Low-Power, Low-Voltage SRAM Circuit Designs For Nanometric CMOS Technologies

by

Tahseen Shakir

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

Waterloo, Ontario, Canada, 2011

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

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

Abstract

Embedded SRAM memory is a vital component in modern SoCs. More than 80% of the System-on-Chip (SoC) die area is often occupied by SRAM arrays. As such, system reliability and yield is largely governed by the SRAM's performance and robustness. The aggressive scaling trend in CMOS device minimum feature size, coupled with the growing demand in high-capacity memory integration, has imposed the use of minimal size devices to realize a memory bitcell. The smallest 6T SRAM bitcell to date occupies a $0.1um^2$ in silicon area.

SRAM bitcells continue to benefit from an aggressive scaling trend in CMOS technologies. Unfortunately, other system components, such as interconnects, experience a slower scaling trend. This has resulted in dramatic deterioration in a cell's ability to drive a heavily-loaded interconnects. Moreover, the growing fluctuation in device properties due to Process, Voltage, and Temperature (PVT) variations has added more uncertainty to SRAM operation. Thus ensuring the ability of a miniaturized cell to drive heavily-loaded bitlines and to generate adequate voltage swing is becoming challenging. A large percentage of state-of-the-art SoC system failures are attributed to the inability of SRAM cells to generate the targeted bitline voltage swing within a given access time.

The use of read-assist mechanisms and current mode sense amplifiers are the two key strategies used to surmount bitline loading effects. On the other hand, new bitcell topologies and cell supply voltage management are used to overcome fluctuations in device properties. In this research we tackled conventional 6T SRAM bitcell limited drivability by introducing new integrated voltage sensing schemes and current-mode sense amplifiers. The proposed schemes feature a read-assist mechanism. The proposed schemes' functionality and superiority over existing schemes are verified using transient and statistical SPICE

simulations. Post-layout extracted views of the devices are used for realistic simulation results.

Low-voltage operated SRAM reliability and yield enhancement is investigated and a wordline boost technique is proposed as a means to manage the cell's WL operating voltage. The proposed wordline driver design shows a significant improvement in reliability and yield in a 400-mV 6T SRAM cell. The proposed wordline driver design exploit the cell's Dynamic Noise Margin (DNM), therefore boost peak level and boost decay rate programmability features are added. SPICE transient and statistical simulations are used to verify the proposed design's functionality.

Finally, at a bitcell-level, we proposed a new five-transistor (5T) SRAM bitcell which shows competitive performance and reliability figures of merit compared to the conventional 6T bitcell. The functionality of the proposed cell is verified by post-layout SPICE simulations. The proposed bitcell topology is designed, implemented and fabricated in a standard ST CMOS 65nm technology process. A 1.2$\times$ 1.2 mm^2 multi-design project test chip consisting of four 32-Kbit (256-row x 128-column) SRAM macros with the required peripheral and timing control units is fabricated. Two of the designed SRAM macros are dedicated for this work, namely, a 32-Kbit 5T macro and a 32-Kbit 6T macro which is used as a comparison reference. Other macros belong to other projects and are not be discussed in this document.

Acknowledgements

I would like to express my sincere gratitude and appreciation to Professor Manoj Sachdev. It gives me a great pleasure and honor being a student of Professor Sachdev. His insightful and thoughtful guidance, support and encouragement have been invaluable. While closely supervising my research progress through a regular weekly meetings, he provided me with an excellent research environment. The moral support I was getting from him during my ups and downs were invaluable. Thank you Professor Sachdev.

I would like to thank also Professor Bruce Cockburn, Professor James Martin, Professor David Nairn, and Professor Ajoy Opal for kindly accepting to be my examination committee members.

I would like thank the ECE department computing staff: Pual Ludwing Phil Regier, and Fernando Hernandez for their help whenever needed. Phil, in particular, was my "hero" on many occasions. To our department graduates studies staff, you have been wonderful and helpful people, thank you all.

My dear friends in CDR group, thank you all for the cheerful memories and moments we spent together. Thanks to our supervisor, we had such a pleasant moments during our "annual" lunch meetings. In Particular, Dr. David Rennie, I am deeply thankful to his help during many hard times I had. Dr. Rennie willingness to help was beyond words. Adam , David Li, Pierce, Jaspal and Tasreen, I am so happy having you in my life.

I am indebted to my wife, Karama, my kids: Nassar, Manar, and Mohamed. At some points, I felt guilty for the hard time I gave them during my "downs", which were many! But you were always there with your support and care! My Mother, parents inlaw, brothers, sisters and inlaws, I am grateful for your care.

Dedication

To my beloved father, a live or dead; the man who sacrificed everything for us. To my dear mother, the person who dedicated her life for us; may God keep her safe and healthy. To every one who prayed and wished me success in my life. To my family: Karama, Nassar, Manar, and Mohamed, I dedicate this thesis.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Symbol	Description
ALU:	Arithmetic Logic Unit
ASIC:	Application-Specific Integrated Circuit
Bl(b):	Bitline(Complement)
BWL:	Boosted Wordline
CFS:	Constant Field Scaling
CVS:	Constant Voltage Scaling
CMOS:	Complementary Metal Oxide Semiconductor FET
DNM:	Dynamic Noise Margin
DRD:	Destructive Read Operation
DRAM:	Dynamic Random Access Memory
EOT:	Effective Oxide Thickness
FBB:	Forward Body Bias
FET:	Field Effect Transistor
FIR(W):	Failure in Read(Write)
FOM:	Figure-of-Merit
HK-(MG):	High-Dielectric-(Metal-Gate)
ITRS:	International technology Roadmap for Semiconductor
L(H)V_{TH}:	Low (High) V_{TH}
LSI:	Large Scale Integration
LER:	Line Edge Roughness
MOSFET:	Metal-Oxide Semiconductor Field Effect Transistor
NMOS:	N-Type MOS
PMOS:	P-Type MOS

Symbol:	Description:
PDP:	Power Delay Product
PVT:	Process, Voltage and Temperature Variations
RDM:	Read Noise Margin
RBB:	Reverse Body bias
RDF:	Random Dopant Fluctuation
RA-SA:	Read-Assist Sense Amplifier
RA-WRBK-SA:	RA-Write-Back-SA
SAE:	SA Enable Signal
SA:	Sense Amplifier
SNM:	Static Noise Margin
SVNM(SINM):	Static Voltage (Current) Noise Margin
SEU:	Single Event Upset
SoC:	System-on-Chip
SRAM:	Static Random Access Memory
ST:	STMicroelectronics
TFT:	Thin-Film Transistor
TSMC:	Taiwan Semiconductor Manufacturing Company
V_{TH}**:**	Transistor's Threshold Voltage
V_{DDmin}**:**	SRAM Cell Minimum Operating Voltage
VLSI:	Very Large Scale Integration
VTC:	Voltage Transfer Function
WLE:	Wordline Enable Signal
WR(RD)bl:	Write(Read) Bitline
WRM:	Write Noise Margin

Chapter 1

Introduction to Embedded Memories

1.1 Introduction

The early 1970s was the starting point of the era of large scale integration (LSI) and semiconductor memory mass production. The first sale of a 1 Kbit dedicated Dynamic Random-Access Memory (DRAM) and the extensive use of semiconductor memory chips in IBM mainframe computers were the most remarkable events of that time. From those days, the increase in memory chip capacity has skyrocketed, owing to the ever-increasing scaling in Complementary Metal-Oxide Semiconductor (CMOS) technologies. Furthermore, consistent research, studies, and technology developments have led to a substantial and dramatic improvement in high-density integration [10]. **Figure 1.1** illustrates the growing trend in device count in Intel microprocessors. The transistor number, for example, doubles in each subsequent CMOS generation.

Embedded memories, which often occupy a significant portion of the die area, are the cornerstone of many state-of-the-art system-on-a-chip (SoC) applications. On-chip cache

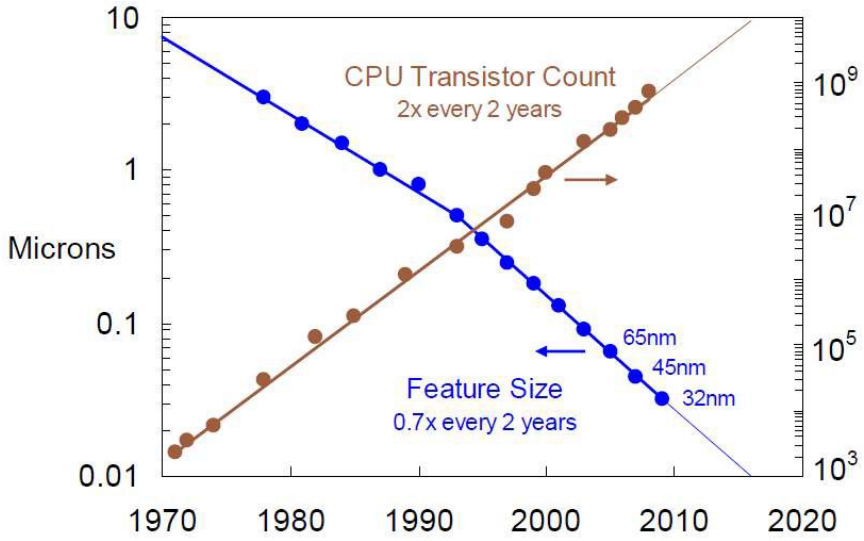


Figure 1.1: Trend in Device Count Per Chip and Minimum Feature Size [2].

memory, a vital component of any high performance microprocessor [2], is a good example to highlight the importance of embedded memories. Microprocessor cache memory is used to reduce the divergence in speed between the processor's Arithmetic Logic Unit (ALU) and the system's main dedicated off-chip DRAM. The system speed gradually decreases from very high-speed microprocessor registers to a relatively low-speed main DRAM. As such, microprocessor cache memory is usually built in a hierarchial structure. The first-level cache (L0), which is the microprocessor's registers, is a very high-speed high-performance, yet low-density, memory array; whereas, the top-level cache (L2) is relatively low-speed and high-density. In between these two levels, the L1 cache is designed to match the speed difference between L0 and L2 speed and density. The microprocessor cost/performance trade-off is decided based on its L2 cache. Hence, high performance microprocessors can be optimized for low cost by minimizing the L2 cache density.

Because of the minimal number of transistors needed to realize a single data bitcell, the

one transistor (1T) DRAM bitcell was the ideal choice for high-capacity cache memories in the early 1970s. However, limited speed and high power consumption were an increasing concerns when using 1T DRAMs. The high standby leakage current and related need to refresh the stored data imply significant power consumption in the 1T DRAM cell. Static power consumption in battery-operated SoCs is a major design concern, therefore, the development of a potential replacement for the 1T DRAM bitcell was inevitable.

In contrast, the unique features and characteristics of the Static Random-Access Memory (SRAM) bitcell have made it a preferable choice in many state-of-the-art SoC applications. Compatibility with the standard CMOS logic and its symmetrical and differential topology are among many features that make the SRAM a suitable candidate for embedded cache applications. The SRAM bitcell's compatibility with standard CMOS logic eliminates the need for different fabrication masks when integrated in a system. Additionally, the cell lithography symmetry makes it reliable and easy to fabricate.

Despite the fact that SRAM bitcell integrity was a concern in high-density embedded applications due to the large number of transistors needed to realize a single data bitcell, SRAM has dominated the market of the embedded memory applications in many Application-Specific Integrated Circuits (ASIC), owing to aggressive scaling in the CMOS industry.

1.2 CMOS Technology Scaling Trends

The revolution in CMOS technology started in 1963. Since then, this technology has become the preferred digital circuit design platform due to its scalability and low power consumption. In 1965 Gordon Moore, later on a co-founder of Intel, predicted the growth in device integration for the then foreseeable future and a law, named after Moore, was

established and has been followed ever since. According to Moore's law, CMOS device minimum feature size (poly gate) is predicted to scale by half in each subsequent CMOS generation, and thereby device and system performance are expected to double every two years (see **Figure 1.1**).

Classical MOSFET transistor scaling approach [11] suggests that the MOSFET device is scaled by transformation in three variables: dimension, voltage, and substrate doping rate. The device's physical dimensions include: gate oxide thickness, drain and source diffusion area, and gate width. Accordingly, CMOS device scaling has been performed in two approaches:

(1) Constant Voltage Scaling (CVS)

Down to 0.8 μm CMOS technology, CVS was an acceptable way to improve CMOS device performance. In this approach, device physical dimensions are scaled down approximately two times per subsequent generation (two years) without scaling the operating supply voltage V_{DD} [1]. This scaling approach leads to greater integration density, higher-speed operation and lower power consumption (at the circuit level). More importantly, this approach maintains the CMOS device's compatibility with other semiconductor devices requiring higher power supply voltages. However, as the CMOS device dimension continues to scale into submicrometer regime, the device started to deviate from its classical long channel behavior. For example, short channel effects, such as velocity saturation, gate dielectric breakdown, and gate leakage became significant limiting phenomena, so further device feature size scaling no longer enhances device performance.

Table 1.1: Scaling in CMOS Device [1].

Parameter	Relation	CFS	CVS
L, W, t_{ox}		1/S	1/S
V_{DD}, V_{TH}		1/S	1
N_{SUB}	V_{DD}/W_{depl}^2	S	S^2
$Area$	WL	$1/S^2$	$1/S^2$
C_{gate}	$C_{ox}WL$	1/S	1/S
K_n, K_p	$C_{ox}W/L$	S	S
I_{on}	$C_{ox}WV_{DD}$	1/S	1
delay	$C_{gate}V_{DD}/I_{on}$	1/S	1/S
Intrinsic Power	$I_{on}V_{DD}$	$1/S^2$	1

(2) Constant Field Scaling (CFS)

Beyond 0.8 μm process, further scaling in device minimum feature size starts to degrade the device's performance due to the high electric field induced over the small device area. It was to address this issue that the CFS scaling approach was developed. In this scaling approach, the power supply voltage-to-device feature size ratio remains constant. Thus, the induced electric field remains constant. Using this scaling approach results in improvements in device integration and performance and reduces the overall chip power consumption. **Table 1.1** shows how different CMOS device parameters are scaled by a factor of S according to the above scaling approaches.

Down to the 100-nm CMOS generation, classical device scaling continues to enhance device and system performance. However, although further device scaling is still possible owing to the advancement in the CMOS industry and fabrication facilities, further scal-

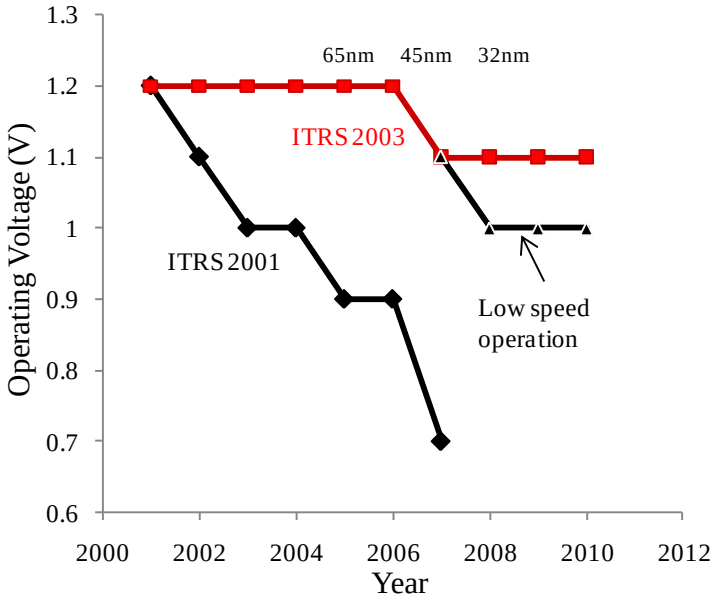


Figure 1.2: Supply Voltage Scaling Shift in Modern CMOS Technologies [3].

ing of the supply voltage starts to deteriorate the digital circuits and systems operation. Additionally, the impact of process and mismatch variations becomes more pronounced at low operating voltages. Thus, power supply scaling in 90-nm CMOS technology node and beyond no longer follows the device scaling trend.

Figure 1.2 shows the International Technology Roadmap for Semiconductors (ITRS) projected supply voltage scaling trend in 2001 in contrast to the actual supply voltage used in the industry and the projected scaling trend in 2003. As can be seen from this figure, while a constant voltage supply was maintained over three generations (130 nm to 65 nm), the supply voltage scaling now faces a 1.0 V barrier even though device continues to scale [3].

The use of a relatively high voltage over a tiny channel area degrades the carrier's

Table 1.2: Intel's Device Scaling Using HK and HK-MG Technologies: Reproduced from [2].

Process	technology	Channel length	Contacted gate	EOT	Supply voltage
LP65nm	HK	60nm	220nm	1.7nm	1.2V
LP45nm	HK-MG	40nm	160nm	1.0nm	1.0V

mobility and brings gate oxide breakdown, gate tunneling leakage, and subthreshold leakage current problems back onto the scene. In fact, CMOS devices in these technology nodes (90 nm and beyond) are even more susceptible to the high electric field challenges. Microscopic variations in the number and location of dopant atoms in the channel region of the device are highly affected by the channel's electric field. This leads to a high degree of uncertainty in the electrical properties of the fabricated device's figures of merit (speed, leakage and reliability) [12]. Thus, CFS cannot be considered to be a suitable approach in modern CMOS technologies. Therefore, new device-level solutions have been introduced by many CMOS industry leaders like Intel, Toshiba, and TSMC, to reduce the impact of the high electric field induced in nanometric CMOS devices.

The first device-level solution introduced, to accommodate the high electric field in the device's channel region, was the use of high dielectric constant material (High-K) in the gate region to reduce the gate leakage and increase device reliability. This solution seems to be feasible to some extent; however, further device scaling has resulted in further increases in the channels's electric field. Consequently, this excessive increase in the electric field started causing polysilicon gate depletion and dopant penetration which can lead to erroneous gate activation and, thereby, device failure. Therefore, the polysilicon gate, traditionally used in CMOS devices, has been replaced by a metal gate electrode. This

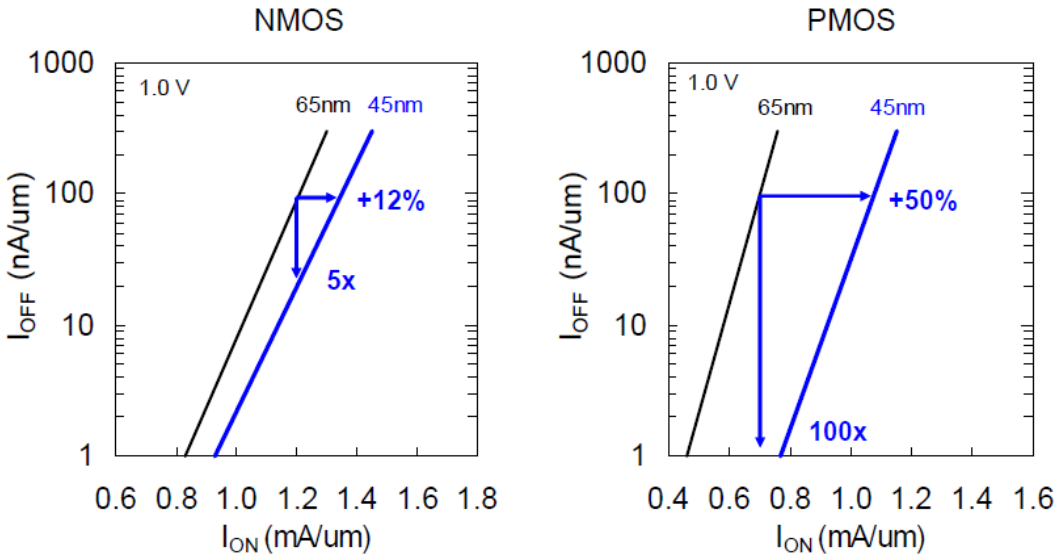


Figure 1.3: CMOS Device Performance Enhancement Using HK-MG [2].

device generation, which has been considered a formidable leap in the modern CMOS device industry, was developed by Intel in 2007 and used in their 45 nm CMOS process.

Table 1.2 shows the scaling benefits of using of HK-MG device technology compared to HK device technology as used in two generations of Intel's state-of-the-art CMOS devices [2]. The use of a relatively high supply voltage over very small device dimensions, as shown in **Table 1.2**, results in tremendous device performance improvement. **Figure 1.3** shows the performance improvement of NMOS and PMOS devices implemented in 65 nm HK compared to 45 nm HK-MG processes operating at 1.0 V. As seen in **Figure 1.3**, at least 5X leakage current reduction and 12% saturation current improvement are achieved by the move to HK-MG technology.

1.3 Nanometric CMOS Device Performance

CMOS device performance and reliability continue to benefit from growing advancements in CMOS technology. An approximate transistor performance metric used in the industry is: $C_{gate} \times V_{DD}/I_{dsat}$; where, C_{gate} is a device process parameter; V_{DD} is the device operating supply voltage and I_{dsat} is the device saturation current measured at 100 nA/μm I_{off}. The last two parameters are usually used as transistor performance metrics [4].

The device's on state current (I_{on}) is strongly depending on the device's threshold voltage (V_{TH}). As such, nanometric CMOS device performance is modulated by variations in V_{TH}. **Equation 1.1** indicates that V_{TH} variation is inversely proportional to the square root of device physical width (W) and length (L), and directly proportional to gate oxide thickness (EOT). Therefore, according to the data given in **Table 1.2**, V_{TH} variation in miniaturized devices used in high density VLSI systems is expected to be wide due their minimal dimensions. Another source of device V_{TH} variation are variations in process, voltage and temperature (PVT). Process variations in nanometric CMOS technology are predominantly caused by two sources: random dopant fluctuation (RDF) and line edge roughness (LER) [13].

$$\sigma V_{TH} = \frac{EOT}{\sqrt{W \times L}} \tag{1.1}$$

1.4 SRAM Bitcell Performance

Like any other CMOS system, SRAM cell area and performance have benefited from the trend toward aggressive scaling in CMOS device's minimum feature size. If fact, since its first appearance in 1972, the well-known six-transistor (6T) SRAM bitcell continues to

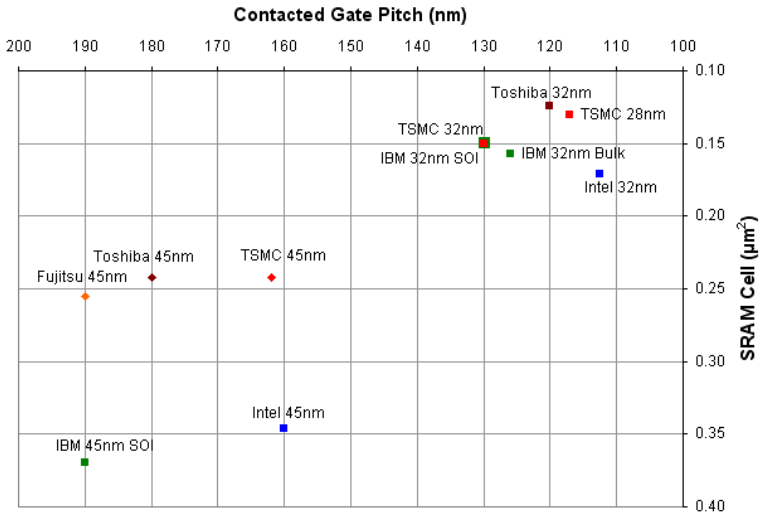


Figure 1.4: 6T SRAM Bitcell Area Scaling Trend in Nanometeric Regime [4].

follow Moore's law and scales by two, approximately every two years. Whereas the earliest CMOS 6T SRAM cell reported in 1972 occupied more than 5700 μm^2 [14], a state-of-the-art 6T SRAM cell reported in 2008 occupied only 0.1 μm^2 of silicon area. **Figure 1.4** summarizes the advancement in typical 6T SRAM bitcell area scaling in state-of-the-art CMOS generations as reported by different industrial foundries.

The two key SRAM cell design metrics are performance and reliability. SRAM cell performance is measured in terms of ability to drive a bitline and to generate adequate bitline differential voltage in a given time interval. The cell's reliability, on the other hand, is measured in terms of ability to retain stored data indefinitely under different operating conditions. The SRAM cell design is based on a delicate balance of transistor size and electrical properties. Its static nature is ensured by the use of an active latch structure that exhibits a positive feedback mechanism.

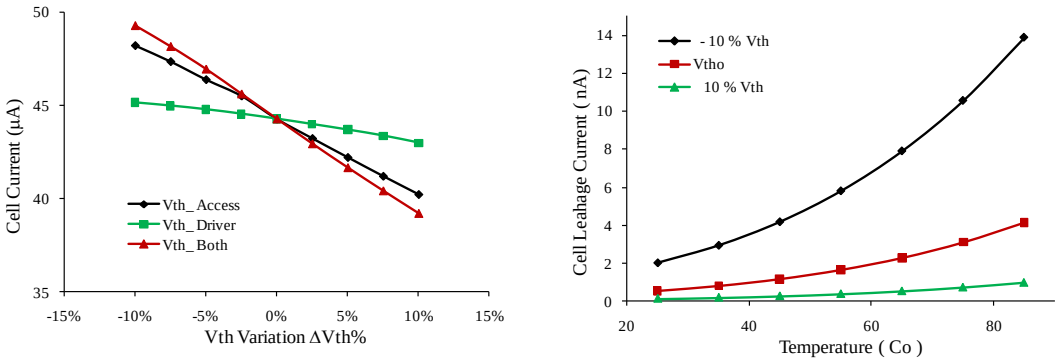
Whereas V_{TH} variation in logic circuits is not especially crucial and can be mitigated by

proper transistor sizing or can average out through multi-stages designs, V_{TH} variation in SRAM circuits can be amplified by the circuit structure and may lead to cell malfunction. Furthermore, in logic circuits, device performance degradation due to V_{TH} variation and/or reduced operating supply voltage produces system speed limitations but not functionality failure. V_{TH} variation in SRAM circuits can, however, be exacerbated and thereby the cell loses the ability to retain data.

In order to ensure a minimum impact of V_{TH} variation on cell stability, the cell's operating voltage must not be below a well know industry figure called V_{DDmin}. SRAM cell operation beyond V_{DDmin} is a very important power consumption metric. Even though SRAM cell failures beyond the V_{DDmin} are soft failures and the cell can recover by increasing the supply voltage, the overall SoC performance and the chip power management are highly affected by V_{DDmin} value.

In nanometric CMOS technologies, a 15% spread in V_{TH} variation is typically expected. V_{TH} fluctuations of neighboring SRAM cell transistors, due to process and mismatch variation, can result in a considerable degradation in cell performance. This has been confirmed with SPICE simulation results performed on a 6T SRAM cell in ST 90-nm CMOS technology. **Figure 1.5(a)** shows a ± 10% V_{TH} variation of cell's access and driver transistors in an SRAM cell and its impact on the cell's drivability. As can be seen in the figure, within this range of V_{TH} variation, the cell drivability varies within ± 12%. The combined impact of device V_{TH} and operating temperature variations is depicted in **Figure 1.5(b)**. As can be seen in the figure, the leakage current almost doubles when device V_{TH} is reduced by 10% at room temperature.

The ever increasing demand for high-capacity SRAM integration has imposed the use of minimum or near-minimum transistor feature size in memory cell design. Unfortunately, other system parameters' scaling do not improve system performance as well as the scaled



(a) Cell Current Variability

(b) Cell Leakage Current Variability

Figure 1.5: V_{TH} Variation Impact on SRAM Cell Performance.

transistor does. For example, dense and thin low level interconnects can deteriorate system performance due to their high resistance. Similarly, high-level long, wide, and thick interconnect layers can affect the SRAM cell's operation due to the high parasitic resistance and capacitance. Thus, miniaturized transistors used to realize the SRAM cell in the nanometric CMOS regime can no longer cope with interconnect loading effects, particularly at V_{DDmin}.

SRAM cell failures due to heavily loaded interconnects are soft failures in nature. These failures happen due to low operating supply voltage or high operating frequency and are aggravated by the presence of process and mismatch variations. Generally speaking, SRAM soft failures are classified as:

- Failure in read (FIR): defines a cell's inability to develop adequate signal to indicate the stored data value.
- Failure in write (FIW): defines a cell's inability to write new data in response to write

operation.

- Failure in data retention: defines a cell's inability to retain the stored data. This especially occurs when cell supply voltage is reduced.

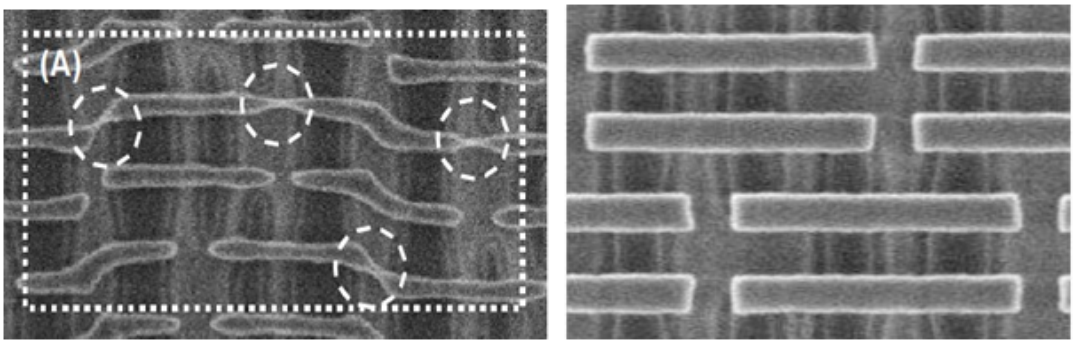
1.5 Existing SRAM Enhancement Techniques

Unlike SRAM bitcell hard (permanent) failures, SRAM soft failures are non-permanent and cell functionality can be recovered by increasing the operating supply voltage above V_{DDmin} or lowering the operating frequency. However, in order to maintain the same scaling/performance improvement trend in modern CMOS technologies, process-level and circuit-level solutions have been developed to cope with the limited cell's drivability.

1.5.1 Process-Level Solutions

From a device perspective, process-level solutions adopted by industry to cope with SRAM cell reliability issues include the introduction of HK and HK-MG device technologies, as presented in Section 1.3 and shown in **Figure 1.4**. From an interconnect perspective the introduction of high count metal layers, such as the nine layers used in Intel's 45-nm process, and the replacement of aluminum by copper have been used in the last few years as the means of increasing conductivity and improving electromigration resistance.

Recently, a significant reduction in wire parasitic capacitance has been achieved through the use of a very low-K dielectric. Low wire capacitance is a great asset in active power reduction and system speed enhancement [2]. Furthermore, the use of accurate lithography patterning has contributed to cell failure reduction. **Figure 1.6** shows a micrograph of a



(a) Single Exposure Lithography (b) Double Exposure Lithography

Figure 1.6: Conventional 6T SRAM Cell Micrograph in 32-nm CMOS Technology Using Different Lithography Technologies.

conventional 6T SRAM cell realized in a standard 32-nm CMOS process. **Figure 1.6(b)** illustrates the advantage of using double exposure lithography technology compared to the conventional single exposure lithography used in **Figure 1.6(a)**.

1.5.2 Circuit-Level Solutions

We will use the conventional 6T SRAM bitcell circuit schematic in our circuit-level solutions discussion since it is the core of this study. As seen in the 6T cell circuit diagram, shown in **Figure 1.7**, the cell's ability to drive the bitline Bl (or Blb) is determined by the drivability of the access and driver transistors M2, M6 (or M1, M5). Cell-level solutions focus on these two transistors to control cell drivability.

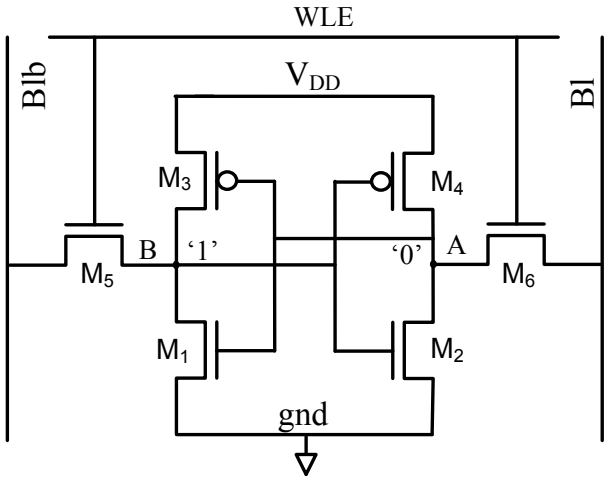


Figure 1.7: Conventional 6T SRAM Bitcell Schematic Diagram.

(1) 6T Bitcell Design Adjustment

Cell drivability can be enhanced by improving the cell's access and driver transistors' drivability. This can be accomplished either by increasing the transistor's physical channel width or by increasing the transistor's overdrive voltage.

- Increasing Device Channel Width:

  This is a straightforward approach in which the cell is sized-up to meet the performance requirements irrespective of area and power overhead. This approach is usually adopted in low operating supply voltage conditions and no other cell enhancement techniques are used. **Figure 1.8** shows SPICE simulation results conducted to find the relationship between cell operating supply voltage V_{DD} and both cell area and operating frequency. As can be seen in the figure, in order to satisfy the same operating conditions, a cell area increase of 22 times or two orders of magnitude reduction in operating frequency is needed.

- Increasing Transistor Overdrive Voltage:

  From a high-density, high-speed design perspective, increasing the cell area or decreasing the operating frequency seems inappropriate due to the associated cell area overhead and speed degradation. Therefore, changing the transistor overdrive voltage has become an interesting research topic in state-of-the-art-SRAM circuit designs. In this context, circuit-level techniques proposed in the literature suggest controlling the driver and the access transistors' (M1, M2 and M5, M6) overdrive voltage to control either their absolute and/or relative drivability.

  One way to do this is by using a dynamic power supply voltage (V_{DD}) [15][16]. In this approach, the cell power supply voltage V_{DD} is made dynamic so that the driver transistor overdrive voltage can be controlled based on the intended memory operation. For example, in order to retain the data during retention mode, the driver transistor does not need to be strong, hence the cell V_{DD} is lowered. This can help in standby power reduction without compromising data stability.

  On the other hand, during a cell read operation the driver transistor needs to be strong enough to drive the heavily-loaded wires. The operating supply voltage V_{DD} must therefore be raised in order to increase the driver's overdrive voltage. This enhances the cell drivability and increases its noise immunity.

  Access transistor drivability can also be controlled by controlling the level of the wordline enable (WLE) signal (see **Figure 1.7**). The WLE signal is activated when the cell is in access mode *,i.e.,* the cell is performing either a read or a write operation (modes of operation will be discussed in Chapter Two). Wordline suppress [17][18] and wordline boost techniques are used to control the access transistor's drivability and thereby enhance SRAM cell performance and reliability [19][20][21]. Details on

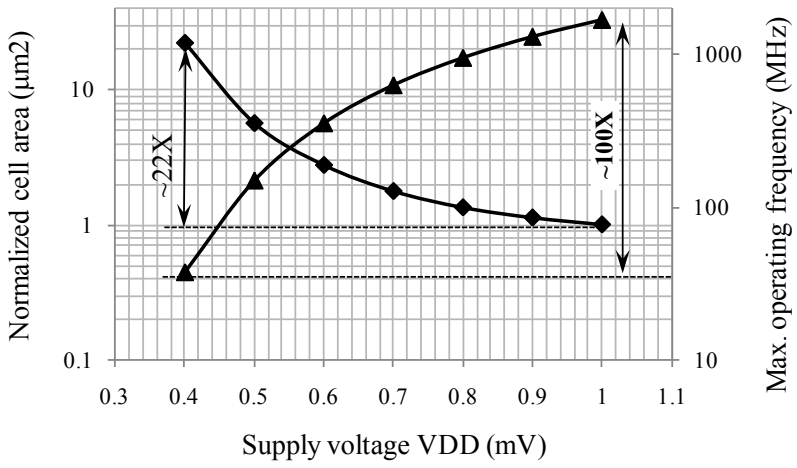


Figure 1.8: 6T SRAM Cell Area and Operating Frequency as a Function of Cell Operating Supply Voltage V_{DD}.

WL boost circuits will be presented in Chapter Four.

(2) Using Alternative SRAM Bitcell Topologies

Read and write operations in the conventional 6T SRAM cell (**Figure 1.7**) are interdependent due to the use of a common port to perform both operations. This interdependency creates a read/write conflict so that transistors' oversizing for reliable read operation can hurt write operation reliability and *vice versa*. Recently, alternative SRAM cell topologies, as summarized in **Figure 1.9**, have been reported. These topologies are mainly meant to break down the read/write interdependence in the conventional 6T SRAM cell by using separate read and write ports [8][6][5][7]. As seen in **Figure 1.9**, these topologies employ extra transistors to isolate the read and write operation ports.

