OVERCOMING THE CIRCUIT DESIGN CHALLENGES IN NANOSCALE SRAMs

by

Praveen Elakkumanan

A dissertation

submitted to the Department of Computer Science and Engineering of the State University of New York at Buffalo

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

September 2006

ACKNOWLEDGMENT

This section is devoted to all the people who helped me in completing this dissertation. My advisor, Prof. Sridhar, had played a very important role in motivating me both academically and personally throughout my PhD program. His critical comments and suggestions during the various stages of this program has played a vital role in its completion. I am particularly indebted to his constructive criticism and non-academic support.

I sincerely thank Prof. Sakurai, University of Tokyo for taking time off his tight schedule and commenting on my work as an outside reader. I also thank Prof. Upadhyaya and Prof. Scott for participating as members of my dissertation committee and for reviewing my thesis report at a very short notice.

I would like to extend my special thanks to Dr. Kevin Nowka and Dr. J. B. Kuang, IBM Austin Research Lab for providing me the opportunity to work with them. I also thank Mr. Thomas Sandwick and Dr. J. O. Plouchart at IBM Semiconductor Research and Development Center.

This work was partly funded by the New York State of office of Science, Technology and Academic Research (NYSTAR) through Micro-electronic Design Center (MDC). I thank my research group members Ashok, Geetha, Manjari, Srikanth, Charan and Kishan for offering me a cordial environment to work in. Working amidst them was definitely an enriching experience. I thank Dr. Mishra, Compsys Technologies Ltd. for her guidance and input in the initial stages of my research. I am thankful to TK, Mangalam, Richa, Ashok, GT, Ninju, Looney, Nara, Sam, Koyan and Param for the memorable moments outside my work and during my

stay in Buffalo. Thanks guys!

I am greatly indebted to my parents and sister, Lavan for their constant motivation and moral support throughout the period of my study. None of this would have been possible without the guidance and inspiration from my uncle, Dr. Bala.

ABSTRACT

Most microprocessors use large on-chip SRAM caches to bridge the performance gap between the processor and the main memory. Due to their growing embedded applications coupled with the technology scaling challenges, considerable attention is given to the design of low-power and high-performance SRAMs. However, there are many challenges in the design of both embedded and stand-alone SRAMs, such as, the estimation and optimization of stand-by power, design of high-speed peripheral circuits, and design of robust circuits for low-voltage operation.

Further, as the technology continues scaling into the nanometer domain, controlling the variation in device parameters during fabrication becomes a great challenge. Variations in process parameters, such as, oxide thickness, channel length, channel width and dopant concentration can result in large variations in threshold voltage. This in turn is expected to severely affect the functionality of the minimum geometry transistors that are commonly used in SRAM designs. Our studies of new memory and peripheral circuits have shown significant promise in terms of power, speed and robustness.

In this research, we address the following problems:

- Circuit techniques to estimate and simultaneously reduce gate leakage and sub-threshold leakage
- Process variations tolerant design approaches to reliably sense and amplify the bitlines with a minimum discharge providing a fast and accurate readout at low power

- Failure analysis to understand the impact of process variations, soft errors, leakage and noise on different memory fault mechanism to help in the design of variation tolerant low power and high performance memories
- Design of test structures for CMOS process tuning and variation control, and improvement of SRAM reliability by predicting the design yield early in the product cycle.

In short, this dissertation characterizes the issues in nanoscale memory design, which will have a ubiquitous presence in commercial electronic market. It is important for these systems to be reliable, fast and consume less power, thereby, increasing battery life. Design techniques to achieve these goals are presented.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Considerable attention has been given to the design of low-power and high-performance SRAMs since they are critical components in both high-performance processors and hand-held portable devices. The design of high-performance computer systems require SRAMs with cycle times below 5ns for the cache and control memories. With the process technology pushing well into the ultra deep sub-micron (UDSM) arena, IC designers can now integrate significant densities of memory and logic together in the same chip. Such an embedded SRAM market is even larger than stand-alone SRAM market [1] and this memory on chip reduces cost with improved speed performance. Design of higher speed and higher density SRAMs is necessary because of their growing embedded applications.

1.1 SRAM Design Issues

The ever-increasing levels of on-chip integration of sub-100*nm* SRAMs pose serious design challenges in terms of power and speed performance. The following are the major challenges in the design of an efficient SRAM.

1.1.1 Power Consumption

In recently presented reduced-power processors, nearly half of the total system power consumption is attributed to the memory circuits [2, 3, 4]. Hence, reducing the power dissipation in memories can significantly improve the system power-efficiency, performance, reliability and overall costs. Historically, the primary source of power dissipation has been the dynamic energy due to the charging/discharging of load capacitances when a device switches. Partitioned memory arrays and hierarchical word lines reduce the total capacitance that is switched per access [5]. As we delve deeper into the sub-micron region, scaling of both supply voltage and threshold voltage of transistors enables high-speed and low-power operation. However, it causes a significant increase in the sub-threshold static leakage current due to its exponential relation with the threshold voltage. This results in increased leakage (static) power dissipation that is almost 44% of the total power consumed in the recent Intel's Pentium III processor [6].

Due to the increasing fraction of chip area devoted to memory structures, state-of-art on-chip cache designs have unacceptably large leakage power dissipation [7]. Recent energy estimates for a $130nm$ process indicate that 30% of L1 cache energy and 80% of L2 cache energy is contributed to leakage power [8].

1.1.2 Read and Write Access Times

Embedded memory applications require techniques for maximizing the access speeds of static memories with minimal power consumption to optimize the overall system performance. The performance of the address decoders, sense amplifiers and the periphery I/O circuitry need to be simultaneously improved for

achieving this goal. The decoder delay contributes to nearly half the access time in a memory circuit. Hence, design of fast address decoders that consume minimal power are required for high-performance memories. Many techniques are presently available for improving the access times with or without significant penalties on other parameters.

Sense Amplifier is one of the most critical memory peripheral circuits. They strongly influence the memory access times as they are used to retrieve the stored data from the memory array by amplifying the small signal variations on the bitlines. Our preliminary studies and analysis show that major speed improvements are possible when using current-mode sensing techniques as opposed to conventional voltage mode sensing. The key to this approach is to reduce both the impedance at the sensing point and the voltage swings on long bitlines with the use of low-resistance current-signal sense amplifier circuits.

1.1.3 Reliability

CMOS technology scaling trends, applications and operating conditions have added both reliability and robustness as design metrics in addition to the traditional metrics of power, speed, area and cost. Reliability is normally defined as the immunity to hard failures such as electromigration, hot carrier effects, or dielectric breakdowns. Design robustness is the ability of a circuit to operate efficiently under varying process, temperature, voltage, and noise conditions.

Transient faults due to neutron or alpha particle strikes pose a serious reliability threat to nanoscale memories. Radiation-induced transient errors increase with increasing altitudes and reducing voltages, and thus affect the system robustness. In addition, these radiation-induced errors pose a significant obstacle

to increasing processor transistor counts in future technologies.

Transistor mismatch, or the inability to form tiny transistors which are electrically alike, has become a major robustness issue for sub-100nm CMOS technology due to the statistical nature of semiconductor processes. As technology scales, understanding manufacturing variation becomes essential to effectively design robust high performance memories.

1.1.4 Interconnect Delay

Interconnect performance issues for future technology nodes specified by the ITRS '05 is of major concern due to the increasing latency (RC delay) of global wires in sub-100*nm* technologies. It has been observed that over the past four decades interconnect scaling has increased the distributed RC product, thus resulting in larger latency for a given interconnect length. The interconnect hierarchy organizes the interconnect in local, intermediate and global levels to provide a solution for wiring complexity and higher on-chip frequencies. With respect to memories two kinds, of interconnects have been identified [9]: inter-memory interconnect connects memories and functional units to each other and intra-memory interconnect refers to the lines inside the memories *i.e.* mainly bitlines and wordlines.

It has been reported that intra-memory interconnect [9] dominates the energy consumption and will be contributing to more than half of the total interconnect energy consumed. This suggests that any technique that reduces the intramemory interconnect length will not only significantly reduce latency but also will save power. Further, physical variations in bitline interconnect lead to interconnect electrical-parameter (RCL) variation, which in turn leads to even more potential

variation in the actual performance (and power consumption) of the memory chip.

1.2 Dissertation Objectives

This dissertation provides circuit and architectural solutions to increase the efficiency of SRAM caches and thus help in addressing the processor-memory bottleneck problem. Design approaches for reducing the leakage power consumption and increasing the access speeds of memories are presented. Techniques to improve the reliability and robustness of SRAM designs are also be proposed.

Specific contributions of the dissertation are as follows:

- Design of novel low leakage memory cells that saves the state even during the sleep state.
- Low power and robust current mode techniques to sense the current difference in the accessed cell through the bitlines.
- Failure analyses and modeling of systematic process variations that result in memory failures.
- A detailed study on the effect of radiation induced upsets on SRAM functioning.
- SRAM ring test structures to charecterize the technology and to provide variation control.

1.3 Organization

The dissertation has been organized as follows:

- Chapter 2 gives an overview of the components of CMOS SRAMs and explains their functioning. The general circuit and architectural techniques involved in their design has also been explored.

- Chapter 3 deals with low leakage static memories. The existing methods for reducing leakage power in memory circuits are presented. The constraints of these techniques and the significance of our designs in handling these limitations are discussed in detail.

- The concept of current-mode operation for bitline sensing is introduced in Chapter 4. We also present two current sense amplifier designs, LPCSA and WTA, which consumes less power and operates at a higher speed as compared to the existing designs.

- Chapter 5 presents our failure analyses study on local bitline access schemes. The analyses were performed in an industry standard $65nm$ process technology using hardware models. The findings from the study are used to determine the design window available for a given array size.

- Chapter 6 discusses the problems due to radiation induced particle strikes on the functioning of SRAM circuits. In particular, we discuss about the radiation induced soft errors in memories and how Silicon-On-Insulator (SOI) technology provide an effective solution for the same.

- Chapter 7 presents the SRAM ring oscillator test structures that are designed to charecterize the technology and provide variation control early in the product cycle.

- Finally, the concluding remarks of this dissertation is presented in Chapter 8. We highlight the major contributions of this research work and also give pointers and specific directions for future work.

Chapter 2

Overview of CMOS SRAMs

SRAMs have experienced a very rapid development of low-power, low-voltage memory design during recent years due to an increased demand for notebooks, laptops, hand-held communication devices, and IC memory cards. A randon-access memory (RAM) is one in which the time required for storing (writing) information and retreiving (reading) information is independent of the physical location (within the memory) in which the information is stored. Static RAMs (SRAMs) utilize static latches as the storage cells and can hold their stored data indefinitely, provided the power supply remains on.

2.1 SRAM Organization

The bits on a memory chip are individually addressable, or addressable in groups of 4, 16 or 32 bits (a *word*). The bulk of the memory chip consists of the memory cells in which the bits are stored. Each memory cell is an electronic circuit capable of storing one bit. The componenets of a memory cell and its functioning are discussed in the following section. For easier addressing of the stored

information, it is desirable to physically organize the storage cells on a chip in a square or a nearly square matrix. Figure 2.1 illustrates such an organization [10]. The memory array has 2^M rows and 2^N columns, for a total storage capacity of 2^{M+N}. For example, a 4 Kb memory array would have 64 rows and 64 columns ($M = N = 64$). Each cell in the array is connected to one of the 2^M row lines, known as *wordlines*, and to one the 2^N column lines, known as the *bitlines*. Typically, each memory cell would be connected to one wordline and two complementary bitlines. A particular cell is selected for reading or writing by activating its wordline and bitlines.

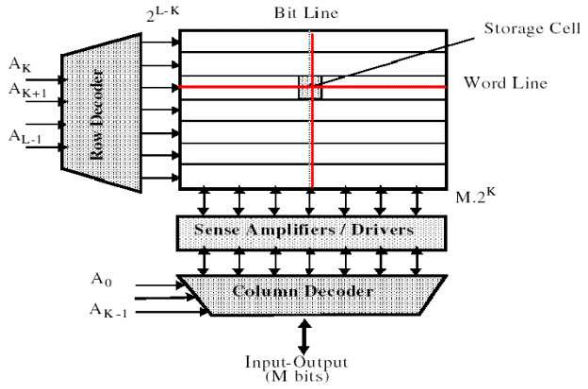

Figure 2.1: Static RAM Organization

One of the 2^M wordlines is activated by the *row decoder*, which is a combinational circuit that raises the voltage of the wordline whose M-bit address ($A_0A_1...A_{M-1}$) is applied to the decoder input. When the Kth wordline is activated for, say, a read operation, all the 2^N cells in row K will provide their contents to their respective bitlines. These contents, in the form of a small read-out signal would then be sensed by a *Sense Amplifer* connected to the bitlines. There is a

sense amplifier for each pair of bitlines to provide full-swing digital signal at its output. This signal, together with the output signals from all the other cells in the selected row, is then delivered to the *Column Decoder*. The coulmn decoder selects the signal of the column whose N-bit address is applied to the decoder input ($A_M A_{M+1} ... A_{M+N-1}$) and causes the signal to appear on the chip input/output (I/O) data line. Address decoders and Sense amplifiers are discussed in detail in Chapter 4.

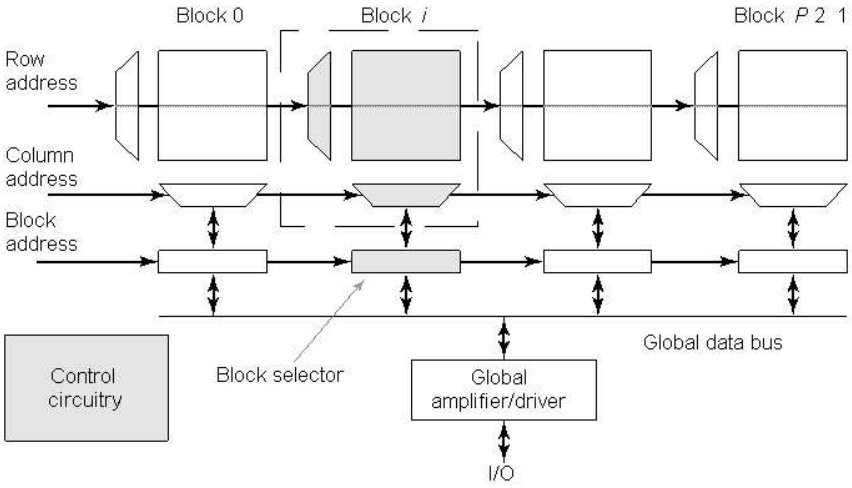

Figure 2.2: Static RAM Architecture

The architeure of Fig. 2.1 works well for smaller memories up to a range of 64 Kbits to 256 Kbits [11]. However, larger memories start to suffer from a serious speed degradation as the length, capacitance, and resistance of the word and bit lines become excessively large. Larger memories have consequently gone one step further and added one extra dimension to the address space, as illustrated in Fig. 2.2. The memory is partitioned into P smaller blocks. The composition of each of these individual blocks is identical to one of Fig. 2.1. A word is selected

in the basis of the row and column addresses that are broadcast to all the blocks. An extra address word called the block address, selects one of the P blocks to be read or written.

2.2 Static Memory Cell (6T-Cell)

The memory cell is the basic building block of any static memory system. Figure 2.3 shows a conventional 6-transistor static memory cell in CMOS technology [11]. The circuit is a flip-flop comprising two cross-coupled inverters and two access transistors, $Q5$ and $Q6$. The access transistors are turned on when the wordline is selected and connect the memory cell to the column bitlines (B or $\overline{B}$). They act as transmission gates allowing bideirectonal current flow between the flipflop and the two bitlines. Both the bitlines carry complementary data and connect all the memory cells in a single column. For performing a read or write operation on the memory cell, the wordline should be high to connect the memory cell and the bitlines. The fact that each memory cell has two bitlines is used to distinguish between a memory read or write operation. The following two subsections discusses the read and write operations of a memory cell in brief.

2.2.1 Read Operation

Before the read operation begins, the bitlines, B and $\overline{B}$ are precharged to a high value, usually V_{dd}. When the wordline is selected, the access transistors are turned on. This will cause a current to flow from V_{dd} through the pMOS pull-up tranistor of the node storing 1 and the access transistor onto the bitline B, charging the capacitance of line B, C_B. On the other side, current will flow from

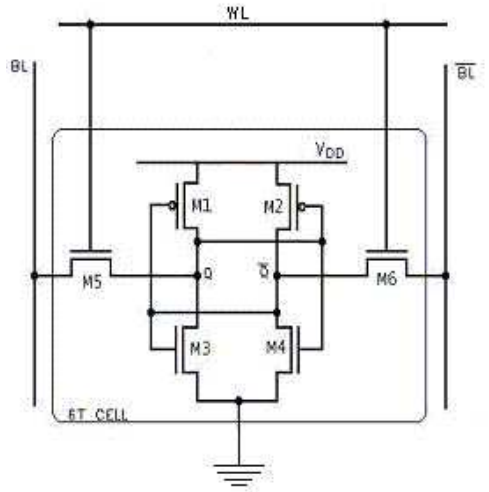

Figure 2.3: Conventional 6-T Static RAM Cell

the precharged $\overline{B}$ line through $Q5$ and $Q1$ to ground, thus discharging $C_{\overline{B}}$. From the above description. we note that the voltage across C_B will rise and that across $C_{\overline{B}}$ will fall. Thus, a differential voltage $v_{B\overline{B}}$ develops between the bitlines B and $\overline{B}$. This small potential differnce between the bitlines is sensed and amplified by the sense amplifers at the data output.

2.2.2 Write Operation

For a write operation, one of the bitlines is lowered to 0 V and the other one is raised to V_{dd}, and of course the cell should be selected by raising the word line. Depending on the value we are writing into the cell one of the bitlines is pulled low and the other is pulled high. The 0 on one of the bitlines will overpower the pMOS pullup and the flip-flop would change its state. However, during a read operation a 1 on the bitline should not overpower the nMOS pull down. If this happens read operation would destroy the state of the memory cell and operation

is no longer valid. These requirements would dictate the (W/L) ratios of all the transistors in the memory cell. Typically, the inverters are designed so that the pullup and pull down transistors are pitch matched. The access transistors are however made two to three times wider than the other transistors.

2.3 Sources of SRAM Power

There are different sources of active and stand-by (data retention) power present in SRAMs. The active power is the sum of the power consumed by the following components.

- Decoders
- Memory array
- Sense amplifiers
- Peripheral (I/O circuitry, write circuitry, etc.) circuits

The total active power of an SRAM with mxn array of cells can be summarized by the following expression [12].

$$P_{active} = (mi_{active} + m(n-1)i_{leak} + (n+m)fC_{DE}V_{INT} + mi_{DC}\Delta t f + C_{PT}V_{INT}f + I_{DCP})V_{dd} \quad (2.1)$$

where i_{active} is the effective current of the selected cells, i_{leak} is the effective data retention current of the unselected memory cells, C_{DE} is the output node capacitance of each decoder, V_{INT} is the internal power supply voltage, i_{DC} is the dc current consumed during read operation, Δt is the activation time of the dc current consuming parts (sense amplifiers), f is the operating frequency, C_{PT} is the

total capacitance of the CMOS logic and the driving circuits in the periphery, and I_{DCP} is the total static (dc) or quasistatic current of the periphery.

The stand-by power of an SRAM has a major source represented by $i_{leak}mn$ because the static current from other sources is negligibly small (sense amplifiers are disabled during this mode). Therefore, the total stand-by power can be expressed as:

$$P_{standby} = mni_{leak}xV_{dd} \tag{2.2}$$

2.4 Summary

A broad overview of static CMOS RAMs has been presented in this chapter. The organization of a SRAM memory array explaining the different components within the circuit is described. The basic building block of a SRAM, a static RAM memory cell, is then described in detail. The two possible operations on a memory cell are also discussed. Finally, all the different sources of power in a SRAM are summarized.

Chapter 3

Low Leakage SRAM Cells

Nearly half of the total system power consumption in recent low power processors is attributed to the memory circuits [3, 4]. Hence, reducing the power dissipation in memories can significantly improve the system power-efficiency, performance, reliability and overall costs. Historically, the primary source of power dissipation has been the dynamic energy due to the charging/discharging of load capacitances when a device switches. Partitioned memory arrays and hierarchical word lines reduce the total capacitance that is switched per access [5]. As we delve deeper into the sub-micron region, supply voltage scaling and threshold voltage scaling help in achieving high-speed and low-power operation. However, it causes a significant increase in both the sub-threshold static and gate leakage currents due to short channel effects and direct tunneling current due to low oxide thickness. This results in increased leakage (static) power dissipation that is almost 44% of the total power consumed in the recent Intel's Pentium IV processor [6]. Due to the increasing fraction of chip area devoted to memory structures, state-of-art on-chip cache designs have unacceptably large leakage power dissipation [7].

Energy estimates for a 130*nm* process indicate that 30% of L1 cache energy and 80% of L2 cache energy is due to leakage [8].

3.0.1 Leakage in SRAM

Technology scaling and hence lowering V_t impacts design of integrated circuits, especially SRAMs heavily, because of the increase in static power consumption due to the leakage in bitlines and degradation in cell-data stability.

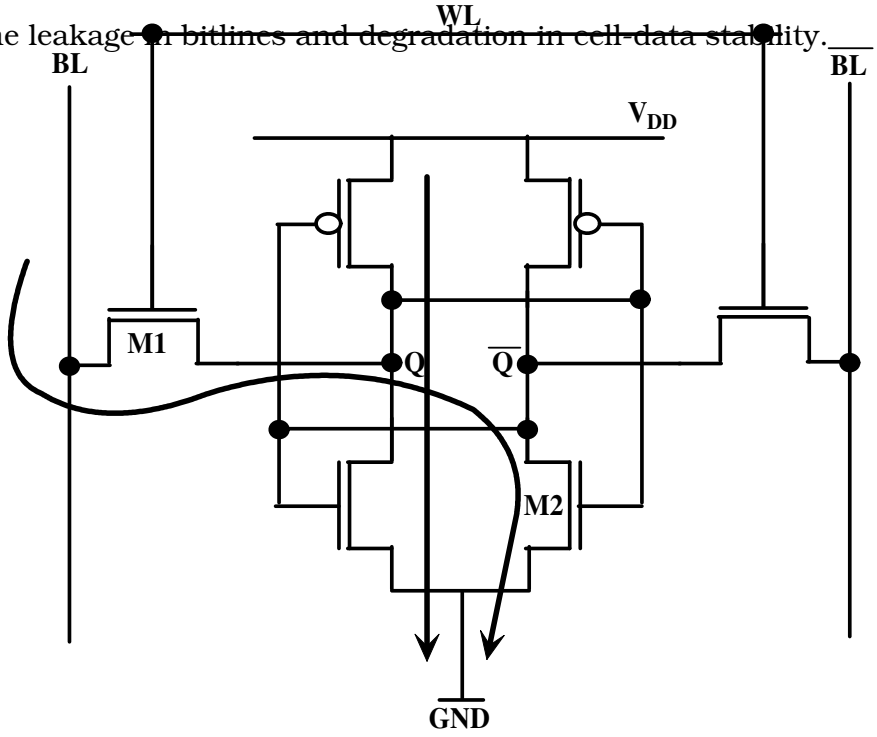

Figure 3.1: Two Dominant Leakage Paths in SRAM

In addition, stand-by dissipation increases due to leakage within the memory cells. Until now, for CMOS devices with large oxide thickness, sub-threshold leakage was the most dominant component of leakage in CMOS circuits. However, as the ITRS predicts, for sub-70*nm* devices, the gate oxide thickness reduces to a value of 1.1-1.6*nm*. At such low oxide thickness, MOSFET gate tunneling current

increases significantly. Thus, at ultra-thin gate oxide regime, gate tuneling current becomes appreciable and dominates the total "off" state leakage current of the transistor along with sub-threshold leakage [13]. Figure 3.1 shows the dominant leakage components for a conventional 6T SRAM cell when it is not accessed (stand-by mode). As we could observe, the two dominant sub-threshold leakage paths are: i) V_{dd} to ground and ii) bitline to ground leakage paths. Together, they make up 93% of the total leakage in SRAMs for larger gate oxide devices.

The gate direct tunneling current increases exponentially with decrease in the oxide thickness and increase in voltage across the oxide. The gate leakage component of any transistor depends on the voltage difference across its terminals. The voltages at which the bitlines are precharged determine the leakage through the access transistors, whereas, the internal node and supply voltages determine the gate leakage in the 4 transistors in the cross-coupled inverter.

There have been many efforts both in architectural and circuit level to solve the leakage problem in the SRAM [7, 8, 14, 15, 16, 17, 13]. However, these techniques either completely turn off circuits by creating a high-impedance path to ground (gating) or trade off increased access times for reduced static power consumption. In addition, many of these techniques do not address the problem of gate leakage in nanometer era. In some cases, these techniques can be implemented entirely at the circuit level without any changes to the architecture or may involve simple architectural modifications. The following sections discuss a few important existing techniques for leakage reduction in static memories, the constraints of these techniques and the significance of our approach in handling these limitations.

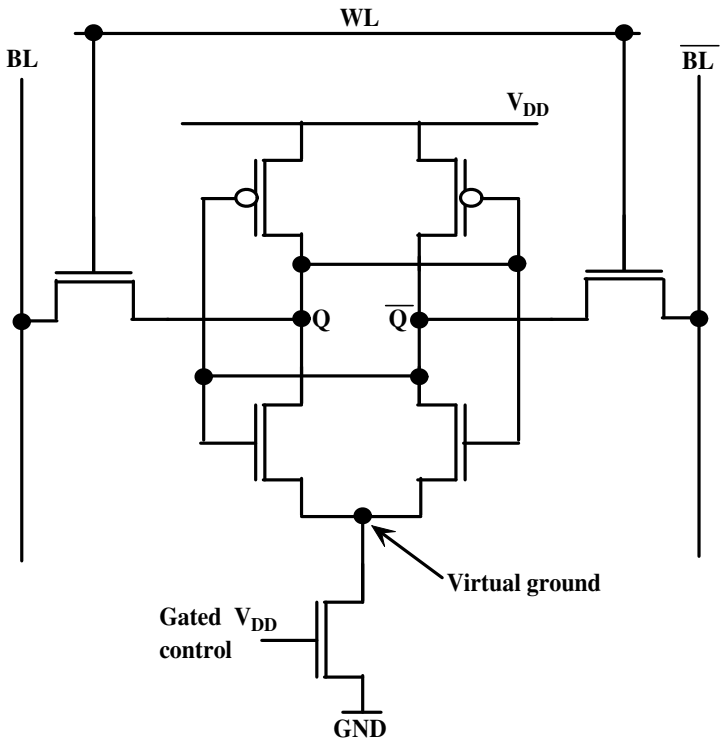

Figure 3.2: SRAM with an nMOS Gated-V_{dd}

3.0.2 Previous Work and their Limitations

As shown in Fig. 3.2, the dynamically resizable (DRI) I-cache presented in [7] uses a gated-ground nMOS transistor between the ground and SRAM cell to turnoff the unused sections of the cache and thus reduces the leakage power significantly. This approach for reducing the leakage power by varying the size of the cache using power-gating technique has high savings. However, the data in the memory cell is lost because of the floating nodes in the cross-coupled latch inside the 6T-SRAM memory cell. This will significantly increase the access times affecting the performance significantly.

Subsequent work (ABB-MTCMOS) [14] suggests dynamic increase in the threshold voltage (thus reducing leakage) when the cell is in the sleep mode to save the state of the memory cell. However, since this technique requires that the voltage of the N-Well and supplies be changed when the cell enters/leaves the sleep mode, energy and time required for this state transition can be high. More recent techniques such as dynamic V_t SRAM [8] and drowsy caches [15] also have certain limitations.

Dynamic V_t SRAM that increases the threshold voltage dynamically by body biasing, pose cost (due to twin well process) requirement and reliability problems. It also increases the energy and time needed to switch from/to sleep mode because of the high well capacitance. Also, the effectiveness of body biasing is reduced for low-V_t devices due to increased diode depletion resulting in degradation due to body effect. Drowsy caches may not lead to optimal reduction as active-mode leakage is not addressed and bitline leakage through the NMOS pulldown transistors is not reduced significantly. However, this approach retains the value of

the data stored in the memory cells and does not affect the read or write access times.

An asymmetric SRAM cell family is introduced in [16] to reduce the sub-threshold leakage power while maintaining low access latency. The SRAM cell has an asymmetric structure and the devices use dual V_t technology. An improvement of the above is proposed in [18] to reduce gate leakage power and increase the stability of the SRAMs. However, both these work assume a strong bias towards zero and work well only for caches that have large number of zeros stored in them.

A single V_t Data Retention Gated-Ground Cache (DRG-Cache) design and architecture is proposed in [19] to reduce power by setting the unused portions of the memory core to a low leakage mode. DRG-Cache uses an nMOS transistor as a gated-ground transistor to turn off the supply voltage. However, for data retention in the sleep mode this approach requires proper sizing of the gated-Ground transistor and an optimum value for the transistor threshold voltage. Moreover, when the gated-Ground transistor is turned off, the virtual ground node is left floating and this may result in a noise source degrading the stability of the data stored in the cell. Also, as mentioned by the authors, if the sub-threshold resistance of the gated-Ground transistor is very less, the data is either lost or the leakage savings is not optimal. The proposed design techniques overcome these limitations and simplifies the design and control of the memory cells when they are not used.

3.1 NC-SRAM Cell - Design and Analysis

This section presents an N-Controlled SRAM (NC-SRAM) design for maximum leakage reduction in cache and embedded memories without affecting the performance significantly. The data stored in the memory cells are retained even when the memory is operating in stand-by mode, thus ensuring that the read/write access times will not be affected during the normal mode of operation. Dual-threshold voltage (dual-V_t) process technology [20] allows integrating transistors with two different threshold voltages in the same circuit. These designs typically use high-V_t and V_{dd} devices for the transistors in the leakage critical paths and use low-V_t and V_{dd} devices for transistors in the performance critical paths. In our approach, we use high-V_t transistors in certain key leakage-prone parts of the NC-SRAM cell. In addition, we use supply voltage gating in a way that the data stored is not lost nor susceptible to noise that can destabilize the stored value, to achieve maximum leakage savings. As proved by simulation results, this integrated circuit solution offers key advantages over the available dual-V_t and supply voltage gating techniques.

3.1.1 NC-SRAM Cell: Circuit Details

The proposed method uses Dynamic Voltage Scaling (DVS) to reduce the subthreshold and gate leakage power of the cache cells and also retains the data stored during the inactive state. The key idea of the NC-SRAM is the use of two pass-transistors (Fig. 3.3) that provide different ground supply voltages to the memory cell for normal and sleep modes. These pass-transistors provide a positive ground supply voltage when the cell is inactive and connect the cross-coupled

inverters to the ground supply during normal operation to function as a conventional 6T-cell. The positive voltage during the stand-by state has 3 important effects: (i) Negative body to source potential causes more body effect resulting in increased threshold (ii) Gate to source voltage becomes negative (iii) Drain to source potential decreases resulting in less Drain Induced Barrier Lowering (DIBL). These effects combined together results in low sub-threshold and gate leakage currents.

