power consumption. The bit-lines have the largest capacitance and their voltage swing during write operations has significant power consumption. Some strategies have been proposed in the literature to reduce active power consumption by reducing the bit-line capacitance ($C_{BL}$) such as hierarchical bit-line and local sense amplifiers [47] or by reducing the bit-line discharge voltage [48].

### 2.7.5 Bit-cell Read Current

The SRAM read current ($I_{read}$) corresponds to the source current from the bit-line into the SRAM node that stores a 0. During a read operation, the current is responsible for discharging the pre-charged $C_{BL}$ to a value greater than the sense amplifier offset ($V_{OS}$) so as to obtain a correct evaluation. Assuming constant $I_{read}$, the read access time ($T_{acc}$) is given as

$$T_{acc} = C_{BL}.V_{OS}/I_{read} \tag{2.6}$$

In fact, the actual $I_{read}$ should also take into account the leakage current from the inactive bit-cells sharing the same bit-line. By definition, the access-time is the smallest time for which the sense amplifier will execute a successful read. This definition, however, does not take into account the variability in read access operation due to variability in the bit-cell.

### 2.7.6 Offset in Sense Amplifiers

The total read access time is a function of bit-line discharge delay and the sense amplifier sense delay. In order to insure correct data read-ability, sense amplifier offset calls for an increased bit-line differential requirement which means increased bit-line discharge delay. The trade-off between device up-sizing and offset is well known, both with regards to $V_{th}$ mismatch and geometry mismatch [49], [50], [51]. The increased area in order to maintain a constant offset voltage also causes increased delay and this trade-off is a major limitation in sense amplifier area scaling [52]. The increased area may help in soft error robustness by increasing the node capacitance, but it may not meet the target performance metrics.

## 2.8 Soft-Error-Robust SRAMs

The SEU phenomenon results in corruption of the data in memory cells. Design hardening can be approached either at the circuit level or process level. First, its possible to reduce the amount of collected charge in the substrate by modifying the process, such as by using SOI. Second, it is possible to reduce the sensitivity towards SETs or increase the critical charge of the memory element by adding resistance in the feedback path as shown in Fig. 2.13 (adapted from [22]). Alteratively, a capacitor can also be added to increase the critical charge as shown in the Fig. 2.14 (adapted from [23]). The addition of resistance or capacitance [53] only improves the tolerance towards particle-induced transients to a certain degree, but it does not provide immunity. [54]

Another approach is to design for immunity such that the SRAM cell is immune to SEs. Immunity comes at a cost of increased area and/or access time. An SRAM cell which is inherently robust to SEU is presented in [24], [25]. In [24], a dual interlocked cell (DICE) is proposed which is shown in Fig.2.15 (adapted from [24]). In this cell, the logic value is stored on four nodes: $X1$, $X2$, $X3$, and $X4$. At a given moment, two of the four nodes store identical logic value e.g., if $X1 = 0$ and $X2 = 1$ then $X3 = 0$ and $X4 = 1$. During a particle strike if one of the nodes gets affected, then there are two consecutive nodes that have values 1 and 0. The affected node can be restored by the unaffected hold nodes. The DICE cell provides very good soft error immunity at a cost of approximately 100% area overhead and increased word line drive capability.

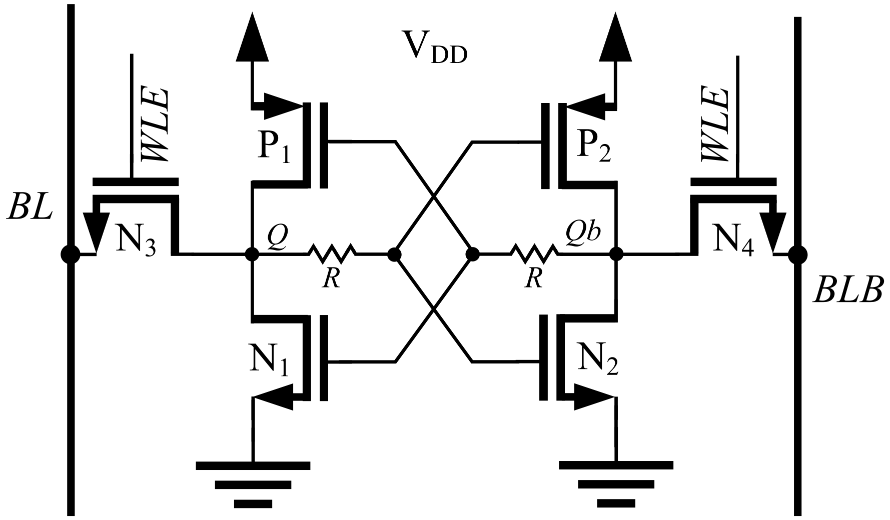


Figure 2.13: 6T SRAM bit-cell with feedback resistance.

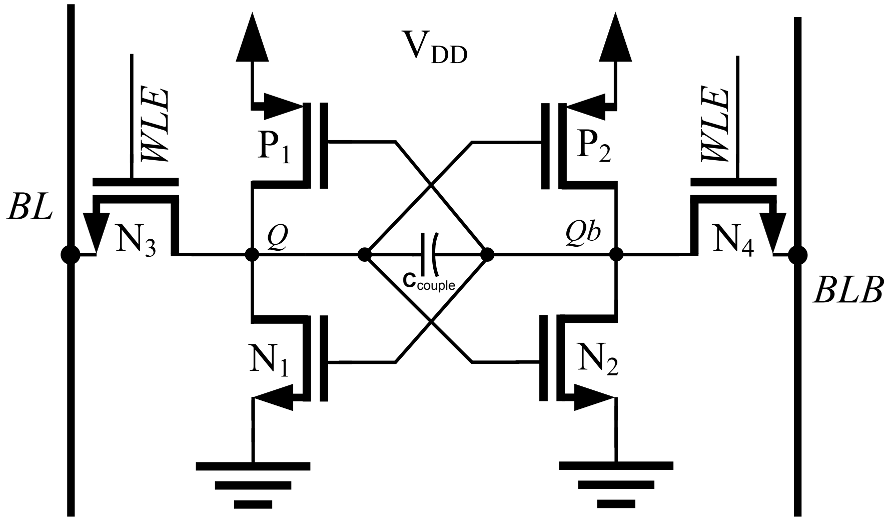


Figure 2.14: 6T SRAM bit-cell with a capacitor in the cross couple.

In [25], a quatro-10T cell is proposed which is shown in Fig. 2.16 (adapted from [25]). There are four storage nodes: *Q*, *Qb*, *Q*2, and *Q*2*b*. Each of the nodes is connected to an NMOS and a PMOS transistor, their gates are connected to two different nodes. If a node is pulled down (up) by an SET, the node voltage is restored by the on PMOS (NMOS) transistor connected to the unaffected node. The layout area of quatro cell is 2.57 times the layout area of the 6T cell and shows 98% less SEs as compared to the 6T cell [25].

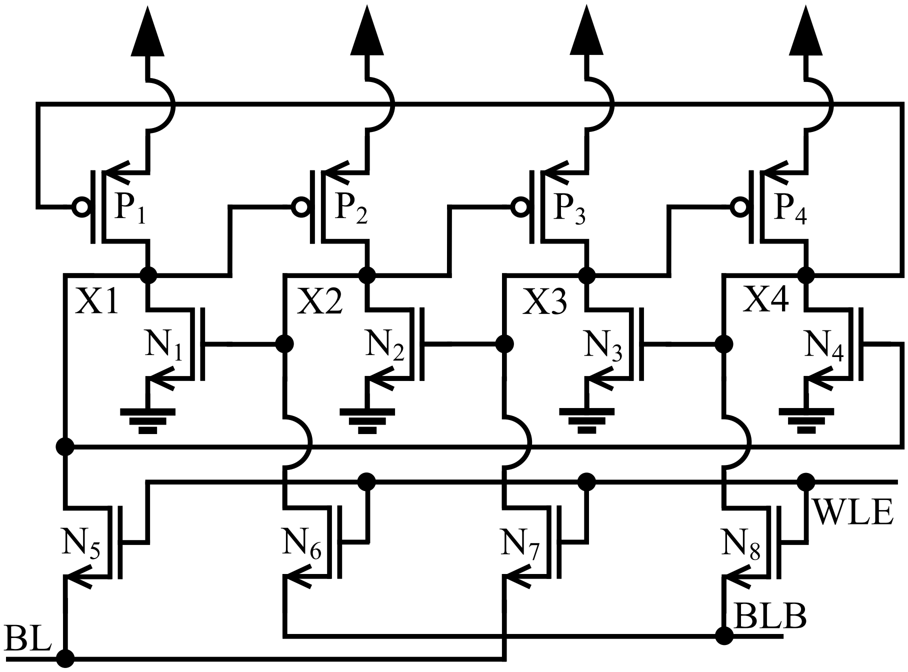


Figure 2.15: Dual-interlock storage cell (DICE).

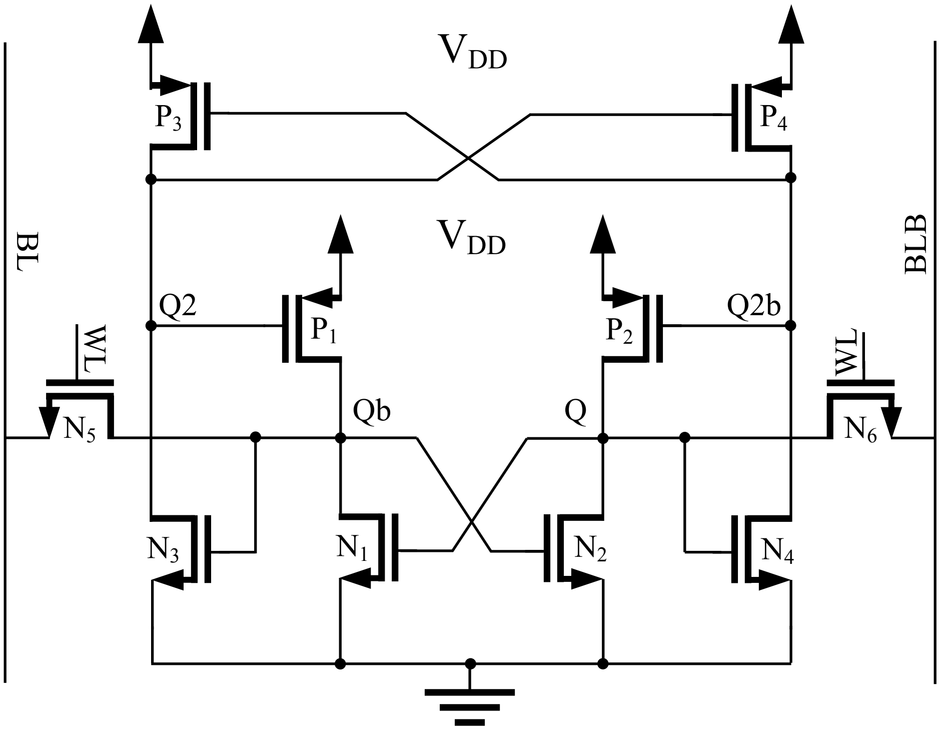


Figure 2.16: Quatro-10T cell.

## 2.9 Summary

In this chapter, a typical 6T SRAM cell and architecture was discussed. Fundamental design constraints of the bit-cell i.e., read and write operations were reviewed. The common building blocks of the SRAM such as write drivers, row and column decoders and timing schemes, have been analyzed. Low power design techniques and challenges were investigated and some of the existing design approaches were reviewed. Key issues and tradeoffs in the design of sense amplifiers were introduced. Existing soft error robust solutions in SRAM were introduced and some comparisons were made. The low power techniques discussed typically results in higher soft error rates. Also, it was found that the existing solutions have area overhead which necessitates the requirements of an area-efficient low-power soft-error-robust solution paving the way for the following chapters.

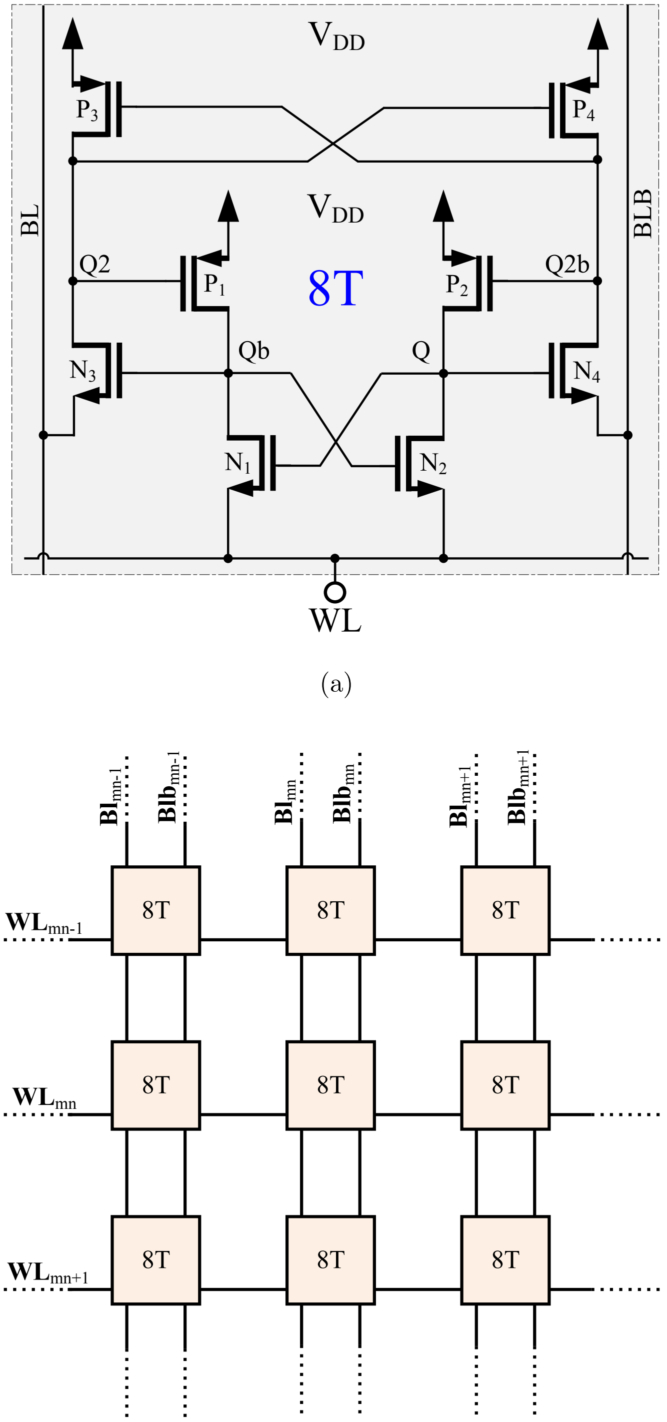
# Chapter 3

# Low-Voltage Soft-Error-Robust SRAM

## 3.1 Soft-Error-Robust 8T SRAM

The proposed eight transistor soft error robust SRAM bit-cell (8T SER SRAM / 8T bit-cell) is shown in Fig. 3.1a. This configuration comprises of four NMOS and four PMOS transistors. It provides four storage nodes Q, Qb, Q2, and Q2b. The nodes Q (Q2) and Qb (Q2b) store complimentary logic states. The 8T bit-cell can be accessed differentially using the source of transistors $N_3$ and $N_4$ to perform read and write operations.

In Fig. 3.1b, a $3 \times 3$ array is shown which is a part of an array of m rows and n columns. The bit-lines BL and BLB are shared by the cells in a given column. The word-line signal is shared by the cells in a given row. Based upon the address from an address decoder, a given row and column is selected to perform a read or write operation. Any operation in the cell is performed using the bit-lines and the word-line. Hence, a control over the bit-lines is required with the ability to connect them to the power supply or ground. The control can be achieved with the bit-line transistors $ND_1$ and $ND_2$ to connect the bit-lines to ground, $PU_1$ and $PU_2$ to connect the bit-lines to a power supply as shown in Fig. 3.2. Additionally, bit-line voltage ($V_{BL}$) can also be controlled to reduce the static power, as explained in the later sections.



(a)

(b)

Figure 3.1: a) Proposed 8T SRAM bit-cell and b) 8T bit-cell in a 3 × 3 array.

The four transistors $ND_1$, $ND_2$, $PU_1$, and $PU_2$ are managed using the signals $PD_{BL}$, $PD_{BLB}$, $PU_{BL}$, and $PU_{BLB}$, respectively. The signals $PD_{BL}$, $PD_{BLB}$, $PU_{BL}$, and $PU_{BLB}$ are generated by timing and control circuitry. Table 3.1 summarizes the states of these signals in various modes of operation.

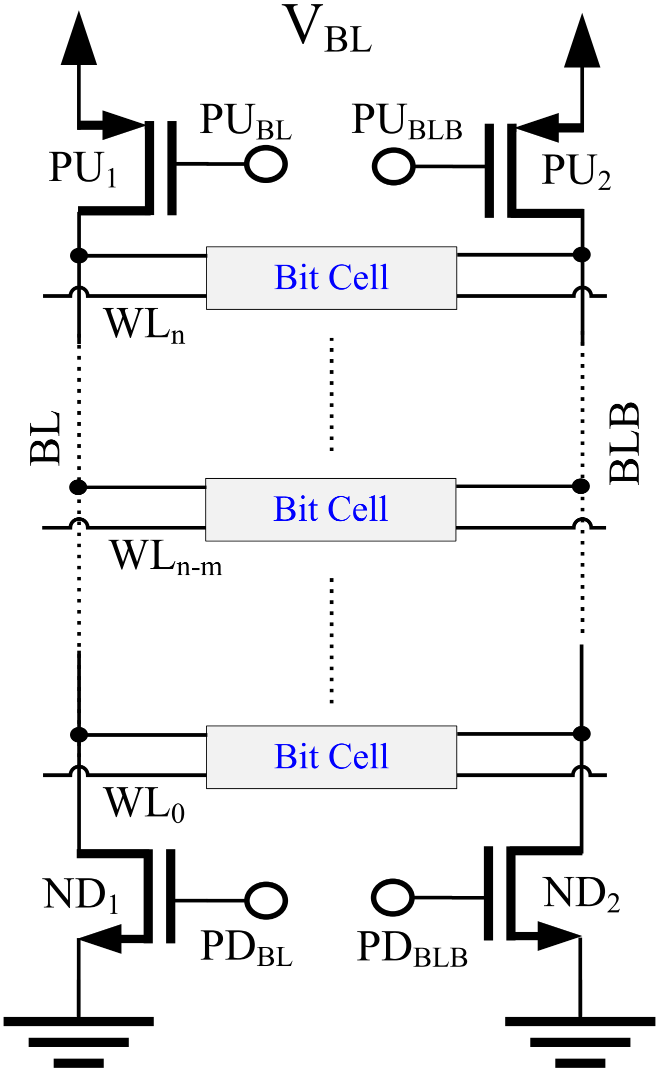


Figure 3.2: 8T SRAM cell: a typical column.

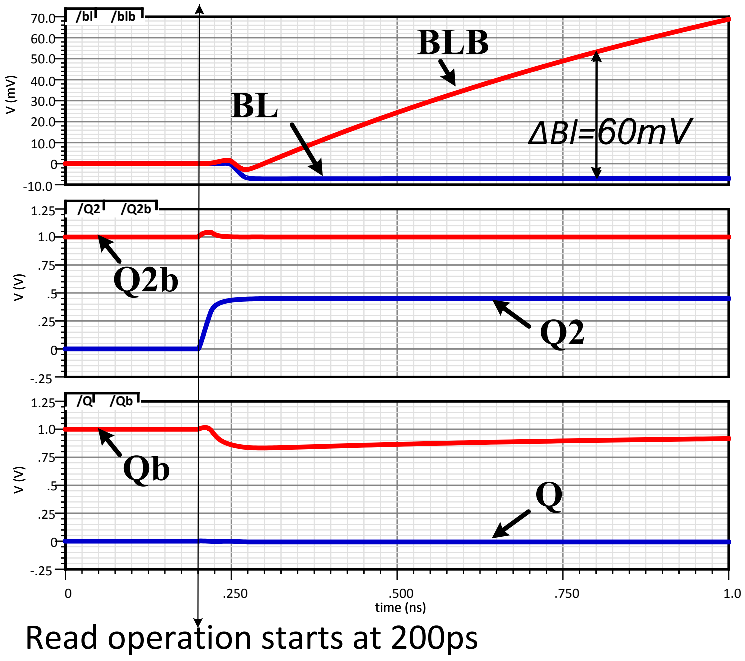
Table 3.1: 8T bit-cell modes of operation

| Row Address | Column Address | Read /Write | Data | $PU_{BL}$ | $PD_{BL}$ | $PU_{BLB}$ | $PD_{BLB}$ | Mode of operation |
|---|---|---|---|---|---|---|---|---|
| 0 | X | X | X | 1 | 1 | 1 | 1 | Retention |
| 1 | 0 | X | X | 1 | 1 | 1 | 1 | Retention |
| 1 | 1 | 0 | X | 1 | 0 | 1 | 0 | Read |
| 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | Write 0 |
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | Write 1 |

## 3.2 Read Operation

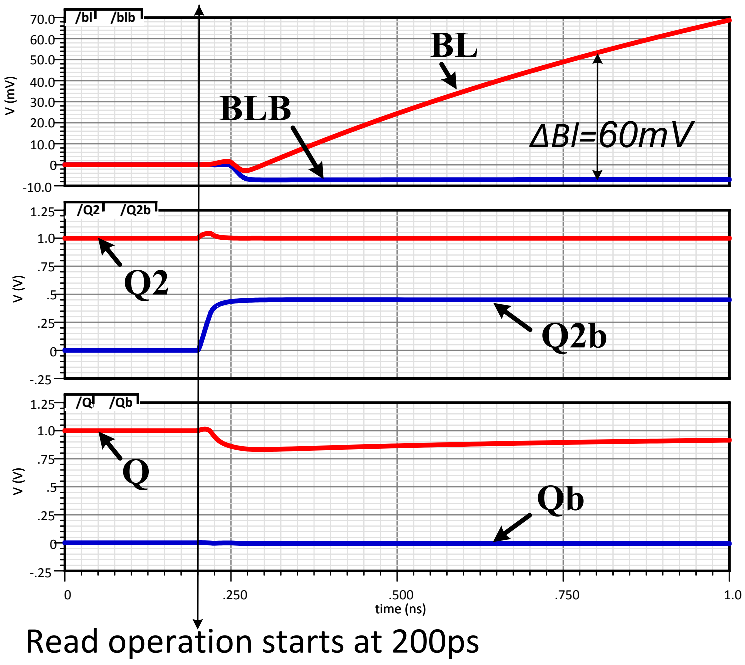
In order to explain a read operation of the 8T bit-cell, the cell is assumed to store a logic 1. Further, the supply voltage ($V_{DD}$) is assumed to be 1 V such that the nodes $Q$ and $Q2$ are at 1 V and the nodes $Qb$ and $Q2b$ are at 0 V. The differential bit-line pair BL and BLB is pre-charged to 0 V by turning on the NMOS transistors $ND_1$ and $ND_2$ through signals $PD_{BL}$ and $PD_{BLB}$ and then allowed to float and keeping the transistors $PU_{BL}$ and $PU_{BLB}$ off through the signals $PU_{BL}$, and $PU_{BLB}$. The word-line voltage ($V_{WL}$) is raised to a read voltage. The $V_{WL}$ is raised to a value greater than the $V_{th}$ of the transistors $N_3$ and $N_4$. As the $V_{WL}$ is raised, the voltage at node $Qb$ is also raised to $V_{WL}$.

The value of $V_{WL}$ is chosen to maintain a non-destructive read operation which is a tradeoff between read current and read data stability. Analysis of the tradeoff is presented in Section 3.6. The increased voltage at node $Qb$ causes transistor $N_3$ to be weakly turned on. The current flowing into the bit-line $BL$ through transistors $P_3$ and $N_3$ results in a voltage increase on $BL$. Once a sufficient voltage difference, $\Delta V_{BL}$, is developed between $BL$ and $BLB$, a sense amplifier circuit is enabled to expedite the read operation. Since the cell is symmetrical in nature, if the nodes $Q$ and $Q2$ are at 0 V and the nodes $Qb$ and $Q2b$ are at 1 V, a similar read operation develops $\Delta V_{BL}$ with voltage on $BLB$ to be greater than $BL$. The outcome of the sense amplifier determine if the 8T bit-cell is storing a logic 1 or logic 0. In Fig. 3.3 and Fig. 3.4, the waveforms show a read '0' and a read '1' operation.



Read operation starts at 200ps

Figure 3.3: 8T SER SRAM cell: read 0 operation.



Read operation starts at 200ps

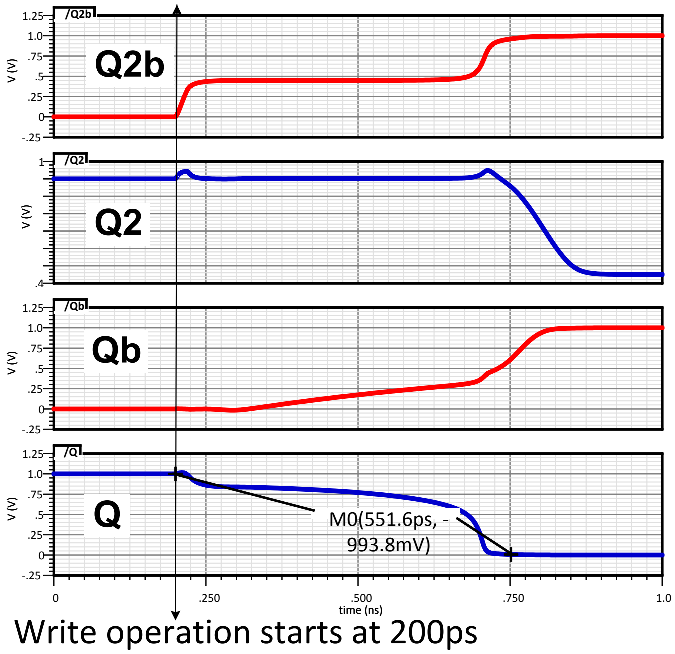
Figure 3.4: 8T SER SRAM cell: read 1 operation.

The relevant row address, column address, and bit-line control signals are set to an appropriate level, as explained above, but not shown in the figures. The 8T bit-cell is connected to the bit-lines and allowed to float at 200 $ps$. In about 600 $ps$, approximately 60 $mv$ of $\Delta V_{BL}$ has developed. The sense amplifier can detect this $\Delta V_{BL}$ and amplify it to a full swing output.

## 3.3 Write Operation

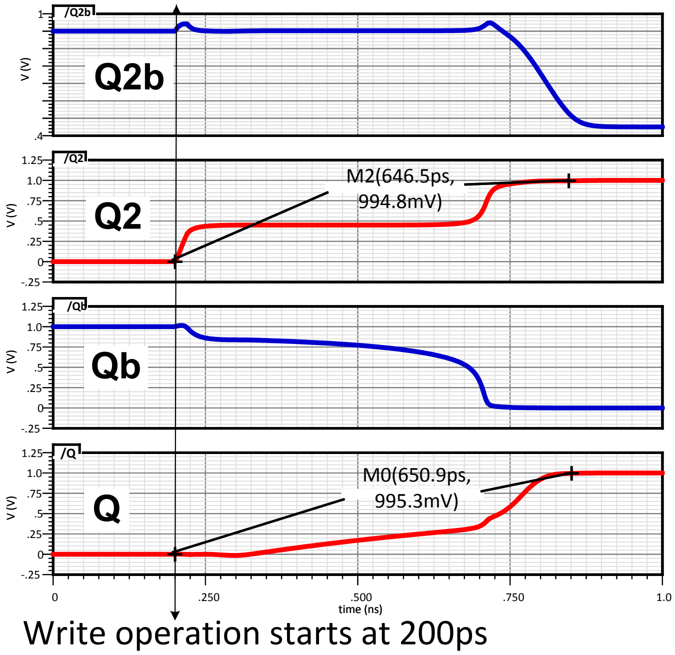
In order to explain a write 0 operation for the 8T bit-cell, the cell is assumed to store a logic 1. Further, the supply voltage is assumed to be 1 V such that the nodes $Q$ and $Q2$ are at 1 V and the nodes $Qb$ and $Q2b$ are at 0 V. To write a logic 0 into the 8T bit-cell, the differential bit-line pair BL and BLB is set to 0 V and 1 V, respectively.

The voltage of 0 V on bit-line $BL$ is achieved by turning on the NMOS transistor $ND_1$ through the signal $PD_{BL}$ and turning off the PMOS transistor $PU_1$ through the signal $PU_{BL}$. Similarly, the voltage of 1 V on bit-line $BLB$ is achieved by turning on the PMOS transistor $PU_2$ through the signal $PU_{BLB}$ and turning off the NMOS transistor $ND_2$ through the signal $PD_{BLB}$. The word-line voltage $V_{WL}$ is raised to a write voltage $V_{write}$. The $V_{WL}$ is raised to a value greater than the $V_{th}$ of the transistors $N_3$ and $N_4$. As the $V_{WL}$ is raised, the voltage at node $Qb$ is also raised to $V_{WL}$. The increased voltage at the node $Qb$ causes the transistor $N_3$ to be weakly turned on. As the voltage at bit-line $BL$ is 0 V, it will allow node $Q2$ to discharge through transistor $N_3$. On the other side, node $Q$ is not affected by $V_{WL}$ and keeps transistor $N_4$ on. As the bit-line $BLB$ is at 1 V, the on transistor $N_4$ begins charging up the node $Q2b$ from 0 V. After some time, the voltage at node $Q2b$ has reached to the point where the $V_{gs}$ (gate to source voltage) of transistor $N_4$ is less than its $V_{th}$ and, thus, turns it off. By this time, the voltage on nodes $Q2$ and $Q2b$ changes to 0 V and 1 V, respectively and internal feedback takes over which effectively forces nodes $Q$ and $Qb$ to 0 V and 1 V, respectively. Afterwards, the bit-line and word-line voltages are returned to their retention mode levels.



Write operation starts at 200ps

Figure 3.5: 8T SER SRAM cell: write 0 operation.



Write operation starts at 200ps

Figure 3.6: 8T SER SRAM cell: write 1 operation.

To write a logic 1 into the 8T bit-cell, the operation is identical to a Write 0 as explained above. It is assumed that in the beginning nodes $Q$ and $Q2$ are at 0 V and the nodes $Qb$ and $Q2b$ are at 1 V. Then, to write a logic 1, the bit-lines $BL$ and $BLB$ are set to 1 V and 0 V, respectively, and the $V_{WL}$ is raised. Thus nodes $Q2$ and $Q2b$ are charged and discharged to 1 V and 0 V, respectively. Then internal feedback takes over and completes the write operation and all the control signals can return to the retention mode levels.

In Fig. 3.5, the relevant row address, column address, and bit-line control signals are set to appropriate levels as explained above (but not shown in the figure) and logic '0' is written in about 550 ps into 8T bit-cell. Similarly, in Fig.3.6, a logic 1 is written into the 8T bit-cell.

## 3.4 Half-Selected Cells

During a write operation, one of the bit-lines which is normally connected to ground in retention mode is now connected to $V_{BL}$. For the selected row, the word-line is driven to $V_{WL}$ to complete the write operation. In the case when the cell in the same column having nodes Q<2>, Qb<2>, Q2<2>, and Q2b<2> is holding a logic 1 (which means Q<2> = 1, Q2<2> = 1, Fig. 3.7) and we are writing a logic 0 in a different row, the BL is connected to ground and BLB is connected $V_{BL}$. In other words, the half-selected column cell has BL = 0 and BLB = $V_{BL}$. Since transistor $N_3$ is off, BL = 0 does not affect the stored data at node Q2<2>. On the BLB side, $N_4$ is on, which will raise Q2b<2> to $(V_{BL} - V_{thN_4})$ which cannot discharge Q<2>. Hence, the cell recovers to the retention mode once the write operation is complete. Under the assumption that $P_2$ and $P_3$ do indeed turn off, the node Q2<2> is now holding the logic 1 on the drain and gate capacitances of $P_3$ and $P_4$ respectively. Since there is no path to discharge the capacitance, the cell will still recover at the end of the write operation.

In a read operation, the half-selected row cell's word-line signal is raised to $V_{WL}$; however, the BL and BLB are still connected to ground. Assuming that the cell in the same row having nodes Q<1>, Qb<1>, Q2<1>, and Q2b<1> is holding a logic 1 (which means Q<1> = 1, Q2<1> = 1, Fig. 3.7), activating the word-line will raise Qb<1> to $V_{WL}$, but

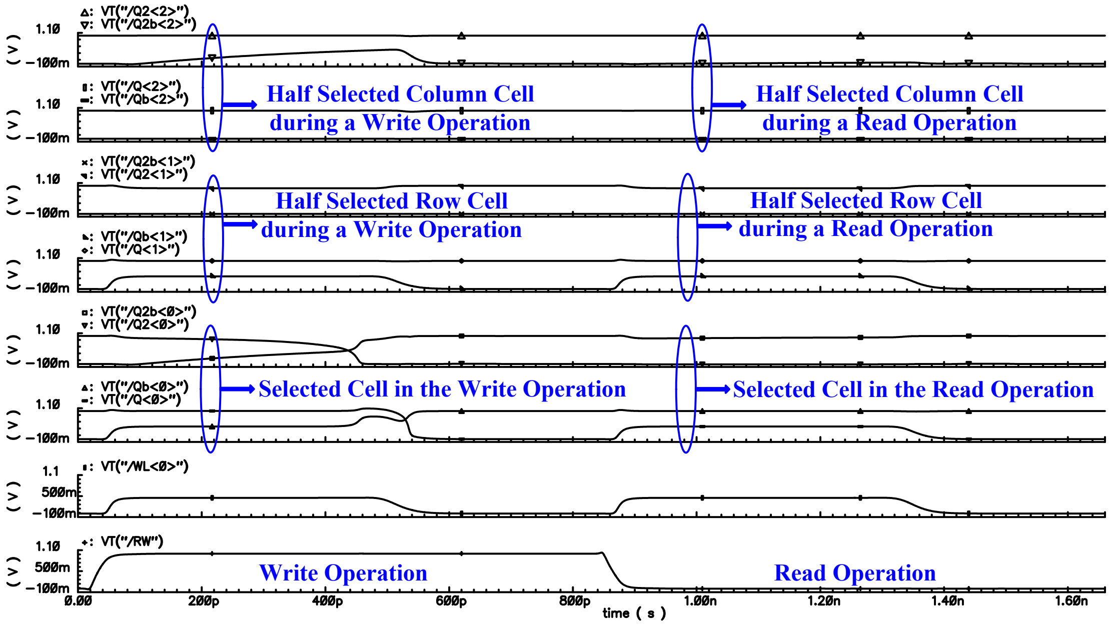


Figure 3.7: Simulations showing a write and read operation in a selected bit cell. Also shown in the figure is a half-selected row cell (Q<1>, Qb<1>, Q2<1>, and Q2b<1>) and a half-selected column cell (Q<2>, Qb<2>, Q2<2>, and Q2b<2>). Both during the read and write operations half-selected cells faithfully hold the data.