The penalty associated with these topologies takes the form of area overhead due to

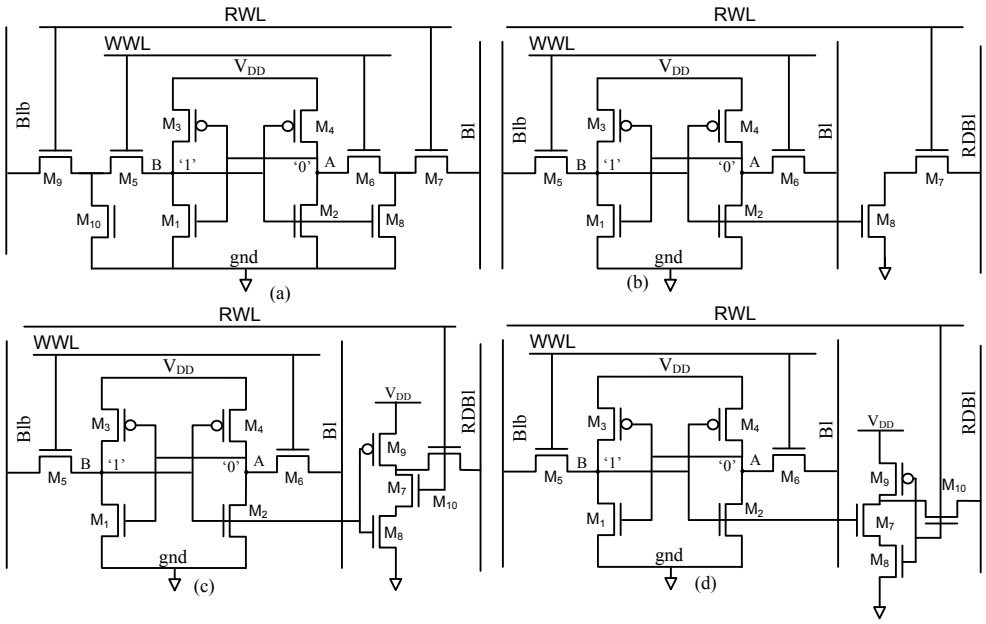


Figure 1.9: State-of-the-Art Multi-Port SRAM Bitcell Topologies Proposed by [5][6][7][8], Respectively.

the use of extra transistors and power overhead due to the use of extra control lines. Additionally, some of these topologies abandon the symmetrical layout structure which is one of the great advantages of the conventional 6T cell. A promising potential replacement for the conventional 6T cell is the 8T cell proposed in [5].

(3) SRAM Peripheral Circuit Designs

Aside from the memory cell itself, another important component in the SRAM memory array that contributes to a successful read operation is the sense amplifier. The sense amplifier is a complementary component in SRAM memory array that can help the cell to perform successful read operation by sensing and amplifying a very small voltage swing the the memory cell generates during a read operation. Despite the fact that the sense amplifier

itself is subject to process and mismatch variations, high-sensitivity sense amplifiers can support the scaled SRAM memory cell by sensing diminishing bitline voltage swings.

For this reason, there have been a number of sense amplifier schemes reported in the literature. In general, the core circuit in most sense amplifier schemes is the basic differential amplifier. Due to its differential nature, this sense amplifier is vulnerable to mismatch variation and overcoming its intrinsic offset is a key reliability issue. One step ahead in the use of high sensitivity sense amplifiers is the introduction of sense amplifiers which feature read-assist mechanisms. This kind of circuit solution allows the cell to develop high voltage swing during a read operation by creating an additional current path to increase the bitline discharge rate so that the sense amplifier can make a reliable decision.

The fact that high-capacity SRAM arrays are highly in demand has created tended to increase the number of cells per bitline and hence increase the bitline loading. Therefore, the memory-developed bitline voltage swing continues to diminish and conventional voltage sense amplifiers are no longer able to sense this low voltage level, especially in the presence of process and mismatch variations which create a high sense amplifier offset voltage. Thus, the need for robust offset-insensitive sense amplifiers became inevitable.

One effective solution for reliable high-speed sensing is to employ a current mode sense amplifier. The operation of this kind of sense amplifier is based on the fact that, during a read operation, the memory cell generates a differential current under any circumstance. As the basic differential sense amplifier scheme was the core for many conventional voltage sense amplifiers, the sense amplifier scheme proposed in [22] is also considered a core for most current mode sense amplifiers used in SRAM arrays. Further details on the sense amplifiers are presented in Chapters Two and Three.

1.6 Motivation and Thesis Outline

System performance continues to benefit from the aggressive scaling trend in CMOS devices. However, process and mismatch variations in nanometric CMOS devices are becoming proportionally more significant and they can cause device properties to deviate from the designed values. Therefore meeting design specifications under worst-case conditions is getting harder then ever. SRAM circuits, in particular which are designed based on a delicate balance of transistor size and properties, are more susceptible to process variations than logic circuits. Furthermore, the stability of conventional SRAM bitcells operating at low supply voltage adds additional memory reliability challenges.

There have been many ongoing attempts at both the process and circuit levels to introduce new schemes capable of providing robust and reliable SRAM arrays without compromising cell integration and performance. In general SRAM circuit-level solutions focus on three circuit aspects: 1) the memory bitcell topology, 2) memory bitcell supply voltage management, and 3) the memory sense amplifier. This research work seeks to introduce new circuit-level solutions to enhance SRAM bitcell reliability and performance. We focused on the aforementioned three aspects in the SRAM array and proposed new circuit-level designs to enhance reliable SRAM memory operation. We will adopt the conventional 6T SRAM cell as a circuit under test whenever we introduce new peripheral circuit designs.

The rest of this thesis is organized as follows: Chapter Two reviews SRAM array architecture and conventional 6T SRAM bitcell characterization and analysis. Chapter Three presents new SRAM sensing schemes which feature a read-assist mechanism. Simulation results that demonstrate the proposed schemes' functionality and performance are also presented in this chapter. In Chapter Four, we propose a new wordline boosting technique

that is capable of supporting low voltage operating 6T SRAM bitcell without compromising the cell functionality or stability. The proposed circuit design scheme is presented along with supporting simulation results and analysis. A novel 5T SRAM bitcell aimed toward power-efficient embedded SRAM applications is presented in Chapter Five. Chapter Six concludes this research and summarizes the achievements.

1.7 Summary

In this chapter we presented the importance of embedded memories in SoC applications. A brief introduction to CMOS device and power supply voltage scaling trends was presented. CMOS device performance enhancement due to technology advancements, such as the introduction of high-K and HK-MG technology in state-of-the-art CMOS processes, was also highlighted.

We highlighted the SRAM bitcell and the scaling trends of key parameters and their effect on SRAM cell reliability. Furthermore, the influence of process and mismatch variations on miniaturized nanometric CMOS device properties and the impact these have on SRAM cell reliability is presented. Variations in transistor threshold voltage V_{TH}, in particular, cause unpredictable fluctuations in an SRAM bitcell's ability to drive a heavily loaded bitline. Existing circuit-level solutions used to mitigate SRAM drivability limitations were discussed including the use of new bitcell topologies and the use of offset-insensitive sense amplifiers.

Chapter 2

SRAM Architecture and Bitcell Circuit Design

2.1 SRAM Architecture

A memory bitcell is the primary building component in the memory unit. Each bitcell is capable of storing a single binary digit, known as "bit". The SRAM bitcell stores data in a complementary fashion at two nodes. A number of bits (typically 8, 16, 32, or 64) constitute a "word". A "row" is a high-level memory structure which is used to connect a number of words to a common control signal, known as a wordline enable (WLE). Vertically, a number of bitcells are stacked on top of each other, and share a pair of control signals known as "bitlines". A set of cells that share a pair of bitlines constitute a "column". The bitline that corresponds to the data node, is referred to as Bl, whereas the other bitline which corresponds to the complement data is referred to as Blb. From a top-level view, the memory unit can be thought of as an ($N \times M$) element array which has "N" rows and

"M" columns in which each bitcell is assigned by a (row, column) address (similar to point coordinates in the x,y plane).

In a multi-word per row architecture, a column interleaving technique is usually used. In this technique, the i^{th} bits of all words are laid out adjacent to each other in a patch. A column multiplexer is used to activate one column in the patch to manipulate its bitline variations. A column interleaving technique allows the use of a single sense amplifier per column patch. The sense amplifier layout can thereby occupy X times the column layout pitch, where X represents the number of words per row. Another advantage of column interleaving is the immunity against multi-bit errors caused by layout catastrophic defects or soft errors due to cosmic ray bombardment. A defect in one location can lead to a single bit error in the X words instead of causing X errors in one word. The single bit error can be either tolerated or fixed as opposed to two or more bit errors that can result in storage failure.

A high-capacity memory unit is usually divided into a number of blocks or banks. All the blocks share the same control signals but are individually addressed by the address decoder. In other words, when the k^{th} row is selected, all rows of the same order (K) in all blocks are selected but only the one on the selected block is activated. Multi-block architecture requires additional address bits in the address bus, such as $K = 2^Z$, where Z is the number of blocks.

In addition to the memory array described above, the memory unit has other peripheral circuits used to access and manipulate the data of each individual bitcell. **Figure 2.1** shows a typical SRAM unit architecture. In the following subsections, a brief introduction to each circuit in the memory unit is presented.

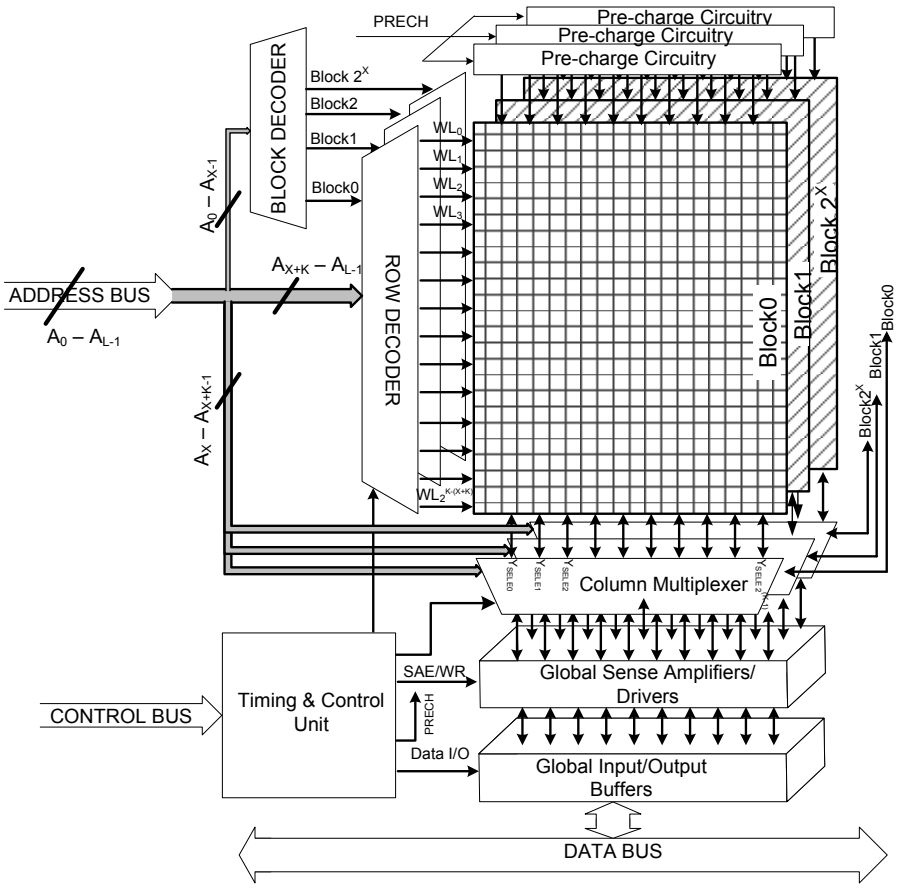


Figure 2.1: Typical Multi-Block SRAM Unit Architecture.

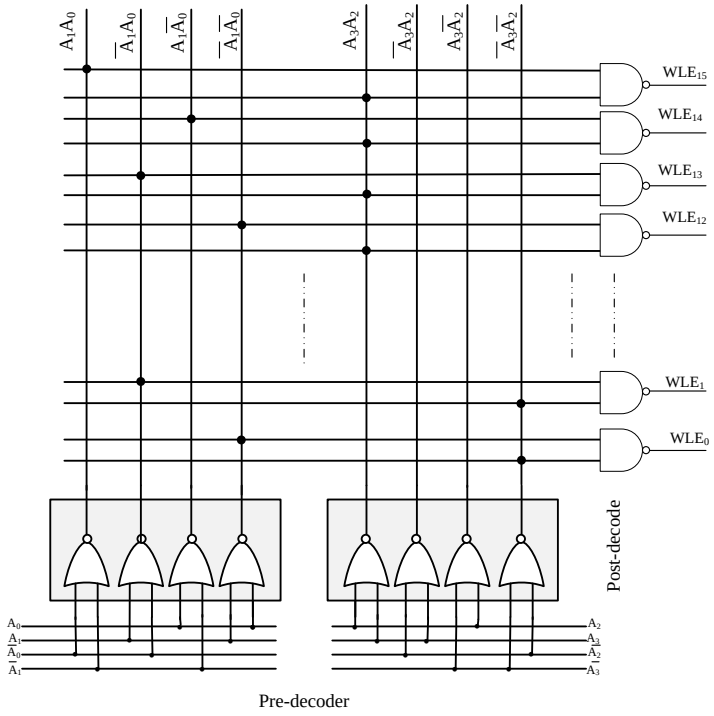


Figure 2.2: Two Stage 4-16 Row Decoder Implementation.

2.1.1 Row Address Decoder and Column Multiplexer

A row address decoder is used to activate one out of N rows in the array. The address bus width required to address N rows is A_0 to A_{n-1}, where 2^n=N. High-density SRAM arrays, (*i.e.,* with a large number of rows) use multiple stage row decoders. In such a case, the output of the first stage decoder (pre-decoder) is multiplexed with the output of the second stage decoder (post-decoder). **Figure 2.2** illustrates a simple 4-16 two stage row decoder. As can be seen in the figure, the maximum number of inputs of the NAND gates used in this decoder is limited to two. This is significant in nanometric CMOS technologies where a stack of more than two transistors is not recommended due to transistor body effects.

A column decoder, on the other hand, is used to address a number of columns (equal

to the number of bits used in each word); therefore, relatively few address bits A_n to A_m, where 2^m= number of word/row, are needed. The column decoder, which is usually referred to as a column multiplexer, generates a Y_{SEL} signal to activate the addressed columns. As a result, the total number of address bus bits required to address each word in the array is (n+m). The following example explains how to address four words in a 32-Kbit SRAM array.

A 32-Kbit memory array can be built as 256 rows X 128 columns. The 128 bits can be divided into four 32-bit words. In order to address one row out of the 256, the row address decoder needs 8 address bits, that is 2^8=256. This can be accomplished by using an (8-256) row address decoder. The column multiplexer, on the other hand, needs to address one out of four words. Therefore, a (2-4) decoder, requiring 2 address bit, is used as a column multiplexer. The required address bus length, in this case, is 10 bits (A_0 to A_9). The first 8 bits (A_0 to A_7) are used for the row decoder and the last two bits (A_8 to A_9) are used for column multiplexer.

2.1.2 Timing and Control Unit

An SRAM bitcell is a synchronous system, so each memory activity starts and ends according a restricted timing scheme based on a system clock signal (CLK). Failure to meet a specified timing constraint results in a cell malfunction. The timing block is a crucial component of the SRAM memory macro. The main objective of the timing block is to synchronize different control and activity signals so that no signal leads or lags the time it is designed for.

The first rising edge of the CLK signal triggers the timing block to generate a precharge control signal to deactivate the precharge circuit and to start the evaluation phase. At the

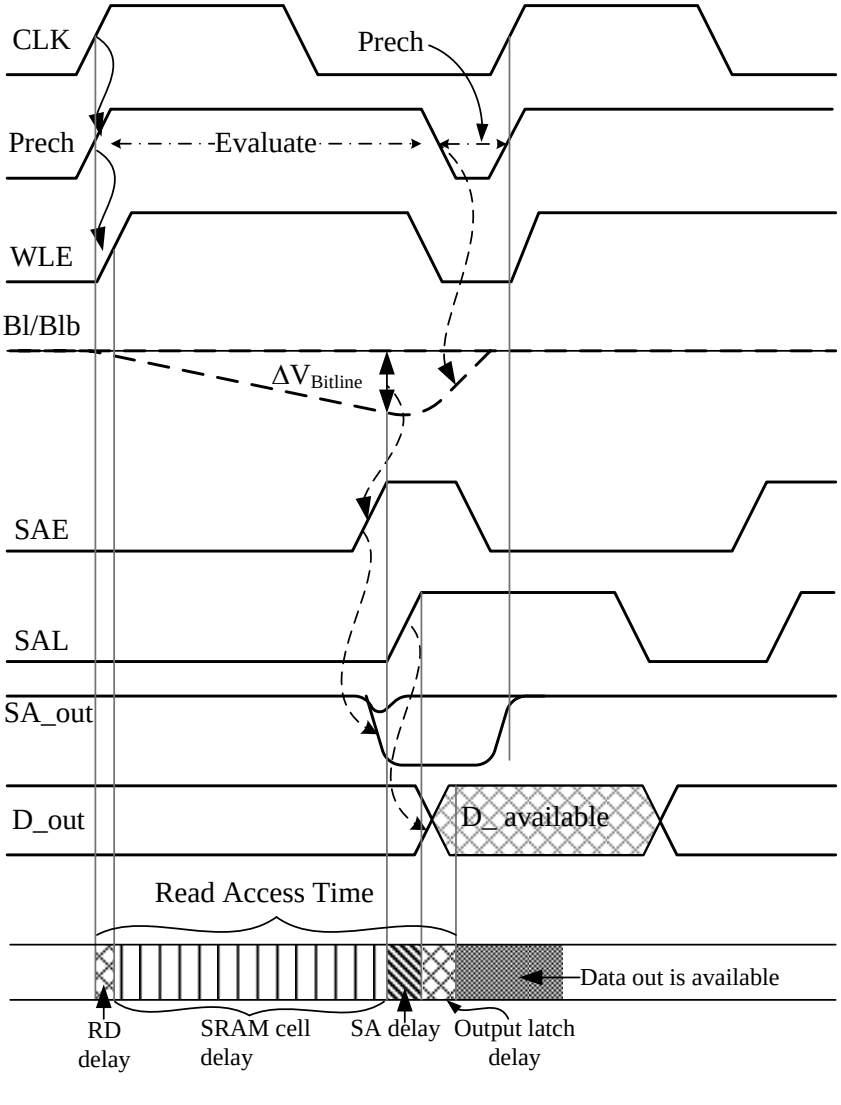


Figure 2.3: Typical 6T SRAM Timing Scheme.

same time, the row decoder assigns a row driver to generate the WLE signal according to the input address. The timing block ensures that the WLE signal is completely contained within the evaluation time period in order to avoid a direct current path from the precharge circuit to ground through the cell. Once accessed, the cell starts communicating with the bitlines and developing a bitline differential voltage in a time interval Δt.

A read/write control signal (WR) tells the timing block to activate either the sense amplifier (a read operation, WR low), or the write driver (a write operation, WR high). When WR is low, the timing block asserts a sense amplifier enable signal (SAE) after a time delay Δt. The sense amplifier makes a decision based on its differential input voltage ($\Delta V_{Bitline}$) and generates a full swing output signal. After some delay time, the timing block asserts a sense amplifier output latch (SAL) signal to activate the sense amplifier output latch and have the read data ready at the output data bus.

The time it takes the cell to perform a successful read operation is known as "read-access time". This time interval is defined as starting from the CLK rising edge until the data-out latch latches the sense amplifier output and buffers it to the data bus. The timing block activates the write driver, when WR is high, to perform a write operation. After the memory activity is accomplished, the timing block disables the WLE signal, isolating the written cells on the addressed row, and enables the precharge to start a precharge phase. **Figure 2.3** shows a complete memory read operation timing scheme with read-access time definition highlighted.

2.2 SRAM Column Structure

The column is a main building block of a memory array. Typical SRAM column structure comprises the following components: (1) a number of SRAM bitcells, (2) a precharge and

equalization circuit, (3) a write driver, and (4) a sense amplifier. The bitline pair behaves as communication media between the column's components, on the one hand, and the outside world, on the other hand. **Figure 2.4** illustrates a typical architecture of a conventional SRAM column.

2.2.1 Precharge Circuit

The precharge circuit is responsible for providing the bitlines' initial conditions using a precharge signal. During the precharge phase, the memory cell is not being accessed and the bitlines are usually held and equalized to a high voltage level (usually V_{DD}). Based on the memory bitcell design, the precharge voltage level could be any reference voltage. **Figure 2.5** depicts two traditional precharge circuits used in an SRAM array.

Whereas PMOS transistors are used to precharge the bitlines to V_{DD}, NMOS transistors are used to precharge the bitlines to a voltage level other than V_{DD}. The maximum precharge level an NMOS transistor can provide is V_{REF}-V_{THn}, where V_{REF} is a reference voltage level and V_{THn} is the transistor threshold voltage. The transistor sizes in the precharge circuits are selected based on the expected bitline loading. Heavily-loaded bitlines require relatively strong precharge transistors. The equalizer transistor M3 is usually of minimum size and is used to equalize the bitlines' initial voltage. The two extra PMOSs in **Figure 2.5** (M5 and M6) are used for bitline leakage compensation; hence they are usually made weak (minimum size).

2.2.2 Write Driver

The write driver is the data input device in a memory unit. It transfers the data from the data bus to the addressed cell via the bitline pair. The write driver's function is to pull

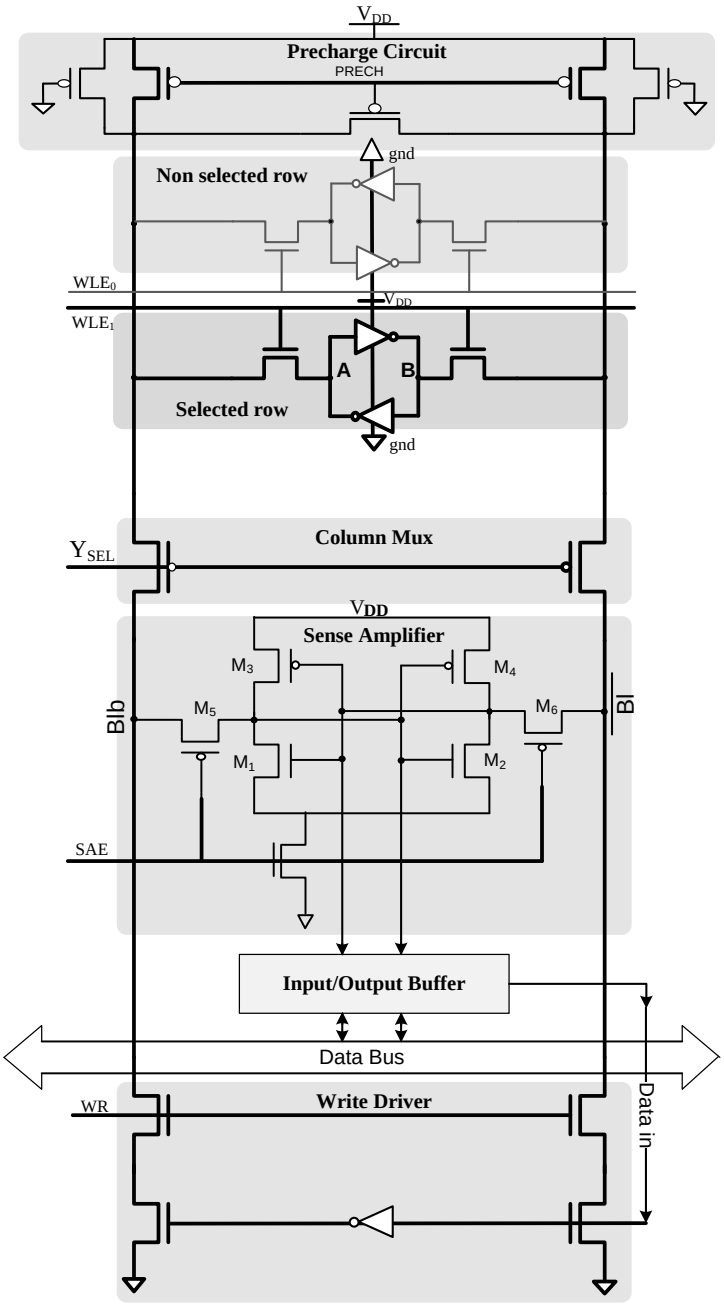


Figure 2.4: Typical SRAM Column Structure.

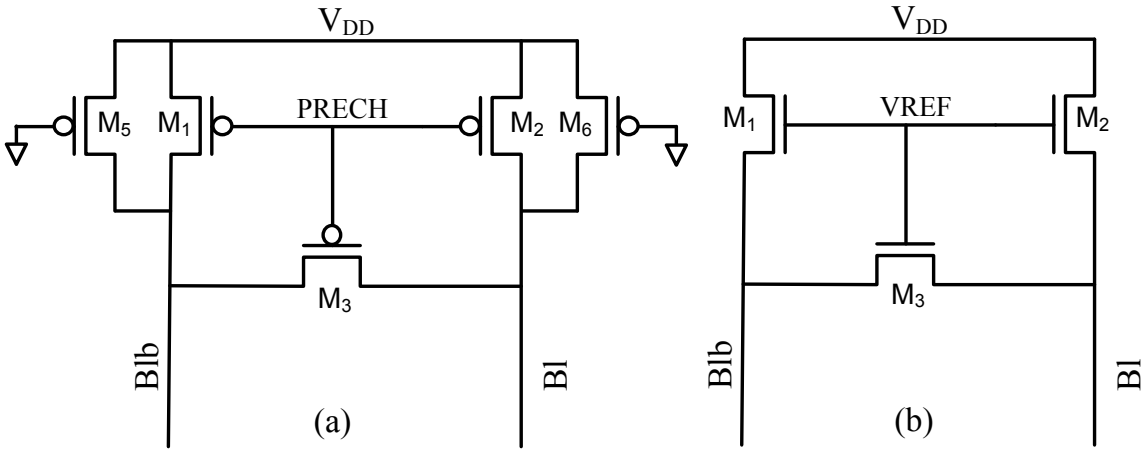


Figure 2.5: Traditional Precharge Circuits.

one of the two bitlines down to ground according to input data and the WR control signal. For simplicity, a write driver is usually designed to manipulate Bl according to the input data *,i.e.,* if the input data is "0", Bl is pulled down to write a "0", but if the input data is "1" then the Bl stays at a high voltage level while the Blb is pulled to ground.

The simplest way to implement a write driver is by performing an AND operation (multiplex) between the WR control signal and the input data. **Figure 2.6** shows two approaches to implementing an AND gate write driver commonly used in SRAM arrays. **Figure 2.6(a)** shows a pass transistor AND gate implementation which requires small area but has slow response due to the use of a series NMOS transistors. In **Figure 2.6(b)**, the AND operation is performed separately and the output drives the bitline by a single NMOS transistor. This approach is faster but this speed comes at a cost of area overhead.

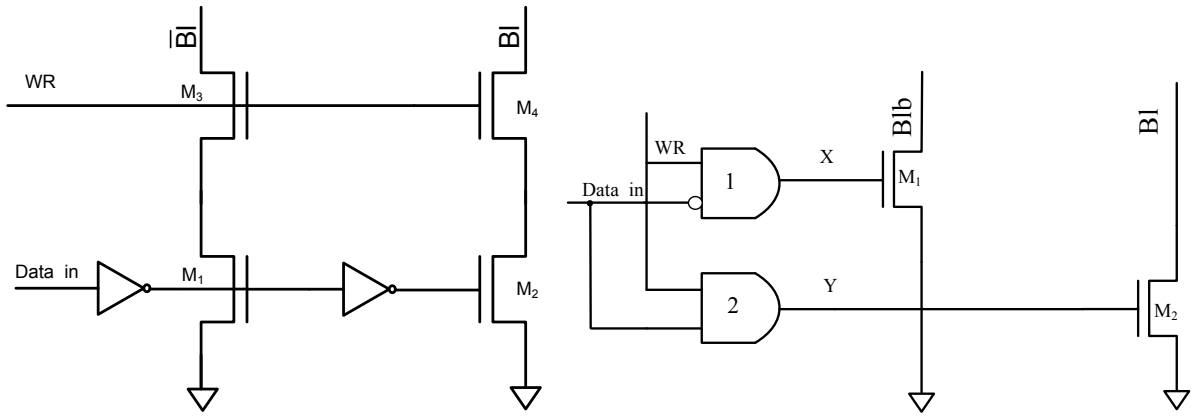
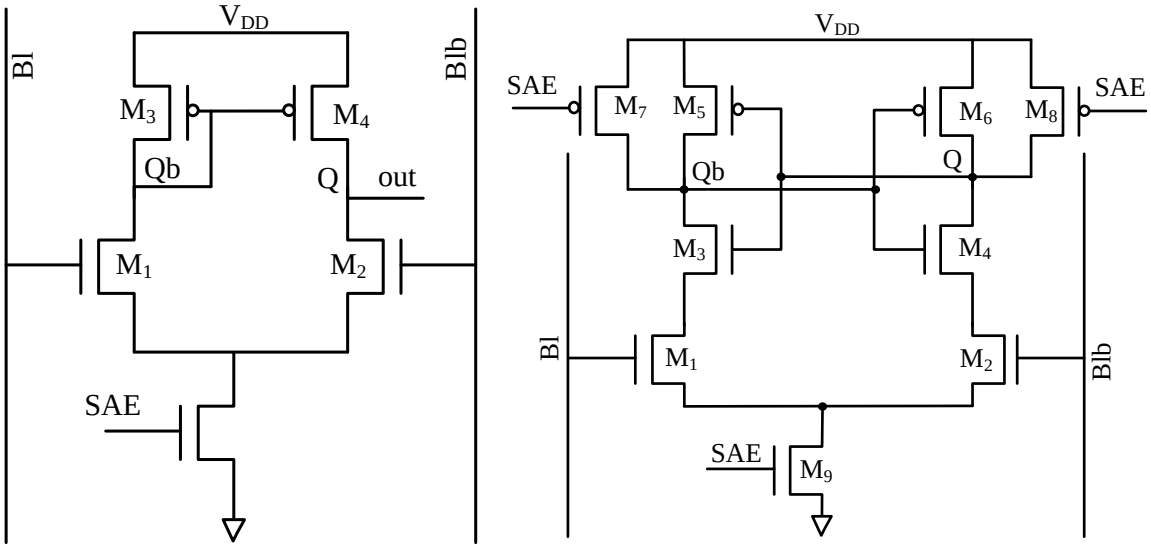


Figure 2.6: SRAM Write Driver Circuits.

2.2.3 SRAM Sense Amplifiers

The primary objective of a sense amplifier in an SRAM array is to amplify a small bitline differential voltage swing to a full-swing logic output. Because of the 6T SRAM bitcell's differential nature, the 6T bitcell-based SRAM arrays usually employ a differential sense amplifier. The conventional differential voltage sense amplifier with a current-mirror active load or with a current-latch cross-coupled inverter load, shown in **Figure 2.7(a)** and **(b)**, were typical choices for 6T SRAM arrays in the older CMOS technologies (up to CMOS 180-nm process). This kind of sense amplifier is easy to implement and operates with reasonable speed and power consumption.

Transistors M1 and M2 are the amplifier's differential input pair. The current-mirror active load, comprises PMOS transistors M3 and M4, is used to increase the amplifier gain by increasing the output impedance as defined in **Equation 2.1**. According to **Equation 2.1**, the amplifier's gain "G" be increased by increasing the differential pair transconductance (g_m) by widening transistors M1 and M2. A Larger differential input pair M1, M2 not only increases the amplifier gain but also contributes to the amplifier's offset voltage



(a) Differential Voltage Sense Amplifier (b) Current Latch Sense Amplifier

Figure 2.7: Conventional Differential Voltage Sense Amplifier.

reduction by reducing the impact of V_{TH} variations as defined in **Equation 1.1**.

However, this comes at an expense in area and power overhead. The key constraint for the sense amplifier layout pitch is the column's narrow width. Therefore, a sense amplifier has a restricted area constraint. However, if column interleaving technique is used, the sense amplifier layout pitch can be relaxed to span a number of column layout pitch.

$$G = -gm_2 \times (r_{o2}//r_{o4}) \tag{2.1}$$

In modern CMOS technologies, process and mismatch variations in the sense amplifier differential input transistor pair create a large input offset voltage. At the same time, the cell's ability to drive heavily-loaded bitlines to develop adequate bitline voltage swing in a given time Δt is significantly diminishing (**Equation 2.6**). This degrades the reliability

of any sensing scheme that requires the development of a sufficient differential signal to initiate the sensing operation. The key strategy to overcoming the offset voltage limitation is to use a high sensitivity sense amplifier that can make a right decision with a very small bitline signal swing.

One effective solution for accurate power-efficient, high-speed sensing is the use of a current sense amplifier. The current sense amplifier's operation is based on the bitline differential current created by a cell read operation, irrespective of bitline voltage swing [22][23][24]. The first offset voltage insensitive sense amplifier was proposed by [22]. This scheme provided a more than 66% improvement in delay compared to a conventional voltage sense amplifier.

In addition, the most important feature of current mode sense amplifiers is that its operation is offset voltage insensitive, *i.e.,* it can sense a very small bitline differential voltage swing; therefore, the sense amplifier delay is almost independent of bitline loading. Another current mode sense amplifier, proposed in [25], seems more attractive in terms of having fewer transistors, low power consumption and high speed. It is a single-stage amplifier and can be used per column due to its compact layout pitch. The performance of this scheme will be discussed in Chapter Three when it is compared to a proposed current mode sense amplifier.

Even though current mode sense amplifiers seem to be attractive for high-performance, low-power applications, the limited cell drivability and column leakage current can compromise the advantages of using current sense amplifiers, especially over a heavily-loaded bitline. Therefore, cell read-assist techniques have garnered more attention in the last few years. The read-assist mechanism is used to assist the cell in developing a targeted bitline swing or differential current. The main idea of the read-assist circuit is to provide another current path to discharge the bitline during a read operation.

Read-assist and write-back features can be added to the sense amplifier to enhance the cell operation in the nanometric CMOS regime. The voltage latch sense amplifier, used in [9] for example, provides a read-assist through the bitline/sense amplifier coupling so that, when the sense amplifier latches the read data, it discharges the bitline at the same time. If the sense amplifier is designed in a way to fully discharge the bitline, then the SRAM cell undergoes a write operation. This kind of sense amplifier is known as a "read-assist write-back" sense amplifier. Details on sense amplifiers will be presented in Chapter Three.

2.3 SRAM Bitcells: An Overview

The first appearance of a full CMOS 6T SRAM bitcell was in 1972 [26]. The storage mechanism of this cell is based on the operation of the active Flip-Flop structure as shown in **Figure 1.7**. High- stability, noise immunity, negligible static power consumption, and compatibility with standard CMOS logic are among many other outstanding features that characterize this topology. Despite these exceptional features and characteristics, the transistor count of the 6T topology (cell area) was a prime concern in high-density integrated systems. As such, many other topologies emerged in the field in the following years to reduce the 6T cell area. For example, an asymmetrical 5T cell [14] and a 4T with Poly-silicon, or Thin-Film Transistor (TFT) load cell were proposed in the early 1970s as potential area-efficient replacement topologies.

Because of its high integrity, the 4T with a poly-silicon load SRAM cell topology, shown in **Figure 2.8(a)**, succeeded in dominating the market of dedicated SRAM memories for awhile. However, the high static power consumption during idle mode prevented the use of this cell in power-conservative systems. The next generation of the 4T cell employed a TFT as a high resistance active load to reduce static power consumption [10]. The main

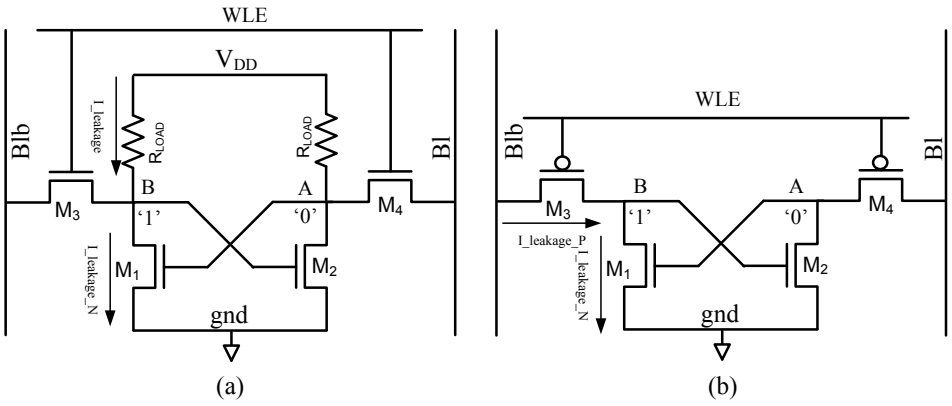


Figure 2.8: Conventional SRAM Cells, a) 4T With Resistive Load, and b) 4T Loadless.

concern in the use of a TFT was the necessity of using more than one fabrication mask due to the existence of two different processes in one design. In addition to its complexity and costly implementation, the ability of this cell topology to retain data at low operating voltage is exceptionally poor.

Recently, a full CMOS four transistor (4T) loadless SRAM bitcell, shown in **Figure 2.8(b)**, has been reported [27]. This cell topology has been considered a very attractive choice for high integration applications due to its compatibility with standard logic CMOS and the small number of transistors required to realize it. The cell consists of two cross-coupled NMOSs serving as drivers and two PMOSs serving as access transistors to link the cell's storage nodes to bitlines.

Data retention in this topology is carried out via a leaky PMOS access transistor, so the PMOS leakage current is made at least two orders of magnitude greater than the NMOS leakage, otherwise the cell loses the data after some time. In order to balance the two leakage current components, the cell is implemented in a dual V_{TH} process. The PMOS transistor is made low LV_{THp}, and the NMOS is made high HV_{THn}. Static power consump-

tion associated with this topology is considerable which makes this topology unsuitable for low-power battery-operated systems. Furthermore, read and write operations in this topology are relatively slow because of the use of PMOS transistors as access transistors.