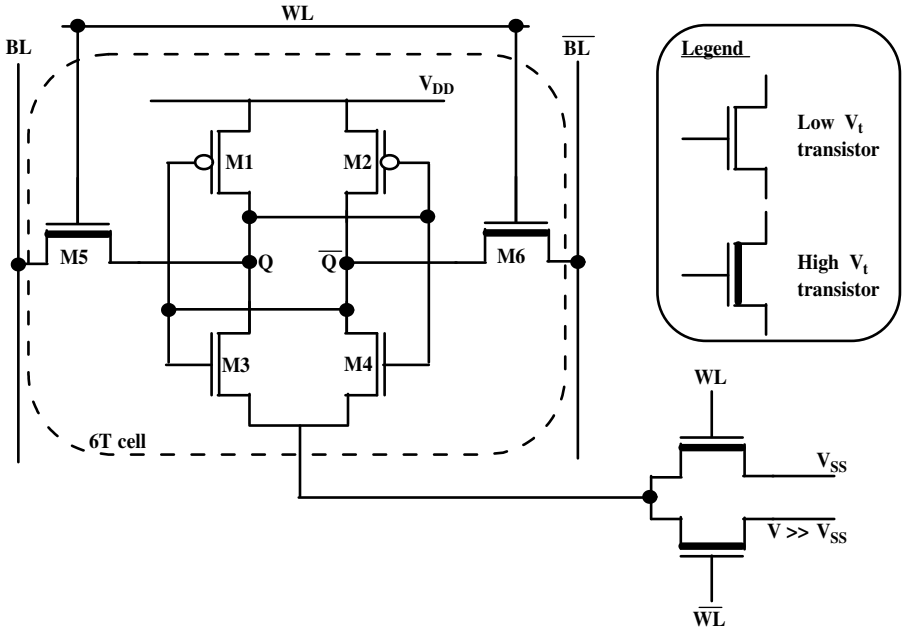

Figure 3.3: NC-SRAM Cell: Pass transistors control threshold voltages of the nMOS transistors in the cross-coupled inverter to reduce leakage power

The operating voltages of an array of memory cells are varied to switch between the active and stand-by modes thus reducing the leakage power significantly. Both the access transistors ($M5$, $M6$) are high-V_t devices to further reduce the bitline leakage. Each of the pass transistors used to control the source voltages of the nMOS transistors in the cross-coupled inverter is also a high-V_t device to

control the leakage current through these two pass transistors (from the positive control voltage v to ground). None of the nodes is left floating when the cell is not in use and this ensures the stability of the stored data with no additional complexity or circuitry. Since the capacitance of the ground supply lines is significantly less than that of the wells, this approach has improved transition time and energy as compared to [8] and [14]. Moreover, since the source voltage, as opposed to substrate voltage, is used to control the V_t of the nMOS transistors during the sleep mode, the inherent problems associated with body bias are fully eliminated. The leakage reduction is also significantly greater compared to other technologies since both cell current and voltage are reduced.

Figure 3.4 depicts the schematic of an NC-SRAM cache line. When the cache line is not in use, the source terminals of the nMOS transistors in the cross-coupled inverter can be switched to a positive voltage v through the pass-transistor. This increases the threshold voltages of the nMOS transistors dynamically and thus reduces the sub-threshold leakage in the dominant paths within the cell. The gate leakage of the nMOS transistors in the cross-coupled inverter is reduced significantly due to the increase in the source voltages of transistors, M3 and M4. In addition, as high-V_t devices are used for the access transistors that connect the memory's internal inverters to the read/write lines, the leakage through the bitlines is reduced considerably. The two pass transistors that control the threshold voltages of the nMOS transistors can be shared among multiple SRAM cells to reduce the area overhead. In our design, the wordlines from the row decoder logic are used to control the gates of these pass transistors. The cells are connected to the ground as in a conventional 6T-cell, only when the data is

being written to the row or when the row is being read.

To minimize the requirement of a larger gate capacitance associated with the pass transistors, we use divided wordline technique first proposed by Yoshimoto et al. in [21]. In this approach, a typically large row is partitioned into a number of smaller identical sub blocks and a pair of nMOS pass-transistors controls each of these blocks. The cells in the blocks that are being accessed operate in the normal mode, whereas all the other cells operate in the standby node with their sources connected to a positive voltage, thus reducing the leakage. Further, even within this block, only the memory cell being accessed should operate in the normal mode to obtain maximum leakage savings, All the other cells in the entire memory block should have their threshold voltages dynamically increased through the pass transistor and operate in a low-leakage mode.

3.1.2 Leakage Power: Analysis and Comparisons

In this section, we present experimental results on the leakage energy savings of the NC-SRAM design compared to the conventional 6T design and other existing techniques for leakage reduction in memory circuits. First, we discuss the method used in our circuit evaluation. Second, we present detailed circuit results corroborating the effectiveness of our design as compared to the available techniques. Third, we briefly discuss the impact of changing certain design parameters of NC-SRAM on the overall power and leakage power consumption. Finally, we present leakage power savings on our scaled down netlist for 100-*nm* and 70-*nm* technologies using Berkeley Predictive Technology Model (BPTM) [22].

Our evaluation circuit consists of a memory core with a fixed number of rows and columns (16 and 8). This circuit was designed using different memory cells,

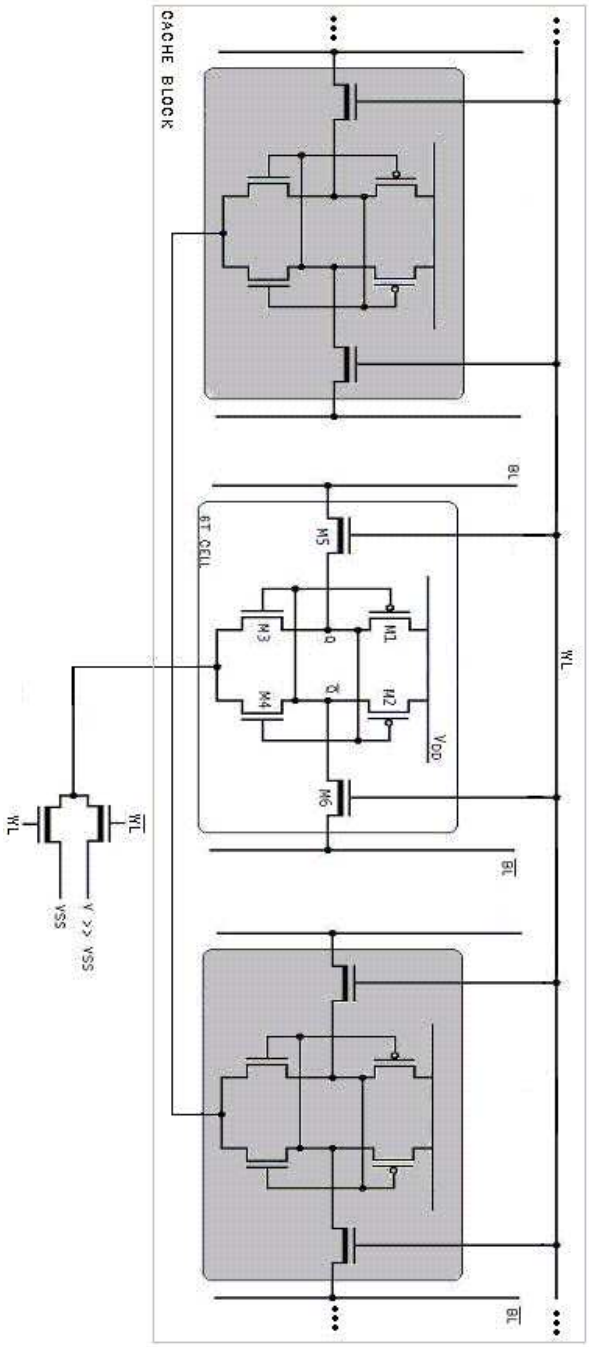

Figure 3.4: NC-SRAM Design: The two pass transistors are common to one cache block. A single row is made up of many such blocks

namely, Conventional 6T, NC-SRAM, High-V_t access transistors [8], Gated-V_{dd} [7] and DVS [15]. A fixed number of read and write operations were performed on all these circuits and the leakage power savings in all the above design techniques compared to that of a conventional 6T cell design were calculated. The results obtained are shown in Table 3.1.

Design Style	Shared By	Total Power (μW)	Leakage Power (μW)	Leakage Savings (%)
Conv. 6T	N/A	502	71.4	-
High Access V_t	N/A	495	67.6	5.32
Gated-V_{dd}	1 Row	493	54.8	23.25
NC-SRAM	1 Row	468	39.15	50.17
NC-SRAM	1 Block	340	0.53	99.25

Table 3.1: Leakage Energy Savings

The second column of Table 3.1 shows the number of memory cells sharing the pass-transistors (NC-SRAM) or the gated-ground transistor. The last two rows of Table 3.1 show that the NC-SRAM design has better leakage savings compared to the available techniques. In fact, when the pass-transistors are shared by only one block of memory cells, the number of transistors operating in the normal mode per access is reduced dramatically, thus almost eliminating the leakage power dissipation. Though similar results are expected for the DVS technique [15], the area overhead is high as pMOS transistors are used to gate the supply voltage, and leakage through the nMOS transistors are not reduced significantly.

Leakage power savings and total power consumed were analyzed for different control voltages of the pass-transistors. Figure 3.5 shows that as the control voltage is increased, the total power consumed also increases significantly with

the increase in leakage power savings (45%-70%). From the graph, we determined an optimal control voltage of $0.3V$ for the pass-transistors and that yields a leakage reduction of more than 50%.

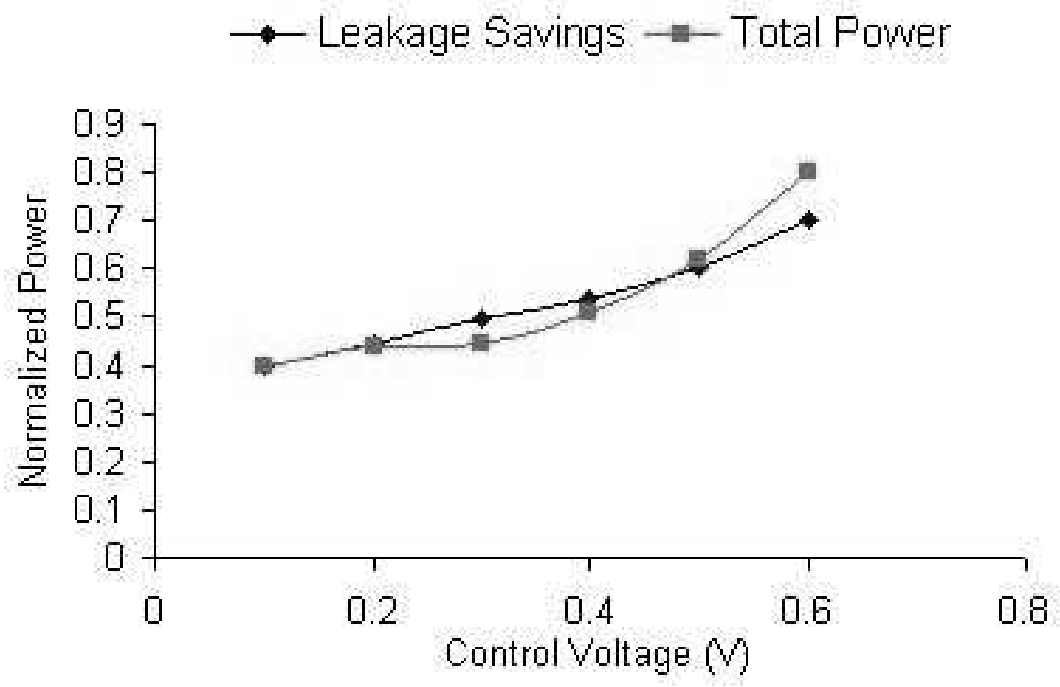

Figure 3.5: Power trends for different control voltages

Table 3.2 shows the leakage savings for NC-SRAM as compared to a conventional 6T design in $100nm$ and $70nm$ technologies. The V_{dd} used for these technologies are $1.2V$ and $1.0V$, respectively. As the results suggest, the percentage of leakage power increases significantly as we move towards the ultra deep submicron regime.

The results indicate that NC-SRAM almost eliminates leakage with the right type of partitioning and yielded a leakage reduction between 45%-70% depending on the control voltages used.

Technology	Total Power (μW)	Leakage Power (μW)	Leakage Savings (%)
$100nm$	187.9	8.64	77.34
$70nm$	159.7	13.4	55.77

Table 3.2: Impact of Technology Scaling on NC-SRAM

3.1.3 Gate Leakage Analysis

In addition to reducing sub-threshold leakage, the proposed technique is efficient in reducing gate leakage too. As mentioned earlier, the gate leakage across any transistor depends on the voltage difference across the gate-drain, gate-source and gate-body terminals. During the stand-by mode. a small positive voltage is given to the nMOS sources. This acts as an increased ground potential reducing the most of the voltage differences and hence reducing the gate leakage across all the transistors in the cross-coupled inverter pair. Though the additional pass-transistors introduce some leakage, it is neglibile compared to the leakage savings in the other components.

The NC-SRAM was simulated in $65nm$ predictive technology at room temperature for different values of t_{ox} and its gate leakage power was compared with that of a conventional SRAM [23]. Table 3.3 shows the different components of gate leakage at an oxide thickness of $1.7nm$ for the different transistors in the cross-coupled invereter pair. It also shows the additional leakage introduced due to the pass-transistors in the proposed design. As one can observe from the Table, the proposed NC-SRAM design reduces almost every leakage component present in the conventional SRAM cell.

Type	$M5$ (nA)	$M6$ (nA)	$M3$ (nA)	$M4$ (nA)	$M1$ (nA)	$M2$ (nA)	$P1$ (nA)	$P2$ (nA)	Total (nA)
Conv. SRAM	2.14	1.07	2.22	6.37	0.05	0.04	NA	NA	11.90
NC-SRAM	2.12	1.06	0.88	2.43	0.02	0.01	0.04	1.17	7.74

Table 3.3: Comparison of Gate Leakage Components in $65nm$ Technology (t_{ox}=$1.7nm$, V_{dd}=$0.8V$)

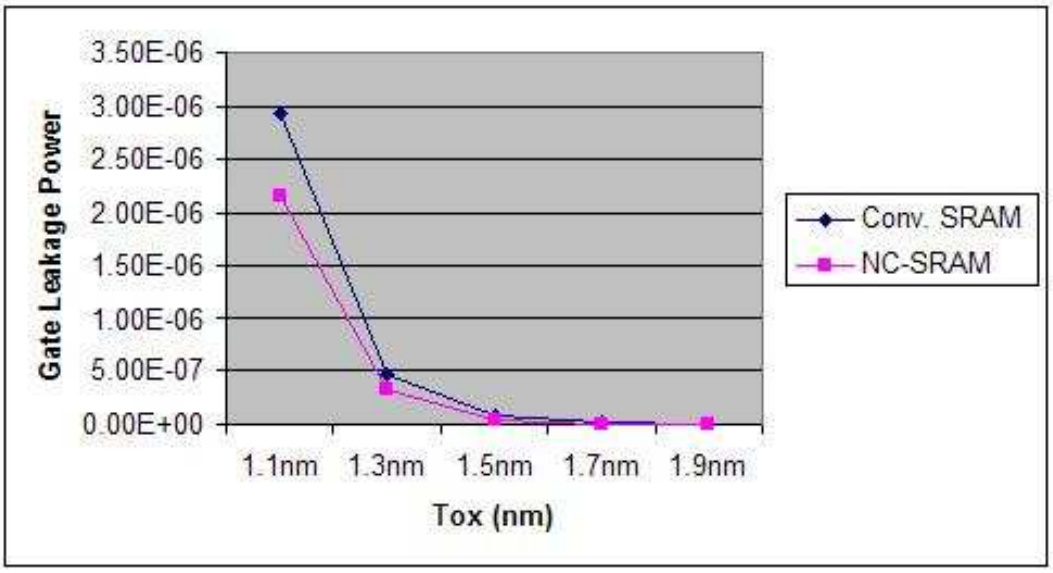

Figure 3.6: NC-SRAM Gate Leakage compared to conventional SRAM

In Table 3.3, $M5$ and $M6$ denote the access transistors connected to the bit-line and bitline-bar. The four transistors comprising the cross-coupled inverter pair are denoted by $M1 - M4$. In addition, the two additional pass-transistors in the proposed NC-SRAM circuit are denoted by $P1$ and $P2$ transistors. The total gate leakage power for varying oxide thickness in $65nm$ technology for both conventional SRAM and NC-SRAM designs is shown in Fig. 3.6. The total gate leakage power savings of NC-SRAM as compared to that of conventional SRAM is presented in Fig. 3.7. We can observe that NC-SRAM saves around 60% of gate

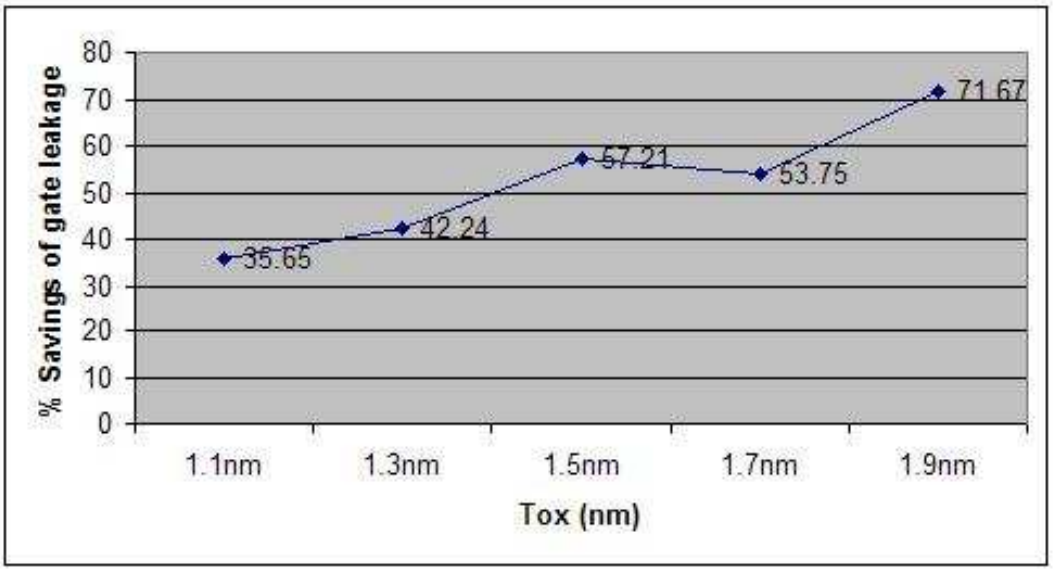

Figure 3.7: Gate Leakage power savings of NC-SRAM compared to conventional SRAM

leakage power as compared to that of conventional SRAM.

3.1.4 Static Noise Margin

Static noise margin is one of the important metrics to determine the stability of the cell design. The static noise margin (SNM) of an SRAM cell is defined as the minimum DC noise voltage necessary to flip the state of the the cell [24]. The SNM of both the conventional and NC-SRAM were computed through simulation. Figure 3.19 shows the superimposed static transfer charecteristics of two inverters in a single cell for both conventional and NC-SRAM designs in $65nm$ technology. The SNM is the noise voltage that corresponds to the maximum width of the enclosed square in the superimposed voltage transfer curves of V(Q) and V(Qbar). Under nominal conditions, the SNM of conventional and NC-SRAM was found to be $0.247V$ and $0.245V$, respectively. This result shows that the NC-SRAM design

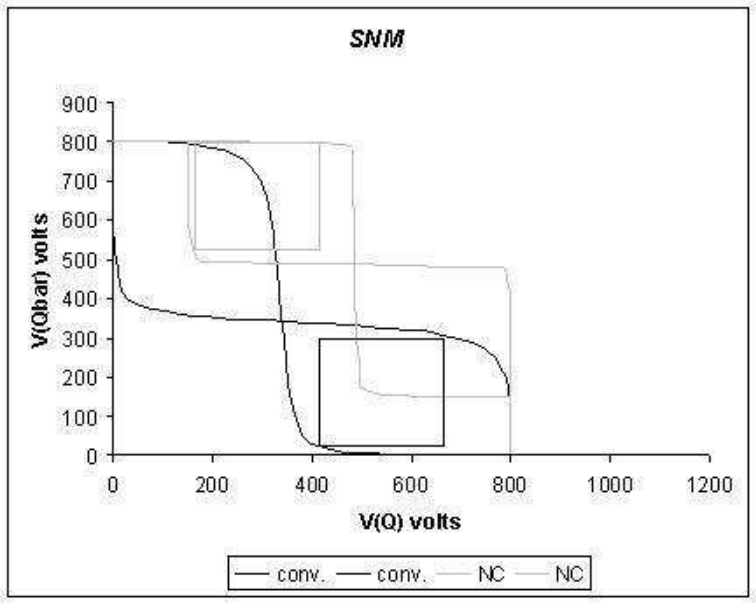

Figure 3.8: Static Noise Margin of NC-SRAM

does not suffer from static noise margin problem.

3.1.5 Read and Write Performance

Due to the presence of additional nMOS transistor between the cross-coupled inverter and ground in the NC-SRAM design, the read and write times are degraded slightly. However, this degradation, which ranges from $1\% - 3\%$ in a $65 - nm$ technology could be compensated and also improved by increasing the performance of the peripheral circuits. The proposed techniques for improving the performance of the address decoders and sense amplifiers are presented in the next section.

At the cell level, the total write time is determined by the flip time, which is the time taken for the memory cell to flip values between zero and one. The increase in cell flip times for NC-SRAM as compared to that of a conventional SRAM is shown in the Fig. 3.9. There is an increase of around 2% in a $65nm$

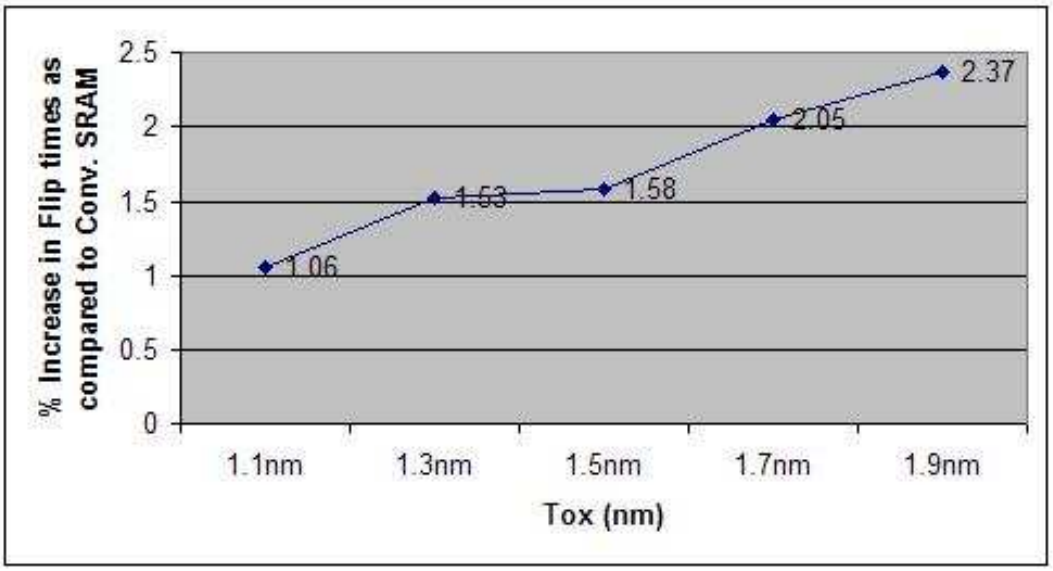

Figure 3.9: Increase in Cell Flip Times (write times) for NC-SRAM

technology with an oxide thickness of $1.7nm$. The read access time at the cell level is determined by the time taken for the bitlines to develop a potential difference of about $100mV$ between them. The read access times for varying oxide thickness in $65nm$ technology is presented as bitline discharge time in Fig. 3.10. We note that for higher oxide thickness, the increase in discharge time of NC-SRAM design as compared to that of conventional SRAM is minimal.

Thus, the proposed NC-SRAM design has increased sub-threshold and leakage savings as compared to other existing designs, with minimal effect on read and write performance. The static noise margin of the NC-SRAM design is also found to be similar to that of a conventional SRAM.

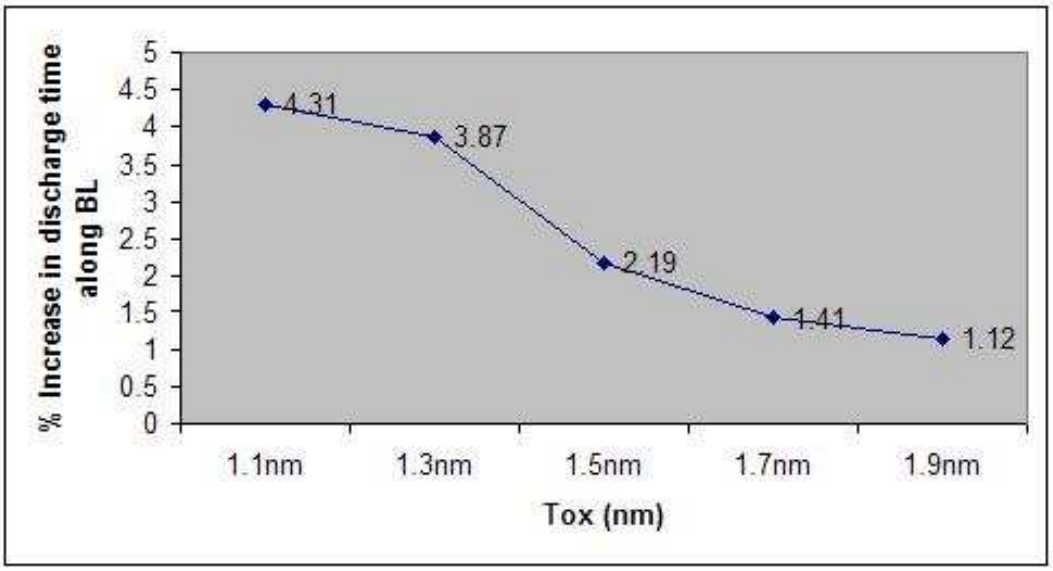

Figure 3.10: Increase in Discharge Time on the bitline (read time) for NC-SRAM

3.2 Gate Leakage - A Discussion

In 65*nm* BPTM [22], the gate tunneling leakage is modeled using voltage controlled current sources to account for the following components.

a. Gate to channel current (I_{gc}), part of which goes to source and the remaining goes to the drain (I_{gcs} and I_{gcd})

b. Edge Direct Tunneling (EDT) components between the gate and the Source Drain Extension (SDE) region (I_{gs} & I_{gd})

c. Gate to substrate leakage current (I_{gb})

Of these, gate to substrate leakage current can be neglected since it is many orders smaller than edge direct tunneling and channel currents [25]. EDT components depend on the terminal voltages regardless of the ON/OFF conditions of the MOS devices. Channel currents are the major source of gate tunneling when MOS devices are ON, whereas they cease to exist when they are turned OFF.

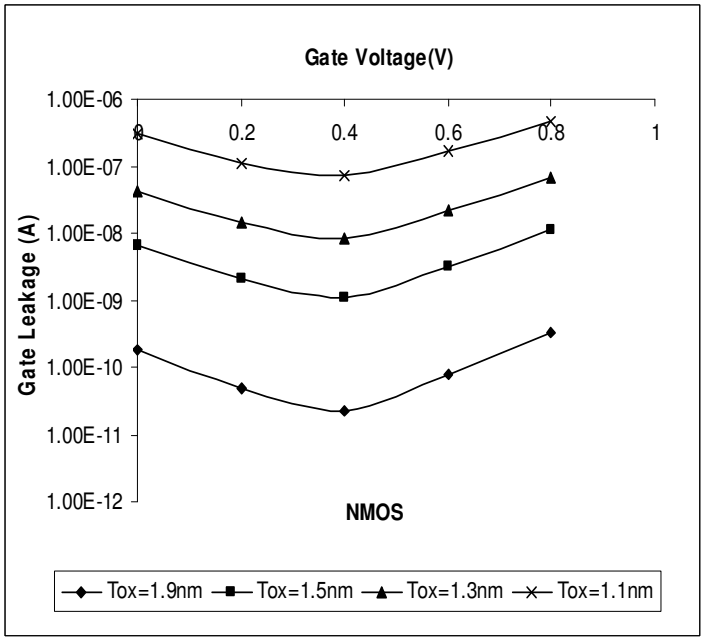

Figure 3.11: Dependence of Gate Leakage on Gate Voltage for NMOS

Figures 1 and 2 indicate the dependence of gate leakage on both gate voltage and T_{ox} in $65nm$ technology for NMOS and PMOS devices, respectively. In the case of an NMOS device, the drain and source voltages were held constant at V_{dd} and GND, respectively. The gate bias was then varied from 0 to V_{dd}. As the gate voltage is increased from 0 to V_{dd}, due to the decreasing V_{gd}, there is a reduction in I_{gd}, whereas, an increase in V_{gs} causes I_{gs} and I_{gc} to increase. In addition, gate leakage is increased by an order of magnitude for each 2% decrease in T_{ox}, indicating the exponential dependence of gate leakage on the oxide thickness. We can also observe from these figures that the gate leakage current of NMOS is 4-5 times greater than that of PMOS for the same oxide thickness [26]. This is due to the fact that the electron tunneling from conduction band (ECB) is the dominant gate leakage for an NMOS device, whereas, the hole tunneling from valence (HVB) band is the dominant one for a PMOS device.

In addition, the gate leakage current has a minimum value for a gate voltage of roughly $V_{dd}/2$. As we could observe, this minimum value occurs when the gate voltage is increased (from 0V) or decreased (from V_{dd}) by roughly 0.4V. This trend in gate leakage of these two devices is exploited in our DG-SRAM design.