$P_1$ and $N_3$ are off and are further supported by $P_2$ and $N_4$. Thus, the half-selected row cell is not affected. The half-selected row cells during a write operation behave in an identical manner and similar conditions ensure the cell stability. Thus, the 8T bit-cell shows high stability for half-selected row and column cells.

## 3.5 Analysis of Soft Error Robustness

A SET occurring in a cell or a group of cells has a high probability of altering the stored data. In order to mimic the event of particle strike, an exponential current pulse is injected [55] into various locations in the cell. The pulse has a short rise time (10 $ps$) and a long decay time (100 $ps$). The critical charge is calculated from numerical integration of the injected current pulse that just caused a bit-flip. The equation associated with a current pulse is

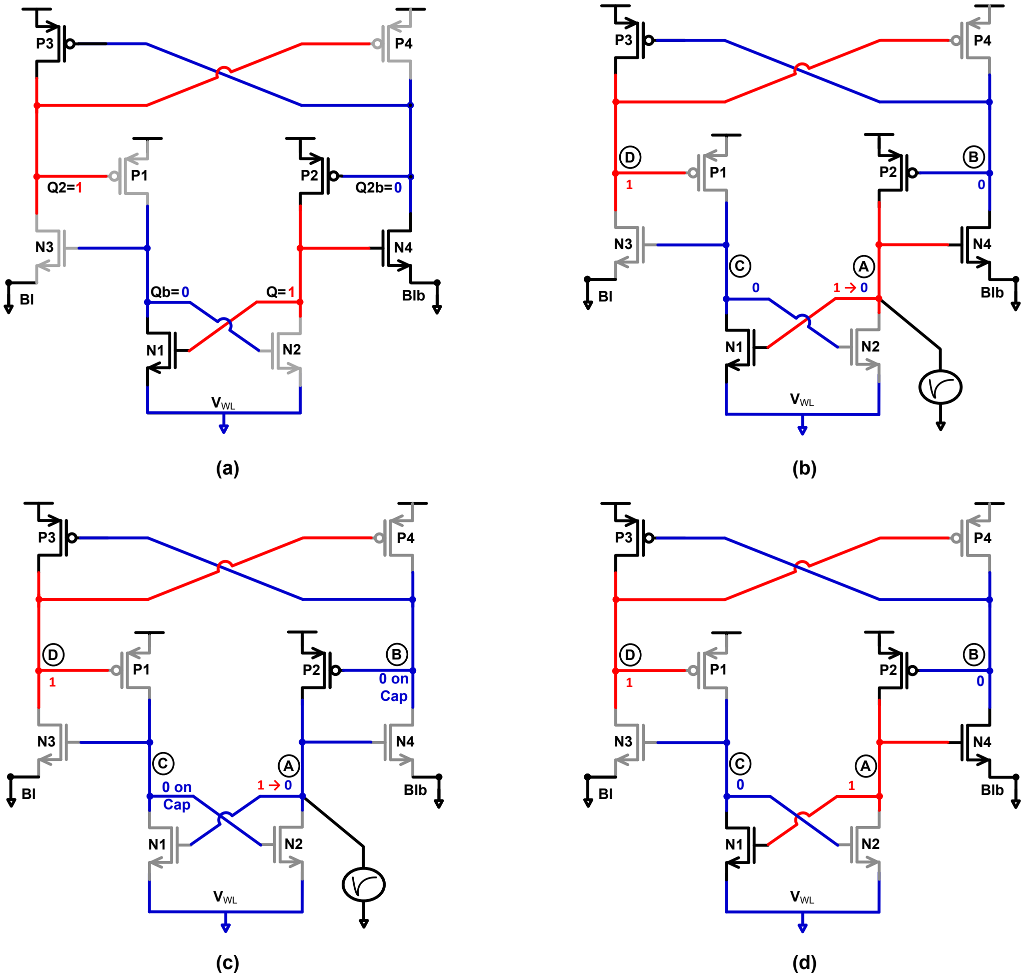
$$i_{set}(t) = \frac{Q}{\tau_f - \tau_r}\left(e^{\frac{-t}{\tau_f}} - e^{\frac{-t}{\tau_r}}\right) \tag{3.1}$$

where $\tau_f$ and $\tau_r$ are the fall and rise time of the injected current pulse, respectively.

### 1 → 0 Analysis

Fig. 3.8a shows the 8T SER SRAM cell in steady state storing a logic 1. Note that nodes $Q$, $Q2$ and $Qb$, $Q2b$ are holding logic 1 and logic 0, respectively. Thus, transistors $N_2$, $N_3$, $P_1$, and $P_4$ are off. If an SET affects the node $Q$ or $A$ as shown in Fig. 3.8b, such that it is a 1 to 0 event, it can potentially turn off transistors $N_1$ and $N_4$ making nodes $C$ and $D$ to hold the logic levels on the drain capacitance of the PMOS and NMOS transistors. Assume that the incoming SET overcomes the critical charge of the node $A$, it will turn off transistors $N_1$ and $N_4$, as shown in Fig. 3.8c. The SRAM cell goes to a steady state restoring logic 1 at nodes Q, Q2 and logic 0 at nodes Qb, Q2b once the incoming transient diminishes, as shown in Fig.3.8d.

However, a $1 \rightarrow 0$ transition at node $Q2$ (Fig. 3.8a) can potentially turn on the transistor $P_1$. In effect, the node $Q2b$ can experience a $0 \rightarrow 1$ transition. Once the transition occurs it will be held by the feedback. Thus, $P_4$ will turn on leading to a $0 \rightarrow 1$ transition at node $Q2b$. Also, $Qb = 1$ will turn on transistor $N_2$ resulting in a $1 \rightarrow 0$ transition at $Q$. In steady state, the 8T bit-cell can get overwritten. Thus, the proposed cell is completely robust to a $1 \rightarrow 0$ transition at one node ($Q$) and has some vulnerability at another node ($Q2$).

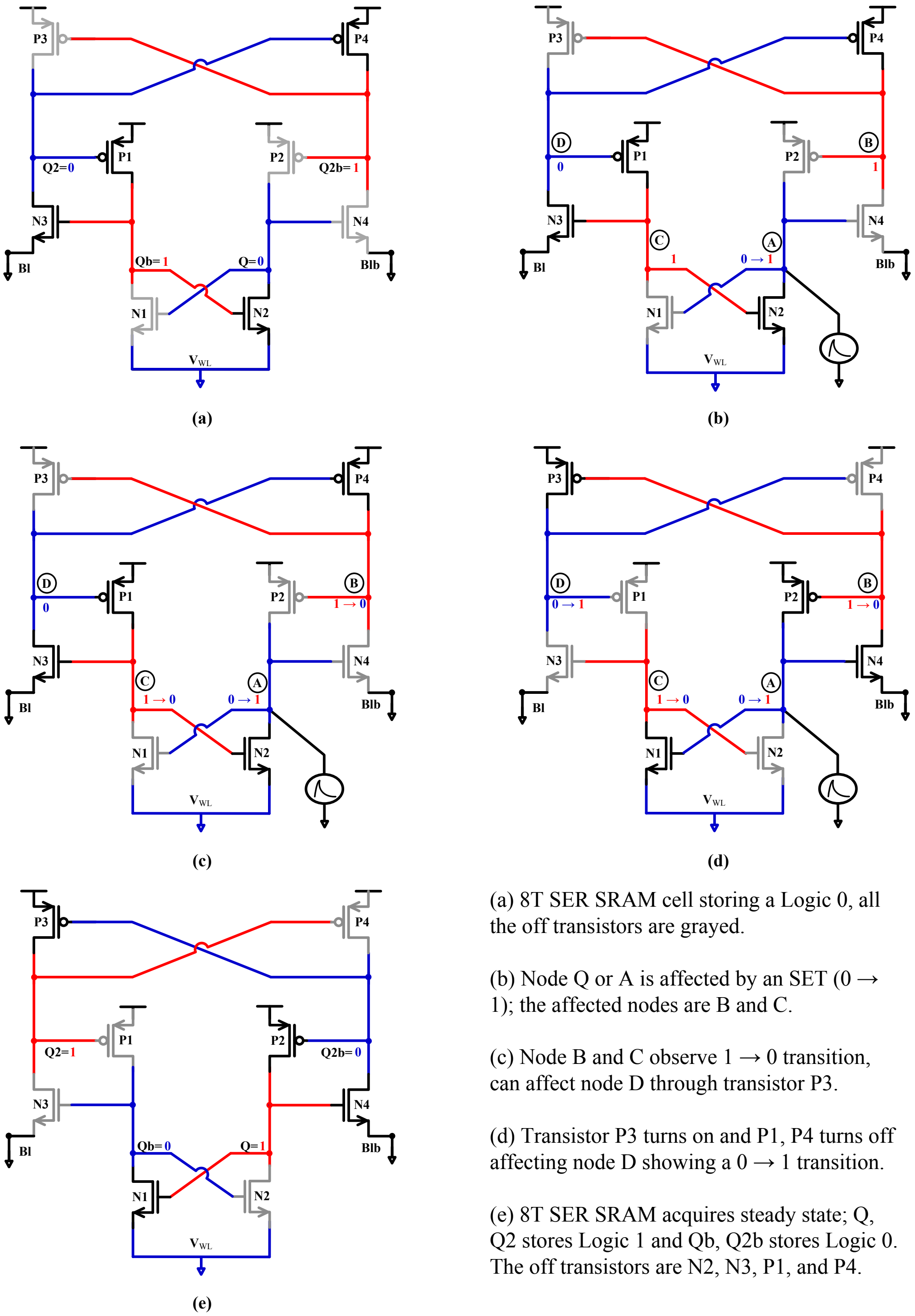


(a) 8T SER SRAM cell storing a logic 1, all the off transistors are gray.
(b) Node Q or A is affected by an SET (1 → 0). If the SET deposits charge equal to the critical charge, additional affected nodes are B and C.
(c) Transistors N1 and N4 turn off. Nodes B and C holds the logic '0' on the gate and drain capacitances.
(d) 8T SER SRAM acquires steady state by restoring logic levels to pre-SET state i.e., Q, Q2 stores logic 1 and Qb, Q2b stores logic 0. The off transistors are N2, N3, P1, and P4, as before.

Figure 3.8: 8T SER SRAM cell: Effect of an SET 1 → 0.

### 0 → 1 Analysis

Fig. 3.9a shows the 8T bit-cell in a steady state storing a logic 0 at nodes $Q$, $Q2$ and a logic 1 at nodes $Qb$, $Q2b$. Thus, the transistors $N_1$, $N_4$, $P_2$, and $P_3$ are off. If an SET affects node $Q$ or $A$ as shown in Fig.3.9b such that it is a 0 → 1 event, it can potentially turn on transistors $N_1$ and $N_4$ making the nodes $C$ and $D$ vulnerable to a 1 → 0 transition.



(a) 8T SER SRAM cell storing a Logic 0, all the off transistors are grayed.

(b) Node Q or A is affected by an SET (0 → 1); the affected nodes are B and C.

(c) Node B and C observe 1 → 0 transition, can affect node D through transistor P3.

(d) Transistor P3 turns on and P1, P4 turns off affecting node D showing a 0 → 1 transition.

(e) 8T SER SRAM acquires steady state; Q, Q2 stores Logic 1 and Qb, Q2b stores Logic 0. The off transistors are N2, N3, P1, and P4.

Figure 3.9: 8T SER SRAM cell: Effect of an SET 0 → 1.

Assuming the incoming SET overcomes the critical charge of the node $A$, it can turn on transistors $N_1$ and $N_4$ leading to a 1 → 0 transition at nodes $C$ and $D$, as shown in Fig. 3.9c. In Fig. 3.9d, logic 0 at the node $B$ can turn on transistor $P_3$ which can create a 0 → 1 transition at the node $D$ turning off transistor $P_1$. The 8T bit-cell goes to a steady state storing a logic 1 at nodes Q, Q2 and a logic 0 at nodes Qb, Q2b as shown in Fig. 3.9e. In this case, a 0 → 1 transition at the node $Q$ is capable of upsetting the bit-cell. However, a 0 → 1 transition at the node $Q2$ (Fig. 3.9a) can only affect the node $Q2$. Since, it does not turn on any of the bit-cell transistors, once the pulse diminishes the 8T bit-cell can recover to the pre-SET level. Thus, the proposed bit-cell is completely robust to SET at one node ($Q2$) and has some vulnerability to SETs at another node ($Q$).

## 3.6 Comparison and Design Tradeoffs

The comparison between different SRAM cells is a challenging task because the SRAM design is a multidimensional problem. There is no single metric which can be used as a reference between different designs. For example, if speed is the metric, the trade off can be small vs. large cell area where a smaller cell will take longer to read. If the leakage power is an important metric; smaller cell means higher leakage as compared to one with large cell area (i.e., using transistors with larger than minimum length). Yield is always important for SRAM cells, the trade off in this case is area vs. manufacturablity where a larger cell will have better yield. The minimum operating voltage of the SRAM cell can take different values depending upon the state of the cell i.e., retention or active mode. In retention mode, the gate-oxide tunnel leakage and gate-induced drain leakage are the main components of the leakage. In this work, the results are evaluated by designing the bit-cells to operate at 1.2 GHz speed at a supply voltage of 1 V in 65-nm general-purpose CMOS process.

### 3.6.1 Leakage and Read Current

The power dissipation in memories is only a fraction of the overall power budget during active mode. As SRAM must remain powered to hold their stored data, a large number of transistors in an on-die SRAM draws leakage power. The standby power becomes substantial owing to the large size of the memory array. Reducing the leakage power is hence essential. A typical tradeoff is area vs. leakage as increasing the length of the transistor reduces the leakage current. Fig.3.11 shows the leakage current and read current of different bit-cells. These results are obtained by using standard threshold voltage (s$V_{th}$) transistors. Even though simulation results for the 6T bit-cell are included, it is more appropriate to compare the proposed 8T with other soft error robust memory cells. In the 8T bit-cell, the length of the transistors can be optimized to reduce the leakage (Fig. 3.10). The leakage current ($I_{leak}$) of 8T is 78 % higher than the 6T, 26 % smaller than the DICE and is slightly better than the 10T SRAM cell. The read current of the 8T cell is 8.4× smaller than the 6T, 9.3× smaller than the 10T cell, and 17.4× smaller than the DICE cell.

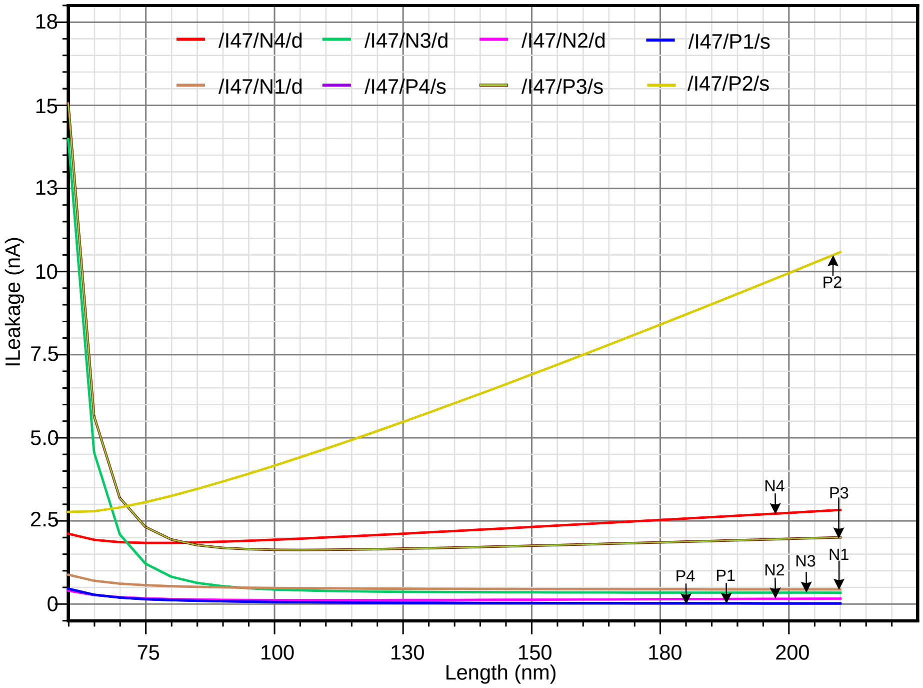
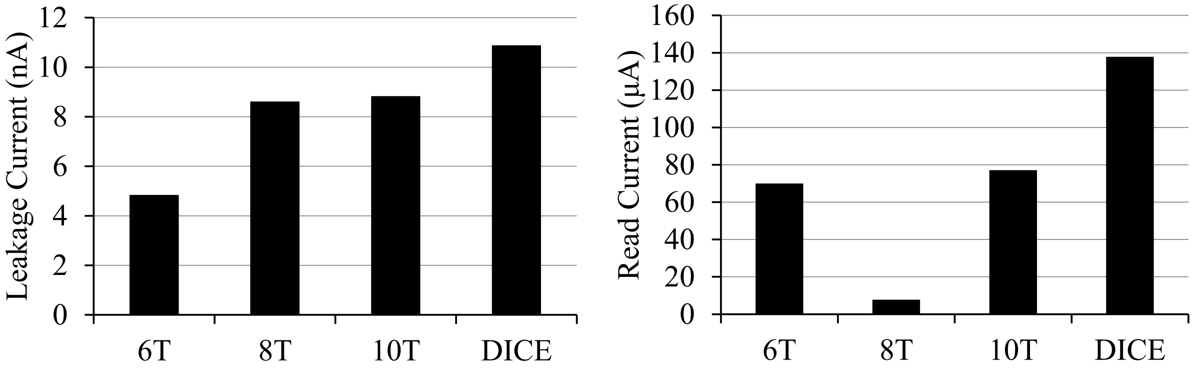


Figure 3.10: 8T bit-cell: leakage current vs. channel length of transistors.



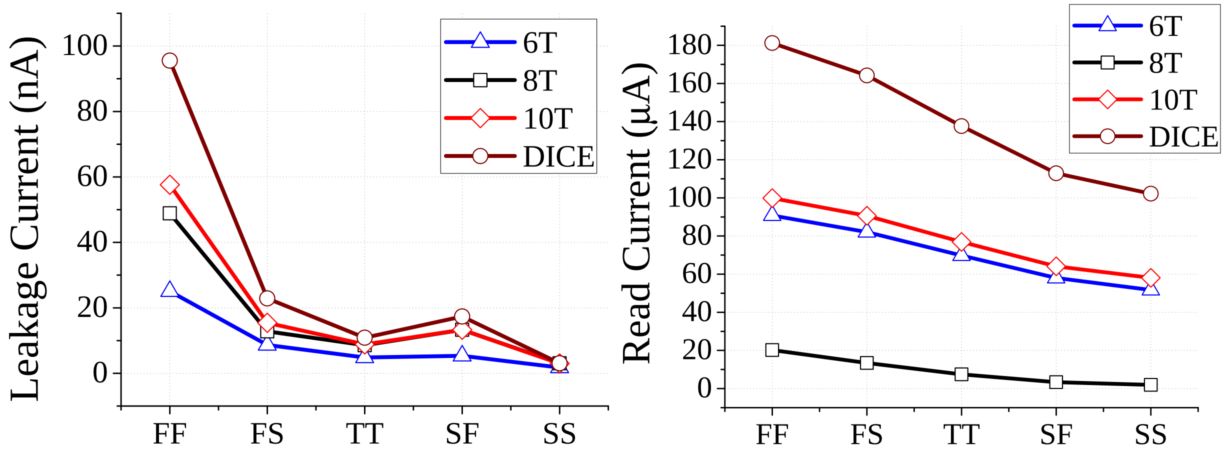
(a) Leakage current

(b) Read current

Figure 3.11: Read current and leakage current comparison of 6T, 8T, 10T, and DICE SRAM cells.

The 8T has minimum-length access transistors ($N_3$ and $N_4$ in Fig. 3.1a) for the results shown in Fig. 3.11. If the 8T access transistor length is increased, $I_{leak}$ is reduced at the cost of smaller ΔBL development. The lost differential bit-line swing can be gained back by using h$V_{th}$ transistors. If the PMOS transistors $P_1$, $P_2$, $P_3$, and $P_4$ of Fig. 3.1a are replaced with h$V_{th}$ transistors, we obtain reduced leakage current and higher bit-line swing. In the proposed 8T bit-cell, if the PMOS transistors (ref. Fig. 3.1a) could be h$V_{th}$ transistors and the other transistors $N_3$, $N_4$ (access transistors) and the transistors $N_1$ and $N_2$ could be s$V_{th}$. This resulted in 30% leakage current reduction.

Further, a comparison is carried out at iso-speed for read and leakage current consumption. The leakage variation between the FF and SS corners is 14x for 6T, 16x for 8T, 19x for 10T and 31x for the DICE cell, as shown in Fig. 3.12a. The 8T performs favorably when compared with other soft error robust cells. In the TT process corner, the 8T has a read current of 7.47$\mu$A which is 9.4x smaller than the conventional 6T, 10.3x smaller than the 10T, and 18.4x smaller than the 12T DICE cell (Fig. 3.12b).



(a)

(b)

Figure 3.12: Variations in the leakage current and read current across different process corners.

### 3.6.2 Bit-line and Word-line Voltage Scaling

In standard CMOS logic, the trade-off between power and delay dominates other metrics, such as functional robustness which is relatively easy to achieve. In memories, the need for large storage density makes area a dominant metric as well. To reduce the area SRAM compromises some important properties of CMOS logic, e.g., noise margins. Unfortunately variations in state-of-the-art processes cause circuit parameters to vary. For example, the closely placed transistors with identical layout can have different threshold voltages. This means that the adjacent memory cells can exhibit different behavior. The more important issue is a tradeoff between power consumption and functional robustness. The goal of lowering the power consumption is constrained by functionality such as read, write, retention, and soft errors. The possible solution is to reduce the supply voltage or change the bias voltage of the transistors or it can be a combination of the two techniques. Any acceptable approach must retain the data reliably. The bit-line voltage can be scaled to reduce the active power consumption during a write operation. The effect of reduced bit-line voltage $V_{BL}$ is shown in Fig. 3.13 for the 8T bit-cell.

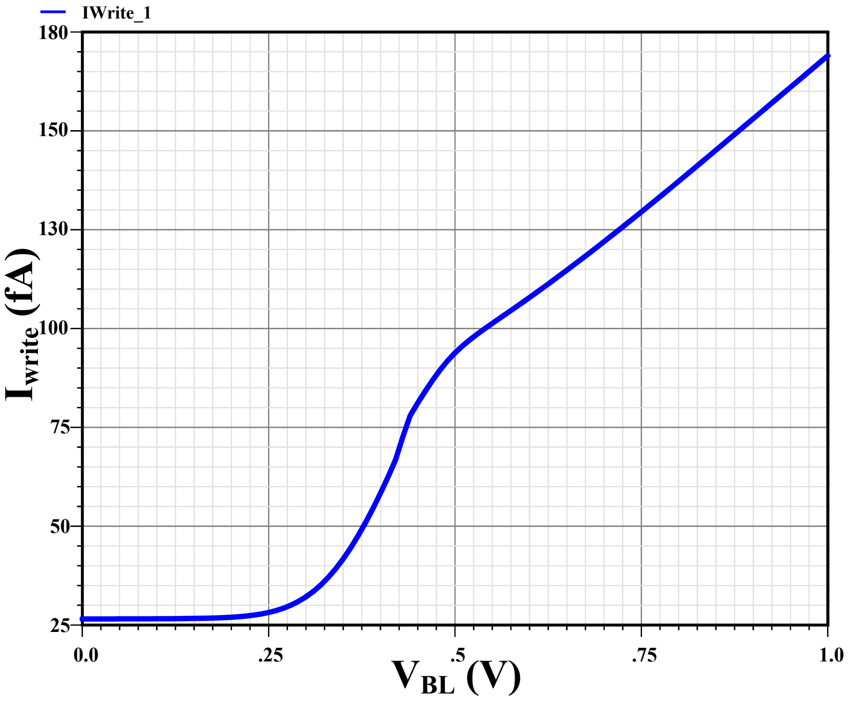
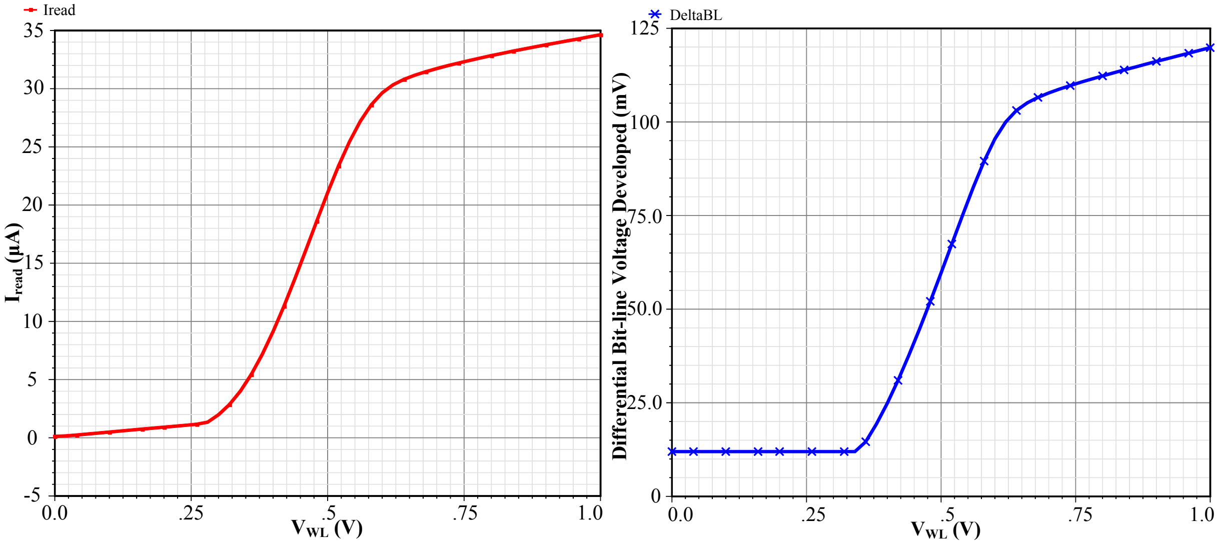


Figure 3.13: Effect of bit-line voltage on the write current.

By reducing the $V_{BL}$ from 1.0 V to 0.5 V, the write current saving is 50%. The applied word-line voltage $V_{WL}$ during a read operation has an impact on the read current, as shown in Fig. 3.14a and $\Delta V_{BL}$, the bit-line differential voltage developed, as shown in Fig. 3.14b. The increased read current stems from the increase in word-line voltage $\Delta V_{WL}$. The increased voltage in effect increases the drain current of transistor $N_3$ (ref. Fig. 3.1a) allowing node $Q2$ to discharge faster and hence build up voltage on the bit-line $BL$.



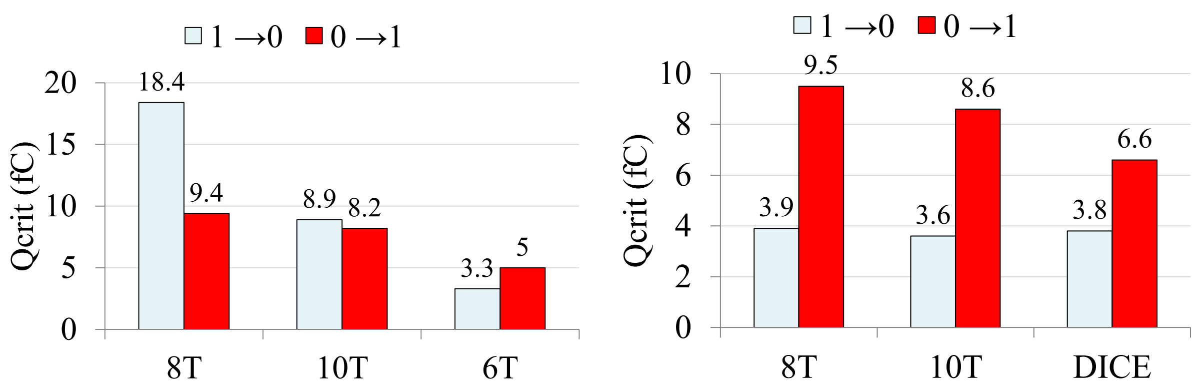
(a) $I_{read}$ vs $V_{WL}$

(b) $\Delta BL$ vs $V_{WL}$

Figure 3.14: Effect of $V_{WL}$ on read current and differential bit-line voltage generation.

### 3.6.3 Soft Error Robustness

The soft error robustness of the 8T cell is compared with the 6T, 10T and 12T DICE cells in Fig. 3.15a and Fig. 3.15b. The simulations were carried out for worst case $1 \rightarrow 0$ and $0 \rightarrow 1$ scenarios. In particular, the 8T bit-cell shows 5.6× improvement over the 6T bit-cell and 2.1x improvement over the 10T bit-cell for a $1 \rightarrow 0$ transition for single node upset. For $0 \rightarrow 1$ transition, 8T bit-cell shows 1.9x improvement over the 6T bit-cell and is slightly better than the 10T bit-cell. It has been reported that the drain of an off NMOS transistor, which means a node holding a 1, is more sensitive to an SET ($1 \rightarrow 0$) [2] and the 8T bit-cell shows high robustness in this case. For double-node upsets, the 8T bit-cell is 10% better than the 10T bit-cell and shows 44% improvement over DICE for a $0 \rightarrow 1$ transition. The 8T bit-cell shows high robustness for both single and double-node upsets.



(a) Critical charge for single-node upset

(b) Critical charge for double-node upset

Figure 3.15: Critical charge comparison for 6T, 8T, and 10T cells towards single-node upset; critical charge comparison for 8T, 10T, and DICE cells towards double-node upset.

## 3.7 Measurement Results

A 32-kbit block consisting of four 32 bit words with 256 rows per column was designed in 65-nm CMOS technology with a nominal supply voltage of 1.0 V. The 8T bit-cell layout complying with logic design rules is shown in Fig. 3.16. The 8T array shown in Fig. 3.17 was designed to operate at 1.2 GHz at 1.0 V.

After designing the array, simulations were carried out with the nodes loaded with post-layout extracted capacitance values. The tradeoff between power consumption and functional robustness was determined by varying the bit-line voltage and the word-line voltage. For each combination of $V_{BL}$ and $V_{WL}$, a read and write operation was evaluated and the results are shown in the shmoo plot of Fig. 3.18. Note that the 8T is fully functional over a wide range of bit line and word-line voltage levels. In particular, the array can be read for $V_{WL}$ as low as 290 mV and, for the same $V_{WL}$, the cell can be written using a $V_{BL}$ as low as 630 mV. Scaling the bit line voltage also results in reduced active power consumption during a write operation. Simulations show that reducing the $V_{BL}$ from 1.0 V to 630 mV results in 4x write power savings without an impact on the speed.

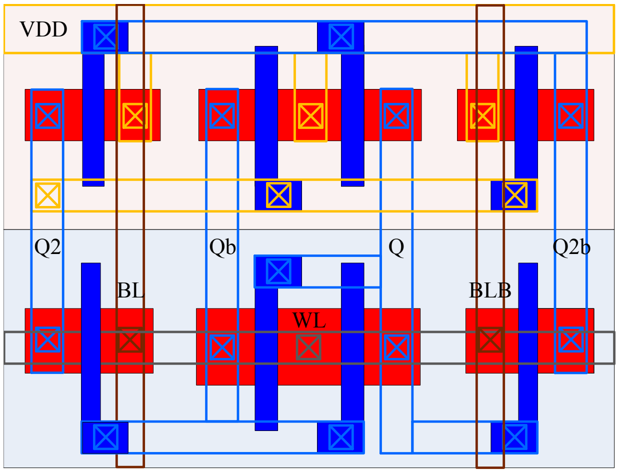


Figure 3.16: Layout of the 8T bit-cell.