Further advances in CMOS technology and aggressive scaling in device minimum feature size have allowed high-density integration due to minimized bitcell area. Therefore, conventional 6T integration is no longer as a concern as before. In fact, the high integration capabilities have allowed the use of more than six transistors to realize a bitcell that has the ability to mitigate process and mismatch variations associated with the conventional 6T bitcell, and, therefore, to enhance the bitcell's reliability without compromising integration.

Figure 1.9 summarizes state-of-the-art non-6T SRAM bitcell topologies. These topologies are characterized by two common features: first, they all employ the conventional 6T SRAM bitcell as a core storage element, which is also used to perform a write operation. Second, they all utilize extra transistors to separate read/write ports from each other for performance enhancement. The basic idea behind these topologies is to isolate the cell's storage nodes from the bitline pair to eliminate any loading impact on cell stability. As a result, these bitcells are process variations-tolerant and are capable of operating at an extremely low supply voltage (subthreshold).

Even though the two extra transistors added in the 8T cell result in about 30% area overhead, this cell topology is considered the most promising replacement candidate for the conventional 6T [28] [29]. Whereas the conventional 6T SRAM cell fails at low operating voltages, a 200 mV 8T SRAM cell has been successfully realized and reported [30][2]. In fact, the resulting area overhead is acceptable considering the performance improvement achieved with the use of an 8T cell. Furthermore, it has been reported that under the same operating conditions, the 8T and 6T areas start to cross over at 32 nm node, since, at this

technology cell and beyond, the 6T cell can no longer reliably operate with minimal size devices.

Nevertheless, these topologies suffer some drawbacks. The use of the conventional 6T as a core storage element has moved the 6T stability problems to these topologies. For example, although read and write ports are separated from each other, half-selected cells on the same row still perform a read operation on their 6T cell, and, since the 6T cell is not designed for robust read operation, extra caution must be taken to ensure the stability of half-selected cells.

In [31] the 8T half-selected cell problem is addressed by using a "byte write" technique [32]. In this technique the write wordline signal is gated to turn "on" only one selected block of a certain number of cells, then the entire selected cells on that block perform a write operation, *i.e.,* there are no half-selected cells. Another approach is followed in [29] by using a write-back scheme. In this scheme, both read and write wordline are activated during each memory activity. The half-selected cells then perform a read operation and the sensed data is used to write-back the cell. Noticeably, power consumption associate with both schemes ([31][29]) is considerably large.

The ten-transistor (10T) symmetrical bitcell topology proposed in [5] seems a feasible solution for the 8T bitcell half-selected problem. In this topology half-selected cells are separated from selected ones by adding extra pass gate transistor. This transistor is turned "on" during a write operation and kept "off" during read operation. Advantages of this scheme are the absence of a dedicated read bitline and the bitcell's symmetrical structure. Disadvantages, however, are represented by the need for two extra transistors (area overhead) and to activate both read and write wordline during a write operation (power overhead).

In conclusion, the conventional 6T SRAM bitcell is the foundation for all bitcell topologies, and stable and reliable cell design is crucial in embedded systems reliability. The outstanding features that the 6T topology exhibits continue to make it the topic of innumerable studies and research. This bitcell topology has, therefore, been adopted in this work as a benchmark and will be emphasized in the following sections.

2.4 Six-Transistor (6T) SRAM Background

The core storage element in the basic 6T SRAM cell, shown in **Figure 1.7**, is a two back-to-back inverter structure comprising transistors M1-M3 and M2-M4. The NMOS transistor of each inverter, M1 (M2), is called the driver, and the PMOS transistor M3 (M4) is called the load. This architecture acts as an active latch to statically preserve the state of the cell. The two storage nodes, A and B, are linked to the outside world via two pass gate transistors, M5 and M6, known as access transistors.

In general, the cell has two modes of operation: retention mode and access mode. During retention mode, the WLE signal is deactivated and the storage nodes are isolated from the bitlines. The latch action during this mode helps retain data as long as the cell is powered with very low static power consumption.

The cell enters access mode when the WLE signal is activated. The "on" access transistors allow the cell to communicate with the bitlines. The bitline/storage node interaction depends on the intended operation. If a read operation is intended, the bitlines transfer data from the cell's storage nodes to the outside world. On the other hand, the bitlines transfer data from the outside world to the cell's storage nodes when a write operation is forced by a low-impedance write driver.

2.4.1 6T SRAM Cell Characterization

Reliable 6T bitcell design must ensure the cell's ability to perform the aforementioned modes of operation under worst-case operating conditions. Furthermore, the cell must be capable of performing read and write operations with adequate voltage margins. Therefore, 6T bitcell design is based on a balanced transistor size. Successful read and write operations require proper transistor ratios. The cell's driver-to-access transistor defines the "cell ratio" (CR or β) and the cell's load-to-access transistor defines the cell "pull-up ratio" (PR or α). These two ratios are key elements in cell stability and reliability.

2.4.2 Read Operation

Initially, the column precharge circuit holds the bitlines at a high voltage level (typically V_{DD}). A read operation is initiated upon the activation of the WLE signal. The WLE signal turns the access transistors M5 and M6 "on" and a high voltage ("1") at node B turns the driver transistor M2 "on"; whereas, the low voltage level ("0") at node A keeps the second driver M1 "off". Accordingly, the cell discharges the bitline (Bl) below V_{DD}, whereas the opposite bitline (Blb) stays high (V_{DD}). Consequently, a differential voltage $\Delta V_{Bitline}$ is created between the two bitlines.

Figure 2.9 highlights the cell's schematic diagram and the corresponding cell transient response under read access mode. The lightened transistor's symbol signifies a transistor in an "off" state, whereas the dark symbol signifies a transistor in an "on" state. The resistance of the "on" driver and access transistors form a potential divider. The cell current passing through this potential divider creates a voltage level above zero at node A, known as a "zero level degradation" (Δ).

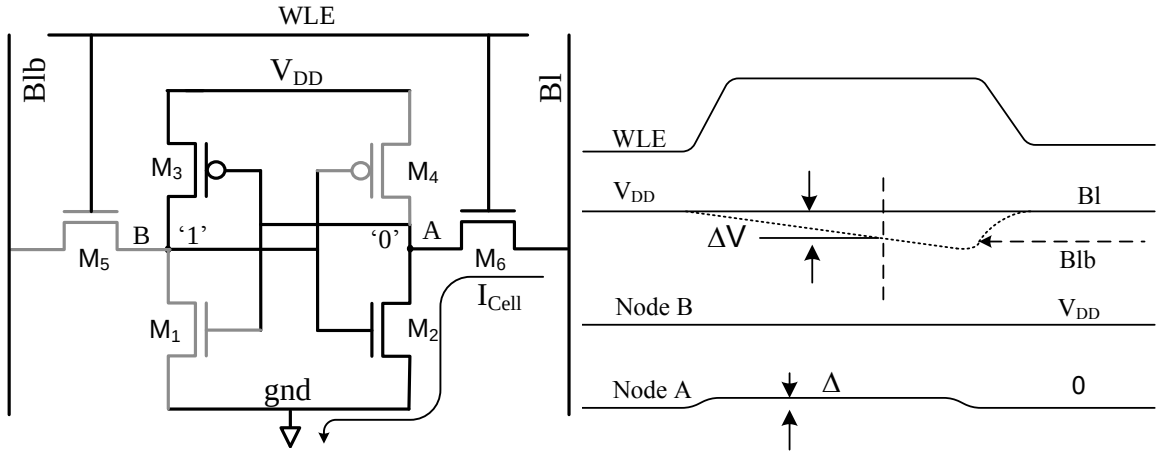


Figure 2.9: 6T SRAM Cell Behavior During a Read Operation: Schematic and Timing Diagrams.

In order to ensure a successful read operation, Δ must stay as low as possible. The cell ratio β determines the zero level degradation degree such that a high cell ratio (wide driver and narrow access) results in a low Δ and thereby a stable read operation. If (Δ) exceeds V_{THn}, M1 can gradually turn "on" by the rising potential at node A. The positive feedback configuration of the crossed-coupled inverter exacerbates the voltage level degradation and maximizes the loop gain until the cell flips. This process is known as a "destructive read operation (DRD)" and has to be avoided.

Equation 2.2 gives the relationship between cell ratio β and level degradation Δ. **Figure 2.10(a)** shows SPICE simulation results obtained during a cell read operation to show the zero level degradation dependency on the cell ratio β. According to the simulation results, if V_{THn} is assumed to be 200 mV, then a cell ratio equal to or less than 1.2 can lead to a destructive read operation. In other words, a safe read operation requires a driver transistor that is at least 1.2 times stronger (wider) than the access transistor. Furthermore, large, β, and thereby small Δ, increases the voltage margin the cell can

tolerate without losing the stored data, *i.e.,* higher cell stability. This is verified by the SPICE simulation results shown in **Figure 2.10(b)**.

$$\Delta = \frac{V_{DSAT} - \beta(V_{DD} - V_{THn}) - \sqrt{{V_{DSAT}}^2(1+\beta) + \beta^2(V_{DD} - V_{THn})^2}}{\beta} \tag{2.2}$$

where β is the cell ratio, which is given by **Equation 2.3**:

$$\beta = \frac{(W/L)_{Driver}}{(W/L)_{Access}} \tag{2.3}$$

The cell speed is determined by the time it takes the cell to generate a targeted bitline differential voltage. This time is variable and depends on the bitline loading and the cell's drivability. The sense amplifier is activated at time instant t_{SA} to sense the bitline differential voltage and generate a full swing logic output signal that reflects the sensed data. When the sense amplifier outputs are recovered to a full swing logic, the sense amplifier latch (SAL) latches the data and buffers it to the outside world through the data bus (see **Figure 2.4**). In conclusion, from a read operation perspective, larger drivers (M2, M1) and small access transistors (M6, M5) produce a higher cell ratio β and hence a high stable cell.

2.4.3 Write Operation

Unlike a read operation, a write operation is initiated by activating the write driver first to discharge Blb then activating the WLE. Once the WLE is activated, node B discharges down toward gnd and node A charges up toward V_{DD} via the access transistors M5 and M6, respectively. The positive feedback mechanism of the cross-coupled structure accelerates the voltage-level degradation and flips the cell. In order to do so, the voltage level at node

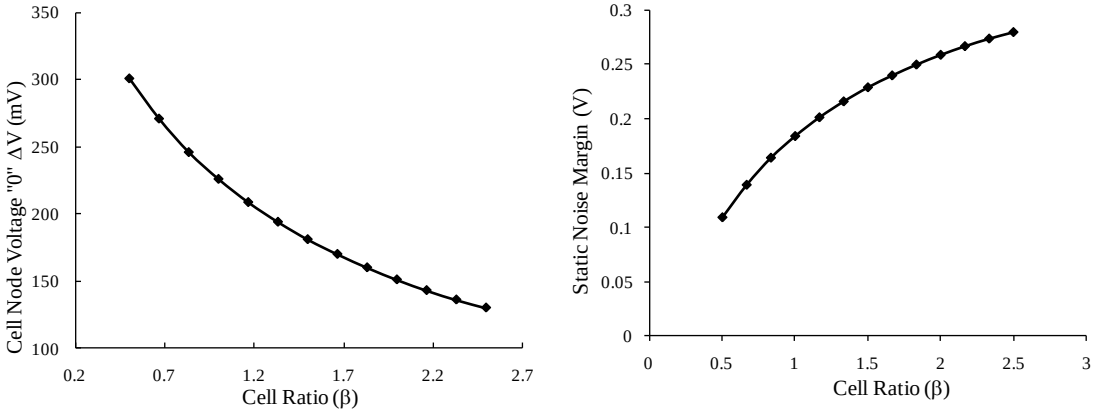


Figure 2.10: Zero Level Degradation (Δ) and Cell Voltage Margin as a Function of Cell Ratio β.

B must be lowered below the inverter M2-M4 trip point. The voltage level at node B is determined by the M3-to-M5 ratio which is set by the cell's pull-up ratio α. **Figure 2.11** illustrates cell behavior under a write operation condition.

Equation 2.4 signifies the relationship between the voltage level at node B and the cell pull-up ratio α. This relationship is verified by the SPICE transient simulations depicted in **Figure 2.12**. If the inverter trip voltage is assumed to be 400 mV, then α values of up to 3 are acceptable. Due to the mobility difference in NMOS and PMOS transistors (μn $\simeq$ 2μp), same size NMOS drivability is approximately double that of the PMOS. Thus, the use of minimal size load and access transistors results in an α ratio that equals approximately 1.0, but the effective PMOS-NMOS strength is approximately 1.5-2.0. This relaxes the cell devices' balance and gives more design flexibility to strengthen the load and/or weaken the access transistor. A strong load transistor increases the cell's ability to retain the data, whereas a weak access transistor increases the β and hence decreases the zero level degradation.

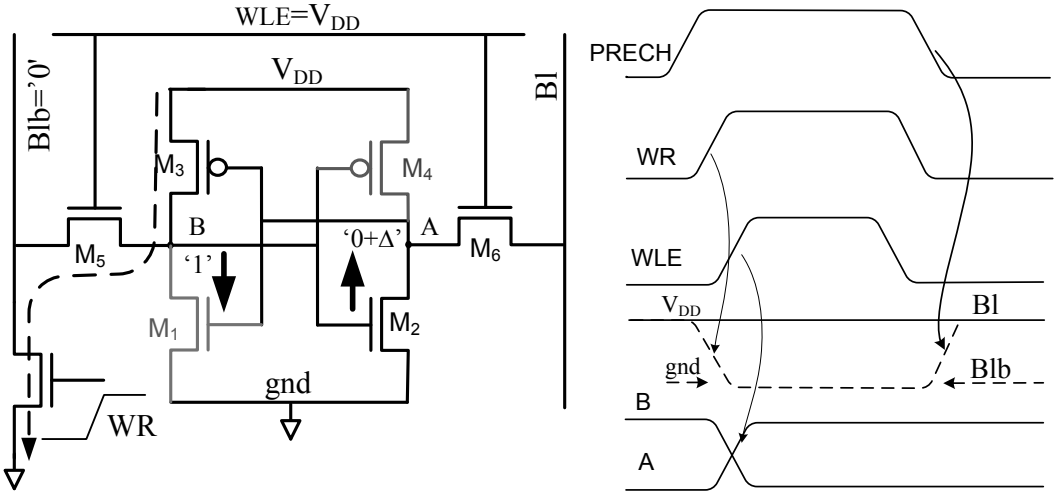


Figure 2.11: 6T SRAM Cell Behavior During a Write Operation: Schematic and Timing Diagrams.

$$\Delta = (V_{DD} - V_{THn}) - \sqrt{(V_{DD} - V_{THn})^2 - 2\frac{\mu_p}{\mu_n}\alpha(V_{DD} - V_{THp})V_{DSATp} - \frac{V_{ASATp}^2}{2}} \quad (2.4)$$

where α is given by:

$$\beta = \frac{(W/L)_{Load}}{(W/L)_{Access}} \quad (2.5)$$

It is worth mentioning that during a write operation, all the cells located on the accessed row respond to the WLE signal activation. Whereas only those cells located on the selected columns undergo a write operation, cells located on non-selected columns, known as half-selected cells, undergo a normal read operation since their bitlines are floating during this operation.

As we can see, read and write operations in a conventional 6T SRAM cell are interrelated and contradict each other. Stable read operation requires a large driver and a weak

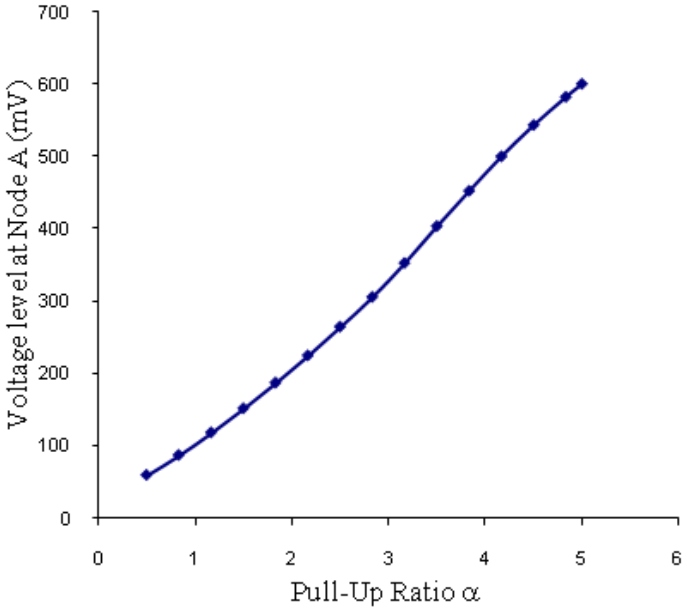


Figure 2.12: 6T SRAM Cell Node High Voltage as a Function of Cell Pull-Up Ratio α.

access transistor; on the other hand, successful write operation requires a strong access transistor and a weak load transistor. Additionally, data retention requires a reasonable load transistor strength to hold the data. As such, a delicate device sizing approach must be adopted to ensure a stable and functional SRAM cell with sufficient read, write and retention voltage margins.

2.5 6T SRAM Figures of Merit

The 6T SRAM cell reliability and performance are measured in terms of a number of metrics used as figures of merit (FOM). These figures are usually used to analyze, characterize, and assess a bitcell topology and to compare alternative SRAM bitcell topologies. Some of these figures highlight the bitcell performance; others highlight bitcell reliability. For

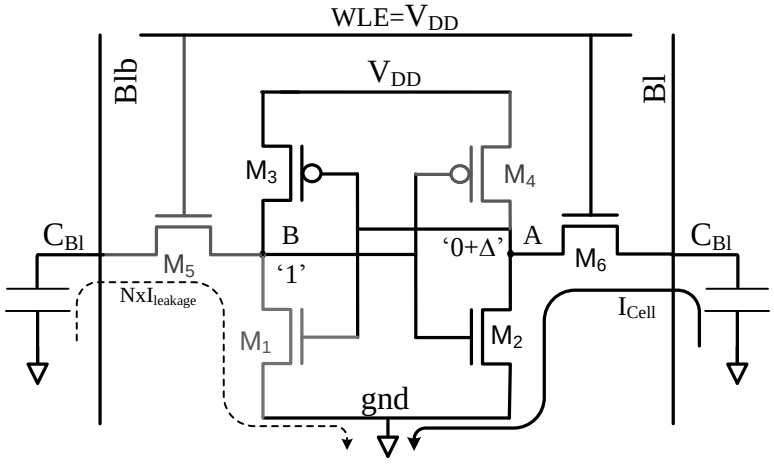


Figure 2.13: 6T SRAM Operation: Cell Drivability.

example, speed and power consumption (active and standby) are performance metrics; on the other hand, noise, read, and write voltage margins are reliability metrics. In this section, we will define each of these metrics and identify design strategies and solutions to maintain or improve each of them.

2.5.1 Cell Speed

SRAM cell speed is measured in terms of the time (Δt) required to generate a targeted bitline differential voltage ($\Delta V_{Bitline}$) during a read operation. SRAM read speed depends on the bitline's loading and the cell's drivability measured by the cell current. During a read operation, the cell current can be thought of as the bitline capacitance discharge current, I_{CBl}, that results in a $\Delta V_{Bitline}$ voltage drop across C_{Bl}, in a time interval Δt, as shown in **Figure 2.13**, and defined in **Equation 2.6**.

$$I_{Cell} \geq C_{Bitline} \times \frac{\Delta V_{Bitline}}{\Delta t} + N \times I_{leakage} \tag{2.6}$$

where, $C_{Bitline}$ is the total bitline loading parasitic capacitance, $V_{Bitline}$ is the targeted bitline differential voltage in a Δt time interval, and $N \times I_{leakage}$ is the total leakage current resulting from N cells attached to the opposite bitline.

I_{Cell} equals the driver's drain-to-source current $I_{DSdriver}$, which is the same as the access transistor current $I_{DSaccess}$ since they are effectively connected in series. In accordance with cell operating voltages, the driver operates in the linear region; whereas, the access transistor operates in the saturation region. However, due to their minimum feature size and high gate-to-source voltage, the short channel effect is likely to drive these two transistors to the velocity saturation region. Therefore, a generic NMOS transistor drain current formula (**Equation 2.7**) can be used to calculate the cell current I_{Cell} [1].

$$I_{Cell} = K_n \times (W/L) \times [(V_{GS} - V_{TH})\, V_{min} - V_{min}^2/2] \tag{2.7}$$

Here: K_n is a device technology parameter equals: $\mu n Cox$; μn is the electron's mobility and Cox is a technology parameter equals to the gate oxide per-unit area capacitance, W/L is the transistor width-to-length ratio, and V_{min} is the minimum of transistor overdrive voltage V_{ov}, velocity saturation voltage V_{DSAT}, or actual drain-source voltage V_{DS}.

Considering **Equations 2.6** and **2.7**, the cell current and, hence, the cell speed can be increased in two ways: first, reducing the bitline's loading (C_{BL}), and second, increasing the transistor W/L ratio. The first option can be achieved by reducing the number of cells attached to the bitline, thereby reducing the bitline's physical length. This, as a result, reduces the memory capacity. The second option results in bitcell area overhead, due to the physical increase in device dimensions, which degrades the cell array density. For high-density, high-performance SRAM applications, the first option seems feasible since some design techniques, such as the segmented column architecture, can be used to mitigate the

bitline loading limitation.

2.5.2 Cell Noise Immunity

The cell's ability to retain data under different operating conditions is a key element in SRAM stability and hence reliability. A cell is considered stable if it can retain the data indefinitely and can perform successful read and write operations under the worst operating conditions. Noise coming from different sources threatens cell stability whether in retention or accessed mode. Noise immunity in a 6T cell is determined in terms of the amount of noise that the storage node can tolerate without causing the cell lose stored data. Under nominal operating supply voltage, the likelihood of the cell losing data during retention mode is rare since both storage nodes are driven to one power rail or another (active latch). One noise source that can endanger cell retention stability is soft errors that might exist because of cosmic rays or photon bombardment in high radiation operating environments. This kind of noise and cell failure is beyond the scope of this study.

During access mode, the cell/bitline interaction is the major noise source due to the resultant zero level degradation. The zero level degradation makes the cell susceptible to stability problems, specifically when it is combined with other fluctuation factors such as PVT variations. The bitline's influence on cell stability is traditionally estimated by inserting two equal but opposite DC voltage sources (one at each storage node) and sweeping these voltages to observe the DC voltage level at which the cell loses the stored data (flips). The cell's forward and backward inverter voltage transfer characteristics (VTC) are superimposed to generate the cell's VTC curves, also known as the "butterfly curves". Since the applied voltage in this measurement is a DC voltage, the injected noise is considered static and hence the measured figure is called the Static Noise Margin (SNM). This SNM

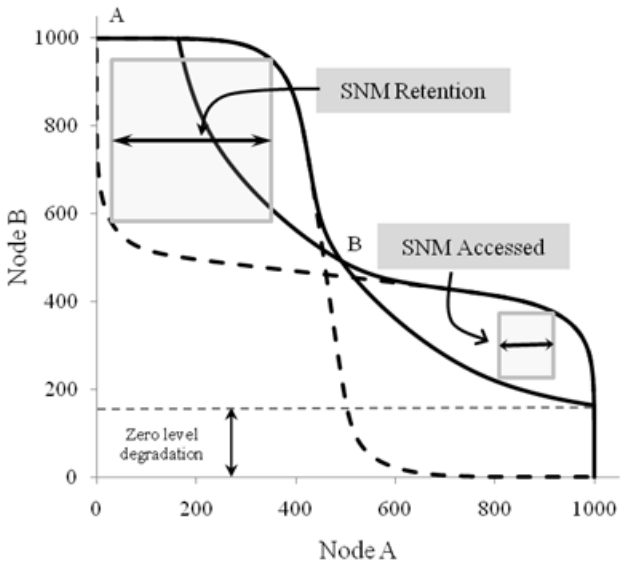


Figure 2.14: Standard 6T VTC Butterfly Curves.

measurement technique was proposed in [33] and is considered to be the main technique to study SRAM cell stability.

Figure 2.14(a) shows standard 6T cell butterfly curves during retention and access modes. The SNM is measured as the side length of the biggest square that can fit in the butterfly curve eye. As can be seen in **Figure 2.14**, the biggest square during retention mode is bigger than that during access mode. A 180-mV zero level degradation results in 60% SNM reduction. It is worth noting that asymmetrical SRAM bitcells, like the 5T or the asymmetrical 6T cells [14] [34], produce asymmetrical butterfly curves due to asymmetrical cell structure. Therefore, the SNM in such cases is measured based on the maximum square that can fit in the bigger eye of the cell's VTC diagram. In fact, the objective of the asymmetrical cell structure is to bias the cell's transfer characteristics to one side and thus maximize the butterfly curves' eye opening.

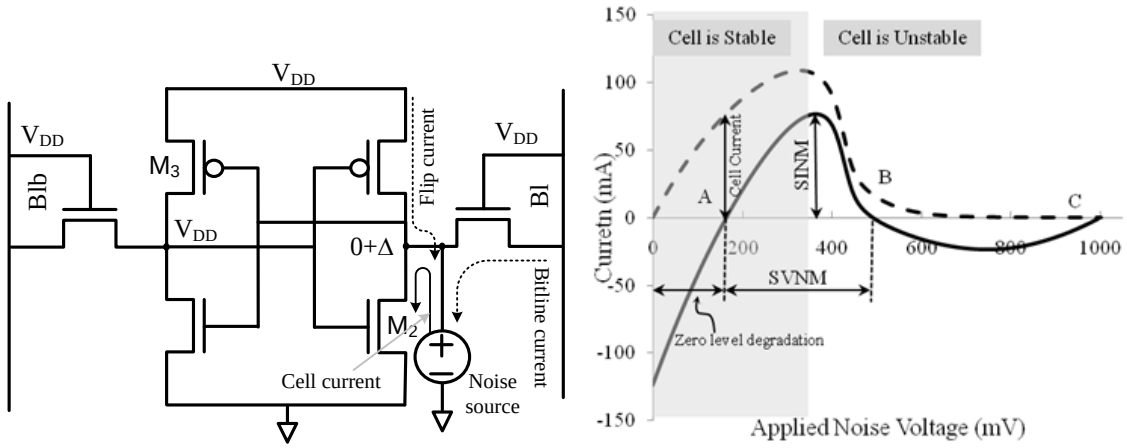


Figure 2.15: The 6T N-Curve Characteristics: Circuit Setup and b) N- Curve Simulation Results.

Recently, a more accurate technique has been reported in which the noise voltage is injected into one node by a voltage source connected to that node. Instead of observing the voltage variation at the opposite node, the current supplied or driven by this source is monitored. This current takes a letter N shape, hence this technique is known as the "N-curve". **Figure 2.15** shows the N-curve measurement circuit setup and the resulting current curve. The solid curve (N-curve) denotes the current sourced or sunk by the noise source, while the dashed curve denotes the cell's drive current resulted from the injected noise voltage at node (0+Δ). As can be seen, the N-curve conveys many cell parameters. Static voltage and current noise margins (SVNM and SINM), the cell current and the zero level degradation can all be calculated from the N-curve simulation results.

Variations in device properties in the nanometric CMOS regime have resulted in shrunken SRAM cell margins. Fluctuations in the few hundred doping atoms used in the channel of state-of-the-art CMOS devices can have an enormous impact on device behavior [12]. In addition, the existence of various transitional noise sources, like power supply noise,

substrate noise and single event upset (SEU) soft errors, in modern CMOS technologies requires different techniques to characterize the dynamic behavior of the device and the memory cell. As a result, the concept of the dynamic noise margin (DNM) came into play.

The DNM is used to analyze cell stability in the presence of variable amplitude and duration noise sources. The basic concept of the DNM evolved from the dynamic behavior of the cell. Data corruption occurs when the cell's storage node capacitance charges (discharges) to a high (low) voltage level which brings both nodes to the cross-coupled structure meta-stable point and the cell become unstable. However, if the storage node RC time constant is made larger than that of the applied noise, the cell can recover the data and return to its stable points ("1" or "0"). It has been shown that the SRAM cell can tolerate high noise levels (more that the estimated SNM) when the cell's access time is shortened [35].

2.5.3 Read and Write Margins

Read and write voltage margins in an SRAM cell determine the voltage limits at which the cell is able to function properly. The cell's read margin (RDM) determines the cell's ability to conduct a successful read operation, *i.e,* the ability to generate a targeted bitline differential, $\Delta V_{Bitline}$, in a given time period Δt. These two parameters, $\Delta V_{Bitline}$ and Δt, are related to cell drivability according to **Equation 2.6**. Thus, for a given cell drivability, I_{Cell}, the cell RDM can be extended either by reducing $\Delta V_{Bitline}$ or by relaxing Δt.

Furthermore, $\Delta V_{Bitline}$ and Δt relate the cell's drivability to the sense amplifier. First, $\Delta V_{Bitline}$ must be large enough to overcome the sense amplifier's offset voltage V_{SA}. Second, Δt must not be greater than the sense amplifier's activation time t_{SA}. As such, the cell's RDM cannot be defined in isolation from the sense amplifier used in the column structure.

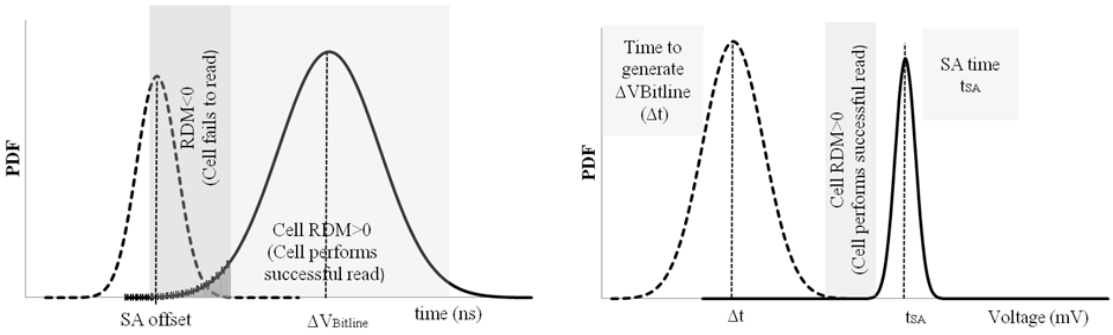


Figure 2.16: 6T SRAM Cell Read Margin Definition.

Figure 2.16 illustrates RDM margin definition as the relationship between the cell and the sense amplifier. In terms of $\Delta V_{Bitline}$ and Δt, in order to maintain sufficient RDM, the probability density function distribution of $\Delta V_{Bitline}$ and the sense amplifier offset voltage V_{SA} must not overlap. Similarly, the probability density function distribution of the cell's delay Δt and sense amplifier delay t_{SA} also must not overlap. If the cell parameters overlap with the sense amplifier parameters, a cell read failure can occur.

Considering the process variation and low operating voltage, satisfying a reasonable RDM in a miniaturized SRAM cell is a major design challenge. In addition to their impact on SRAM device properties, and thereby $\Delta V_{Bitline}$, process variations can manifest as sense amplifier offset voltage V_{SA} variations as well as bitline loading C_{BL} variations due to layout mismatch and non-uniform metal line edges. This wide spread in device properties in modern CMOS technologies due to process variations imposes the use of multiples of the standard deviation (σ) of the cell's parameters to design a reliable SRAM cell. The number of σs required for proper cell design takes into account variations in major cell parameters. The "Z" number, as defined in **Equation 2.8**, sums up the variation in the cell's parameters to determine the required number of σs that need to be covered in designing an SRAM cell [3].

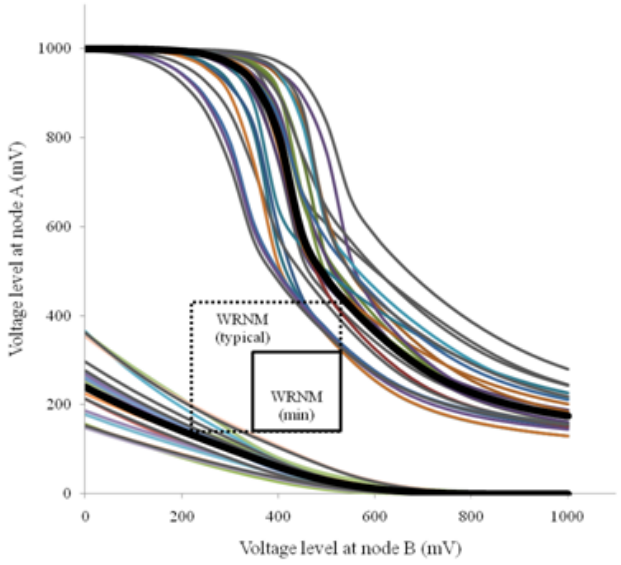


Figure 2.17: 6T SRAM Cell Write Margin Definition.

$$Z = \frac{1}{\sqrt{\left(\frac{\sigma I_{Cell}}{\mu I_{Cell}}\right)^2 + \left(\frac{\sigma V_{SA}}{\mu V_{SA}}\right)^2 + \left(\frac{\sigma \Delta_t}{\mu \Delta_t}\right)^2}} \tag{2.8}$$

During write access mode, the cell write margin (WRM) defines the voltage limit required to flip the cell. This can be accomplished by reducing either the bitline voltage or the cell's supply voltage V_{DD}. In either case, WRM is defined as the lower voltage level required to flip the cell [3]. Graphically, WRM can be quantified by calculating the side of the maximum square that can be embedded between the read and write VTC curves, as shown in **Figure 2.17**. The existence of process and mismatch variations can cause a cell false write operation. **Figure 2.17** illustrates the impact of process variations on cell WRM. A zero or negative WRM is obtained when the two curves touch or cross over each other which indicates that the PMOS load transistor is strong and holds the high storage

node at some non-zero level even if the corresponding bitline is completely discharged to zero and the cell fails to write [36].

2.6 Summary

In this chapter, we presented a typical SRAM array top-level architecture and other peripheral circuits used in the array. Conventional circuits using a typical SRAM column were explored with a brief introduction to each. More importantly, the SRAM cell was reviewed in detail. Different kinds of SRAM bitcell designs were explored; however, the majority of the discussion was devoted to the conventional 6T cell. The basic steps of 6T cell design were presented with definitions for important cell figures of merit. 6T design challenges are investigated and existing solutions were reviewed. The importance of the sense amplifier in SRAM cell operation was highlighted. The superiority of the current mode sense amplifier was justified.

Chapter 3

High-Performance SRAM Sensing Schemes

3.1 Introduction

As we mentioned earlier, system reliability in modern SoC is largely governed by the robustness of embedded SRAM memory. Whereas SRAM bitcells continue to benefit from an aggressive scaling trend in CMOS technologies, interconnect follows a slower scaling trend. Additionally, the bitline capacitive loading is increasing due to the increasing demand for high-density SRAMs. This has resulted in dramatic deterioration in cell drivability due to increased interconnect loading. Moreover, the growing fluctuation in device properties due to PVT variations has added more uncertainty to SRAM operation. Thus, ensuring the ability of a miniaturized cell to drive heavily-loaded bitlines and to generate an adequate voltage swing is becoming challenging. A large percentage of state-of-the-art SoC system failures are attributed to the inability of the SRAM cells to generate the targeted bitline

voltage swing in a given access time which is denoted by failure in read FIR [9].

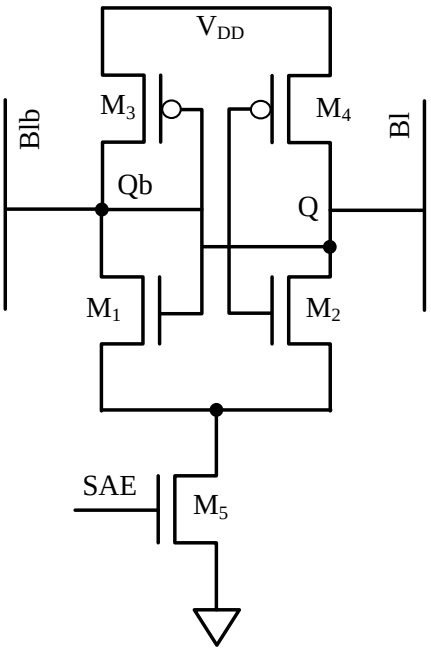
The use of read-assist mechanisms and current mode sense amplifiers are the two key strategies used to surmount bitline loading effects. In the first approach, a read-assist technique is used to reduce the bitline's loading effect by providing additional bitline discharging current path during a read operation. A current-mode sense amplifier is used to sense the bitlines differential current, which is independent of bitline loading, instead of the differential bitline voltage.