3.3 RG-SRAM Cell Design

In this section, we present the RG-SRAM design, which exploits the dependence of gate voltage of both the NMOS and PMOS transistors on the total gate leakage power. As shown in Fig. 3.13, RG-SRAM has two additional PMOS pass transistors (GP1, GP2) connected between the cross-coupled inverters. GP1 is connected such that the output of inverter (L1, D1) drives the gate of GP1. One end of GP1

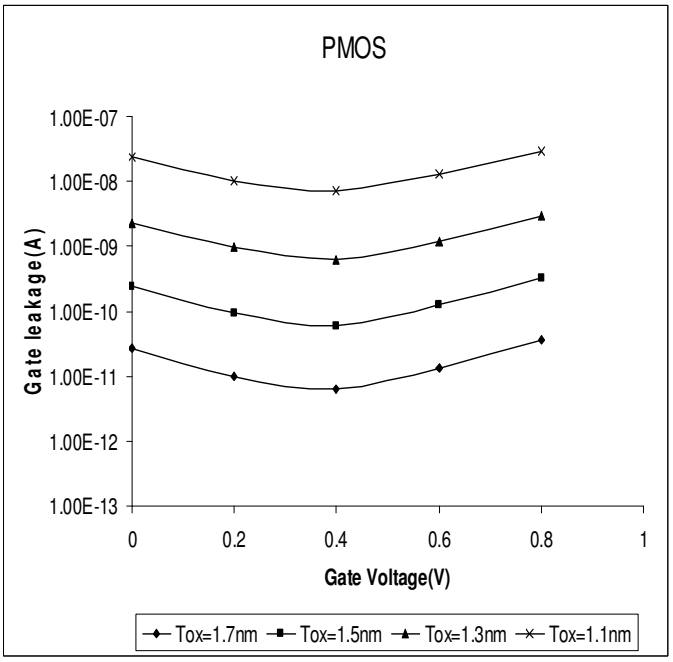

Figure 3.12: Dependence of Gate Leakage on Gate Voltage for PMOS

drives the input of the inverter (L2, D2) and the other end of GP1 is connected to ground. GP2 is connected between the inverters (L2, D2) and (L1, D1) in a similar fashion. The substrate terminals of these transistors are connected to V_{dd} to minimize the body effect. PMOS transistors do not pass a perfect '0' due to their intrinsic nature and this fact is used in varying the gate voltages of different SRAM components. The reduction in gate leakage of the SRAM during different cell states is explained below.

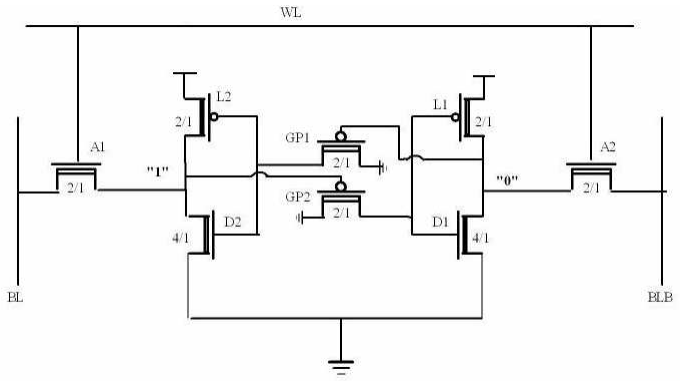

Figure 3.13: RG-SRAM cell

When the cell is storing a '1', GP1 is always turned ON and produces a voltage slightly higher than ground potential at the gates of L2 and D2. As mentioned earlier, GP1 does not pass a perfect zero due to the fact that a PMOS transistor intrinsically is a poor conductor of zero. This results in the increase of gate voltages of both L2 and D2 as compared to a conventional SRAM cell. A decrease in $|V_{gs}|$ and $|V_{gd}|$ of L2 reduces both EDT and direct tunneling components of this load transistor. Similarly, the increase in gate voltage of D2, i.e., a decrease in $|V_{gd}|$ and an increase in $|V_{gs}|$ reduces $|I_{gd}|$ but increases $|I_{gs}|$ and $|I_{gc}|$. However, as simulation results show, this increase in $|I_{gs}|$ and $|I_{gc}|$ components are negligible

compared to the significant reduction in the $|I_{gd}|$ component. GP1 dissipates negligible $|I_{gd}|$, since the drain and gate terminals of GP1 are at the same potential (ground), while it dissipates small amount of $|I_{gs}|$ and $|I_{gc}|$ due to the voltage difference in its gate and the source. However, this additional gate leakage through the GP1 transistor is much smaller as compared to the gate leakage savings achieved in the other components.

On the other hand, GP2 is switched OFF by the output of the inverter (L2, D2) when the cell is storing a '1'. However, the voltage at the input of the inverter (L1, D1) (i.e. output voltage of GP2) rise to a voltage slightly less than V_{dd}, due to the weak leakage currents through GP2. (bulk is connected to V_{dd}). This is similar to the increase in voltage at the node storing a '0' in stand-by mode of the DRG-cache due to the leakage currents through the gating transistor [27]. We need to note that all transistors are sized appropriately such that the trip voltages of the inverters is well below the source voltages of the pass transistors. In addition, as shown by the simulation results, the cell stores a value and also retains it irrespective of its previous state. Consequently, there is a decrease in the gate voltages of both L1 and D1 as compared to a conventional SRAM. A decrease in $|V_{gs}|$ and $|V_{gd}|$ of D1 reduces both EDT and the direct tunneling components associated with D1. Similarly, the increase in gate voltage of L1, i.e., a decrease in $|V_{gd}|$ and an increase in $|V_{gs}|$, reduce $|I_{gd}|$, and increases $|I_{gs}|$ and $|I_{gc}|$. However, as in the earlier case, this increases $|I_{gs}|$ and $|I_{gc}|$ components are negligible compared to the reduction in $|I_{gd}|$. GP2 dissipates negligible $|I_{gd}|$, since the source and gate terminals of GP1 does not differ by much, while it dissipates small amount of $|I_{gs}|$ and $|I_{gc}|$ due to the voltage difference in its gate and drain. In addition, the

gate tunneling current through a PMOS transistor is less compared to an NMOS transistor.

When the cell is storing a '0', the gate leakage in different transistors is reduced in the same way as above. As mentioned earlier, the leakage overhead introduced in the proposed design is much smaller than the gate leakage reduction achieved, as it reduces almost every leakage component present in the conventional SRAM cell. The width of the two NMOS drive transistors are doubled for improved stability and current drive. In addition, increasing the widths of the pass transistors also result in improved noise margin. However, all the other transistors are of minimum size. As illustrated in the simulation section, varying the sizes of the pass transistors (or the drive transistors), results in a tradeoff between cell immunity to noise and leakage savings.

3.4 DG-SRAM Cell Design

3.4.1 Circuit Description

In this section, we present the DG-SRAM design, which exploits the dependence of gate voltage of both the NMOS and PMOS transistors on the total gate leakage power. As shown in Fig. 3.14, DG-SRAM has two additional NMOS transistors (DC1, DC2) connected between the cross-coupled inverters in diode fashion (i.e gate and drain terminals tied together). DC1 is connected such that the output of inverter (P1, N1) drives the gate/drain of DC1. The source of DC1 drives the input of inverters (P2, N2). DC2 is connected between the inverters (P2, N2) and (P1, N1) in a similar fashion. The substrate terminals of these transistors are connected to ground to minimize the body effect. NMOS transistors do not pass a perfect '1'

due to their intrinsic nature and this fact is used in varying the gate voltages of different SRAM components. The reduction in gate leakage of the SRAM during different cell states is explained below.

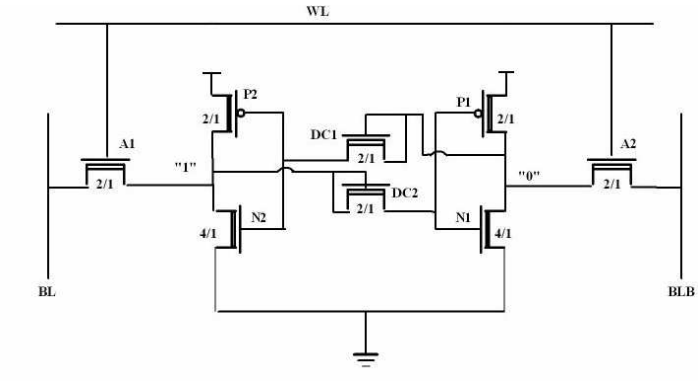

Figure 3.14: DG-SRAM cell

When the cell stores a '1', DC2 is always turned ON and produces a voltage slightly lower than V_{dd} at the gates of P1 and N1. As mentioned earlier, DC2 does not pass a perfect '1' due to the fact that a NMOS transistor intrinsically is a poor conductor of one. This results in the decrease of gate voltages of both N1 and P1 as compared to a conventional SRAM cell. A decrease in $\|V_{gs}\|$ and $\|V_{gd}\|$ of N1 reduces both EDT and direct tunneling components associated with N1. Similarly, the increase in gate voltage of P1, *i.e.*, a decrease in $\|V_{gd}\|$ and an increase in $\|V_{gs}\|$, reduce $\|I_{gd}\|$ but increases $\|I_{gs}\|$ and $\|I_{gc}\|$. However, this increases $\|I_{gs}\|$ and $\|I_{gc}\|$ components are negligible compared to the reduction in $\|I_{gd}\|$. DC2 dissipates negligible $\|I_{gd}\|$, since the drain and gate terminals of DC2 are tied together, while it dissipates small amount of $\|I_{gs}\|$ and $\|Igc\|$ due to the voltage difference in its gate and the source.

On the other hand, DC1 is switched OFF by the output of the inverter (N1, P1) when the cell stores a '1'. However, the voltage at the input of the inverter

(P2, N2) (i.e. output voltage of DC1) reduces to a voltage just above the ground potential due to the weak leakage currents from source to drain of DC1. This voltage depends on the size of DC1. This is similar to the increase in voltage at the node storing a '0' in stand-by mode of the DRG-cache due to the leakage currents through the gating transistor [27]. We need to note that all transistors are sized appropriately such that the trip voltages of the inverters is well above the voltage at the input of the inverter (P2, N2) while the cell stores a '1'. In addition, as shown by the simulation results, the cell stores a value and also retains it irrespective of its previous state. Consequently, there is an increase in the gate voltages of both N2 and P2 as compared to a conventional SRAM. A decrease in $\|V_{gs}\|$ and $\|V_{gd}\|$ of P2 reduces both EDT and the direct tunneling components of this load transistor. Similarly, the increase in gate voltage of N2, *i.e.*, a decrease in $\|V_{gd}\|$ and an increase in $\|V_{gs}\|$ reduces $\|I_{gd}\|$ and increases $\|I_{gs}\|$ and $\|I_{gc}\|$. However, as simulation results show, this increase in $\|I_{gs}\|$ and $\|I_{gc}\|$ components are negligible compared to the significant reduction in the $\|I_{gd}\|$ component. DC1 dissipates negligible $\|I_{gd}\|$, since the drain and gate terminals of DC1 are at the same potential (ground), while it dissipates small amount of $\|I_{gs}\|$ due to the voltage difference in its gate and source. However, this additional gate leakage through the DC1 transistor is much smaller as compared to the gate leakage savings achieved in the other components.

When the cell stores a '0', the gate leakage in different transistors is reduced in the same way as above. As mentioned earlier, the leakage overhead introduced in the proposed design is much smaller than the gate leakage reduction achieved, as it reduces almost every leakage component present in the conventional SRAM

cell.

We can also replace the diode connected NMOS transistors by PMOS transistors for increased savings in gate leakage. The gate leakage reduction process in PMOS-connected DG-SRAM remains the same except that when the cell stores a '1' DC1 is always turned ON and DC2 is switched OFF. The pros (increase in leakage savings) and cons (loss in read and write performance) of replacing DC1 and DC2 by PMOS transistors is illustrated in Section 6. In addition, increasing the widths of the diode connected transistors also results in improved noise margin. As illustrated in Section 6, varying the sizes of the diode connected transistors results in a tradeoff between cell immunity to noise and leakage savings.

3.4.2 Data Retention Capability Of DG-SRAM

Conventional SRAM cell stores the data as long as the power supply is ON. This is because the cell storage nodes at '0' and '1' are firmly strapped to power rails through conducting devices (by pull down NMOS in one inverter and a pull-up PMOS in the other inverter). Fig. 4 shows a single cell schematic of our DG-SRAM. When the cell stores a '1' the output of the inverter (P2, N2) turns OFF DC1. However, it also cut off the opportunity to strap the cell node at '1' firmly to V_{dd}. Node storing '0' remains firmly strapped to ground as long as input to the pull-down NMOS (N1) remains above the trip point of the inverter.

There is an issue with data retention if leakage/discharging currents through DC1 are not strong enough to bring its output from its previous state (near V_{dd}) to near ground during a write operation. The leakage current in MOSFET depends on various process parameters, terminal voltages and the quiescent state of the circuit. The BSIM model uses the following simplified leakage equations.

$$I_{leak} = Ae^{\frac{q}{nkT}(V_{GS}-V_{THO}-\gamma V_{SB}+\eta V_{DS})}(1-e^{-\frac{qV_{DS}}{kT}}) \tag{3.1}$$

where,

$$A = \mu_0 C_{ox}\frac{W}{L}(\frac{kT}{q})^2 \tag{3.2}$$

From the above equations, one could observe that leakage currents are directly proportional to the size/width of MOSFET. All the above leakage currents are discharging in the case of DC1 because of the ground potential at its gate and drain terminals. This enables us to size the diode connected transistors (DC1, DC2) appropriately with negligible leakage overhead to ensure that their output voltage is well below the trip point of the inverter when they are turned OFF. Alternatively, we could size the inverters forming the SRAM latch for a well balanced trip point. The optimized size for all the transistors in DG-SRAM is shown in Fig. 5.

3.4.3 Gate Leakage Components - Comparative Analysis

The different gate leakage components of all the transistors for both the conventional SRAM and DG-SRAM cells are presented in Table 3.4. In this Table, I_{gc} is the gate to channel current and is determined by Electron tunneling from Conduction Band (ECB) for NMOS transistors and Hole tunneling from Valence Band (HVB) for PMOS transistors. I_{gs} represents the gate tunneling current between the gate and the source diffusion region, while I_{gd} represents the gate tunneling current between the gate and the drain diffusion region. I_{gs} and I_{gd} are determined by ECB for NMOS transistors and HVB for PMOS transistors.

The gate leakage through any transistor depends on the voltage difference across gate-drain, gate-source and gate-body terminals. The addition of the two

pass transistors in DG-SRAM design reduces most of the above voltage differences and hence results in overall reduced gate leakage across transistors N1, N2, A1 and A2, as shown in Table 3.4. The savings in gate leakage is also quite significant compared to the increase in the gate leakage of some components and that of the additional transistors. We could also observe that the major savings is from one of the driver transistors, whose gate potential is reduced by the additional NMOS pass transistor that is turned ON. The exponential relationship between the gate voltage and the tunneling currents can also be observed from the simulation results.

Table 3.4: Gate Leakage Components of Conventional SRAM & DG-SRAM (All currents in *nA*)

Design	I_{gc}	I_{gs}	I_{gd}	I_{gc}	I_{gs}	I_{gd}
	N1			N2		
Conv. SRAM	18.94	9.49	9.49	0.0	0.0	13.6
DG-SRAM	0.286	0.55	0.327	2.16	1.31	1.28
	P1			P2		
Conv. SRAM	0.0	0.0	0.316	0.0	0.239	0.239
DG-SRAM	0.0	0.041	0.013	0.0	0.011	0.062
	A2			A1		
Conv. SRAM	0.0	0.0	6.57	0.0	6.57	6.55
DG-SRAM	0.0	6.57	2.73	0.0	6.57	0.0001
	DC1			DC2		
Conv. SRAM	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
DG-SRAM	0.0	0.25	0.0	0.002	0.01	0.0

3.5 Simulation Results

3.5.1 RG-SRAM

In this section, we present detailed Spectre simulation results for RG-SRAM in BPTM 65nm technology at 27^{o}C. The total leakage savings of RG-SRAM, which considers both the sub-threshold and gate leakage, is summarized in Table 3.5. At high values of T_{ox}, sub-threshold leakage dominates and there is no discernible decrease in total leakage. The pronounced effect of sub-threshold leakage at higher T_{ox} is further illustrated by the fact that 62.6% reduction in gate leakage at 1.9nm is reflected as mere 2.9% reduction in total leakage. As T_{ox} scales down, the reduction in gate leakage is reflected more on total leakage of the cell. As T_{ox} is lowered to 1.1nm, the total leakage is just 45.4% of the conventional cell total leakage.

Table 3.5: Simulation Results of RG-SRAM

T_{ox} (nm)	Leakage savings $(\%)$	Change in Discharge Time $(\%)$	Change in Bit-Flip Time $(\%)$
1.9	2.9	3.1	27.1
1.7	24.5	2.8	26.8
1.5	47.1	2.2	27.9
1.3	50.9	1.6	28.1
1.1	54.6	1.3	28.3

The read access time at the cell level is determined by the time taken for the bitlines to develop a potential difference of 100mV. When the cell is storing a '0', the bitline discharge along BL takes longer due to D2's low conductance. The discharge time along BL, which is only a fraction of the total read access time,

is only 3.1% longer than when the T_{ox} is 1.9nm. The time taken for the memory cell to flip values between zero and one, in the worst case, produced a 28.3% compared to the conventional SRAM. This is due to the increased delay through additional PMOS transistors to turn ON the respective transistors in the cross-coupled inverter pair.

3.5.2 DG-SRAM

In this section, we present detailed Spectre simulation results for DG-SRAM in BPTM 65nm technology at 27^{o}C.

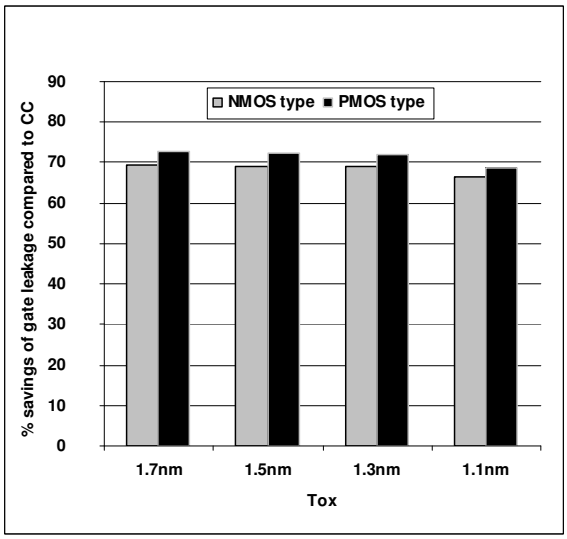

Figure 3.15: Gate Lekage Savings Compared to CC

Gate Leakage

Figure 3.15 shows the percentage savings in gate leakage of DG-SRAM as compared to a conventional cell at different values of T_{ox}, 1.7nm-1.1nm for both NMOS type (DC1,DC2 are NMOS) and PMOS type (DC1,DC2 are PMOS) DG-SRAM. It is

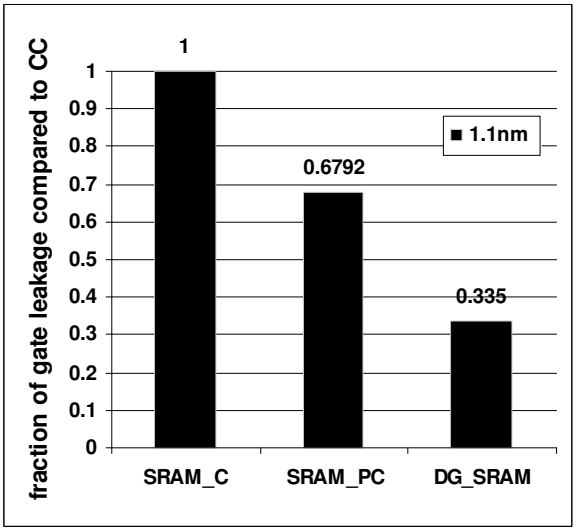

Figure 3.16: Comparison between DG, PC, CC

clear that maximum gate leakage savings of 68.8 % (69.8 %) is obtained for NMOS (PMOS) type DG-SRAM at 1.3nm. The increase in gate leakage savings for PMOS type DG-SRAM is due to the fact that NMOS transistors exhibit 4-5 times more gate leakage than its counterpart PMOS. DG-SRAM design also has better gate leakage savings than PC-SRAM [28], as seen in Fig. 3.16.

Total Leakage Benefits

The total leakage savings of DG-SRAM, which considers both the sub-threshold and gate leakage, is summarized in Table 3.6. Band to Band tunneling and Gate Induced Drain Leakage (GIDL) are neglected since they form only negligible portion of the drain current [16].

At high values of T_{ox}, sub-threshold leakage dominates and there is no discernible decrease in total leakage. The pronounced effect of sub-threshold leakage at higher T_{ox} is further illustrated by the fact that 69.4% reduction in gate

leakage at 1.7nm is reflected as mere 30.2 % reduction in total leakage for NMOS type DG-SRAM. As T_{ox} scales down, the reduction in gate leakage is reflected more on total leakage of the cell. As T_{ox} is lowered to 1.1nm, the total leakage for NMOS type DG-SRAM is just 33.5% of the conventional cell total leakage.

Table 3.6: Total Leakage Savings of DG-SRAM

T_{ox} (nm)	Leakage Savings $(\%)$ (NMOS type)	Leakage Savings $(\%)$ (PMOS type)
1.7	30.2	31.6
1.5	64.6	67.5
1.3	68.8	69.8
1.1	66.5	68.6

Read Performance

The read access time at the cell level is determined by the time taken for the bit-lines to develop a potential difference of 100mV. The percentage increase in bitline discharge time as compared to a conventional cell, for varying oxide thickness in 65nm technology is presented in Fig. 3.17. When the cell is storing a '0', the bit-line discharge along BL takes longer due to N2's low conductance. The discharge time along BL, which is only a fraction of the total read access time, is only 2.86% longer than conventional SRAM for PMOS type DG-SRAM when the T_{ox} is 1.7nm. The same results hold good for increase in BLB when the cell stores a '1', since the cell is symmetrical. The discharge time of PMOS type DG-SRAM is significantly larger than NMOS type DG-SRAM due to the intrinsic speed of NMOS devices.

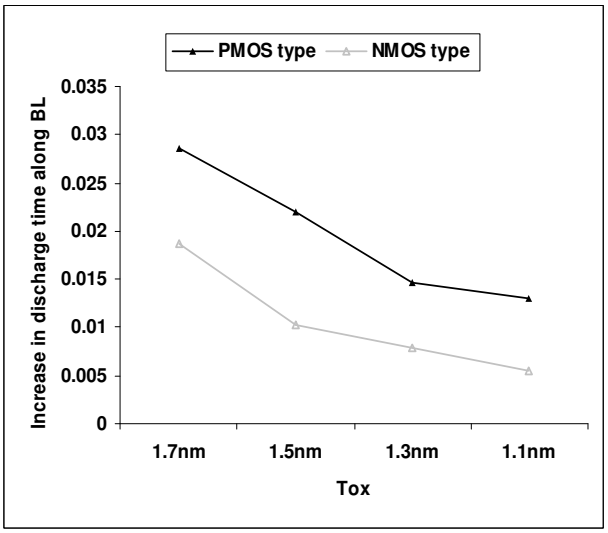

Figure 3.17: Increase in Discharge time along BL compared to CC

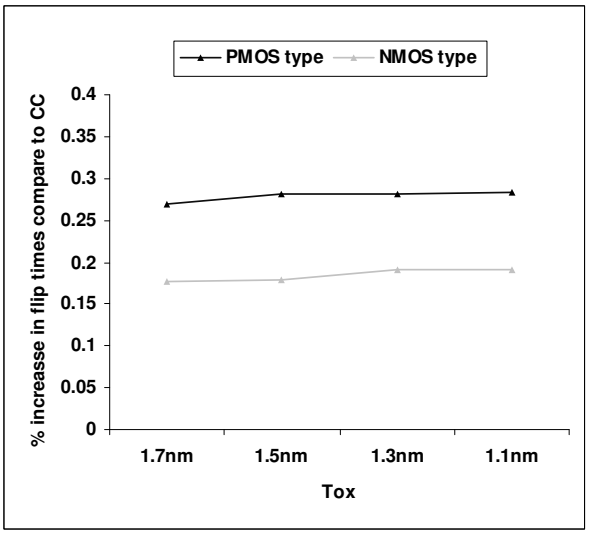

Figure 3.18: Increase in flip times compared to CC

Write Performance

To evaluate the write performance, we calculated the percentage increase in bit-flip times, the time taken for the memory cell to flip values between zero and one, and the results are shown in Fig. 3.18. In the worst case, PMOS type DG-SRAM produced a 28.3% increase in flip times as compared to the conventional SRAM. This is due to the increased delay through additional PMOS transistors to turn ON the respective transistors in the cross-coupled inverter pair. However, since flip time is a very small portion of the total write time, 28.3% increase does not reflect entirely on the write delay. A 28.3% increase is only a $16ps$ increase in the total write access time (about 2% increase). The write performance of NMOS type DG-SRAM is better than its counterpart PMOS type. This again is due to the intrinsic speed of NMOS transistors. Thus, from the above illustrations, we could observe a performance/leakage savings trade-off between NMOS-connected and PMOS-connected DG-SRAM.

3.5.3 Static Noise Margin

We analyzed the static noise immunity of both DG-SRAM and RG-SRAM using the approach presented in [29]. The static noise margin of CMOS SRAM cell is defined as the minimum DC noise voltage necessary to flip the state of the cell [30]. The conventional and NMOS type DG-SRAM static transfer characteristics during stand-by mode for 65nm technology are shown in Fig. 3.19. The Static Noise Margin (SNM) equals the noise voltage necessary at each of the cell storage nodes to shift the static characteristics of the two cell inverter vertically or horizontally along the side of the maximum nested square so that they intersect

at only one point. From Fig. 3.19, we observe that the SNM of DG-SRAM cell is around 55% of conventional cell. This is because of the node driving one of the inverters is strapped to a voltage slightly above ground potential by the weak leakage currents. However, the inverters are designed for a trip point less than the voltage available at the output of the OFF transistor (DC1 when cell stores '1') and thus the data is always retained.

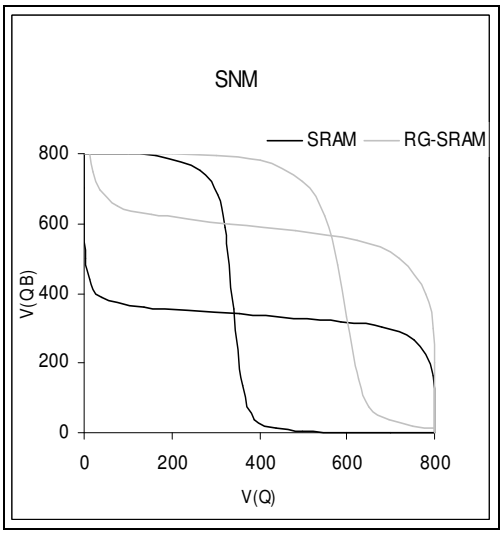

Figure 3.19: SNM analysis of DG-SRAM

The SNM can also be improved by increasing the width of the devices DC1(GP1), DC2(GP2). This increase in width enhances the weak leakage currents that strap the output voltage near ground potential when the transistor is OFF for a NMOS type DG-SRAM. It can be seen from Table 3.7 and Fig. 3.20, for NMOS type DG-SRAM that there is a 50% increase in SNM and 3-5% reduction in leakage savings when the additional NMOS transistor widths are increased by three times. Similarly increase in GP size increases the SNM of RG-SRAM with negligible decrease in leakage savings as shown in Table 3.8.

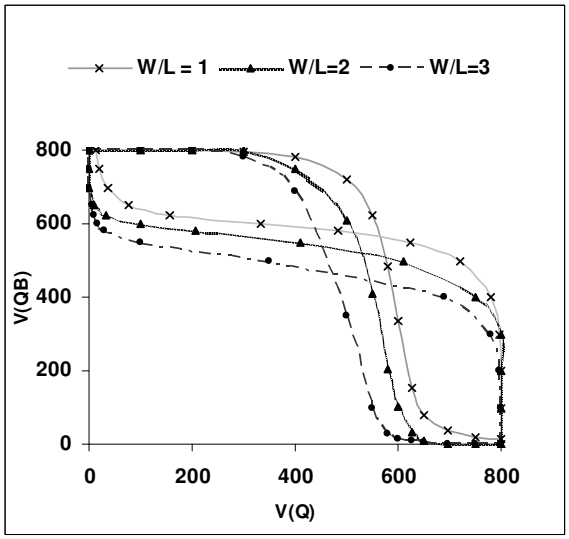

Figure 3.20: SNM variation with DC size

Table 3.7: DG-SRAM: Impact of DC size on Leakage and SNM

DC Size	Leakage Savings (NMOS type) $(\%)$	Conv-Sram (SNM) (mV)	DG-Sram (SNM) (mV)
1	66.5	210	115
2	65.8	210	135
3	64.5	210	150

3.6 Summary

This chapter explored three different circuit-level techniques for reducing the leakage power in deep submicron caches. Comparison of leakage power savings in other contemporary cache designs with the proposed designs was performed. Based on the simulation results, we found that the proposed memory designs has key advantages over the other existing techniques for reducing leakage in SRAM circuits. One of the notable features of this work is that the our design achieves large leakage power savings and at the same time retains the data stored in the

Table 3.8: RG-SRAM: Impact of GP size on Leakage and SNM

GP Size	Leakage Savings (%)	Conv-Sram (SNM) (mV)	RG-Sram (SNM) (mV)
1	54.6	210	110
3	53.2	210	132
5	51.5	210	160

memory cell with no additional overheads. We evaluated and presented simulation results from implementing the design in different technologies using a dual-V_t approach. The results indicated that NC-SRAM almost eliminates leakage with the right type of partitioning and yielded a leakage reduction between $45\% - 70\%$ depending on the control voltages used. We also simulated the proposed design in $100nm$ and $70nm$ technologies to study the impact of technology scaling and achieved promising results in terms of leakage power savings.

In addition, we presented two designs, RG-SRAM and DG-SRAM, for suppressing the gate leakage current. In order to reduce the gate leakage current, we used two additional PMOS/NMOS devices which change the gate voltages of the transistors forming the inverter latch in SRAM. One of the notable features of the proposed work is that it achieves significant leakage savings irrespective of the state of the cell and the value stored in the cell. Simulation results show 66.5% reduction in total leakage at 65nm technology with Tox at $1.1nm$ with only around 2.86% degradation in discharge time for NMOS type DG-SRAM.

Chapter 4

Robust and High Speed Peripheral Circuits

4.1 Sense Amplifiers

Sense amplifier is one of the key peripheral circuits in the memory system as it significantly influences the memory access times. It retrieves the stored data from the memory array by amplifying the small differential signal on the bitlines. Figure 4.1 shows the critical path of the memory syste m during read operation. Over 60% of the delay during the read operation is attributed to the SRAM column capacitance, column multiplexer and sense amplifier. Consequently, any innovation in the memory critical path will considerably improve the system performance.