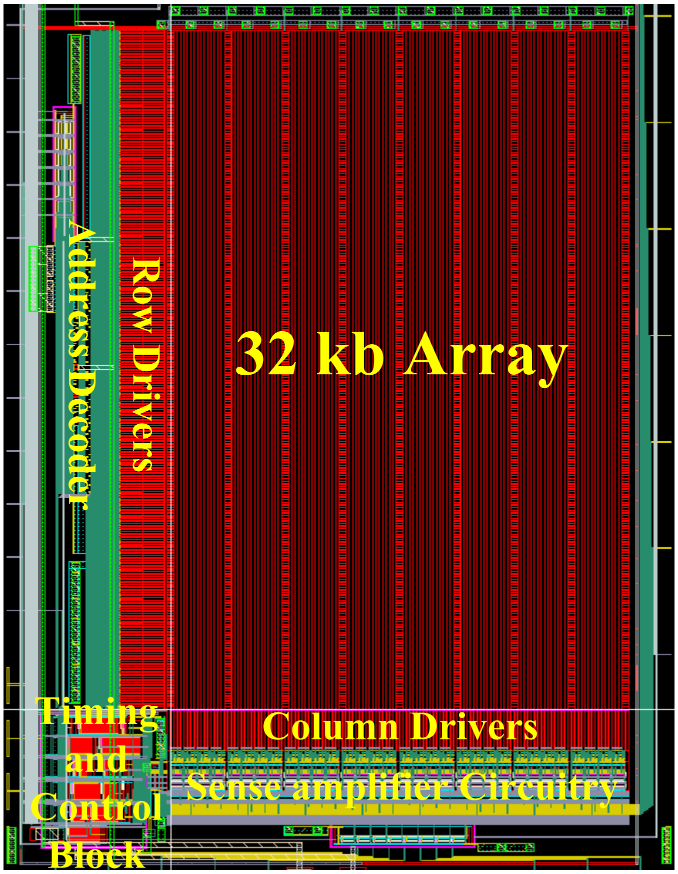


Figure 3.17: Layout of the 32-kb 8T array designed in 65-nm bulk CMOS technology.

Subsequently, an analysis was carried out to determine the read current by sweeping the word-line voltage of a selected bit cell for a successful read operation. Half-selected row cells were observed during this experiment to analyze the possibility of a destructive read. Since the read operation is independent of the $V_{BL}$, the findings are consistent with the results presented in Fig. 3.18. The smallest word-line voltage to make a correct read decision is 300 mV. Thus, a 14x reduction in read current is observed when the word-line voltage is decreased from 1.0 V (highest $V_{WL}$ with a successful read) to 300 mV. Thus the read margin of the cell is approximately 700 mV. Therefore, it has been shown that the proposed 8T bit-cell and the access transistor-less architecture provides wide read and write margins in addition to soft error robustness.

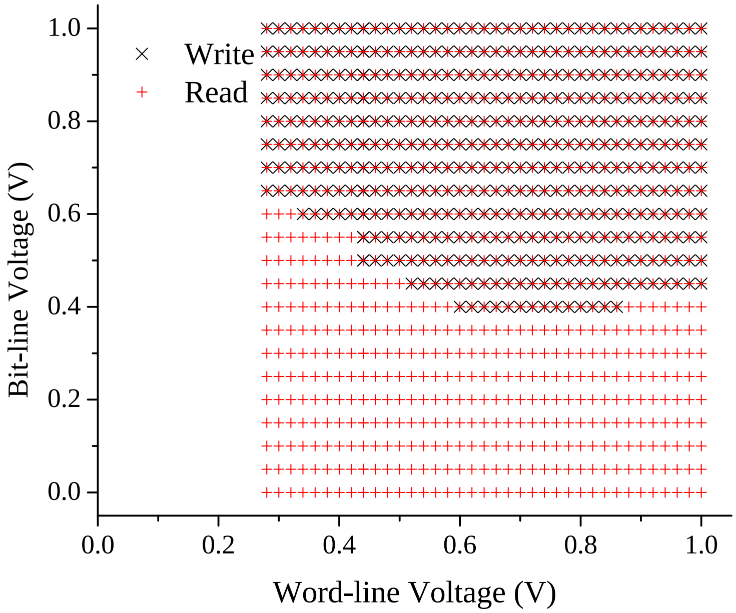


Figure 3.18: Simulated shmoo plot of read and write operations of the 8T with $V_{BL}$ vs. $V_{WL}$. Blanks indicate a fail, a star ($\star$) indicates full functionality, a plus (+) indicates only writes are operational, and a cross (X) indicates only reads are operational. Since reads are independent of $V_{BL}$ they are operational even for $V_{BL}$=0.

The measurement of the test chip involved a few steps. First, a printed circuit board (PCB) was designed to perform various measurements. The resulting four-layer PCB used the top and bottom layers as signal layers, and the second layer as a power plane ($V_{DD}$),

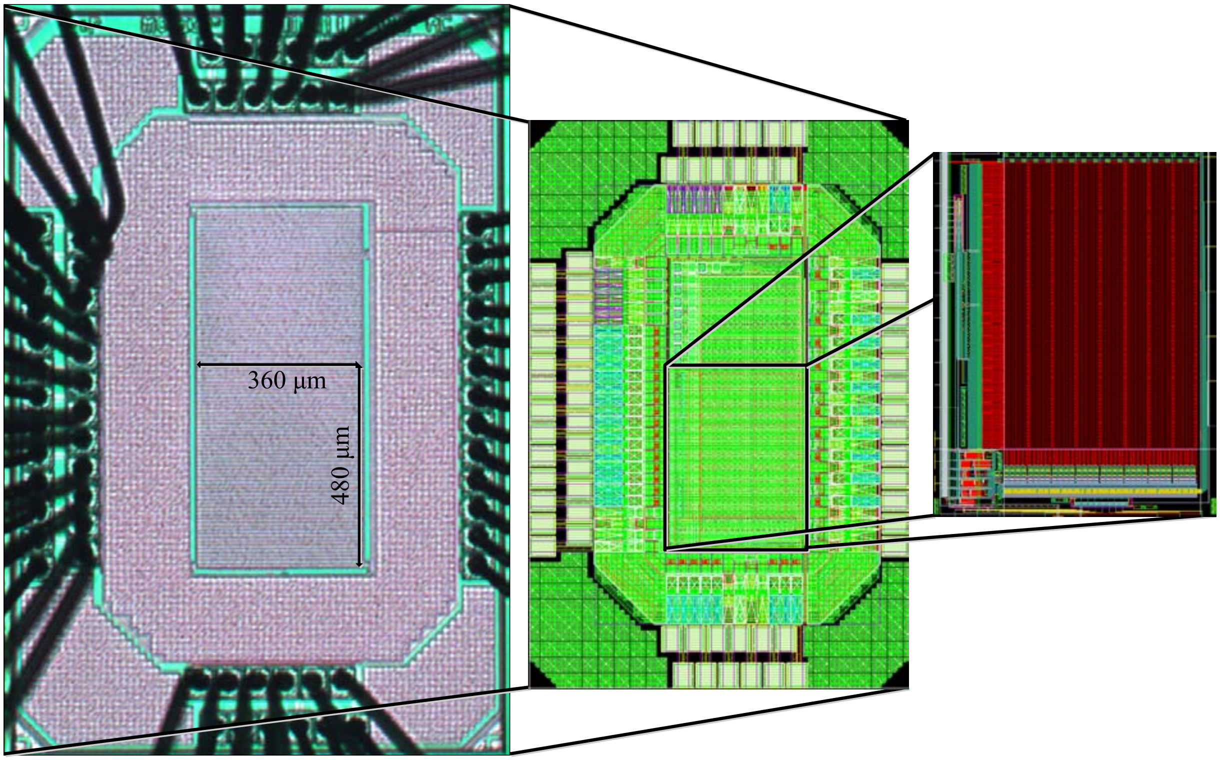


Figure 3.19: Die photo, chip-level layout, and the array layout of the test chip designed in 65-nm CMOS technology.

and the third layer as a ground plane ($V_{SS}$). This implementation enabled better $V_{DD}$ and $V_{SS}$ contacts and higher component density. Additionally, the inputs, outputs and control were planned in such a way that would enable radiation testing at a later date. Specifically, the radiation test involved placing the test chip/ PCB in front of a neutron beam while the chip was powered and connected through long cables. The PCB with the necessary components and the test chip is shown in Fig. 3.20. The test chip was packaged in an 80-pin ceramic quad flat package (CQFP). All the signals were generated on chip, but some control and reference signals, such as block select, word-line voltage, bit-line voltage, and reference voltage for timing delay, were generated by potentiometers on board. The address, data input, read/write signals were generated using jumpers (acting as toggle switches) during initial stages of testing, and using a data generator later for complete testing. Provisions were made to feed these signals through ribbon cables sockets. The chip was designed with a 32-bit data word and, the design being pad-limited, multiplexers and latches were used to siphon the 8-bit data out in four clock cycles. The chip was designed to work at 1.2 GHz. All the timing and control signals were edge driven. The measurement of the test chip was carried out at 100 MHz. Even though the measurements were carried out at 100 MHz, internally the read write operations were completed at a speed of 1.2 GHz.

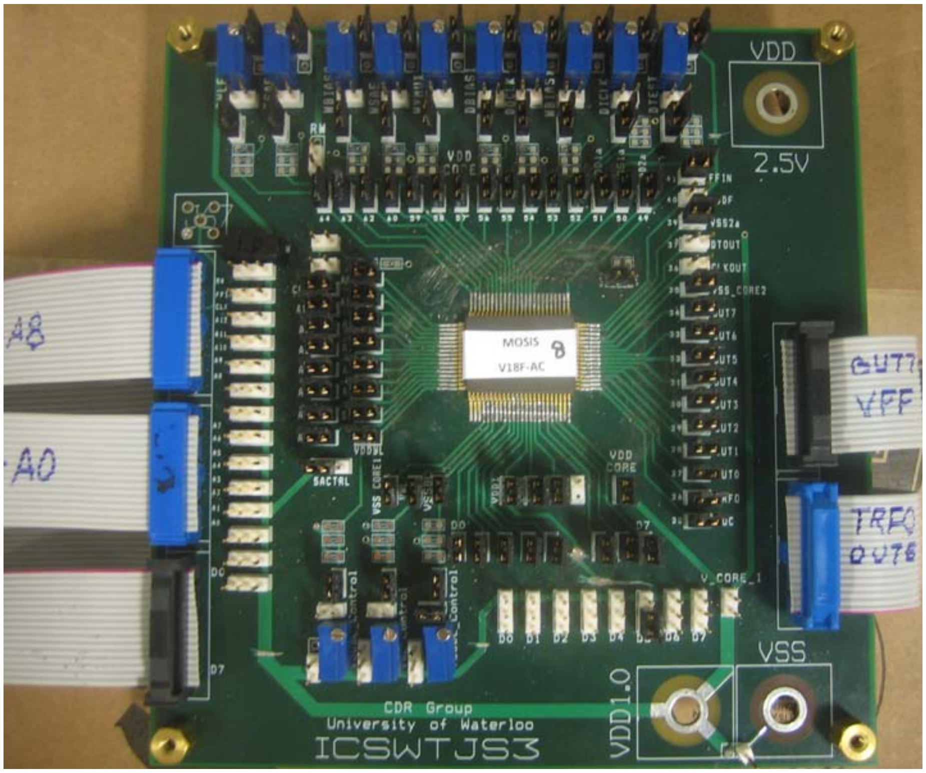


Figure 3.20: Photograph of the PCB used to test the chip containing the 8T array at TRIUMF.

The chip was tested in two stages. In the first stage, the functional and performance measurements were carried out at the test lab of CMOS Design and Reliability Group at the University of Waterloo. In the second stage, the soft error robustness was evaluated by irradiating the test chip at Canada's National Laboratory of Particle and Nuclear Physics also called Tri-University Meson Facility (TRIUMF) located in Vancouver, British Columbia.

In power and performance measurements, the leakage and active power consumption was measured at different operating voltages and clock frequencies. The following test equipment was used during tests.

- To supply power to the PCB/test chip : Precision DC Power Supply (Agilent E3631A, BK Precision 1760A)
- To generate the clock and address signals : Data Generator (Tektronics DG 2020A)
- For data evaluation : Logic Analyzer (Tektronics TLA 5201)
- To observe clock and output signals : Oscilloscope (LeCroy WaveRunner 6100)
- For voltage and current measurements : Precision Multimeter (Fluke 189, Fluke 8846A)

The $V_{DD}$ of the array was a separate pin on the test chip which enabled the measurement of active and leakage power. However, the power measurements does include the contribution of peripheral circuits surrounding the array. In order to measure the leakage current, a multimeter was used as an ammeter in series with the pin supplying voltage to the memory array. During the leakage measurements the array was kept in the retention mode. Similarly, active power was measured by measuring the current consumed during a read and write operation and multiplying it by the operating voltage.

In Table 3.2, the measurement results of the 32-kb test chip containing the 8T bit-cell are compared with the work from literature in a similar technology node. The work presented by Arnaud [56], Utsumi [57] used a low-power process and Wang [58] used an ultra-low-power process. The proposed 8T bit-cell designed with logic design rules in a general-purpose CMOS process has $3-5\times$ higher area than SRAM work from the literature which is designed with SRAM design rules. The measurement results reported for the 8T array operating at 1 V are at a clock frequency of 100 MHz. The leakage current of 8T is comparable to Utsumi. The read current of 8T is 20× smaller than Arnaud, which is operating at 0.9 V. Also, the 8T bit-cell exhibits better write margins among the reported results. In Fig. 3.21, the 8T bit-cell is evaluated for a range of bit-line and word-line

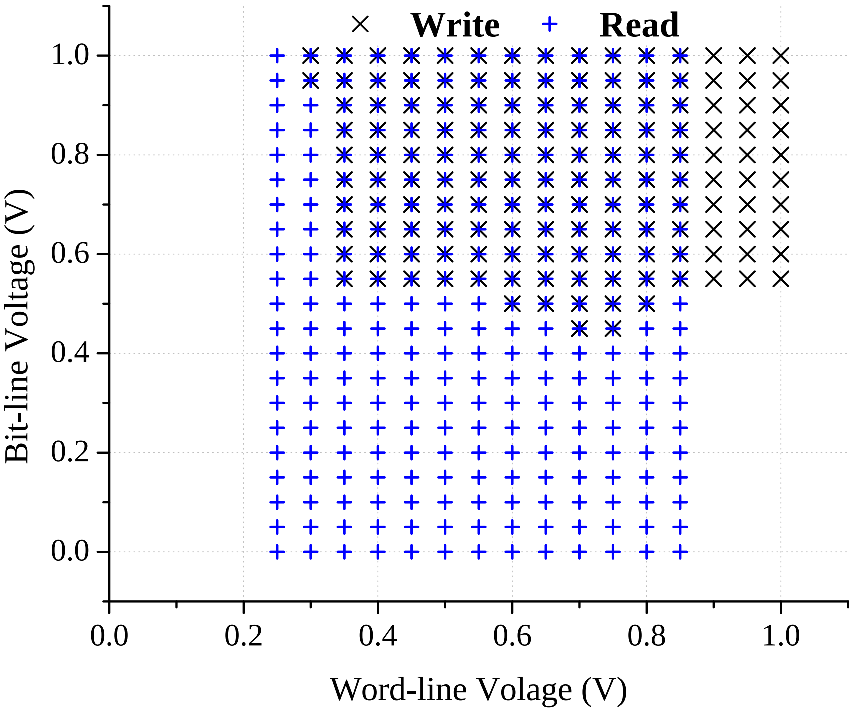


Figure 3.21: Measured Shmoo plot of the read and write operations of the 8T with $V_{BL}$ vs. $V_{WL}$. Blanks indicate a fail, a star (∗) indicates full functionality, a plus (+) indicates only reads are operational, and a cross (X) indicates only writes are operational. Since reads are independent of $V_{BL}$, reads are operational even for $V_{BL}$=0.

voltages. The write is operational for a large range of word-line voltage. Typically, a write operation is successful when the $V_{WL}$ changes from 0.3 V to 1.0 V; however, even with the scaled bit-line voltage of 0.55 V, the 8T bit-cell can be written from 0.35 V to 1.0. V. Recall that in retention mode, the bit-line rests at $V_{SS}$. Thus the read operation being independent of the bit-line voltage, is functional when the word-line voltage varies from 0.25 V to 0.85 V. In a realistic implementation, the word-line voltage varies around the threshold voltage of the NMOS transistor. Even with the consideration of 6$\sigma$ variation of $V_{th}$ which may require higher $V_{WL}$ there will not a read upset.

Based on the simulation and measurement results as presented in Fig. 3.18, Fig. 3.21, and Table 3.2, the 8T is fully functional over a range of $V_{DD}$, $V_{BL}$, and $V_{WL}$ voltages. In particular, array is fully functional for $V_{DD}$ as low as 0.55 V. At 0.55 V, the leakage and

Table 3.2: Comparison of SRAMs

| Features | This work (2012) | Arnaud (2003) [56] | Utsumi(2005) [57] | Wang (2007) [58] |
|---|---|---|---|---|
| Memory Size | 32 kb | 4 Mb | 7 Mb | 1 Mb |
| Technology | 65 nm GP-CMOS | 65 nm LP-CMOS | 65 nm LP-CMOS | 65 nm ULP-CMOS |
| Area ($\mu m^2$) | 2.42 | 0.69 | 0.495 | 0.667 |
| $V_{DD}$ Core | 1 V to 0.55V | 0.9 V | 1.2 V | 1.2 V to 0.5 V |
| Speed | upto 1.2 GHz | - | - | 1.1 GHz |
| $I_{Leakage}/bit$ | 5.38 nA @ 1.0 V<br>1.33 nA @ 0.55 V | - | 5.5 nA | 0.012 nA @ 0.5V |
| $I_{Read}/bit$ | 1.153 $\mu A$ @ 1.0 V<br>6 nA @ 0.55 V | 23 $\mu A$ | - | - |
| Write Margin | > 400 mV | > 300 mV | - | - |

read current per bit are reduced by 4× and 192×, respectively, when compared with $V_{DD}$ of 1.0 V. In other words, the 8T can be safely operated at a lower voltage which will result in a smaller active and standby current consumption and over all low power.

## 3.8 Radiation Test Results

The soft error robustness of the chip was evaluated after successful functional verification at the University of Waterloo. The chip was radiated with accelerated neutrons at the TRIUMF Neutron Facility (TNF) according to Joint Electron Device Engineering Council (JEDEC) standards [59]. The neutron beam has the energy spectrum shown in Fig. 3.22. The neutron beam had an average fluence of 1.959 x $10^6$ $n/cm^2 - s$, which is approximately 3.646 x $10^8$ times higher than the neutron fluence at sea level in New York City (NYC). As a consequence, the neutron beam enabled cosmic neutron-induced SER measurements with a much shorter irradiation time.

The summary of the radiation test procedure followed at TRIUMF according to the JEDEC standard is below:

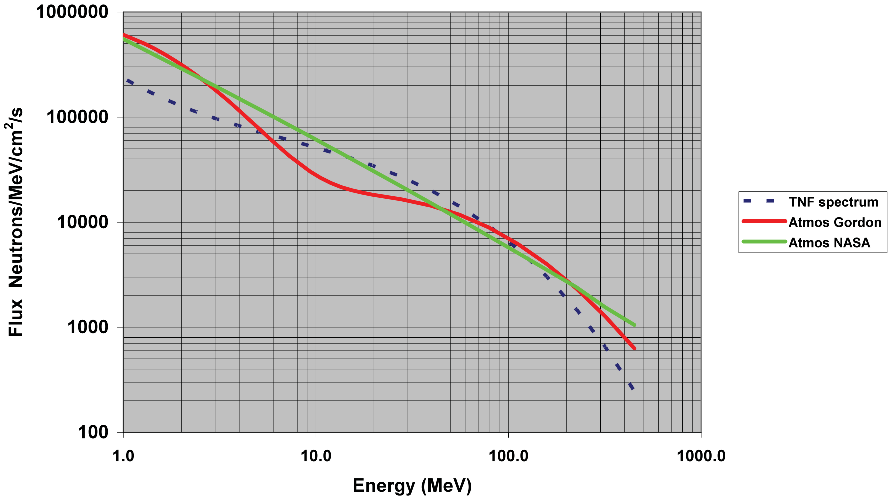


Figure 3.22: Neutron spectrum at TNF compared to the atmospheric spectrum.

1. Set up the equipment and verify connectivity. Further, set up the power supply and ground as the PCB is approximately 7 m away.

2. Perform functional tests (Read/Write) on the chip with the neutron beam on, while the PCB is not in the irradiation path.

3. Irradiate the chip and note the neutron fluence at the Neutron Monitor.

4. Write 1 to the entire address space and read entire address space to verify that the data is written correctly.

5. For two and half hours, read the entire address space every 30 minutes using the Logic Analyzer. If there are any errors (1 → 0) over this time, the erroneous data are captured by the Logic Analyzer. Analyze the data and count the errors. These errors are referred to as as total errors (1 → 0).

6. After two and half hours of data acquisition, check the chip for any hard errors. A '0' is written over the entire address space followed by a '1' and then, the entire address space is read. If there are some 0s, some hard errors have occurred.

7. In the case of zero hard errors, use the same PCB for further testing; otherwise, use another PCB and repeat steps 1 through 6.

8. Subsequently, write ‘0’ in the entire address space and find soft errors (0 → 1) in two and half hours.

During the test, no hard errors were recorded and thus, the same PCB was used throughout the radiation test. The test was carried out for both 0 → 1 and 1 → 0 transitions. Since the bit-cell is symmetrical in nature and both of the data values (1 and 0) are part of the cell, thus the probability of an upset was identical. The data from radiation test was recorded every 30 minutes using the Logic Analyzer which was consulted at the end of the experiment. The SER in FIT was calculated using the following equation.

$$SER = \frac{1}{a_t a_n} \times 10^9 \times \text{Number of Errors}, \tag{3.2}$$

where $a_t$ is the time of neutron irradiation and $a_n$ is the neutron fluence acceleration factor. The fluence is defined as the number of neutrons per unit area per unit time. The neutron monitor (NM) at the TRIUMF facility was used to calculate $a_n$. For a given time, the fluence is given as ratio of NM count without and with the design under test in front of the beam times the counted neutrons multiplied by the calibration factor (CF). The CF is the TNF calibration factor, which at the time of test was $2.7 \times 10^3$. The $a_n$ is the ratio of fluence at TRIUMF to the atmospheric neutron fluence at New York City. The SER in FIT was then calculated using equation (3.2) once number the bit error count was known from the Logic Analyzer.

The summary of the SER performance of the chip is presented in Table 3.3. The proposed SRAM has zero SBUs at 1.2 V while Clerc [60] reported 147 and Autran [61] has 21. The FIT/Mb for the proposed work is 24× smaller than a conventional SRAM and is 2× smaller than the Quatro cell [25] designed in 90-nm. Even at scaled voltage the SER for the 8T increases only from 0.975 FIT to 1.34 FIT. No multi-bit upsets were observed in the 8T in this experiment.

## 3.9 Soft Error Rate and Critical Charge

Typically, SRAM vulnerability to soft errors is evaluated with the help of critical charge ($Q_{crit}$). The $Q_{crit}$ depends upon many factors such as transistor size, substrate doping,

Table 3.3: Comparison of Radiation Test Results of SRAMs

| Source | Year | Technology | VDD(V) | Bit-cell | SBU | MBU | FIT | FIT/Mb |
|---|---|---|---|---|---|---|---|---|
| This work | 2012 | 65 nm | 1.2 | 8T | 0 | 0 | 0 | 0 |
| | | | 1.0 | | 2 | 0 | 0.975 | 31.2 |
| | | | 0.8 | | 4 | 0 | 1.34 | 42.94 |
| Clerc [60] | 2012 | 65 nm | 1.2 | 6T | 147 | - | - | - |
| | | | 0.35 | | 1155 | - | - | - |
| Autran [61] | 2012 | 40 nm | 1.1 | SP-RAM1 | 21 | 19 | - | 759 |
| | | | | SP-RAM2 | 20 | 36 | - | 747 |
| | | | | DP-RAM | 5 | 3 | - | 459 |
| Fuketa [62] | 2011 | 65 nm | 1.0 | 10T | $r$[a] | 0 | - | - |
| | | | 0.3 | | $7.8r$ | 0 | - | - |
| Jahinuzzman [25] | 2009 | 90 nm | 0.9 | 10T | - | 0 | - | 60 |

[a] where $r$ is the number of SBUs at 1.0 V.

carrier mobility, the voltage at the collecting node and the nodes in the periphery of collecting node [63], [64], [65]. The SER exhibits exponential relationship with $Q_{crit}$ [64] and is given by the following empirical model:

$$SER \propto FA \times exp\left(-\frac{Q_{crit}}{Q_s}\right) \tag{3.3}$$

where $F$ is the neutron flux, in $particles/cm^2 - s$; $A$ is the sensitive area of the the circuit, in $cm^2$; and $Q_s$ is the charge collection efficiency of the device, in $fC$. The charge collection depends upon the process parameters, node capacitance and supply voltage and hence, is an important information for SER estimation. Equation 3.3 can be written as:

$$SER = KFA \times exp\left(-\frac{Q_{crit}}{Q_s}\right) \tag{3.4}$$

where $K$ is a proportionality constant. For a given technology node, $K$ and $Q_s$ are constant. If the SER of an SRAM bit-cell is known through radiation test for different $Q_{crit}$ values, the SER of another bit-cell in the same technology can be estimated.

The unknowns of the equation (3.4)can be extracted by taking natural logarithm on both sides of (3.4) and rearranging,

$$ln\left(\frac{SER}{FA}\right) = \left(-\frac{1}{Q_s}\right) Q_{crit} + lnK \tag{3.5}$$

Equation (3.5) is of the form $y = mx + c$, where $m$ and $c$ are the unknowns $(-1/Q_s)$ and $(\ln K)$ and which can be extracted from the plot of (3.5).The SER estimation with this procedure require a few data points. This option is expensive in terms of chip area and hence it was not exercised in the test chip. The results of the 8T as presented in Table 3.3 and compared with other hardened cells such as one proposed by [25] shows a significant improvement in the SER.

## 3.10 Summary

This chapter presented an improved SRAM architecture and an 8T bit-cell. The cell metrics were evaluated and its soft error robustness was analyzed. The 8T demonstrated higher soft error robustness, smaller leakage and read currents. Test chip measurement results show that the 8T bit-cell can be operated at a $V_{DD}$ as low as 0.55 V. Additionally, the bit-line and word-line voltage scaling can be used to reduce power. Radiation test results show that the 32 kb 8T SRAM has zero FIT at 1.2 V and a FIT of less than 1 at 1.0 V at an improved cost in area as compared to other robust SRAM bit-cells.

# Chapter 4

# Robust Sense Amplifiers for Low-Voltage SRAM

## 4.1 Introduction

In SRAMs, the memory cells can generate only small current and voltage signals. Hence, a sense amplifier is employed to read and amplify the signal stored in the selected memory cell. Factors that determine the suitability of an SA include sensing delay, power consumption, die area and resolution [52]. Amongst all these factors, the sensing delay and resolution of the read operation are the most important [66]. Scaling continues to have a profound impact on the design, packing density and operational speed of SRAMs. However, scaling has also resulted in increased process variation due to random dopant fluctuation, line edge roughness, oxide thickness fluctuations, and proximity effects [67], [68], [69], [70], [71]. These factors lead to within-die variations and matched pairs of transistors are affected. Simply increasing SA transistor sizes to reduce mismatch ([49], [50]) will increase the capacitive loading and thus can slow down the sensing. When minimum sensing delays are required, it has been shown that the current latch-type SA (CLSA) is preferred over the voltage-mode SA (VSA) [72], [73] which are shown in Fig. 4.1a and Fig. 4.1b respectively. When high resolutions are required, it is important to minimize the SA's input referred offset voltage ($V_{OS}$). The $V_{OS}$ of an SA is largely determined by the threshold mismatches of the sensing and input transistors [74], [75], [76], [77], [78].

Unfortunately, aggressive device scaling has resulted in increased device variations, which leads to increased threshold mismatches [71], [70], [69], [68], [67]. Consequently, to enable the design of SAs with minimum sensing delays and high resolutions, the effect of $V_{th}$ mismatches in CLSAs must be reduced.

The $V_{OS}$ is defined as the voltage that must be applied between the two inputs of a differential amplifier to obtain zero volts at the output. In the particular case of a sense amplifier, $V_{OS}$ is the minimum magnitude of the difference in the bit-line voltages to reliably generate the correct output. Consequently, the sense amplifier's $V_{OS}$ determines the sense amplifier's resolution. Input referred offset voltages arise from mismatches in the gain factor, the drain current, the threshold voltage and the layout of the devices used in the SA [79], [80], [81], [82]. Among these contributors, $V_{th}$ mismatch has been identified as the dominant contributing factor to large $V_{OS}$ [74], [83], [84], [85]. In particular, the $V_{th}$ mismatch between the input transistors is known to cause read failures in SRAMs [75], [76], [77], [78], [86].

The effect of $V_{th}$ mismatches in a CLSA (Fig. 4.1a) is illustrated in Fig. 4.2. The SA is reset when the sense amplifier enable is low. During this time, $P_1$ and $P_2$ are on and $N_5$ is off. This causes the output nodes, $V_1$ and $V_2$, to go high, setting both outputs to logic 1 and reducing the currents in all the other devices to zero. Then, during sensing, $P_1$ and $P_2$ are turned off, releasing the output nodes, while $N_5$ is turned on to power the sense amplifier and the bit-lines, $BL$ and $\overline{BL}$, are connected to the gates of $N_1$ and $N_2$. Consequently, a differential voltage on the bit-lines causes an imbalanced current to flow in the cross-coupled inverters formed by $N_3$ and $P_3$ and by $N_4$ and $P_4$, which then quickly amplifies the imbalance to the full logic levels due to positive feedback. Ideally, $N_1$ matches $N_2$, $N_3$ matches $N_4$ and $P_3$ matches $P_4$. The effect of $V_{th}$ mismatches between these pairs of devices is shown in Fig. 4.2 and Fig. 4.3. As can be seen in the figure, it is very important to minimize any $V_{th}$ mismatches in $N_1$ and $N_2$, the input devices and in $N_3$ and $N_4$, the sensing devices. The reason that mismatches between $P_3$ and $P_4$ are relatively unimportant is that by the time these devices turn on, the decision has largely already been made by $N_3$ and $N_4$. Consequently, $V_{th}$ mismatches between the input devices and between the sensing devices largely determine the SA's $V_{OS}$.

The mismatch in $\beta$ will affect the rate at which the output is developed, but it does

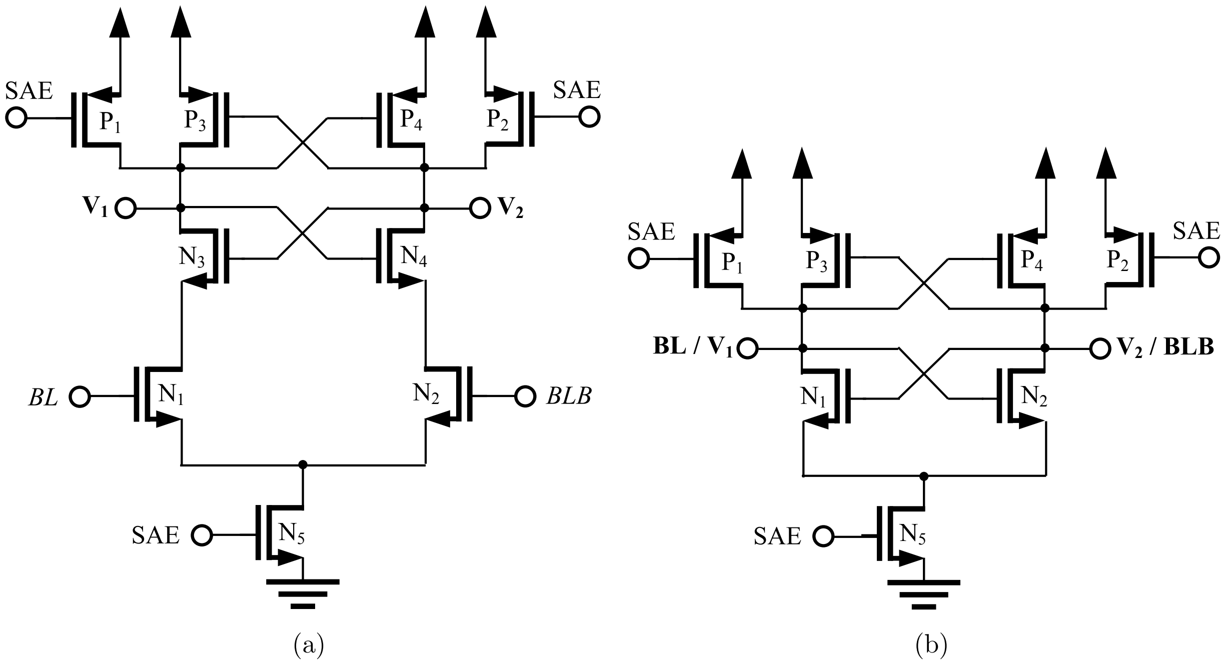


Figure 4.1: (a) Current latch-type sense amplifier schematic. (b) Voltage mode sense amplifier schematic.

not affect the decision making ability of the sense amplifier. The $I_{ds}$ mismatch affect is not significant as initially only subthreshold current flows when the sense amplifier is enabled. Thus, the $\beta$ and $I_{ds}$ mismatch is not significant for analyzing the smallest bitline swing that a sense amplifier will require to produce a correct decision.

The simplest way to reduce a SA's $V_{OS}$ is to reduce $V_{th}$ mismatches by increasing the size of the devices [49], [50]. Unfortunately, $V_{th}$ mismatches are inversely proportional to the square root of the effective channel area (i.e. $1/\sqrt{WL}$ ). Also, increased area increases the gate capacitance and hence the input capacitance, which negates the effect of bit-line delay reduction. Consequently, significant increases in die area, bit-line loading and power dissipation are required to achieve a meaningful reduction in the SA's $V_{OS}$.