3.2 Existing Sense Amplifier Schemes

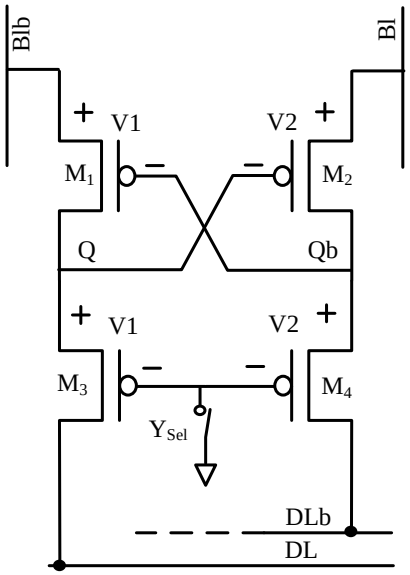
3.2.1 Read-Assist Techniques

One straightforward way to assist the cell during a read operation is to physically reduce the effective bitline loading. This can be accomplished by reducing the physical column length using a segmented bitline with a local sense amplifier technique. In this technique, the cell drives local short bitlines and a local sense amplifier is used to drive heavily-loaded global bitlines. In this case a global sense amplifier is required to amplify the global bitline differential voltage. This solution is beneficial when combined with a dynamic power supply scheme [15]. This allows for powering a selected segment only with a full swing supply voltage, and non-selected segments are kept at a reduced supply voltage swing.

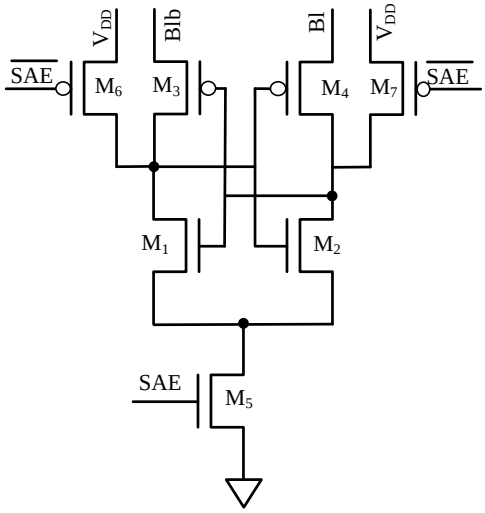
In order to further extend the benefit of bitline segmentation and the use of local sense amplifiers, [9] used the voltage latch local sense amplifier shown in **Figure 3.1(a)**. This scheme provides read-assist and write-back features to enhance the cell's performance. The write-back action is used to eliminate the zero level degradation and maintain cell stability [29].



(a) Voltage-Latch SA [9]



(b) Current-Mode SA [22]



(c) Current-Mode SA [25]

Figure 3.1: Conventional Current-Mode Sense Amplifier.

The local voltage-latch sense amplifier is comprises two cross-coupled inverters (M1-M3 and M2-M4) and a sense enabled NMOS transistor (M5). The sense amplifier output nodes, Q and Qb, are directly coupled to the bitlines so that they can track the voltage variation over the bitlines. If the stored data is "0", the cell discharges the bitline (Bl) below V_{DD} and creates a bitline differential voltage that is applied directly to Q and Qb. Consequently, upon the activation of the SAE signal, the cross-coupled inverter configuration helps Q and Qb to resolve to "0" and "1", respectively and the two "on" NMOS transistors (M2 and M5) discharge the bitline (Bl). A fully discharged Bl resembles write operation conditions; therefore, the cell actually undergoes a rewrite operation.

This scheme has a minimal number of transistors and it is easy to design. However, there are two disadvantages associated with this scheme. First, the full swing bitline discharge leads to a 70% increase in read operation power consumption [9]. Second, the sense nodes-bitlines direct coupling degrades the sense amplifier speed as the number of cells per segment increases. Additionally, the sense nodes-bitlines direct coupling causes both nodes Q and Qb to discharge momentarily upon the activation of the SAE signal and then resolve. Therefore, under low bitline voltage swing and in the presence of mismatch variations, the sense amplifier is likely to make a wrong decision.

3.2.2 Current-Mode Sense Amplifiers

The notion of maximizing memory capacity has resulted in ever diminishing bitline voltage swing over long, heavily-loaded bitlines. The main challenge to overcome in using conventional voltage sense amplifiers is the sense amplifier's inherent offset voltage. Traditionally, the sense amplifier's intrinsic offset voltage is reduced by sizing the sense amplifier driver transistors up to minimize V_{TH} deviations due to PVT variations. Although this seems a

feasible solution to some extent, the higher power consumption and area overhead associated with this approach has imposes restrictions on its application in nanometric CMOS technologies.

The fact that the cell generates a bitline differential current under any circumstance has motivated researchers to develop a current-mode sense amplifier capable of sensing very small bitline differential currents irrespective of voltage swing [22]. The most commonly used current-mode sense amplifier is shown in **Figure 3.1(b)** [22]. This scheme consists of four identical PMOS transistors (M1-M4). The precharge conditions of bitlines (Bl and Blb), which is V_{DD}, and datalines (DL and DLb) which is gnd, bias the four PMOSs into the saturation region. Upon the activation of the sense amplifier (Y_{Sel} goes low), the current passing though M1, M3 (or M2, M4) is the same since they are connected in series. This current depends on the transistor's V_{GS}, as a result the voltage level at both bitlines is set at V_{GS}+V_{GS} (V1+V2), *i.e.,* $\Delta V_{Bitline} \simeq 0$. The current conveyer, therefore, has the ability to convey the bitline's differential current to the datalines without the need to develop a differential bitline voltage. A second stage is, therefore, used to sense the developed dataline differential voltage [37].

Another current-mode sense amplifier, proposed in [25] and used in [38], is shown in **Figure 3.1(c)**. This scheme eliminates the need for two separate sensing stages and amplifies the bitline differential current via a crossed-coupled NMOS transistors similar to the voltage-latch scheme shown in **Figure 3.1(a)**. In this scheme, sensing nodes Q and Qb are initially precharged to V_{DD} through PMOS transistors M6 and M7. Upon the activation of the SAE signal, the ability of the two PMOS transistors M3 and M4 to hold the corresponding sensing nodes at V_{DD} is determined by the bitlines' differential voltage and current. So, if the cell is performing a read "0" operation, the Bl voltage and current become lower than that of Blb. Consequently, the current supplied to node Qb through

M3 is higher than the current supplied to node Q through M4.

As a result, a positive feedback action of the cross-coupled configuration takes place and the sensing nodes resolve. The advantages of this scheme are that it is a single stage and needs fewer transistors to realize, so this scheme can be used for local application (one sense amplifier per column). However, the sensing node precharge condition makes this scheme vulnerable to mismatch variations. At a low-level bitline voltage swing $\Delta V_{Bitline}$, the V_{TH} variation of PMOS transistors M3, M4 can result in differences in their drivability and consequently can lead to the sense amplifier making a decision.

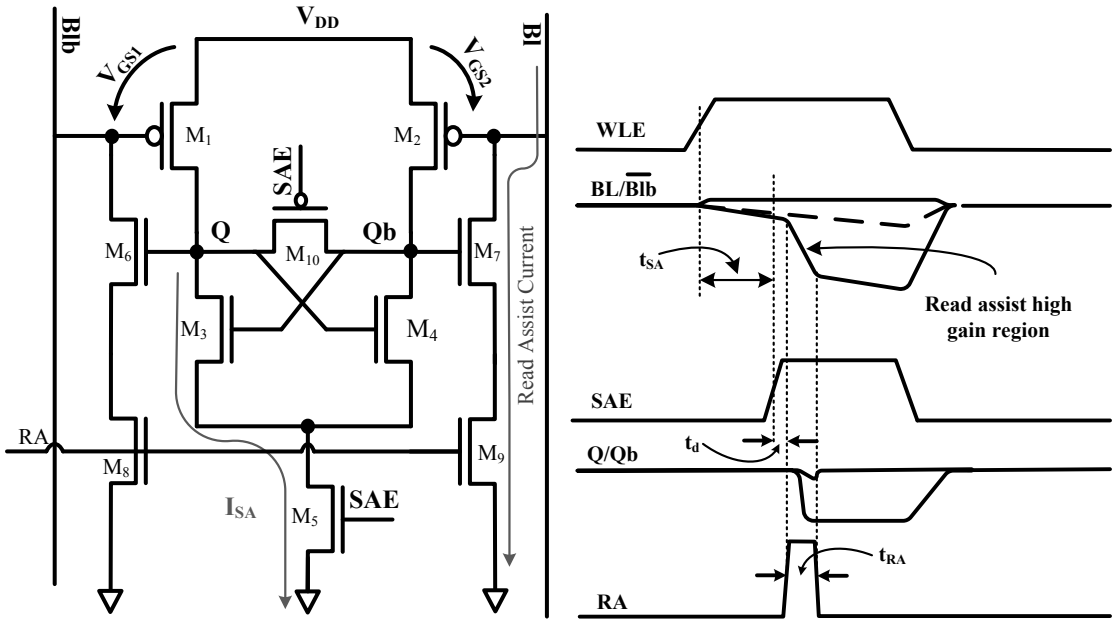
3.3 Proposed Sense Amplifier Schemes

Driven by the benefits of using read-assist and write-back mechanisms, we propose new sensing schemes to overcome some of the disadvantages of conventional sense amplifiers while improving system performance. The first two schemes are differential voltage sense amplifiers with read-assist and write-back features. The third scheme is a current-mode sense amplifier with a read-assist feature.

3.4 Read-Assist Voltage Sense Amplifier (RA-SA): Scheme I

3.4.1 Circuit Description

Figure 3.2(a) shows the proposed sense amplifier schematic diagram. The input differential pair PMOS transistors M1 and M2 along with column bitlines are employed to



(a) Schematic Diagram (b) Timing Scheme

Figure 3.2: Proposed Read-Assist Voltage Sense Amplifier.

precharge nodes Q and Qb to V_{DD} The precharge level of these nodes is equalized through transistor M10. In order to keep M1 and M2 "on", the reference precharge circuit described in **section 2.2.1**, **Figure 2.5** is utilized with high threshold voltage (HV_{THn}) NMOS transistors to precharge the bitlines Bl and Blb to $V_{DD} - HV_{THn}$.

Consequently, M1 and M2 are biased at the edge of the conduction region with source-gate voltage of $V_{SG1} = V_{SG2} = HV_{THn}$. The high voltage level (near V_{DD}) at Q and Qb keeps the cross-coupled NMOS transistors M3 and M4 and read-assist transistors M6 and M7 "on". The sense amplifier is activated by an active high SAE signal applied to NMOS transistor M5. The read-assist mechanism is invoked by enabling a read-assist signal (RA) applied to the gate of transistors M8 and M9.

The sensing operation is performed in two phases. In the first phase the sense amplifier is activated by enabling a SAE signal. In the second phase, a read-assist action is invoked by enabling a read-assist pulse (RA). Since the early activation of the RA signal can lead to instantaneous discharge for both bitlines, which can deteriorate the developed bitline differential voltage, the second phase has to be made to lag behind the first phase by a time delay t_d. **Figure 3.2(b)** illustrates the timing scheme used in the proposed sense amplifier and the anticipated bitline and sense nodes response to a read operation.

3.4.2 Circuit Operation

Upon activation of a WLE signal, the voltage level difference between the cell's storage nodes and the bitlines makes cell develops a differential bitline voltage across the bitlines. The high level node (V_{DD}) charges the Blb up above $V_{DD}-HV_{THn}$, while the low level node (" 0") discharges the Bl below $V_{DD}-HV_{THn}$. This imbalanced distribution in bitline voltage shifts the operating points of transistors M1 and M2 toward the cut-off and saturation regions, respectively. Consequently, M2's drivability becomes higher than that of M1 due to the difference in their V_{GS} voltage. Once the SAE signal is asserted, both sensing nodes Q and Qb tend to drop down; however, the current difference in the differential pair helps M2 to hold Q at a high voltage level, whereas Qb continues to drop to ground. The positive feedback created by the M3-M4 cross-coupled configuration increases the loop gain until Q and Qb resolve.

The read-assist signal is turned "on" just after enabling the SAE signal to activate the two read-assist transistors M6 and M7. Consequently, additional positive feedback is created between the sensing node Q (Qb) and the Bl (Blb) to speed up the Bl discharging process. In order to reduce read operation power consumption, the read-assist action can

be deactivated by turning M8 and M9 off (RA signal goes low). As the read operation is accomplished, the SAE goes high and the sense amplifier is precharged again for the next read operation.

The use of the proposed scheme exhibits the following advantages:

- The sense amplifier does not need precharge transistors.
- Owing to its precharge scheme, the accessed memory cell creates differential bitline voltage in opposite directions, *i.e.,* one bitline charges up as the other charges down. Even though the bitline charge-up process might not be noticeable as an increase in the bitline voltage level, this can significantly contribute to column leakage current compensation.
- The opposite change in the bitline voltage minimizes the V_{THp} difference between the sense amplifier input differential pair. Therefore, due a process mismatch, if V_{THp1} is lower than V_{THp2}, the simultaneous decrease in V_{THp1} and increase in V_{THp2} counterbalance the mismatch in V_{THp}.

3.4.3 Circuit Implementation and Simulation Results

The proposed circuit was designed and implemented in ST 90-nm standard CMOS technology and simulated on a 256-cell 6T SRAM column. One cell out of the 256 is accessed by activating its WLE signal, whereas the rest of the cells are made non-selected by tying their WLE signals to ground ("0"). To verify the proposed scheme's functionality in a realistic environment, SPICE transient simulations were carried out using post layout extracted instances for the sense amplifier and the column. Timing and control signals were generated from a control unit designed for that purpose. The SAE signal is activated

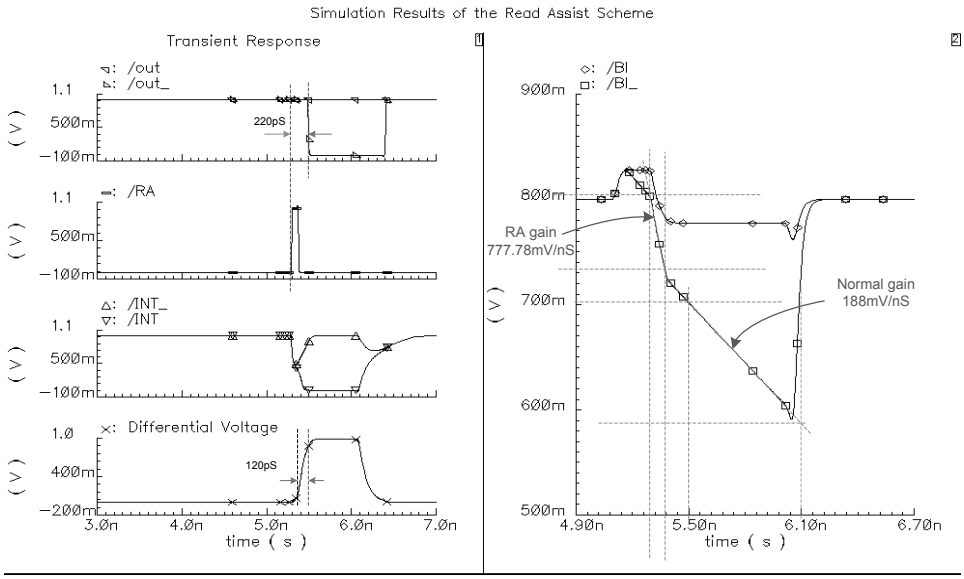


Figure 3.3: Proposed Read-Assist Post Layout Simulation Results.

at a 100 mV bitline differential. **Figure 3.3** shows the obtained post layout transient simulation results.

Table 3.1: Proposed RASA Schematic and Post Layout Simulation Results Comparison

	RA pulse width (ps)	SA Diff. Voltage rise time (ps)	RA gain (V/ns)	Normal gain (mV/ns)	RA delay (ps)
Schematic	118	105	0.82	166.66	173
Post layout	125	120	0.778	188.0	220

As can be seen in the figure, the Bl discharge process has three distinct regions. Region I is the region where the bitline discharges normally through the selected memory cell driver and access transistors. Region II exhibits the bitline discharge acceleration (high gain) due

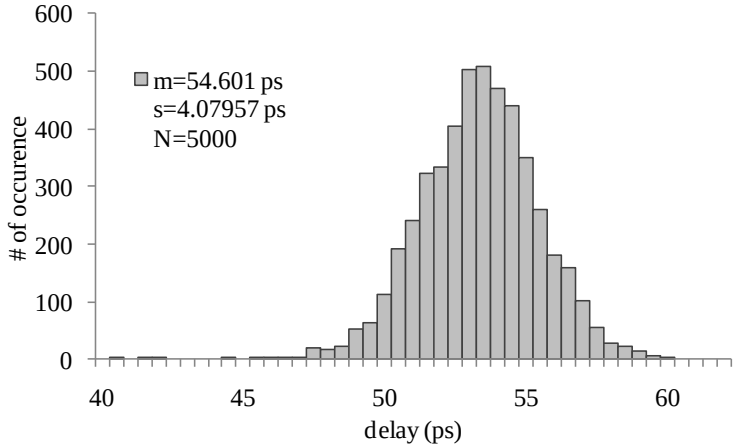


Figure 3.4: Proposed Read-Assist Scheme Monte Carlo Simulation Results.

to the activation of the sense amplifier and the read-assist action. In this region, the bitline gain ($\Delta V_{Bitline}/\Delta t$) is increased as a result of positive feedback created by the read-assist transistors M7 and M9. In Region III, the RA is disabled and the bitline returns to its normal discharging gain. The sense amplifier speed (delay) is measured between the SAE signal's rising edge and the 50% point of the sense amplifier's differential output voltage. **Table 3.1** provides a comparison between schematic and post layout simulation results.

The proposed scheme's robustness against mismatch variations was verified by Monte Carlo simulations. In order to reduce the computation time, the post layout extract view of the proposed scheme was used in the simulation test bench along with a dummy capacitance of 250 fF that mimics a 256-cell bitline extracted loading capacitance. **Figure 3.4** indicates that the proposed scheme is functional in the presence of mismatch variations with less than 10% deviation in its nominal delay value.

3.5 Read-Assist Write-Back Sense Amplifier (RA-WRBK-SA): Scheme II

Proposed read-assist write-back sense amplifier RA-WRBK-SA (Scheme II) operates in a manner complementary to RA-SA (Scheme I) presented in the previous section. This second scheme features a read-assist mechanism and it is designed to perform a write-back operation when needed. The write-back property of this scheme is an exaggerated read-assist operation. In other words, the read-assist feature of this scheme can be maximized to fully discharge the bitline and thereby it performs a write-back operation.

3.5.1 Circuit Description

Figure 3.5 shows the proposed scheme's circuit diagram. The sensing nodes Q and Qb are precharged to "0" through NMOS transistors input differential pair, M1 and M2, and the column's bitline pair Bl and Blb. transistor M8 is used to equalize Q and Qb. This precharge scheme eliminates the need for a sense amplifier precharge transistors. The pre-discharge nodes keep the two cross-coupled PMOS transistors M3, M4 "on" and the two read-assist NMOS transistors M6, M7 "off". The NMOS/PMOS combination on each side of the amplifier (M1-M3 and M2-M4) is skewed toward the PMOS transistor. **Table 3.2** provides the transistor sizing used in this circuit.

3.5.2 Circuit Operation

This circuit is designed to be activated independent of the WLE signal, *i.e.*, the sense amplifier can be activated early even if the cell has not yet developed the target bitline

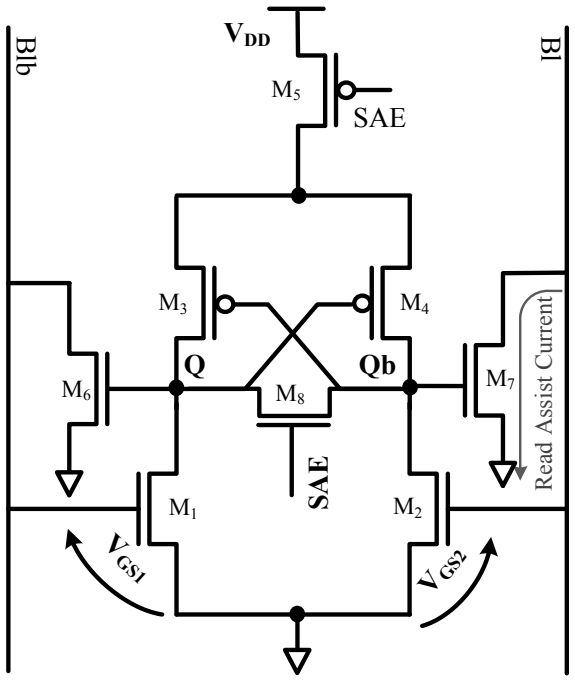


Figure 3.5: Proposed RA-WRBK Sense Amplifier Schematic Diagram.

differential voltage. Upon activation of the SAE signal, both nodes (Q and Qb) charge up at the same pace, but the developed voltage drop at Bl creates a gate-source (V_{GS}) difference between M2 and M1. The low voltage level at BL makes transistor M2 weaker than M1. Consequently, node Qb charges up to V_{DD} faster than node Q. While the cross-coupled positive feedback mechanism accelerates the voltage level degradation at the sensing nodes, the positive feedback action through M7 speeds up bitline discharge until the sensing nodes resolve.

The read-assist transistors, M6 and M7, can be sized according to the desired sense amplifier operation. Wider read-assist transistors strengthen the write-back operation; otherwise, these two transistors are only used for read-assist by providing additional positive feedback path to discharge the bitlines.

Table 3.2: Proposed RA-WRBK Sense Amplifier Transistor (W/L) in μm.

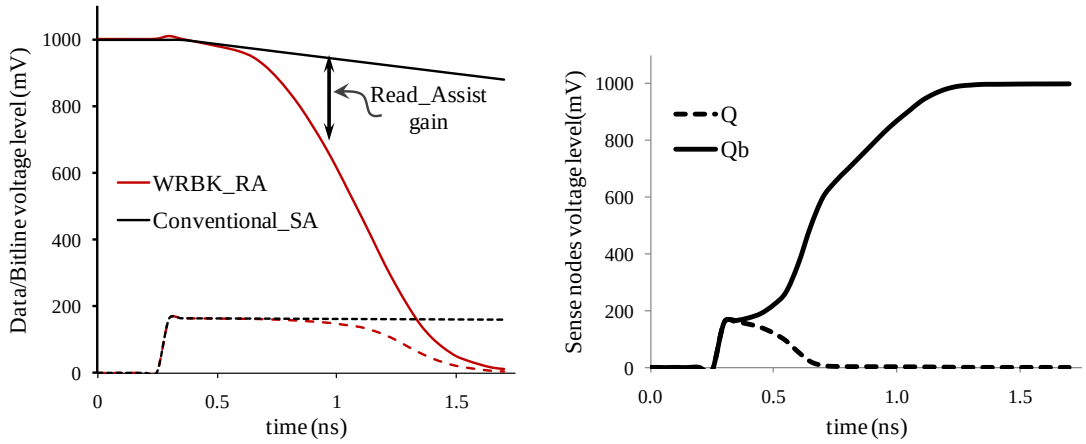
Drivers	PMOS Loads	Read-assist	Equalizer
M1, M2	M3, M4, M5	M6, M7	M8
0.4/0.1	1.6/0.1	1.0/0.1	0.15/0.1

Even though the proposed scheme shows timing independent behavior when proper transistor sizing is used, delaying the SAE signal to allow the memory cell to develop bitline differential voltage is recommended in order to overcome any V_{TH} mismatch in the NMOS input differential pair M1 and M2.

3.5.3 Circuit Implementation and Simulation Results

Figure 3.6 depicts SPICE transient simulation results obtained with and without the presence of read-assist. **Figure 3.6(a)** shows Bl and the cell's stored data behavior during a read operation. The solid curves signify the response with read-assist and the dashed line without read-assist. As seen in the figure, the Bl discharge rate is accelerated until it is fully discharged and the data is rewritten into the cell when a read-assist is used, as opposed to a more gradual bitline discharge and persistent data level degradation (the zero level data remains above its nominal value as long as the WLE is active) when read-assist is not used. **Figure 3.6(b)** shows the sense amplifier sensing node's transient response. As can be seen, both nodes initially attempt to charge up, but they ultimately resolve as M2 drivability decreases due to the bitline Bl voltage drop. The write-back property is extremely important to maintain data stability under low-voltage operating conditions, as we will present later in Chapter Four.

In order to study the impact of transistor mismatch on the proposed scheme's per-



(a) Bitline and Data Response W and w/o Read-Assist

(b) SA Differential Output

Figure 3.6: Proposed RA-WRBK Sense Amplifier Transient Simulation Results.

formance, Monte Carlo simulations are performed. **Figure 3.7** indicates that the sense amplifier works properly in the presence of mismatch variations with less than 10% standard deviation.

3.5.4 Performance Comparison

The proposed schemes' performance is compared to a reference local sense amplifier proposed in [9]. The comparison is based on sense amplifier speed and power delay product (PDP). The three schemes were simulated with a typical 256-cell SRAM column and triggered at a 100 mV differential bitline voltage. A column post layout extracted view is used to establish a realistic bitline loading effect. In order to ensure a fair comparison, the three schemes are designed to occupy relatively the same layout area.

Figure 3.8 shows the transient simulation results obtained under these conditions for

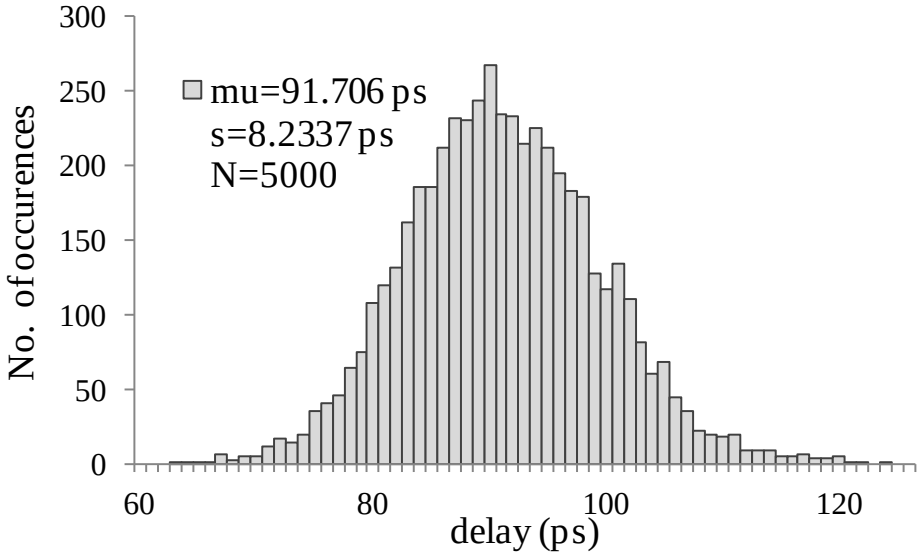


Figure 3.7: RA-WRBK-SA (Scheme II) Monte Carlo Simulation Results.

the three schemes' differential output voltage. As seen in the figure, the conventional sense amplifier speed is lower than either of the proposed schemes.

Power consumption in an SRAM read operation is attributed to sense amplifier activity and the need to restore the bitlines to their precharge levels after each read operation. Because of the limited bitline voltage swing, read operation power consumption is relatively low compared to write operation. Power consumption in sense amplifier circuits is measured in terms of PDP. Therefore, the sense amplifier power consumption can be managed by increasing the sensing speed. Bitline recovery power consumption is also manageable because of the limited bitline voltage swing. However, the use of a write-back sense amplifier leads to high read power consumption due to the fact that the write operation is a by-product of a successful read operation.

The write-back feature is an exaggerated way of assisting the memory cell in performing

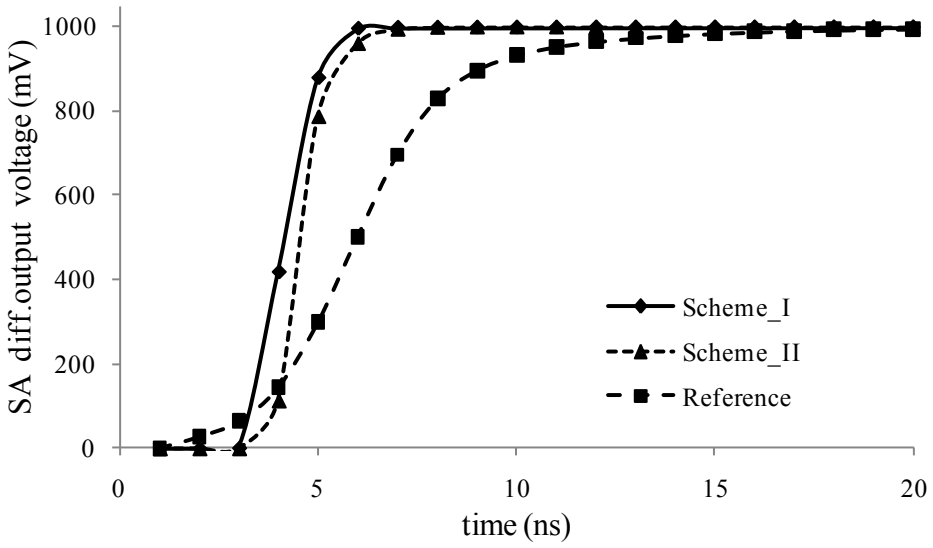


Figure 3.8: Proposed Schemes Performance Compared to Voltage-Latch SA [9].

a successful read operation. Therefore, if the read-assist action is limited to a specific time window, the cell can perform successful read operations without needing to fully discharge the bitline. This can manage the high power consumption associated with the bitline recovery process.

As such, a read-assist window control feature was added to Scheme I to reduce column power consumption. Whereas full bitline discharge is inevitable in [9], Scheme II can gradually discharge the bitline, thereby providing the cell with the required read-assist without needing to fully discharge the bitline. As such, column power consumption can be reduced by 50% or more depending on the bitline discharge level.

Table 3.3 provides post layout simulation results of the proposed schemes compared to the scheme used in [9]. This comparison is based on time delay and PDP at the cell and column level. Two time delays components are calculated, t_{d1} and t_{d2} are calculated when

Table 3.3: Post Layout Simulation Comparative Results.

Scheme	Delay (ps)		SA(μW)	Col(mW)	Col PDP(pJ)	SA PDP(fJ)	No.of Tran.
	t_{d1}	t_{d2}					
Scheme I	83.4	115.7	1.92	21	2.6	0.16	10
Scheme II	112.2	145.3	1.96	42.5	4.76	0.219	7
Reference	215.0	367.5	2.01	85.2	18.3	0.425	5

the SA differential output SA_{diff} reaches 200 mV and 500 mV, respectively. Column and sense PDP is calculated based on the time required to develop 200 mV bitline differential voltage because the reference sensing nodes and the bitlines are directly coupled. Otherwise, power consumption in the column that utilizes the voltage latch scheme used in [9] would be very high due to the fully discharged bitline. As **Table 3.3** indicates, Scheme I and Scheme II are 2.5 times (2.5X) and 2.0 times (2.0X), respectively, faster then the reference. This reflects as column PDP saving in Scheme I and Scheme II of 7X and 3.8X, respectively.

Table 3.3 depicts the column power consumption during a successful read operation. The power consumption is calculated based on the bitline voltage swing at the end of a read operation. Scheme II is designed to discharge the bitline to 50% of the supply voltage V_{DD}. As can be seen, the proposed Schemes I and II provide up to 75% and 50% column read power savings, respectively, compared to the reference.

The dependence of sense amplifier operation on bitline loading is a key factor in sense amplifier performance. As such, we investigated the impact of bitline loading ($C_{Bitline}$) on the delay of the proposed sense amplifiers compared to the reference. **Figure 3.9** shows the delay of the sense amplifier output SA_{diff} as a function $C_{Bitline}$ for the proposed schemes compared to the reference. As expected, the sensing node/bitline direct coupling in the

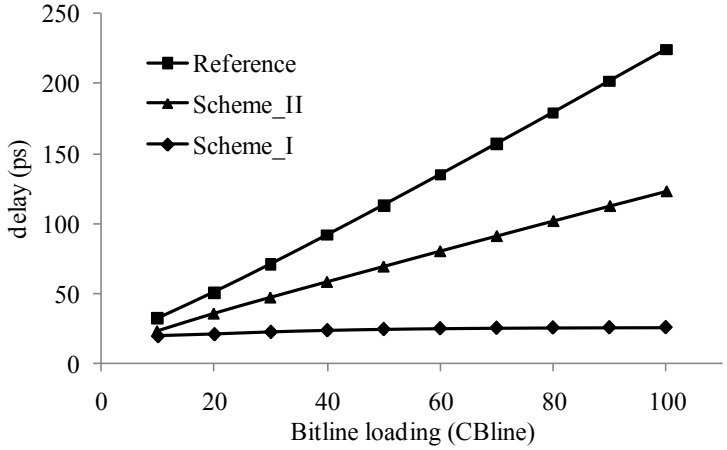


Figure 3.9: Sense Amplifier Delay as a Function of Bitline Loading ($C_{Bitline}$).

reference scheme makes the delay directly proportional to $C_{Bitline}$. However, the decoupled sensing node/bitline configuration adopted in the two proposed schemes makes them less dependent on $C_{Bitline}$. Furthermore, owing to the precharge configuration employed in Scheme I, the bitline loading impact on the sense amplifier speed is marginal.

3.6 Test Chip Design

In order to verify the obtained post layout simulation results with silicon measurements, a test chip was designed and fabricated in CMOS 90-nm technology in the March 2008 run. Due to limitations in area and number of pads available, only Scheme I was implemented. The designed test chip contains a 256-cell SRAM column with the proposed Scheme I and other required peripheral circuits, such as timing control unit, leakage and read/write control units, and a column write driver in addition to input/output buffers. **Figure 3.10** shows the designed test chip block diagram. **Table 3.4** shows the input/output and control

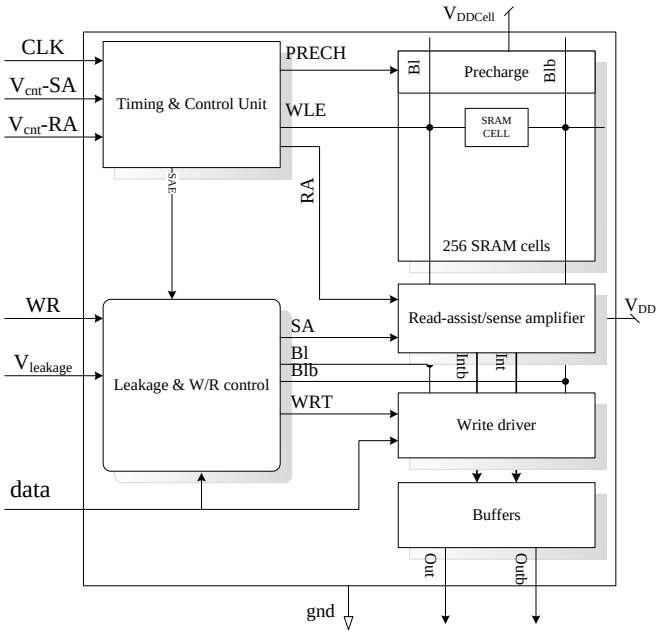


Figure 3.10: Test Chip Block Diagram.

input/output pins and the signal type used in this chip.

The sense amplifier and read-assist control signal, V_{cnt-SA} and V_{cnt-RA}, are used to control the SAE timing signal and the read-assist signal time window. This is accomplished by controlling the delay in the delay line used in the timing and control unit. In order to explore the proposed scheme's robustness against bitline leakage current increase, an NMOS transistor is attached to each bitline with a controlled gate voltage so that increasing the gate voltage can mimic a bitline leakage current increase. The leakage control voltage $V_{leakage}$ is used for that purpose. It is worth mentioning that the chip is designed so that the cell is capable of performing read and write operations of both "0" and "1". The leakage current control mechanism was therefore added to both bitlines. Even though post layout simulation results show a solid agreement between both the schematic and extracted

Table 3.4: Test Chip Control Signals.

Pin	Signal type
CLK	AC
Read/Write WR	DC
$Data_{In}$	DC
$Data_{Out}$	DC
V_{cnt-SA}	DC
V_{cnt-RA}	DC
$V_{leakage}$	DC

views' simulation results, the silicon measurement outcomes, unfortunately, did not reflect the anticipated results.

3.7 Proposed Body Bias-Based Current-Mode Sense Amplifier

Device miniaturization in high-density very-large scale integration (VLSI) systems has made transistor characteristics susceptible to temperature variations and process imperfections [12]. A 15% fluctuation in transistor V_{TH} has become normal in modern CMOS technologies. V_{TH} variation in SRAM arrays manifests as variations in cell drivability. Therefore, the cell's ability to generate adequate bitline voltage swing in a given access time cannot be guaranteed (see **Section 2.5.3**). Additionally, V_{TH} variations in conventional differential voltage sense amplifiers creates an offset voltage that compromises the cell-developed bitline differential voltage. Thus, V_{TH} variations can impact SRAM cell

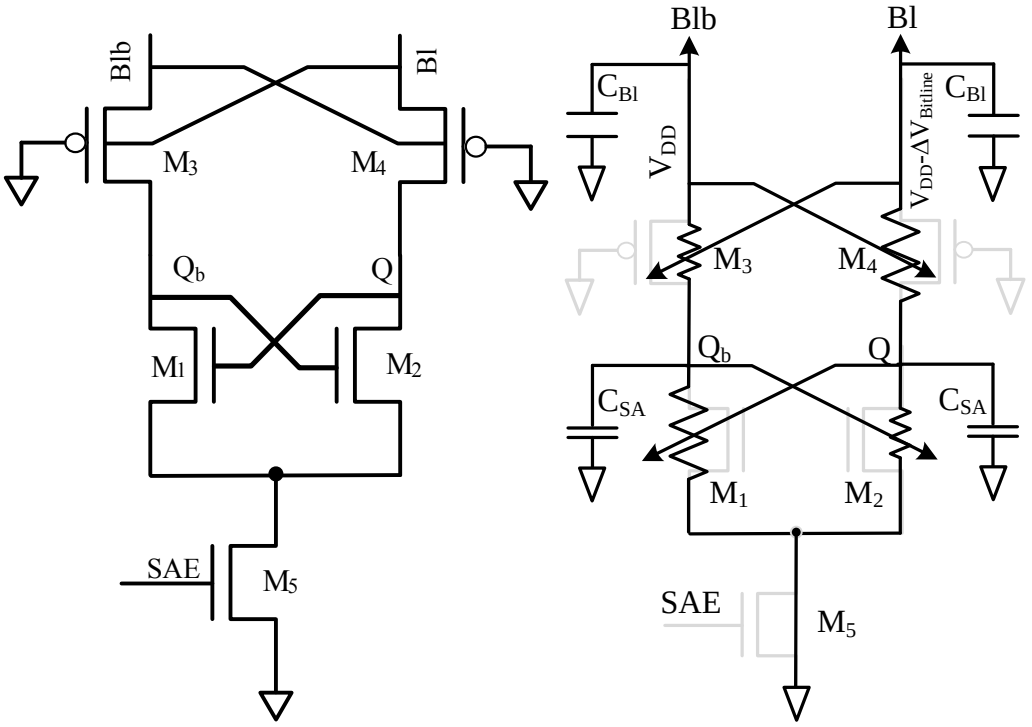
reliability in two opposite ways. On one hand, it degrades the cell's ability to generate the required differential input voltage on the bitlines so that the differential voltage sense amplifier of the column can make a right decision. On the other hand, it increases the sense amplifier offset voltage.