One of the key challenges that limits the performance of the sense amplifiers as we scale down into the nanometer domain is the increasing bitline capacitances [31]. Increased bitline capacitance results in increased time to develop the differential bitline voltage and limits the efficiency of the traditional voltage mode

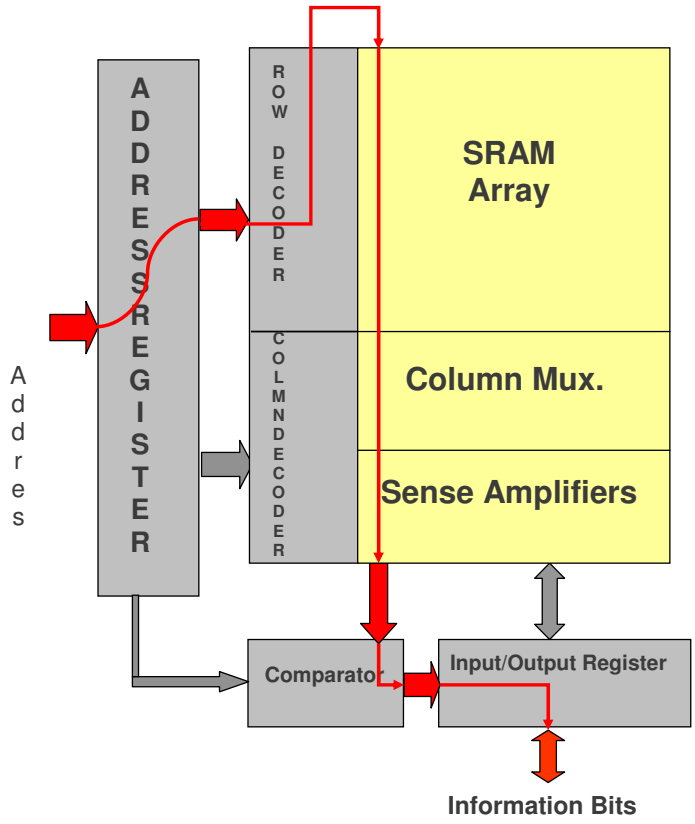

Figure 4.1: SRAM Critical Path

sensing schemes. Apart from the increased bitline capacitance, the time to develop differential bitline voltage is also increased due to the minimum sized transistors being used in the memory cell design of high density nanoscale SRAMs. An alternative strategy is to sense the current difference in the accessed cell through the bitlines. Current sensing [32, 33, 34] does not depend on differential discharging of the heavily loaded bitlines and hence provides considerable speed improvement.

Another factor that limits the performance in memory and microprocessor design is the systematic and random variations in process, supply voltage and temperature (P, V, T) [35, 36]. Technology scaling below $100nm$ results in higher levels

of device parameter variations, such as, variations in threshold voltages and effective channel lengths, as these change the design problem from deterministic to probabilistic [37]. The impact of these process variations are more pronounced on sense amplifiers, as they are designed to be electrically balanced and symmetric circuits [38] and any small variation in the device parameters would adversely affect the circuit functionality and performance. In addition, the extent of variations, specifically variations in effective channel lengths, are gaining significance due to the use of near minimum transistor sizes in nanometer memory cell designs. This alters the value of bitline differential signal supplied to the sense amplifier and in the worst case, may even result in a read failure.

In this Chapter, we present two novel, process variation tolerant, high performance, low power, current-mode sense amplifiers that provide reliable sensing in the nanometer domain. We focus on diminishing the bitline swing to reduce both the delay and the energy during charging and discharging of the bitlines, and on developing a novel amplifier topology.

The first sense amplifier [39] uses a Winner-Take-All approach (WTA) [40] and has a current sensing stage followed by an amplification stage. We use a four transistor current conveyor circuit, in the sensing stage to sense the differential cell current while maintaining a virtual short across the bitlines [32]. This helps in reducing the sensing delay, power consumption and the input impedance. This differential current is then amplified by a novel winner take all circuit explained in Section 3. The output nodes would be pulled high and low by this amplifier depending on who the winner is between BL and $\overline{BL}$ with respect to the magnitude of currents. This circuit offers high immunity to process variations during

worst case analyses by keeping the sensitive amplification stage simple with less number of transistors.

We also present a novel low power current-mode sense amplifier (LPCSA) [41], which has a similar current sensing stage followed by a cross coupled amplification stage. Similar to the previous amplifier we use a four transistor current conveyor circuit, in the sensing stage to sense the differential cell current. This differential current is then amplified by a high-gain positive feedback cross coupled inverter pair. We use a novel reset scheme during the pre-sensing phase that results in considerable power savings as explained in the simulation results section. The low input and output impedance of the sense amplifier greatly reduces the charging and discharging time of the bitline capacitances.

We compare our WTA and LPCSA designs with a standard voltage mode sense amplifier (CCIL) and two popular current mode sense amplifiers: the clamped bitline sense amplifier (CBLSA) [33] and the Izumikawa sense amplifier (ICSA) [42] for variation tolerance, speed and power consumption. We also analyze the impact of worst case threshold voltage mismatch [29], effective channel length and supply voltage variations [34] on the functioning of all these sense amplifiers. The simulations were performed in $70nm$ predictive technology and the results are presented.

4.2 Previous Work and their Limitations

In general, sense amplifiers have two stages of operation: the sensing stage and the amplification stage. Majority of the existing sense amplifiers utilize a cross coupled transistor topology for amplification and differ primarily in the type of

signal sensed and their sensing circuits. This section describes the working of a few key voltage and current mode sense amplifier circuits and their limitations.

4.2.1 Cross-Coupled Inverter Latch (CCIL)

This is one of the most commonly used sense amplifier circuit and has two cross coupled inverters with very high gain to provide fast amplification [43]. The bitlines are directly terminated at the sense amplifier outputs through the read enable transistors. When sufficient voltage difference in the bitlines develops, the sense amplifier is turned on and the amplifier latches onto the value stored in the memory. The main drawback is that the working of the amplifier depends on the timely discharge of the bitline capacitances to sense the differential voltage. As technology scales down and the number of memory cells per column increases, the time to develop the differential voltage in the bitlines increases significantly. This results in a considerable increase in the sensing time irrespective of how fast the amplification process may be. To overcome this limitation, current sensing techniques [34, 44] that is independent of the bitline capacitance have been proposed. In addition, to obtain significant energy savings during the read operation, we must minimize the bitline swings as much as possible.

4.2.2 Clamped Bitline Sense Amplifier (CBLSA)

The clamped bitline sense amplifier, one of the first current sense amplifier circuits, was proposed by Blalock and Jaeger [33]. The bitlines are terminated in a low impedance node which is isolated from the sense amplifier output. The current difference through the bitlines flows through the low impedance nodes

and then translates to a small differential voltage between the outputs of the inverter pair. The cross coupled latch amplifies this difference to full rail levels. The main limitation of this circuit is that the bitlines are pulled down considerably from their precharge state through the low impedance NMOS termination. This results in significant amount of energy consumption in charging and discharging the highly capacitive bitlines. Also, the presence of two NMOS transistors in series with the cross coupled amplifier results in an increase in the speed of amplification.

A modified clamped current sense amplifier (MDCSA) was proposed recently by Sinha *et.al* [44]. Four equally sized PMOS transistors act as a current transporter with unity gain. The input and output nodes of the sense amplifier are separated to prevent any coupling. The cell current is directly used as a signal and is detected by the sense amplifier. Though there is a speed improvement as compared to the traditional voltage mode sense amplifiers, it has considerable static power consumption.

4.2.3 Izumikawa Current SA (ICSA)

The current sense amplifier proposed by Izumikawa et al [42] utilizes a combination of two PMOS and two NMOS transistor current conveyor circuit to sense the current difference. Two PMOS transistors that form the current conveyor are also a part of the cross coupled inverter amplifier while the other two NMOS transistors predischarge the outputs of the sense amplifier to ground. Hence, this maintains the two PMOS transistors in ON state, through which the bitline differential current flows. After the sense amplifier is enabled, the cross coupled latch quickly amplifies the difference. Although the circuit provides considerable power

savings over CBLSA, it predischarges the output nodes to ground which results in static power dissipation during the sensing phase. In addition, the current conveyor circuit used in the sense amplifier is a modification of that proposed by Seevinck et. al [32]. The modified conveyor circuit fails to maintain a virtual short across the bitlines and hence the sense amplifier consumes more energy per read operation with increasing bitline capacitance.

As most of the existing current sense amplifiers are similar to or minor variations of the above the three sense amplifiers, we restrict our performance comparisons to these three techniques mentioned above. Earlier current sensing techniques [45] [46] utilize similar current conveyor circuits and exhibit performance independent of the bitline capacitance. However, the sensing and amplification schemes involve complicated circuitry making them extremley sensitive to process variations. Hence, the advantages offered by these circuits are negated in the nanometer regime. There have also been other recent sense amplifier circuits [34, 44] but only provide marginal improvements in performance over the three established techniques and at the cost of degraded performance in the presence of process variations.

4.3 WTA Current Sense Amplifier

There are two stages in the proposed Winner Take All (WTA) current sense amplifier: the current sensing stage consisting of a unity gain current conveyor [32] and a novel amplification stage that amplifies the current difference. The circuit realization of the proposed current sense amplifier circuit is shown in Fig. 4.2.

Transistors $P1-P4$ forms the current conveyor circuit that senses the differential current on the bitlines. Transistors $N1-N5$ constitute the amplification stage followed by two inverters ($P5$, $N6$ and $P6$, $N7$) connected to the outputs of the amplifier. The two inverters drive a $10fF$ load and ensure a rail to rail output.

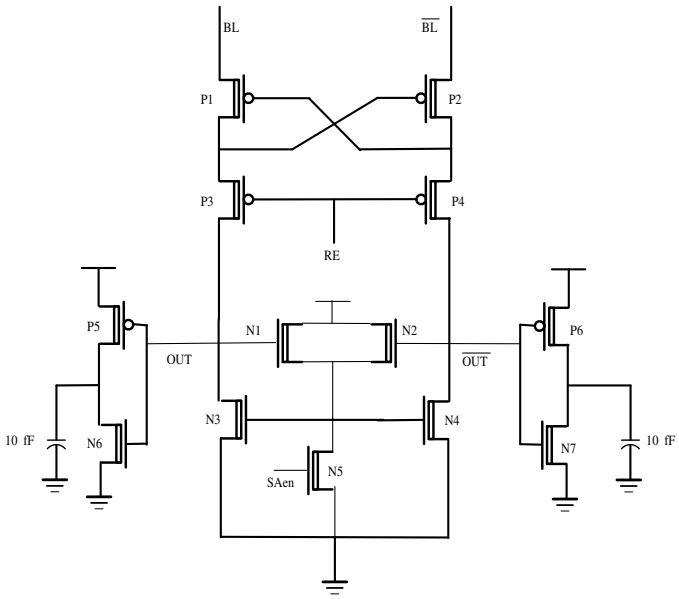

Figure 4.2: WTA Sense Amplifier (Note: all transistors are normal MOSFETs)

The circuit functions as follows: The previous read state is reset by pulling the *SAen* signal low. The four transistor current conveyor circuit is then turned *ON* to sense the differential current. The current conveyor acts as a virtual short across the bitlines and thus reduces the influence of increasing bitline capacitance on the sense amplifier. The amplifier proceeds to amplify the current difference. To illustrate the circuit behavior, consider the case of reading a 1 from the memory cell. This means that more current flows from *BL* to device $N3$ than to device $N4$ from $\overline{BL}$. Assuming matched transistors $N3$ and $N4$, which have identical gate

source overdrive, the voltage at the drain of $N4$ reduces and is less than the voltage at the drain of device $N3$. The differential current from the bitlines is now effectively translated to a differential voltage at the drains of $N3$ and $N4$. Consequently, device $N1$ now has a greater gate source overdrive than $N2$ and draws more current from $N5$ than $N2$. This increases the voltage difference between the drains of $N3$ and $N4$ further, and the device $N4$ begins to move out of saturation. The amplifier then gets into a feedback loop and amplifies the outputs to CMOS levels.

The four transistor current conveyor ensures that there is no differential discharging of the heavily loaded bitlines. Hence, the sensing delay is almost independent of the bitline capacitances and consequently the number of rows in the memory array. Moreover, the current conveyor isolates the outputs of the sense amplifier from the bitlines, preventing the amplifier's outputs from affecting the bitline voltages. The bitline differential current flows through devices $N3$ and $N4$, which are maintained in the *ON* state by the tail transistor $N5$, and offer a low impedance to the input differential current. Transistors $N3$ and $N4$ are a part of the amplification circuitry, while also providing a low impedance termination to the bitlines. Whereas, other current sensing techniques [33, 44] have additional circuitry in the amplification stage to provide a low impedance termination. This feature of the WTA circuit offers significant improvement in sensing speed and robustness to process variations over other existing current sense amplifiers. The speed of amplification can be controlled to a certain extent by the bias provided to the tail transistor $N5$. The tail transistor $N5$ acts as a constant current source and also maintains a constant voltage at the gates of devices $N3$ and $N4$. This voltage

should be significantly large to turn devices $N3$ and $N4$ ON and maintain device $N5$ in saturation, while being sufficiently low to provide adequate gate source overdrive to $N1$ and $N2$. Through simulations we determined the optimal bias voltage for the tail transistor to be $450mV$. This infact serves to reduce the sense amplifier offset as observed in [47], as the gates source overdrive of N1 and N2 reduces.

The important feature of the proposed circuit that offers significant improvement in sensing speed (around 75%) is at the beginning of the amplification process. When the bitline differential current is presented to devices $N3$ and $N4$, it effects a differential voltage at the drains of these devices. The rate at which the drain voltage drops initially is a function of the channel length modulation. The saturation current equation for device $N4$ can be given as

$$I = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{gs} - V_{th})^2(1 + \lambda V_{ds}) \tag{4.1}$$

or,

$$I = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{gs} - V_{th})^2(1 + \frac{\Delta L}{L}) \tag{4.2}$$

Hence, for a transistor in saturation with a constant gate source overdrive, a change δI in current causes a corresponding change in the drain voltage V_d. As we can see from Eqn 1, the magnitude of this change in V_d depends on the extent of channel length modulation for that particular technology, as V_d is directly proportional to $\Delta L/L$. As we scale down to the nanometer domain, the $\Delta L/L$ factor increases, thereby effecting a larger change in V_d for the same change in δI. Consequently, when reading a 1 (or a 0), scaling down causes the drain voltage of N4 (or N3) to drop rapidly. technology scaling causes the drain voltage of $N4$ to drop

rapidly. This provides faster amplification by immediately driving device $N4$ out of saturation resulting in considerable improvement in the access speed. When combined with the current conveyor, the amplifier achieves significant improvement in sensing speed and can be expected to offer excellent performance when we scale down further, as channel length modulation, which aids the amplification process, increases with technology scaling.

4.4 Low Power Current Sense Amplifier (LPCSA)

Our LPCSA current sense amplifier design consists of two stages: a unity gain current conveyor circuit that senses the current difference and a cross coupled inverter latch that amplifies the current difference. Figure 4.3 shows the circuit realization of the new current sense amplifier. The four transistor current transporting circuit (P1 - P4) [32], with positive feedback senses the differential current while maintaining a virtual short across the bitlines and the cross coupled inverter pair (P5, P6, N1, N2) amplifies the difference. Two CMOS inverters (P7, N3 and P8, N4) that drive a $10fF$ load are connected to the output of the cross coupled inverter pair to produce a rail to rail signal at the output. Though this load might be a little lower than what one might in see in current technologies, it provides a common platform for comparison against different sense amplifier designs.

The operation of the circuit is as follows. Before the start of the read operation, the sense amplifier outputs contain the pervious read state and hence needs to be reset. So, the enable transistor (N5) is turned ON through the SAen signal to bring the voltages at the output nodes closer to each other. When the Read

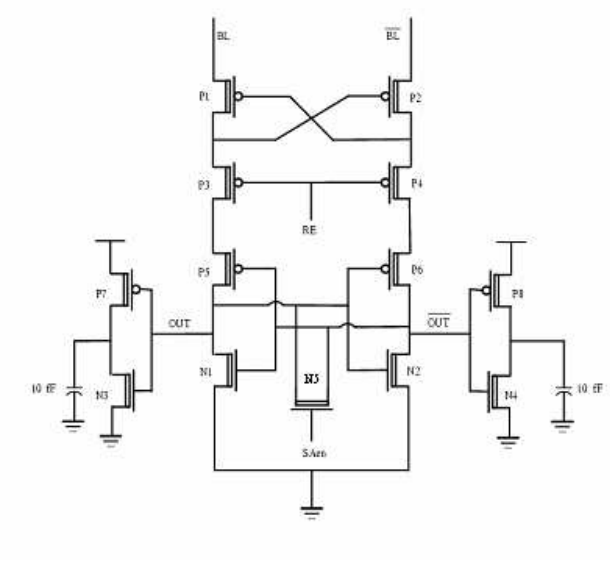

Figure 4.3: Schematic of Low Power Current Sense Amplifier (Note: all transistors are normal MOSFETs)

Enable (RE) signal is pulled low, the current conveyor circuit senses the current difference in the bitlines and this current difference flows through the two legs of the cross coupled inverter pair. All the four transistors N1, N2, P5 and P6 are in saturation as they are diode connected through transistor N5. After sufficient current difference is developed, the transistor N5 is turned off and the amplification process begins. The differential current flowing through the two legs of the sense amplifier causes a differential voltage to be developed at the sense amplifier outputs. Due to the high gain of the cross coupled latch, the sense amplifier outputs are quickly amplified. The two inverters present at the output of the sense amplifier ensures that the outputs are either a perfect V_{dd} or Ground.

The four transistor current conveyor ensures that there is no differential discharging of the heavily loaded bitlines and hence the sensing delay is almost independent of the bitline capacitances. Since all the transistors in the latch are

ON and in saturation during the sensing stage, the cross coupled latch provides a low input impedance to the bitlines. This is in direct contrast with other existing current sense amplfiers which have additional transistors in the amplification stage [33, 44] to provide a low impedance termination to the bitlines. In addition, the circuit arrangement ensures that the bitlines are isolated from the sense amplifier output, removing any influence of the output voltage on the bitlines. The power saving feature of this sense amplifier circuit is in the pre-sensing phase. Existing sense amplifier circuits either precharge their outputs to V_{dd} (CBLSA, CCIL) or predischarge them to ground (ICSA), before the read operation, and they further proceed to pull one of the output nodes to V_{dd} or Gnd. In our design, we move the output voltages of the sense amplifier closer to each other by turning ON transistor N5. By pulling the output voltages of the sense amplifier towards each other, we ensure that all the transistors are in saturation and the bitline differential current flows directly through the amplifier. This results in considerble energy savings during the read operation and hence a reduction in sensing delays.

4.4.1 Process Variations

As the technology scales, the difficulties in the fabrication process is expected to create parameter variations in the designed circuits and ultimately lead to a loss in performance. Variations in some of the key parameters, such as, the effective channel length L_{eff} [48] and threshold voltage V_t [29] significantly impacts the circuit performance. The impact of V_t and L_{eff} variations on sense amplifier circuits is even more pronounced as differential sense amplifiers are designed to be electrically balanced symmetric circuits [38]. Hence, even a slight mismatch in the threshold voltages or effective channel length of the two supposedly matched

transistors can lead to a significant increase in the sensing delay and ultimately a loss in functionality. Further, variations in V_t and L_{eff} alter the I-V characteristics of devices and make fast devices slow and vice-versa. The deviation in the threshold voltage results from a number of factors, such as, variation in the geometry of devices, random dopant number fluctuation, and mobile charges in gate oxide [47, 29]. Whereas, imperfections in the photolithography process is expected to cause variations in the effective channel length. In this section, a complete analysis to determine the worst case threshold voltage mismatch combination in both the WTA and LPCSA sense amplifiers for a particular read value is presented. A similar analysis is performed on other sense amplifier circuits as well.

Ideally, in traditional sense amplifier circuits, the *BLC* (Bitline Complement) can be infinitesimally less than the *BLT* (V_{dd}) for the sense amplifier output to be latched to the correct value. However, perfect matching of transistors is almost impossible and hence, *BLC* typically needs to be less than *BLT* by a finite amount for correct reading. This finite minimum value (typically around $150mV$) of voltage difference between the two bitines for the sense amplifier to evauate correctly forms the offset voltage of the sense amplifier. Due to threshold voltage mismatches between the differential pair input transistors and mismatches in the cross-coupled inverter pair of the sense amplfier circuit, functional failures in the read operation may happen. Such failures are mainly due to the shift in the offset voltage of a sense amplifier. The effects of threshold voltage variations in both the memory circuit and the sense amplifiers are summarized in Fig 5.2.

The process variation for each device that results in the worst case delay of the WTA sense amplifier is presented in Table 4.1. The devices are classified as those

connected to BL, devices connected to $\overline{BL}$ and common devices. An Inc and Dec represents a increase and decrease in either V_t or L_{eff}. The effect of varying either the threshold voltage or the effective channel length will have similar impact on the device. This is due to the fact that both have a similar relation to the drain current. The bitline BL is assumed to be high while the $\overline{BL}$ is pulled low. This means that more current flows through the transistors connected to BL than $\overline{BL}$. Hence, if devices $P1$ and $P3$ were made slower by increasing V_t and L_{eff}, it would increase the time taken for the bitline differential current to develop. For a similar reason, devices $P2$ and $P4$ are made faster. To read a 1 from the memory cell, the node out of the sense amplifier should be charged to V_{dd} while $\overline{out}$ is discharged to ground. This means that transistors $N1$ and $N3$ are ON while transistors $N2$ and $N4$ are OFF. Therefore, to obtain the worst case mismatch we need to increase the V_t and L_{eff} of both $N1$ and $N3$ to make them slower, and also make transistors $N2$ and $N4$ faster by decreasing V_t and L_{eff}. Similarly, in the two inverters at the output, transistors $N6$ and $P6$ are ON and hence are made slower, while devices $N7$ and $P5$ are OFF and are made faster. Transistor $N5$ is a common device to both BL and $\overline{BL}$ and is made slower to reduce the bias current, which adversely impacts the speed of amplification

A similar analysis for the LPCSA is performed and is summarized in Table 4.2, again by assuming the bitline BL is pulled high, while the $\overline{BL}$ is pulled low. This means that more current flows through the transistors connected to BL than $\overline{BL}$. Hence, as before, devices $P1$ and $P3$ were made slower by increasing V_t and L_{eff} Similarly devices $P2$ and $P4$ are made faster. To read a 1 from the memory cell, the node out of the sense amplifier should be charged to V_{dd} while $\overline{out}$ discharges

Devices connected to BL	Process Variation	Devices connected to $\overline{BL}$	Process Variation	Common Devices
P1	Inc	P2	Dec	N5(Inc)
P3	Inc	P4	Dec	
N1	Inc	N2	Dec	
N3	Inc	N4	Dec	
P5	Dec	P6	Inc	
N6	Inc	N7	Dec	

Table 4.1: Worst Case Process Variation in WTA

to ground. This means that transistors $P5$ and $N2$ are ON while transistors $P6$ and $N1$ are OFF. Therefore, to obtain the worst case mismatch we need to increase the V_t and L_{eff} of both $P5$ and $N2$ to make them slower, and also make transistors $P6$ and $N1$ faster by decreasing V_t and L_{eff}. Similarly, in the two inverters at the output, transistors $N3$ and $P8$ are ON and hence are made slower, while devices $N4$ and $P7$ are OFF and are made faster. Again, as before, transistor $N5$ is a common device to both BL and $\overline{BL}$ and is made slower to reduce the bias current and the speed of amplification.

Devices connected to BL	Process Variation	Devices connected to $\overline{BL}$	Process Variation	Common Devices	Process Variation
P1	Inc	P2	Dec	N5	Inc
P3	Inc	P4	Dec		
P5	Inc	P6	Dec		
N1	Dec	N2	Inc		
P7	Dec	P8	Inc		
N3	Dec	N4	Inc		

Table 4.2: Worst Case Process Variation in LPCSA

Though Monte Carlo analysis can give better understanding of variation tolerance, we chose to use our approach for two reasons: 1) We can customize the variations of each circuit to its worst case performance and thus get a better picture of the robustness. 2) The variations in W and L_{eff} result in variations in V_t which we have modeled.

Our claim is that making transistors that are fully ON at the sensing period weaker and vice versa results in the worst case scenario for threshold voltage mismatch. This is a different approach of finding the failing point of a circuit and the worst case variation condition will be different for each circuit depending on its structure and functionality. We preferred this deterministic approach as it could also perform variation analyses in opposite directions simultaneously. This kind of analysis may be well suited for sense amplifier like circuits where matching of neighboring transistors are critical for the functioning. A similar analysis with large number of sample points will provide an improved flavor of Monte Carlo analysis.

The impact of worst-case V_t and L_{eff} variation on the sensing speed of different designs is presented in the next section.

4.4.2 Simulation Results

Simulation Setup and Methodology

The simulation setup that we used has a single column of memory cells connected to the bitlines as shown in Fig. 4.4. The number of memory cells connected to the bitlines are varied from 64 to 256 to reflect the increasing bitline capacitance. The bitline interconnect parasitics are modeled as a 3π RC network [44] using the

Berkeley predictive models for interconnects [22].

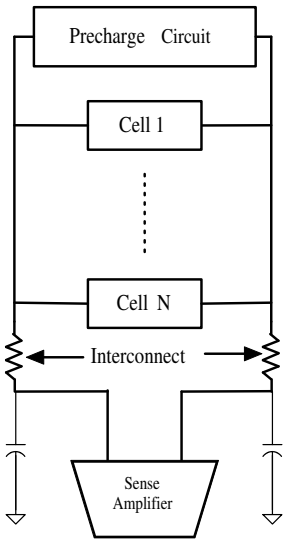

Figure 4.4: Simulation Setup with Precharge Circuitry and Memory Column

The simulations were then performed using Cadence suite for the 70nm Berkeley Predictive Technology Model. A V_{dd} of $1V$ was used for all simulations and the transistor sizing in the different sense amplifiers are made the same for fair comparison. All the sense amplifiers were enabled after a voltage difference of 50mV or current difference of 20μA is developed. These numbers aim to provide a common and fair means of comparing the performance of both the voltage mode and current mode sense amplifiers [44]. The sensing delay was calculated as the time required for the sense amplifier output to reach 90% of V_{dd} after enabling the word line. The timing diagram for the different clock signals is shown in Fig. 4.5. The time elapsed between word line enable and SA enable constitues the sensing delay, while the amplification delay is time between the SA enable and its output reaching 90% of V_{dd}. The impact of V_t variation on sensing delay is simulated by

varying the V_{th0} value in the model file. The effective channel length in the nominal model of a $70nm$ BPTM technology node is $38nm$. To account for the variation in L_{eff} the parameter *Lint* in the model file is varied by 20% on either side of the nominal value to account for the 40% variation. In addition, the effect of power supply variation on the sensing delay of all sense amplifiers is evaluated.

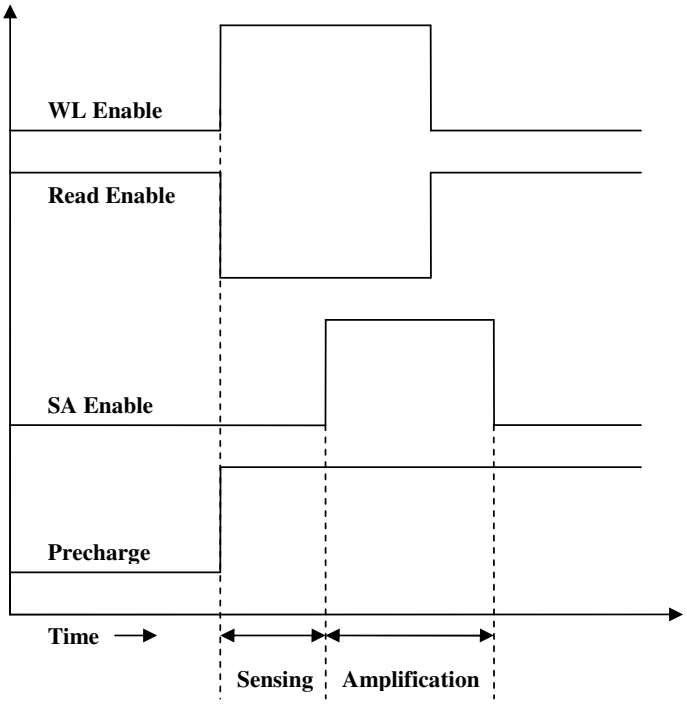

Figure 4.5: Timing waveform and Delay calculation

Effect of Bitline Capacitance

Figure 4.6 shows the delay measurements for the different sense amplifiers under the influence of increasing number of cells per memory column. From the slopes, it can be observed that unlike the voltage mode sense amplifier, the sensing delay of the current mode sense amplifiers is almost independent of the bitline capacitance , specifically for WTA and LPCSA.

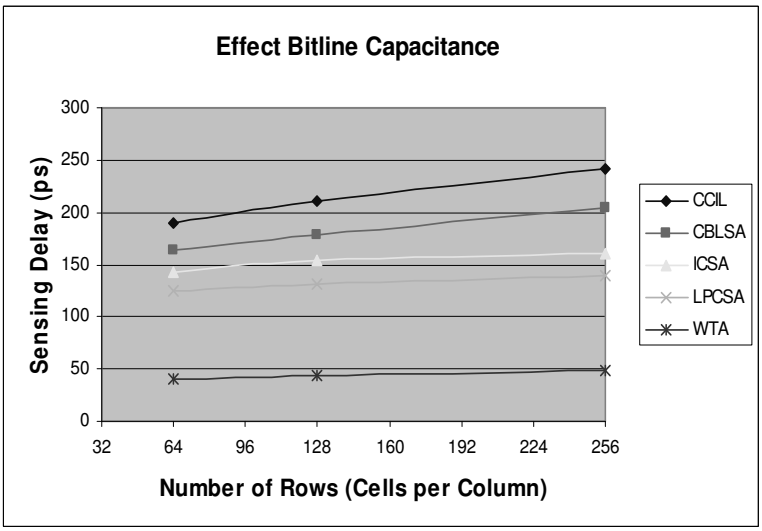

Figure 4.6: Effect of Bitline Capacitance

The WTA offers significant improvement in sensing speed over all the other sense amplifiers that are compared. The performance speed-up of WTA over CCIL, is a 78.3% for a column of 64 memory cells and goes upto 80.2% for a column of 256 cells. The corresponding speed up figures for LPCSA are 34% for 64 rows and 42.5% for 256 rows over CCIL amplifier. This clearly indicates the advantage of using current mode sensing and degradation of the performance of voltage mode sense amplifier in the face of increasing bitline capacitance. When contrasted with other current sensing techniques, WTA offers about 75-76.5% savings over CBLSA and 67.2-70.2% over ICSA. This considerable improvement in the sensing speed is attributed to the winner take all amplification stage which offers high sensitivity to the bitline differential current. Moreover, the delay performance of WTA experiences the least degradation as the bitline capacitance increases. This is evident from the fact that the delay of WTA increases only by a marginal $7ps$, as opposed to other current mode techniques which experience a $40ps$ (CBLSA)

and $19ps$ (ICSA) increase. Correspondingly, LPCSA offers about 23.8-31.9% savings over CBLSA and 12-13.6% over ICSA. The delay performance of LPCSA, also experiences lesser degradatios, than both CBLSA and ICSA, as the bitline capacitance increases. This is evident from the fact that the delay of LPCSA increases only by a marginal $14ps$, as opposed to other current mode techniques which experience a $40ps$ (CBLSA) and $19ps$ (ICSA) increase. Thus, WTA circuit exhibits best immunity to the increasing bitline capacitance and offers maximum performance speed-up followed by LPCSA, ICSA, CBLSA and CCIL, in that order.