A number of more practical methods have been proposed in the literature to address the $V_{th}$ mismatch problem in sense amplifiers. One approach is to add additional devices to

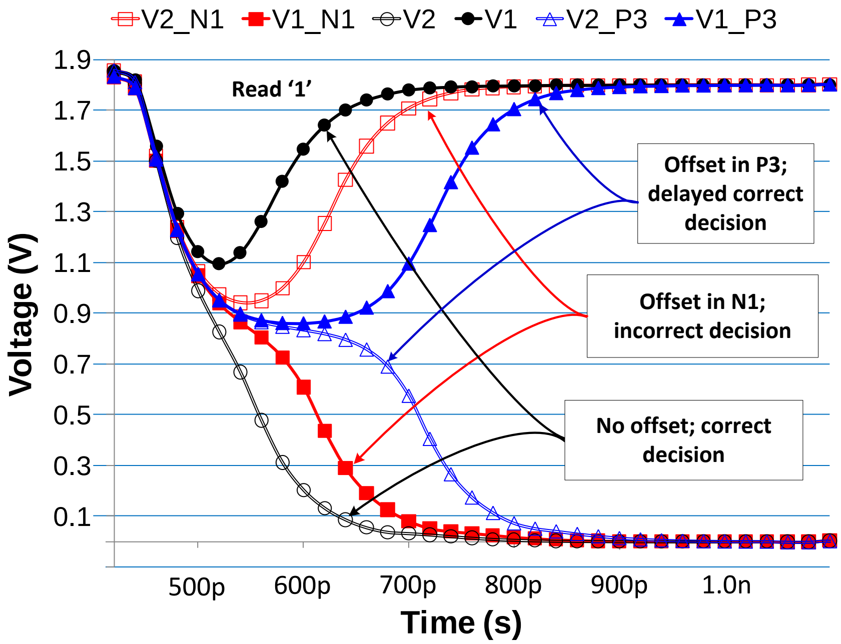


Figure 4.2: Simulation results for the CLSA under different conditions of offset between transistors pairs.

provide a feedback mechanism that reduces the SA's sensitivity to $V_{th}$ mismatches [87]. A dynamic current offset calibration sense amplifier is implemented using capacitors, current mirrors and bias circuits by J. Takahashi et al. [88]. The capacitors are charged before the sense operation begins and in the sensing window they maintain the gate voltage. The sense amplifier makes a decision by detecting the current difference. Y. Watanabe et al. in [84] presented a method to compensate the offset voltage caused by the threshold voltage mismatch at the input of a current mirror sense amplifier. In this scheme one of the bit-lines ($BL$) is connected to the output of the sense amplifier during pre-charge while the other is connected to a reference voltage, thus $BL$ is pre-charged to a level different from $\overline{BL}$ before sense operation starts and thus compensates for the mismatch. A direct sense nMOS only sense amplifier for DRAMs by T. Kawahara et al. in [89] compensate the threshold voltage mismatch by discharging data line capacitance corresponding to the $V_{th}$ of the sense transistors. Additional pair of transistors connect the drains of the sense transistors to their gates which are already connected to the bit-lines ($BL, \overline{BL}$). The

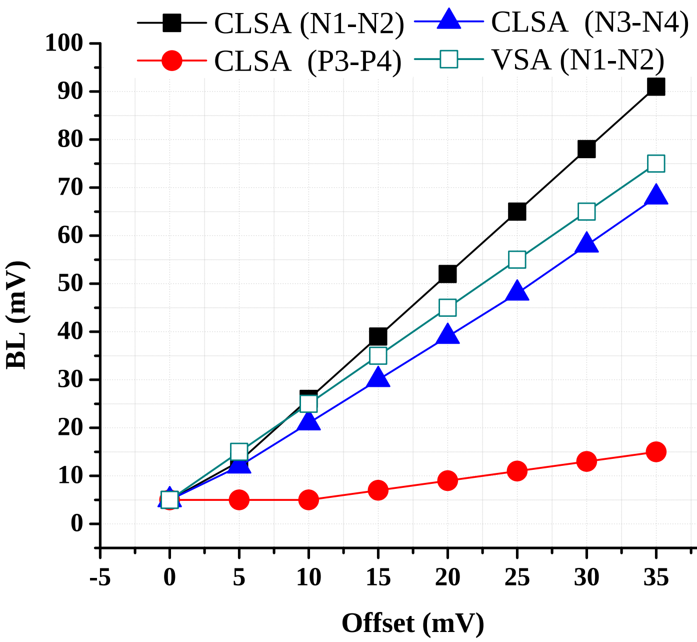


Figure 4.3: Simulation results showing CLSA and VSA outputs for different levels of mismatch between various transistor pairs ($N_1$ and $N_2$, $N_3$ and $N_4$, and $P_3$ and $P_4$). Y-axis shows the smallest bitline differential voltage required to make a correct decision for a given offset between a transistor pair.

effective mismatch reduction depends on the time for which transistors are diode connected. Threshold voltage mismatch of the paired sense transistors is compensated by T. Furuyama et al. [90] by diode connecting them to bit-lines thereby adjusting the bit-line pre-charge levels corresponding to mismatch.

In [91], K. Ishibashi et al. employed a closed loop differential amplifier to implement an offset-voltage-insensitive current sense amplifier. In this implementation, bitlines are pre-charged to a reference voltage ($V_{dd} - V_{th}$). As long as the sensing transistors are in saturation, the sense amplifier is insensitive to offset voltage in the differential amplifier. An offset compensation technique that slows the rise time of the sense enable signal in a latch type sense amplifier [92] is presented by R. Singh et al. in [93]. M. Bharavgava et al. introduced a post silicon digital offset compensation technique in [94] using a pair of registers, transistors, and capacitors for a latch type sense amplifier [92] and a transistor and register pair for a strongARM sense amplifier [95]. For the latch type sense amplifier, switched capacitors are used to control the regenerative feedback in the cross coupled transistors. The capacitor slows down the faster side by increasing the capacitance that need to be discharged. In the strongARM implementation the weaker NMOS sense transistor is assisted with a parallel device controlled through a register. In [96], M. Sharifkhani et al. presented a circuit technique to cancel the $V_{th}$ mismatch between column mux transistors. The technique works in three stages: pre-amplification, access and evaluation through which it balances the $g_m$ of the column mux transistors and thus, compensates for the offset.

The techniques of compensation reported in the literature focus on a certain aspect of offset in the design e.g., capacitance, pre-charge, column mux transistors and decision making pairs. The offset compensation usually involves cost in terms of timing, area and design complexity. Additional transistors and/or capacitors comprising the compensation circuitry can have inherent offset which can compromise the effectiveness of a given solution scheme. Often, VSA is employed with long bitline development time to tolerate sense amplifier offset at the cost of a reduced read speed; however, CLSA achieves faster reads as compared to VSA [72]. In the proposed offset cancelling sense amplifier (SAOC), we describe a current mode area-efficient offset cancellation scheme that takes into account the offset between the sensing transistors. Further, the proposed scheme do not incur any timing penalty.

## 4.2 An Offset Cancelling Sense Amplifier

The ground-referenced configuration of the proposed SAOC is shown in Fig. 4.4 and the device sizes are listed in Table 4.1. The proposed SAOC can also be implemented in a supply-referenced configuration, for which the bit-lines are expected to be at or near the positive supply, $V_{DD}$. The ground-referenced configuration was chosen for this work to allow the $V_{th}$ of the input transistors to be controlled through the devices' n-well potentials in Section 4.2.2 and Section 4.2.3.

The SAOC's timing diagram is shown in Fig. 4.5 and a conceptual schematic to generate the timing signals is shown in Fig. 4.6. A read cycle begins with a pre-charge phase (PC)

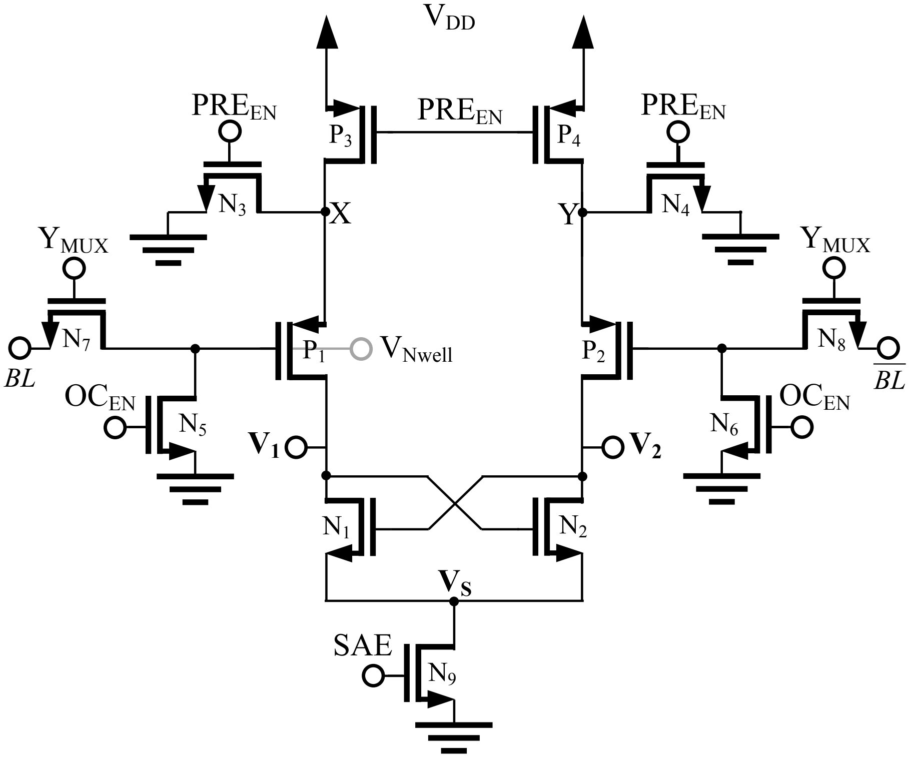


Figure 4.4: Offset cancellation sense amplifier schematic.

followed by a pre-discharge phase (PD). During the PC and PD phases, the bit-lines are pre-charged and the bit-line voltage is allowed to develop. Consequently, the PC and PD phases do not add time to the read operation. After this, the sensing or evaluation phase (EV) occurs. The EV phase includes a short data acquisition phase ($Y_{MUX}$) followed by sufficient time for the SAOC to resolve the data and cancel the offset of the sense transistors.

While the offset cancellation operation occurs during the EV phase, the PC and PD phases are first necessary to initialize the SAOC's node voltages. During the PC phase, nodes $V_1$ and $V_2$ are pre-charged to $V_{DD}$ by keeping $OC_{EN}$ high and $PRE_{EN}$ low. Then, during the PD phase, $PRE_{EN}$ goes high, turning off transistors $P_3$ and $P_4$ and turning on

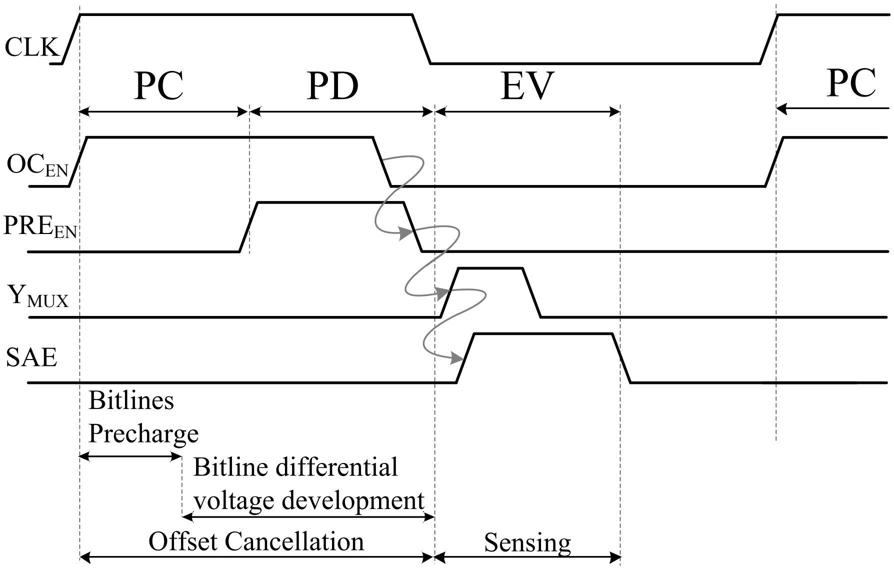


Figure 4.5: Timing waveform for the SAOC (not to scale). Typical delays between the falling edges of $OC_{EN}$ and $PRE_{EN}$ and the falling edge of $PRE_{EN}$ and the rising edge of $Y_{MUX}$ and the rising edges of $Y_{MUX}$ and $SAE$ are 30 ps.

transistors $N_3$ and $N_4$. Thus, nodes $V_1$ and $V_2$ are connected to ground through transistors $P_1$, $N_3$ and $P_2$, $N_4$, respectively. Note the roles of $P_1$'s and $P_2$'s drains and sources are reversed during this time. Nodes $V_1$ and $V_2$ discharge until $P_1$ and $P_2$ turn off. At this point, node $V_1$ will be at $-V_{thP1}$ or $|V_{thP1}|$ where $V_{thP1}$ is the threshold voltage of P1. Also, node $V_2$ will be at $-V_{thP2}$ or $|V_{thP2}|$ where $V_{thP2}$ is the threshold voltage of $P_2$. Hence, before the EV phase, the gates of the sensing devices, $N_1$ and $N_2$, are pre-charged with the threshold voltages of $P_2$ and $P_1$, respectively. Subsequently, in the EV phase the bit-lines are connected to the input devices by turning on $N_7$ and $N_8$ with the $Y_{MUX}$ control signal.

While the above procedure only compensates for the mismatch in the input transistors, as shown in Fig. 4.3, the sense amplifier's offset is dominated by the input transistors. Furthermore, transistors $N_3 - N_6$ along with $P_3$ and $P_4$ are only used as switches and are off during the decision making process, hence, their mismatches do not affect the decision.

Table 4.1: Typical Device Sizes for the SAOC Implementation

| Device Name | Size (nm) |
|---|---|
| $P_1$, $P_2$, $N_3$, $N_4$, $N_5$, $N_6$ | 500 |
| $N_7$, $N_8$ | 750 |
| $P_3$, $P_4$, $N_1$, $N_2$, $N_9$ | 1000 |

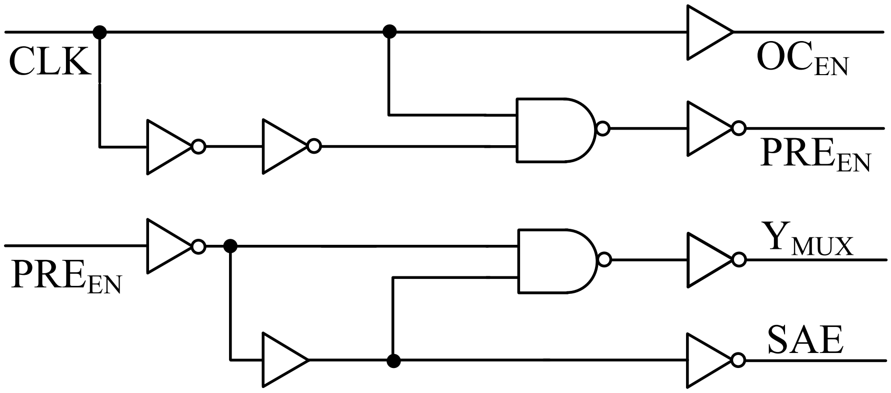


Figure 4.6: Timing schematic used to develop the waveforms of Fig. 4.5.

### 4.2.1 Analysis

The simulations and experimental results are provided in the following sections to illustrate the effectiveness of the offset cancellation operation, a small-signal analysis is used in this section to provide some insight into the key parameters of the SAOC. The analysis depends on the sense amplifier's node voltages at the start of the EV phase and on the circuit's response to any voltage imbalances.

Prior to the start of the EV phase, the PC and PD phases have pre-charged nodes $V_1$ and $V_2$ to $|V_{thP1}|$ and $|V_{thP2}|$, respectively. Then, at the start of the EV phase, devices $P_3$ and $P_4$ pull nodes X and Y to $V_{DD}$, device $N_9$ pulls node $V_S$ to ground and devices $N_7$ and $N_8$ connect the sense amplifier's inputs (i.e., the gates of $P_1$ and $P_2$) to the bit-lines. Consequently, the sense amplifier can be simplified to the circuit shown in Fig. 4.7.

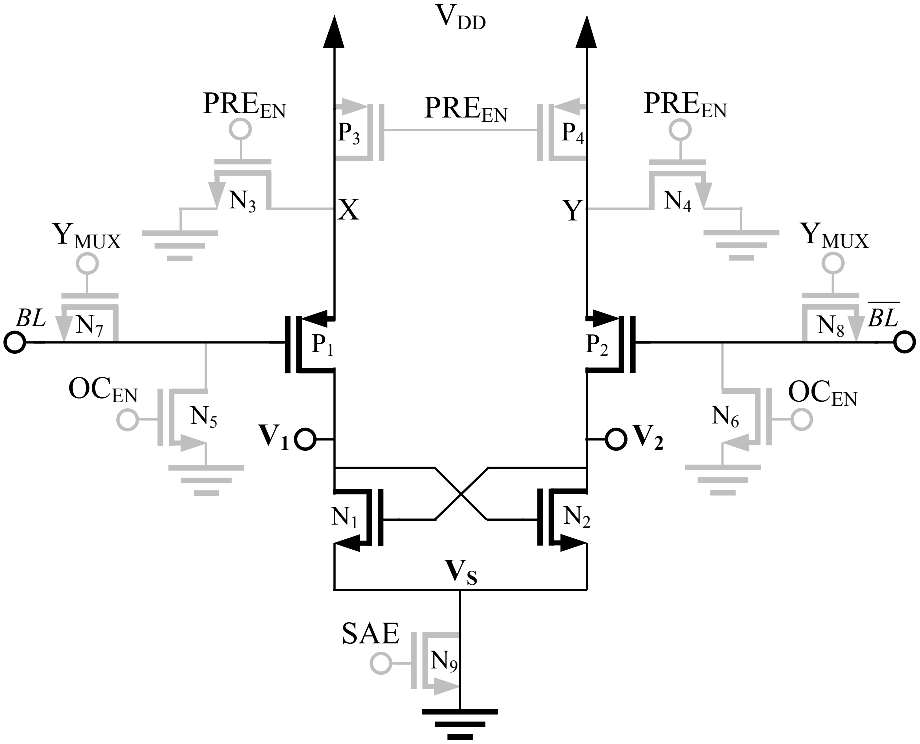


Figure 4.7: Voltages at the core of sense amp at the start of the evaluation phase.

To simplify the analysis, the bit-line voltages and the threshold voltages of $P_1$ and $P_2$ can be re-expressed using differentials. For the bit-lines,

$$V_{BL} = V_B + \frac{\Delta V_B}{2} \tag{4.1}$$

and

$$V_{\overline{BL}} = V_B - \frac{\Delta V_B}{2} \tag{4.2}$$

where $V_B$ is the average or common-mode bit-line voltage,

$$V_B = \frac{(V_{BL} + V_{\overline{BL}})}{2} \tag{4.3}$$

and $\Delta V_B$ is the difference between the bit-line voltages,

$$\Delta V_B = V_{BL} - V_{\overline{BL}} \tag{4.4}$$

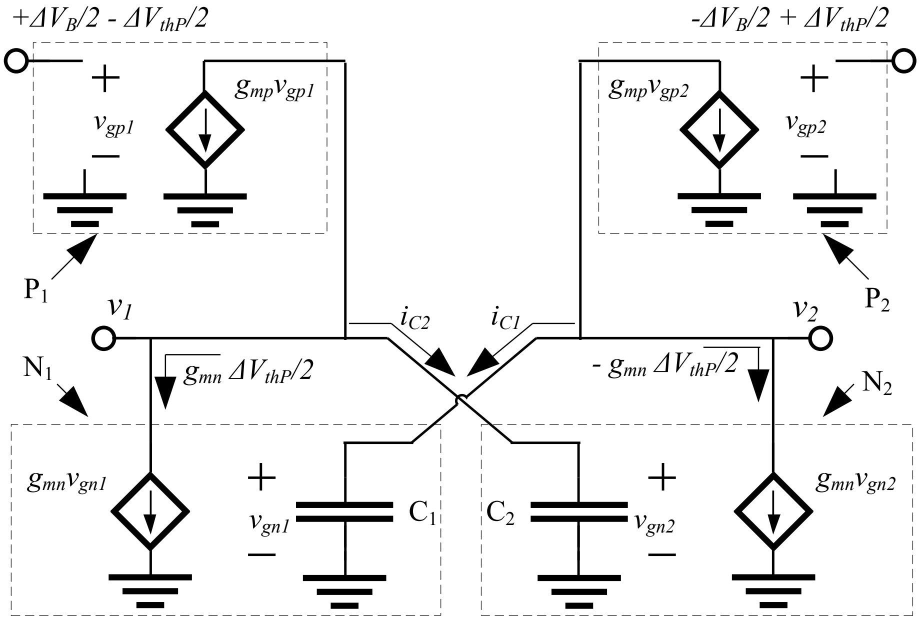


Figure 4.8: Small signal model of the sense amplifier at the beginning of the evaluation phase.

Similarly, for the threshold voltages,

$$V_{thP1} = V_{thP} + \frac{\Delta V_{thP}}{2} \tag{4.5}$$

and

$$V_{thP2} = V_{thP} - \frac{\Delta V_{thP}}{2} \tag{4.6}$$

where $V_{thP}$ is the average or mean threshold voltage,

$$V_{thP} = \frac{(V_{thP1} + V_{thP2})}{2} \tag{4.7}$$

and $\Delta V_{thP}$ is the difference between the threshold voltages,

$$\Delta V_{thP} = V_{thP1} - V_{thP2} \tag{4.8}$$

Equations 4.1 through 4.8 allow the common-mode and difference voltages to be analyzed separately.

When only the common-mode signals are considered, the sense amplifier, as shown by Fig. 4.7, is perfectly balanced. The sense amplifier's inputs are biased at $V_B$ and the sense amplifier's outputs are biased at $-V_{thP}$ or $|V_{thP}|$. These biasing levels can be used to determine the parameters of the sense amplifier's small signal model.

The small-signal model of Fig. 4.7, including the threshold mismatches of $P_1$ and $P_2$, is shown in Fig. 4.8. Based on equations 4.1 and 4.2, the gates of $P_1$ and $P_2$ will see small-signal voltages of $+\Delta V_B/2$ and $-\Delta V_B/2$ respectively. In addition, based on equations 4.5 and 4.6, the small-signal voltage at node $V_1$, $v_1$, will be $-\Delta V_{thP}/2$ and the small signal voltage at node $V_2$, $v_2$, will be $+\Delta V_{thP}/2$. In the small signal model, it has been assumed that $N_1$ and $N_2$ match. Hence, $g_{mn1} = g_{mn2} = g_{mn}$. Furthermore, $P_1$ and $P_2$ are assumed to be identical, except for a threshold voltage mismatch. Consequently, by setting $g_{mp1} = g_{mp2} = g_{mp}$, the threshold mismatch can be accounted for by adding a small-signal voltage equal to $-\Delta V_{thP}/2$ to the gate of $P_1$ and a small-signal voltage equal to $+\Delta V_{thP}/2$ to the gate of $P_2$. Included in the small-signal model are capacitors $C_1$ and $C_2$ which model the total capacitance at nodes $v_1$ and $v_2$ respectively.

In Fig. 4.8, the currents due to $N_1$, $N_2$, $P_1$ and $P_2$ at time zero are indicated. The currents due to $P_1$ and $N_1$ determine the current flowing onto $C_2$ or the gate of $N_2$. At time zero, $P_1$ supplies a current $i_{P1}$ of value

$$i_{P1} = g_{mp}\left(\frac{\Delta V_{thP} - \Delta V_B}{2}\right) \tag{4.9}$$

into node $v_1$. Note that the current supplied by $P_1$ is determined by both the input signal (i.e., $+\Delta V_B/2$) and the share of threshold voltage mismatch on $P_1$ (i.e., $-\Delta V_{thP}/2$). $N_1$ draws a current $i_{N1}$ of value

$$i_{N1} = g_{mn}\frac{\Delta V_{thP}}{2} \tag{4.10}$$

out of node $V_1$. Note that the current drawn by $N_1$ is determined by the share of the threshold voltage mismatch on $P_2$ (i.e., $+\Delta V_{thP}/2$). Consequently, the net current flowing onto $C_2$, $i_{C2}$ is found to be

$$i_{C2} = (g_{mp} - g_{mn})\frac{\Delta V_{thP}}{2} - g_{mp}\frac{\Delta V_B}{2} \tag{4.11}$$

where it can be seen that if $g_{mp} = g_{mn}$ the charging current is only determined by the bit-line signal. Similarly, the current flowing onto $C_1$, $i_{C1}$ is found to be

$$i_{C1} = g_{mp}\frac{\Delta V_B}{2} - (g_{mp} - g_{mn})\frac{\Delta V_{thP}}{2} \tag{4.12}$$

Once again, if $g_{mp} = g_{mn}$ the charging current is only determined by the bit-line signal. If the pre-discharge phase is not used, nodes $v_1$ and $v_2$ will both be equal to $V_{DD}$ and the charging current will be found to be

$$i_C = g_{mp}\left(\frac{\Delta V_B - \Delta V_{thP}}{2}\right) \tag{4.13}$$

Consequently, the SAOC effectively cancels the effects of a threshold mismatch in the input devices.

The above analysis (equations 4.11 and 4.12) shows that if $g_{mp} = g_{mn}$ the charging current is only determined by the bit-line signal. Unfortunately, it is highly unlikely that the two transconductances will match. To determine the range over which the proposed threshold voltage mismatch cancellation scheme provides a lower offset than the un-cancelled circuit, it is necessary to ensure that

$$\left|(g_{mp} - g_{mn})\frac{\Delta V_{thP}}{2}\right| \leq \left|g_{mp}\frac{\Delta V_{thP}}{2}\right| \tag{4.14}$$

which can be re-arranged to yield the condition

$$g_{mn} \leq 2g_{mp} \tag{4.15}$$

Consequently, the proposed SAOC will provide offset reduction over a wide range of device parameters.

### 4.2.2 Simulation Results

To validate the effectiveness of SAOC, the behavior of the SAOC is compared with the conventional sense amplifier (CONV). The CONV is realized by removing transistors $N_3$

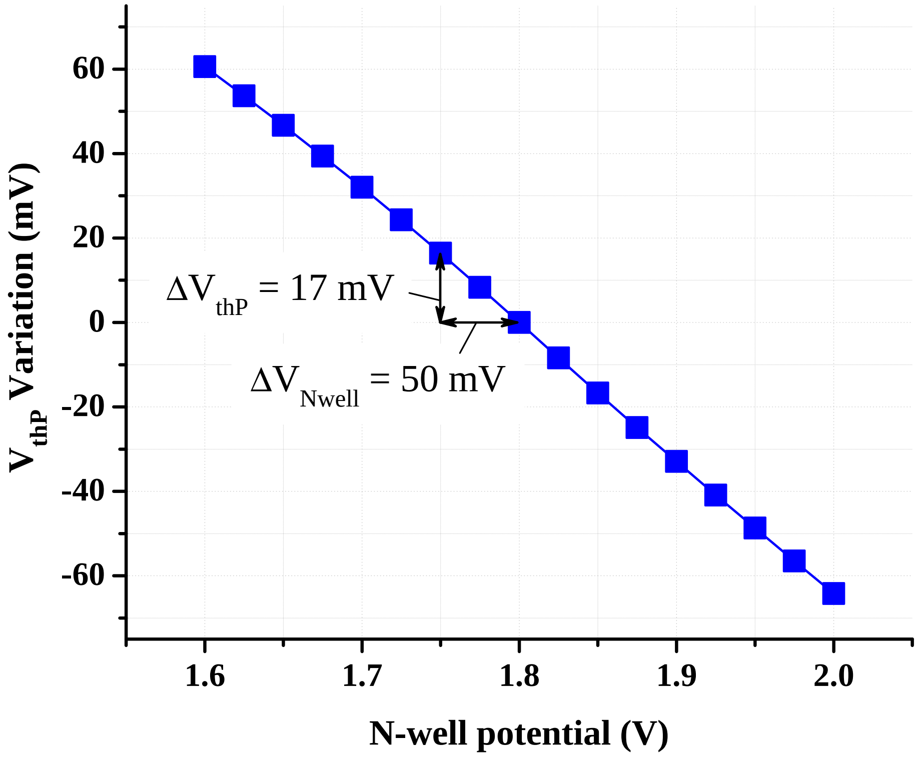


Figure 4.9: Effect of n-well potential on the $V_{th}$ of a PMOS transistor, $V_{DD} = 1.8\ V$.

and $N_4$ from the SAOC as they are only used for offset cancellation. For the simulations, a 180-nm CMOS process was used. Initially, simulations were carried out to determine the effect of an offset in the input transistors by observing the output nodes $V_1$ and $V_2$. Second, the resolution of the SAOC and the CONV were evaluated (the minimum magnitude of the difference in the bit-line voltages to reliably generate the correct output).

To introduce an offset, the n-well potential of one of the sensing transistors ($P_1$ of Fig. 4.4) was controlled. Based on SPICE simulations, a 50 $mV$ change in the n-well potential changes the $V_{thP}$ by approximately 17 $mV$ (Fig. 4.9). This simulation approach was used to enable direct comparisons between the simulations and the measured results later in Section 4.2.3. The SAOC is simulated, with a clock rate of 1 GHz and an offset induced into $P_1$. The n-well potential (having a nominal value of 1.8 $V$) was set to 1.7 V introducing an offset of approximately 34 $mV$. The bit-lines were pre-charged and initialized such that

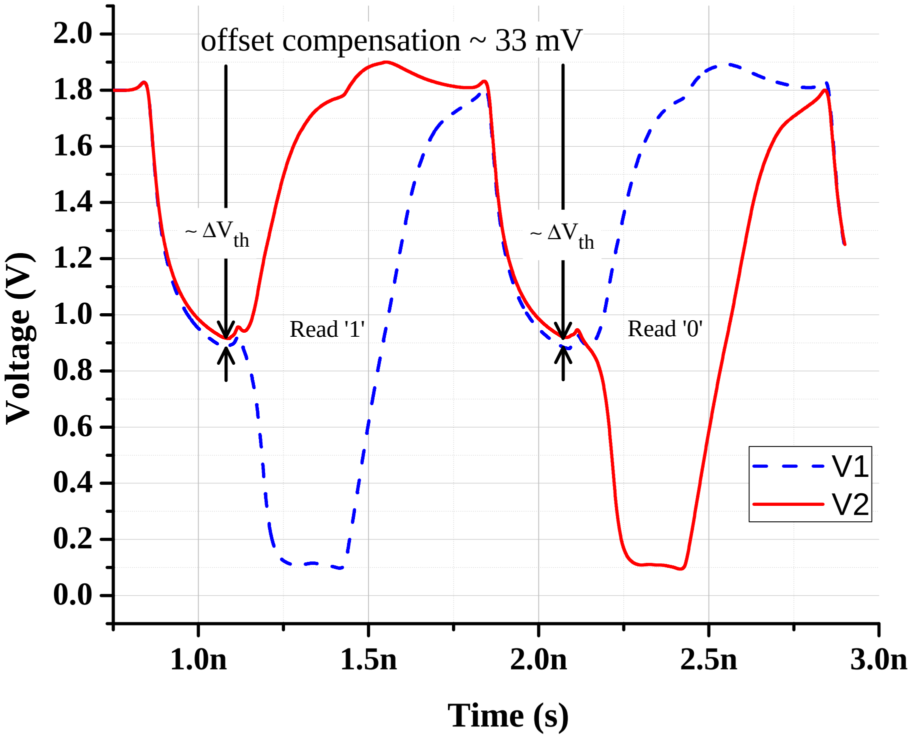


Figure 4.10: SAOC cancelling the induced offset of approximately 34 $mV$ and delivering a correct decision. The offset is induced by setting the n-well potential to 1.7 $V$. The simulations were carried out at 1 GHz.

$BL > \overline{BL}$ in the first clock cycle and $\overline{BL} > BL$ in the next clock cycle so as to eliminate the memory effect at the sensing nodes. The nodes $V_1$ and $V_2$ behave as expected as shown in Fig. 4.10. At the end of the PC/PD phases, $V_1 - V_2$ approximately equals the $V_{th}$ mismatch. In particular, node $V_1$ is set at a lower potential to compensate for the induced offset at the end of the PD phase. The compensated offset is approximately 33 $mV$ which is very close to the introduced offset of approximately 34 $mV$. In Fig. 4.11, the ability of the SAOC to capture the $V_{th}$ mismatch of the input transistors on the sensing nodes ($V_1$ and $V_2$) for a range of $V_{th}$ mismatches is illustrated. It can be observed that the $V_{th}$ mismatch of the input transistors is tracked reasonably well.