One way to overcome the limited bitline voltage swing is to employ an offset insensitive sensing scheme. The current-mode sense amplifier is an important solution in high density SRAM applications. In this context, we propose a new current-mode sense amplifier that exhibits competitive performance figures.

3.7.1 Circuit Description and Operation Principal

The proposed current-mode sense amplifier scheme is shown in **Figure 3.11(a)**. It consists of a five transistors, so it can be used as a dedicated local sense amplifier for each column in the memory array. The sources of two permanently "on" PMOSs, M3 and M4, are attached to the bitlines; therefore, the sensing nodes Q and Qb are precharged to V_{DD} through the bitlines' precharge circuitry. The sense amplifier input current is the bitline differential current created by the memory cell during a read operation. The body contact (substrate) of each PMOS is cross-coupled to the bitlines to control the transistor's body voltage. Another cross-coupled configuration is established using two NMOS transistors, M1 and M2. The sense amplifier operates at active high SAE signal that is applied to the gate of NMOS transistor M5.

During a read operation, if the stored data is "0", the SRAM cell discharges the bitline Bl below V_{DD} and creates a bitlines voltage swing $\Delta V_{Bitline}$. The two sensing nodes Q and Qb track the bitline voltage change and modulate the cross-coupled NMOS pair operating point, *i.e,* change $V_{GS1,2}$. The body-source voltage difference (V_{SB}) of transistor M3 makes



(a) Schematic Diagram

(b) Resistance Equivalent Schematic Diagram

Figure 3.11: Proposed Current-Mode Sense Amplifier.

this transistor forward-body biased (FBB), with V_{SB}= $\Delta V_{Bitline}$. Similarly, the V_{SB} of transistor M4 makes it reverse-body biased (RBB), with source-body voltage difference V_{SB}= -$\Delta V_{Bitline}$. The bitline-generated body bias voltage modulates the PMOS transistors' drivability [39]. According to [1], **Equation 3.1** indicates that a FBB decreases the PMOS transistor's V_{TH}, whereas a RBB increases V_{TH}.

The V_{TH} variation due to the body bias can be modeled as a variation in the transistor's on resistance. such that transistor M3 (FBB) can be represented as a small resistor and transistor M4 (RBB) can be presented as a relatively high resistor. Similarly, the two

cross-coupled NMOS transistors can be represented as a resistor that depends on their overdrive voltage ($V_{GS} - V_{TH}$). As such, M1 can be thought of as high resistor and M2 as small resistor. **Figure 3.11(b)** gives the amplifier's resistance equivalent circuit during a read operation.

$$V_{TH} = V_{THo} + \gamma \left(\sqrt{|-2\Phi_F + V_{SB}|} - \sqrt{|-2\Phi_F|} \right) \tag{3.1}$$

Upon the activation of the SAE signal, the resistances of the potential divider created by the PMOS-NMOS combination on the bitline Bl (M2-M4) drops the voltage level at node Q lower than that at node Qb due to the difference in the resistances of the potential divider of the PMOS-NMOS combination on the bitline bar Blb (M1-M3). The positive feedback of the two cross-coupled configurations (M1-M2 and M3-M4) accelerates the convergence of the two sensing nodes Q and Qb to "0" and "1", respectively. In other words, the weak transistor M1 (low V_{GS}) and the strong transistor M3 (low V_{TH}) hold Qb at high voltage level whereas the strong transistor M2 (high V_{GS}) and the weak transistor M4 (high V_{TH}) allow Q to discharge to "0". When node Q goes low, the Bl continues to discharge through the "on" PMOS transistor M4 and assists the cell to perform a successful read operation. The bitline discharge level is determined by the path resistance of the series combination of transistors M2 and M4. The difference in $C_{Bitline}$ and the sense amplifier parasitic capacitance (C_{SA}) determine the discharge RC time constant. The small parasitic capacitance at node Q (C_{SA}) discharges to zero via the small resistance of M2, whereas the large bitline capacitance $C_{Bitline}$ stays at a relatively high voltage level because of the relatively high resistance of the M4, M2 series combination.

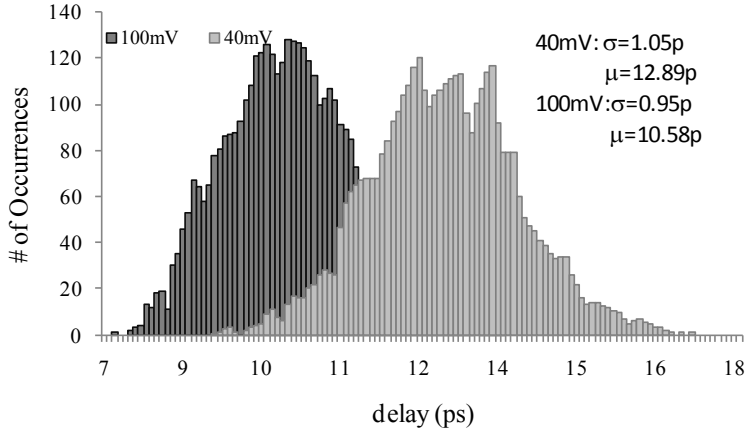
3.8 Simulation Results

The proposed SA scheme was implemented in a ST 65-nm CMOS design kit and simulated with a 256-cell 6T SRAM column. Monte Carlo simulations were performed to verify the proposed scheme's reliability under conditions of low bitline voltage swing and high operating temperature in the presence of process and mismatch variations. **Figure 3.12** confirms the proposed scheme's functionality when the bitline voltage swing is reduced from 100 mV to 40 mV with a marginal shift in the output delay. Additionally, the proposed scheme's functionality under typical and high operating temperature (27 and 100 C^o) is also verified.

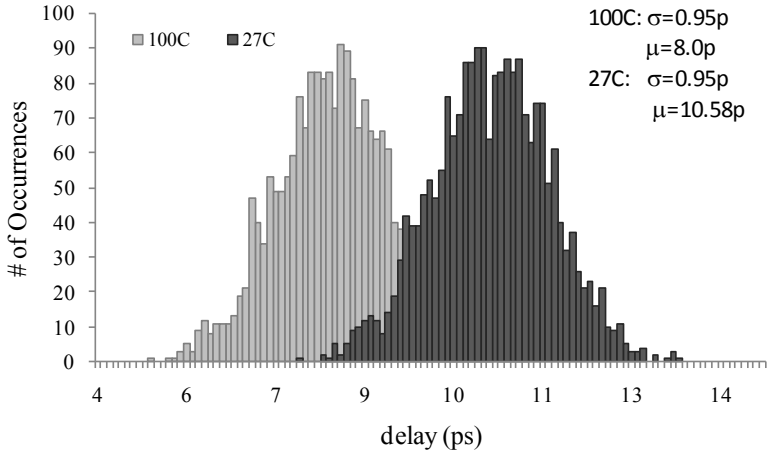
3.9 Performance Comparison

In order to verify the proposed scheme's performance advantages, two conventional schemes were also implemented [9][25] for comparison purposes. In the following discussion we will refer to these references as S1 and S2, respectively. Even though the SA scheme S1 is not a pure current-mode sense amplifier, it has been used here as a reference to compare the read-assist feature with the proposed scheme. Excluding the two PMOS transistors used to precharge the sense amplifier S2, the three schemes have the same number of transistors. However, the actual area required for optimized performance is different. Each scheme has been optimized for the most optimal performance and area.

Each scheme is employed in a 256-cell 6T SRAM column operating under read operation conditions. **Figure 3.13** depicts the sense nodes of each sense amplifier scheme and the corresponding bitlines' responses to a read operation, bearing in mind that the sense nodes and the bitlines in scheme S1 are the same due to the direct couple configuration. As



(a) Output Delay at Different Bitline Swings



(b) Output Delay at Different Operating Temperatures

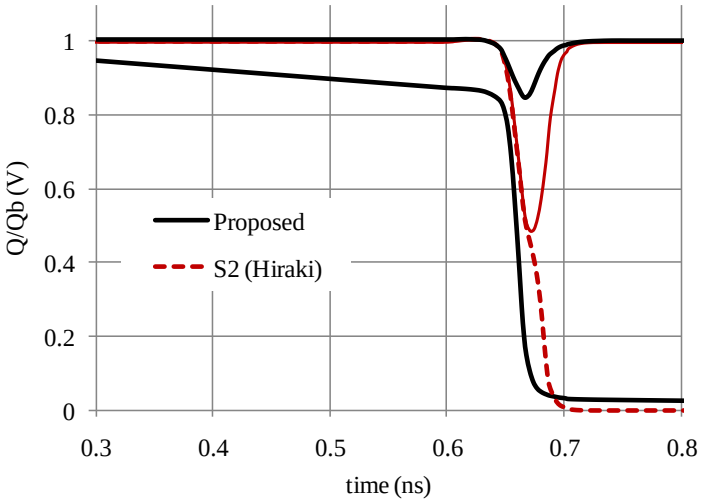
Figure 3.12: Proposed Current-Mode Sense Amplifier Monte Carlo Simulation Results.

can be seen in **Figure 3.13(a)**, the sense nodes in the proposed scheme track the bitline differential voltage. In contrast, the sense nodes in S2 are kept high (precharge level) until SAE is activated. The moment SAE is activated, the sense nodes in the proposed scheme resolve smoothly, whereas in the conventional schemes S1 and S2 they track each other and resolve after some time delay. In the presence of process and mismatch variations, this behavior can cause incorrect sensing decision.

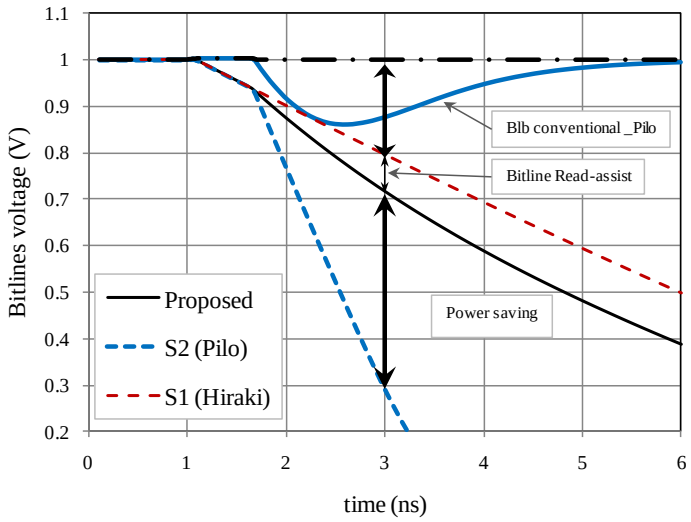
The bitline response when the sense amplifier is activated is shown in **Figure 3.13(b)**. In scheme S1, the sense nodes (Q/Qb) and the bitlines are attached; therefore, the sense amplifier output and the bitline response are the same. In order to meet the speed requirement, the sense amplifier's pull-down path is made strong. This causes both bitlines to discharge upon SAE activation and adds more delay to the sense amplifier's differential output. Even though fast bitline response provides the memory cell with a read-assist mechanism, the power consumed during this process is significant. Unlike S1, in both the proposed scheme and S2, the sense amplifier output does not follow Bl due to the sense node/bitline isolation. Since the sense amplifier speed is measured at the sense amplifier output, the bitline response is not necessarily fast.

Owing to its permanently "on" PMOS transistor, the proposed scheme provides a reasonable read-assist mechanism without high power consumption, as shown in **Figure 3.13(b)**. This behavior helps to conduct a successful read operation while avoiding the excessive energy consumption associated with unnecessary full swing bitlines, as in S1. Moreover, limiting the bitline voltage drop is necessary to avoid the latch up in the opposite PMOS due to the high body voltage level.

Another performance comparison made here is the sense amplifier delay and probability of correct decision making as a function of the bitlines' differential voltage swing. **Figure 3.14(a)** depicts a comparison of the three schemes' respective delays as a function of bitline



(a) SA Sense Nodes Q/Qb Response Compared to [25]



(b) Read Operation Bitline Response Compared to [25][9]

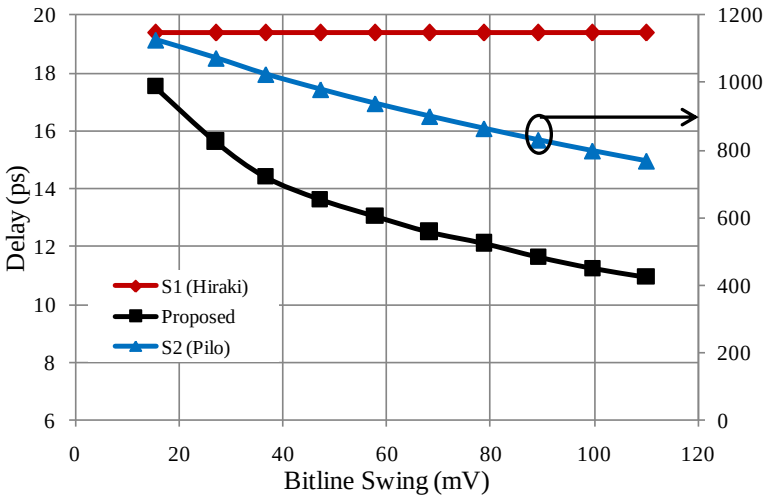
Figure 3.13: Proposed Sense Amplifier Performance Comparison.

voltage swing. As seen in the figure, the output delay of the proposed scheme and S1 can be improved by increasing the bitline voltage swing, whereas S2 has a relatively constant delay. This is due to the fact that the bitlines and sense nodes in the proposed scheme and S1 are directly coupled, whereas they are isolated (precharged) in scheme S2. Scheme S1 has a relatively large delay time compared to the proposed scheme and S2 due to the fact that the sense nodes in scheme S1 are directly coupled to the bitlines, therefore bitline loading has a direct impact on the sense amplifier response. Moreover, considering a 50-mV bitline swing, the proposed scheme shows a 31% speed improvement compared to S2, as can be seen in **Figure 3.14(a)**.

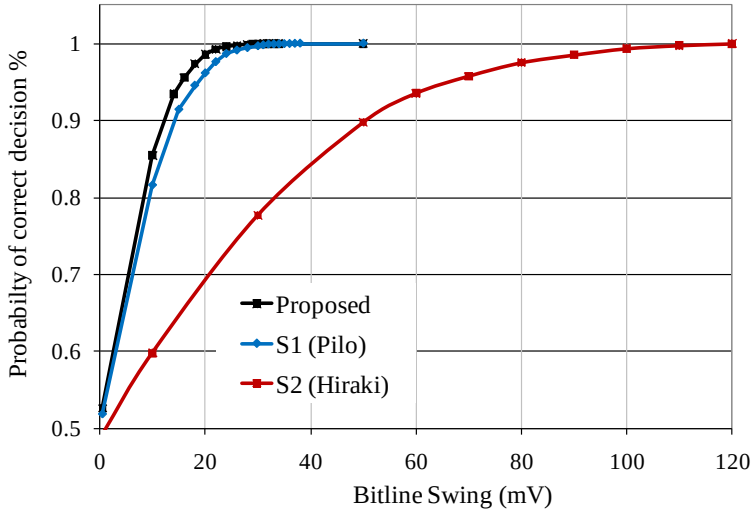
The probability of read failures is also examined as a function of bitline voltage swing. Monte Carlo simulations were conducted at different levels of bitline voltage swing and a 100% pass rate was targeted as an indication of no read failures. Read failures were predicted based on the SA's inability to make a right decision for a given bitline's swing. **Figure 3.14(b)** indicates that the proposed scheme yields zero failures for bitline voltage swings of up to 25 mV compared to 120 mV and 40 mV voltage swing required for S2 and S1, respectively. That is 37% reduction in bitline voltage with respect to [9]. Read failure reduction achieved at 25 mV is 3.3% and 28.4% compared to [9] and [25], respectively.

Operating supply voltage V_{DD} lowering and its impact on the required bitline swing for reliable SA operation is depicted in **Figure 3.15(a)**. The proposed scheme shows the ability to properly operate (with zero read failures) at 600 mV supply voltage with a 45 mV bitline differential. Compared to the 115 mV required by S2, this represents 2.5X less bitline swing requirement at a low operating voltage.

Finally, the advantage of the body bias use is verified by calculating the probability of read failure with and without body bias. Simulation results shown in **Figure 3.15(b)** indicate that a 7.5% improvement in read failure reduction can be achieved when operating

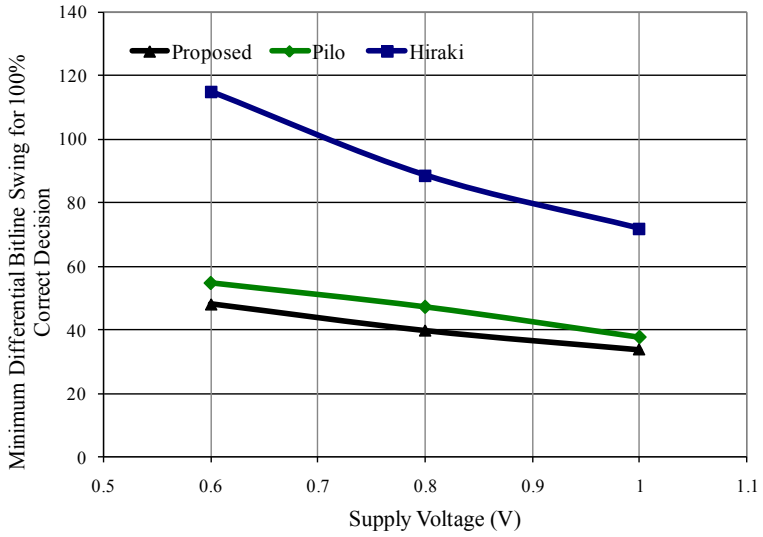


(a) SA Output Delay Comparison

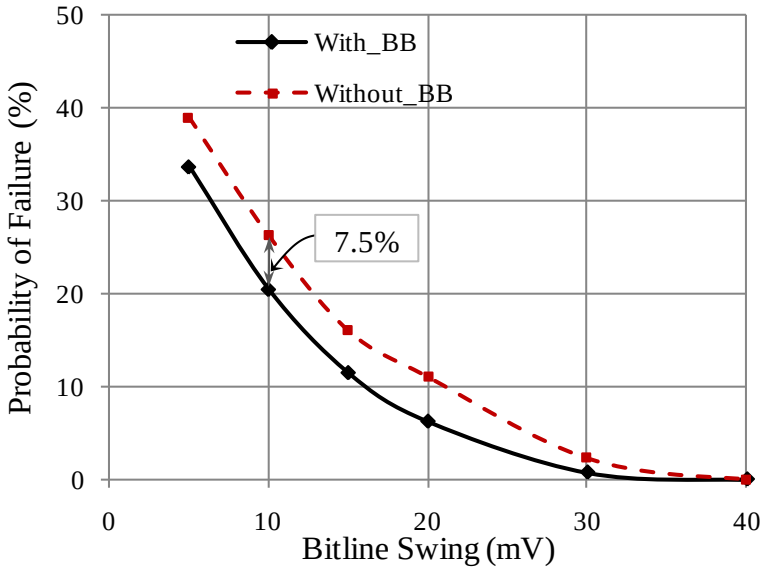


(b) Probability of Correct Decision Making

Figure 3.14: Sense Amplifier Performance as a Function of Bitline Swing ($\Delta V_{Bitline}$).



(a) Yield at Reduced V_{DD}



(b) Yield Improvement Associated With the Body Bias

Figure 3.15: Sense Amplifier Performance as a Function of Supply Voltage (V_{DD}) and the Impact of Body Bias.

at a 10-mV bitline swing.

3.10 Summary

In this chapter, we presented three new sense amplifier schemes. The first two schemes involved a differential voltage sense amplifier working with a controllable read-assist mechanism. In addition to their superior operational speed, these schemes provide significant read power savings. The third scheme involved a body-bias-based current-mode sense amplifier that can sense with minimal bitline voltage swing compared to conventional schemes. The body bias effect is employed to enhance the sense amplifier operation and reliability. Furthermore, this scheme provides a read-assist mechanism as well. Monte Carlo simulations were used to verify the proposed schemes' functionality and reliability in the presence of process and mismatch variations. With the current-mode sense amplifier scheme, read failure reductions of 3.3% and 28.4% were achieved for a 25-mV bitline swing. At the bitline swing voltage a speed improvement of 21% was realized compared to conventional schemes. Simulation results were used to show the advantage of using the transistor body bias. At very low bitline voltage swing a 7.5% reduction in the probability of read failures was achieved.

Chapter 4

Programmable Wordline Boost Driver for Low-Voltage Operated SRAM Cell Reliability Enhancement

4.1 Introduction

SRAM arrays dominate the majority of the area and account for the most of the transistors in the modern System-on-Chip (SoC). For such SoCs, the chip yield is determined by the SRAM array reliability. The ever-increasing demand in high-speed battery-operated devices requires the use of low-voltage, high-density SRAMs. While advances in CMOS technologies in the nanometric regime allow the use of minimum-size transistors to realize SRAM cells, reduction in the cell's supply voltage has faced the barrier of a well known industry term V_{DDmin}: the lowest voltage at which cell ability to meet design requirements deteriorates.

The SRAM's cell drivability is the key element in SRAM array yield. The cell's drivability, defined in **Section 2.5.1**, depends on the cell's transistors' current (see **Equation 2.6**). In the low-supply voltage regime (near V_{TH}), the transistor current changes exponentially with the supply voltage. Therefore, transistor V_{TH} variation is becoming a detrimental element in low-voltage operated devices.

Key SRAM functional parameters, such as deviations in I_{Cell}, speed (Δt), and SA_{offset}, are all directly affected by V_{TH} variation. To guarantee the functionality of millions of SRAM cells in an embedded memory instance, a reliable SRAM bitcell design has to cover a span of more than six standard deviations ($Z \geq 6$, see **Equation 2.8**) for given parameter variations. SRAM yield degradation in state-of-the-art CMOS technologies is increasingly dominated by soft failures (mentioned in **Section 1.4**) which are mainly caused by V_{TH} variation. The impact of V_{TH} variations on SRAM yield is even worse when operating at a low-supply voltage.

According to **Equation 2.6**, which is restated below, for a given operating speed (Δt), there are three factors that can degrade cell drivability and thereby reduce cell voltage margins: 1) weak cell driver and access transistors (low I_{Cell}), 2) heavily-loaded bitlines (high $C_{Bitline}$), and 3) high-leakage current due to a large number of cells per bitline (N). Because $I_{leakage}$ reduces with the lowered supply voltage, and $C_{Bitline}$ reduction is not a design option in this case, the cell current I_{Cell} is the only key player in cell drivability degradation due to reduced operating voltage.

$$I_{Cell} \geq C_{Bitline} \times \frac{\Delta V_{Bitline}}{\Delta t} + N \times I_{leakage} \tag{4.1}$$

One straightforward way to surmount the problem of poor cell margins is to increase cell area and pursue appropriate design constraints (cell ratios α and β). **Figure 4.1**

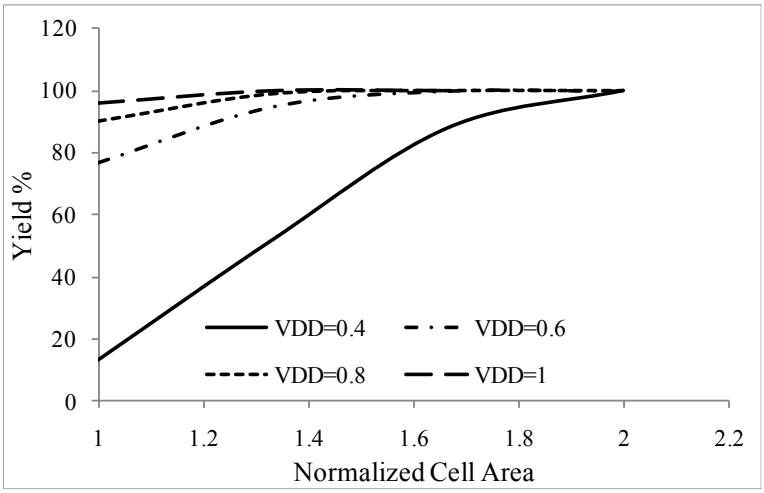


Figure 4.1: Conventional 6T SRAM Yield as a Function of Supply V_{DD}.

shows yield Monte Carlo simulation results conducted on a conventional 6T SRAM cell to investigate the required cell area increase to achieve a 100% yield at different operating supply voltages.

As can seen in **Figure 4.1**, at V_{DD}=0.4 V, at least two times the cell area is required to satisfy a 100% yield, knowing that these simulations are conducted under light bitline loading. Considering a realistic bitline loading ($C_{Bitline}$) of, about, 150 fF, which is an extracted post layout bitline loading capacitance of a typical column of 256 SRAM cells in a 65-nm process, more than 10 times an increase in cell area would still not produce a 100% yield (The same conclusion can be drawn from **Figure 1.8**, where a cell area of 22 times bigger than the nominal area is needed to meet design specifications at 0.4 V.) Thus, the direct cell area increase is not a practical choice in high-density SRAM applications.

4.2 Low-Voltage Operated SRAM Circuits

Despite the area and power overhead, increasing the transistor's channel width is the primary way to widen the 6T cell margins (RDM, WRM, and SNM) and to reduce the impact of V_{TH} variations [40]. For small capacity SRAM applications, enlargement of the cell's transistors allows lower V_{DD} operation; however, in high-density SRAM applications requiring the use of a miniaturized cell area, the search for alternative techniques to enhance the cell's drivability is inevitable.

Even though both the driver and access transistors in a conventional 6T SRAM cell play a key role in cell functionality, the cell's reliability is highly governed by the access transistor (see **Figure 1.5**). The access transistor is the main source of noise in a 6T cell because it represents the communication link with the outside world. Managing the access transistor's operating voltages (V_{DS}, V_{GS}), and thereby transistor drivability, is generally the means used in state-of-the-art embedded SRAM applications to overcome the growing variation in device V_{TH} [41].

Technically, the operation of the 6T SRAM cell is based on three power supply voltages: 1) the cell supply voltage V_{DD}, 2) the bitline precharge voltage, and 3) the WL voltage. Managing the supply voltage of these sources is an effective way to tackle the driver and access transistors' V_{TH} variations. One way to control the access transistor's drivability is to use a low LV_{TH} transistor and a negative WL voltage [42]. Whereas a negative WL voltage during retention mode reduces the cell's leakage current, a LV_{TH} access transistor and active WL signal increase the cell's drivability (I_{Cell}) and enhance cell RDM.

In addition to its high drivability, a LV_{TH} transistor is less vulnerable to V_{TH} variations [41]. Thus, this approach addresses all three of the performance enhancement factors; namely, high cell current, low leakage current, and less V_{TH} variation. Alternately, a high

V_{TH} access transistor with boosted WL is used in [17]. In this approach, cell RDM is enhanced via a WL boost, while leakage current reduction is achieved by the use of a HV_{TH} access transistor.

Non-DC WL boost approaches are reported as possibilities for eliminating the need for a dual V_{TH} process. Step-down and two-step WL boost techniques are proposed in [20][21][43]. In [43], for example, proper control of wordline pulse width and lower bitline voltage are used to improve the cell's stability. However, limited WL pulse width degrades the cell's write-ability, so a write operation in this scheme requires two phases. The first WL phase is a narrow pulse during which the bitline discharges to preserve cell stability, whereas the second WL phase is a wide pulse and used to write the cell. In addition to the timing complexity associated with this scheme, the use of two-phase write operation limits cell speed.

Nanometric low-voltage-operated SRAM cells suffer stability issues even with normal WL operation; so current thinking has moved instead toward suppressing the WL [30][17][18], thereby rendering the use of conventional WL boost techniques questionable. Although wordline suppression technique increases the 6T cell's SNM and thereby enhances cell stability, it causes a dramatic cell drivability degradation. Cell drivability limitation in a low supply voltage-operated SRAM cell is detrimental to the cell's minimum operating voltage V_{DDmin}. As such, designers usually resort to increasing the cell's area to maintain an acceptable cell drivability [16].

Dynamic and dual power supply SRAM designs are used to reduce the read and write margin interdependency [15]. In this approach, the SRAM cell operates at two V_{DD} levels based on the intended operation. To increase the cell RDM and enhance the cell's stability (high SNM), a high voltage supply (V_{DDH}) is used during read operations. On the other hand, a low supply voltage (V_{DDL}) or a floating V_{DD} is used when the cell performs a write

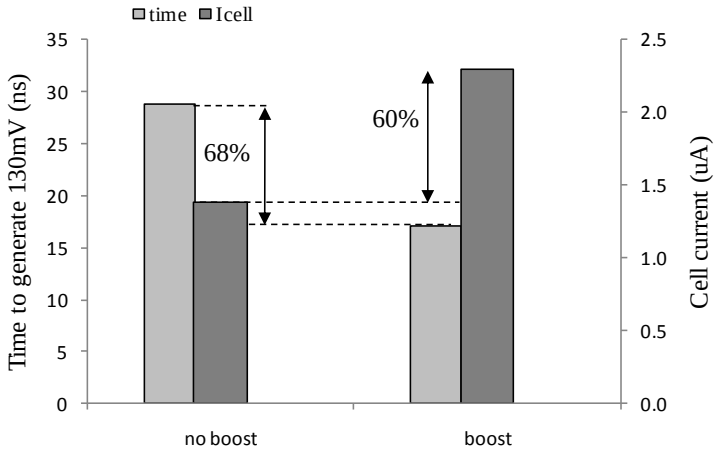


Figure 4.2: 400-mV 6T SRAM Cell Drivability and Speed Improvement Owing to a 100-mV DC WL Boost.

operation in order to improve the cell's WRM. In addition to the two required voltage supplies , the use of floating or low V_{DD} during a write operation highly threatens the stability of the half-selected cells on the same row.

Recently, new SRAM cell topologies have been reported as memory cell-level solution techniques (see **Section 1.5.2** and **Figure 1.9**). These topologies are mainly meant to break down the RDM/WRM interdependence in the conventional 6T SRAM cell [8][6][5][44][7]. This has been achieved by using separate read and write ports. The penalty associated with these topologies is area overhead. Additionally, some of these topologies abandon the symmetrical layout structure which is actually one of the great advantages of the conventional 6T cell.

A most promising potential replacement for the conventional 6T cell is the 8T cell proposed in [6]. Separating the read and write ports from each other gives flexibility to design the cell's α and β ratios separately. Even though this technique provides outstand-

ing reliability and performance improvement, the associated area and power overhead are significant. More importantly, this cell is incompatible with a column interleaving architecture due to the fact that half-selected cells are not vulnerable to noise during a write operation.

4.3 Wordline Boost: The Motivation

As presented in the previous section, increasing the access transistor strength helps to increase the cell's RDM and WRM margins. This can be accomplished by increasing the access transistor width or overdrive voltage. As shown in **Figure 4.2**, a DC 100-mV WL boost can improve the 6T cell's drivability and speed by 60% and 68%, respectively. Increasing the cell drivability without the need to oversize the actual cell area is an important advantage of the WL boost. This is particularly beneficial for SRAM cells operating at the minimum voltage (V_{DDmin}), where cell area increase is necessary to maintain cell stability and to satisfy other performance specifications.

However, a DC WL level boost can degrade the cell's SNM significantly. In fact, a DC WL boost is usually used to measure the cell's stability in read access mode by increasing the WL signal level above the cell's supply voltage V_{DD} and observing the point at which the cell fails. This voltage level defines a cell stability parameter that signifies the maximum tolerable DC voltage rise on the WL before causing a read upset, which known as wordline read retention voltage WRRV [36]. **Figure 4.3** shows the RDM and WRM of a 400-mV operated 6T SRAM cell as a function of WL DC boost. As can be seen from **Figure 4.3**, a steady improvement in both RDM and WRM margins can be achieved if the boost level is kept bellow 100 mV (25% above nominal). However, if the boost level is increased beyond 130 mV, *i.e.,* 33% above the nominal 400-mV WL voltage, the RDM drops significantly

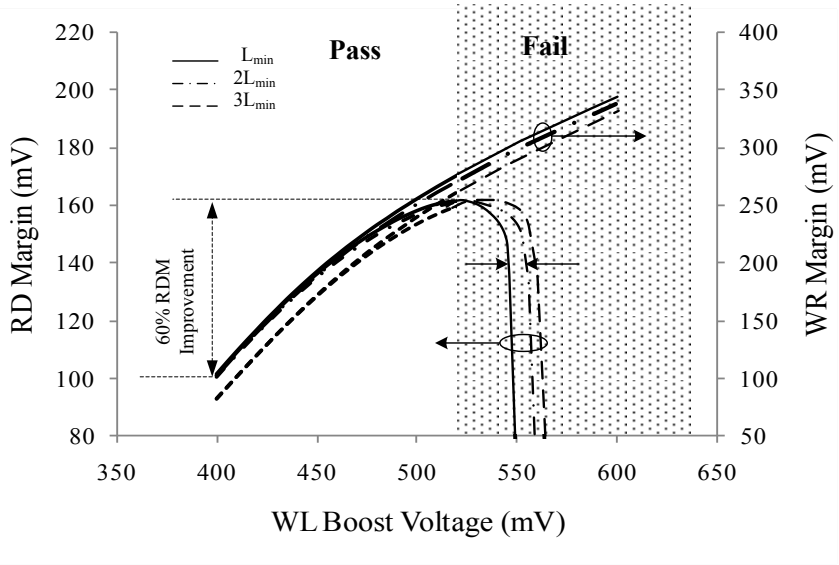
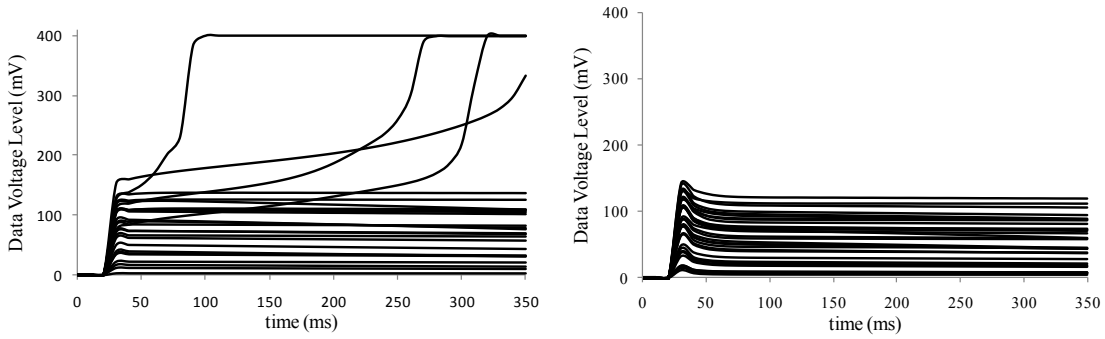


Figure 4.3: SRAM Cell RD and WR Margin Improvement as a Function of WL Boost Level.

and results in read failures, whereas WRM continues monotonically increasing as the boost level increases.

Within this boost range, as indicated in **Figure 4.3**, both the RDM and WRM can be improved by 60 mV and 140 mV, respectively, compared to normal WL operation. As is further shown in **Figure 4.3**, channel lengths that are twice the minimum length ($2L_{min}$) can add an additional 10 mV boost peak without degrading the WRM or RDM. Additionally, the increased access transistor channel length adds another advantage to using a WL boost by reducing the cell leakage current, which improves the cell's overall performance.

Nevertheless, cell stability under low operating supply voltage conditions is vulnerable to process and mismatch variations even under normal WL operating conditions. **Figure**



(a) Conventional WL

(b) Decaying WL Boost Scheme

Figure 4.4: Transient Simulation Results Showing Data Zero Level Degradation in the Presence of Process Variations.

4.4 shows transient Monte Carlo simulation results of the 6T cell highlighting storage node (*i.e.* data) stability in the presence of process variations. **Figure 4.4(a)** verifies that, under normal WL operation, the cell exhibits some destructive (DRD) failures because of process and mismatch variations. Thus, more failures are expected if the WL signal is DC boosted. Recently, researchers investigated the concept of the dynamic noise margin (DNM) and its application in SRAM cell stability [45][35][46]. They argued that an SRAM cell is a dynamic system with retention and access modes. The cell stability can be enhanced if the access time is reduced.