Figure 4.7 shows the performance speed-up offered by WTA over its closest competitor ICSA. The word lines for both the sense amplifiers is pulled high at the same time. From the figures, we can see that the *SAen* signal for WTA is activated before the *SAen* for ICSA and that the output nodes switch values faster in WTA. A correspoding performance speed-up graph for LPCSA is shown in Fig. 4.8. From the figure, it is evident that LPCSA is faster, though not as much as WTA, than ICSA.

Energy Consumption

The energy consumption of all the sense amplifiers was measured for varying number of cells per column and the results are shown in Fig. 4.9. Figure 4.10 shows the total energy consumed during the read operation, including both the energy consumed during the precharge phase and the energy consumed during amplification. The energy consumption for LPCSA is the least due to the power saved in the pre-sensing phase and is also almost independent of the bitline capacitance as the current conveyor always maintains the bitlines close to V_{dd}. Results indicate that LPCSA offers 46% savings in energy consumption over CCIL

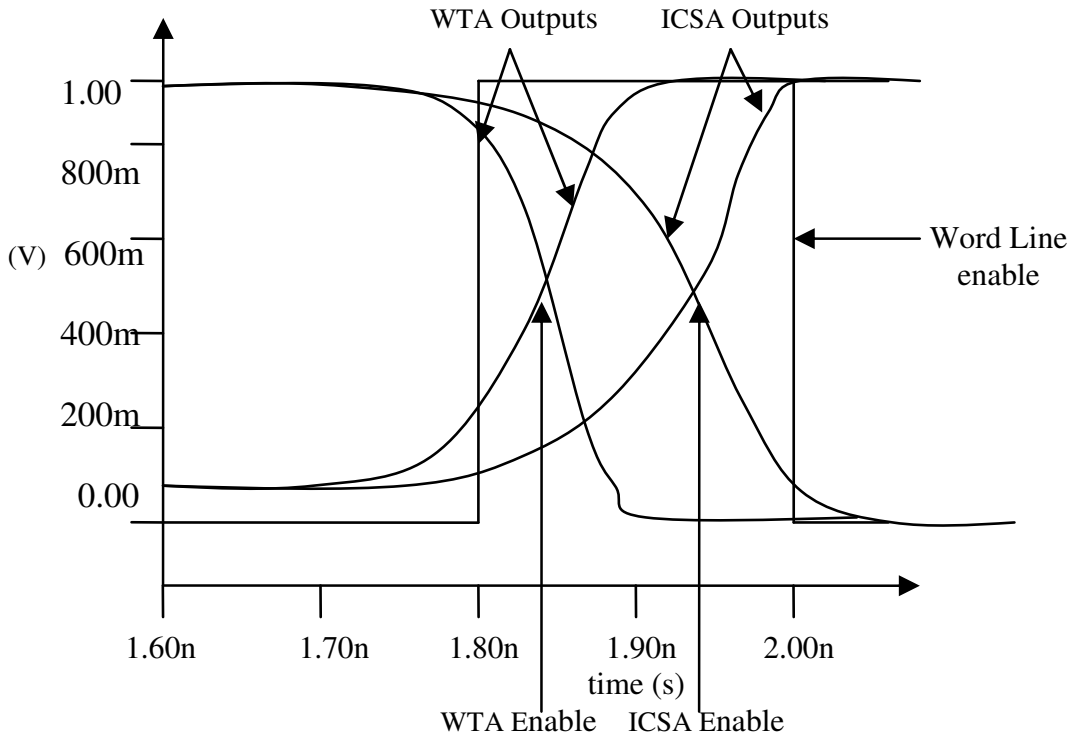

Figure 4.7: Delay Comparison of ICSA and WTA

for a 64 cell column and about 55.3% savings for a 256 cell column. This shows the increase in energy consumption of voltage mode techniques with increasing number of cells per column, as the energy is dissipated in charging the bitlines. The corresponding figures for WTA, is that it offers 15.2% savings in energy consumption over CCIL for a 64 cell column and about 28.6% savings for a 256 cell column, which reinforces the conclusion reached above. In addition, LPCSA consumes 86.3% and 54.9% less energy than CBLSA and ICSA, respectively, for a 256 cell memory column. Whereas WTA consumes 73.9% and 28% less energy than CBLSA and ICSA respectively for a 256 cell memory column. Though ICSA reduces the bitline swing it fails to maintain a virtual short across the bitlines and predischarges its outputs to ground and hence consumes more energy then

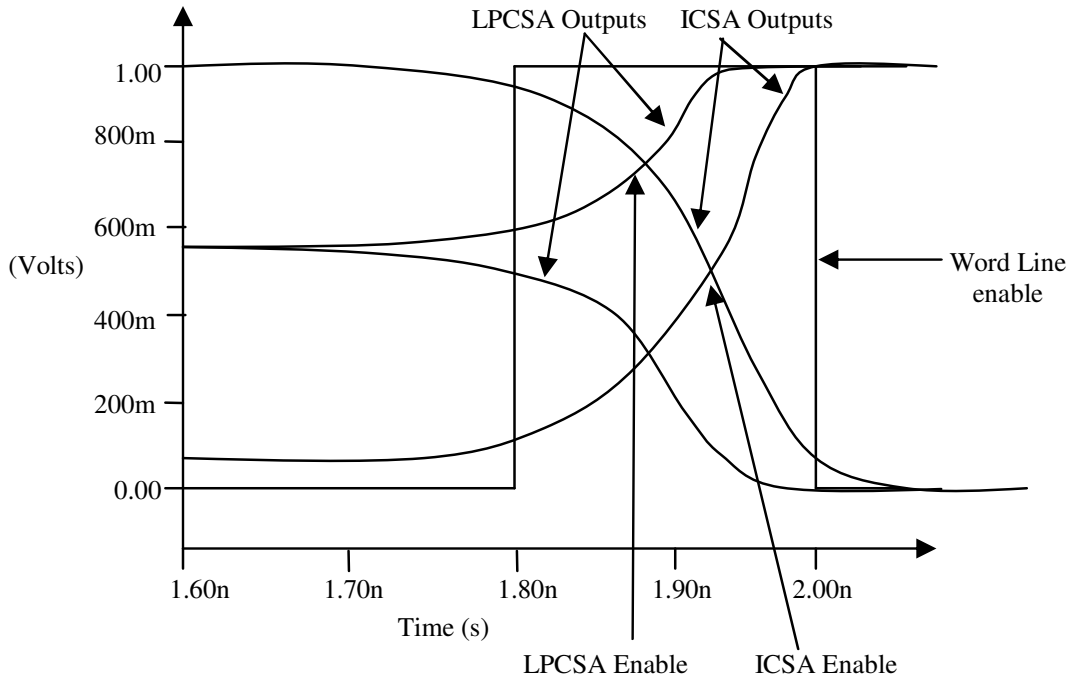

Figure 4.8: Delay Comparison of ICSA and LPCSA

LPCSA. Among the other sense amplifiers the CBLSA circuit consumes maximum energy as it clamps the bitlines through its low impedance nodes. Both CCIL and ICSA exhibit low energy consumption, when contrasted with CBLSA, and a similar trend is observed with increase in bitline capacitance.

Threshold Voltage Variation

Table 4.3 shows the analyses of our worst case variation in threshold voltage that results in the maximum sensing delay. For variations that exceed the presented values, the circuits may stop functioning. The worst case mismatch combination of the devices for all the other sense amplifiers is determined by performing an analysis similar to that explained in Section 5. In Table 4.3, the first column indicates the maximum variation that the specific circuit could tolerate without

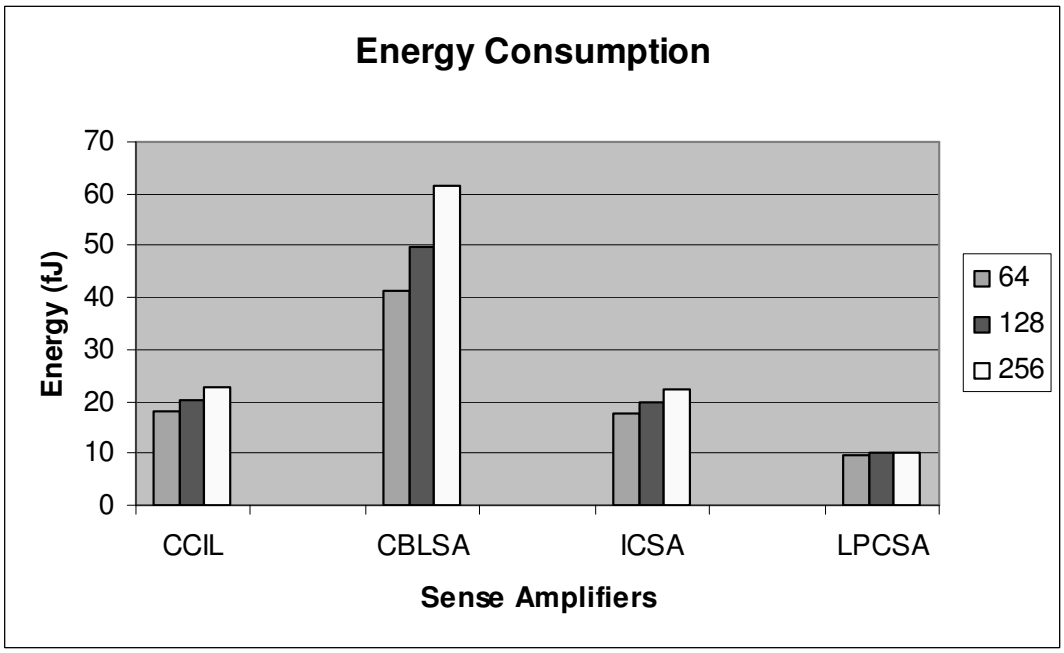

Figure 4.9: Energy Consumption of Sense Amplifier per Read Operation

losing the functionality and the corresponding sensing delays are presented in the second column. Of course, for lower variations the sensing delay will reduce significantly.

Sense Amplifier	% V_t Variation	Delay (ps)
CCIL	21	422
CBLSA	1	262
ICSA	8	247
WTA	35	464
LPCSA	8	224

Table 4.3: Impact of V_t Mismatch on Sensing Delay. *WTA functionality does not fail beyond 35% variation.

The results are consistent with the general rule that V_t mismatch tolerance reduces for circuits with increased complexity. The proposed WTA requires just two additional transistors in the sensing path over the basic cross coupled latch

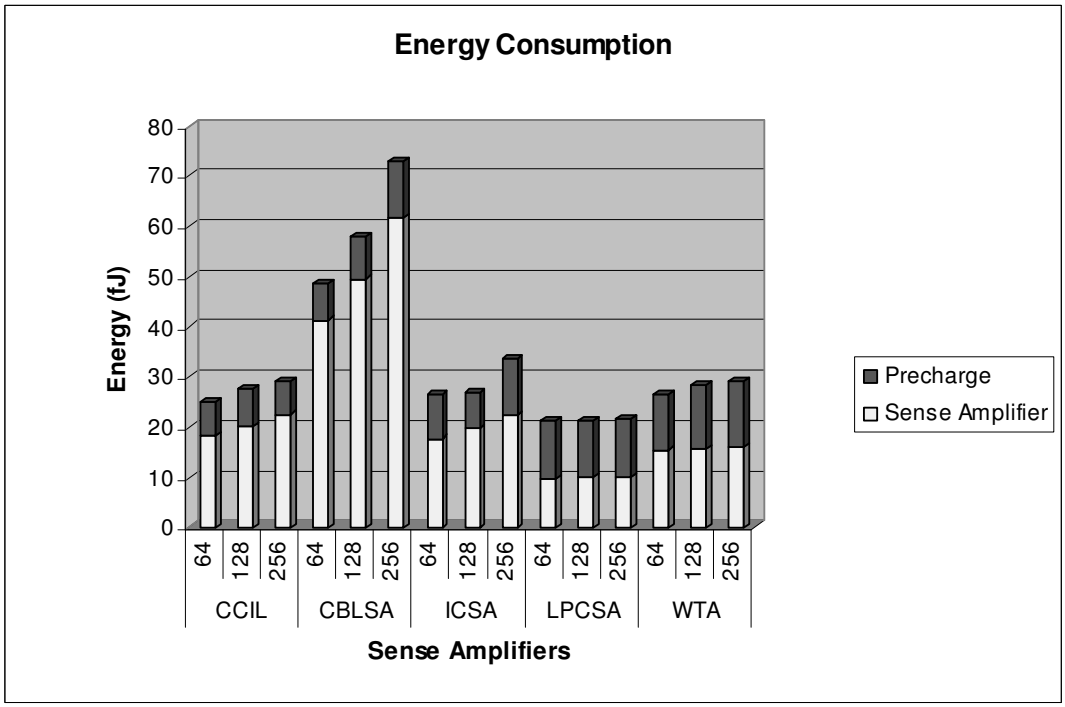

Figure 4.10: Total Energy Consumption per Read Operation

and hence tolerates up to 35% variation in V_t. We also need to note that due to its inherent working principle, WTA will amplify and provide the correct outputs as long as there is a differential in the bitline currents. One possible reason for this robustness is the simple amplification stage that offers excellent tolerance to wide range of variations. LPCSA also requires only two additional transistors in the sensing path and tolerates upto 8% variation in V_t. Among current sense amplifiers with cross coupled amplifier topology, LPCSA circuit is one of the most simple, utilizing minimum number of transistors, and offering adequately robust performance in presence process variations. CBLSA displays least tolerance to V_t mismatch as it has two additional transistors in the amplification stage as opposed to ICSA and WTA. The numbers provided in Table 4.3 indicate only the worst case scenario, under maximum possible variations in V_t. The combination

of increasing and decreasing variations of neighboring transistors may not even happen in real systems.

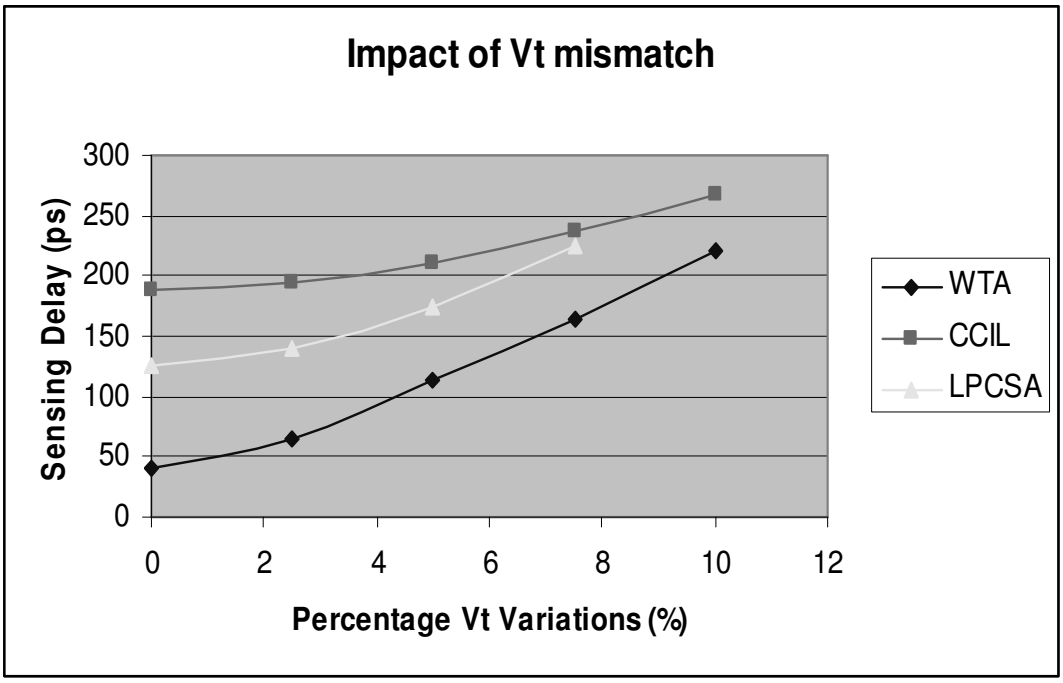

Figure 4.11: Impact of V_t Variation on Sensing Delay

It is also useful to analyse how current sense amplifiers perform against voltage mode sense amplifiers in the presence of increasing V_t mismatch. Figure 4.11 shows how the sensing delay of WTA and LPCSA are affected by increasing V_t mismatch as against CCIL. Though the sensing delay of both WTA and LPCSA are numerically less than the sensing delay of CCIL, the trend observed, may indicate that the speed penalty under V_t mismatch is more pronounced on current sense amplifiers than voltage mode sense amplifiers. However, these are under worst case assumptions and the speed-up gains offered by current sensing for variations less than 10strategy in nanoscale memory systems. Further, WTA does not fail functionally even at 35to full rail values at a performance penalty. Whereas,

CCIL sense amplifier fail to function for larger variations.

Effective Channel Length Variation

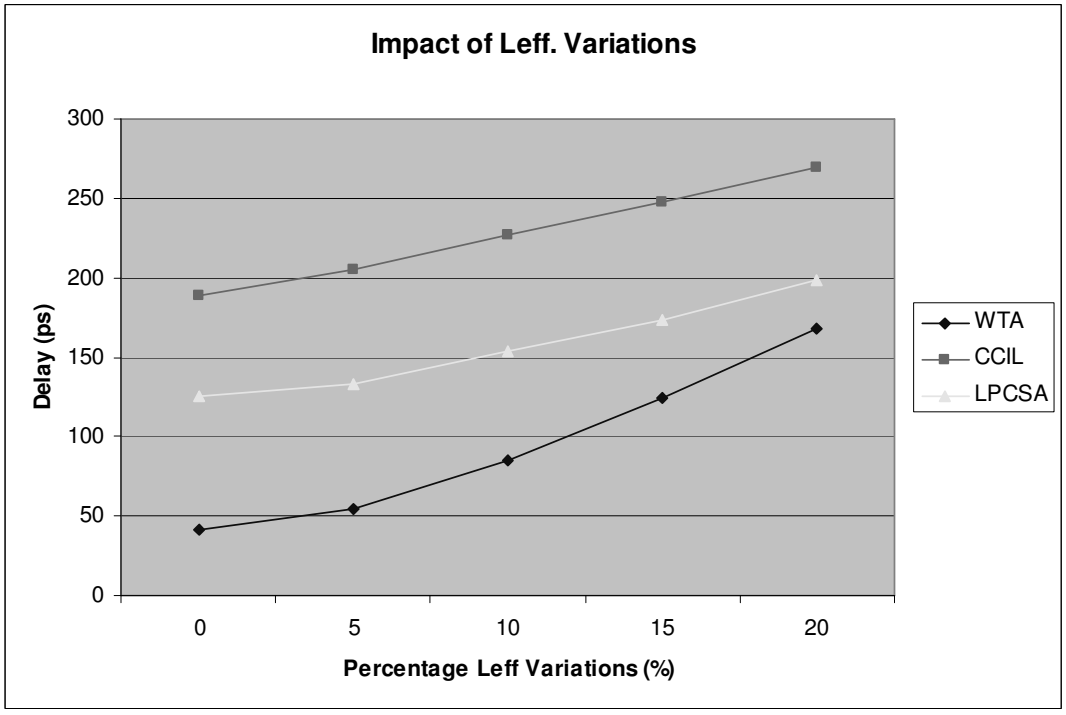

Figure 4.12: Impact of L_{eff} Variation on Sensing Delay

We compare the delay performance of WTA and LPCSA, in the presence of variations in L_{eff}, with the most robust sense amplifier, CCIL. The worst case mismatch combination in L_{eff} is determined as explained in the previous section. All three sense amplifiers tolerate upto 20% variation and have an expected degradation in sensing delay as shown in Fig. 4.12. However, the WTA continues to offer about 37.5% savings and LPCSA offers 26% savings, over CCIL for a 20% variation in L_{eff}.

Power Supply Variation

A maximum of 12% variation in V_{dd} is expected in the $70nm$ process. Table 4.4 presents the performance of the different sense amplifiers in the presence of variations in V_{dd}. The simulations were performed with the sense amplifiers having their V_{dd} lower than the 1V that is provided to the other cells. CCIL, LPCSA and WTA tolerate a 12% variation in V_{dd}, ICSA withstands a 11% variation while CBLSA displays the least tolerance of only 3% to variations in V_{dd} . All sense amplifiers experience a similar percentage degradation in sensing delay with variation in the supply voltage.

Sense Amplifier	% V_{dd} Variation	Delay (ps)
CCIL	12	271
CBLSA	3	267
ICSA	11	219
WTA	12	135
LPCSA	12	207

Table 4.4: Impact of V_{dd} Variations on Sensing Delay

4.5 Summary

In this chapter, we discussed the functioning of an important class of memory peripheral circuits, namely, the sense amplifiers. We differentiated the sense amplifiers into two groups - voltage mode and current mode, depending upon how they sense the bitline differential signal. We presented the earlier work in both current mode and voltage mode sense amplifiers, and pointed out their limitations. Finally, Two novel robust high performance current mode sense amplifier

with improved power consumption for nanoscale SRAM Memories is presented. One sense amplifier uses a winner take all approach to provide fast amplification while the other design, based on a cross coupled latch, focusses on low power operation. The WTA sense amplifier is highly robust to mismatch in threshold voltage and tolerates upto 10% variation in V_t with acceptable degradation in the sensing delay, while LPCSA tolerates upto 8% variations. Simulation results show that our designs are also tolerant to variations in the effective channel length and supply voltage. WTA offers around 70-80% speed improvement and the LPCSA around 12-32%, when compared to other voltage and current mode sense amplifiers. Such large improvements are possible due to the inherent design and amplification mechanism of the sense amplifier design. In addition, unlike other current sensing techniques, we do not have excessive bitline swings or additional circuitry in the amplification stage. Consequently, this results in significant speed improvement and tolerance to process variations. Since it does not precharge/predischarge the output nodes to V_{dd}/ground, LPCSA also consumes the least power among the sense amplifiers considered. Thus, the performance of both WTA and LPCSA is least affected, in terms of both sensing speed and energy consumption, in the presence of increasing bitline capacitance and process variations.

Chapter 5

SRAM Reliability: Process Variations

5.1 Impact of Variability on SRAM Designs

Technology scaling has enabled us to integrate both memory and logic circuits on a single chip. However, the performance of embedded memory and its peripheral circuits can adversely affect the speed, reliability and power of the overall system. One of the key challenges that limits the performance in memory and microprocessor design is the systematic and random variations in process, supply voltage and temperature (P, V, T) [36]. Conseqently, technology scaling beyond $90nm$ causes higher levels of device parameter variations, such as, variations in threshold voltages and effective channel lengths, thus changing the design problem from deterministic to probabilistic [49]. Further, variations in V_t and L_{eff} alter the I-V characteristics of devices and make fast devices slow and vice-versa. This deviation in the threshold voltage results from a number of factors, such as, variation in the geometry of devices, random dopant number fluctuation, and mobile

charges in gate oxide [47, 29].

The impact of these threshold voltage variations are more pronounced in minimum geometry transistors commonly used in area-constrained circuits such as memory cells. They also significantly affect the functioning of circuits, such as, sense amplifiers, which are designed to be electrically balanced and symmetric circuits [38] and any small variation in the device parameters would adversely affect the circuit functionality and performance. Hence, a thorough understanding of the dynamic stability under process parameter variations is crucial to determine the design window available for existing read/write circuit styles. Further, a failure in any one of the cells in a column of the memory array will make that column faulty. If the number of such columns exceeds the available redundant columns, then the chip is considered a fauty chip. Consequently, the failure probability of the cell is directly related to the yield of the chip [50]. Hence, estimating both the read and write failure probability of the memory cell for different peripheral circuit style is necessary in the design phase to ensure a good yield.

In this chapter, we analyze the different dynamic failure mechanisms of SRAM designs in an advanced $65nm$ technology to understand the impact of process parameter variations on the stability of memory cells [51]. The findings from these failure analyses can be used in the early design cycle to optimize the design for yield enhancement. We consider three sense amplifier designs and two different write architectures in our analyses. We finally present detailed simulation results for the above designs and discuss the impact of the different design styles and architectures on the cell failure probability and memory yield. The SRAM cell used in our analyses has a small aspect ratio, reflecting the industry trend of thin

cell images [52, 53]. The analyses were performed using simulation models for an advanced $65nm$ technology [54].

5.2 Failure Mechanisms in SRAMs

The principal source of variation is the intrinsic fluctuation of V_t due to random dopant effect [29]. Therefore, in this work, we focus mainly on the threshold voltage variations. This section first investigates the failure mechanisms under threshold voltage variations for the local bit lines in a memory system, comprising SRAM cells, sense amplifiers and the write drivers. Then, the different read and write circuits considered in our analyses are described in detail.

Process variations in SRAM cells (Fig 5.1) may result in [50] four different failure mechanisms - read access failure, flipping read failure, write failure and hold failure. Both flipping read failure and hold failure occur only in the presence of excessive variations coupled with increased disturbance to the cell and very low supply voltages. The robustness of the cell against threshold voltage variations during the read and write operation defines the dynamic stability of the cell. Write failure and read access failure may result even in the presence of slight variations and thus have to be analyzed carefully. These failures are of even more concern when we employ aggressive timing and low supply voltages.

Let us assume that the SRAM cell shown in Fig 5.1 stores a value of zero (*i.e* V_R=0 and V_L=1). During a read operation, the access transistor S_R and the NMOS pull-down transistor N_R form a resistor-voltage-divider between the nodes *BLT* and *R*. This results in a slight increase in the voltage V_R, thereby discharging the node *L* from '1' and reducing the strength of N_R. There may also be a decrease

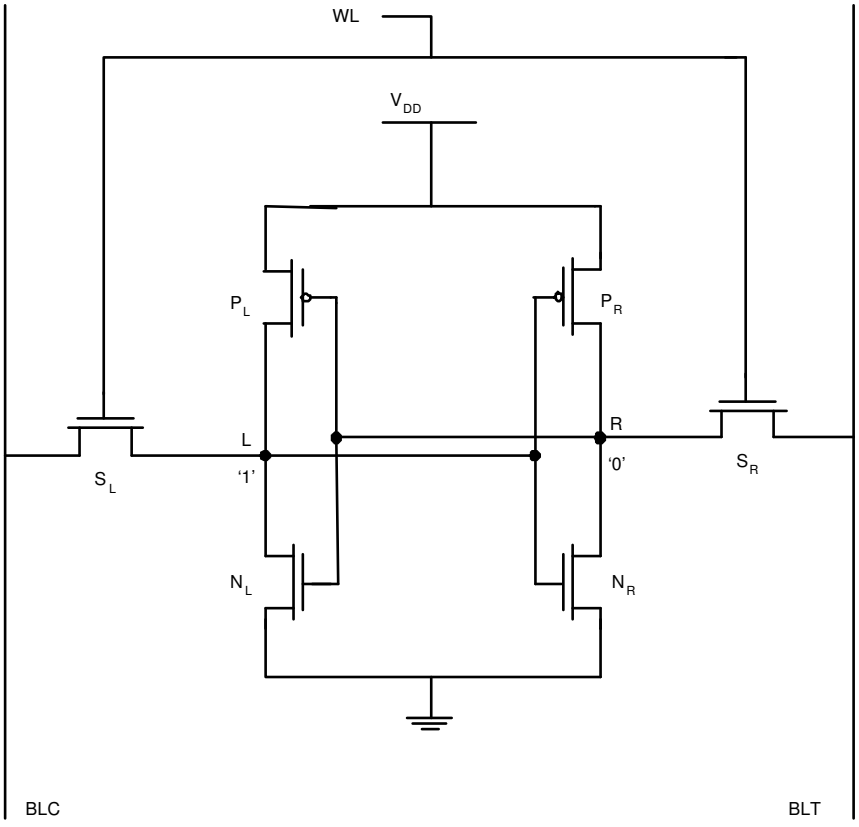

Figure 5.1: 6T-SRAM Cell

in the current that discharges the bitlines during the read operation due to weak (i.e. high V_t) access transistor S_R. Due to the above mentioned mechanisms, the volatge difference between the bitlines when the sense amplifier samples them may be less than the required value, which may result in wrong evaluation of the value stored in the memory cell.

Ideally, in traditional sense amplifier circuits, the *BLC* can be infinitesimally less than the *BLT* (V_{dd}) for the sense amplifier output to be latched to the correct value. However, perfect matching of transistors is almost impossible and hence, *BLC* typically needs to be less than *BLT* by a finite amount for correct reading. This finite minimum value (typically around $150mV$) of voltage difference between the two bitines for the sense amplifier to evauate correctly forms the offset voltage of the sense amplifier. Due to threshold voltage mismatches between the differential pair input transistors and mismatches in the cross-coupled inverter pair of the sense amplfier circuit, functional failures in the read operation may happen. Such failures are mainly due to the shift in the offset voltage of a sense amplifier. The effects of threshold voltage variations in both the memory circuit and the sense amplifiers are summarized in Fig 5.2.

5.3 Small Signal Read Circuits

Sense amplifier is one of the important peripheral circuits in the memory system as it strongly influences the memory read access times. It retrieves the stored data from the memory array by amplifying the small differential signal on the bitlines. In general, sense amplifiers have two stages of operation: the sensing stage and the amplification stage. Majority of the existing sense amplifiers utilize

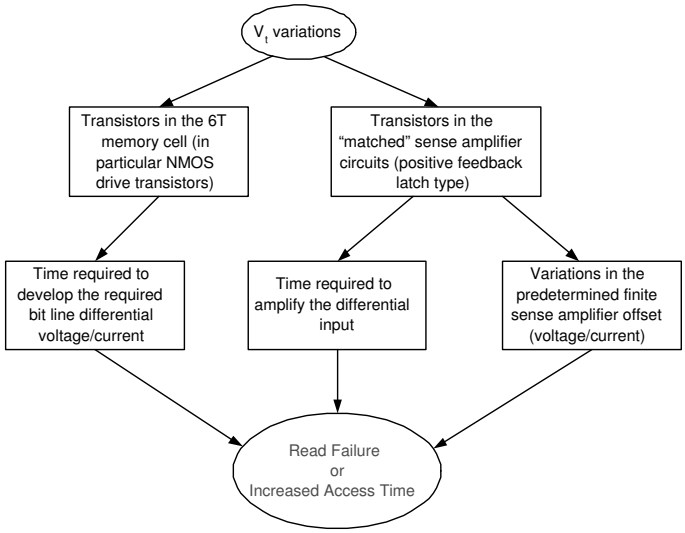

Figure 5.2: Read failure mechanisms due to V_T variations

a cross coupled transistor topology for amplification and differ primarily in the type of signal sensed and their sensing circuits. This section describes the three different sense amplifier circuits considered in our variation analyses.