The simulation results in Fig. 4.12 show the effect of a $V_{th}$ mismatch in the input devices for both the SAOC and the CONV on the sensing ability of the SAs. The SAOC makes a correct decision for significantly smaller differential bit-line voltages compared to

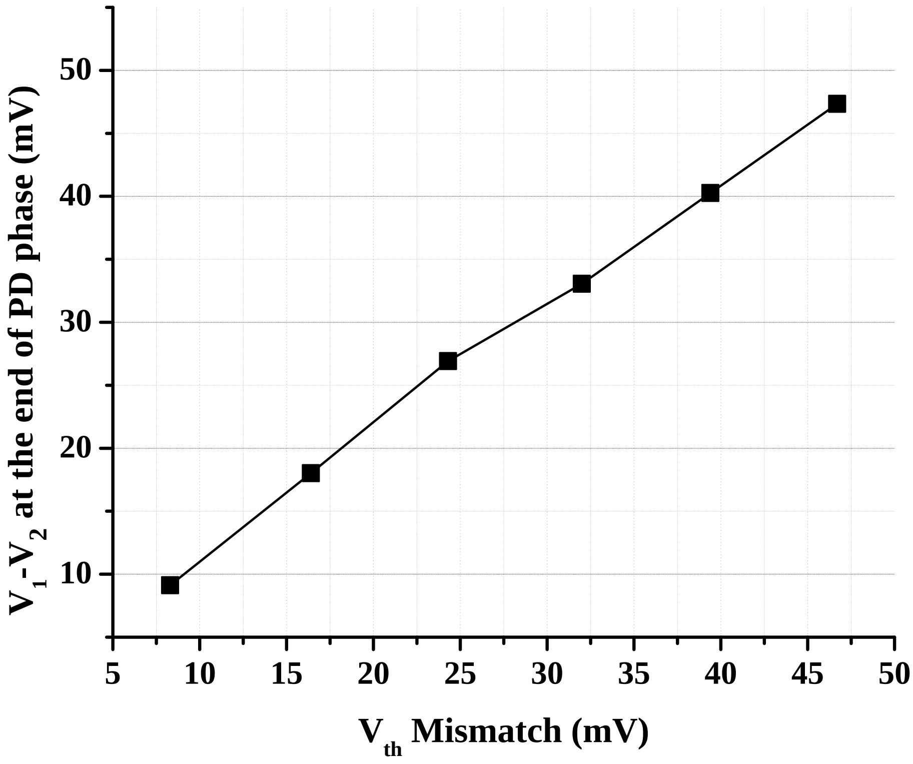


Figure 4.11: Simulation results showing the offset is effectively captured on the sensing nodes ($V_1$ and $V_2$) of the SAOC.

the CONV scheme. Thus, based on simulations, not only is the mismatch of the input devices captured on the sensing nodes, the proposed sense amplifier shows significantly lower offsets and better resolutions.

### 4.2.3 Measurement Results

To verify the advantages of the proposed sense amplifier, a test chip was implemented using a commercially available 180-nm, n-well CMOS process. The SAOC and CONV were implemented on the same die. The chip was then bonded in a CQFP package and mounted on a PCB for testing.

To enable a variable threshold mismatch, the n-well of one of the input transistors in each SA was controlled off-chip. The independent n-well control resulted in a slight area

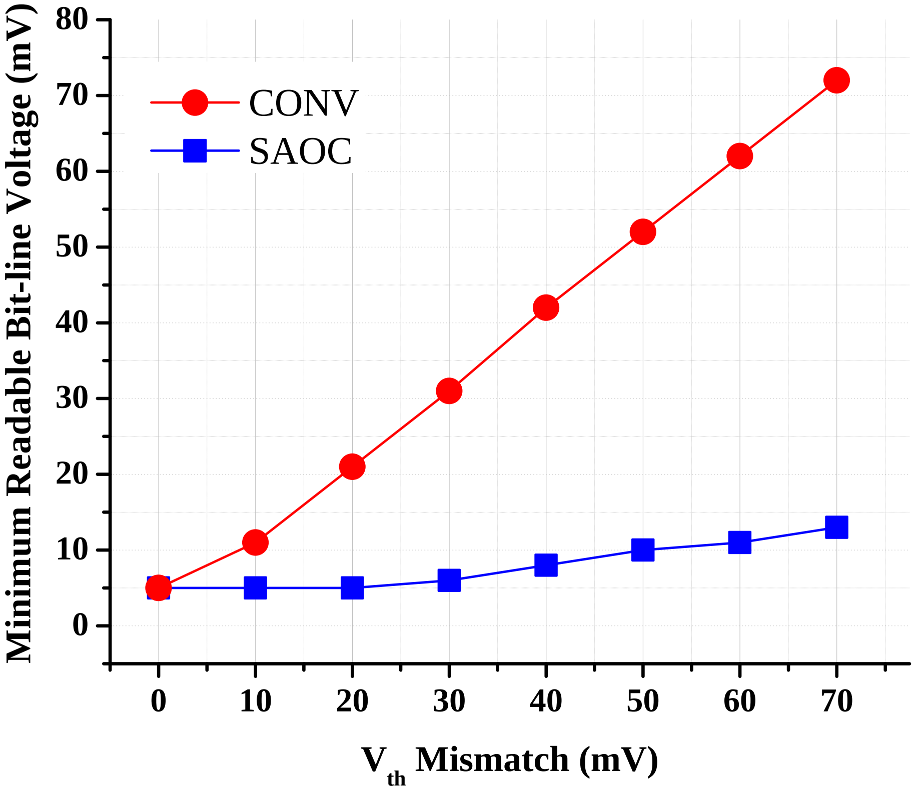


Figure 4.12: Simulation results at 1 GHz showing that the smallest correctly read differential bit-line voltage for varying $V_{th}$ mismatch values.

overhead which was similar for the two designs. The n-wells and bit-lines were driven directly from off-chip voltage sources. A single external clock drove the logic circuit (Fig. 4.6) which generated all necessary clock phases on-chip.

The layout of the SAs and the associated routing was done such that both of the designs see the same parasitics and load. Furthermore, the chip pads were laid out in such a way that they saw similar bond wire lengths and the PCB traces were drawn symmetrically. The SA outputs were buffered but, not latched, to allow the timing behaviour of the SAs to be observed directly on an oscilloscope. The die area was 1 x 2 $mm^2$ and the die photo is shown in Fig. 4.13.

All the measurements were carried out at a 100-MHz clock frequency. The choice of frequency was constrained by the equipment at the test facility. The differential input

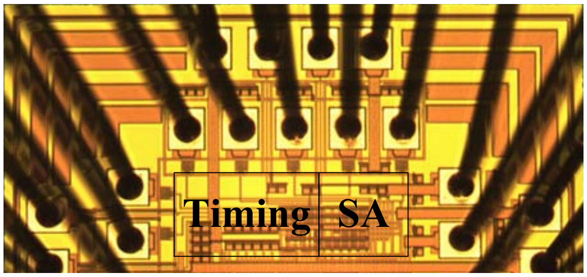


Figure 4.13: Die photo of the 180nm CMOS test chip.

voltages were set through variable potentiometers on the PCB. For each measurement, the $BL$ and $\overline{BL}$ were set with an accuracy of 0.1 $mV$. For a given n-well potential, the SAs are analyzed in two ways. Initially, a higher potential (100 $mV$) is applied on the $BL$ while keeping the $\overline{BL}$ at ground, the SAOC and CONV outputs were analyzed. Then, keeping the $\overline{BL}$ at ground potential, the $BL$ potential was reduced in steps of 1 $mV$. The SAOC and CONV outputs were recorded for different differential input voltages. Once $BL$ reaches ground, it was held there, while $\overline{BL}$ was increased in steps of 1 $mV$ until it reached 100 $mV$. This measurement is termed *$BL$ decreasing and $\overline{BL}$ increasing*. Subsequently, we carried out complementary steps for which the $\overline{BL}$ potential was reduced from the initial voltage of 100 $mV$ in the steps of 1 $mV$ until it reached ground, while the $BL$ voltage was kept at ground. Then $\overline{BL}$ was held at ground and the $BL$ potential was increased and outputs were recorded. This procedure continued until $BL$ reached 100 $mV$ and is termed *$\overline{BL}$ decreasing and $BL$ increasing.*

In Fig. 4.14, results are presented for the SAOC and CONV SAs. The measurements were carried out with an n-well potential of 1.8 $V$ which is the nominal case thus, providing a base-line. The measured offset voltage for the SAOC, $V_{OS-SAOC}$ was approximately $+4mV$ while the measured offset voltage for the CONV, $V_{OS-CONV}$ was $-67mV$. By adjusting the n-well potentials, a $V_{th}$ mismatch shift of approximately $+33\,mV$ was induced in both the CONV and SAOC. For the CONV, a correct decision was made for a differential bit-line voltage of $-35\,mV$. The $-35\,mV - (-67\,mV) = +32\,mV$ shift in offset corresponding to a $+33$ $mV$ shift in $V_{th}$ mismatch of the input devices indicates that the CONV is highly sensitive to $V_{th}$ mismatches. The SAOC on the other hand made a correct decision

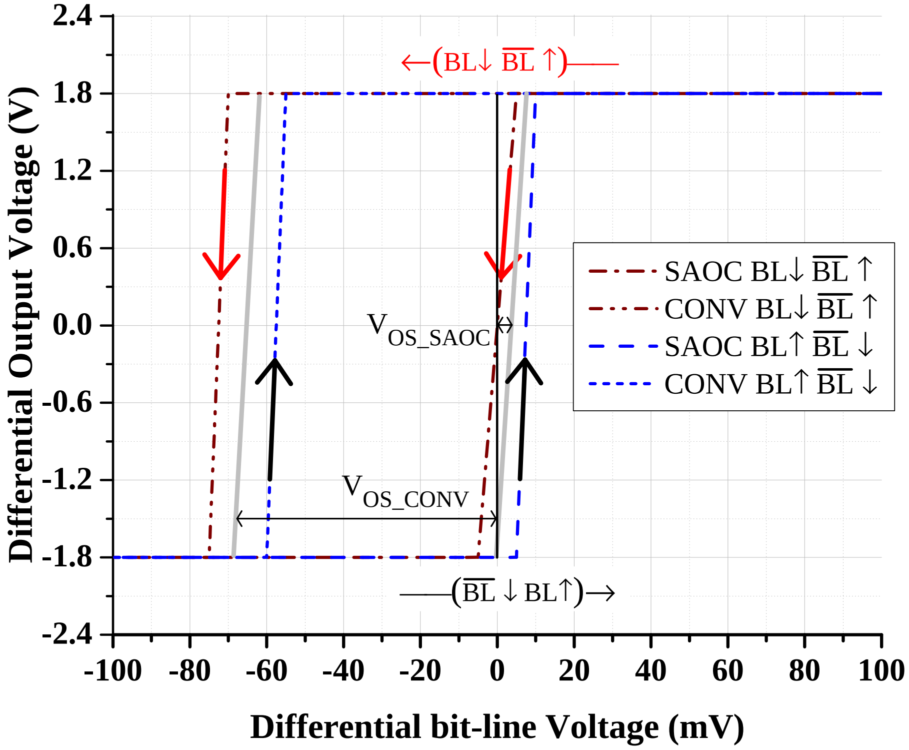


Figure 4.14: Measurement results for the SAOC and CONV schemes for n-well potential of 1.8 $V$.

for a differential input voltage of $+5\ mV$. The 5 $mV$ $-$ 4 $mV$ or 1 $mV$ shift in offset corresponding to a 33 $mV$ shift in the $V_{th}$ mismatch of the input devices indicates that the SAOC is largely insensitive to $V_{th}$ mismatches in the input devices. Based on these observations it is clear that adding the offset cancellation feature in the same sense amplifier improves its resolution.

Further measurements were carried out starting with an n-well potential of 1.7 $V$ that was increased in steps of 0.05 $V$ until it reached 1.9 $V$ implying that the change in $V_{th}$ was approximately 68 $mV$. The measurement methodology was identical to the one presented in Fig. 4.14 and the offset was measured in each case. In Fig. 4.15, the measurement results are summarized for the proposed and conventional schemes for different values of mismatch voltages. The results are based on measurements from three test chips. It is observed that both the SAs displayed an offset. For the SAOC the offset was small and

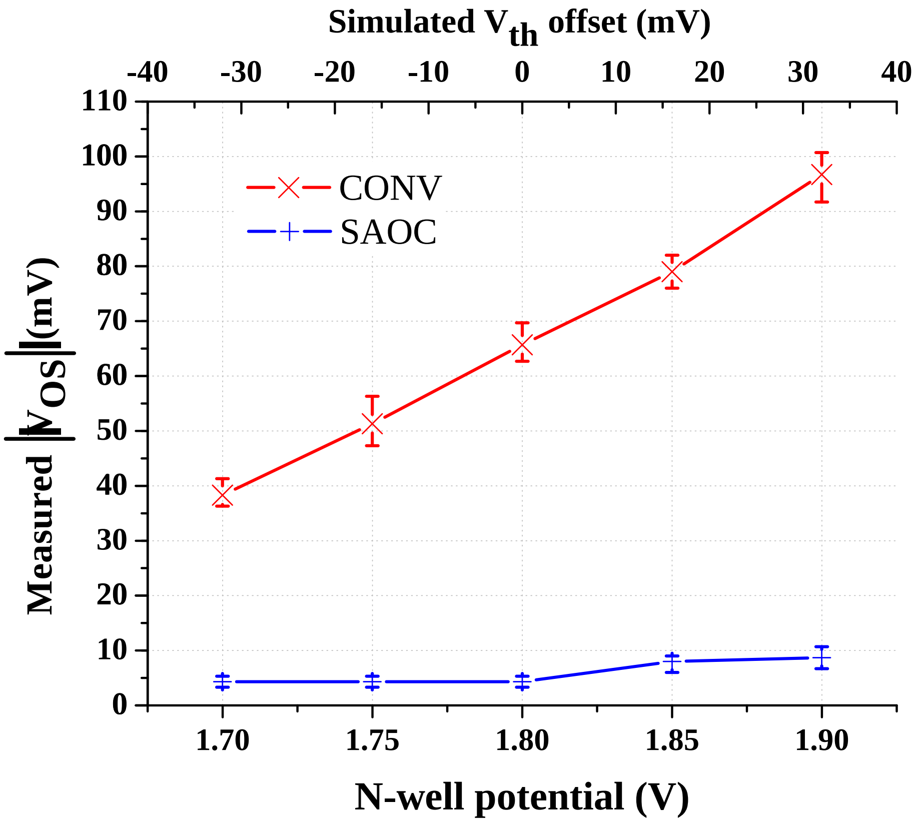


Figure 4.15: Measured $V_{OS}$ for SAOC and CONV for a range of n-well potentials.

largely insensitive to the threshold variations. The CONV had both a large offset and was significantly more sensitive to threshold variations. Thus, the SAOC can reliably detect significantly smaller small bit-line differences than the CONV.

The simulated offsets (Fig. 4.11) and the measured offsets (Fig. 4.15) are fairly similar. For a 200 $mV$ change in the n-well potential, both the simulated and measured offset of the CONV changed by approximately 60 $mV$ while the simulated and measured offset of the SAOC changed by approximately 10 $mV$. Consequently, there is good agreement between simulation and measurement for both SAs, thereby allowing designers to simulate the expected offset reduction provided by the SAOC, with confidence.

The SAOC has two additional transistors that are twice the minimum size, leading to a 7.5% larger area than the CONV. This area increase is significantly smaller than would be required if one were to simply increase the device sizes to minimize the mismatches. The

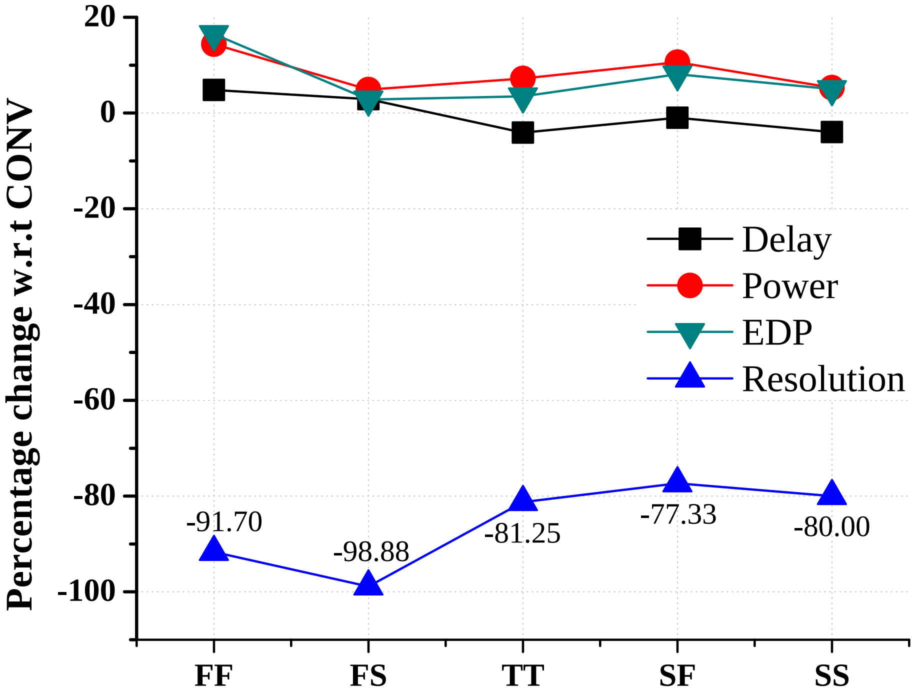


Figure 4.16: Performance comparison of the proposed w.r.t the conventional SA.

additional devices and clocks also lead to a power increase, which when compared directly to the CONV, appears relatively high. However, the additional power consumption in the SAOC does include the cost of generating the $OC_{EN}$ and $PRE_{EN}$ signals. Nevertheless, when the total read power on a per-bit basis is compared, the proposed SA displays only a marginal increase (0.1%) over that of the conventional SA. In addition, due to the order of magnitude improvement in the proposed SA's resolution, smaller bit-line voltages can be used to reduce the required read power. Finally, the proposed SA does not incur a delay penalty. Thus, the proposed SA provides a significant improvement in resolution at minimal cost when compared with the conventional sense amplifier.

Illustrated in Fig. 4.16 is the comparison between the SAOC and the CONV. The SAOC has comparable delay, power, and energy numbers, however, it requires much lower differential signal ($-77.33\%$ in the worst case) to make a correct decision. Two additional transistors in the SAOC lead to 7.5% larger area than the CONV which is significantly smaller than would be required if one were to simply increase the device sizes to minimize

Table 4.2: Comparison of Sense Amplifiers

| Design | Area | Design Node | Offset Reduction | Energy | Sensing Delay |
|---|---|---|---|---|---|
| SAOC | 13T | 0.18 $\mu m$ | ± 35 mV | 0.212 pJ | 0.70 ns |
| Seno [87] | 31T | 0.35 $\mu m$ | 50 mV | 3.734 pJ | 4.40 ns |
| Takahashi [88] | 14T + 2C | 0.5 $\mu m$ | 20 $\mu A$ | 1.62 pJ | 6.50 ns |
| Ishibashi [91] | 27T | 0.25 $\mu m$ | 50 mV | 4.813 pJ | 1.75 ns |
| Bhargava [94] | 9T+2C+2FF | 45 $nm$ | 55 mV | - | - |
| Sharifkhani [96] | 14T | 0.18 $\mu m$ | 40 mV | 0.84 pJ | 1.10 ns |

the mismatches. However, the additional power consumption in the SAOC does include the cost of generating the $OC_{EN}$ and $PRE_{EN}$ signals and is only a fraction of the read power of an SRAM cell. In addition, due to the order of magnitude improvement in the proposed SA's resolution, smaller bit-line voltages can further reduce the required read power. Finally, the SAOC does not incur a delay penalty.

The results from three test chips showed that both the SAs displayed an offset, however, the SAOC was largely insensitive and the CONV had both a large offset and was significantly more sensitive to threshold variations. Thus, the SAOC can reliably detect significantly smaller small bit-line differences than the CONV. For a $200 - mV$ change in the n-well potential, both the simulated and measured offset of the CONV changed by approximately 60 $mV$ while the simulated and measured offset of the SAOC changed by approximately 10 $mV$. Consequently, there is good agreement between simulation and measurement for both SAs, thereby allowing designers to simulate the expected offset reduction provided by the SAOC, with confidence.

In Table 4.2, the SAOC is compared to other SAs designed for improved resolution. While all of the proposed techniques improve the SA's resolution, the proposed SA does so with significantly fewer devices, smaller delay and energy. Consequently, the proposed SA is an area-efficient offset cancellation method for SRAM sense amplifiers in scaled technologies.

An SRAM bit-line sense amplifier is proposed with offset cancellation capability. Measurements at 100 MHz for a 180-nm CMOS test chip show that the proposed sense amplifier makes a correct decision with 10 $mV$ differential inputs with an induced offset of 35 $mV$ when the CONV requires an input of 101 $mV$. The ability of the SAOC to detect a small bit-line swing translates into smaller read currents which leads to power saving, enhanced stability and improved yield.

## 4.3 Dual-Input Sense Amplifier Architecture

In a read operation, the contents of the memory are determined by a sense amplifier. In SRAM, there are two methods of sensing, voltage and current. For an SRAM bit-cell, the differential column architecture is more common and thus differential sense amplifiers are more prevalent. The differential sense operation begins with pre-charging the bit-lines to $V_{DD}$ and then they are allowed to float. During this time, a current is drawn by the selected cell the side where a 0 is stored. The cell current $i_{cell}$ discharges the bit-line capacitance and causes a voltage drop of $\Delta BL$ on that bit-line. In effect, there is a differential voltage developed between the bit-lines given by $V_{DD} - \Delta BL$. Ideally, there will be no current flowing for the other bit-line connected to side of the bit-cell where a 1 is stored. A sense amplifier in this case can be as simple as an differential voltage amplifier followed by an inverter to provide a full swing. Two major issues with voltage sensing are:

- Relatively slow because a large bit-line capacitance has to be discharged.
- Read time depends upon the size of the array.

Despite these drawbacks, voltage sensing is commonly used [97].

In current sensing, on the other hand, the current from the memory cell as an input signal and provides an output voltage proportional to the cell current. The output voltage ($V_{out}$) is now evaluated by a voltage sense amplifier. The current sensing stage and the voltage sense amplifier forms a current sense amplifier. The current sensing stage needs a low input resistance and a bias current. The bias current causes a voltage drop across the bit-line and thus pre-charge circuitry is needed. The advantage in terms of speed is attractive but, it comes at the expense of additional power consumption by the bias current and additional area which makes the design relatively less popular. In this section a new sense amplifier architecture is presented which requires a smaller differential voltage to make a sensing decision and thus, increases speed and enhances the reliability of the voltage sense amplifiers.

### 4.3.1 Dual-Input Classic Sense Amplifier

In Fig. 4.18 and Fig. 4.19, a classic sense amplifier (CSA) and a dual-input classic sense amplifier (DICSA) are shown. The timing diagram associated with the SAs is shown in Fig. 4.17. When the sense amplifier enable (SAE) is low, output nodes $V_1$ and $V_2$ are pre-charged to $V_{DD}$ in CSA while in DICSA the nodes are balanced with the source of transistors $P_3$ and $P_4$. At the time of reading from an SRAM bit-cell, i.e., after waiting for a stipulated amount of time to allow bit-line differential voltage development, one of BL and BLB will be at a lower potential than the other. The nodes ($V_1$ and $V_2$) are exposed to differential inputs through column multiplexer transistors $P_5$ and $P_6$ for the CSA and $P_5$, $P_6$, and $P_7$, $P_8$ for the DICSA through the control signal $Y_{MUX}$. The SAs are turned on by a rising transition of SAE. The potential at both nodes ($V_1$ and $V_2$) falls simultaneously towards the ground or $V_{SS}$. While the nodes $V_1$ and $V_2$ for CSA were pre-charged to $V_{DD}$, $N_3$ and $N_4$ both turn on, but due to the voltage difference between $V_1$ and $V_2$, the gate connected to the lower terminal voltage will have lower conductivity and finally, one transistor will enter cut-off while other remains on. In the case of DICSA, before the time SAE sees a rising transition, nodes $V_1$ and $V_2$ are not balanced, but have differential inputs being fed through transistors $P_5$, $P_6$ and through the transistor pairs $P_7$, $P_3$ and $P_8$, $P_4$, respectively, which in effect bias the transistors $N_3$, $N_4$ favorably to make a correct decision. Once SAE has a rising transition, based on the differential inputs, one of the NMOS transistors shuts off. The buffers at the outputs $V_1$, $V_2$, which are typical for a sense amplifier, will produce the full swing output.

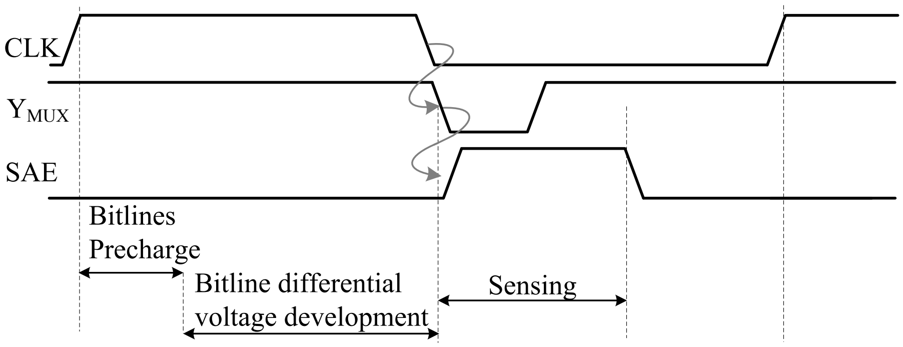


Figure 4.17: Timing for the sense amplifiers.

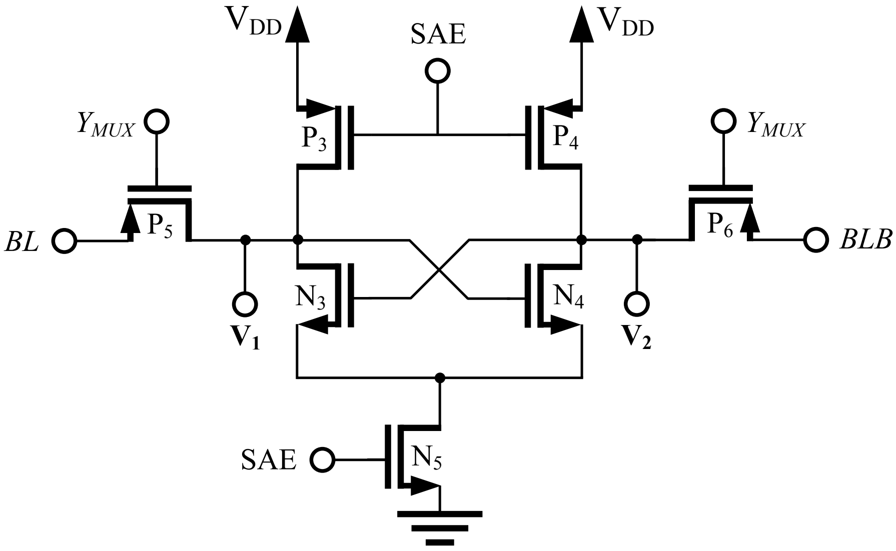


Figure 4.18: Classic sense amplifier.

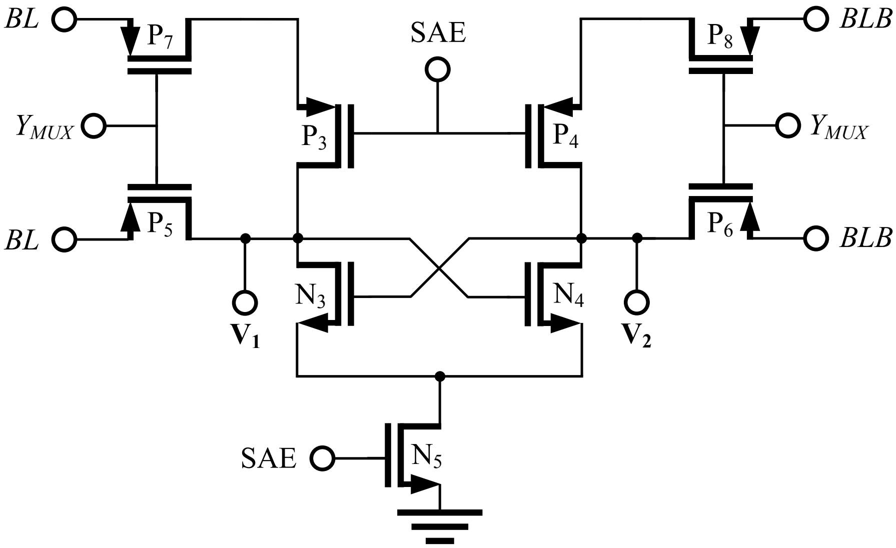


Figure 4.19: Dual-input classic sense amplifier.

### Simulation Results

In Fig. 4.20 the Monte Carlo results for 10,000 runs for CSA and DICSA are presented. The simulations were carried out at 1.0 V in 65 nm at 1 GHz frequency of operation. The nodes were loaded with the capacitance of a typical SRAM column. The results show that DICSA has 10 % higher correct results at the smallest differential input simulated in this experiment.

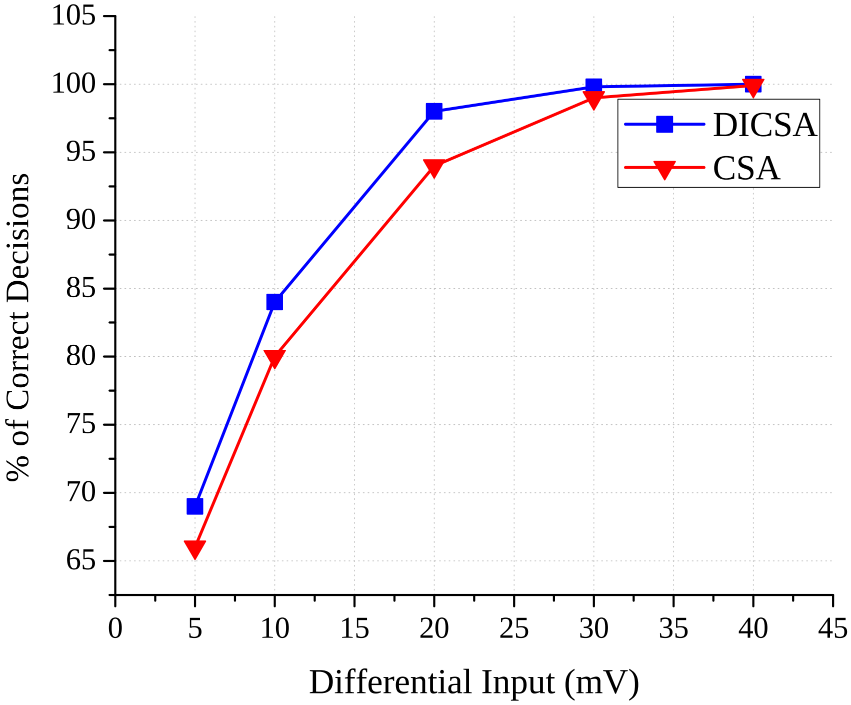


Figure 4.20: Monte Carlo results for the CSA and DICSA for increasing differential input signals.

### Measurement Results

A test chip was designed in a commercial 65-nm bulk CMOS process. The chip included the proposed and reference sense amplifiers and was wire bonded in a CQFP package and mounted on a PCB. A timing block triggered by an external clock signal generated the necessary control signals on-chip. The $BL$ and $BLB$ were directly controlled through input pins. The layout of the sense amplifier and the associated routing was done symmetrically.

Single-ended latched outputs were observed on an oscilloscope. The measurements were carried out at 100-MHz clock frequency. The choice of frequency was constrained by the equipment at the test facility. The differential input voltages are DC values set through variable potentiometers on the PCB. For each measurement the $BL$ and $BLB$ were set with an accuracy of 0.1 mV. The SAs were analyzed in two ways.

Initially, a potential $(V_{DD} - 100mV)$ was applied on the $BL$ while keeping the $BLB$

at $V_{DD}$, and then the SA output was analyzed. Then, keeping the $BLB$ at $V_{DD}$, the $BL$ potential was increased in steps of 1 $mV$. The output was recorded for different differential input voltages. Once $BL$ reached $V_{DD}$, it was kept there. Then, $BLB$ was decreased in steps of 1 mV until it reached $(V_{DD} - 100mV)$. This measurement was termed as *$BL$ increasing and $BLB$ decreasing*.

Subsequently, complementary steps was carried out. The BLB potential was increased from the initial voltage of $(V_{DD} - 100mV)$ in steps of 1 mV until it reached $V_{DD}$ while the BL voltage is kept at the nominal voltage of $V_{DD}$. Then, BLB was held at $V_{DD}$ and the BL potential was decreased and its outputs were recorded. The current procedure continued until BL reached was $(V_{DD} - 100mV)$. This measurement was termed as *$BLB$ increasing and $BL$ decreasing*. The BLs were swept from $V_{DD}$ to $V_{DD} - \Delta V$ and then back to $V_{DD}$ in order to eliminate any memory effect on the sensing nodes and to account for any offset. The procedure was repeated for different values of $V_{DD}$ potential. Multiple sets of measurements were carried out for different differential input voltages and different $V_{DD}$ potentials. In all measurements, for a given $V_{DD}$, a highest functional frequency range was selected.

For the DICSA implementation, the measurement results are shown in Fig. 4.21. At a clock frequency of 100 MHz and a $V_{DD}$ of 1.0 V, DICSA required 45% smaller differential input when compared with CSA in order to resolve correctly. While scaling the $V_{DD}$, the DICSA and CSA are functional at $V_{DD}$ as low as 0.4 V at a variety of operational frequencies. The DICSA required 6× smaller inputs than CSA at 10 MHz and a $V_{DD}$ of 0.5 V, 5.6× smaller input at 1 MHz and a $V_{DD}$ of 0.4 V to sense the inputs correctly. It is interesting to note that DICSA is completely functional at a $V_{DD}$ as small as 0.2 V with a clock frequency of 350 kHz.