We exploit this property of an SRAM cell to devise a programmable transient WL boost scheme to improve low-voltage-operated 6T SRAM cell yield. The proposed boost action takes place at the commencement of the cell's access mode and transitionally decays. During this transitional period, the cell is expected to endure high noise levels and to discharge a considerable amount of bitline charge. This has been verified with Monte Carlo transient simulation results, shown in **Figure 4.4(b)**. As seen in the figure, a transitional

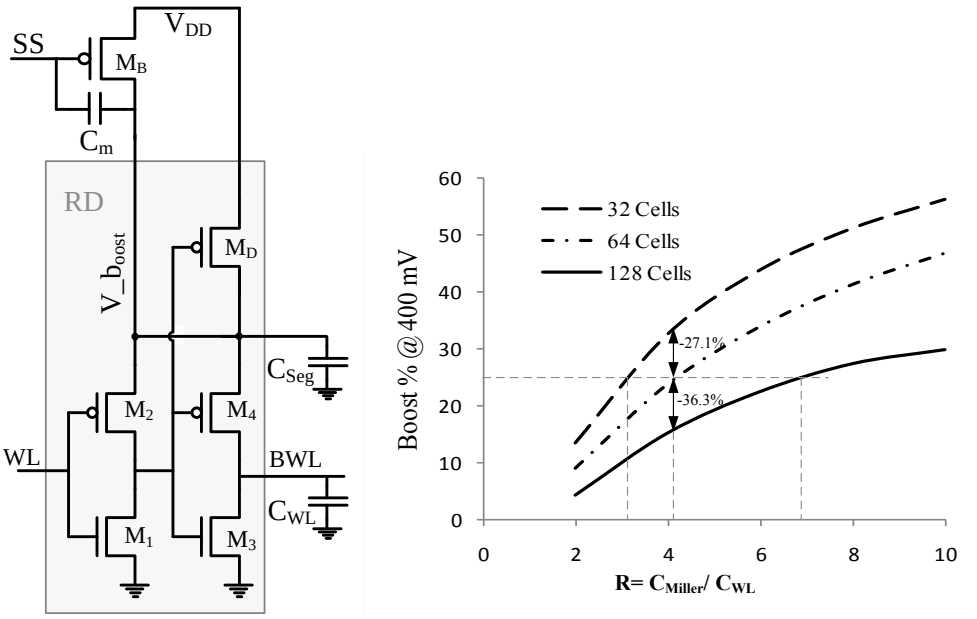
100-mV peak WL boost eliminates DRD failures existing in **Figure 4.4(a)** and maintains cell stability.

Two conclusions can be drawn from the simulated results shown in **Figure 4.4(b)**. First, the strengthened (overdriven) access transistor increases cell drivability and helps the cell to discharge a considerable amount of the bitline charge during the boost interval. This reduces the bitline's impact on the stored data (low zero level degradation). Second, charge feed-through at the complementary storage node (the node that stores "1") increases the voltage level at this node above V_{DD} which in turn increases the overdrive voltage of the driver transistor.

Boost peak and interval are two fundamental components in this scheme. Therefore, a programmability feature is added to the proposed WL driver to control these two components. Exploiting the WL boost adds the benefit of a cell leakage current reduction by optimizing the cell's access transistor channel length without compromising cell performance. In addition, we employed the read-assist write-back sense amplifier proposed in Chapter Three (Scheme II) to further enhance the DRD failures. A 4-Kbit SRAM subarray of the conventional 6T SRAM cell is used as a circuit under test (CUT) to investigate the effectiveness of the proposed scheme.

4.4 Proposed Programmable WL Boost Driver

A programmable WL boost circuit is desirable so that the programmed settings can be optimized for different PVT conditions while ensuring that SRAM instances do not suffer from soft failures. Arguably, several different shapes of transient WL boost signal may be realized through circuit means. However, the shape of the signal must be a compromise between circuit simplicity, its effectiveness to enhance the performance, and the stability



(a) Schematic Diagram (b) Boost Level Dependency on Capacitance Ratio

Figure 4.5: Proposed Boosted WL Row Driver (RD).

of the SRAM. A boost signal with short rise time and relatively large fall (decay) time was found to be optimal.

Figure 4.5(a) shows the basic circuit diagram of the proposed boosted WL driver. The row driver (RD) is a buffer consisting of two inverters. The source of PMOS transistors M2 and M4 are connected to the boost node (V_{boost}). The V_{-boost} node is shared by a number of row drivers. Therefore, transistor MB and Miller capacitance (C_m) are shared by a segment of N rows. The output of the second inverter drives the wordline. The Segment Select (SS) signal is decoded from the row address decoder.

This WL boost circuit is designed based on the charge feed-through (Miller) concept. The boost capacitor C_m is initially charged to V_{DD} through the PMOS transistor M_B.

Upon the assertion of the SS signal, the rising edge injects charge to the RD parasitic capacitance C_{Seg} owing to charge feed-through via the Miller capacitance C_m and raises the voltage level at node V_{boost}. At the same time, the row decoder output activates one row driver signal WL. The row driver boosted WL (BWL) output signal charges to a higher than nominal V_{DD} voltage.

In order to limit the boost level impact, PMOS transistor M_D is used to damp out the boosted voltage at node V_{boost} and exponentially bring back BWL to V_{DD}. The BWL boost level is determined by the C_m/C_{WL} ratio and the number of cells per segment attached to the boost circuit, represented by C_{Seg}; where, C_{WL} and C_{Seg} are the parasitic capacitance loading of the WL and the segment's diffusion capacitance, respectively. **Figure 4.5(b)** depicts the boost level's dependence on the C_m/C_{WL} ratio and the segment loading C_{Seg} (typical values of 32, 64, and 128 rows per segment were investigated). For this particular experiment, the WL driver is designed to drive a segment of 32 rows. The boost level decay rate is determined by the time constant of the RC circuit at node V_{boost} which comprises the segment's diffusion capacitance C_{Seg} and the on resistance of transistor M_D.

In compliance with the cell's DNM concept, the reliability of the 6T SRAM cell operating under boosted WL conditions depends on two factors: the boost level and the boost interval. A controllable WL boost peak and interval can enhance SRAM cell reliability by means of fine tuning the boosted WL signal to minimize failures, so the proposed WL driver is designed to support multiple boost levels and different time intervals.

Multiple boost levels are achieved by using cascaded boost circuits that allow adding or removing of boost capacitance as needed. This has been accomplished by the use of the circuit illustrated in **Figure 4.6**. Different combinations of three parallel boost capacitances (C_{m1}, C_{m2}, and C_{m3}) are used to generate multiple boost levels. Each capacitance is invoked to the circuit via a control signal C_n. A three-bit control data pattern provides

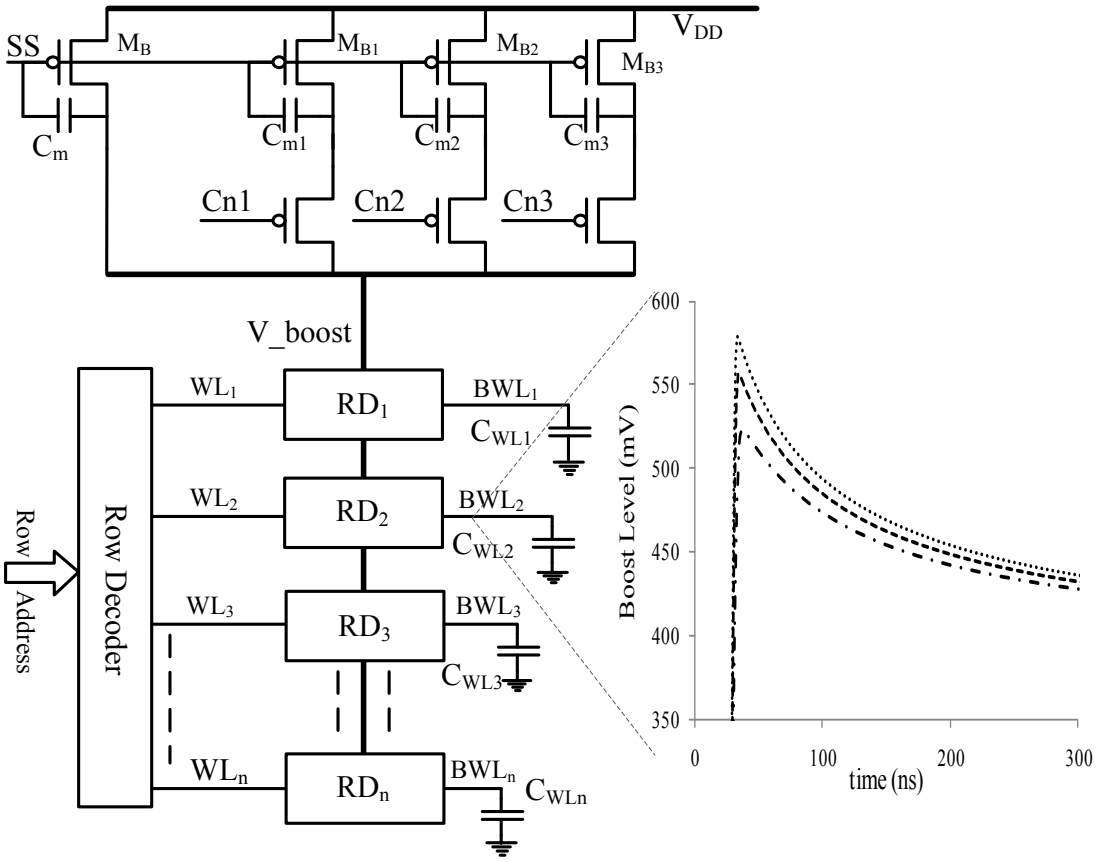


Figure 4.6: Proposed Multiple Level WL Boost Driver with Output WL Signal Simulation Results.

Table 4.1: Capacitance Ratio and Boost Level Control Data Pattern.

C_{n1}	C_{n2}	C_{n3}	C_m/C_{WL}	$V_{boost}(mV)$
1	1	1	2.0	100
0	1	1	2.5	120
1	0	1	3.0	140
0	0	1	3.5	155
0	0	1	4.0	165
1	0	0	4.5	175
0	0	0	5.0	190

eight different boost levels. The WL loading capacitance (C_{WL}) was extracted from the layout of a 128-cell row and found to be about 100 fF. So, for simulation purposes, C_m is chosen to be 200 fF which gives a C_m/C_{WL} of 2. Considering this capacitance ratio, simulation results show that a boost peak level of 100 mV can be achieved. This boost level is considered the default boost peak. The controlled boost capacitance values were selected as follows: C_{m1}=50 fF, C_{m2}=100 fF, and C_{m3}=150 fF. **Table 4.1** gives the control data pattern along with the associated capacitance ratio and simulated WL boost levels. **Figure 4.6** further shows the output boosted WL signal simulation results overlaid onto the proposed multi-level boost WL driver.

The boost interval is determined by the boost signal decay rate. In order to add more flexibility, a boost interval programmability option is added to the proposed driver. The boost interval is controlled by controlling the RC time constant of the boost level damping circuit comprises C_{Seg} and the "on" resistance of the PMOS transistor M_D. For a given C_{Seg} value, this resistance can be modulated by the M_D transistor current. A controllable

Table 4.2: Decay Rate Control Data Pattern

V_{cn1}	V_{cn2}	V_{cn3}	$decayrate(ns)$
0	0	0	16.0
0	0	1	12.0
0	1	1	8.0
1	1	1	5.0

current mirror circuit, shown in **Figure 4.7**, can be used to control the RC time constant.

In this experiment three transistors are used with three control signals (V_{cn1}, V_{cn2}, and V_{cn3}). Different monotonic decay rates are achieved by using a thermometer data pattern of the three control signals from 000 to 111. This data pattern allows us to generate four decay rates. Additionally, in order to eliminate current bleeding associated with the current starved transistor, a segment select pass gate is used to break the current path to ground when the segment is not selected. The control signal data patterns and the obtained decay rate simulation results of the proposed circuit are shown in **Table 4.2**. **Figure 4.7** illustrates the complete driver circuit with boosted WL output signal simulation results.

4.5 Employing The RA-WRBK-SA

Under low voltage operation, the time it takes the SRAM cell to generate an adequate bitline differential voltage is relatively large (low frequency operation). The long lasting zero level degradation could result in a destructive read operation and threaten the cell's stability. This becomes even worse when a high-level WL boost is used, as shown in **Figure 4.8(a)**. In order to prevent this, the read-assist write-back SA proposed in **Section 3.5**

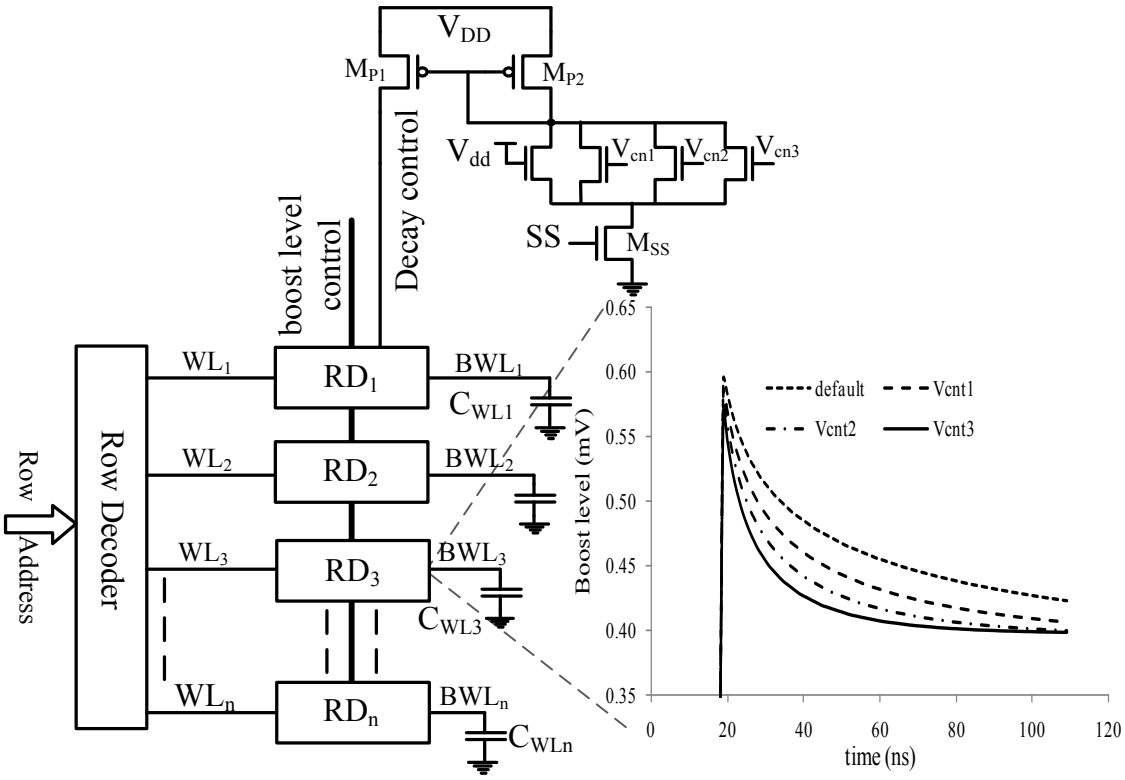
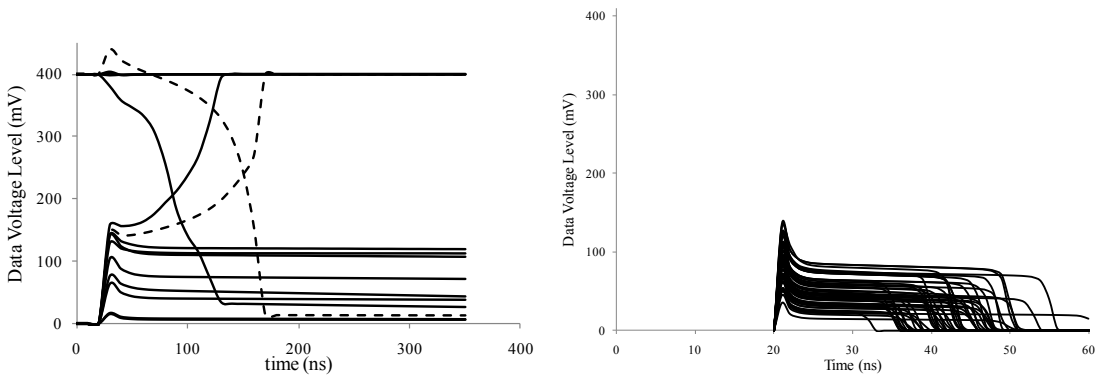


Figure 4.7: Decay Rate Control Circuit Diagram and Generated WL Boost Output Signal Simulation Results.

is used. The proposed SA serves two functions: to assist the cell during read operations by providing a positive feedback path to accelerate the bitline discharge process, and to rewrite the data back to the cell.

Figure 4.8(a) shows the results of Monte Carlo simulations of 400-mV 6T cell data stability during the read access mode. As can be seen, a relatively high boost level results in DRD failures. DRD failures happen because of the high WL level and long lasting zero level degradation associated with the read operation. If a read-assist mechanism is added to speed up the bitline discharging process, data-level degradation can be lowered and the



(a) High Boost Level Without RA-WRBK-SA (b) Data Recovery Using RA-WRBK-SA

Figure 4.8: Advantage of Using RA-WRBK Sense Amplifier in Elimination of DRD Resulted from High WL Boost Level.

cell will correctly retain the data.

Moreover, if the bitline is completely discharged to ground during a read operation, the cell can recover the stored data in a write-back operation. A read-assist write-back sense amplifier (RA-WRBK-SA) is usually designed to perform this operation [9][47]. **Figure 4.8(b)** depicts the data stability of the 6T cell operating under the same conditions but with the aid of the RA-WRBK-SA. As can be seen, the zero level degradation value and the interval have been reduced due to the read-assist and write-back operations, respectively.

The early activation of the RA-WRBK-SA provides the cell with a continuous read-assist action through the NMOS positive feedback loop. As such, upon the activation of the WLE signal, both the sense amplifier and the memory cell are working together to discharge the bitline. Accordingly, data zero level degradation stays low until the cell write-back when the bitline discharge completely. **Figure 4.9** shows a comparison of the bitlines' response with performance enhancement provided by the proposed scheme to that

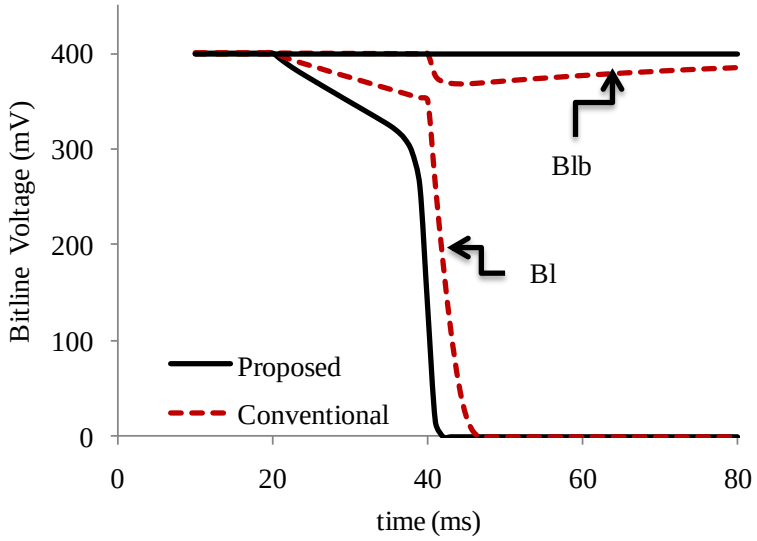


Figure 4.9: Bitline Response Comparison: Solid Line Proposed, Dashed Curves [9].

used in [9]. As we can see, the bitline discharge trend is faster when RA-WRBK-SA is utilized.

4.6 Simulation Results and Discussion

The proposed WL driver scheme was designed in ST 65-nm CMOS technology to operate on a 32-Kbit 6T SRAM array. The SRAM macro was designed to operate at a 400-mV supply voltage. Each column is segmented into eight 32-cell segments. Post-layout simulations were used to verify the proposed scheme's functionality. The extracted 128-cell WL loading capacitance was found to be 100 fF. The default boost capacitor was correspondingly set to 200 fF, *i.e.*, (C_m/C_{WL})=2.

Additional boost capacitances are set to 50, 100, and 150 fF (schematic instance).

This corresponds to (C_m/C_{WL}) ratios of 2.5-5.0, depending on the control signal pattern given in **Table 4.1**. Control signals C_{m1}, C_{m2}, and C_{m3} are selectively used to invoke the required value of C_m for the required boost level. Boost level control simulation results are illustrated in **Figure 4.6**. The boost levels shown correspond to capacitance ratios of 2.5-4.5. Using different combinations of the control signals C_m, boost levels ranging from 25% to 90% are achieved. This provides a flexibility to test the cell stability under different stress levels.

As for the boost interval, the control signals C_{n1-3} are used to obtain different decay rates for given boost levels. Simulation results shown in **Figure 4.7** indicate that a decay rate ranging from 5 ns to 16 ns is achieved using the control signal data patterns given in **Table 4.2**. The maximum decay rate corresponds to the default state in **Figure 4.7** where none of the control signals is active, whereas the minimum rate corresponds to the case where all control signals are high. The decay rate is calculated as the time for the boost level to fall 50% below the maximum. The segment select NMOS transistor is activated only when the corresponding segment is selected. Table 4.2 gives the control signal pattern and the corresponding WL signal decay rates.

4.7 Performance and Yield Analysis

The proposed WL driver was used to drive a segment of 32 rows in a 4-Kbit (32x128) SRAM sub-array laid out in ST's standard CMOS technology. Post-layout simulations were used to extract the 32-cell column segment and 128-cell row loading capacitance. A conventional WL driver was used to drive another 4-Kbit sub-array to compare the cell performance and stability in two different environments using Monte Carlo simulations.

Performance simulations are used to investigate the cell's figures of merit, such as cell

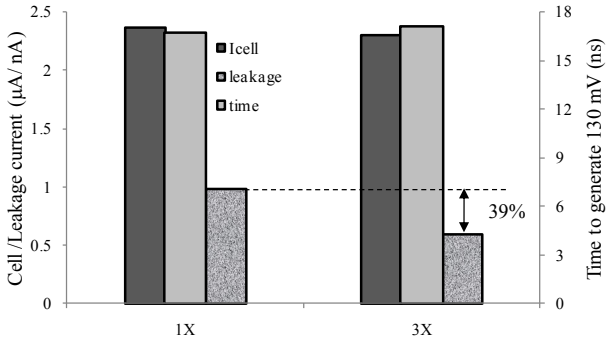


Figure 4.10: Leakage Current Reduction Associated with Three Times Increase in Access Transistor Channel Length.

current, leakage current and the mean value of the developed bitline differential voltage. A nominal boost level of 100 mV with a decay rate of 16 ns is used. **Figure 4.4(a)** depicts a 6T cell's stability when a 100-mV boost is used, as opposed to an unstable cell operating with a conventional non-boosted WL, **Figure 4.4(b)**. However, increasing the boost level to 155 mV causes some DRD failures, as shown in **Figure 4.8(a)**. These failures are eliminated by using RA-WRBK-SA sense amplifier, as confirmed in **Figure 4.8(b)**.

As stated in **Section 4.2**, and shown in **Figure 4.3**, increasing the access transistor channel length under boosted WL operation results in cell leakage reduction without degrading performance. Simulation results, shown in **Figure 4.10**, confirm this and show that, by using three times the minimum channel length for the access transistor, a 39% leakage current reduction is achieved with only minor changes in other cell parameters (cell current and speed). In addition, the simulation results shown in **Figure 4.11** indicate that a 28.5% bitline differential mean value improvement is achieved when WL boosting is used.

Reliability simulations are used to explore cell stability as a function of WL boosting.

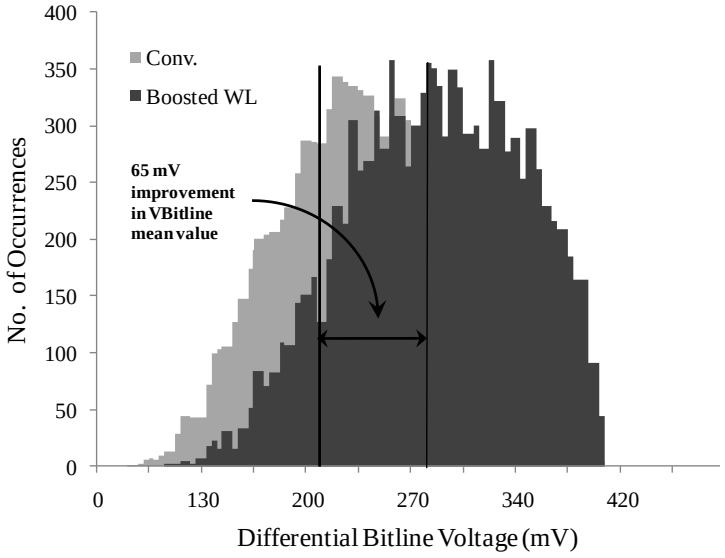


Figure 4.11: Improvement in Bitline Differential Voltage as a Result of Using 100-mV/16-ns WL Boost.

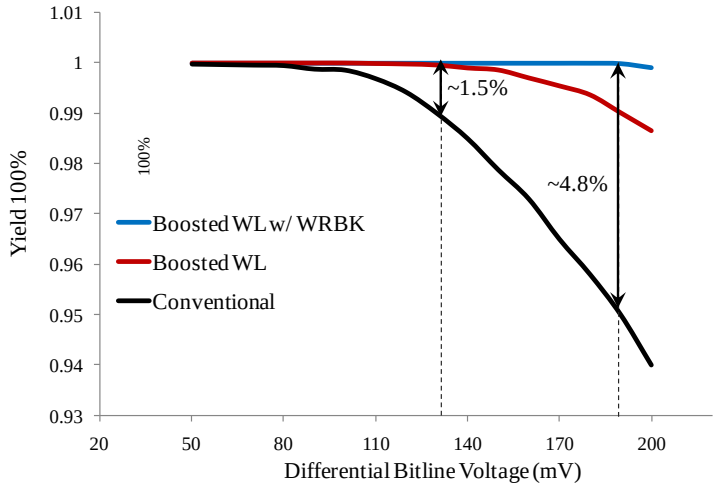


Figure 4.12: SRAM FIR Rate Improvement Using Boosted WL and RA-WRBK-SA Compared to Conventional WL.

The main reliability metric used here is the FIR. The pass/fail criterion in FIR analysis is based on the cell's ability to generate a targeted bitline differential in the presence of process and mismatch variations. FWR failures are excluded since the impact of WL boosting is expected to be favorable to WRM.

Monte Carlo simulations are conducted for an SRAM cell under normal WL, 100 mV/16 ns boosted WL and boosted WL with RA-WRBK-SA. **Figure 4.12** shows that when the targeted bitline differential voltage is set to 130 mV, the use of the proposed WL boost technique reduces the FIR rate by up to a 1.5% compared to normal WL operation. This rate is further improved when a higher bitline voltage is targeted and RA-WRBK-SA is employed. As can be seen in **Figure 4.12**, a 4.8% reduction in FIR rate compared to normal WL operation is achieved.

Moreover, the read-assist mechanism of the RA-WRBK-SA helps the cell to develop a higher bitline differential in a given time interval. Simulation results presented in **Figure 4.13** show that the use of the RA-WRBK-SA contributes an extra 10% improvement in the bitline differential mean value.

4.8 Summary

Low-voltage operated 6T SRAM cell reliability was discussed in this chapter. Traditionally, WL boost was used to overdrive the gate-to-source voltage of the cell access transistor. However, DC WL boost can cause an increase in destructive read rate. Therefore, a level/interval programmable boost WL design was presented. A 400-mV 6T SRAM cell performance and yield were investigated utilizing the proposed scheme in the presence of process and mismatch variations. High-level boost can cause an increase in the destructive

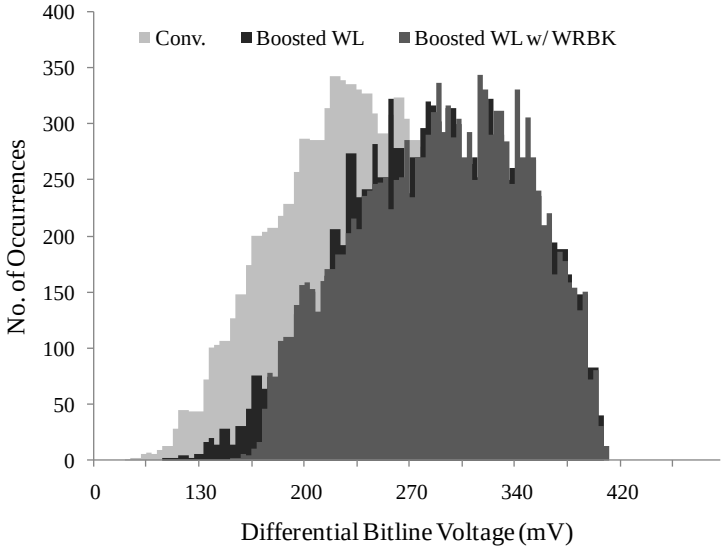


Figure 4.13: Differential Bitline Voltage Improvement as a Result of Boost WL and RA-WRBK-SA.

read rate; therefore, we employed the RA-WRBK-SA proposed in Chapter Three to eliminate any DRD failures that may arise due to unexpected fluctuations in the WL boost peak or interval.

The proposed WL driver was used to drive a segment of 32 rows in a 4-Kbit (32x128) SRAM sub-array laid out in ST standard CMOS technology. Post-layout simulations are used to extract the 32-cell segment and 128-cell row loading capacitance. A conventional WL driver is used to drive another 4-Kbit sub-array to compare the cell performance and stability in two different environments using Monte Carlo simulations. Monte Carlo simulations are conducted to validate the proposed scheme's functionality in the presence of process and mismatch variations. A yield improvement of 4.8% is achieved when a combination of the proposed WL boost driver and RA-WRBK-SA are used. The mean value of the bitline differential voltage is improved by 38% compared to a conventional WL

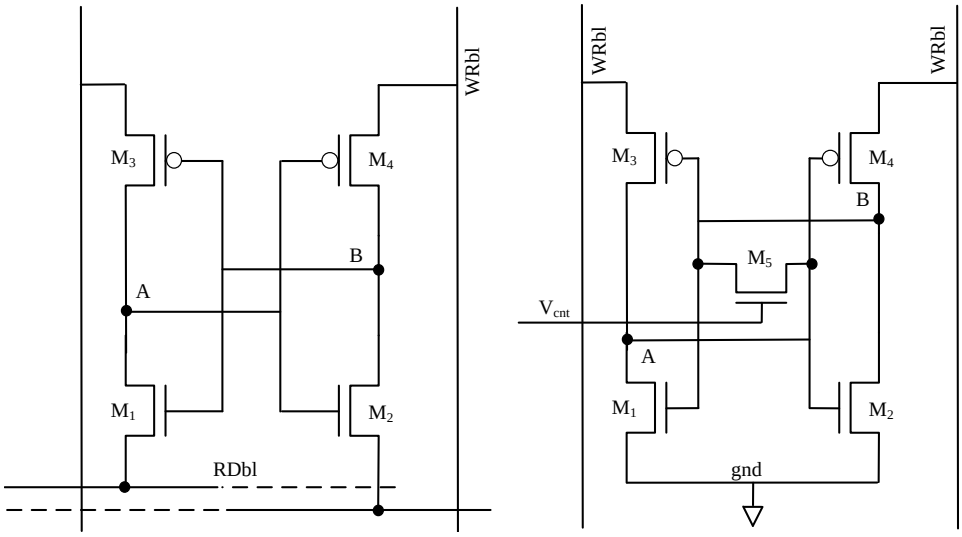
driver. Additionally, a leakage current reduction of 39% is obtained by doubling the access transistor channel length.

Chapter 5

New Five-Transistor 5T SRAM Bitcell Topology for Low Power Applications

5.1 Introduction

For decades now, the conventional six-transistor 6T SRAM bitcell, shown in **Figure 1.7**, has been considered the workhorse for embedded memory applications. However, in the nanometric CMOS regime, designing a reliable, low-voltage operated 6T SRAM array has proved challenging [9]. The use of a common port to perform both read and write operations creates a 6T cell design conflict. Design for reliable read operation with high RDM and SNM results in low WRM and *vice versa*. For this reason, alternative bitcell topologies with separated read/write ports have been proposed (refer to **Section1.5.2**) [8][6][5][7]. These topologies are, in general, based on a performance-area trade-off. Furthermore, they



(a) 4T Bitcell [48] (b) 5T Bitcell [49]

Figure 5.1: Conventional "Access-Less" 4T and 5T SRAM Bitcell Topologies.

all utilize the conventional 6T bitcell as a core storage element. Therefore, the 6T bitcell design reliability issues have been shifted but not solved.

Since the access transistors in the conventional 6T bitcell have no contribution in the data storage mechanism, another way to treat the cell storage node-bitline interaction is by eliminating the cell's access transistors. State-of-the-art "access-less" SRAM bitcells, shown in **Figure 5.1**, are reported in [48][49]. The operation of this kind of SRAM bitcell is based on the idea of eliminating the access transistor and using the load and/or driver transistor as an access transistor in addition to its main duty as load or driver transistor.

In [48], an area efficient four-transistor (4T) cell is reported, however a considerable cost is introduced into the array interface in a form of the different voltage levels and multiple clock phases needed to perform reliable RD and WR operations. For example, in order to read the cell, the voltage level at the read port is raised above its nominal

value but must not exceed a certain limit, otherwise, the cell may lose the stored data. Similarly, a write operation is also based on certain level changes in bitlines to ensure the stability of the half-selected cells. Wieckowski in [49] proposed a five-transistor (5T). Although only five transistors are used to implement a single data bitcell, the deviation of the designed cell parameters from conventional 6T cell is significantly large. For example, for iso-cell drivability design (I_Cell), a seven times (7X) cell area overhead is required. Also, the iso-cell area design results in 6X and 23X degradation in cell drivability and SNM, respectively.

SNM degradation in the 6T is mainly attributed to zero level degradation created by the access transistor during access mode. This level degradation can be exacerbated by the positive feedback gain of the cross-coupled configuration and thus lead to a destructive read operation. Takeda [44] proposed a 7T bitcell topology to eliminate the zero level degradation influence and improve cell SNM. In this topology the closed-loop positive feedback gain is controlled via an additional transistor added to the conventional 6T cell. This transistor, along with a control signal, isolates the cell's storage nodes and eliminates the impact of the zero level degradation on cell stability (SNM).

Although this cell topology provides significant improvement in SNM, area and power overhead is not negligible. In addition to a 13% increase in the cell's area, the proposed cell's functionality requires the use of separate read/write WL, plus extra control signal to control the closed-loop gain. More importantly, this topology converts the 6T cell from differential to single-ended bitline signalling. In contrast, the use of an asymmetrical 6T cell configuration [34] can provide the same SNM improvement without a need to using extra transistor and control signals. In this chapter we present a new "access-less" five-transistor (5T) SRAM bitcell that shows promising performance improvements compared to the aforementioned bitcells and the conventional 6T cell.

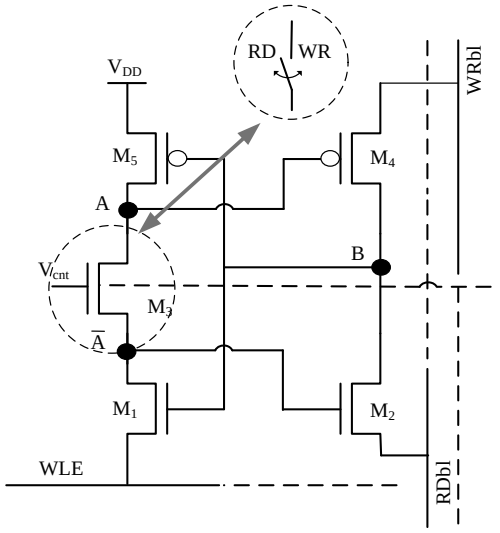
5.2 Proposed 5T SRAM Bitcell

5.2.1 Cell Concept and Operation

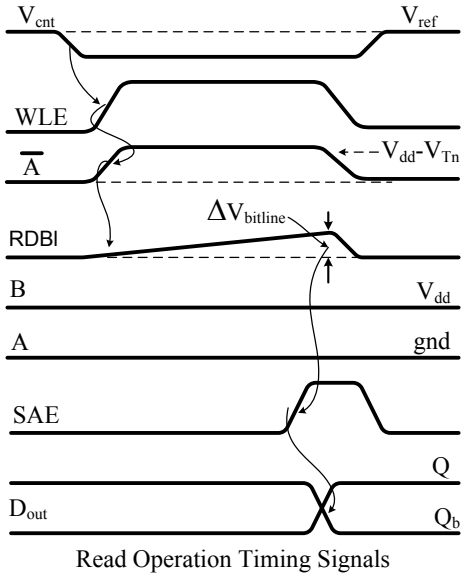
The purpose of the proposed bitcell topology is to isolate the read and write operations from each other and to eliminate the unnecessary two access transistors. The first objective is accomplished by selectively controlling a closed-loop positive feedback gain, while the second objective is achieved by using a specialized controlling (timing) scheme that allows the inverter's driver or load transistor to behave as an access transistor under certain operating conditions. **Figure 5.2(a)** illustrates the proposed 5T schematic along with the timing scheme used to perform read or write operations.