5.3.1 Cross-Coupled Inverting Latch (CCIL)

This is one of the most commonly used sense amplifier circuit and has two cross coupled inverters with very high gain to provide fast amplification [43] (see Fig 5.3). The bitlines are directly terminated at the sense amplifier outputs through the read enable transistors. Initially, both the bitlines and the sense amplifier outputs are precharged. When sufficient voltage difference in the bitlines develops, the sense amplifier is turned on and the amplifier latches onto the value stored in the memory. The main drawback is that the amplifier functioning depends on the discharge of the bitline capacitances through the access and the

NMOS drive transistors connected to the node storing zero to sense the differential voltage. As technology scales down and the number of memory cells per column increases, even a slight variation in the strengths of these transistors or that of the sense amplifier transistors significantly affects the the time to develop the differential voltage in the bitlines increases significantly. This results in a considerable increase in the sensing time or may even result in incorrect readout irrespective of how fast the amplification process may be.

5.3.2 Mid Rail Low Power SA

The mid rail sense amplifier shown in Fig 5.4 is a low power version of the cross coupled inverter latch type sense ampifier. In this design, to obtain significant savings during the read operation, the bitline swings are reduced as much as possible. Minimal swing in the bitlines reduces the charging and discharging of the bitline capacitance, while continuing to provide sufficient signal differential to the amplifier. The differential voltage between the two bitlines is sensed using two PMOS transistors ($P1$ and $P2$) and the cross coupled inverter pair ($P3$, $P4$, $N1$, $N2$) amplifies the difference. Two CMOS inverters that drive a $10fF$ load are connected to the output of the cross coupled inverter pair to produce a rail to rail signal at the output.

Before the start of the read operation, the sense amplifier outputs contain the previous read state and hence needs to be reset. So, the enable transistor ($N3$) is turned ON through the SAE signal to bring the voltages at the output nodes closer to each other. The circuit arrangement ensures that the bitlines are isolated from the sense amplifier outputs, removing any influence of the output voltage on the bitlines. In this design, the output voltages of the sense amplifier

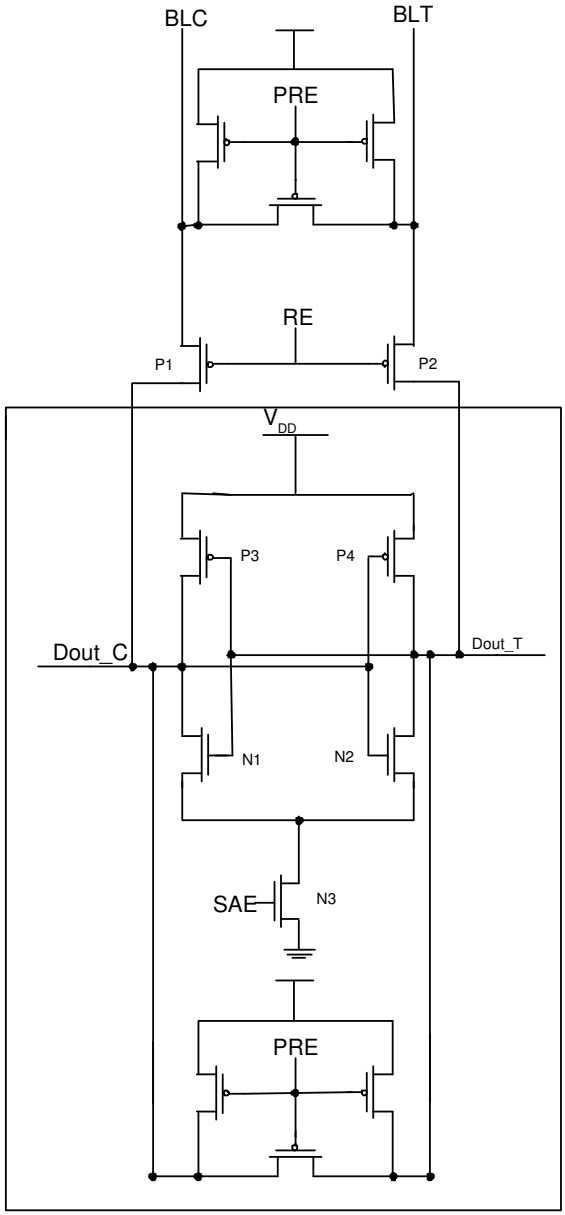

Figure 5.3: CCIL type sense amplifier where complementary bitlines are precharged to high

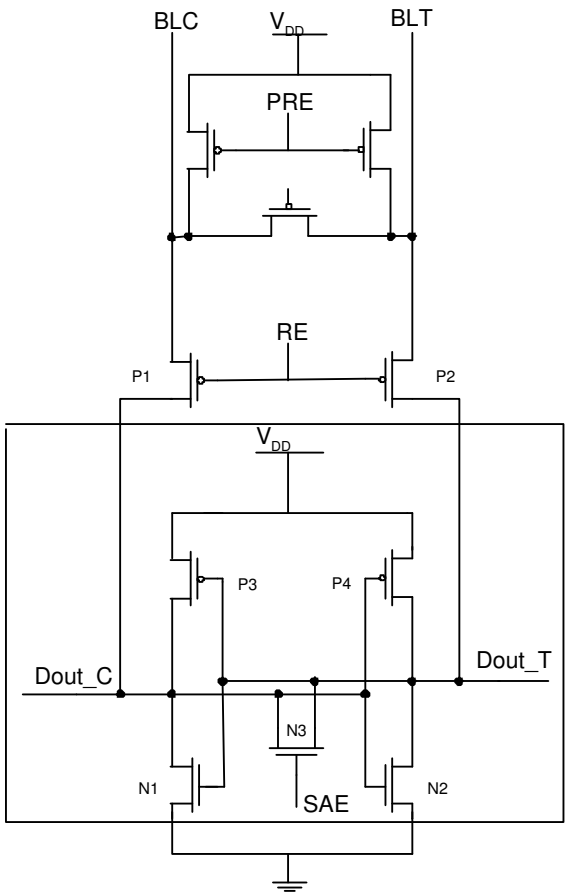

Figure 5.4: Low power sense amplifier with bitlines precharged to high and SA outputs precharged to midrail voltage

are moved closer to each other by turning ON transistor $N5$. Though this design has good power saving features it is highly immune to process variations due to the minimal swings in the bitlines. As shown by the simulation results, a slight mismatch in the threshold voltages of the transistors would disrupt the bitline differentail voltage and affect the functionality of the read operation.

5.3.3 Gate Sense Current SA

This variation tolerant current sense amplifer [55](Fig 5.5) is based upon a conventional gate-terminal input sense and latch circuit [56]. The current flow of two NMOS transistors, $N3$ and $N4$, whose gates are connected to the two bitlines, controls the serially connected latch circuit. A small difference of current through these two NMOS transistors converts into a large output voltage. Cross coupled PMOS transistors ($P3$, $P4$) are added to reduce the sensitivity to V_t mistracking between both $N1$-$N2$ and $N3$-$N4$. $P3$ and $P4$, directly inject BLC/BLT signals to the SA internal nodes, $net0$ and $net1$, during precharge before SA enable (SAE) is asserted [55].

Due to the BL and SA output isolation, BL precharge starts once SA is set. In addition, by keeping the sense amplifier enable signal, SAE, high, the SA output is extended to the next cycle. In this design, both the bitlines and the sense amplifier outputs are precharged high using dedicated precharge circuitry.

We restrict our failure analyses to the three sense amplifier techniques mentioned above. Though there have been other recent sense amplifier circuits [34, 44] they are slight variations of the three established techniques and provide marginal improvement at the cost of degraded performance in the presence of process variations.

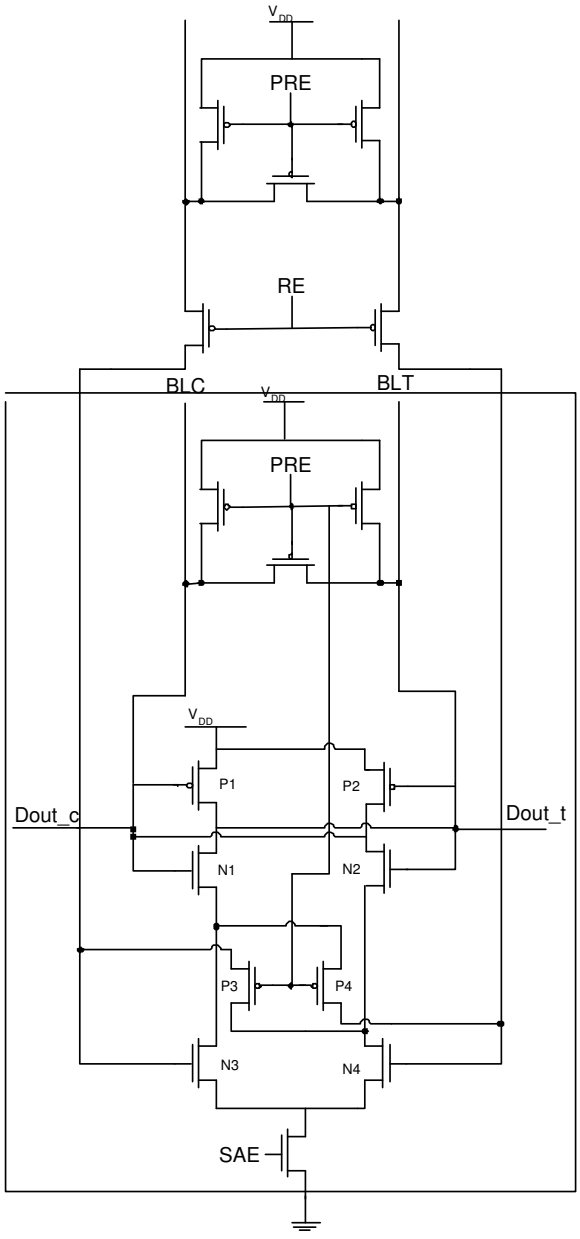

Figure 5.5: Gate sense current sense amplifier

5.4 Simulation Setup and Failure Criteria

The failure analyses were performed for both the read and write operations of $65nm$ SOI 6T-SRAM cell with floorplan information taken into consideration. An aggressive timing of $1ns$ cycle time was used for both the read and write operation. A range of supply voltages varying from $0.8V$ to $1.2V$ were used to analyze the impact of supply voltage variations. We then observed the V_{dd} region in which the cell is stable when accessed for each unit of variation in the threshold voltage. A larger stable region indicates a better robustness against failures due to process variation. Hardware data obtained from the SRAM testsites were used in introducing specific variations to the functionaly critical transistors (not shown here). Notice that in this study, parameters are treated as independednt and variations are applied in the worst case directions, thus representing the worst case scenario. We now discuss the simulation setup, circuit architecture and the failure criteria used for both read and write operation.

5.4.1 Read Operation

Each bitline spans 64 thin cell images with small aspect ratio. Four blocks of 64 cells each are considered and one sense amplifier is shared by four such bitline pairs (as shown in Fig 5.6). Timing skew between near-end and far-end is a maximum of $3pS$. Critical timing signals, WL, SAE and RE are generated manually with external margin control considering the circuit topologies of the three different sense amplifiers. The bitines of the accessed cell are precharged high and the sense amplifier outputs are precharged high or mid rail depending on the circuit type. The bitline interconnect parasitics are modeled as a 3π RC network

using the $65nm$ wire models for interconnects. Different supply voltages were used for the memory cells and the transistor sizing in the different sense ampifiers are made similar for a fair comparison. All the sense amplifiers are enabled after a voltage difference of $150mV$ is developed across the bitlines. The sensing delay was calculated as the time required for the sense amplifier output to reach 90% of V_{dd} after enabling the word line. We need to note that the side that undergoes a strong pull-down, when the sense amplifier is enabled, is the critical transition for timing.

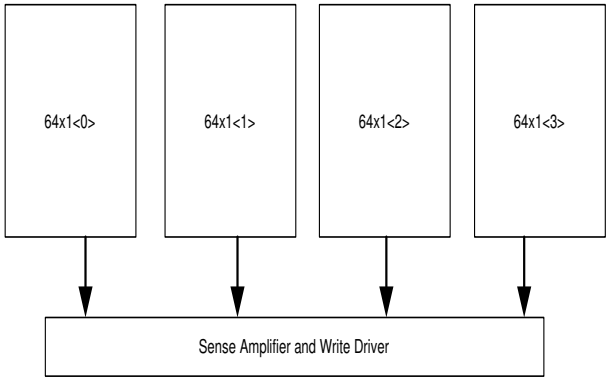

Figure 5.6: Simulation Setup for a Read Operation

In all the considered sensing schemes, complementary bitlines develop a small voltage differential that triggers the large voltage swing to rails when the sense amplifier is enabled. The failure criteria for a read operation is then determined based on the voltage difference developed across the bitlines of the accessed cell. The read operation fails when a threshold differential voltage is not developed after word line activation within a predetermined time interval. The threshold differential voltage may or may not be scaled proportional to the supply voltages. Detailed simulation results and our observations for the different sensing schemes under varying process variations and transistors strengths are presented in the next

section.

5.4.2 Write Operation

During a write operation, one of the bitlines is pulled low and the other pulled high to flip the contents of the memory cell. Two different write driver circuit topology was considered. Both the write schemes use a domino style fast rail-to-rail write. The first has write switches and column circuitry common to 64 cells. The longer bitline dictates that the write drivers are sized large enough to flip the contents of the memory cell within the specified $1ns$ cycle time. The second write topology uses write in a short bitline architecture (16 cells). A simplified cross section diagram for write operation is shown in Fig. 5.7. A group of local bitline pairs are common to a global bitline pair within a single column. All the local bitlines are preset to high, whereas, the global bitlines are preset to low.

Due to variations in the threshold voltages, the strengths of the access transistors and the trip point of the inverter may deviate from the nominal values resulting in a write failure. A write failure happens when the memory cell holds its previous state without flipping its value.

5.5 Simulation Results

5.5.1 Corner Analyses and Failure Trends

A sequence of two write and read operations were performed on all these circuits and the overall failure trend is determined using corner analyses. Four different product corners that characterize the failure regions were considered to identify ranges for stable device parameters in a 6T-cell: Typical, worstcase, strong NFET -

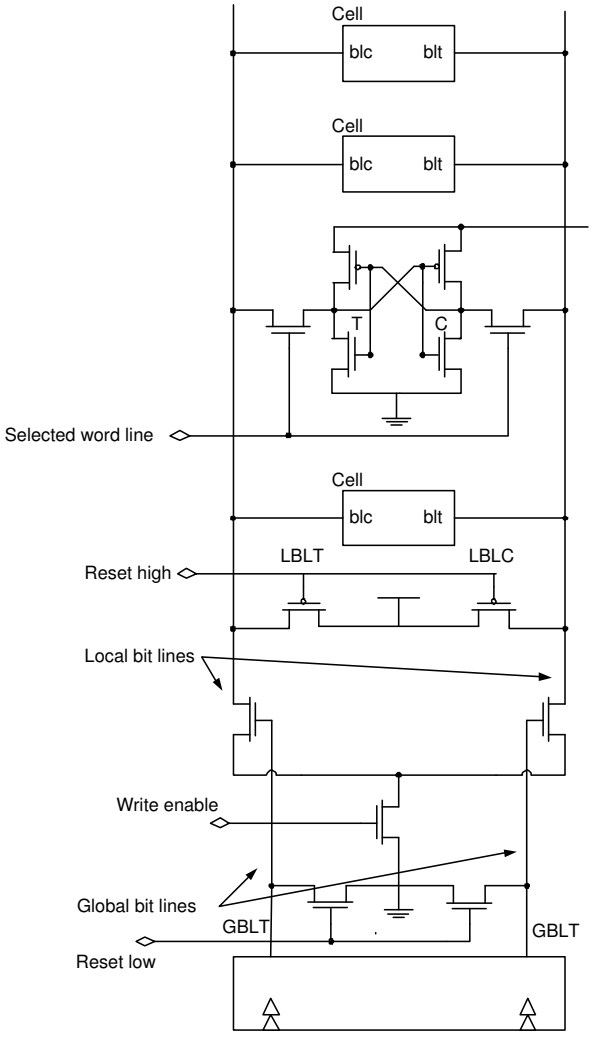

Figure 5.7: A simplified write cross section

weak PFET (sNwP) and strong PFET - weak NFET (sPwN). The sPwN/sNwP corners make all the transistors in the circuit stronger or weaker by 3σ. The last two corners provide different supply voltages to the memory cells and the other parts of the design.

Based on the corner study, the transistors that significantly affect the read and write functionality due to variations are identified. We found that the NMOS drive transistors and the NMOS access transistors in the 6T-cell are most critical for read/write operations. As expected, we also observed that with reduction in the supply voltage there is a significant decrease in the functional reliability. The read and write access delay trends are also studied in all the different corners. It was found that a typical design corner with strong P-devices and weak N-devices are highly susceptible to variations, followed by the worst case corner. In addition, the best performance was observed in the corner with strong N-devices.

Figures 5.8 and 5.9 show the comparative analysis of the write access times for two type of write architectures (short bitline and long bitline). As we can see, there are no write failures for a short bitline write up to 3σ variation. The sPwN corner performs worse below $1V$ supply voltage. Long bitline write fails for supply voltages $0.8V$ and $0.9V$ for a worstcase corner, and for $0.8V$ for sPwN corner. Also, the sNwP corner (3σ favorable variation) has the best write access time for both the circuits.

The bitline differential voltage developed during a read operations in all the corners for both the CCIL sense amplifier and gate sense current sense amplifier are shown in Fig 5.10 and 5.11, respectively. The more the differential volatge

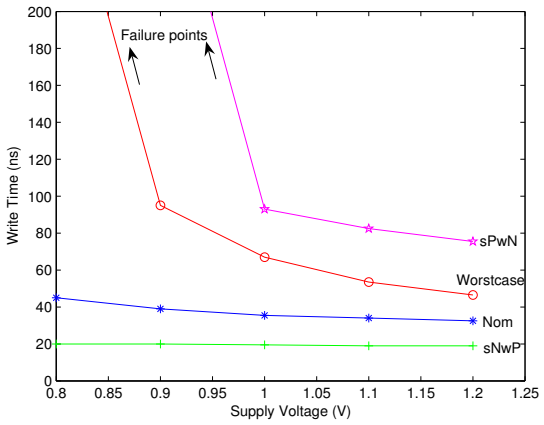

Figure 5.8: Corner Analyses for Write Operation: Long Bitline Subarray

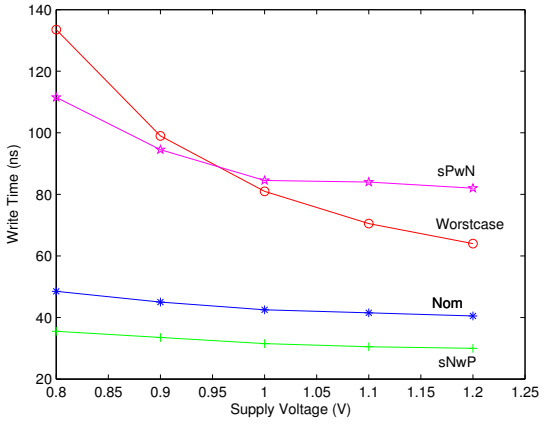

Figure 5.9: Corner Analyses for Write Operation: Short Bitline Subarray

better is the read access time. The analyses again suggests that the NMOS transistors in both the memory cell and the sense amplifier are critical for reliable operation. We also observe that the gate type sense amplifier has worse performance charecteristics as compared to the CCIL sense amplifier, and are also equally susceptible to threshold voltage variations.

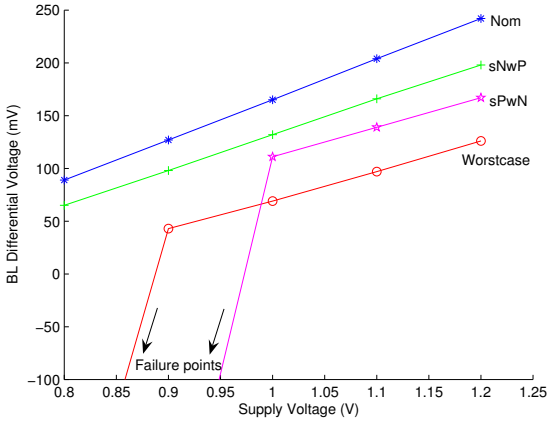

Figure 5.10: Corner Analyses: Precharge High CCIL Sense Amplifier

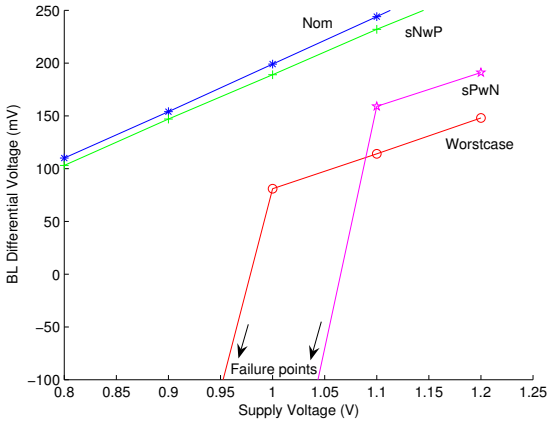

Figure 5.11: Corner Analyses: Gate Sense Amplifier

5.5.2 Specific Variation Analysis

Specific variations were then modeled and applied to the identified functionally critical transistors and random variations ranging from -6σ (stronger) to $+6\sigma$ (weaker) were introduced. Within the memory cell, both the NMOS drive transistor and the access transistor are subject to random threshold voltage variations. The failure probability and the failing point for all the considered circuits are then determined.

Table 5.1 summarizes the observed results for read '0' and a write '0' operations. The write '1' and read '1' operations will have similar failing points and failing probabilities when the symmetric transistors are subject to variations. As seen from the Table, the write operation is more tolerant to threshold voltage variations. This is due to the inherent working principle of the write operation where strong input signals are applied externally to flip the memory contents. The shorter bitline write also has a better variation spread as compared to the longer bitline write. The precharge high CCIL sense amplifier has the highest immunity to threshold voltage variations due to the larger swings in the bitline voltages.

$V_{dd}(V)$	0.8	0.9	1.0	1.1	1.2
Local Bitline Read					
CCIL SA	0.61σ	2.66σ	4.42σ	7.99σ	7.99σ
Gate SA	1.06σ	2.96σ	4.71σ	7.99σ	7.99σ
Mid SA	0.64σ	2.67σ	4.43σ	7.99σ	7.99σ
Local Bitline Write					
Shorter BL	3.52σ	4.15σ	4.72σ	5.16σ	5.57σ
Longer BL	3.20σ	4.25σ	5.17σ	6.32σ	7.99σ

Table 5.1: Failing points for different circuit styles

The simulation data for 1000 random variation in both the NMOS drive transistors and access transistors in the 6T-cell for a write '0' operation in both short and long bitline architectures are shown in Figs 5.12- 5.15. The short bitline write drivers have less failing points and better spread. Even slighlty Weak NMOS drive transitors result in a write failure. Access transistors that are weaker by 5σ or more fails irrespective of the strengths of the drive transistors.

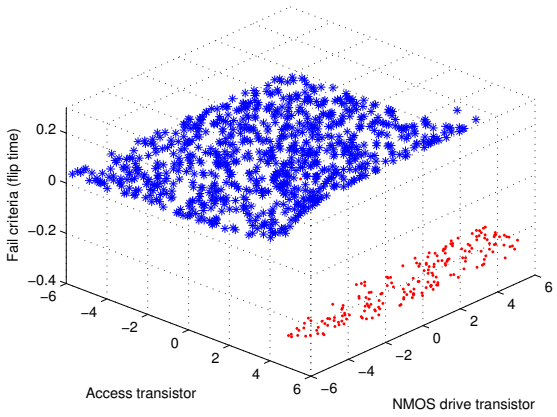

Figure 5.12: 3D Plot for short bitline write

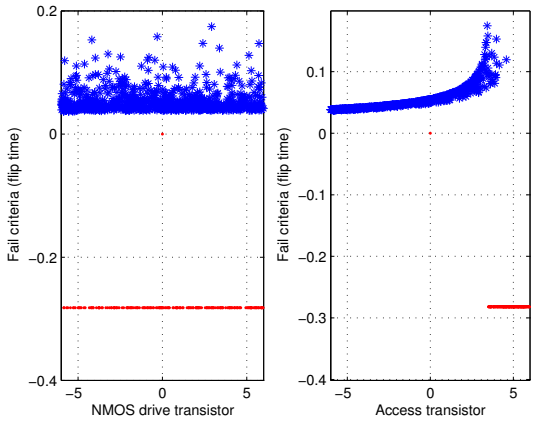

Figure 5.13: 2D Plot for Short Bitline Write

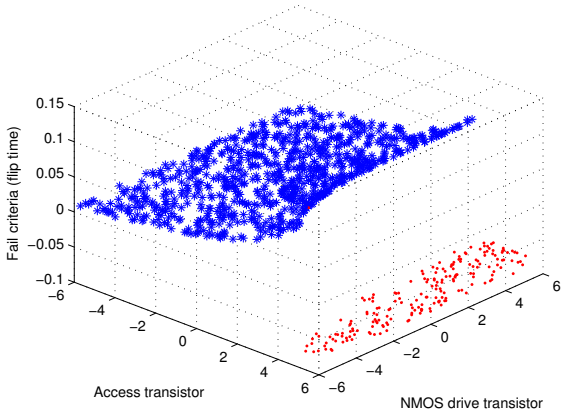

Figure 5.14: 3D Plot for Long Bitline Write

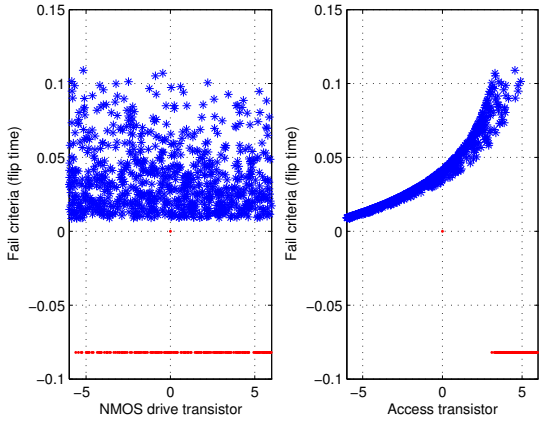

Figure 5.15: 2D Plot for Long Bitline Write

The bitline differential voltage just before the sense amplifier is enabled in a read operation is plotted against the variations in both the NMOS drive and access transistors (Figs 5.16 and 5.17). The developed bitline differential keeps reducing as the access transistors become weaker. This is due to the reduced current drawn by the bitlines. The weak pull down transistor also has a major impact on the bitline differential and hence, on the reliability of teh read operation.

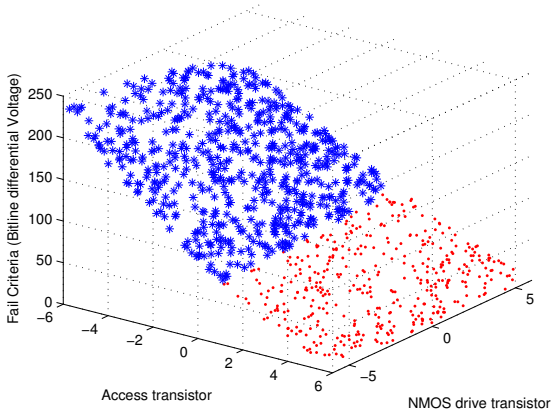

Figure 5.16: 2D plot for precharge SA read

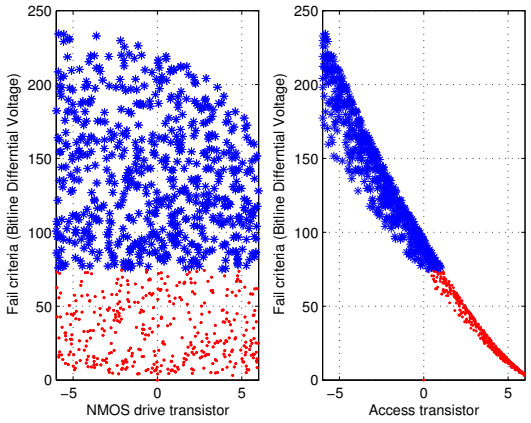

Figure 5.17: 2D plot for precharge SA read

5.6 Summary

Detailed failure analyses to understand the failure mechanisms and trends in the local bitline access schemes of 65*nm* SRAM designs has been performed. Typically, write operation is much more immune to process variations as compared to the read operation. Technology scaling dictates high beta ratios in conventional 6T cells and thus making scaling difficult. Shorter bitline write style has better variation spread and charecteristics as compared to a longer btiline write. Sense amplifier with inherently large bitline swings are much more immune to threshold voltage variations and hence provides better read stability. The NMOS pull down and access transistors in the 6T cell are functionally critical for both the write and read operation. The failure trends and analyses in this study could be used in the early stage of design cycle to decide on the array architecture and read out/write circuit styles to provide better dynamic stability for future memory designs.

Chapter 6

SRAM Reliability: Soft Errors

Future generations of missile guidance systems, interceptors, and smaller, lighter, cheaper spacecraft will require radiation hardened memories with higher densities, lower power operation and higher performance than those available today. In this chapter, we shall discuss the radiation effects on microelectronics and some advanced radiation-hardened memory technologies. The studies discussed here would help in the design of radiation and variation tolerant reliable memories. The following section discusses the problem of radiation in detail.

6.1 Radiation Problems and Environments

In space, radiation from the sun alone can degrade microelectronic circuits and optical components. In addition, nuclear or conventional-weapon induced effects on the battlefield pose direct threat to weapon system operation. This is because the optical, electro-optical componenets and related electronics have the added stress of operating in a cryogenic environment. Hardened microelectronics that

must not only survive but must also continue to function even under highly adverse conditions are required for future computing systems.

For the space and avionics marketplace three types of environments have to be considered: natural space, nuclear weapon and atmospheric. The near earth natural space radiation environment consists of photons, electrons, protons, and heavy ions. The transient space radiation consists of protons and heavy ions from the solar actvity, and galactic cosmic rays (GRCs). The GRCs consists of high energy photons and heavy ions, upto the mass of iron.

The radiation from a nuclear weapon consists primarily of photons (x-rays and γ-rays) and neutrons. This radiation is propotional to the size of the weapon (measured in kilotons, kT) and the radiation flux falls off as the inverse of the distance from the weapon squared ($1/r^2$) for exo-atmospheric bursts. In addition, nuclear weapon bursts in high altitudes releases a high fluence of electrons that can increase the electron flux by a factor of 10 and this effect can last for even months.

At high altitudes, where the magnetic field lines are weak, high energy GRCs can cause a cascade of energetic secondary particles that leads to a high flux of energetic neutrons. These high energy neutrons pose a serios threat to the avionic systems.