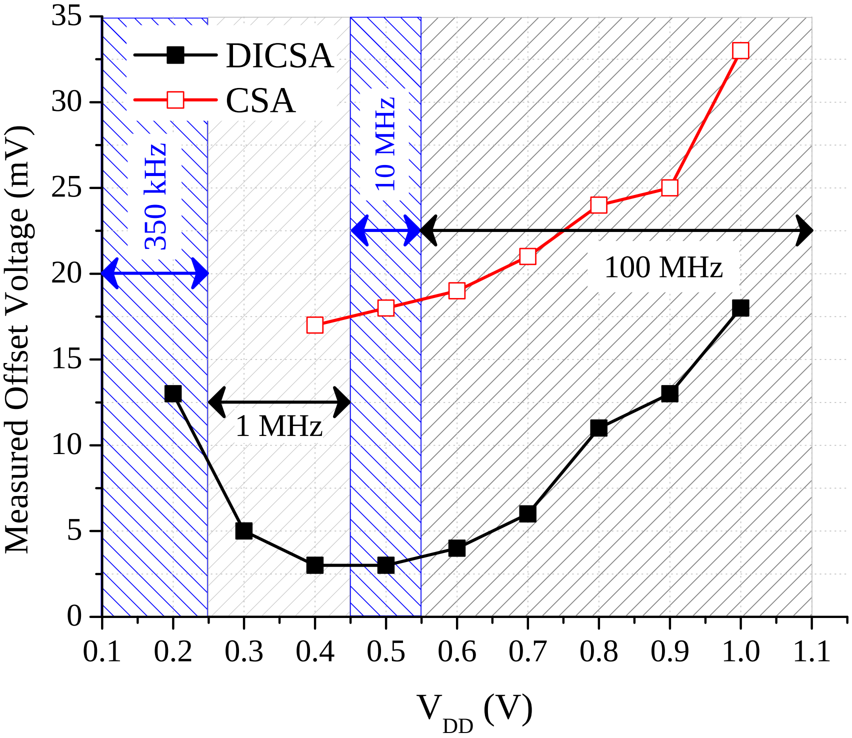


Figure 4.21: Measurement results for DICSA and CSA showing the smallest differential voltage required to make a correct decision at a given $V_{DD}$ for the highest frequency of operation.

### 4.3.2 Dual-Input Voltage Sense Amplifier

In Fig. 4.22 and Fig. 4.23, a conventional voltage latch type sense amplifier (LSA) and a dual-input voltage latch type sense amplifiers (DILSA) are shown, respectively. The timing diagram shown in Fig. 4.17 is also applicable to these SAs. When the sense amplifier enable is low, output nodes $V_1$ and $V_2$ are pre-charged to $V_{DD}$ in LSA and DILSA through transistors $P_1$ and $P_2$. At the time of reading from an SRAM bit-cell i.e., after waiting for a sufficient amount of time to develop enough differential input, one node will be at a lower potential than the other due to bit-line discharge. The nodes ($V_1$ and $V_2$) are exposed to differential inputs through column multiplexer transistors $P_5$ and $P_6$ for LSA and, additionally, have $P_7$, $P_8$ for DILSA through the control signal $Y_{MUX}$. The transistors $P_3$, $P_4$ are off at this time because nodes $V_1$ and $V_2$ are pre-charged to $V_{DD}$, however the difference between the source voltages is the same as the bit-line differential voltage. The SAs are turned on by a rising transition of the signal SAE. The potential at both nodes ($V_1$ and $V_2$) falls simultaneously towards ground or $V_{SS}$. While the nodes $V_1$ and $V_2$ for LSA were pre-charged to $V_{DD}$, $N_3$ and $N_4$ both turn on, but due to the voltage difference between $V_1$ and $V_2$, the gate connected to the lower terminal voltage will have a lower conductivity and, finally, one transistor will go to cut off mode while other remains on. In the case of DILSA, before the time SAE sees a rising transition, nodes $V_1$ and $V_2$ have differential inputs. Additionally, one among $P_3$ and $P_4$ has $V_{SG}$ greater than the other transistor which allows one of $V_1$ and $V_2$ to charge and discharge faster making the sensing environment mode conducive for a correct evaluation. In this case, $N_3$, $N_4$ are the decision making pair, however, $P_3$, $P_4$ aids in attaining a full swing. DILSA will respond better if there is a mismatch between $N_3$ and $N_4$. The buffers at the outputs $V_1$, $V_2$, which are typical for a sense amplifier, will give full swing output for further processing of the data.

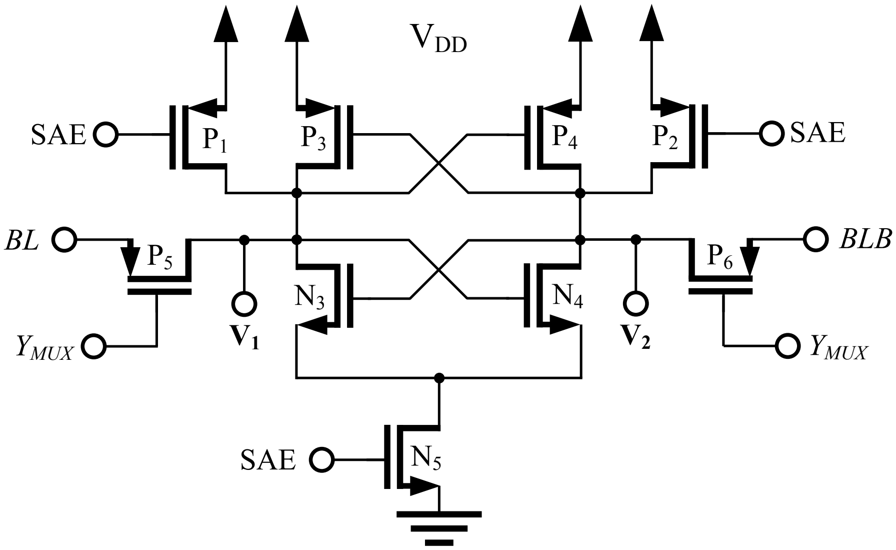


Figure 4.22: Conventional voltage latch type sense amplifier.

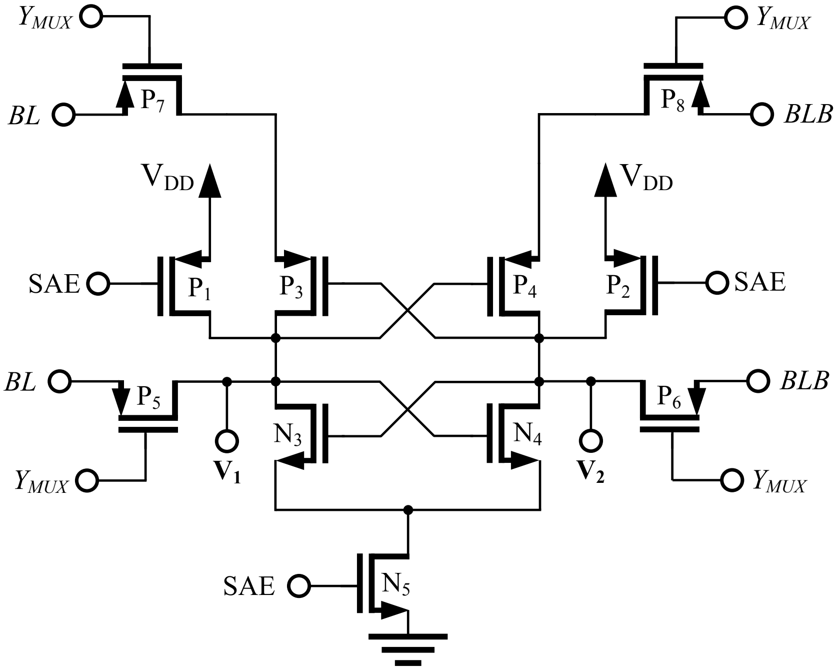


Figure 4.23: Dual-input voltage latch type sense amplifier.

### Simulation Results

In Fig. 4.24 the Monte Carlo results for 10,000 runs for LSA and DILSA are presented. The simulations were carried out at 1.0 V in 65-nm at 1 GHz frequency of operation. The nodes were loaded with the parasitic capacitance of a typical SRAM column. The results show that DILSA has a very small improvement over LSA in this implementation. These results are consistent with the findings of Fig. 4.2, Fig. 4.3 where we saw that the mismatch between PMOS transistors pair does not have big impact on the sensitivity of the sense amplifier.

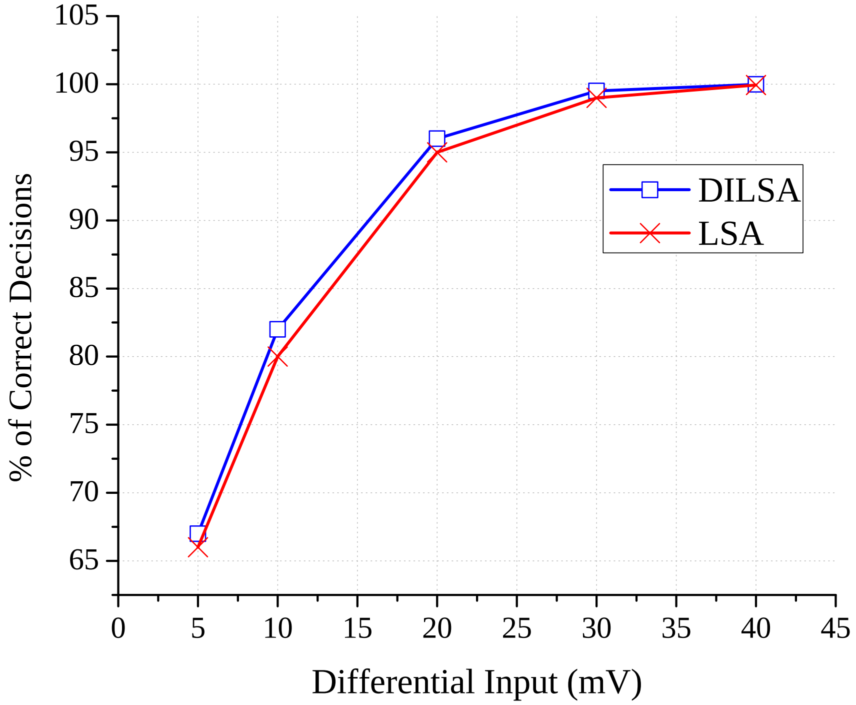


Figure 4.24: Monte Carlo results for the LSA and DILSA for increasing differential input signals.

### Measurement Results

For the DILSA and LSA implementations, the measurement results are shown in Fig. 4.25. The measurement procedure is identical to the one described in the Section 4.3.1. At a clock frequency of 100 MHz and a $V_{DD}$ of 1.0 V, DILSA required 9% smaller differential input when compared with LSA in order to resolve correctly. While scaling the $V_{DD}$, the DILSA and LSA are functional at a $V_{DD}$ of 0.4 V and at 10 MHz; however, only DILSA is fully functional at a reduced clock frequency of 1 MHz and 0.3 V while requiring a differential input of 65 $mV$. The LSA was not evaluating correctly at 0.3 V and hence is assumed non functional at this $V_{DD}$.

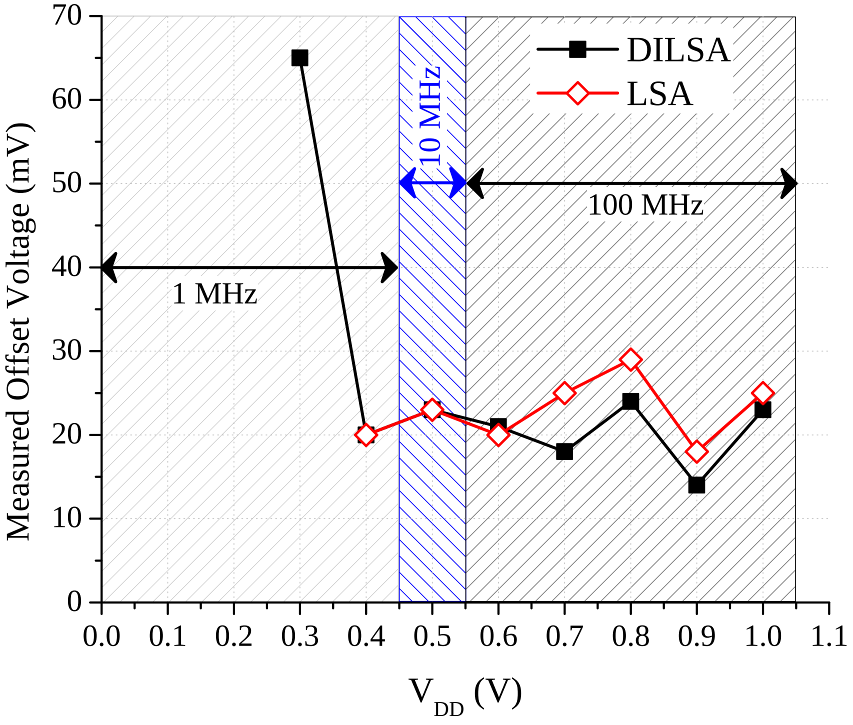


Figure 4.25: Measurement results for the DILSA and LSA showing the smallest differential voltage required to make a correct decision at a given $V_{DD}$ for a highest frequency of operation.

### 4.3.3 Dual-Input Current Latch Type Sense Amplifier

In Fig. 4.26 and Fig. 4.27, a conventional current latch type sense amplifier and dual-input current latch type sense amplifier (DICLSA) are shown. The timing diagram shown in Fig. 4.17 is also applicable to these SAs. When the sense amplifier enable is low, output nodes $V_1$ and $V_2$ are pre-charged to $V_{DD}$ in CLSA and DICLSA through transistors $P_1$ and $P_2$. At the time of reading from an SRAM bit-cell i.e., after waiting for a stipulated amount of time to provide enough differential input, one node will be at a lower potential than the other due to the bit-line discharge. The nodes ($V_1$ and $V_2$) are exposed to differential inputs through column multiplexer transistors $P_7$ and $P_8$ for DICLSA through the control signal $Y_{MUX}$ while the gates of input transistors $N_1$ and $N_2$ see differential inputs through $P_5$ and $P_6$ for both CLSA and DICLSA. These sense amplifiers combine positive feedback with a high resistive input. Having nodes $V_1$ and $V_2$ of DICLSA exposed to differential inputs before enabling SAE biases $N_3$ and $N_4$ in an environment favorable for a correct evaluation. The SAs are turned on by rising transition of the signal SAE. The potential at both nodes ($V_1$ and $V_2$) falls simultaneously towards ground or $V_{SS}$. The current flow through differential input transistors $N_1$ and $N_2$ enables the latch circuit. The drain currents of $N_1$ and $N_2$ discharge the outputs $V_1$ and $V_2$, respectively. With a differential voltage at the gates of $N_1$ and $N_2$, their drain currents are different and these currents control the speed at which $V_1$ and $V_2$ discharge. It is interesting to note that $V_1$ and $V_2$ were pre-charged to $V_{DD}$ for CLSA and to a differential input for DICLSA before enabling SAE, and the PMOS transistors $P_3$ and $P_4$ remained off until one of the nodes $V_1$ or $V_2$ discharges below $(V_{DD} - V_{thP})$. The discharge happens faster for DICLSA because of the initial conditions. At this time, the positive feedback takes over bringing one of the nodes among $V_1$ and $V_2$ to $V_{DD}$ and the evaluation of the SAs is completed when one of transistors among $N_1$ and $N_2$ turns off. The buffers at the outputs $V_1$, $V_2$, which are typical for a sense amplifier, will give full swing output for further processing of the data.

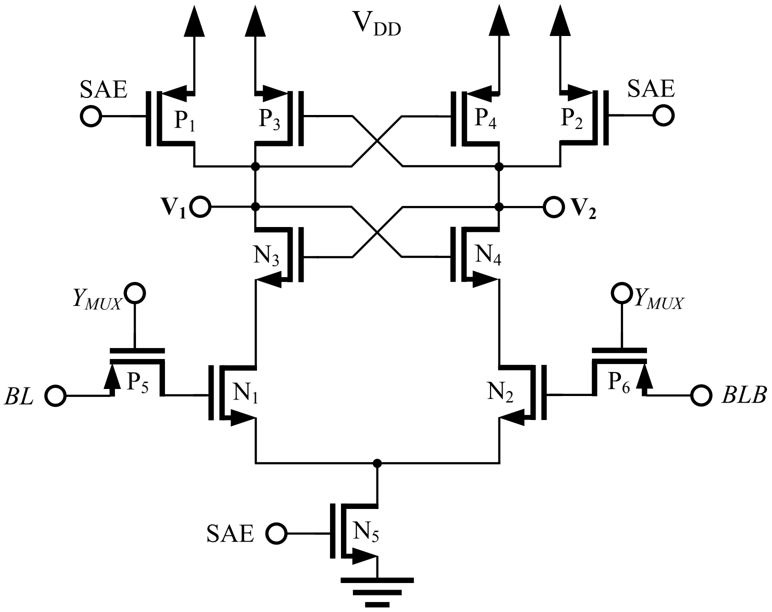


Figure 4.26: Current latch sense amplifier.

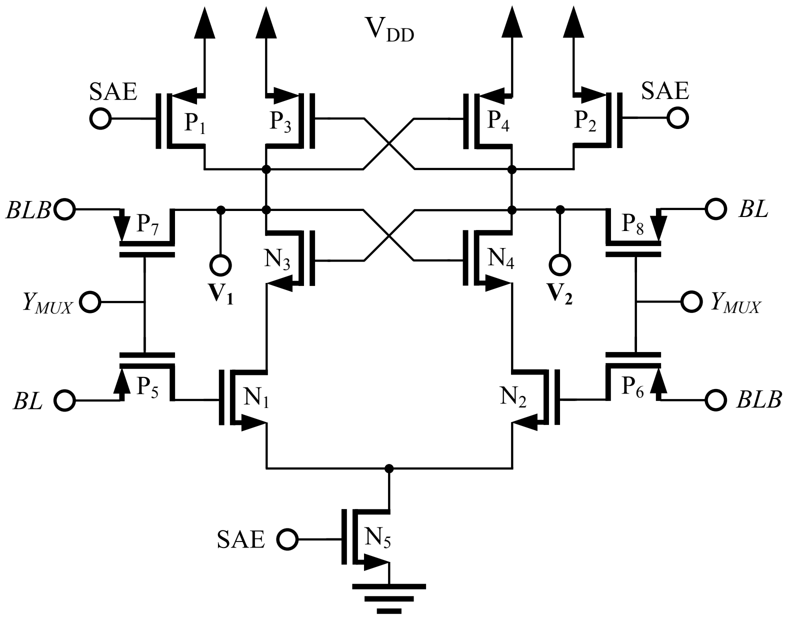


Figure 4.27: Dual-input current latch sense amplifier.

### Simulation Results

In Fig. 4.28 the Monte Carlo results for 10,000 runs for CLSA and DICLSA are presented. The simulations were carried out at 1.0 V in 65-nm at 1 GHz frequency of operation. The nodes were loaded with the parasitic capacitance of a typical SRAM column. The results show that DICLSA has 40% higher correct results at the smallest differential input simulated in this experiment.

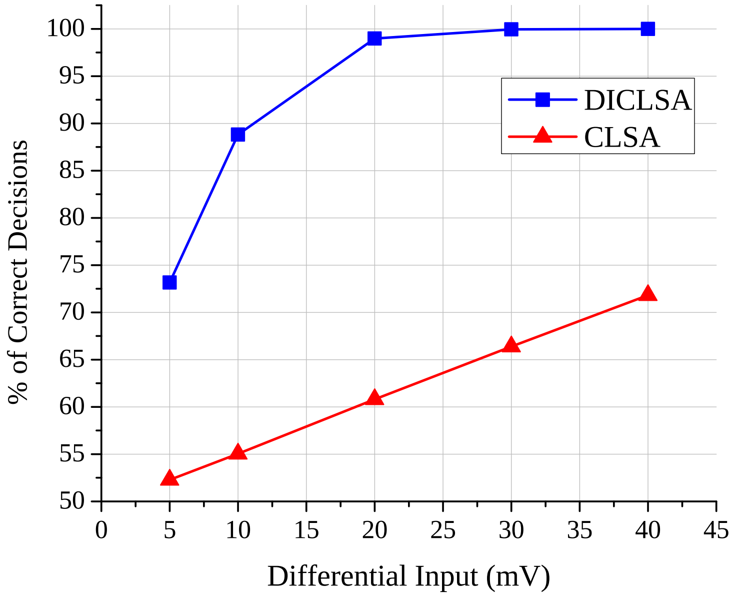


Figure 4.28: Monte Carlo results for the CLSA and DICLSA for increasing differential input signals.

### Measurement Results

For the DICLSA implementation, the measurement results are shown in Fig. 4.29. The measurement procedure was identical to the one described in the Section 4.3.1. At a clock frequency of 100 MHz and a $V_{DD}$ of 1.0 V, it required 2.5× smaller differential input when compared with the CLSA in order to deliver a correct sensing decision. While scaling the $V_{DD}$, the DICLSA and CLSA are functional for $V_{DD}$ as low as 0.2 V at a varied range of operational frequency. The DICLSA required 1.47× smaller input than CLSA at 10 MHz and a $V_{DD}$ of 0.5 V, 1.54× smaller differential input at 1 MHz and $V_{DD}$ of 0.3 V, and it takes 40% smaller differential input than the CLSA at 300 kHz.

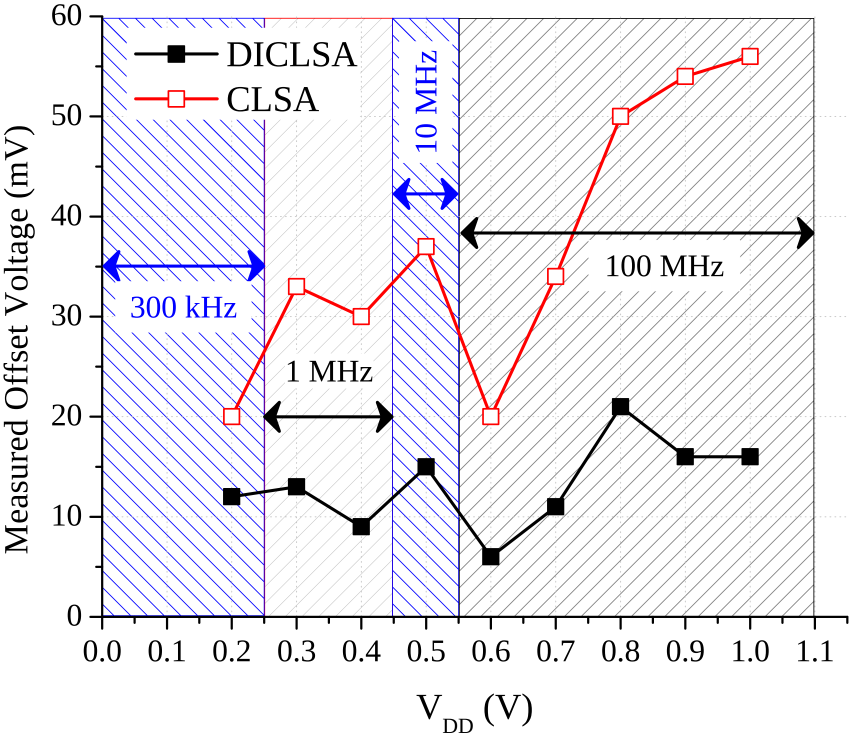


Figure 4.29: Measurement results for the DICLSA and CLSA showing the smallest differential voltage required to make a correct decision at a given $V_{DD}$ for the highest frequency of operation.

## 4.4 Comparison

In this section, the results of the proposed and reference sense amplifiers are compared.

### Simulations Results

The proposed and reference SAs are compared in Fig. 4.30. The comparison is based upon the percentage of correct decisions for a given value of differential input. All the bold lines represent proposed schemes and the dotted lines with hollow symbols represent the reference SAs. The DICLSA has the highest probability to make a correct decision for small values of differential inputs. Typically, it showed a 40% better chance of making a correct decision over CLSA, and a 9% improvement over LSA. Additionally, it has 6% higher probability to resolve correctly when compared with DICSA and 9% better than DILSA. In fact, DILSA which has lowest probability to evaluate correctly among proposed schemes, shows slightly better probability than the best reference SA.

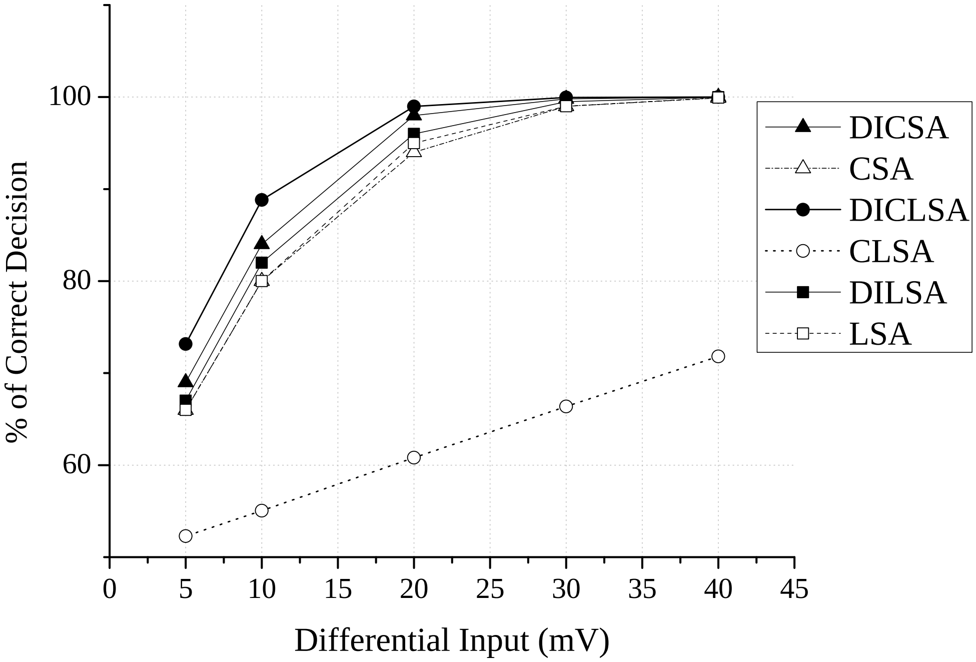


Figure 4.30: Comparison between the proposed and conventional SAs to make a correct decision based upon 10,000 Monte Carlo simulations for increasing differential inputs.

### Measurement Results

The proposed and reference SAs are compared in Fig. 4.31 based upon measurement results. The comparison is based upon the offset voltage for each sense amplifier; in other words, for the smallest value of differential inputs required to resolve correctly, for a range of $V_{DD}$ values. Once again, all the bold lines represent proposed SAs and the dotted lines with hollow symbols represent the reference SAs. All of the proposed schemes performed better than the reference SAs with the exception that DILSA has a smaller offset only over the range when $V_{DD}$ is reduced from 1.0 V to 0.7 V. At 0.6 V, DICSA and DICLSA show 3.3× improvement over the reference SAs including DILSA. The DICSA has the smallest offset among all the SAs and at ultra-low $V_{DD}$ of 0.2 V, its offset is comparable to DICLSA.

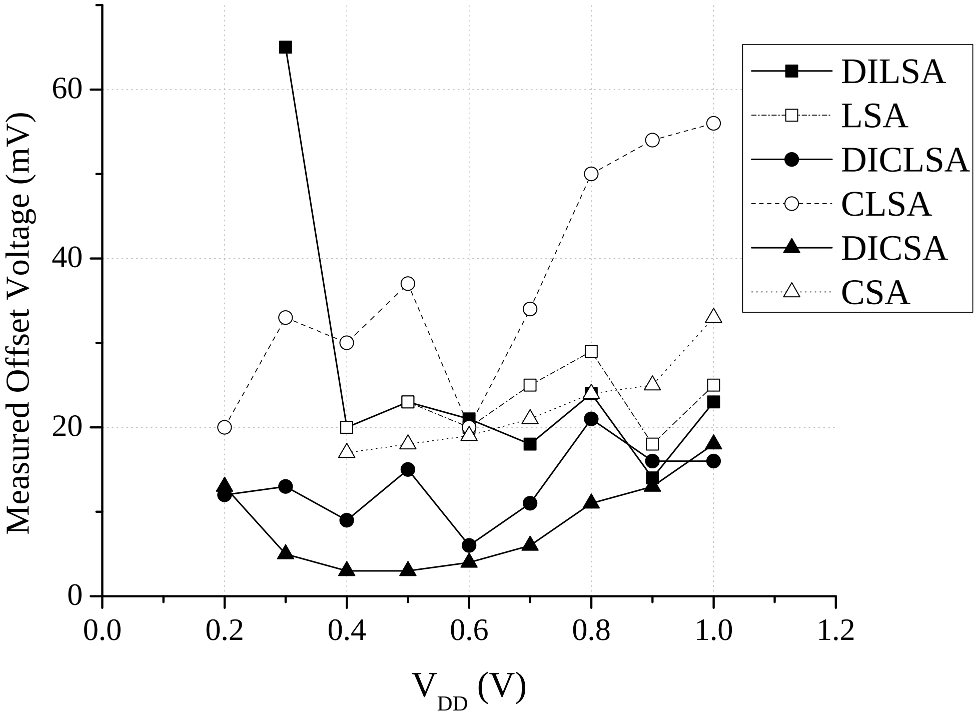


Figure 4.31: Comparison between the proposed and conventional schemes based upon measurement results showing the smallest differential voltage required to make a correct decision at a given $V_{DD}$ for the highest frequency of operation.

The proposed dual-input sense amplifiers in general have smaller offsets compared to the reference SAs and are well suited for low voltage applications.

## 4.5 Summary

In this chapter, the source of offset in SRAM sense amplifiers were discussed and a solution was proposed. It was shown that the threshold voltage mismatch between the input transistor can result in an incorrect evaluation by the sense amplifier. Theoretical analysis showed that the proposed offset cancellation scheme in the sense amplifier is effective over a range of design parameters. The simulation results of the proposed scheme were verified with measurements. Another sense amplifier architecture is proposed in this work which requires smaller differential inputs over a range of supply voltage and frequencies of operation. Monte Carlo results demonstrated that the probability of a correct decision is higher for different implementations of the proposed idea. Finally, measurement results of a test chip in 65 nm showed that the sense amplifier indeed require smaller differential inputs in order to resolve correctly.

# Chapter 5

# Conclusions

Embedded SRAM constitutes more than 50% of the die area for state-of-the-art microprocessors and SoCs and is expected to increase in the future. To achieve higher reliability, robust SRAM design is necessary. This work analyzes SRAMs with two objectives: (1) to make them soft-error-robust with minimum area and power cost, and (2) at the architectural level in the periphery to facilitate a reliable operation under optimal energy conditions. Soft error robustness can be achieved through process, circuit or architectural techniques. Process techniques being have cost overhead and architectural techniques have timing overhead. On the other hand, circuit level techniques do not have these constraints, which allows effective scaling of the idea in advanced technologies.

## 5.1 Summary of Contributions

This work is expected to make the following contributions: The first contribution is the low-voltage soft-error-robust SRAM. Details of the contributions are summarized below:

**Low-Voltage Soft-Error-Robust SRAM**

- A cost-effective access-transistor-less architecture.
- Proposal of an area-efficient soft-error-robust 8T bit-cell.

- Analysis of 8T bit-cell operating margins, read current, leakage current, low-voltage operation, and soft error robustness.
- Development and testing of an 8T test chip in 65-nm GP CMOS incorporating the proposed bit-cell array and demonstrating its operation down to 0.55 V.
- Radiation test of the test chip according to the JEDEC standard including a FIT calculation procedure.

### An Offset-Cancelling Sense Amplifier

The proposed 8T cell has shown promising results in terms soft error robustness. As opposed to a 6T bit-cell, where during a read operation the bit-cell is read by sensing the difference between $V_{DD}$ and $(V_{DD} - \Delta V)$, in the 8T bit-cell, the read operation is carried out by reading the difference between $V_{SS}$ and $(V_{SS}+\Delta V)$. Thus, there is a need for a robust read operation taking into account process variations such as $V_{th}$ offset. This leads to the second contribution, which is the analysis and development of an offset cancelling sense amplifier. Details of this contribution are summarized below:

- Proposal of an offset cancelling sense-amplifier (SAOC) circuit which can sense small differential voltage in spite of $V_{th}$ mismatch between the input transistors. The SAOC is also compatible with the 8T bit-cell.
- The SAOC is analyzed for mismatch between input transistors, in terms of design space through small-signal analysis. The effectiveness of the offset cancellation technique is demonstrated over a wide range of mismatch values.
- The SAOC was prototyped in 180-nm CMOS technology where it was compared with a conventional sense-amplifier. The offset cancellation capability of SAOC was shown to resolve 77% smaller differential signal in the worst case and it is achieved in an energy efficient manner (that is, a 75% reduction from a comparable number in the literature).

### Dual-Input Sense Amplifier Architecture

The read operation is always critical in SRAM figures of merit. In an effort to increase speed and save power, the cell access time is reduced which results in smaller bit-line differential voltage development. At the same time, the reduced operating voltage will also result in smaller sensing margins during a read operation. As a natural consequence, a sense amplifier circuit is required which would provide robust sensing at reduced differential inputs and at reduced operating voltages. This leads to the third contribution which is the development of the dual-input sense amplifier architecture. Details of this contribution are summarized below:

- Proposal of dual-input sense amplifier (DISA) circuits, which can sense small differential voltage to resolve correctly.
- DISA circuits namely: the dual-input classic sense amplifier, dual-input voltage sense amplifier, and dual-input current latch type sense amplifier are analyzed and compared with the classic sense amplifier, voltage sense amplifier, and current latch type sense amplifier, respectively. The comparison is based upon Monte Carlo simulations where DISA circuits performed favourably where they showed higher probability of a correct decision even at smaller differential inputs.
- DISA circuits were prototyped in 65-nm GP CMOS technology where they were compared with the conventional counterparts. DILSA requires 9% smaller differential input as compared to LSA and is functional at a power supply of 0.3 V. DICSA required 45% smaller differential input and is completely functional at a $V_{DD}$ of 0.2 V where it resolved a differential input of $13\,mV$. The DICLSA works at 2.5× smaller differential input as compared to CLSA at 1.0 V and 40% smaller input at a power supply of 0.2 V.