The cross-coupled inverters (M2-M4 and M1-M3-M5) form the core storage element of the proposed cell. The existence of the active latch configuration (in a cross-coupled inverters configuration) allows for the storage of data in a complementary fashion and ensures the static nature of the proposed cell. The two inverters in the 5T bitcell are referred to as a "read inverter" and a "write inverter". The data is stored at the output of inverter M2-M4 (node B) and this inverter is assigned to perform a read operation, so we refer to it as the "read inverter". The complementary data is stored at the output of the second inverter M1-M3-M5 (node A). This inverter is dedicated to performing the write operations, so we refer to it as the "write inverter". Transistor M3 transfers the voltage level at node A to node $\overline{A}$, therefore the voltage level at this node is a copy of the data complement at node A. This node we refer to as the "access node".

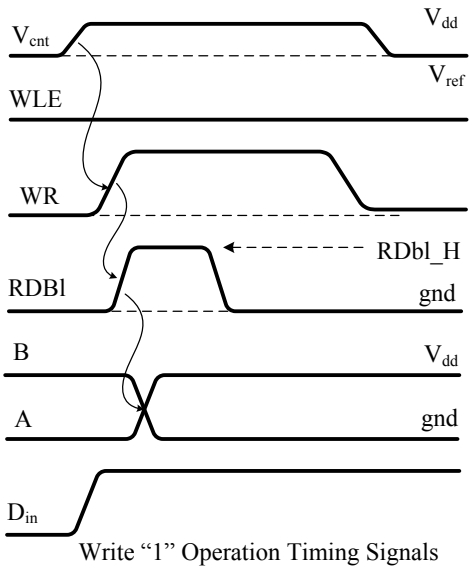
The WLE, read bitline (RDbl), write bitline (WRbl) and the control signal (V_{cnt}) are signals that are used to control cell operation. WLE, RDbl and WRbl are not merely control signals: they also act as the cell supply voltage under certain operating conditions.



(a) Schematic Diagram



(b) Read Operation Timing



(c) Write Operation Timing

Figure 5.2: Proposed 5T Schematic Diagram and Read/Write Operation Timing Scheme.

The default state of the WLE signal is "0" (low) which provides a ground path to the write inverter. The WRbl and RDbl default states are V_{DD} and gnd, respectively and they are used as the read inverter supply voltages. The control signal V_{cnt} is set at some reference voltage level (typically $V_{DD}/2$) and used to control the closed-loop feedback gain of the cross-coupled inverter configuration with the aid of transistor M3. As any other SRAM bitcell, the proposed 5T cell has two modes of operation: retention mode and access mode. During retention mode the cell must be stable (static) and can retain the data as long as it remains powered. During access mode, on the other hand, the cell performs either read or write operation.

5.2.2 Modes of Operation

(1) Retention Mode

In this mode of operation, all control signals remain at precharge voltage levels. WLE is precharged low (gnd) to serve as the ground for the write inverter; meanwhile, WRbl and RDbl are precharged to high (V_{DD}) and low (gnd) to serve as the read inverter's power supply rails V_{DD} and gnd, respectively. The mid-level of the control signal V_{cnt} keeps the pass transistor M3 in its high impedance state (partially on) to complete the pull-down path to ground. Thus, the cell schematic looks like the conventional asymmetrical cross-coupled inverters configuration shown in **Figure 5.3(a)**. The data stored at node B and its complement at node A are retained as long as the cell is powered. If node B is high, then M1 is "on" and pulls down the access node $\overline{A}$, thereby holding node A to "0". If node B is low, then M5 is "on" and holds node A at high. The access node $\overline{A}$ in this case is a fraction of the high voltage level at node A (V_{DD}) but is high enough (higher than the threshold voltage V_{TH2}) in order to keep transistor M2 "on" to hold the low data at node

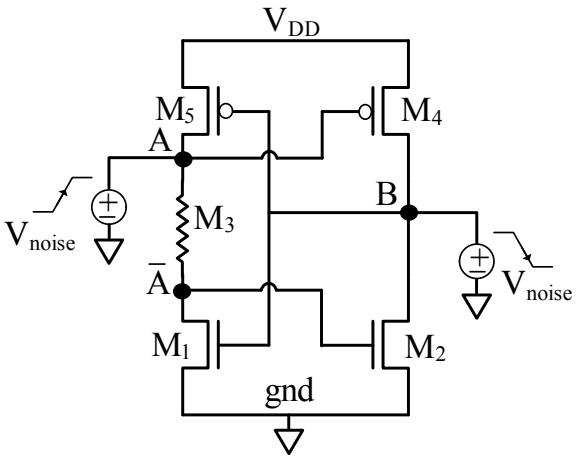
B.

The cell's Voltage Transfer Characteristics (VTC) curves during the retention mode can be established following the same procedure used in the 6T bitcell. **Figure 5.3(b)** depicts the VTC of the cell inverters during retention mode. The read inverter VTC (**Figure 5.3(b)**) is drawn by considering a noise voltage that is injected at node A when the stored data is high. The write inverter VTC (**Figure 5.3(c)**), on the other hand, is drawn by considering the noise injected at node B. The impact on node $\overline{A}$ due to the level degradation at node A is limited by the potential divider constituted by M3 and M1. The on state resistance of transistor M1 is much smaller than that of transistor M3 because of the difference in their V_{GS} voltage.

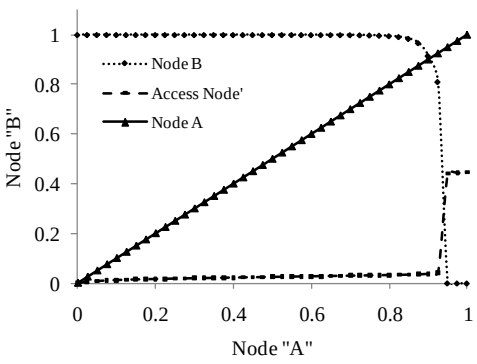
The cell's VTC curves can be found by superimposing the read and write drives VTC curves in a similar way to that used in a conventional 6T bitcell. To investigate the impact of control signal V_{cnt}, the read and write inverters' VTC are found for different values of V_{cnt}. **Figure 5.4** depicts the cell's VTC curves during retention mode as a function of V_{cnt} in contrast to conventional 6T VTC curves. Because of its single-ended nature, the SNM of the proposed cell is estimated based on the side of the maximum square that can fit in the bigger eye of the butterfly curves [33].

(2) Access Mode

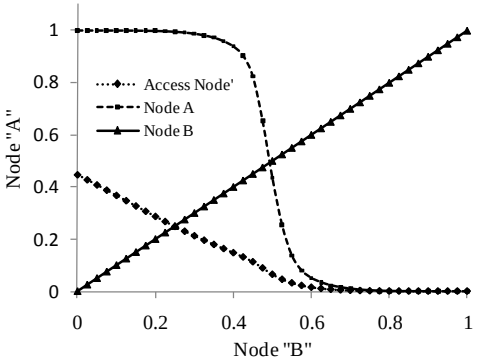
During access mode, the status of the control signals, WLE, RDbl, and WRbl, depend on the intended operation. If the cell performs a read operation, then V_{cnt} stays unchanged (reference level), or it can make a reference-to-low transition for enhanced performance as we will see later, while WLE makes a low-to-high transition. At the same time, RDbl is made to float and WRbl stays at V_{DD}. If the cell performs a write operation, then



(a) Cell Equivalent Schematic



(b) Read Inverter VTC



(c) Write Inverter VTC

Figure 5.3: Read and write Inverter Voltage Transfer Characteristics.

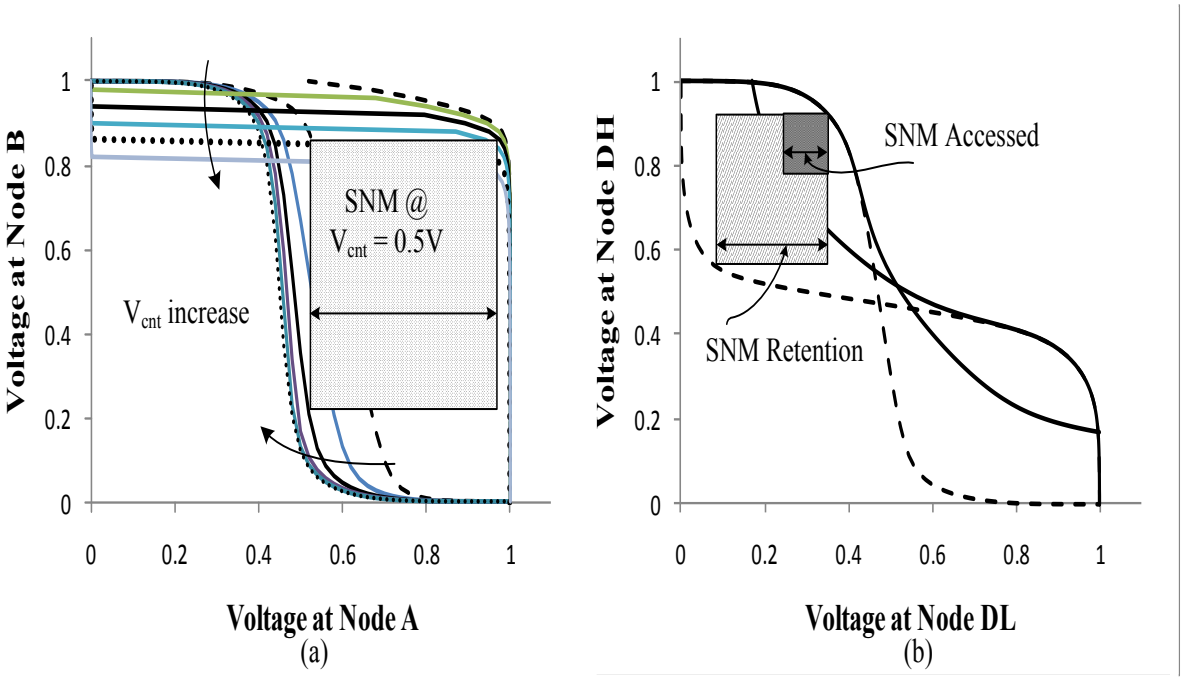


Figure 5.4: Proposed Cell VTC Under Retention Mode (a) in Contrast to Conventional 6T Cell (b).

V_{cnt} makes a reference-to-high transition, WLE stays low "0" and the RDbl, and WRbl conditions vary depending on data to be written ("0" or "1").

(A) Read Access Mode

A read operation starts by disabling the RDbl precharge circuit and enabling the WLE signal. The WRbl and V_{cnt} stay at V_{DD} and V_{ref}, respectively. If the stored data is high, transistor M1 is in the triode region "on" and both the voltages of node A and $\overline{A}$ are initially low. By asserting the WLE signal to V_{DD}, transistor M1 behaves as an access

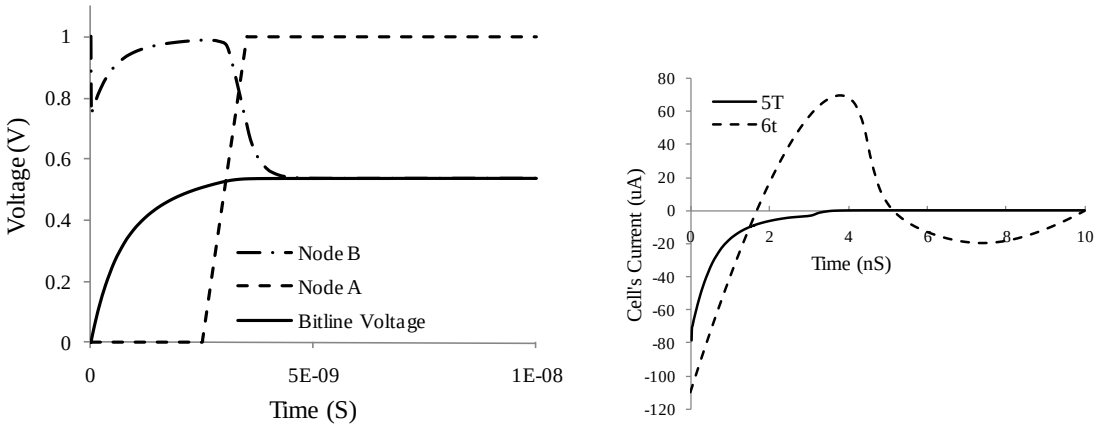
transistor and passes a weak high voltage (V_{DD}-V_{THn}) to the access node $\overline{A}$. This in turn turns on transistor M2 and make it behaves as an access transistor in order to communicate with the RDbl.

The low reference voltage V_{cnt} isolates node A from node $\overline{A}$ which prevents excessive zero level degradation at node A and keeps M4 in the triode region "on". Accordingly, a cell read current, sourced by M4, passes though M2 to charge up a pre-discharged RDbl parasitic capacitance. The highest voltage level RDbl can charge up to is limited to V_{DD}-$2V_{THn}$ at which point the gate-source voltage V_{GS} of transistor M2 reaches V_{THn} and M2 moves to the cut-off region.

At this point the cell current becomes zero and the stored data stays at the high level. In other words, transistor M2 shuts off the cell current and prevents further bitline charging. This not only preserves the data from being corrupted, but it is also stops the cell from bleeding during read operation and reduces the read power consumption. Voltage level degradation at node B does not accelerate because of the broken feedback loop (through the weak M3).

Figure 5.5(a) shows that under maximum RDbl loading and a full-swing WLE signal, the data level at node B and the RDbl capacitance maximum voltage are equalized at $V_{DD}/2$ which means that node B cannot flip under any circumstances. This is further confirmed in **Figure 5.5(b)** which demonstrates that the cell current eventually becomes zero. The cell's current behavior in **Figure 5.5(b)**, is compared to the 6T N-curve in which the change in current direction indicates a change in stored data.

If stored data is "0", transistors M1 and M4 operate in the cut-off region so that activating the WLE signal has no impact on the cells' voltage levels and the cell's output current to RDbl is zero. The only source available to charge up the RDbl in this case is



(a) Developed RDbl Voltage

(b) 5T Cell Current Compared to 6T N-Curve

Figure 5.5: The 5T Cell Stability During Access Mode.

the leakage current from selected and non-selected cells on the same column. However, the low leakage current property of the proposed cell (as we will see later), makes read "1" and read "0" operations clearly distinguishable.

To speed up read operations, transistor M3 can be completely turned off to isolate the zero level data at node A from the access node A and keep M4 fully "on". This can be accomplished by turning V_{cnt} off before activating WLE during a read operation. In this case PMOS transistor M4 has the maximum V_{GS} voltage which makes it capable of sourcing a maximum cell current to RDbl to boost the read operation speed. However, simulation results show that the improvement in read operation speed associated with this operation is not significant considering the extra required timing complexity and power consumption to activate the control signal V_{cnt}. That is mainly because of the zero level degradation by the low reference voltage level of V_{cnt}.

(B) Write Access Mode

A write operation is initiated by asserting the control signal V_{cnt} to a high voltage level (V_{DD}) while keeping WLE at ground potential. The high level V_{cnt} turns transistor M3 fully "on" and maximizes the feedback closed loop gain of the cross-coupled inverters structure. As a result, varying the input voltage of the write inverter will directly reflect its output. Bitlines RDbl and WRbl are used to perform write "1" or "0" operations, respectively.

1. **Write "0" Operation**

   When the stored data is "1", transistors M1, M4 are in the triode region and transistors M2, M5 are in the cut-off region. Write "0" is initiated by asserting the V_{cnt} to a high full swing voltage (V_{DD}) and pulling the WRbl down toward ground while keeping the RDbl bitline low (gnd). Under these conditions, transistor M4 behaves as an access transistor and discharges node B to change the write inverter input voltage level. The PMOS transistor M4 clamps the voltage drop at V_{THp}, so the write inverter trip point must be designed to be within that limit. The high closed-loop feedback gain helps to turn M2 "on" to fully discharge node B to zero. If the stored data is already "0", then transistor M4 is in the cut-off region and discharging the WRbl will not affect the cell's content in any event. **Figure 5.6(a)** illustrates the write "0" operation equivalent schematic diagram and the associated data and control signals.

2. **Write "1" Operation**

   Similar to a write "0" operation, when the stored data is "0", transistors M2, M5 are in the triode region and M1, M4 are in cut-off region. A write "1" operation is initiated by asserting the V_{cnt} to a full swing voltage (V_{DD}) and pulling RDbl

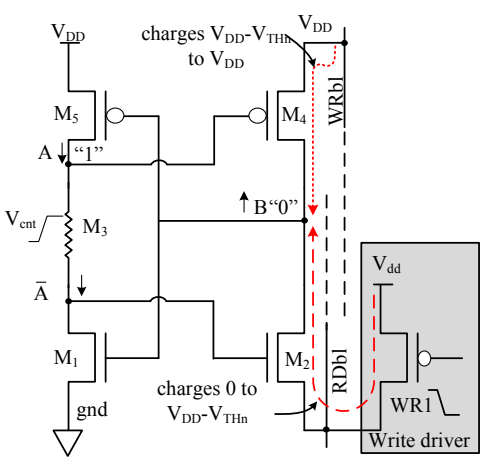
up toward V_{DD}, while holding WRbl to V_{DD}. Transistor M2 behaves as an access transistor and passes the bitline high voltage level to node B. As the voltage level at node B crosses the trip point, the write inverter flips and the data complement at node A changes to "0". Because M2 is an NMOS transistor, the maximum voltage level node B can go to is limited to V_{DD}-V_{THn}. However, the high closed-loop feedback gain accelerates discharging node A to zero and turns M4 "on" to fully pull node B up to "1". **Figure 5.6(b)** illustrates the cell equivalent schematic diagram during a write "1" operation and the associated data and control signals. If the stored data is already "1", transistor M2 is in the cut-off region and change in the RDbl voltage level will not affect the cell's content.

In order to verify the proposed cell write-ability, the cell's WR0 and WR1 margins were investigated in the presence of process and mismatch variations. **Figure 5.7 (a)** and **(b)** shows the carried out statistical simulations to verify the proposed cell write-ability.

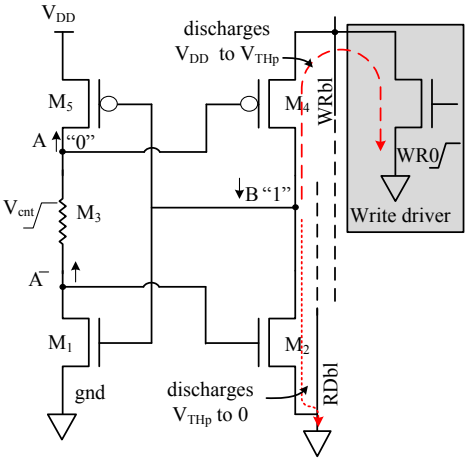
5.3 Cell Design Methodology and Stability Analysis

5.3.1 Read Inverter Design

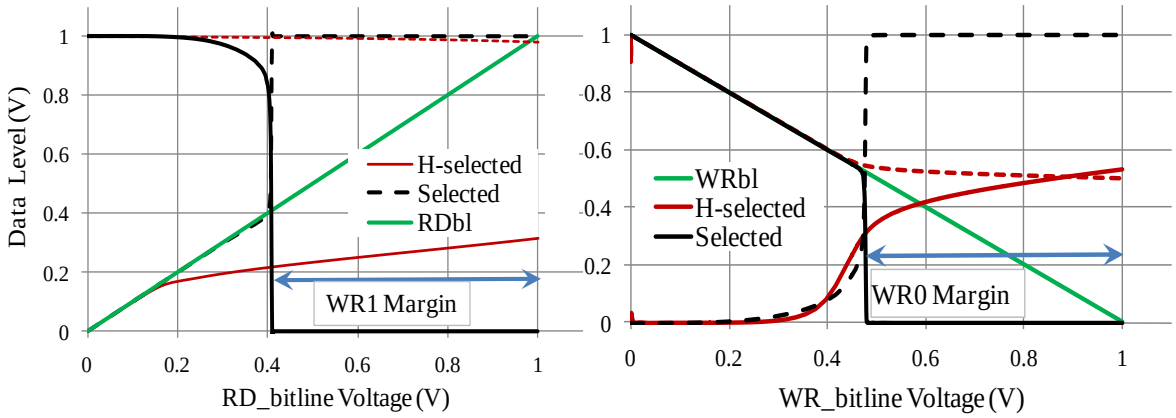
Because of the asymmetrical nature of the proposed cell, its two inverters can be designed independently. The read inverter is mainly designed for a reliable read operation by ensuring data stability and sufficient current to charge up RDbl through M4 and M2 (adequate read margin RDM). From a write operation perspective, however, the read inverter design is not crucial. Similar to the 6T cell's current representation, the proposed cell's current can be represented as the current required to charge up the bitline capacitance $C_{Bitline}$ to $V_{Bitline}$ in Δt time interval as defined in **Equation 2.6**, which is restated below:



(a) Write "0" Circuit Setup

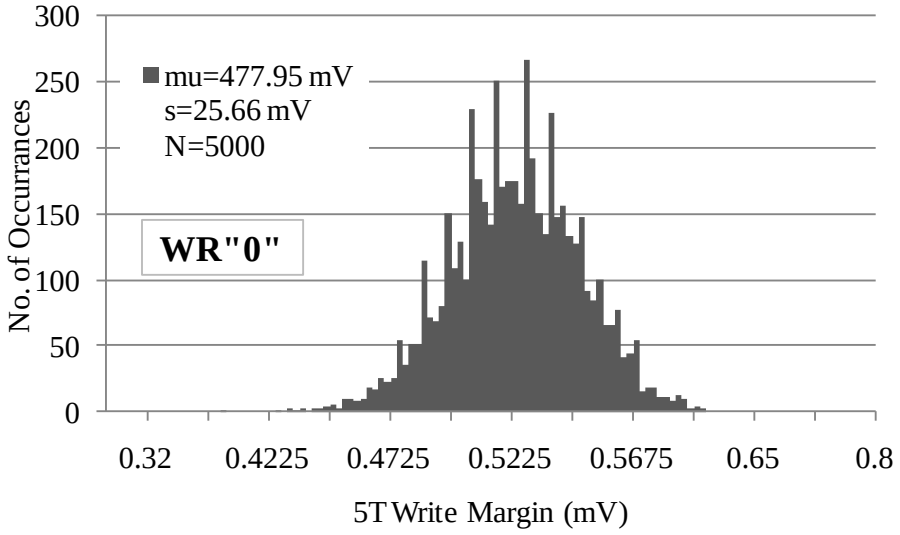


(b) Write "1" Circuit Setup

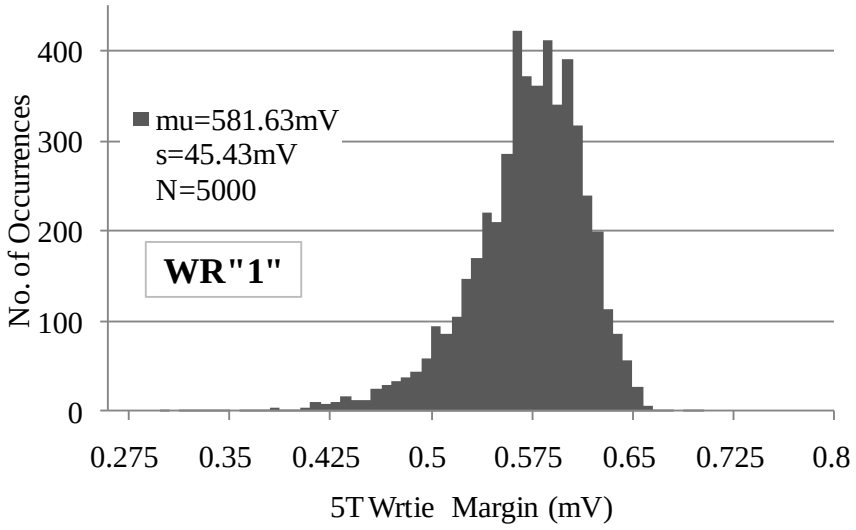


(c) WR"0"and WR"1" Selected and Half-selected Transient Simulation Results

Figure 5.6: 5T Write Stability: Selected and Half-Selected Data Stability During Write Access Mode.



(a) Write "0" WRM



(b) Write "1" WRM

Figure 5.7: 5T Write-ability Statistical Simulation Results in Presence of Process and Mismatch Variations.

$$I_{Cell} \geq C_{Bitline} \times \frac{\Delta V_{Bitline}}{\Delta t} + N \times I_{leakage} \tag{5.1}$$

where $C_{Bitline}$ is extracted from the post layout simulations and was found to be approximately 100 fF for a column with 256 cells assuming ST 65-nm technology. The required bitline voltage $V_{Bitline}$ and the time interval Δt are decided based on the sense amplifier accuracy and the targeted speed, respectively. **Figure 5.8** shows that the cell current is driven from the PMOS load transistor (M4) through the NMOS driver (M2). Since these two transistors are in series, their drain-source currents are equal, *i.e.,* I_{DSp}=I_{DSn} = I_{Cell}. Note that the cell current exists only when the stored data is high. During the read access mode, the voltage levels indicated in **Figure 5.8** suggest that both M2 and M4 are operating in the triode region with the following biasing voltages: V_{GS2}=V_{DD}-V_{TH}, V_{DS2}=V_{DD}-ΔV, $|V_{GS4}|$=V_{DD}, and $|V_{DS4}|$= the data level degradation at node B, ΔV. The high V_{GS2} and V_{DS2} values of the short channel transistor M2 drive it to the velocity saturation regime where the velocity saturation voltage V_{DSAT} is lower than the transistor's overdrive voltage V_{ov}. Using the generic drain current equation defined in **Equation 2.5**, I_{DS4} and I_{DS2} can be expressed by:

$$I_{DS2,4} = K_{n,p} \times (W/L) \times \left[(V_{GS} - V_{TH})\, V_{min} - V_{min}^2/2\right] \times (1 + \lambda V_{DS}) \tag{5.2}$$

where: V_{min} = min $(V_{ov}, V_{DSAT}, V_{DS})$; V_{ov} is transistor overdrive voltage, V_{DSAT} , is the velocity saturation voltage, and V_{DS} is the drain-source actual voltage.

Utilizing **Equation 5.2**, Appendix **A** shows that the allowable level degradation (ΔV) at node B determines the required M4/M2 ratio, which we refer to as the cell ratio "R". The level degradation ΔV results because of the charge sharing between the RDbl capacitance $C_{Bitline}$ and the cell diffusion capacitance at node B. **Figure 5.9** shows the simulation

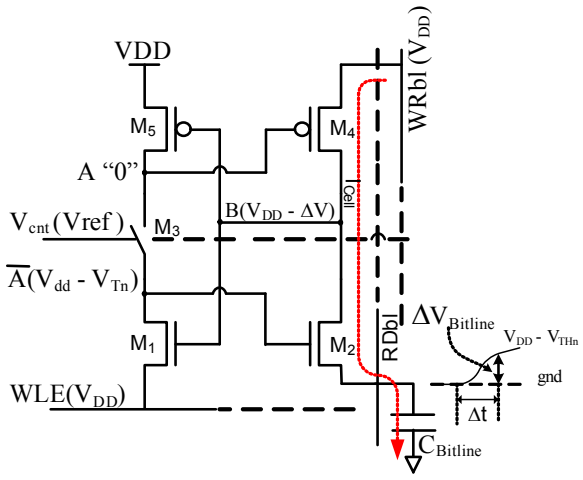
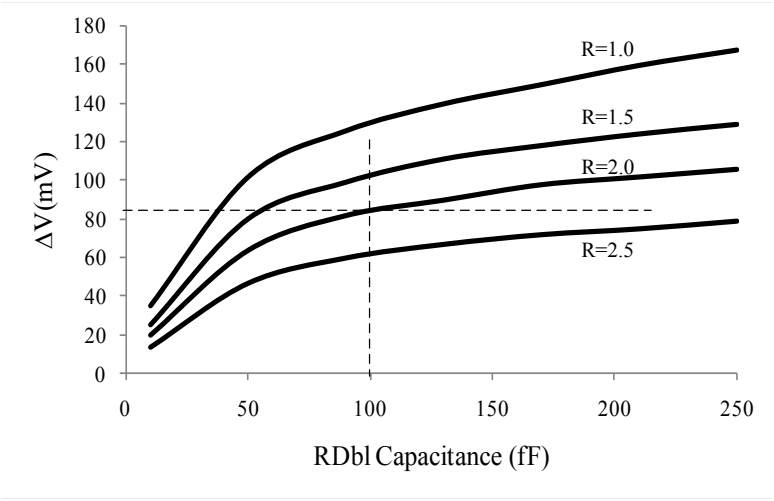


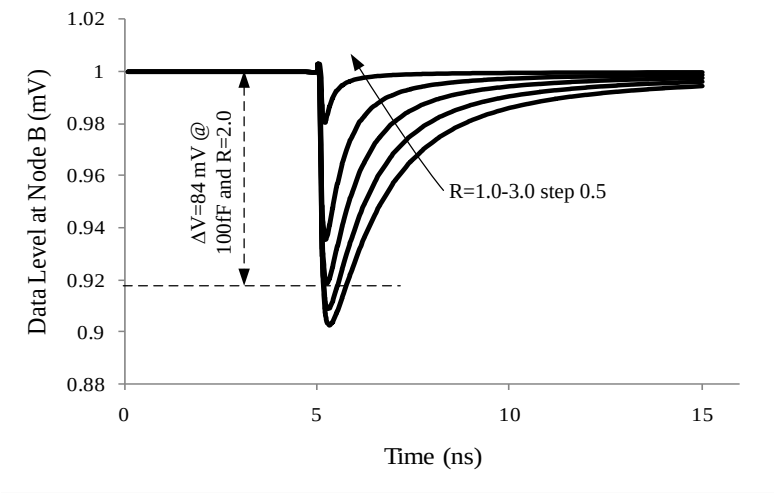
Figure 5.8: The 5T Cell Read Inverter Design Considerations Under Read Access Mode.

results used to investigate the ΔV as a function of $C_{Bitline}$ and R. It can be seen in **Figure 5.9(a)** that higher level degradation is expected as the RDbl capacitance increases.

However, this degradation in the data level at node B is transitional and the cell retrieves the data level shortly after the WLE activation. This exactly resembles the SRAM dynamic noise margin (DNM) principal explained in [35]. Moreover, **Figure 5.9(b)** shows that ΔV is negligible at small RDbl loading values, therefore a cell ratio of R=1 can be used. However, with 100 fF RDbl loading (256 cell/column), a cell ratio of R=2.0 is required to ensure ΔV of 84 mV peak. Given the fact that ΔV is transitional (see **Figure 5.9(b)**) and the cell can tolerate more ΔV for a short period of time (DNM), the cell ratio can be relaxed to reduce the cell area. Therefore, a cell ratio of 1.5 is used which results in a transitional level degradation of ΔV= 100 mV peak. It is worthwhile to mention that modern CMOS technologies (32nmn and below) use stained silicon engineering in which the silicon crystal lattice is compressed to increase holes mobility and thereby to reduce



(a) Transient Data Level Degradation Due to Charge Sharing



(b) Data Level Degradation as a Function of RDbl Loading and Cell Ratio (R)

Figure 5.9: Dynamic Behavior of the Proposed 5T Cell Under Read Access Mode.

the drivability gap between PMOS and NMOS devices [50]. As such, read inverter design can be further relaxed by using R=1. The dependency of the degradation level ΔV peak on the cell ratio R is mathematically verified in **Appendix A**.

5.3.2 Write Inverter Design

The write inverter is designed to perform a successful write operation. To ensure approximately equal WR0 and WR1 margins, this inverter is designed to be symmetrical with a trip point of $V_{DD}/2$. During a write operation, the voltage variation at node B should be within the write inverter dynamic range which is defined by $V_H = V_{DD} - V_{THn}$ and $V_L = V_{THp}$. Therefore, increasing the RDbl above V_{THp} and decreasing the WRbl below $V_{DD} - V_{THp}$ ensure a successful WR1 and WR0, respectively. This indicates that the read and write bitlines are not necessarily full swing signals during a write operation. By limiting the WRbl and RDbl voltage swing, write operation power consumption can be considerably reduced.

Symmetrical inverter design requires equal pull-up and pull-down path strength. Knowing that the drivability of an NMOS is higher than of a PMOS transistor, two equal-width NMOS transistors in series (M3 and M1) is equivalent to one PMOS transistor (M5) of the same width. In order to minimize the proposed cell area, all transistors are chosen to be minimal feature size for a given technology, except for PMOS M4 which is chosen to be 1.5 times minimum size. **Table 5.1** summarizes the designed cell transistors' sizes and **Figure 5.3(c)** (cell VTC curves) reflects the simulation results obtained from the designed inverters.

Table 5.1: 5T vs 6T Bitcell Transistor Sizing in (μm).

	M1	M2	M3	M4	M5	M6
5T	0.15	0.15	0.15	0.25	0.15	NA
6T	0.2	0.2	0.18	0.18	0.15	0.15

5.4 5T Cell Stability Analysis

Memory cell stability during all modes of operation is a crucial reliability issue. During the retention mode the entire SRAM array must be capable of retaining the data. This is usually accomplished by the cross-coupled inverters arrangement. Therefore, the cell's stability under retention mode is not a major design issue since both data nodes are driven to one power rail or another (V_{DD} or gnd). However, the cell's stability under access mode is a major concern in SRAMs. This is mainly due to data level variation resulting from bitline/data interaction.

Figure 5.10 shows an intuitive SRAM array architecture utilizing the proposed 5T bitcell. As can be seen in the figure, column interleaving is doable when the 5T bitcell is utilized. In this example, four bitlines from four words on the same row are interleaved. According to the timing scheme used in the proposed cell, during read access mode, same-row half-selected cells perform a normal read operation like the selected cell; therefore, if the selected cell's stability is proven , the half-selected cells' stability is guaranteed. During write access mode, same-row half-selected cells move deep in retention because of the increase in the crossed-coupled positive feedback gain, *i.e.,* the half-selected cell become more stable.

Unfortunately, this is not the case for same-column half-selected cells. The increased

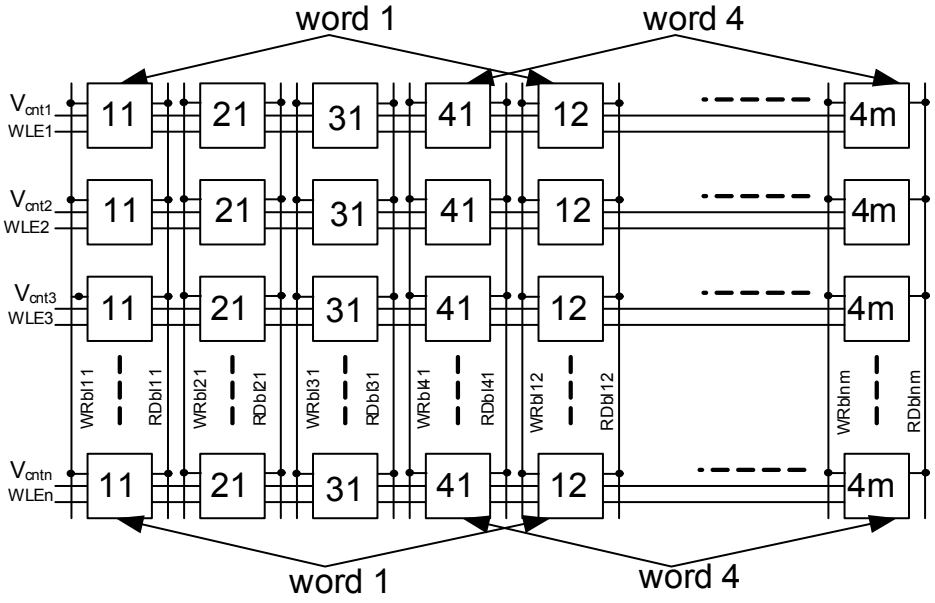


Figure 5.10: The 5T Array Architecture.

voltage level in RD and WR bitlines directly impacts the data stored in same-column half-selected cells. This makes data stored in half-selected cells susceptible to fluctuations. Data fluctuation in half-selected cells on the same column is attributed to voltage variations on the two bitlines (RDbl and WRbl). In particular, cells holding the same data as the selected cell are more vulnerable to level fluctuations. For example, a write "1" operation is accomplished by elevating the RDbl voltage level to upset the selected cell content. However, all same-column half-selected cells holding "0" will perceive the same effect which could upset their contents as well. Thus, verifying the stability of half-selected cells during write operations is a key stability issue.