6.1.1 Radiation-matter Interaction: A Discussion

The interaction of radiation with matter is a very broad and complex topic. In this section, we try to analyze the problem with the aim of explaining, at least qualitatively, the more important aspects that are essential for a physical comprehension of the degradation observed in MOS devices and circuits when they

are irradiated. The manner in which radiation interacts with solid materials depends on the type, mass, charge and kinetic enery of the of the incident particle, and the mass, density and atomic number of the target material.

The affecting particles can be broadly classified into two groups: charged particles and neutral particles. The charged particles interact mainly through Coulomb attraction or repulsion with the electronic clouds of the target system. Protons, heavy ions and electrons fall under this category. Whereas, neutrons and photons that do not experience the Coulomb force form the neutral particles. Neutrons, which are divided into slow, intermediate and fast, interact with the atomic nuclei through nuclear reactions or elastic/inelastic collisions. Photons interact with matter in three different ways, Photoelectric effect, Compton effect and electron-positron effect.

The effects of both charged and neutral particles on matter can be grouped into two classes: ionization effects and nuclear displacement. These phenomena can be caused directly by the incident particle or from secondary phenomena induced by it. They lead to many irradiated events whose proportions depend on the type of the incident particle.

6.1.2 Radiation Effects in ICs (Memories)

Due to reasons which are beyond the scope of this work, MOS transistors are more sensitive to ionization than to displacement damage. The part of the MOS structure which is most sensitive to ionizing radiation is the silicon dioxide (field oxide).

When a ionizing particle goes through MOS transistors [Fig. 6.1(1)], electron-hole pairs are generated. The electron-hole pairs generated in the gate (metal or

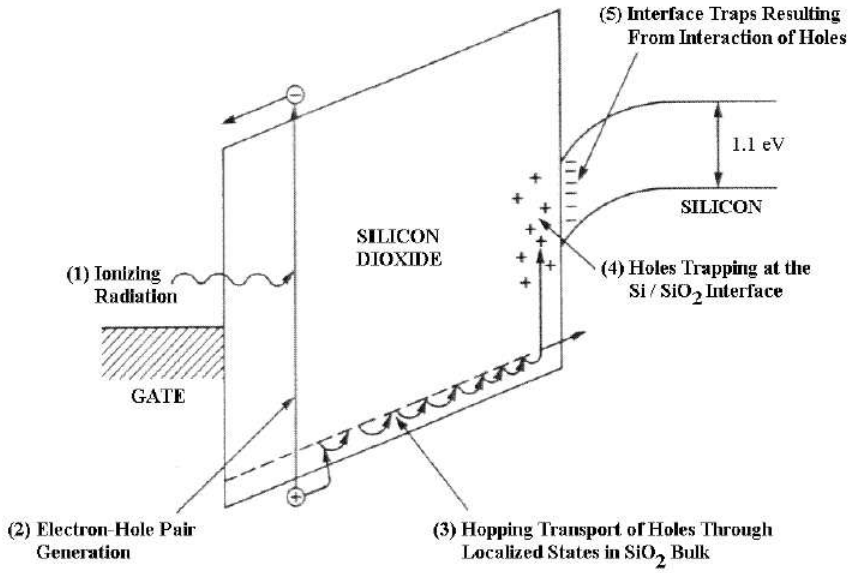

Figure 6.1: Ionizing Radiation Effects in an MOS device with postive gate voltage

polysilicon) and in the substrate disappear quickly due to low resistance. Due to their differing mobilities, electrons and holes behave differently in the insulating oxide layer. A fraction of the radiation-induced electron-hole pairs will recombine immediately after being created. The electron-hole pairs which do not recombine are separated in the oxide by the electric field [Fig. 6.1(2)]. For a positive bias applied to the gate, the electrons drift to the gate in a very short time (order of picoseconds), whereas the holes move towards the $SiO_2 - Si$ interface [Fig. 6.1(3)] with a very characteristic hopping transport phenomenon. Some of the holes may be trapped close to the interface (but remain in the oxide layer) and leads to a fixed positive charge in the oxide layer [Fig. 6.1(4)]. Ionizing radiation also induces the creation of traps at the $SiO_2 - Si$ interface [Fig. 6.1(5)].

This trapped positive charge may anneal out or be compensated over time. As

the holes are transported, they release hydrogen in the oxide that may transport to the $SiO_2 - Si$ interface. As described above, ionizing radiation in MOS produces a trap positive charge and interface traps. The consequences of these radiation effects on the electrical parameters of a MOS transistor are described briefly in the following subsection.

6.1.3 Impacts of Radiation on MOS Transistors

The primary concern for digital CMOS engineers in a space environment is the effect of the positive charge in the isolation of field oxide and the interface traps in the gate oxide. In an n-channel MOSFET, the positive charge may deplete or invert the p-type body and create a leakage path between the source and the drain. In addition, it may also the deplete the p-type region under the field oxide and cause transistor-transistor leakage. The interface traps can degrade the channel mobility reducing the current drive and/or the switching time. The threshold voltage of the MOS transistor also changes when the device is irradiated. Apart from the above effects, radiation may also increase the parasitic and subthrehold currents and decrease the transconductance. Thus, radiation effects pose serious power and performance threats to the MOS transistors.

In the follwing sections, we shall discuss the different memory technologies available that cope with radiation effects in microelectronic circuits. In general, there is some confusion in the space microelectronics marketplace over the terms *radiation hardened* and *radiation tolerant* microelectronics. To overcome this radiation level distinction, researchers have categorized the microelectronics parts into radiations *soft*, *tolerant*, and *hard*.

6.2 Soft Errors in SRAMs: Background and Related Work

Due to higher operating frequencies and lower supply voltages of the today's computing systems, radiation-induced soft errors are of increasing concern in future memory and combinational circuits. Soft errors or transient errors are circuit errors caused due to excess charge carriers induced primarily by external radiation. Radiation directly or indirectly induces a localized ionization capable of upsetting internal data states. While these errors cause an upset event, the circuit itself is not damaged. Further, as the size of the transistor scales into the nanometer domain, circuits are becoming increasingly susceptible to operational disturbances caused by fluctuations in the surrounding environment. Designers routinely use well-known techniques, such as, error detection and correction (ECCs) to cope with soft errors in static memories. However, given the soft error rates and customer expectations in the nanometer domain, protecting just the memory cells may not be sufficient. Effects of soft errors in sequential (flip-flops and latches) and combinational logic must be evaluated and subsequently, effective low power protection mechanisms must be incorporated into the overall SRAM system design.

Typically, both sequential and combinational blocks encode information in the form of a charge stored on a circuit node or as current flowing between any two circuit nodes. Any event which upsets this stored or communicated charge, such as, cosmic ray or alpha particle radiation, can cause erroneous circuit outputs. Due to their nonpermanent, nonrecurring nature, these errors are called soft errors. These radiation induced soft errors on electronic devices have been in existence for a long time now. In 1962, Wallmark [57] pointed out that if the

channel length scales down beyond $1um$, a single cosmic ray particle strike would short-circuit the source and drain terminals of a transistor in the off state and potentially disrupt the circuit. In addition, technology scaling roughly leads to a doubling of clock frequencies, a 30% reduction in supply voltages to reduce power consumption and also a 30% decrease in node capacitances every generation [58]. Due to differing reasons, all these factors result in a significant increase in soft error susceptibility of combinational and memory circuits.

Memories are considered most vulnerable to radiation induced transient errors due to the high density and amount of information they store. Further, these errors are particularly troublesome for memory elements as the stored values of the bits are changed. The frequency of soft errors in SRAMs is becoming a critical issues as technology continues to scale [59, 60, 61]. Specifically soft errors in SRAM memories can be catastrophic in networking applications as a bit flip can result in information packets such as money transfers sent to the wrong account. Trends such as smaller supply voltages and reduced capacitive values at the nodes are potential concerns for the memory cell's susceptibility to soft errors.

The first reports of failures attributed to cosmic rays emerged in 1975 when space borne electronics malfunctioned during a magnetically quiet time, and it was unlikely that these failures were due to spacecraft charging [62]. In 1978, similar problems were observed in dynamic memories at ground level [63]. Alpha particles, high energy neutrons and slow neutrons are the primary sources for these soft errors. When a particle strikes a PN junction, ionization occurs and causes electron-hole pairs which correspond to deposition of a finite amount of charge. The particle loses its energy as it passes through the semiconductor

and this loss in energy is measured in terms of Linear Energy Transfer (*LET*). The recombination of such electron-hole pairs causes a current transient which could disrupt the logic state of a circuit *i.e.* flip the value stored at that junction. This would happen if the charge deposited is greater than the minimum required charge ($Q_{critical}$) for a flip to occur, or equivalently, the energy lost by the particle is greater than the minimum energy that the semiconductor must absorb for a flip to occur (*LET* threshold). Such an event where a particle hit causes the value stored on a junction to flip is called a Single Event Upset (SEU). A comprehensive explanation of this phenomenon can be found in [64].

When such upsets occur at nodes that are part of a regenerative loop (memory cells and latches), the value stored by that cell or latch flips instantly and a soft error occurs. On the other hand, when such upsets occur at combinational nodes, a voltage transient occurs which is frequently called a Single Event Transient (SET). The transient may propagate through the combinational stages and eventually be latched by a sequential element if it arrives at its input during its window of vulnerability. It is only then that this transient causes a soft error. The rate at which SETs get latched as errors depends on the operating frequencies. As mentioned earlier, with rapid technology scaling, the frequencies at which circuits are operated is continuously increasing and thus the probability of SETs getting latched as errors is also increasing. In addition, SETs and SEUs are considered major challenges for low power and high performance microprocessor design [65]. Thus, there is a need for an efficient design of SET and SEU tolerant memory and sequential elements, and many researchers have emphasized its importance [59, 66, 67, 68, 69].

6.2.1 Soft Error Reduction Techniques

One popular approach to reduce the soft error rate is to use pure device material and shield the sensitive circuit from ionizing particles. However, such solutions are generally not effective for the highly penetrative neutron rays besides the additional cost. Yet another well known method to mitigate the soft error rates is to use either space or time redundancy techniques. In the former, the circuit or the block is duplicated in space to provide immunity to soft errors. Whereas, time redundancy techniques are based on the fact that a particle strike does not happen in successive time units and thus reduce soft errors by sampling the outputs at different time instants. The third approach is to exploit the dependence between Q_{crit} and node capacitance and provide soft error protection by increasing the node capacitance. However, care should be taken in increasing the right kind of capacitance. The gate or interconnect capacitance do not affect the charge collecting efficiency. However, by adding more diffusion capacitance, we increase the total diffusion area at the node, which results in an increase of the charge collecting efficiency during a strike. Hence, this could offset the benefits of the increased node capacitance on Q_{crit}. Further, there is a tradeoff between the gain in roubustness and loss in SRAM access times.

6.3 Soft Error Metrics

For a soft error to occur at a specific node in a circuit, the collected charge Q at that particular node should be more than Q_{crit}. If the charge generated by a particle strike at a node generates a charge that is more than Q_{crit}, the generated erroneous pulse is latched on, and results in a bit flip. This concept of critical

charge is generally used to estimate the sensitivity of SER. Hazucha [70] developed a method which models an exponential dependence of SER on critical charge for CMOS SRAM. It is shown as follows:

$$SER \propto N_{flux} * CS * e^{\left(\frac{-Q_{crit}}{Q_S}\right)} \tag{6.1}$$

where N_{flux} is the intensity of the Neutron Flux, CS is the area of crosssection of the Node and Q_S is the charge collection efficiency. Q_{crit} is proportional to the node capacitance and the supply voltage. The Q_{crit} at a node will decrease as voltage or node capacitance reduces.

The nodal capacitance is strongly dependent of how the layout of the design is done. Some well designed layouts offer better immunity against SER than others. An intuitive approach to reduce the vulnerability of a node to soft error is to increase the nodal capacitance. However, clear distinction has to be drawn between the different capacitances. The capacitance from gates or interconnect provide robustness to soft errors, whereas, the same cannot be said about the diffusion capacitance. Adding gate or interconnect capacitance has little impact on the charge collection process during a soft error event and thus increases the robustness of the node to radiation induced soft errors.

However, adding diffusion capacitance increases the total diffusion area at the node, resulting in an increase of the charge collecting efficiency during a strike. Hence, this can offset the benefits of the increased nodal capacitance on Q_{crit}. The value of Q_{crit} can be found by measuring the current required to flip a memory cell and can be derived using the following equation.

$$Q_{crit} = \int_0^{T_f} I_d dt \tag{6.2}$$

I_d is the drain current induced by the radiation, T_f is the flipping time and is defined as the point in time when the feedback mechanism of the cross-coupled inverters in the memory cell will take over from current of the incident ion. In this work, we focus primarily on Q_{crit} in copmaring the SER of different memory cell designs. The charge collection efficiency Q_S is primarily dependent on the doping profile and so is not influenced much by the designs. Further, the cross section of the vulnerable nodes remain the same in all the designs we considered.

6.4 SER Analysis of Standard 6T SRAM Cell

In this section, we present the results of the soft error analysis we did for different SRAM designs considered. We did the soft error susceptibility analysis on the standard 6T-cell [11], loadless 4T-cell [71], resistive load 4T-cell, DRG cache [7] and our NC-SRAM design [72]. All the above SRAM designs were custom designed and simulated using Spectre using 70nm Berkeley Predictive Technology models [22]. For the NC-SRAM cell, the threshold voltage of the access transistors and pass transistors were increased by modifying the v_{th0} parameter in the model files to make the bitline leakage negligible.

For measuring Q_{crit}, the particle strikes were modeled using a piece wise linear current pulse to account for funneling and diffusion charge collection [59]. The current pulse was injected at the node and measured up to a point where the regenerative nature of the memory cell takes over and commits the bit flip. Finally the current pulse causing the bit flip was integrated to get the critical charge of

the node as shown in the earlier section.

6.4.1 Simulation Results and Observations

The node capacitance values for different memory cell designs are presented in Table 6.1. We could observe that the 4T-cell (both without V_{dd} and with resistive load) has much less capacitance as compared to the other designs. Due to this they would have increased susceptibility to radiation induced particle strikes. The pull-up resistors in the 4T-cell were implemented using back-biased PMOS.

SRAM Type	Nodal Capacitance
6T-cell	16.5
4T-cell (no V_{dd})	4.5
4T-cell (resistive load)	7.6
DRG-cache	16.3
NC-SRAM	16.4

Table 6.1: Node Capacitances for Different SRAM Designs

Observing the critical charge values for a 1 to 0 flip for different SRAM designs, we can see that the 4T-cell without V_{dd} has very low critical charge and hence it is highly susceptible to radiation strikes. Due to the absence of the supply voltage in the 4T-cell, the internal node decays naturally with time. In the presence of a transient pulse disturbance, the node discharges faster and loses the stored data. Further, as the value of the stored 1 is already decaying to a 0, a cosmic particle strike will only accelerate the decay and hence, we see very low values for the critical charge. In addition, due to the absence of the restoring power of the cross coupled inverters and a very low nodal capacitance, both the 4T-cell designs are highly susceptible to soft errors.

SRAM Type	$Q_{critical}$ (fC)
6T-cell	15.02
4T-cell (no V_{dd})	$6.2 * 10^{-}5$
4T-cell (resistive load)	5.625
DRG-cache	11.3
NC-SRAM	8.76

Table 6.2: Critical Charge Values of Different SRAMs for 1 to 0 flips

In the gated-ground DRG cache, the array is shut off completely from the ground supply during the sleep state. Therefore, during the stand by state, the virtual ground node does not stay at 0V but charges up to a small positive voltage (say, 0.2-0.4V). This makes the DRG cache design more vulnerable to a bit flip as compared to the standard 6T-cell since a smaller induced charge is enough to trigger the flip. Whereas, our NC-SRAM design due to its reduced voltage across the supply rails in the low leakage mode makes it more susceptible to soft errors as compared to a conventional cell. However, one advantage with this design is that we could trade-off the attained leakage savings to increased soft error tolerance by adjusting the small positive voltage given to the pass transistors.

SRAM Type	$Q_{critical}$ (fC)
6T-cell	320.560
4T-cell (no V_{dd})	23.125
4T-cell (resistive load)	60.260
DRG-cache	130.45
NC-SRAM	86.62

Table 6.3: Critical Charge Values of Different SRAMs for 0 to 1 flips

Table 6.3 shows the critical charge values for a 0 to 1 bit flip. We could clearly observe that the values in Table 6.2 are significantly lower than the values in

Table 6.3. This could be attributed to the stronger influence of N+ diffusion (as compared to P+) in these circuits and hence a 1 to 0 flip is more likely than a 0 to 1 flip. Typically, the memory cells have a stronger pull down for stability reasons and they are more immune to bit flips when storing a 0. This observation also agrees with the soft error rates observed for static latches by Karnik [61]. We can also note the difference in the critical charge values for the 4T cells. In contrast to the natural decay process of the 1 to 0 flip in 4T cells, this flip is against the normal decay. This explains the higher Q_{crit} values for the 4T cells.

6.5 SOI Memories for Soft Error Reduction

In this section, we discuss the technology related approaches to provide immunity to soft errors. The current "power crisis" in the ULSI chips combined with the portable system boom is leading the semiconductor industry to Silicon On Insulator (SOI) technology, as an alternative to conventional silicon technology. Silicon-On-Insulator (SOI) is a semiconductor fabrication technique developed by IBM that uses pure crystal silicon and silicon oxide for integrated circuits (ICs) and microchips. An SOI microchip processing speed is often 30% faster than today's complementary metal-oxide semiconductor (CMOS)-based chips and power consumption is reduced 80%, which makes them ideal for mobile devices.

SOI technology has many advantages over the bulk technologies, such as, total device isolation, speed and density [73]. Furthermore, by limiting the charge collection volume with the buried oxide layer of the SOI system, SOI technologies are tolerant to radiation-induced latch-up and single-event upset phenomena,

making them appealing for space applications [74]. One of the first early applications of SOI has been in memories for space application, since the memories built on SOI were perceived to be more resistant to radiation effects. Soft error rate (SER) refers to upset of data in the memory cells by cosmic rays and background radioactive material. SOI have proven benefits in the reduction of soft-error rate.

SOI refers to placing a thin layer of silicon on top of an insulator such as silicon oxide or glass (see Fig. 6.2. The transistors would then be built on top of this thin layer of SOI [75]. The basic idea of SOI is to reduce the capacitance of the switch for faster operation.

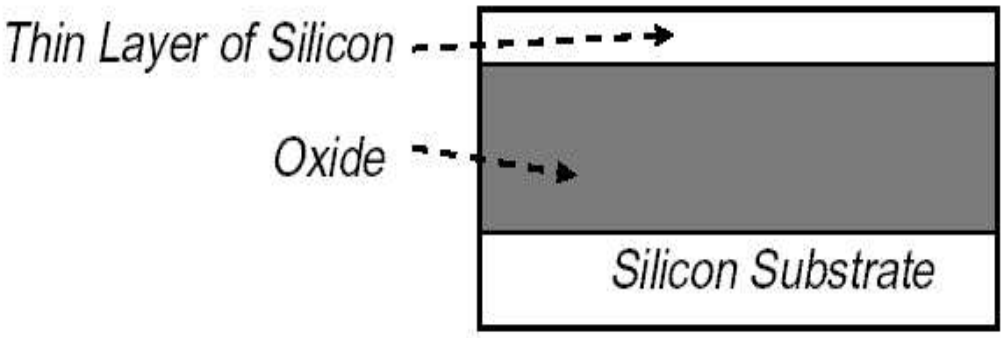

Figure 6.2: SOI Technology

Junction capacitance, which is the area between the impurities in the substrate and the silicon substrate itself, stores charge in a MOS transistor. In order to control the electrical currents needed, the junction capacitance must be discharged and recharged, which takes time. In addition, this also causes the transistors on the chip to heat up. This production of heat limits the speed at which microchips can operate. If a thin layer of an insulator, such as glass, is placed

between the impurities and the silicon substrate, the junction capacitance will be eliminated and the MOS transistor will operate faster due to reduced capacitance. This concept is illustrated in Fig. 6.3.

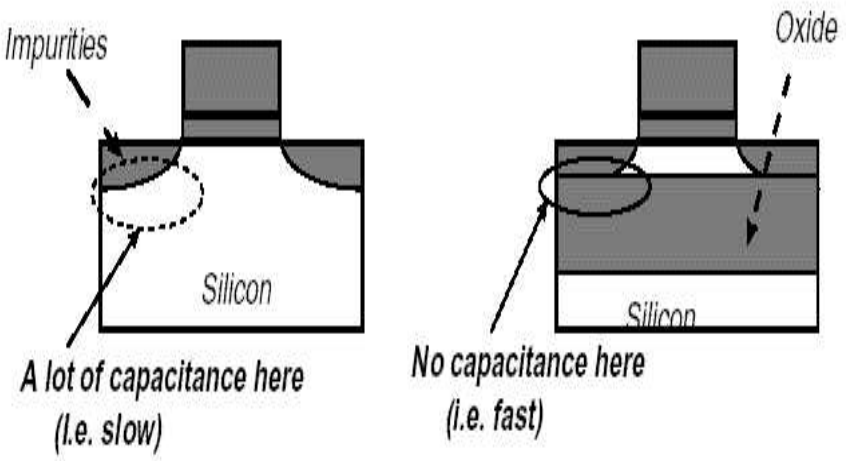

Figure 6.3: Reduced Capacitance in SOI Systems

It is observed that the SOI technology improves performance over bulk CMOS technology by $25-35\%$, equivalent to two years of bulk CMOS advances [75]. In addition, SOI technology also brings power use advantages of $1.7-3$ times. It is expected that SOI will eventually replace bulk CMOS as the most commonly used substrate for advanced CMOS in mainstream memories, microprocessors and other emerging wireless electronic devices requiring low power, high speed and radiation hardening for space applications.

6.6 Summary

In this chapter, we introduced the problem of radiation effects and mentioned the different radiation environments. The impact of radiation effects on memories were discussed in detail. We presented the background and related work on soft errors in SRAM designs. We also discussed the different soft error metrics and reduction strategies before analyzing the impact of raditation errors on different SRAM cell configurations. Finally, we conclude the chapter with the discussions on SOI technology and its importance in designing radiation hardened memories.

Chapter 7

Technology Charecterization: Test Structures

7.1 CMOS Process Tuning and Variability Control

In this chapter, we describe the importance of test structures for charecterizing a particular technology node and then present our SRAM ring macro test structures. These ring macros were developed using an industry standard aggressive process technology and thin cell layout SRAM bitcells. A brief description of the test structure aimed at monitoring variability in the manufacturing line as well as in the product is given. The design techniques described here allow very rapid investigation of the sources of variation in circuit delays.

With advances in silicon CMOS technology and scaling of MOSFET channel lengths to 90 nm and below, process induced variations in circuit delays have begun to significantly impact product performance and power. Variations in circuit delays and leakage power of nominally identical structures may occur locally, across chip, chip reticle, across wafer, from wafer-to-wafer and from lot-to-lot. In

addition, tracking amongst different circuit topologies may also vary in a similar fashion. In a silicon manufacturing line, dc characteristics of single MOSFETs and other structures are monitored on a limited number of sites on a few selected wafers in each lot for tracking and process tuning. These test structures are placed in the scribe line and do not adequately capture all the variations or provide comprehensive insight into their sources.

With the scaling of CMOS and high performance products operating in the 1 to 5 GHz range, it has become increasingly important to rigorously monitor the ac performance of MOSFETs during processing, especially early in the process [76]. This substantially reduces time and cost of optimizing the process and device design for the product. Manufacturing lines employ standard parametric testers for monitoring dc characteristics of MOSFETs and other circuit elements, as well as ring oscillators. Only limited in-line ac testing is conducted using logic or memory testers. The ac functionality tests typically require scan-able latches or registers to stream the test patterns in and out, are complex in design, use at least three to five levels of metal and are limited to a few hundred MHz frequency range. In rare cases, high speed in-line bench tests may also be done.

7.2 Test Structure Methodology and Design

There are two popular kinds of test structures [76]. The first extends the traditional use of ring oscillators from performance measurements on specific gate types to estimating various MOSFET parameters,, wire and parasitic capacitances, layout dependencies and leakage current components [71]. All of these parameters are selfconsistently derived from the same set of circuits operating in

the same frequency range as the products. This is especially critical for MOSFETs in partially depleted (PD) SOI, where floating body and self-heating effects result in significantly different ac and dc behavior.

The second flavor of test structures is aimed at replacing complex high speed bench tests. Such bench tests require a proliferation of test equipment such as pulse generators, sampling scopes, spectrum analyzers, delay lines, and special probes, and may require careful calibration and data analysis [77]. Tests of this kind are generally conducted off the manufacturing floor on an infrequent basis, such as for model-to-hardware correlation, even though ongoing knowledge of the results of such tests would have a strong influence on technology design and development. To be testable with a parametric tester it is essential that these structures be self-timed, selfcalibrating, and fully functional with only dc I/Os. We will discuss the general features of the design of the first flavor of structure and give one specific example that uses SRAM structures to form the ring.

7.3 SRAM Ring Oscillator Macro

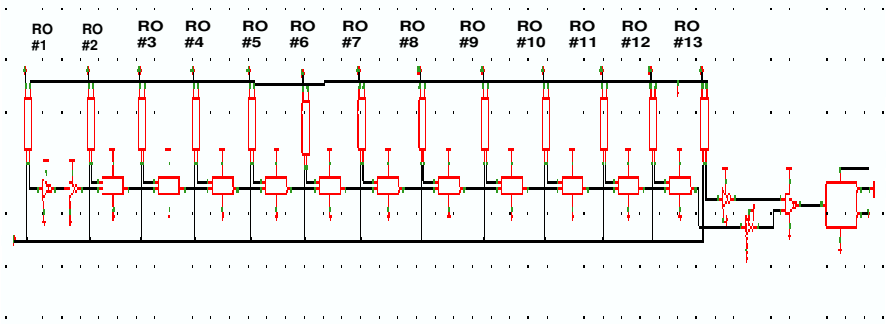

Figure 7.1: SRAM Ring with a NAND2 Gate for Enabling the Oscillations

The ring oscillator design of Fig 7.1 has long been in use for measuring average

gate delay. It has the desirable feature of requiring a dc input to enable the ring and a low frequency output from a divide-by circuit. With an input decoder and a multiplexer in the output, multiple rings can share I/Os. In our macros, we used different circuit configurations from the SRAM 6T bitcell to construct the ring structure. The macro contains experiments to study history in the write operation and read operation of a POR SRAM cell. It also contains an inverting SRAM cell experiment, inverter experiments and inverter plus passgate experiments (the latter two using the inverter and inverter plus passgate from the SRAM cell). In addition, there are several delay chain experiments (SRAM inverter, SRAM inverter plus passgate and inverting SRAM cell) that correspond to the traditional ring osicllator experiments and two capacitor history experiments that relate to SRAM cell function. All these test structures use a SRAM thin cell layout for the base design and the required circuit configuration is formed by splitting the active R_X layer in the layout while maintaining the straight polysilicon orientation.

7.3.1 Circuit Description

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
VDD	SEL	VR1	GND	VR2	VR3	GND	VR4	VR5	GND	VR6	VR7	GND	VR8	VR9	GND	VR10	VR11	GND	VR12	VR13	GND	OUT	VDD	BI

Figure 7.2: IO pad assignments in SRAM ring macro. The pad no. and electrical charecteristics are shown in the top and bottom rows respectively.

This SRAM ring oscillator macro is designed for model-to-hardware correlation

of delays, $IDDQ$, C_{gate}, C_j/C_{ov} and C_{diff} and is placed on a industry standard test-site. The macro comprises 13 SRAM ring oscillators (ROs) with independent V_{DD}s, an output multiplexor and a frequency divider circuit. There are twenty five I/O pads, and their relative location and assignments are shown in Fig. 7.2. An RO is selected by setting its V_{DD} = "1", remaining twelve RO's with V_{DD} ="0". Setting SEL="1" enables the oscillations in the selected ring. The output from all the RO's is multiplexed together and fed to a divide by 256 circuit. The resulting output (OUT) oscillates at a frequency of 1 MHz which can be directly measured with a frequency counter. There are 14 power supply sectors, VR1 to VR13 independently supply power to 13 SRAM rings, and V_{DD} for the control circuits (decoder, multiplexor, frequency dividers, buffers and I/O driver) is on pad 1 and 24.

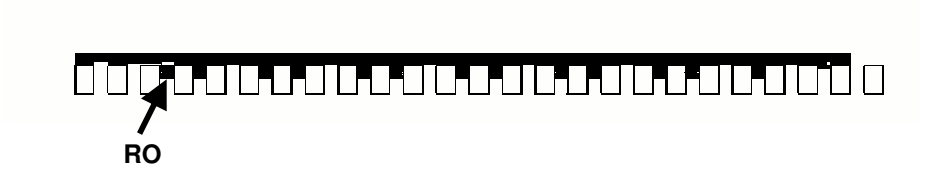

Figure 7.3: Physical layout of the macro with 13 SRAM rings

The physical layout of the designed SRAM macro is shown in Fig. 7.3. The dimensions of the macro are 2500 μm x 140 μm on wafer. The SRAM rings are placed between the V_{DD} and GND pads. Figure 7.4 shows the physical layout of one of the thirteen rings with inter locked supply and ground rails. As mentioned, there are thirteen such rings in the entire SRAM ring macro.

Figure 7.5 shows the circuit schematic of the SRAM ring with 100 identical stages and a NAND2 gate to enable the oscillations. The SEL signal is an input to the ring stage and has to be high for this ring to start oscillating. At any instant,

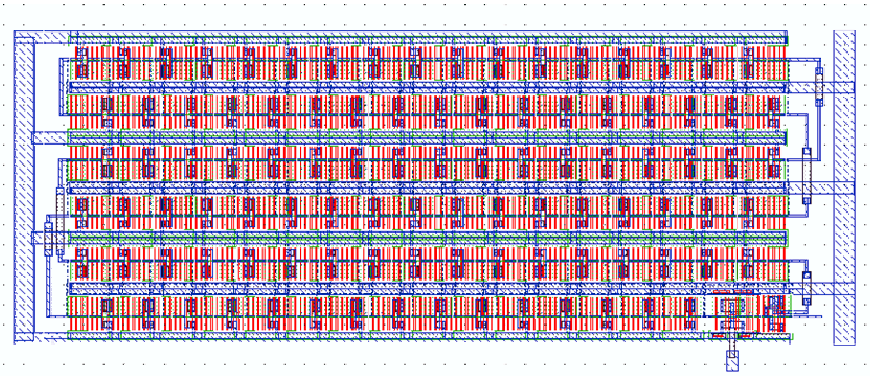

Figure 7.4: Physical layout of a 100 stage SRAM ring macro with thincell base stages

only one of the 13 rings is turned on and this is ensured by the SEL signal. The output of the 100 stage ring is fed to a two input NAND gate before going to the final output. All stages in the 100 stage ring are identical and vary depending on the circuit charecteristic.