## 5.2 Future Work

This work analyzed the key trade-offs associated with soft error robustness and how they relate to area, performance and functionality. There is a tight link between performance and functionality. Soft errors continues to increase with technology scaling and thus, the research in this area can have significant impact in mission critical applications. The low voltage 8T bit-cell has demonstrated soft error robustness at and near $V_{DD}$. With reduced voltage, the soft error rate increases, therefore, evaluation of 8T in ultra-low-power domain and the associated trade-offs are worth investigating. The 8T cell can have applications in robust flip-flops implementations, thus the research can provide tolerant storage units at the system level.

The current technology trends show that process variations will further increase with scaling and more research is required in the area. The proposed offset cancellation techniques work for mismatch between the input transistors, which is shown to be a dominant source of offset; however, the sensing transistors also contribute to offset in the sense amplifiers. Thus, a technique which can address mismatch between the input and the sensing pairs of the sense amplifier is desirable.

The proposed architecture of sense amplifiers (DISA) was prototyped as stand-alone amplifiers loaded with bit-line capacitance at the input nodes. A more realistic design should include a bit-cell array so that the dynamic behaviour of these circuits can be evaluated in a more realistic operating environment.

# Publications from this research

- J.S. Shah, D. Nairn, M. Sachdev, A Soft Error Robust 32kb SRAM Macro Featuring Access Transistor-Less 8T Cell in 65-nm, IEEE International Conf. on VLSI and System-on-Chip (VLSI-SoC), 2012
- J.S. Shah, D. Nairn and M. Sachdev, " An SRAM Sense Amplifier with Offset Cancellation," Circuits and Systems-II: Express Briefs, IEEE Transactions on (under review)
- S.M. Jahinuzzaman, J.S. Shah, D. J. Rennie and M. Sachdev, "Design and analysis of a 5.3-pJ 64-kb gated ground SRAM with multiword ECC," Solid-State Circuits, IEEE Journal of, Vol. 44 (9): pp. 2543-2553, September 2009.
- J.S. Shah, S. M. Jahinuzzaman, D. Li, P. Chuang, and M. Sachdev, A 64-bit, 2.4 GHz adder with SE detection capabilities employing time redundancy, in Proc. Microsystems and Nanoelectronics Research Conf., Ottawa, ON, Oct. 2009, pp. 37-40.

# Appendix A

# Details of Test Chips

## Test chip-1

**Technology** : 65-nm CMOS

**Design Name** : ICNWTTS1

**Idea Implemented** : 32 kb array of the proposed 8T bit-cell

**Functionality** : It was a multi-project chip, unfortunately, there was a short circuit between I/O power supply and ground rails which prevented the communication with the memory array. The current consumption was in a few tens of milli-amperes. Laser correction was attempted to isolate and repair the fault, but this attempt was not successful.

## Test chip-2

**Technology** : 180-nm CMOS

**Design Name** : ICFWTAN2

**Idea Implemented** : Offset cancelling sense amplifier

**Functionality** : It was a multi-project chip and was fully successful. The results are included in the thesis

The test board used in the measurements of the test chip is shown in Fig. A.1. In Fig. A.2a, the screenshot of the oscilloscope shows the single ended buffered output of the SAOC and the CONV. The SAOC (signal C3, which is blue in colour) makes a correct read 1 decision while the CONV (signal C4 which is green in colour) evaluates incorrectly. In Fig. A.2b, the simulation results show the buffered output of the SAOC with respect to clock out of the prototype chip. In simulations, the SAOC output is shown for a read 1 operation. Thus, the shape of the output waveform is explained.

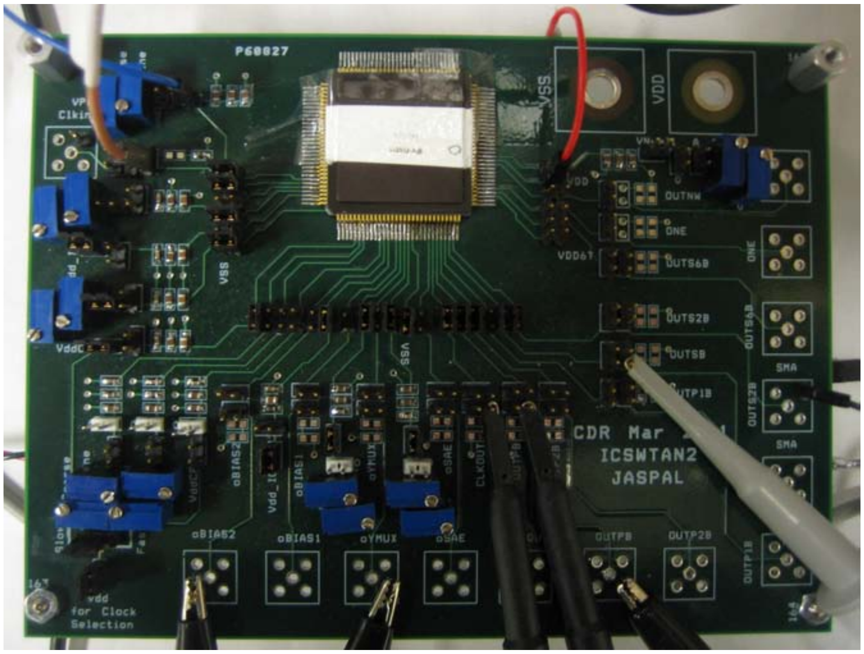
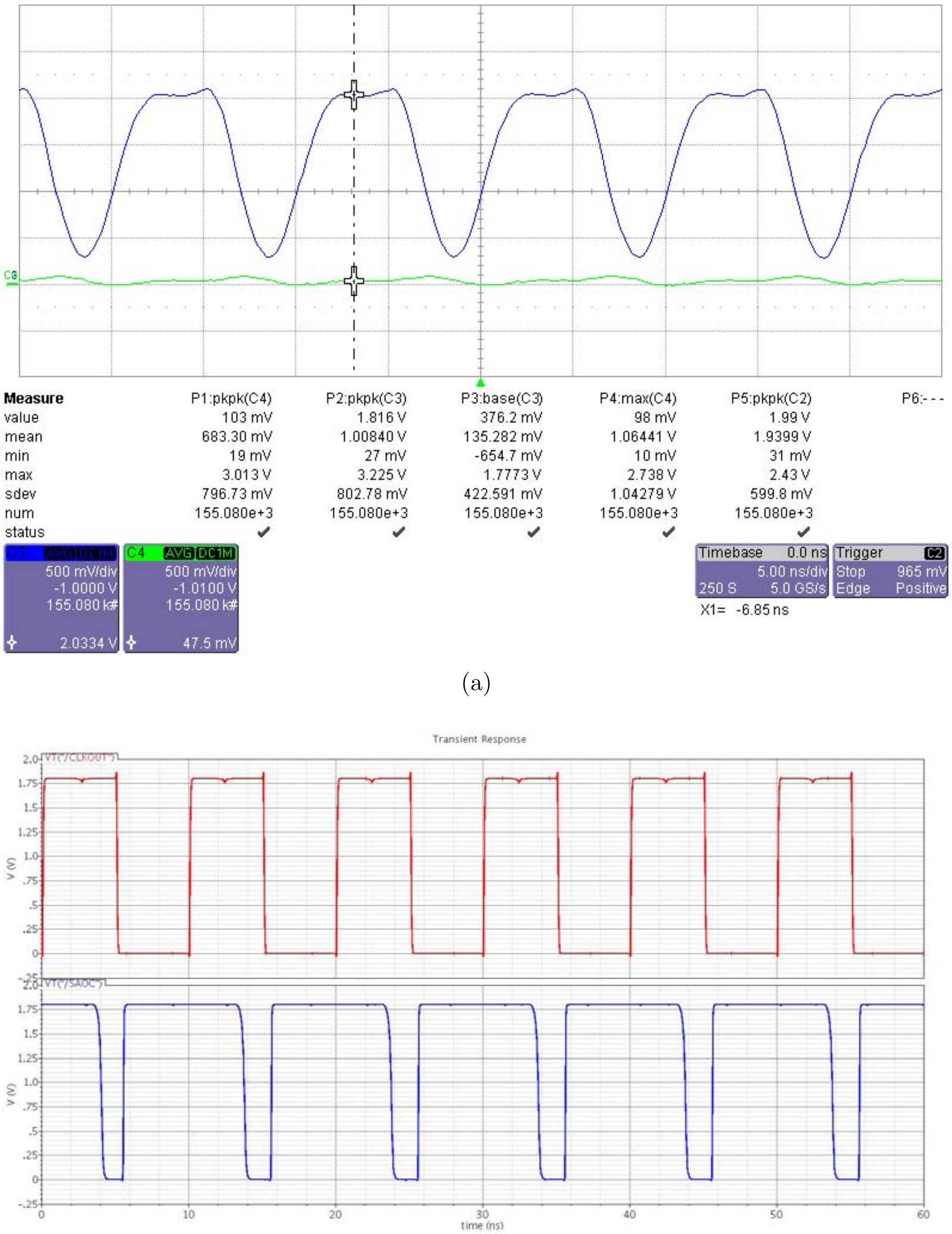


Figure A.1: Test board used in the measurements of Test Chip-2 at the CDR lab.



| Measure | P1:pkpk(C4) | P2:pkpk(C3) | P3:base(C3) | P4:max(C4) | P5:pkpk(C2) | P6:- - - |
|---|---|---|---|---|---|---|
| value | 103 mV | 1.816 V | 376.2 mV | 98 mV | 1.99 V | |
| mean | 683.30 mV | 1.00840 V | 135.282 mV | 1.06441 V | 1.9399 V | |
| min | 19 mV | 27 mV | -654.7 mV | 10 mV | 31 mV | |
| max | 3.013 V | 3.225 V | 1.7773 V | 2.738 V | 2.43 V | |
| sdev | 796.73 mV | 802.78 mV | 422.591 mV | 1.04279 V | 599.8 mV | |
| num | 155.080e+3 | 155.080e+3 | 155.080e+3 | 155.080e+3 | 155.080e+3 | |
| status | ✓ | ✓ | ✓ | ✓ | ✓ | |

```
C3  AVG DC1M        C4  AVG DC1M        Timebase     0.0 ns   Trigger        C2
    500 mV/div          500 mV/div                5.00 ns/div Stop       965 mV
    -1.0000 V           -1.0100 V       250 S      5.0 GS/s   Edge     Positive
    155.080 k#          155.080 k#      X1=  -6.85 ns
✧   2.0334 V        ✧   47.5 mV
```

(a)

(b)

Figure A.2: a) Measurement results of Test Chip-2 and b) Simulation results corroborating the measurements of Test Chip-2.

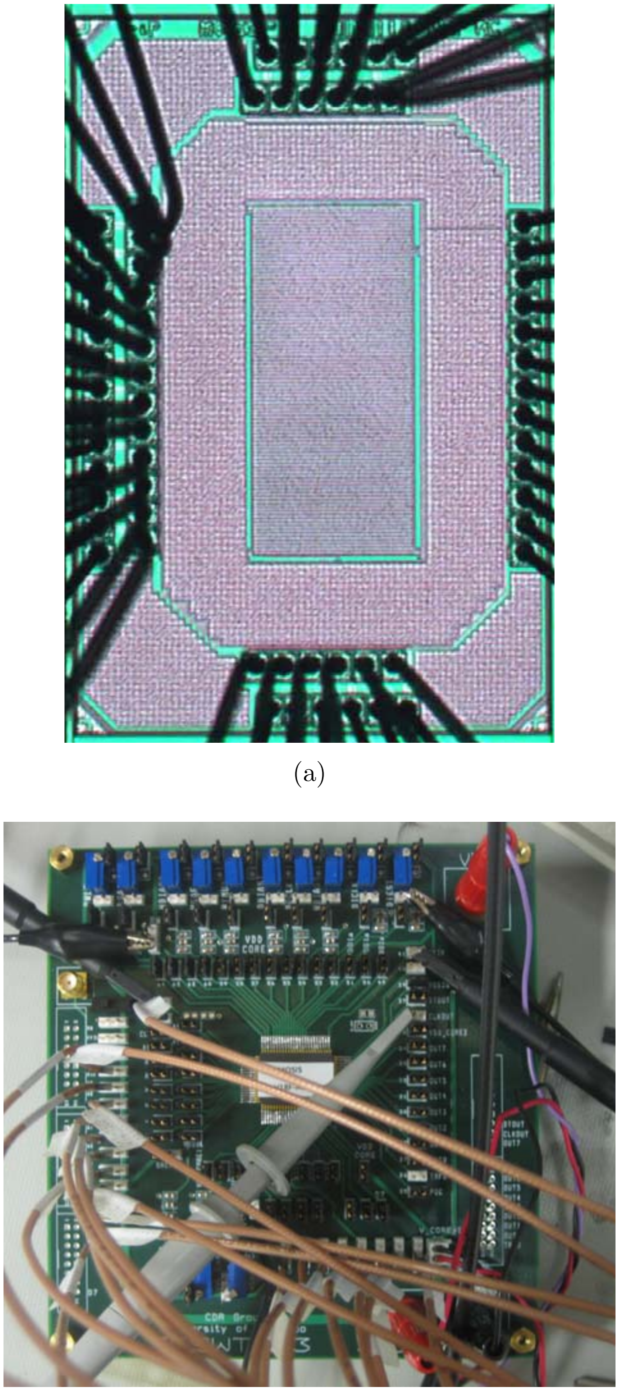
## Test chip-3

**Technology** : 65-nm CMOS

**Design Name** : ICSWTJS3

**Idea Implemented** : 32-kb array of the proposed 8T bit-cell

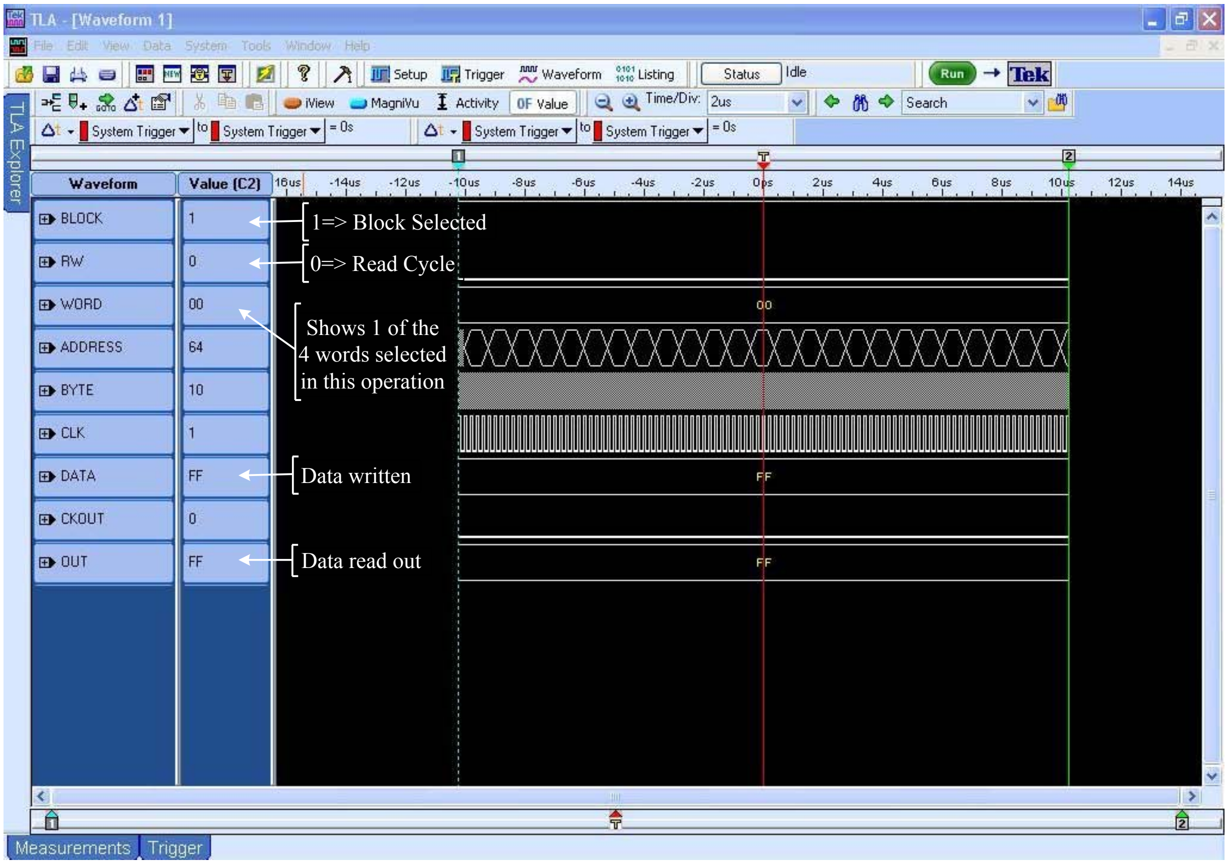
**Functionality** : It was a multi-project chip. In the first submission of this chip, the top metal layer of the pads disappeared mysteriously even though it was just an instantiation of the standard cells provided by the foundry. All efforts to recreate this problem in order to analyze what might have happen did not provide a clue. Later, the pad library was updated by the foundry and the test chip was resubmitted. It was successful and the results are included in the thesis



(a)

(b)

Figure A.3: a) Micrograph of the Test Chip-3 which implements a 32-kb array of 8T bit-cells, and b) Test board used to evaluate the prototype chip at the CDR lab.


TLA - [Waveform 1]
File Edit View Data System Tools Window Help
Setup Trigger Waveform Listing Status Idle Run Tek
iView MagniVu Activity 0F Value Time/Div: 2us Search
System Trigger to System Trigger = 0s
System Trigger to System Trigger = 0s
TLA Explorer
-16us -14us -12us -10us -8us -6us -4us -2us 0ps 2us 4us 6us 8us 10us 12us 14us
00
FF
FF
Measurements Trigger

| Waveform | Value (C2) |
|---|---|
| BLOCK | 1 |
| RW | 0 |
| WORD | 00 |
| ADDRESS | 64 |
| BYTE | 10 |
| CLK | 1 |
| DATA | FF |
| CKOUT | 0 |
| OUT | FF |

1=> Block Selected

0=> Read Cycle

Shows 1 of the 4 words selected in this operation

Data written

Data read out

Figure A.4: A screenshot of the logic analyzer waveforms the 8T array.

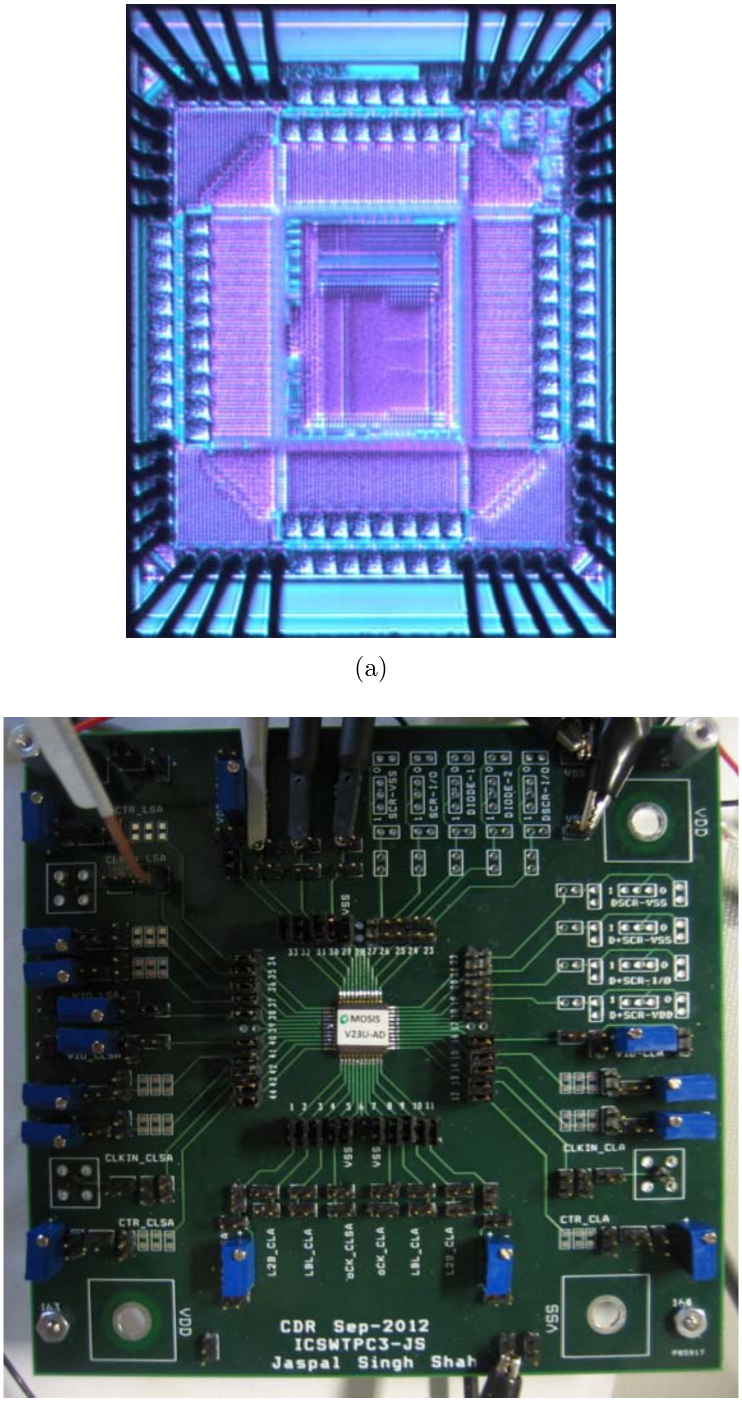
# Test chip-4

**Technology** : 65-nm CMOS

**Design Name** : ICSWTPC3

**Idea Implemented** : Dual-input sense amplifiers

**Functionality** : It was a multi-project chip and was fully successful. The results are included in the thesis

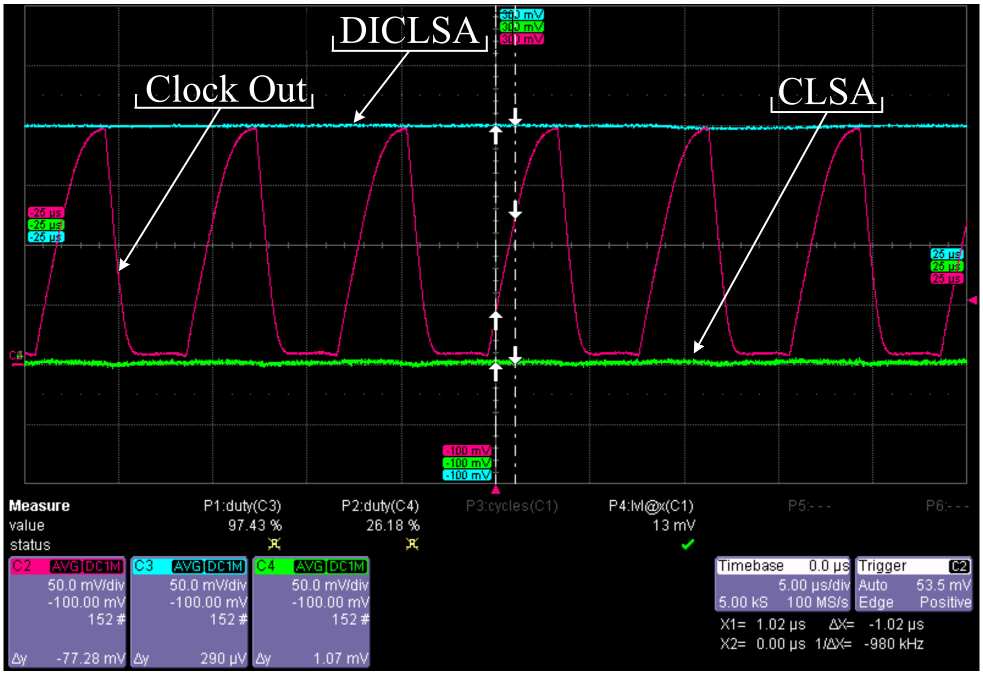


(a)

(b)

Figure A.5: a) Micrograph of the Test Chip-4 implementing Dual-Input Sense Amplifiers and b) Test board used for the measurements of the prototype chip.

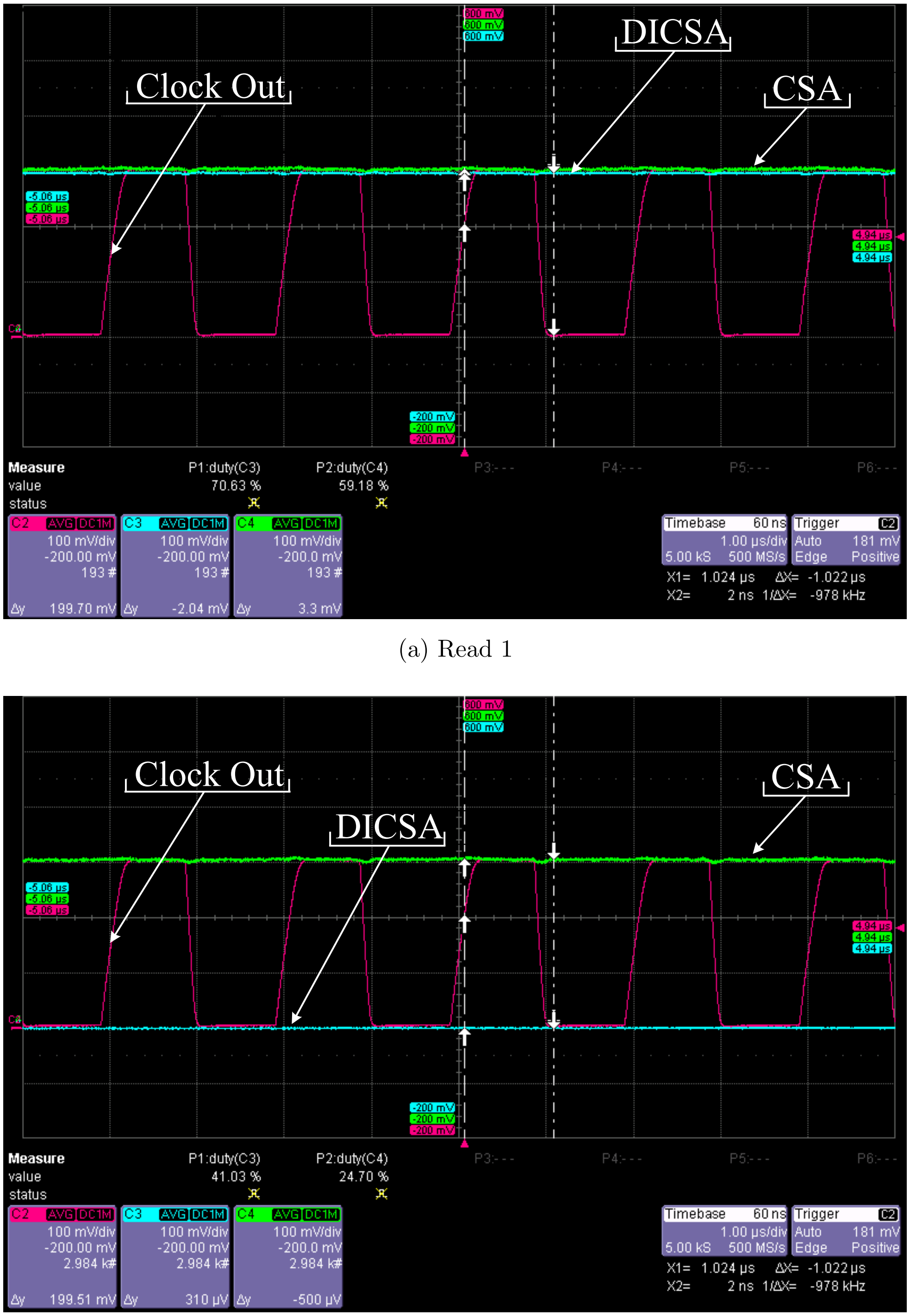
Figure A.6 shows the screenshot of the oscilloscope while measuring CLSA (signal C4 which is green in colour) and DICLSA (signal C3 which is cyan in colour) at $V_{DD}$ of 0.2 V during a read 1 operation. Also shown in the clock out (signal C2 which is magenta in colour) of the chip for reference. The DICLSA is shown to read a 1 correctly while the CLSA is still evaluating to a 0. The CLSA eventually resolved correctly when the level of differential input was increased.



```
Measure     P1:duty(C3)   P2:duty(C4)   P3:cycles(C1)   P4:lvl@x(C1)   P5:- - -   P6:- - -
value           97.43 %       26.18 %                          13 mV
status
C2  AVG DC1M    C3  AVG DC1M    C4  AVG DC1M        Timebase      0.0 µs   Trigger      C2
   50.0 mV/div     50.0 mV/div     50.0 mV/div               5.00 µs/div   Auto    53.5 mV
  -100.00 mV      -100.00 mV      -100.00 mV        5.00 kS    100 MS/s    Edge   Positive
       152 #           152 #           152 #        X1=  1.02 µs    ΔX=  -1.02 µs
Δy  -77.28 mV   Δy    290 µV    Δy   1.07 mV        X2=  0.00 µs  1/ΔX=  -980 kHz
```

Figure A.6: A screenshot of the oscilloscope waveforms for the CLSA and DICLSA at 0.2 V.

Figure A.7 shows the screenshot of the oscilloscope while measuring CSA (signal C4, which is green in colour) and DICSA (signal C3, which is cyan in colour) at $V_{DD}$ of 0.3 V during a read 1 operation. Also shown in the clock out (signal C2, which is magenta in colour) of the chip for reference. In Fig. A.7a, both SAs are at logic 1, however, CSA is stuck at 1 as can be seen in the Fig. A.7b. In fact, DICSA correctly senses a 0 while CSA did not resolve a 0 even for large values of the differential input.



```
Measure      P1:duty(C3)   P2:duty(C4)   P3:- - -   P4:- - -   P5:- - -   P6:- - -
value           70.63 %       59.18 %
status

C2 AVG DC1M        C3 AVG DC1M        C4 AVG DC1M          Timebase     60 ns  Trigger       C2
     100 mV/div         100 mV/div         100 mV/div               1.00 µs/div  Auto      181 mV
   -200.00 mV         -200.00 mV          -200.0 mV        5.00 kS    500 MS/s  Edge    Positive
        193 #              193 #              193 #         X1=  1.024 µs    ΔX=  -1.022 µs
Δy   199.70 mV     Δy   -2.04 mV     Δy     3.3 mV          X2=      2 ns  1/ΔX=   -978 kHz
```

(a) Read 1

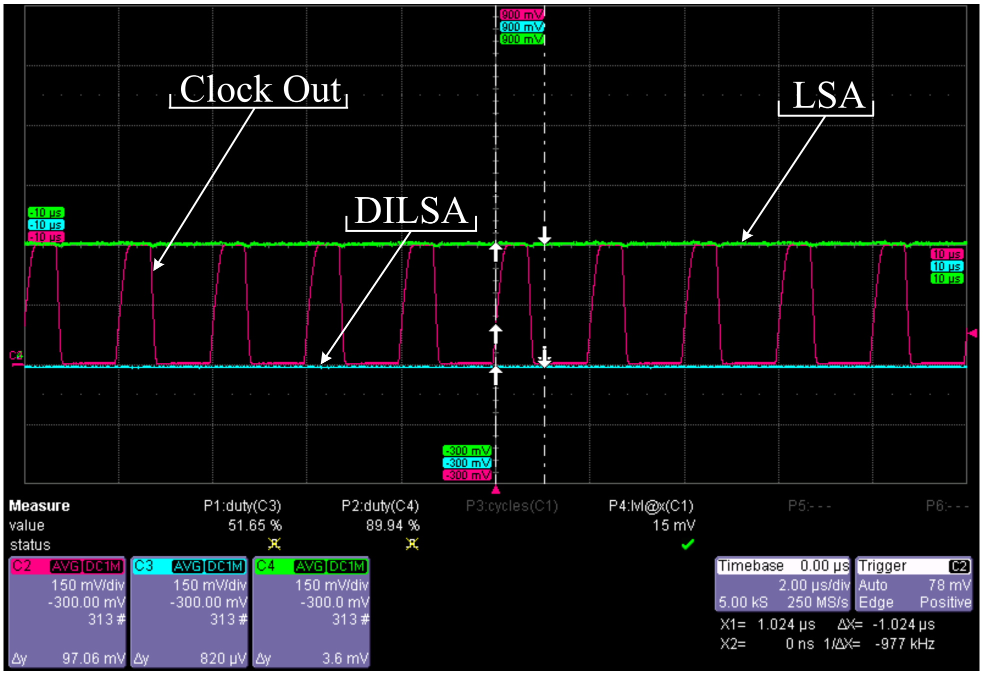
```
Measure      P1:duty(C3)   P2:duty(C4)   P3:- - -   P4:- - -   P5:- - -   P6:- - -
value           41.03 %       24.70 %
status

C2 AVG DC1M        C3 AVG DC1M        C4 AVG DC1M          Timebase     60 ns  Trigger       C2
     100 mV/div         100 mV/div         100 mV/div               1.00 µs/div  Auto      181 mV
   -200.00 mV         -200.00 mV          -200.0 mV        5.00 kS    500 MS/s  Edge    Positive
      2.984 k#           2.984 k#           2.984 k#        X1=  1.024 µs    ΔX=  -1.022 µs
Δy   199.51 mV     Δy     310 µV     Δy    -500 µV          X2=      2 ns  1/ΔX=   -978 kHz
```

(b) Read 0

Figure A.7: A screenshot of the oscilloscope waveforms for the CSA and DICSA.