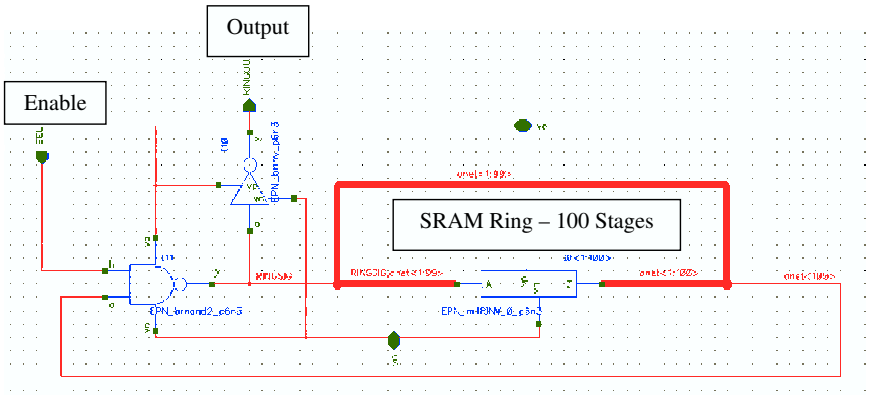

Figure 7.5: Circuit schematic of SRAM ring with 100 identical stages

The schematic of a few circuit configurations implemented are shown in

Fig. 7.6 and Fig. 7.7. They have different gate loads and are all designed maintaining the SRAM cell structure (including straight poly orientation) for lithography purposes. The structures are expected to charecterize the different variations present and will help in analyzing them earlier in the design cycle. It also provides a mechanism for yield analysis and improvement in the development stage.

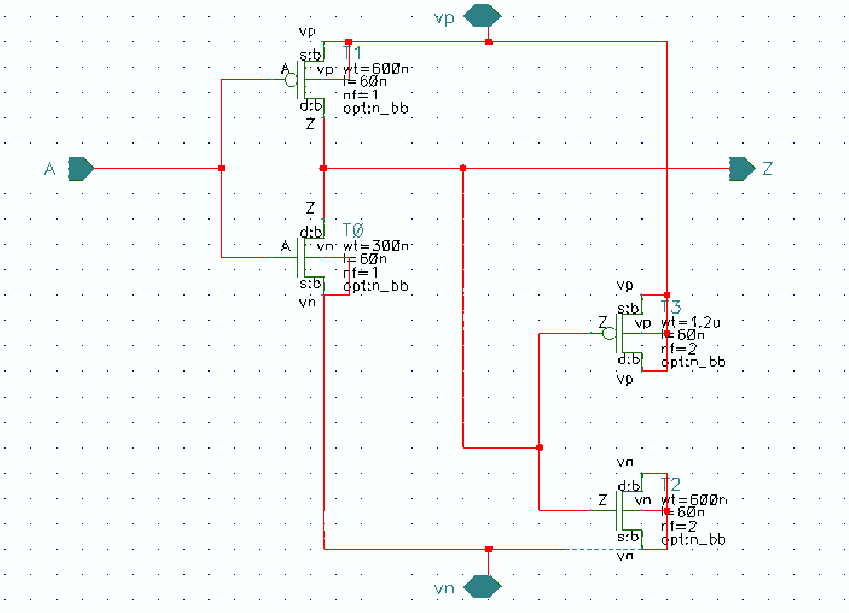

Figure 7.6: Circuit Schematic of a single inverter stage

The above described SRAM experiments will provide a unique view into the history and floating body effects of SRAM cells in SOI technology. Together with the SRAM ring macros, tt would help in better understanding of the key aspects of the performance of the SRAM cells in this technology.

7.3.2 Measurements and Data Analysis

The results are analyzed using the hardware data on several experimenatal lots for the same technology. Each point of the hardware data is randomly selected

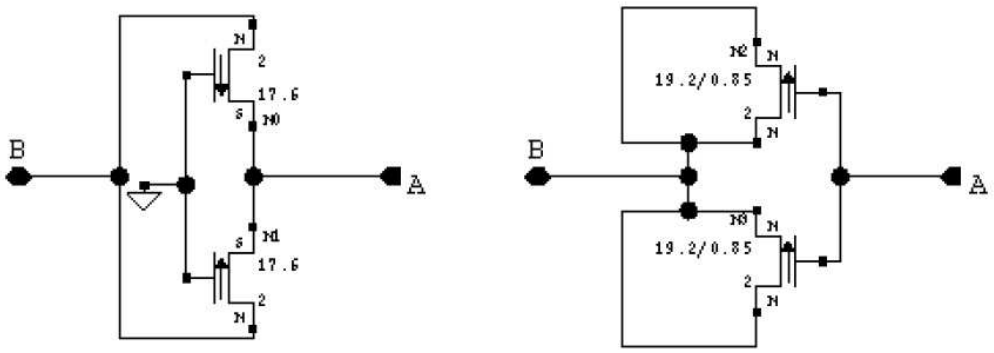

Figure 7.7: Two different NFET capacitor configuration for SRAMCAPS experiments

from measurements made on a single reticle location on specific sites on a wafer. The hardware data captures the trends shown by the models although no attempt is made to do a true model-to-hardware corelation. In a manufacturing line, this type of analysis and offsets between model and hardware provide direct information for process and device monitoring and tuning on an on-going basis. For a more detailed understanding of the MOSFET behavior, standard dc characterization is carried out and correlated to the ac performance.

7.4 Summary

In this chapter, we point of the importance of test structures in charecterizing a technology well before it becomes mature. The SRAM ring structures described in this chapter provide a rapid means of evaluating the effect of variations in key MODFET paramters on switching delays and both active and leakage power of CMOS circuits. In addition to providing basic performance data, the ring oscillators are used to self-consistently extract a wide range of MOSFET

device and circuit parameters. They are also used to measure interconnect resistances and capacitances, to measure optical proximity effects, and to evaluate power/performance trade-offs in product representative paths. Using the SRAM ring design methodology outlined in this chapter, the sources of systematic components of variability in circuit delays across product, across wafer and across hardware vintage as well as with different physical layout styles can be readily identified. These designs could be placed in the scribe line for detailed analysis of MOSFETs as well as being integrated into the product itself.

Chapter 8

Conclusions and Future Work

8.1 Major Contributions

As VLSI technologies shift towards nanometer feature sizes, with interconnect delays and leakage power dominating gate delays and dynamic power, respectively, design of low power and high performance systems face many challenges. One of the key concerns of high speed memory system design is achieving very low cycle times and reducing leakage power consumption. We have not only addressed the design approaches and circuit techniques for low power high speed SRAMs but also provided a detailed failure analysis study to improve the reliability of nanometer SRAMs and predict their yield in an early design stage. The motivation is to provide cost-effective and practical design solutions to develop ultra low power memory systems. The major contribution of this dissertation are as follows:

1. We have explored a circuit-level technique for reducing the leakage power in deep submicron caches. Comparison of leakage power savings in other

contemporary cache designs with the proposed NC-SRAM design was performed. Based on the simulation results, we found that the proposed memory design has key advantages over the other existing techniques for reducing leakage in SRAM circuits. One of the notable features of this work is that the proposed design achieves large leakage power savings and at the same time retains the data stored in the memory cell with no additional overheads. We evaluated and presented simulation results from implementing the design in different technologies using a dual-V_t approach. The results indicated that NC-SRAM almost eliminates leakage with the right type of partitioning and yielded a leakage reduction between $45\% - 70\%$ depending on the control voltages used. We also simulated the proposed design in $100nm$ and $70nm$ technologies to study the impact of technology scaling and achieved promising results in terms of leakage power savings.

2. We have proposed two memory cell designs, RG-SRAM and DG-SRAM, to suppress the gate leakage current. In very deep submicron technology with feature sizes less than 65nm and with low oxide thickness, gate leakage has become the dominant source of leakage and is expected to increase with technology scaling. In the above two designs, we used two additional PMOS/NMOS devices which change the gate voltages of the transistors forming the inverter latch in SRAM to reduce the gate leakage current. One of the notable features of the proposed work is that it achieves significant leakage savings irrespective of the state of the cell and the value stored in the cell. Simulation results show 66.5% reduction in total leakage at 65nm technology with T_{ox} at $1.1nm$ with only around 2.86% degradation in discharge time

for NMOS type DG-SRAM.

3. We have investigated the implementation of high throughput wave-pipelined address decoders. Wave pipelining increases the throughput without additional storage space in the form of registers thereby reducing clock distribution overheads. A wave pipelined $6:64$ address decoder was designed in both TSMC $0.18\mu m$ and IBM SiGe $0.25\mu m$ BiCMOS technologies. The charecteristics of wave pipelined address decoders, such as, equal rise and fall delays, and minimum data dependent path delay variations were exploited to increase the performance of the overall memory system.

4. We have proposed two novel robust high performance current mode sense amplifier with improved power consumption for nanoscale SRAM Memories is presented. One sense amplifier uses a winner take all approach to provide fast amplification while the other design, based on a cross coupled latch, focusses on low power operation. The WTA sense amplifier is highly robust to mismatch in threshold voltage and tolerates upto 10% variation in V_t with acceptable degradation in the sensing delay, while LPCSA tolerates upto 8% variations. Simulation results show that our designs are also tolerant to variations in the effective channel length and supply voltage. WTA offers around 70-80% speed improvement and the LPCSA around 12-32%, when compared to other voltage and current mode sense amplifiers. Such large improvements are possible due to the inherent design and amplification mechanism of the sense amplifier design. In addition, unlike other current sensing techniques, we do not have excessive bitline swings or additional circuitry in the amplification stage. Consequently, this results in

significant speed improvement and tolerance to process variations. Since it does not precharge/predischarge the output nodes to V_{dd}/ground, LPCSA also consumes the least power among the sense amplifiers considered. Thus, the performance of both WTA and LPCSA is least affected, in terms of both sensing speed and energy consumption, in the presence of increasing bitline capacitance and process variations.

5. We have performed detailed failure analyses to study the impact of process induced variations and to understand the failure mechanisms and trends in the local bitline access schemes of $65nm$ SRAM designs. Typically, write operation is much more immune to process variations as compared to the read operation. Technology scaling dictates high beta ratios in conventional 6T cells and thus making scaling difficult. Shorter bitline write style has better variation spread and charecteristics as compared to a longer btiline write. Sense amplifier with inherently large bitline swings are much more immune to threshold voltage variations and hence provides better read stability. The NMOS pull down and access transistors in the 6T cell are functionally critical for both the write and read operation. The failure trends and analyses in this study could be used in the early stage of design cycle to decide on the array architecture and read out/write circuit styles to provide better dynamic stability for future memory designs.

6. We have analyzed in detail the problem of radiation effects and mentioned the different radiation environments. The impact of radiation effects on memories were discussed in detail. We presented the background and related work on soft errors in SRAM designs. We also discussed the different

soft error metrics and reduction strategies before analyzing the impact of raditation errors on different SRAM cell configurations. Finally, we presented the details of designing memories using Silicon on Insulator (SOI) technology and their importance in designing radiation hardened memories.

7. We have studied and designed test structures for charecterizing a technology well before it becomes mature. The SRAM ring structures described in this dissertation provide a rapid means of evaluating the effect of variations in key MODFET paramters on switching delays and both active and leakage power of CMOS circuits. In addition to providing basic performance data, the ring oscillators are used to self-consistently extract a wide range of MOSFET device and circuit parameters. They are also used to measure interconnect resistances and capacitances, to measure optical proximity effects, and to evaluate power/performance trade-offs in product representative paths. Using the SRAM ring design methodology outlined in this chapter, the sources of systematic components of variability in circuit delays across product, across wafer and across hardware vintage as well as with different physical layout styles can be readily identified. These designs could be placed in the scribe line for detailed analysis of MOSFETs as well as being integrated into the product itself.

8.2 Directions of Future Research

Our research provides interesting opportunities for developing design methodologies, algorithms, techniques and tools for the design, test and modeling of nanometer SRAMs. Although a significant amount of research work has been

accomplished in the area of low power nanoscale memory design, new and intriquing unanswered questions will still remain. In the future, we intend not only to enhance the applicability and usefulness of the initial results, but also to extend our research into related areas to find novel applications for the theoritical development. In this section, we shall list a few extensions to this work and directions for future research. More specifically:

- Investigate the issues that would arise when attempting to integrate existing dynamic power reduction techniques with the proposed low leakage power memory cell designs.

- For achieving maximum leakage power savings, only the accessed cell needs to operate in the normal mode as opposed to all the cells in a block or a row. This could be achieved by using both row and column decoder to control the gates of the pass transistors of the NC-SRAM cell. A power X area tradeoff study is in progress to analyze the feasibility of this approach. In addition, improvements to the present design to enable leakage savings in the active mode are also being performed.

- Detailed analysis for the practical implementation of the proposed memory designs need to be explored. This includes analyzing the stability issues for the bitcells, obtaining measurement results by fabricating the designs in aggressive nanometer technologies and performing monte carlo analysis to study the robustness of the proposed designs. More research is needed to investigate the applicability of these designs in an industry standard commerical process node and products.

- The impact of different process parameter variations on the impact of the functioning of the memory in nanometer technologies has to be studied in greater depth. In particular, the other failure mechanisms, such as the flipping read failure and hold failure need to be analyzed. These failures happen when we use very low supply voltages and in the presence of excessive parameter variations. In addition, different directions of failure with varying failing condition need to be analyzed. This may involve going for a more aggressive timing and performing a sensitivity analyses in many directions of failure simulatenously. be explored for compensating for the process induced variations.

  There are many different design approaches for improving reliability through the design of variation tolerant circuits. One could reduce the source of manufacturing variations, reduce the effects of process induced variations at the design stage or reduce the effects of variations post silicon after the chip is fabricated. Careful evaluation of the advantages and limitations of these approaches need to be done before the design stage.

- Design approaches to provide immunity to radiation induced particle strikes and soft errors need to be explored. Especially, at nano technologies and higher frequencies where parasitic capacitances tend to play a major role, techniques that make use of the available parasitics to increase the SET-tolerance of the whole circuit need to be explored. In other words, the adverse effects of technology scaling could be used to our advantage to provide hardening for radiation induced errors. We have devised design techniques to mitigate single event transients (that are based on both space-time

redundancy and time redundancy) in mater-slave flip flop designs. These techniques are capable of tolerating SETs (with widths at most half the clock period) that arrive at its input during its window of vulnerability. These techniques need to be extended to provide soft error immunity to the memory cell designs. The susceptibility of the peripheral circuits, in particular the address decoders and sense amplifiers, to particle strikes was not investigated and need to be addressed in future.

Bibliography

[1] Chipworks, “Comparitive analysis of embedded sram.” Rev 1, August 2002.

[2] R. S. Hauck, K. Anne, J. Bell, G. Cheney, J. Eno, G. Hoeppner, G. Joe, R. Kaye, J. Lear, T. Litch, J. Meyer, J. Montanaro, K. Patton, T. Pham, R. Reis, M. Silla, J. Slaton, K. Snyder, and R. Witek, “ A 200MHz 32b 0.5W CMOS RISC Microprocessor,” *ISSCC Digest of Technical Papers*, pp. 238–239, February 1998.

[3] H. Igura, S. Narita, Y. Naito, K. Kazama, I. Kuroda, M. Motomura, and M. Yamashina, “ An 800 MOPS 100mW 1.5V Parallel DSP for Mobile Multimedia Processing,” *IEEE Journal of Solid–State Circuits*, pp. 1820–1828, November 1998.

[4] M. Margala, “Low Power SRAM Circuit Design,” in *IEEE International Workshop on Memory Technology, Design and Testing*, pp. 115–122, August 1999.

[5] T.Mori, 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 applications to a 64K Full CMOS RAM,” *IEEE Journal of Solid–State Circuits*, pp. 479–484, October 1983.

[6] Intel Corporation, "Pentium II Processor for the SC242 at 450MHz to 1.13 Ghz." Datasheet, 2001.

[7] M. Powell and K. Roy, "Gated-$V-dd$: A Circuit Technique to Reduce Leakage in Cacke Memories," in *Proceedings of the International Symposium on Low Power Electronics and Design*, pp. 90–95, July 2000.

[8] C. H. Kim and K. Roy, "Dynamic V_t SRAM: A Leakage Tolerant Cache Memory for Low Voltage Microprocessors," in *Proceedings of the International Symposium on Low Power Electronics and Design*, pp. 251–254, June 2002.

[9] A. Papanikolau, " Interconnect Exploration for Future Wire Domintaed Technologies," in *Proceedings of the ACM*, 2002.

[10] A. S. Sedra and K. C. Smith, *Microelectronic Circuits*. Oxford University Press, 4th ed., 1998.

[11] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*. Prentice-Hall, 2nd ed., 2003.

[12] M. Margala, "Low power memory circuits," in *The VLSI Handbook* (W.-K. Chen, ed.), IEEE Press, pp. 53–1–53–39, CRC Press, 2000.

[13] A. Agarwal and K. Roy, "A Noise Tolerant Cache Design to Reduce Gate and Sub-threshold Leakage in the Nanometer Regime," in *Proceedings of the International Symposium on Low Power Electronics and Design*, pp. 18–21, August 2003.

[14] K. Nii, H. Makino, Y. Tujihashi, C. Morishima, Y. Hayakawa, and Hahnmano, "A Low Power SRAM Using Auto-Backgate-Controlled MT-CMOS," in

Proceedings of the International Symposium on Low Power Electronics and Design, pp. 260–271, 1998.

[15] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy Caches: Simple Techniques for Reducing Leakage Power," in *International Symposium on Computer Architecture*, pp. 148–157, August 2002.

[16] N. Azizi and F. N. Najam, "Low-Leakage Asymmetric-Cell SRAM," *IEEE Transactions on VLSI Systems*, vol. 11, pp. 701–715, August 2003.

[17] K. Nii, Y. Tsukamoto, T. Yoshikawa, S. Imaoka, Y. Yamagami, T. Suzuki, A. Shibayama, H. Makino, and S. Iwade, "A 90-nm Low-Power 32-kB Embedded SRAM with Gate Leakage Supression Circuit for Mobile Applications," *IEEE Journal of Solid–State Circuits*, pp. 684–691, April 2004.

[18] N. Azizi and F. N. Najm, "An Asymmetric SRAM Cell to lower Gate Leakage," in *Proceedings 5th International Symposium on Quality Electronic Design*, pp. 534–539, March 2004.

[19] A. Agarwal, H. Li, and K. Roy, "A Single-V_t Low-Leakage Gated-Ground Cache for Deep Submicron," *IEEE Journal of Solid–State Circuits*, vol. 38, pp. 319–328, February 2003.

[20] L. Wei, Z. Chen, M. Johnson, K. Roy, and V. De, "Design and Optimization of Low-voltage High-performance Dual-threshold CMOS Circuits," in *Proceedings of the 35th Design Automation Conference*, pp. 489–494, 1998.

[21] M. Yoshimoto, " A 64Kb CMOS RAM with Divided Word Line Structure," in *IEEE International Solid State Circuits Conference, Digest of Technical Papers*, pp. 58–59, 1983.

[22] UC Berkeley Device Group, "Berkeley Predictive Technology Model." Online, http://www-devices.eecs.berkeley.edu/ptm, 2000.

[23] P. Elakkumanan, C. Thondapu, and R. Sridhar, "A Gate Leakage Reduction Strategy for sub-70nm Memory Circuits," in *IEEE Dallas Circuits and Systems Workshop*, pp. 145–149, October 2004.

[24] E. S. Sr., F. J. List, and J. Lohstroh, "Static Noise-margin Analysis of MOS SRAM Cells," *IEEE Journal of Solid–State Circuits*, vol. 22, pp. 748–754, October 1987.

[25] K. M. Cao *et al.*, "BSIM4 gate leakage model including source-drain partition," in *IEEE IEDM Technical Digest*, pp. 815–818, December 2000.

[26] Y. C. Yeo *et al.*, "Direct Tunneling Gate Leakage Current in Transistors with Ultrathin Silicon Nitride Gate Dielectric," in *IEEE Electronic Device Letters*, vol. 21, pp. 540–542, November 2000.

[27] A. Agarwal, H. Li, and K. Roy, "A single-Vt Low-Leakage gated-ground Cache for Deep Submicron," in *IEEE Journal of Solid-State Circuits*, pp. 319–328, February 2003.

[28] N. Azizi and F. Najm, "An asymmetric SRAM cell to lower Gate leakage," in *Proceedings of International Symposium of Quality Electronic Design (ISQED)*, pp. 534–539, March 2004.

[29] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE Journal of Solid–State Circuits*, vol. 36, pp. 658–665, April 2001.

[30] J. Lohstroh, E. Seevinck, and J. Groot, "Worst-case noise margin criteria for logic currents and their mathematical equivalence," in *IEEE Journal of Solid-State Circuits*, pp. 803–806, December 1983.

[31] B. S. Amrutur and M. A. Horowitz, "A replica technique for wordline and sense control in low-Power SRAM's," *IEEE Journal of Solid–State Circuits*, vol. 33, pp. 1208–1219, August 1998.

[32] E. Seevinck, P. J. van Beers, and H. Ontrop, "Current Mode Techniques for High-Speed VLSI Circuits with Application to Current Sense Amplifier for CMOS SRAM's ," *IEEE Journal of Solid–State Circuits*, vol. 22, pp. 525–536, April 1991.

[33] T. N. Blalock and R. C. Jaeger, "A High-Speed Clamped Bit-Line current-Mode Sense Amplifier," *IEEE Journal of Solid–State Circuits*, vol. 26, pp. 542–548, April 1991.

[34] B. Wicht, S. Paul, and D. Schmitt-Landsiedel, "Analysis and Compensation of the Bitline Multiplexer in SRAM Current Sense Amplifiers," *IEEE Journal of Solid–State Circuits*, vol. 36, pp. 1745–1755, November 2001.

[35] S. Borkar, "Parameter Variations and Impact on Circuits and Microarchitecture," in *C2S2 MARCO Review*, pp. 413–416, March 2003.

[36] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of Die-to-die and Within-die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration," *IEEE Journal of Solid–State Circuits*, vol. 37, pp. 183–190, February 2002.

[37] T. Karnik, S. Borkar, and V. De, "Sub-90nm Technologies - Challenges and Opportunities for CAD," in *International Conference on Computer Aided Design (ICCAD)*, pp. 203–206, March 2003.

[38] R. Sarpeshkar, J. John L. Wyatt, N. C. Lu, and P. D. Gerber, "Mismatch Sensitivity of a Simultaneously Latched CMOS Sense Amplifier," *IEEE Journal of Solid–State Circuits*, vol. 26, pp. 1413–1422, October 1991.

[39] S. Sundaram, P. Elakkumanan, and R. Sridhar, "High Speed Robust Current Sense Amplifier for Nanoscale Memories – A Winner Take All approach," in *IEEE International Conference on VLSI Design*, pp. 569–574, January 2006.

[40] J. Lazzaro, S. Ryckebusch, M. Mahowald, and C. Mead, "Winner-take-all networks of O(n) complexity. ," in *Advances in Neural Information Processing Systems*, pp. 703–711, 1988.

[41] P. Elakkumanan, S. Sundaram, and R. Sridhar, "LPCSA: A Novel Low Power Current Sense Amplifier for Nanoscale SRAMs," in *Austin Conference on Integrated Systems and Circuits*, May 2006.

[42] M. Izumikawa, K. Suzuki, M. Nomura, H. Igura, H. Abiko, K. Okabe, A. Ono, T. Nakayama, M. Yamashina, and H. Yamada, "A 400MHz, 300mW, 8Kb CMOS RAM Macro with a Current-Sensing scheme," in *IEEE CICC*, pp. 595–598, September 1994.

[43] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*. Prentice-Hall, 2nd ed., 2003.

[44] M. Sinha, S. Hsu, A. Alvandpour, W. Burleson, R. Krishnamurthy, and S. Borkar, "Low Voltage Sensing Techniques and Secondary Design Issue for sub-90nm Caches," in *European Solid State Circuits Conference*, pp. 413–416, September 2003.

[45] P. Y. Chee, P. C. Liu, and L. Siek, "High-speed hybrid current-mode sense amplifier for CMOS SRAMs ," in *Electronics Letters*, pp. 871–873, April 1992.

[46] J.-S. Wang and H.-Y. Lee, " A New Current-Mode Sense Amplifier for Low-Voltage Low-Power SRAM Design," in *Proceedings of ASIC*, pp. 163–167, September 1998.

[47] R. Singh and N. Bhat, "An Offset Compensation Technique for Latch Type Sense Amplifiers in High-speed Low-power SRAM's," *IEEE Transactions on VLSI Systems*, vol. 12, pp. 652–657, June 2004.

[48] A. Chandrakasan, W. Bowhill, and F. Fox, *Design of high-performance microprocessor circuits*. IEEE Press, 2nd ed., 2001.

[49] S. R. Nassif, "Modeling and Analysis of Manufacturing Variations," in *IEEE Conference on Custom Integrated Circuits(CICC)*, pp. 223–228, May 2001.

[50] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling and Estimation of Failure Probability due to Parameter Variations in Nano-scale SRAMs for Yield Enhancement," in *Symposium on VLSI Circuits Digest of Technical Papers*, pp. 64–67, Jun 2004.

[51] P. Elakkumanan, J. B. Kuang, K. Nowka, R. Kanj, and S. Nassif, "Failure Analyses of SRAM Local Bitline Access Schemes," in *IEEE International Symposium on Quality Electronic Design (ISQED)*, pp. 204–209, March 2006.

[52] K. J. Kim et al., "A Novel 6.4mm Full-CMOS SRAM Cell with Aspect Ratio of 0.63 in High Performance 0.25um generation CMOS Technoogy," in *Symposium on VLSI Circuits Digest of Technical Papers*, pp. 68–69, 1998.

[53] S. Nakai et al., "A 65nm CMOS Technology with High Performance and Low Leakage Transistor, a 0.55um 6T-SRAM cell and Robust hybrid-ULK/Cu Interconnects for Mobile Multimedia Applications," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 285–288, 2003.

[54] E. Leobandung et al., "High performance 65 nm SOI technology with dual stress liner and low capacitance SRAM cell," in *Symposium on VLSI Technology Digest of Technical Papers*, pp. 126–127, Jun 2005.

[55] S. Dhong et al., "A 4.8GHz Fully Pipelined Embedded SRAM in the Streaming Processor of a CELL Processor," in *IEEE International Solid State Circuits Conference(ISSCC)*, pp. 486–487, Feb 2005.

[56] T. Kobayashi et al., "A Current-mode Latch Sense Amplifier and Static Power Saving Input Buffer for Low Power Architecture," in *Symposium on VLSI Circuits Digest of Technica Papers*, pp. 28–29, Jun 1992.

[57] T. Wallmark and M. Marcus, "Minimum size and maximum packaging density of nonredundant semiconductor devices," in *IRE Conference*, pp. 286–298, March 1962.

[58] Y. S. Dhillon, A. U. Diriland, and A. Chatterjee, "Soft-Error Tolerance Analysis and Optimization of Nanometer Circuits," in *Design, Automation and Test in Eurpoe Conference and Exhibition (DATE)*, pp. 288–293, 2005.

[59] N. Seifert, X. Zhu, and L. Massengill, "Impact of Scaling on Soft-Error Rates in Commercial Microprocessors ," *IEEE Transactions on Nuclear Science*, vol. 49, pp. 3100–3106, December 2002.

[60] N. Seifert, D. Moyer, N. Leland, and R. Hokinson, "Historical Trend in Alpha-Particle induced Soft Error Rates of the Alpha Microprocessor," *IEEE International Reliability Physics Symposium*, pp. 259–265, 2001.

[61] T. Karnik, B. Bloechel, K. Soumyanath, V. De, and S. Borkar, "Scaling Trends of Cosmic Rays Induced Soft Errors in Static Latches Beyond $0.18\mu m$," in *IEEE Symposium on VLSI Circuits*, pp. 61–62, June 2001.

[62] D. Binder, E. C. Smith, and A. B. Holman, "Satellite anamolies from galactic cosmic rays ," *IEEE Transactions on Nuclear Science*, vol. 22, pp. 2675–2680, December 1975.

[63] T. C. May and M. H. Woods, "A new physical mechanism for soft errors in dynamic memories," in *IEEE 16th Annual International Reliability Physics Symposium*, pp. 33–42, April–June 1978.

[64] T. Karnik, P. Hazuha, and J. Patel, "Characterization of soft errors caused by single event upsets in CMOS processes ," *IEEE Transactions on Nuclear Science*, vol. 1, pp. 128–143, April–June 2004.

[65] V. De and S. Borkar, "Technology and Design Challenges for Low Power and High Performance," in *IEEE Symposium on Low Power Electronics (ISLPED)*, pp. 163–168, 1999.

[66] H. T. Nguyen and Y. Yagil, "A Systematic Approach to SER Estimation and Solutions," in *IEEE Annual International Reliability Physics Symposium*, pp. 60–70, March–April 2003.

[67] 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," in *Proceedings of the International Conference on Dependable Systems and Networks*, pp. 389–398, June 2002.

[68] M. Baze and S. Buchner, "Attenuation of Single Event Induced Pulses in CMOS Combinational Logic ," *IEEE Transactions on Nuclear Science*, vol. 44, pp. 2217–2223, December 1997.

[69] R. C. Baumann, "The Impact of Technology Scaling on Soft Error Rate Performance and Limits to the Efficacy of Error Correction," in *Error Devices Meeting*, pp. 329–332, December 2002.

[70] M. Hazucha and C. Svensson, "Impact of CMOS Technology Scaling on the Atmospheric Neutron Soft Error Rate," *IEEE Transactions on Nuclear Science*, vol. 47, pp. 2217–2223, December 2000.

[71] N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design : A systems perspective*. Addison-Wesley, 2nd ed., 1992.

[72] P. Elakkumanan, A. Narasimhan, and R. Sridhar, "NC-SRAM - A Low Leakage Memory Circuit for Ultra Deep Submicron Designs," in *IEEE International SoC Conference*, pp. 3–6, September 2003.

[73] A. J. Auberton-Herve, "SOI: Materials to Systems," in *IEEE International Electron Device Meeting Technical Digest*, pp. 3–10, 1996.

[74] F. T. Brady, T. Scott, R. Brown, J. Damato, and N. F. Haddad, "Fully-depleted submicron SOI for radiation hardened applications," *IEEE Transactions Nuclear Science*, vol. 41, p. 2304, December 1994.

[75] IBM, "SOI technology: IBM's next advance in chip design." Whitepaper, http://www-3.ibm.com/chips/bluelogic/showcase/soi/soipaper.pdf.

[76] M. Ketchen, M. Bhushan, and D. Pearson, "High Speed Test Structures for In-Line Process Monitoring and Model Calibration," in *International Conference on Microelectronics and Test Structures*, pp. 33–40, 2005.

[77] M. Ketchen, M. Bhushan, and C. Anderson, "Circuit and Technique for Charecterizing Switching Delay History Effects in SOI Logic Gates," in *International Conference on Microelectronics and Test Structures*, pp. 768–771, 2004.