Figure A.8 shows the screenshot of the oscilloscope while measuring LSA (signal C4, which is green in colour) and DILSA (signal C3, which is cyan in colour) at $V_{DD}$ of 0.3 V during a read 0 operation. Also shown is the clock out (signal C2, which is magenta in colour) of the chip for reference. The DILSA is shown to read a 0 correctly while the LSA was evaluating to a 1. The LSA did not resolve correctly, even when the level of differential input was increased.



| Measure | P1:duty(C3) | P2:duty(C4) | P3:cycles(C1) | P4:lvl@x(C1) | P5:- - - | P6:- - - |
|---|---|---|---|---|---|---|
| value | 51.65 % | 89.94 % | | 15 mV | | |
| status | ☒ | ☒ | | ✓ | | |

| C2 AVG DC1M | C3 AVG DC1M | C4 AVG DC1M |
|---|---|---|
| 150 mV/div | 150 mV/div | 150 mV/div |
| -300.00 mV | -300.00 mV | -300.0 mV |
| 313 # | 313 # | 313 # |
| Δy 97.06 mV | Δy 820 µV | Δy 3.6 mV |

Timebase 0.00 µs
2.00 µs/div
5.00 kS 250 MS/s

Trigger C2
Auto 78 mV
Edge Positive

X1= 1.024 µs ΔX= -1.024 µs
X2= 0 ns 1/ΔX= -977 kHz

Figure A.8: A screenshot of the oscilloscope while measuring LSA and DILSA at 0.3 V.

# Appendix B

# Radiation Tests at TRIUMF

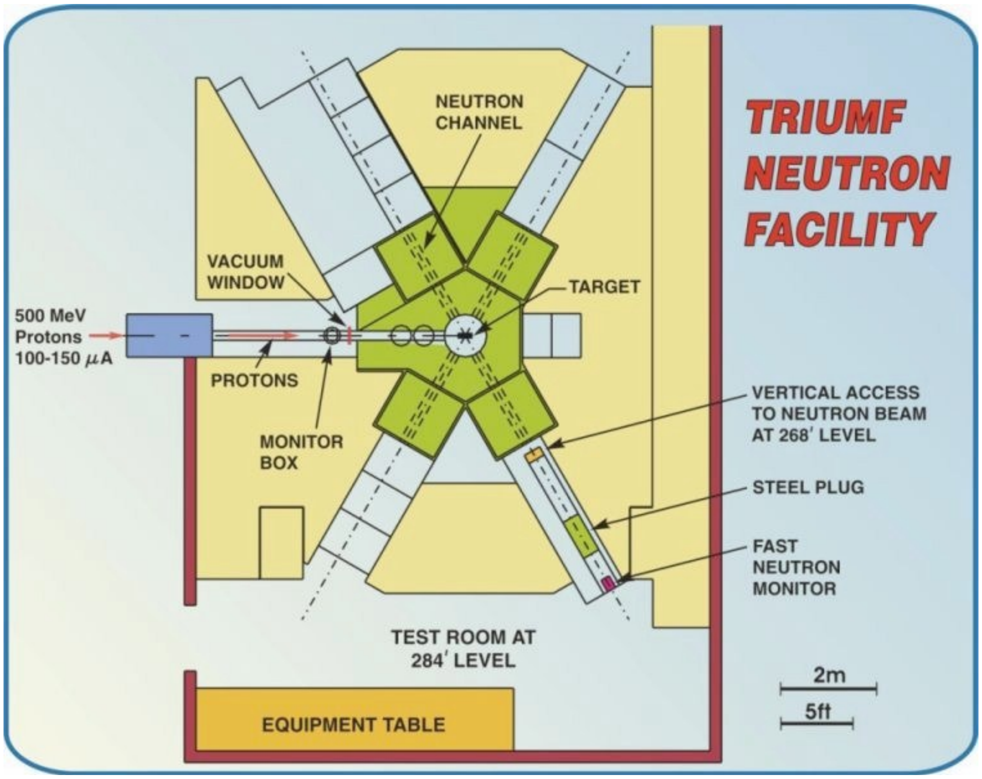
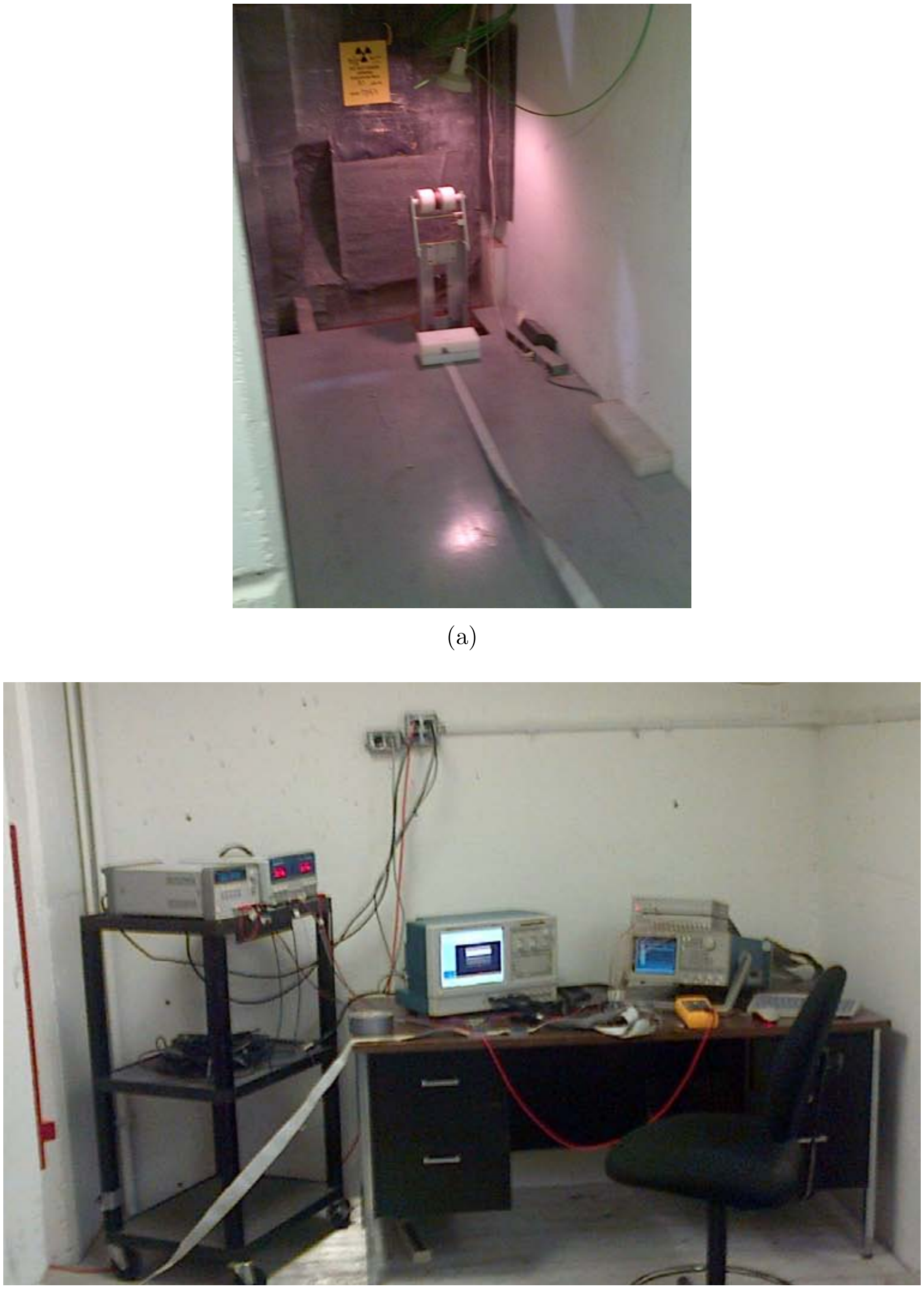


Figure B.1: Layout of the testing area at TRIUMF. The beam is accessed 5 m above the beam by a track system in a vertical slot in a shielding. Key areas to notice are the equipment table and the vertical access to the neutron beam.



(a)

(b)

Figure B.2: a) Vertical access to the neutron beam and b) Equipment table with the necessary setup.

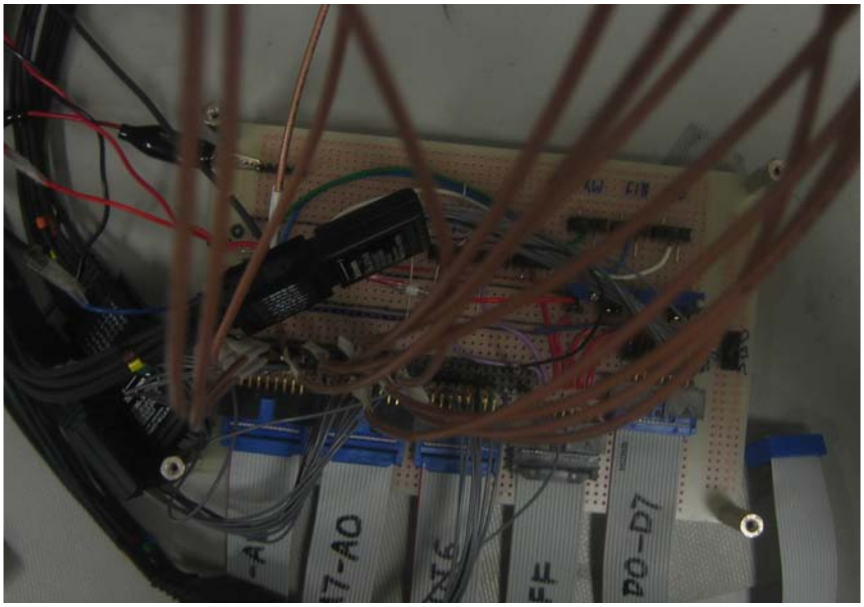


Figure B.3: Support PCB used at TRIUMF for the cables connecting to the test equipment. The other end of the cables was connected to the PCB which mounted Test Chip-4.

# References

[1] R. Baumann, "Soft errors in advanced semiconductor devices-Part I: the three radiation sources," *IEEE Trans. Dev. Mat. Rel.*, vol. 1, pp. 17–22, Mar. 2001. 1, 2, 4

[2] P. Dodd and L. Massengill, "Basic Mechanisms and Modeling of Single-Event Upset in Digital Microelectronics," *IEEE Trans. Nucl. Sci.*, vol. 50, pp. 583– 602, Jun. 2003. 1, 3, 7, 51

[3] J. Wallmark and S. Marcus, "Minimum size and maximum packing density of nonredundant semiconductor devices," *Proceedings of the IRE*, vol. 50, pp. 286–298, Mar. 1962. 1

[4] D. Binder, E. C. Smith, and A. B. Holman, "Satellite anomalies from galactic cosmic rays," *IEEE Trans. Nucl. Sci.*, vol. 22, pp. 2675–2680, Dec. 1975. 2

[5] T. May and M. Woods, "Alpha-particle-induced soft errors in dynamic memories," *IEEE Trans. Electron Devices*, vol. 26, pp. 2 – 9, Jan. 1979. 2

[6] R. Baumann, "Radiation-induced soft errors in advanced semiconductor technologies," *IEEE Trans. Dev. Mat. Rel.*, vol. 5, pp. 305– 316, Sept. 2005. 2, 3, 9

[7] S. Michalak, K. Harris, N. Hengartner, B. Takala, and S. Wender, "Predicting the number of fatal soft errors in Los Alamos National Laboratory's ASC Q supercomputer," *IEEE Trans. Dev. Mat. Rel.*, vol. 5, pp. 329– 335, Sep. 2005. 2, 3

[8] J. Maiz, S. Hareland, K. Zhang, and P. Armstrong, “Characterization of multi-bit soft error events in advanced SRAMs,” *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 21.4.1– 21.4.4, Dec. 2003. 2, 4

[9] D. Radaelli, H. Puchner, S. Wong, and S. Daniel, “Investigation of Multi-bit Upsets in a 150 nm Technology SRAM Device,” *IEEE Trans. Nucl. Sci.*, vol. 52, pp. 2433–2437, Dec. 2005. 2

[10] N. Derhacobian, V. Vardanian, and Y. Zorian, “Embedded memory reliability: the SER challenge,” *Int. Workshop Memory Tech. Design Testing, 2004*, pp. 104– 110, Aug. 2004. 2

[11] W. Atkinson and W. Seidler, “Impact of device scaling and material composition on the soft error rates in avionic systems,” *Proc. IEEE Southeast Conf.*, pp. 601–605, Mar. 2007. 2

[12] P. Meaney, S. Swaney, P. Sanda, and L. Spainhower, “IBM z990 soft error detection and recovery,” *IEEE Trans. Dev. Mat. Rel.*, vol. 5, pp. 419– 427, Sep. 2005. 2

[13] R. Baumann, “The impact of technology scaling on soft error rate performance and limits to the efficacy of error correction,” *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 329–332, 2002. 4

[14] P. Roche and G. Gasiot, “Impacts of front-end and middle-end process modifications on terrestrial soft error rate,” *IEEE Trans. Dev. Mat. Rel.*, vol. 5, pp. 382– 396, Sep. 2005. 4

[15] N. Seifert and N. Tam, “Timing vulnerability factors of sequentials,” *IEEE Trans. Dev. Mat. Rel.*, vol. 4, pp. 516– 522, Sep. 2004. 5

[16] P. Shivakumar, M. Kistler, S. Keckler, D. Burger, and L. Alvisi, “Modeling the effect of technology trends on the soft error rate of combinational logic,” *Proc. Int. Conf. Dependable Syst. Networks*, pp. 389– 398, 2002. 6

[17] J. S. Shah, “Design of soft error robust high speed 64-bit logarithmic adder,” *MASc Thesis: University of Waterloo*, 2008. 6

[18] S.-W. Fu, A. Mohsen, and T. May, “Alpha-particle-induced charge collection measurements and the effectiveness of a novel p-well protection barrier on VLSI memories,” *IEEE Trans. Electron Devices*, vol. 32, pp. 49–54, Jan. 1985. 8

[19] D. Burnett, C. Lage, and A. Bormann, “Soft-error-rate improvement in advanced BiCMOS SRAMs,” *Proc. IEEE Int. Rel. Physics Symp.*, pp. 156–160, Mar. 1993. 8

[20] H. Puchner, D. Radaelli, and A. Chatila, “Alpha-particle SEU performance of SRAM with triple well,” *IEEE Trans. Nucl. Sci.*, vol. 51, pp. 3525–3528, Dec. 2004. 8

[21] E. Cannon, et al., “SRAM SER in 90, 130 and 180 nm bulk and SOI technologies,” *Proc. IEEE Int. Rel. Physics Symp.*, pp. 300–304, Apr. 2004. 8

[22] S. E. Diehl, A. Ochoa, P. V. Dressendorfer, R. Koga, and W. A. Kolasinski, “Error analysis and prevention of cosmic ion-induced soft errors in static CMOS RAMs,” *IEEE Trans. Nucl. Sci.*, vol. 29, pp. 2032–2039, Dec. 1982. 8, 11, 30

[23] F. Ootsuka, M. Nakamura, T. Miyake, S. Iwahashi, Y. Ohira, T. Tamaru, K. Kikushima, and K. Yamaguchi, “A novel 0.20 $\mu$m full CMOS SRAM cell using stacked cross couple with enhanced soft error immunity,” in *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 205–208, 6-9 1998. 8, 11, 30

[24] T. Calin, M. Nicolaidis, and R. Velazco, “Upset hardened memory design for submicron CMOS technology,” *IEEE Trans. Nucl. Sci.*, vol. 43, pp. 2874–2878, Dec. 1996. 8, 11, 30

[25] S. Jahinuzzaman, D. Rennie, and M. Sachdev, “A soft error tolerant 10T SRAM bit-cell with differential read capability,” *IEEE Trans. Nucl. Sci.* 8, 11, 30, 31, 32, 61, 62, 63

[26] C. L. Chen and M. Y. Hsiao, “Error-correcting codes for semiconductor memory applications: A state-of-the-art review,” *IBM J. of Research and Develop.*, vol. 28, pp. 124–134, Mar. 1984. 8

[27] “International Technology Roadmap for Semiconductors (ITRS) Report.” 9

[28] S. R. Corporation, “Gate arrays wane while standard cell soar: ASIC market evolution continues,” 2008. 9

[29] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, “Drowsy caches: simple techniques for reducing leakage power,” *Proc. Int. Symp. Comput. Architecture*, pp. 148–157, 2002. 10

[30] V. Degalahal, N. Vijaykrishnan, and M. Irwin, “Analyzing soft errors in leakage optimized SRAM design,” *Proc. Int. Conf. VLSI Design*, pp. 227– 233, Jan. 2003. 10

[31] N. Azizi, F. Najm, and A. Moshovos, “Low-leakage asymmetric-cell SRAM,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, pp. 701– 715, Aug. 2003. 10, 27

[32] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, “SRAM leakage suppression by minimizing standby supply voltage,” *Proc. Int. Symp. Quality Electron. Design*, pp. 55– 60, 2004. 11, 26

[33] A. Agarwal, H. Li, and K. Roy, “A single-Vt low-leakage gated-ground cache for deep submicron,” *IEEE J. of Solid-State Circuits*, vol. 38, pp. 319– 328, Feb. 2003. 11

[34] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, “SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction,” *IEEE J. of Solid-State Circuits*, vol. 40, pp. 895– 901, Apr. 2005. 11

[35] V. Degalahal, L. Li, V. Narayanan, M. Kandemir, and M. Irwin, “Soft errors issues in low-power caches,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, pp. 1157– 1166, Oct. 2005. 11

[36] D. Krueger, E. Francom, and J. Langsdorf, “Circuit design for voltage scaling and SER immunity on a Quad-Core Itanium processor,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*. 11

[37] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital integrated circuits - A design perspective*. Prentice Hall, 2ed ed., 2004. 17, 18

[38] M. Yoshimoto, K. Anami, H. Shinohara, T. Yoshihara, H. Takagi, S. Nagao, S. Kayano, and T. Nakano, "A divided word-line structure in the static RAM and its application to a 64K full CMOS RAM," *IEEE J. of Solid-State Circuits*, vol. 18, pp. 479 –485, Oct. 1983. 19

[39] T. Hirose, H. Kuriyama, S. Murakami, K. Yuzuriha, T. Mukai, K. Tsutsumi, Y. Nishimura, Y. Kohno, and K. Anami, "A 20-ns 4-Mb CMOS SRAM with hierarchical word decoding architecture," *IEEE J. of Solid-State Circuits*, vol. 25, pp. 1068 –1074, Oct. 1990. 19

[40] B. Amrutur and M. Horowitz, "Fast low-power decoders for RAMs," *IEEE J. of Solid-State Circuits*, vol. 36, pp. 1506 –1515, Oct 2001. 20

[41] S. Schuster, B. Chappell, R. Franch, P. Greier, S. Klepner, F. Lai, P. Cook, R. Lipa, R. Perry, W. Pokorny, and M. Roberge, "A 15-ns CMOS 64K RAM," *IEEE J. of Solid-State Circuits*, vol. 21, pp. 704 – 712, Oct 1986. 24

[42] B. Amrutur and M. Horowitz, "A replica technique for wordline and sense control in low-power SRAM's," *IEEE J. of Solid-State Circuits*, vol. 33, pp. 1208 –1219, Aug 1998. 24

[43] S. Tachibana, H. Higuchi, K. Takasugi, K. Sasaki, T. Yamanaka, and Y. Nakagome, "A 2.6-ns wave-pipelined CMOS SRAM with dual-sensing-latch circuits," *IEEE J. of Solid-State Circuits*, vol. 30, pp. 487–490, Aapr 1995. 24

[44] E. Seevinck, F. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. of Solid-State Circuits*, vol. 22, pp. 748– 754, Oct. 1987. 25

[45] F. Hamzaoglu, Y. Ye, A. Keshavarzi, K. Zhang, S. Narendra, S. Borkar, M. Stan, and V. De, "Analysis of dual-Vth SRAM cells with full-swing single-ended bit line sensing for on-chip cache," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 10, pp. 91 –95, Apr. 2002. 27

[46] K. Osada, Y. Saitoh, E. Ibe, and K. Ishibashi, "16.7-fA/cell tunnel-leakage-suppressed 16-Mb SRAM for handling cosmic-ray-induced multierrors," *IEEE J. of Solid-State Circuits*, vol. 38, pp. 1952– 1957, Nov. 2003. 28

[47] B.-D. Yang and L.-S. Kim, “A low-power SRAM using hierarchical bit line and local sense amplifiers,” *IEEE J. of Solid-State Circuits*, vol. 40, pp. 1366– 1376, Jun. 2005. 29

[48] K. Kanda, H. Sadaaki, and T. Sakurai, “90% write power-saving SRAM using sense-amplifying memory cell,” *IEEE J. of Solid-State Circuits*, vol. 39, pp. 927– 933, Jun. 2004. 29

[49] K. Lakshmikumar, R. Hadaway, and M. Copeland, “Characterisation and modeling of mismatch in MOS transistors for precision analog design,” *IEEE J. of Solid-State Circuits*, vol. 21, pp. 1057–1066, Dec. 1986. 30, 64, 66

[50] M. Pelgrom, A. Duinmaijer, and A. Welbers, “Matching properties of MOS transistors,” *IEEE J. of Solid-State Circuits*, vol. 24, pp. 1433– 1439, Oct. 1989. 30, 64, 66

[51] M. Pelgrom, H. Tuinhout, and M. Vertregt, “Transistor matching in analog CMOS applications,” in *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 915–918, Dec. 1998. 30

[52] K. Zhang, K. Hose, V. De, and B. Senyk, “The scaling of data sensing schemes for high speed cache design in sub-0.18 $\mu$m technologies,” in *Symp. VLSI Circuits Dig. Tech. Papers*, pp. 226 –227, 2000. 30, 64

[53] P. Roche, F. Jacquet, C. Caillat, and J.-P. Schoellkopf, “An alpha immune and ultra low neutron SER high density SRAM,” in *Proc. IEEE Int. Rel. Physics Symp.*, pp. 671– 672, 25-29 2004. 30

[54] S.-M. Jung, H. Lim, W. Cho, H. Cho, H. Hong, J. Jeong, S. Jung, H. Park, B. Son, Y. Jang, and K. Kim, “Soft error immune 0.46 $\mu m^2$ SRAM cell with MIM node capacitor by 65 nm CMOS technology for ultra high speed SRAM,” in *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 11.4.1 – 11.4.4, Dec. 2003. 30

[55] G. Srinivasan, P. Murley, and H. Tang, “Accurate, predictive modeling of soft error rate due to cosmic rays and chip alpha radiation,” *Proc. IEEE Int. Rel. Physics Symp.*, pp. 12–16, Apr. 1994. 42

[56] F. Arnaud, et al., “A functional 0.69 $\mu m^2$ embedded 6T-SRAM bit cell for 65 nm CMOS platform,” in *Symp. VLSI Circuits Dig. Tech. Papers.* 57, 59

[57] K. Utsumi, E. Morifuji, M. Kanda, S. Aota, T. Yoshida, K. Honda, Y. Matsubara, S. Yamada, and F. Matsuoka, “A 65nm low power CMOS platform with 0.495 $\mu m^2$ SRAM for digital processing and mobile applications,” in *Symp. VLSI Tech. Dig. Tech. Papers*, pp. 216 – 217, Jun. 2005. 57, 59

[58] Y. Wang, H. Ahn, U. Bhattacharya, T. Coan, F. Hamzaoglu, W. Hafez, C.-H. Jan, R. Kolar, S. Kulkarni, J. Lin, Y. Ng, I. Post, L. Wel, Y. Zhang, K. Zhang, and M. Bohr, “A 1.1GHz 12 $\mu A$/Mb-leakage SRAM design in 65nm ultra-low-power CMOS with integrated leakage reduction for mobile applications,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers.* 57, 59

[59] J. S. S. T. Association, “Test method for beam accelerated soft error rate,” *JEDEC89-3A*, pp. 1–28, Nov. 2007. 59

[60] S. Clerc, F. Abouzeid, G. Gasiot, D. Gauthier, D. Soussan, and P. Roche, “A 0.32V, 55fJ per bit access energy, cmos 65 nm bit-interleaved SRAM with radiation soft error tolerance,” in *Proc. IEEE Int. Conf. on IC Design Tech. (ICICDT)*, pp. 1–4, Jun. 2012. 61, 62

[61] J. Autran, S. Serre, D. Munteanu, S. Martinie, S. Semikh, S. Sauze, S. Uznanski, G. Gasiot, and P. Roche, “Real-time soft-error testing of 40 nm SRAMs,” in *Proc. IEEE Int. Rel. Physics Symp. (IRPS)*, pp. 3C.5.1 –3C.5.9, Apr. 2012. 61, 62

[62] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, “Neutron-induced soft errors and multiple cell upsets in 65-nm 10t subthreshold sram,” *IEEE Trans. Nucl. Sci.*, vol. 58, pp. 2097 –2102, Aug. 2011. 62

[63] P. Hazucha, C. Svensson, and S. Wender, “Cosmic-ray soft error rate characterization of a standard 0.6-$\mu m$ CMOS process,” *IEEE J. of Solid-State Circuits*, vol. 35, pp. 1422–1429, Oct. 2000. 62

[64] P. Hazucha and C. Svensson, "Impact of CMOS technology scaling on the atmospheric neutron soft error rate," *IEEE Trans. Nucl. Sci.*, vol. 47, pp. 2586 –2594, Dec. 2000. 62

[65] P. Roche, J. Palau, G. Bruguier, C. Tavernier, R. Ecoffet, and J. Gasiot, "Determination of key parameters for SEU occurrence using 3-D full cell SRAM simulations," *IEEE Trans. Nucl. Sci.*, vol. 46, pp. 1354 –1362, Dec. 1999. 62

[66] R. Houle, "Simple statistical analysis techniques to determine minimum sense amp set times," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, pp. 37–40, Sept. 2007. 64

[67] X. Tang, V. De, and J. Meindl, "Intrinsic MOSFET parameter fluctuations due to random dopant placement," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 5, pp. 369–376, Dec. 1997. 64, 65

[68] A. Asenov, S. Kaya, and A. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," *IEEE Trans. Electron Devices*, vol. 50, pp. 1254 – 1260, May 2003. 64, 65

[69] Y. Li, C.-H. Hwang, T.-Y. Li, and M.-H. Han, "Process-variation effect, metal-gate work-function fluctuation, and random-dopant fluctuation in emerging CMOS technologies," *IEEE Trans. Electron Devices*, vol. 57, pp. 437–447, Feb. 2010. 64, 65

[70] X. Yuan, et al., "Transistor mismatch properties in deep-submicrometer CMOS technologies," *IEEE Trans. Electron Devices*, vol. 58, pp. 335–342, Feb. 2011. 64, 65

[71] M. Abu-Rahma, Y. Chen, W. Sy, W. L. Ong, L. Y. Ting, S. S. Yoon, M. Han, and E. Terzioglu, "Characterization of SRAM sense amplifier input offset for yield prediction in 28nm CMOS," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, pp. 1–4, Sept. 2011. 64, 65

[72] M. Sinha, et al., "Low voltage sensing techniques and secondary design issues for sub-90nm caches," in *Proc. European Solid-State Circuits Conf. (ESSCIRC)*, pp. 413– 416, Sept. 2003. 64, 69

[73] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, “A yield-optimized latch-type SRAM sense amplifier,” in *Proc. European Solid-State Circuits Conf. (ESSCIRC)*, pp. 409–412, Sept. 2003. 64

[74] R. Kraus, “Analysis and reduction of sense-amplifier offset,” *IEEE J. of Solid-State Circuits*, vol. 24, pp. 1028–1033, Aug. 1989. 64, 65

[75] A. Bhavnagarwala, et al., “Fluctuation limits and scaling opportunities for CMOS SRAM cells,” in *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 659–662, Dec. 2005. 64, 65

[76] N. Verma and A. Chandrakasan, “A high-density 45 nm SRAM using small-signal non-strobed regenerative sensing,” *IEEE J. of Solid-State Circuits*, vol. 44, pp. 163–173, Jan. 2009. 64, 65

[77] M. J. Lee, “A sensing noise compensation bit line sense amplifier for low voltage applications,” *IEEE J. of Solid-State Circuits*, vol. 46, pp. 690–694, Mar. 2011. 64, 65

[78] M.-F. Chang, et al., “An offset-tolerant current-sampling-based sense amplifier for sub-100nA-cell-current nonvolatile memory,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 206–208, Feb. 2011. 64, 65

[79] S. Lovett, G. Gibbs, and A. Pancholy, “Yield and matching implications for static RAM memory array sense-amplifier design,” *IEEE J. of Solid-State Circuits*, vol. 35, pp. 1200–1204, Aug. 2000. 65

[80] C. Mezzomo, A. Bajolet, A. Cathignol, and G. Ghibaudo, “Drain current variability in 45nm heavily pocket-implanted bulk MOSFET,” in *Proc. European Solid-State Device Research Conf. (ESSDERC)*, pp. 122–125, Sept. 2010. 65

[81] M. Bolatkale, M. Pertijs, W. Kindt, J. Huijsing, and K. Makinwa, “A single-temperature trimming technique for MOS-Input operational amplifiers achieving 0.33 $\mu v/^{\circ}c$ offset drift,” *IEEE J. of Solid-State Circuits*, vol. 46, pp. 2099–2107, Sept. 2011. 65

[82] T. Hook, J. Johnson, A. Cathignol, A. Cros, and G. Ghibaudo, "Comment on channel length and threshold voltage dependence of a transistor mismatch in a 32-nm HKMG technology," *IEEE Trans. Electron Devices*, vol. 58, pp. 1255–1256, Apr. 2011. 65

[83] R. Sarpeshkar, J. Wyatt, J.L., N. Lu, and P. Gerber, "Mismatch sensitivity of a simultaneously latched CMOS sense amplifier," in *IEEE Int. Symp. Circuits and Syst. (ISCAS)*, pp. 2224–2227 vol.4, Jun. 1991. 65

[84] Y. Watanabe, N. Nakamura, and S. Watanabe, "Offset compensating bit-line sensing scheme for high density DRAMs," *IEEE J. of Solid-State Circuits*, vol. 29, pp. 9–13, Jan. 1994. 65, 67

[85] D. Laurent, "Sense amplifier signal margins and process sensitivities," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 49, pp. 269–275, Mar. 2002. 65

[86] Y.-H. Chen, S.-Y. Chou, Q. Li, W.-M. Chan, D. Sun, H.-J. Liao, P. Wang, M.-F. Chang, and H. Yamauchi, "Compact measurement schemes for bit-line swing, sense amplifier offset voltage, and word-line pulse width to characterize sensing tolerance margin in a 40 nm fully functional embedded SRAM," *IEEE J. of Solid-State Circuits*, vol. 47, pp. 969–980, Apr. 2012. 65

[87] K. Seno, K. Knorpp, L.-L. Shu, N. Teshima, H. Kihara, H. Sato, F. Miyaji, M. Takeda, M. Sasaki, Y. Tomo, P. Chuang, and K. Kobayashi, "A 9-ns 16-Mb CMOS SRAM with offset-compensated current sense amplifier," *IEEE J. of Solid-State Circuits*, vol. 28, pp. 1119–1124, Nov. 1993. 67, 85

[88] J. Takahashi, T. Wada, and Y. Nishimura, "A dynamic current-offset calibration sense amplifier with fish-bone shaped bitline for high-density SRAMs," in *Symp. VLSI Circuits Dig. Tech. Papers*, pp. 115–116, Jun. 1994. 67, 85

[89] T. Kawahara, T. Sakata, K. Itoh, Y. Kawajiri, T. Akiba, G. Kitsukawa, and M. Aoki, "A high-speed, small-area, threshold-voltage-mismatch compensation sense amplifier for gigabit-scale DRAM arrays," *IEEE J. of Solid-State Circuits*, vol. 28, pp. 816–823, Jul. 1993. 67

[90] T. Furuyama, S. Saito, and S. Fujii, "A new sense amplifier technique for VLSI dynamic RAM's," in *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, vol. 27, pp. 44 – 47, 1981. 68

[91] K. Ishibashi, K. Takasugi, K. Komiyaji, H. Toyoshima, T. Yamanaka, A. Fukami, N. Hashimoto, N. Ohki, A. Shimizu, T. Hashimoto, T. Nagano, and T. Nishida, "A 6-ns 4-Mb CMOS SRAM with offset-voltage-insensitive current sense amplifiers," *IEEE J. of Solid-State Circuits*, vol. 30, pp. 480–486, Apr. 1995. 68, 85

[92] T. Sakurai, "High-speed circuit design with scaled-down MOSFET's and low supply voltage," in *IEEE Int. Symp. Circuits and Syst. (ISCAS)*, pp. 1487 –1490, May 1993. 68

[93] R. Singh and N. Bhat, "An offset compensation technique for latch type sense amplifiers in high-speed low-power SRAMs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, pp. 652–657, Jun. 2004. 68

[94] M. Bhargava, M. McCartney, A. Hoefler, and K. Mai, "Low-overhead, digital offset compensated, SRAM sense amplifiers," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, pp. 705–708, Sept. 2009. 68, 85

[95] J. Montanaro, R. Witek, K. Anne, A. Black, E. Cooper, D. Dobberpuhl, P. Donahue, J. Eno, W. Hoeppner, D. Kruckemyer, T. Lee, P. Lin, L. Madden, D. Murray, M. Pearce, S. Santhanam, K. Snyder, R. Stehpany, and S. Thierauf, "A 160-MHz, 32-b, 0.5-W CMOS RISC microprocessor," *Solid-State Circuits, IEEE Journal of*, vol. 31, pp. 1703 –1714, Nov. 1996. 69

[96] M. Sharifkhani, E. Rahiminejad, S. Jahinuzzaman, and M. Sachdev, "A compact hybrid current/voltage sense amplifier with offset cancellation for high-speed SRAMs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, pp. 883–894, May 2011. 69, 85

[97] B. Bateman, C. Freeman, J. Halbert, K. Hose, G. Petrie, and E. Reese, "A 450 MHz 512 kB second-level cache with a 3.6 GB/s data bandwidth," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 358–359, Feb 1998. 86