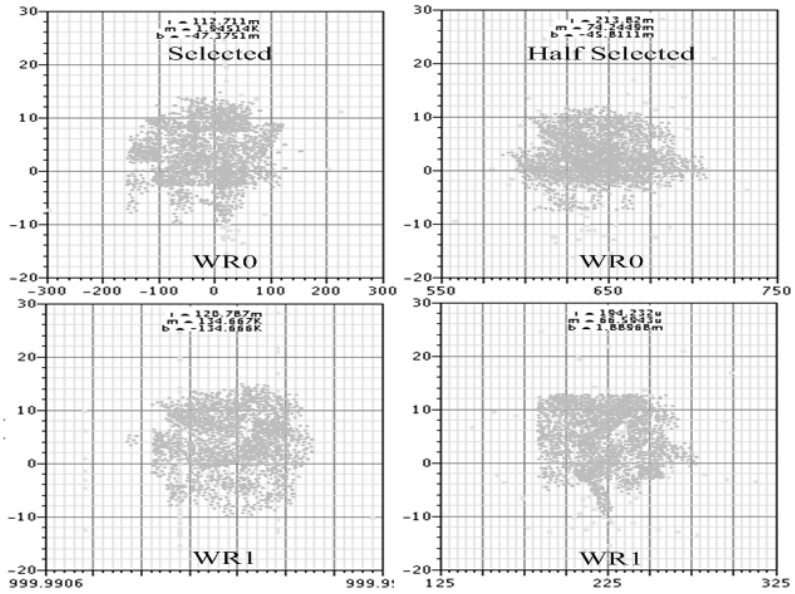
The controlled feedback gain determines the half-selected cells' stability. During a WR1 operation, same-column half-selected cells that hold "0" can retain the data for two reasons:

first the high impedance mode of M3 reduces the closed-loop feedback gain compared to the selected row, and second, low level V_{cnt} clamps the gate voltage of M2 to V_{cnt}-V_{THn} which in turn clamps the voltage rise at node B to V_{cnt}-$2V_{THn}$. This leads to a zero level degradation that must be lower than the trip voltage of the weak write inverter in the non-selected rows (see **Figure 5.6(b)**).

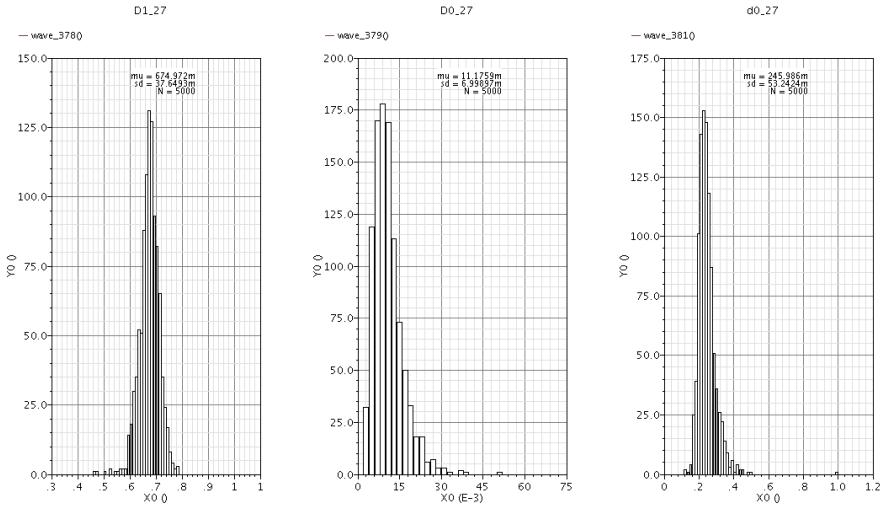
Similarly, same-column half-selected cells that hold "1" are affected during a WR0 operation due to the voltage drop in WRbl. However, these cells are capable of retaining the data because of the limited impact of the voltage level at node B on the access node $\overline{A}$. As such, the voltage level at node $\overline{A}$ is not enough to turn M2 on to fully discharge node B, even though the data level at node A is high enough to turn off M4. The write inverter's VTC in retention mode (shown in **Figure 5.3(b)**) indicates that up to 850 mV level degradation at node A can be tolerated without flipping the cell.

As a result, the data level at node B of the half-selected cells stays high and recovers after the write operation is completed (see **Figure 5.6(a)**). Simulation results shown in **Figure 5.6** verify the stability of the half-selected cells during a write operation. **Figure 5.9(a)** indicates that a 480-mV voltage drop at WRbl is sufficient to write "0" in the selected cell (*i.e.*, the cell's WRM0=520 mV), whereas the half-selected cells retain the stored data ("1") even if the WRbl is completely discharged. Similarly, a 420 mV voltage increase at RDbl is adequate to write "1" in the selected cell (*i.e.*, the cell's WRM1=580 mV), whereas the half-selected cells retain the data even if RDbl is fully charged up to V_{DD}. These values (cell's WRM) were further investigated through statistical post-layout simulations in the presence of process and mismatch variations as shown in **Figure 5.7(a)(a)** and **(b)**.

Figure 5.11 shows Mote Carlo simulation results for data status of the selected and the half-selected cells after WR1 and WR0 operations. The left hand side of **Figure**



(a) Scatter Plot Shows Selected Cell Write-Ability and Half-Selected Cell Stability



(b) Selected and Half-Selected Cells Data Level Distribution: The X-axis signifies voltage level and the Y-axis signifies no. of occurrences

Figure 5.11: Monte Carlo Simulations Over Selected and Half-Selected Cells During a Write Operation.

5.11(a) indicates a successful write operation in the selected cell (WR0 top and WR1 bottom). The right hand side of the figure indicates a limited data level degradation in the half-selected cell. The mean values and the standard deviation in half-selected data level is shown in the histogram Monte Carlo simulation results shown in **Figure 5.11(b)**. As can be seen from this figure, in the presence of process and mismatch variations, the mean value of zero level degradation (the left hand side of the graph) was limited to 245 mV with a standard deviation of 53.24 mV. Similarly, the mean value of the degradation in the "1" level (the right hand side of the graph) was limited to 325 mV (the mean value of the data level at this node drops to 674.97 mV) with a standard deviation of 37.64 mV. The center histogram of the figure indicates a successful WR0 operation of the selected cell. These results verify the proposed cell's stability under the worst-case operating conditions.

5.5 5T-6T Performance Comparison

The performance of the proposed cell was compared to that of the reference 6T SRAM cell. Both cells were laid out in ST 65-nm CMOS technology and each one of them was used to realize a 32-Kbit (256 rows X 128 columns) memory macro. At the cell level, the comparison is based on major SRAM figures of merit, such as SNM, cell current, area, cell leakage current, and energy consumption. At the array level, the comparison is based on overall read/write energy consumption and bitline capacitance loading.

5.5.1 Cell Area and Drivability

Even though the transistor count and transistor size of the proposed cell are smaller than their counterparts in the 6T cell, the actual area of the proposed cell is 6.76% bigger. That

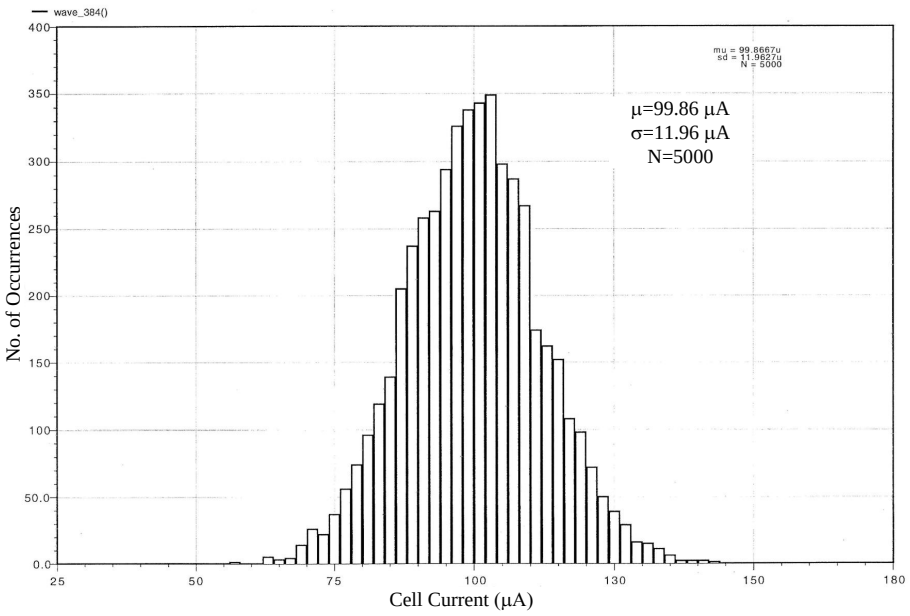


Figure 5.12: Proposed 5T Cell Drivability Monte Carlo Simulation Results During a Read Write Operation.

was because of the asymmetry of the proposed cell's layout which make a horizontal layout overlapping of neighboring cells not possible. Nevertheless, under optimal cell area design and layout the proposed cell drivability (cell current) was higher than that of the 6T cell. Hence, a larger 6T cell is required to match the 5T's cell current increase.

Additionally, the lack of dedicated power supply rails in the proposed cell limits metal layers required for the layout by two layers compared two the three layers used in 6T cell layout. This is important to reduce the parasitic resistance of multiple VIAs used in the layout. Simulation results for optimized 5T and 6T cells show that the proposed 5T cell drivability is about 15% higher compared to the conventional 6T cell. The proposed cell's drivability (cell current) under process and mismatch variations has been verified by 5000

Monte Carlo iterations as shown in **Figure 5.12**.

5.5.2 Leakage Current Calculation

The various leakage current components in the 6T and 5T bitcells are illustrated in **Figure 5.13**. Leakage current components in the 6T cell are data-independent, *i.e.,* each cell produces an equal amount of leakage current from either side (Bl or Blb). Thus, the overall 6T array leakage is given by $n \times I_{leakage}$ cell, where n is the total number of cells in the array. Although all leakage current components in deep sub-micrometer CMOS technology are significant, sub-threshold leakage current and the off state leakage current denoted by solid arrows in **Figure 5.13** are dominant in practice. Therefore, other leakage components, such as gate and substrate leakage (dashed arrows in the figure) can be neglected to simplify hand calculations.

Leakage current components in the 6T SRAM cell can be grouped into two categories: bitline leakage and power supply leakage. In addition to its contribution to overall power consumption during retention mode, bitline leakage affects cell read reliability during read operations (see **Equation 5.1**). Successful read operation requires a cell read current that is orders of magnitude larger than the total leakage current resulting from the half-selected cells on the same column. On the other hand, power supply leakage does not affect the reliability, but it does cause power consumption during an idle condition and hence reduces battery life on portable battery-operated equipment.

According to **Equation 5.3**, the subthreshold leakage current is exponentially proportional to the operating voltages V_{GS} and V_{DS} [51].

$$I_{leakage} = V_T^2 \times \mu_o C_{ox}(W/L)(n-1) \times e^{(V_{GS}-V_{TH})/nV_T} \times \left(1 - e^{-V_{DS}/V_T}\right) \qquad (5.3)$$

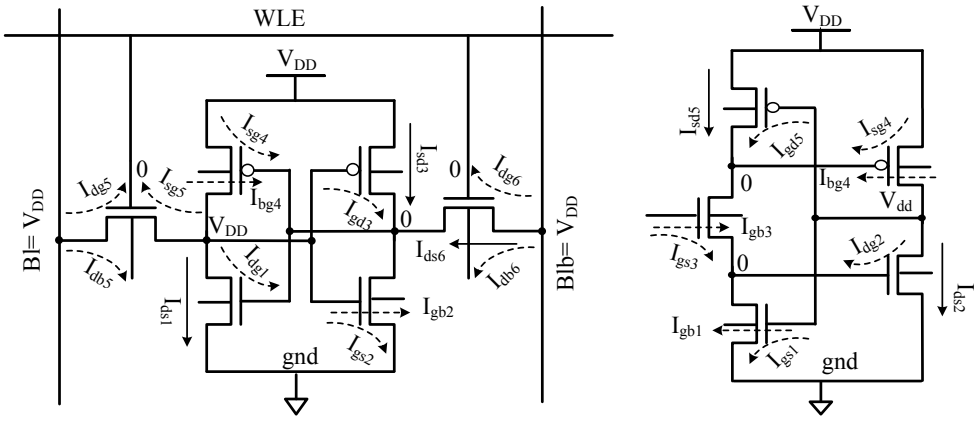


Figure 5.13: Leakage Current Components In 5T and 6T Bitcells.

where, n signifies the gate-to-channel surface voltage ratio known as the subthreshold swing coefficient, and V_T = KT/q is the thermal equivalent voltage.

Subthreshold leakage components, I_{sd3}, I_{ds1} of the 6T cell and I_{sd5}, I_{ds2} of the 5T cell (solid arrows in **Figure 5.13**) are the major supply leakage components to be considered here. Leakage current components I_{sd3}, I_{ds1} and I_{ds2} can be assumed equal due to the equal operating voltages (V_{GS} and V_{DS}). However, I_{sd5} of the 5T is lower because of the high resistive path to ground through M3. The bitline leakage components in the proposed 5T cell are the same as the supply leakage since the bitlines are used to power the cell. I_{ds6} represents the bitline leakage current in the 6T cell. This leakage component is at its maximum due to the maximum operating voltages.

A simple comparison of the leakage components shows that the 6T cell has three major components compared to two components in the 5T cell. Moreover, the 6T cell design requires a relatively stronger driver to ensure an adequate cell ratio. This makes the leakage current higher. Furthermore, the leakage current in the proposed cell is data-dependent

Table 5.2: 5T-6T Figures of Merit Comparison: V_{DD}=1.0 V and 27 C^o.

Metric	Conventional 6T	Proposed 5T	Reduction
Leakage Current(nA)	36.76	15.34	+58%
SNM(mV)	350	$\geq$ 550	-33%
Cell Current (μA)	78.97	91.1	-15.3%
Cell Area (μm^2)	1.24	1.32	+7%

and it is lower when the stored data is "0" because of the lower value of V_{GS2} and V_{GS3}. If the majority of data values stored in a memory array are "0"s, the overall leakage in the proposed bitcell is low compared to a 6T cell where leakage current is data-independent. **Table 5.2** shows the simulation results comparison of the proposed 5T cell compared to the conventional 6T cell.

5.5.3 Energy Consumption

In order to have a rough estimate of energy consumption we utilized **Equation 5.4** to determine the energy consumption associated with read and write operations for both the conventional 6T and the proposed 5T cells.

$$E = \frac{1}{2} C_{Load} \times V^2 \tag{5.4}$$

where, C_{Load} is the expected load capacitance to be driven, and V is the required voltage across the load.

The loading capacitance C_{Load} was extracted from post layout simulations for a column of 256 cells and a row of 128 cells. In the 6T cell, during a read operation, a 200 mV

differential bitline voltage is considered nominal to perform a reliable read operation. A full-swing signal is required at the bitline and WLE to perform write and read operations, respectively. Voltage levels in the proposed 5T cell are different. Because of its single-ended structure, and in order to have a fair comparison, the targeted read bitline voltage is set to 350 mV.

A full-swing WLE signal is required for a read operation, and no WLE activity is required for a write operation. The control signal V_{cnt} is a half-swing ($V_{DD}/2$) signal and it is activated during a write operation and enhanced read operation operation only (usually no V_{cnt} signal activity is required for a read operation). **Table 5.3** tabulates the extracted loading capacitance values for both the 6T and 5T arrays and the associated energy consumption calculation utilizing **Equation 5.4**.

In the 5T array, during a write operation the same-row half-selected cells do not cause any power consumption. Hence, non-selected columns (96 out of 128 in the case of 32 bits/word) stay under retention condition. In contrast, non-selected columns in the 6T array perform a dummy read operation, which means additional power consumption. Energy consumption for the proposed 5T read operation is calculated based on 50% "0" stored data. However, less energy is required when more zeros than ones are stored in the array. Another assumption made when calculating write operation energy consumption is that read and write bitline loading is equal. This assumption is valid since the extracted loading capacitance for both lines was 101.12 fF and 100.096 fF, respectively.

Table 5.3: Loading And Energy Post-Layout Simulation Results Comparison: V_{DD}=1.0 V and 27 C^o.

		Conventional 6T	Proposed 5T	Reduction
Loading (fF)				
Row	WLE	70.4	30.45	-56.71%
	V_{cnt}	NA	39.1	NA
Column	Bitline	118.68	101.12	-14.7%
Energy (fJ)				
Read	WLE	35.24	15.225	-56.71%
	V_{cnt}	0	4.88	NA
	Bitline	293.2	198.2*	-32.47%
Write	WLE	70.4	0	NA
	V_{cnt}	0	4.88	NA
	Bitline	1945.2	815.92**	-58.05%
Total Energy (fJ)		2344.4	1039.13	-55.67%
Leakage Current (μA/Column)				
Data "1"		8.86	3.43	-62.33%
Data "0"		8.86	3.64	-58.42%

* 50% of the cells are assumed storing "0"

** WRbl and RDbl capacitance loading are assumed the same

5.6 Test Chip Implementation and Testing

5.6.1 Test Chip Implementation

In order to verify the proposed 5T SRAM bitcell functionality and performance, a 1.2x1.2 mm^2 test chip was designed and implemented in a standard logic ST 65 nm CMOS fabrication process. The implemented test chip was fabricated through the Canadian Microelectronics Corporation (CMC) in May 2010. The implemented test chip contains three SRAM macros that utilize novel SRAM bitcell schemes and a conventional 6T SRAM reference macro. Each macro is designed as a 32-Kbit array along with the necessary peripheral circuitry. In this section we will discuss the implementation and testing procedure of the fabricated test chip. In particular, we will discuss the implementation of the 5T SRAM array and the associated peripheral circuits. The two other macros are implemented using other SRAM bitcell schemes; namely, 9T and 8T bitcells. These macros were not part of this research and therefore no further discussion is presented.

Figure 5.14 shows a top level floor plan of the fabricated test chip. The 32-Kbit 5T macro occupies the top right corner of the chip. A detailed hierarchical block diagram of the implemented 5T SRAM array along with the timing and control signals used to operate the memory is illustrated in **Figure 5.15**.

A column segmentation technique is used in the test chip where each column is divided into eight 32-rows segment. This technique reduces the stand-by power consumption by ensuring that a cell's supply voltage of non-selected segments is kept at lower supply voltage V_{DDH} (hibernation) compared to the selected segment (full swing V_{DD}) to reduce leakage power consumption. The segment select circuit, shown in **Figure 5.15** is used to switch the segment's local bitlines (LRDbl and LWRbl) between hibernating and active

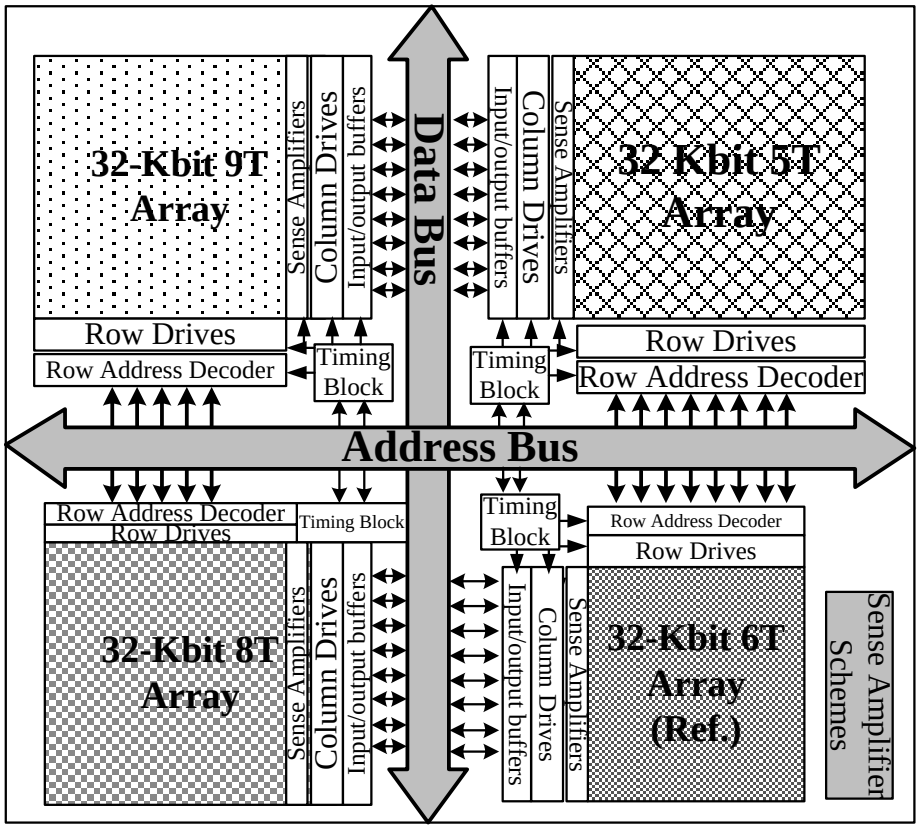


Figure 5.14: The 1.2x1.2 mm^2 Test Chip Top-Level Floor Plan.

conditions. The LRDbl of the non-selected segments (SS is low "0") is connected to gnd through NMOS transistor MN1, while the LWRbl is connected to V_{DDH} through PMOS transistor MP1. The LRDbl and LWRbl of the selected segment (SS is high "1"), on the other hand, is linked to the global bitlines through transistors MN2 and MP2, respectively.

The designed chip has 66 test pins including a 14 bit shared address bus and an 8 bit shared data bus to respectively address and communicate with one array at a time. Two bits of the address bus are dedicated to address the four macros individually using a 2-4 array select decoder. Additionally, a 2-4 data in/out decoder is used to accommodate a

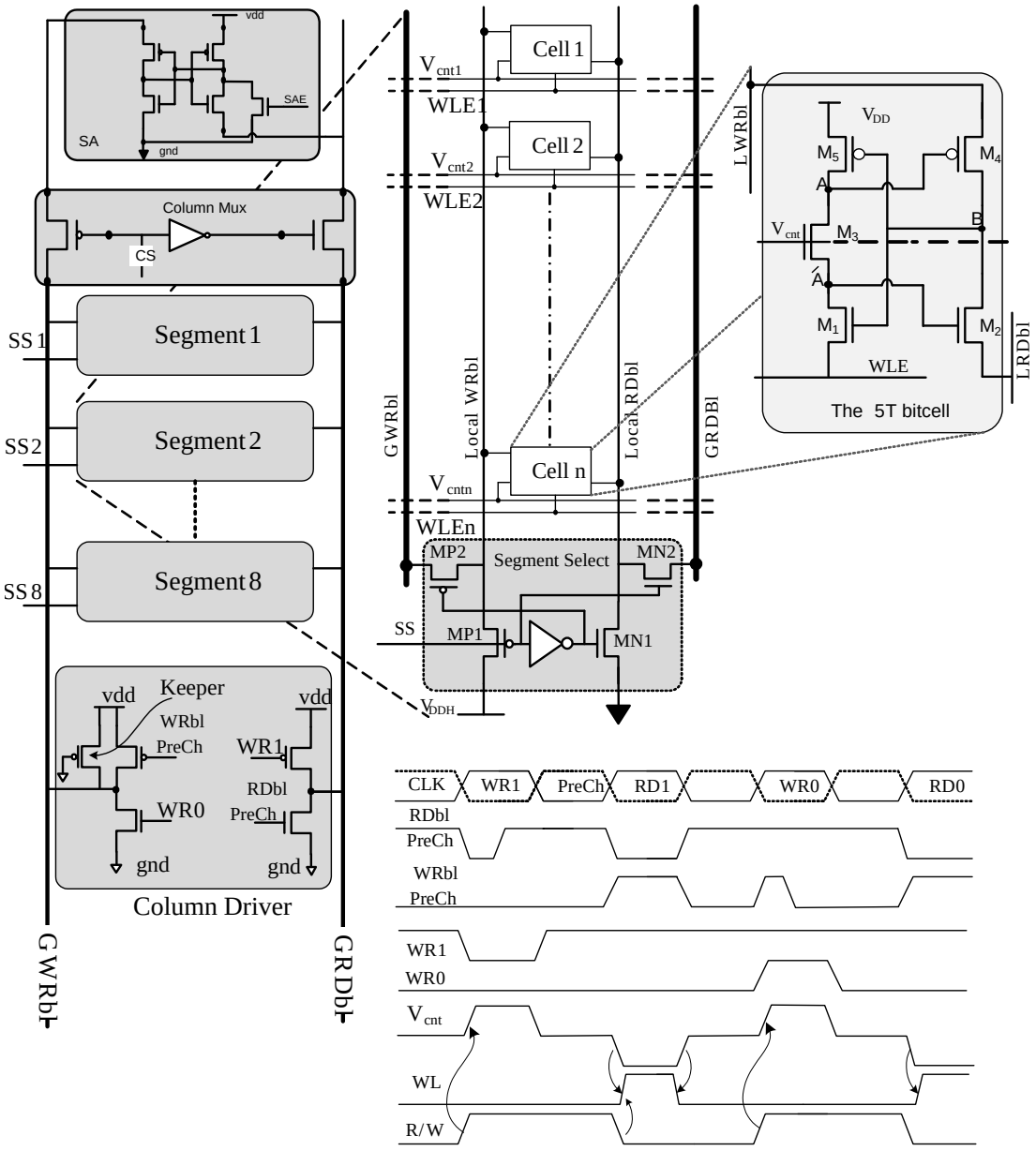


Figure 5.15: Proposed Cell Segmented Column Top-Level Implementation Block Diagram and Associated Timing Signals.

32-bit word on an 8-bit data bus in four clock cycles, as we will see later. In the following sections the structure of each unit used to realize the 5T SRAM macro will be provided. Similar units are used for the other macros.

5.6.2 The Address Bus Construction

Each SRAM array is implemented in a 256 row by 128 column format. Therefore, the first 8 bits of the address bus (A_0-A_7) are used to address one out of 256 possible rows. Since a column segmentation technique is employed, the first three address bits (A_0-A_2) are used to address one out of eight 32-bit segments using a 3-8 segment decoder. The rest of the row address bits (A_3-A_7) are used to address one row in the selected 32-bit segment by a 5-32 row decoder. Two address bits (A_8-A_9) are used to select one out of four 32-bit words using a 2-4 column decoder (multiplexer). Since the data bus capacity is limited to 8 bits , two additional address bits (A_{10}-A_{11}) are used to address an 8-bit data in/out group of the given word. Finally, two array select bits (A_{12}-A_{13}) are used to address one out of four arrays on the chip making the total address bus length used in the experiment 14 bits (A_0-A_{13}).

5.6.3 Row Address Decoder and Row Drivers

(1) Row Address Decoder

Figure 5.16 shows a block diagram of a two stage row address decoder implemented in the test chip. The first stage of the row address decoder (pre-decoder) comprises two units. The first unit is a 3-8 segment decoder used to address one of eight possible column segments. The second unit is a 5-32 decoder used to address one out of 32 possible rows of

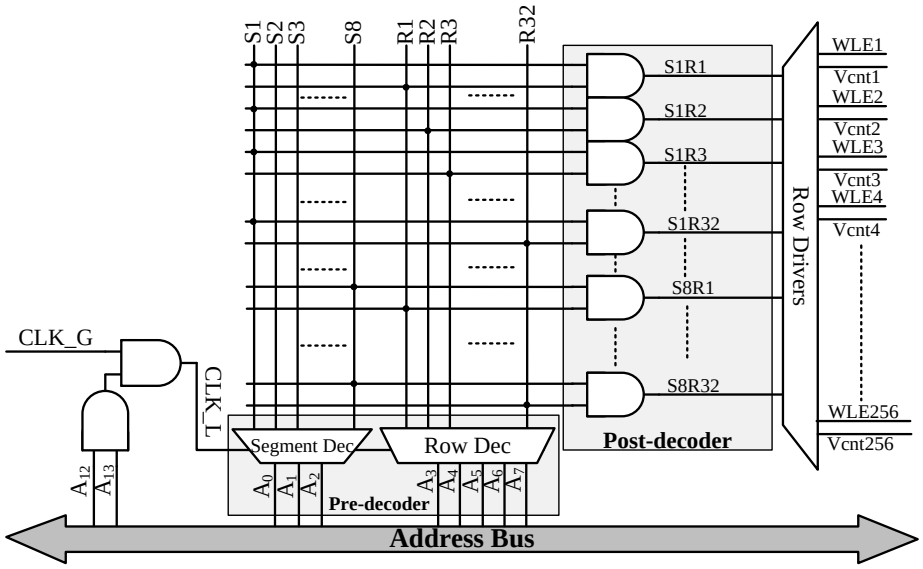


Figure 5.16: A Two-Stage Row Address Decoder Utilized in The Fabricated Test Chip.

the segment. The second stage of the row address decoder is used to multiplex the output of the first stage in order to activate one row in the selected segment.

(2) Row Drivers

The row drivers were designed to generate the cell's operating and control signals WLE and V_{cnt}, as shown in **Figure 5.17**. Since the test chip contained four SRAM macros, the global CLK signal is multiplexed with the array select signal (bits A_{10} and A_{11} of the address bus) to activate a local clock (CLKL) signal of that particular array. The array's CLKL signal along with the read/write operation signal (WR/RD) is used to control the operation of the row drivers. In the presence of the CLKL signal and WR/RD is high (write operation), the row driver holds WLE to ground while V_{cnt} goes to full swing V_{DD}. During a read operation (WR/RD is low), the row driver pulls the WLE high to V_{DD} while V_{cnt} makes a high to low transition (V_{ref} to gnd).

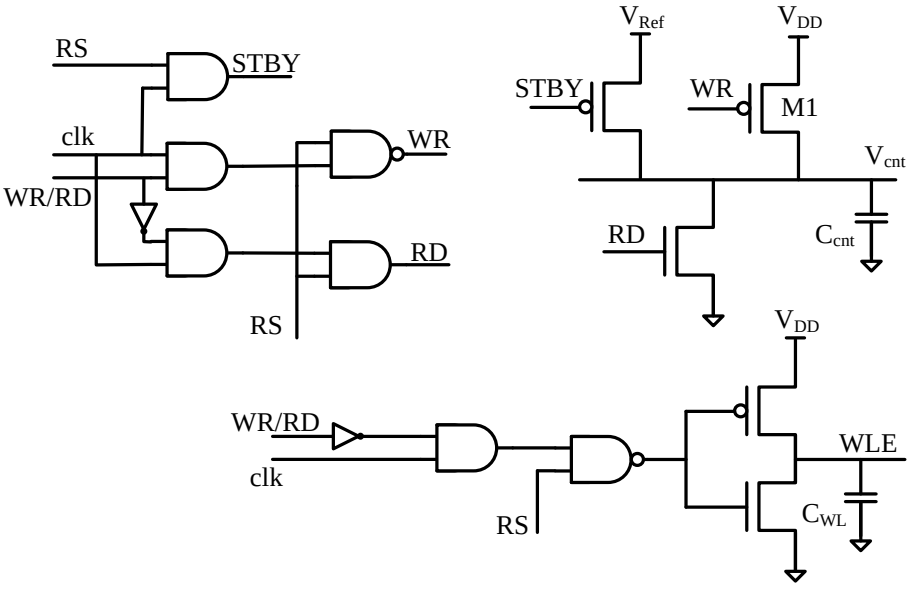


Figure 5.17: Row Driver Circuit Design and the Associated Output Control Signal.

5.6.4 Data Bus

Each row in the implemented memory array is designed to hold four 32- bit words. Data in/out operations are usually performed via a data bus. Due to the limited number of available test pins, data in/out operations are carried out as four 8-bits data bursts,via an 8-bit data bus along with a 2-4 data in/out group decoder. The latency associated with this process is four CLK cycles; in each CLK cycle an 8-bit data burst was input or output. The two additional bits used to address the four data in/out groups form part of the address bus.

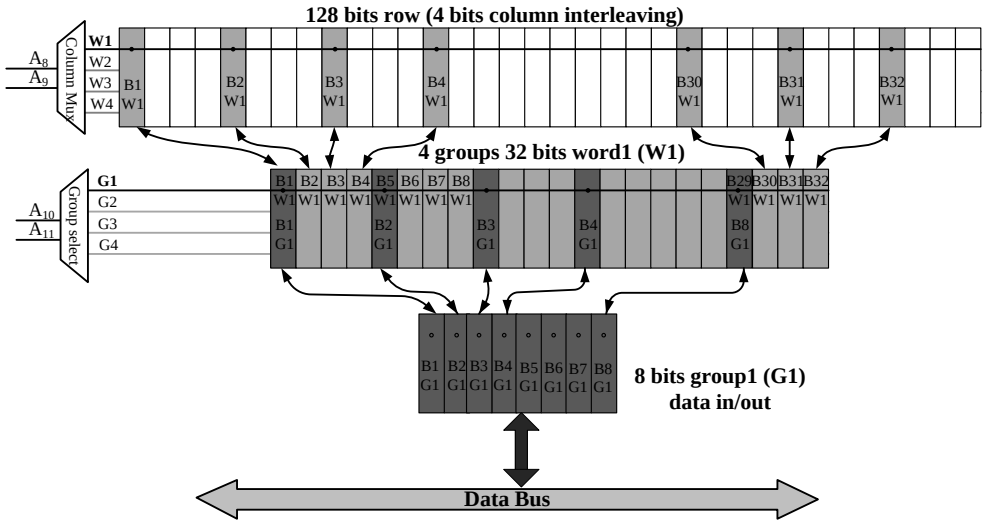


Figure 5.18: Column Interleaving Technique Implementation and Data In/out Multiplexing.

5.6.5 Column Interleaving and Multiplexing

A column interleaving technique is used in the test chip where four bits patch is used such that the first word bits are spread along the 1^{st}, 5^{th}, 9^{th}, and so on until 125^{th} columns (giving a total of 32-bits). When the first word of the selected row is selected (A_8-A_9 are 00), all 32 columns of that word become active. A second level of column interleaving selects one group out of four 8-bit groups of the selected word. A 2-4 group decoder, using address bits A_{12}-A_{13}, is used to select the 1^{st}, 5^{th}, 9^{th}, and so on until 29^{th} columns (8-bits). This data represents the first burst of the output data that correspond to address bits A_{12}-A_{13} set to 00. In the next CLK cycle, bits A_{12}-A_{13} become 01 and the second burst of data is output, and so on. The 8-bit latch in the last stage latches the data and data out buffers are used to buffer the output data to the data bus. **Figure 5.18** illustrates the column interleaving and multiplexing implementation used in the test chip.

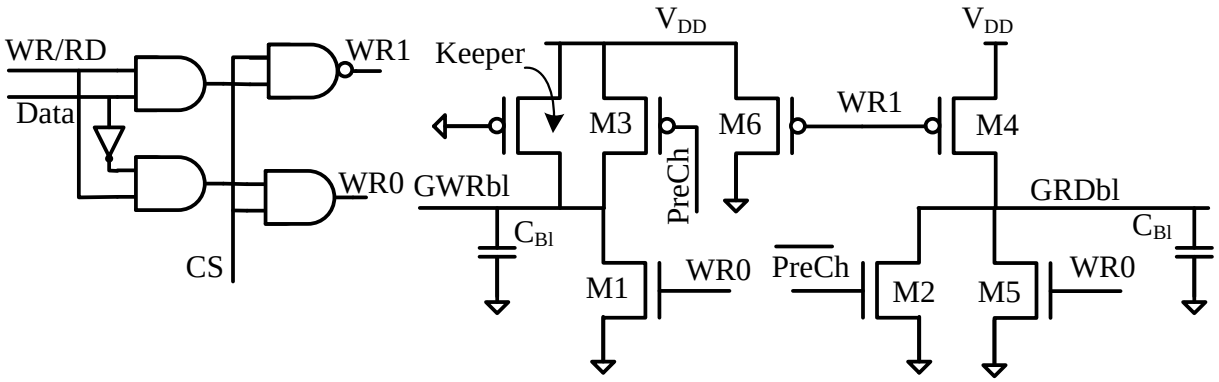


Figure 5.19: The Proposed 5T Bitcell Column Driver.

5.6.6 Column Driver

The column driver, shown in **Figure 5.19**, is designed to be global since the segment's local bitlines are designed to be driven by the segment select circuit, as we mentioned before. In other words, local bitlines of the non-selected segments are kept precharged to the lower power V_{DDH} and ground, where V_{DDH} is the hibernation supply voltage level provided by a dedicated test pin. The precharge control signal (PreCh) is used to set the initial conditions of the global bitlines (GRDbl and GWRbl) such that when the precharge signal is low, GRDbl is pre-discharged to ground while GWRbl is precharged to V_{DD}.

During the evaluation phase (PreCh is high), the cell performs either a read or a write operation and the column driver set the global bitlines' precharge conditions are set accordingly. If a read operation is intended (WR/RD is low "0"), the column driver set the two control signals WR1 and WR0 to "1" and "0", respectively. In this case the GWRbl, which provides the cell's supply voltage V_{DD}, is attached to V_{DD} via a permanently on PMOS transistor (keeper), while the GRDbl is kept floating. If the cell performs a write operation (WR/RD is high "1"), the column driver set the control signals WR1 and WR0

according to the input data (write "1" or write "0"). The column driver set WR1 to "0" and WR0 to "0". In this case PMOS transistor M4 pulls the GRDbl up a high voltage level (typically V_{DD}) while keeping GWRbl attached to V_{DD} through PMOS transistor M6. Similarly, during a write "0", the column driver set WR1 and WR0 to "1". Under these conditions the GRDbl stays grounded through NMOS transistor M5 while the GWRbl is pulled down to low voltage level (typically gnd) through NMOS transistor M1.

In order to investigate the proposed cell's write margin, the GRDbl high voltage level during WR1 operation and the low voltage level of the GWRbl during WR0 operation were made variable. In such a case we would have the flexibility to measure the required voltage drop or rise in order to perform a successful write operation. These two supply voltage were designed to be provided to the test chip via a dedicated pins labeled WRbl-L and RDbl-H.

5.7 Timing and Control Unit

The proposed 5T array is designed to operate at a maximum operating frequency of 1 GHz (1ns CLK signal time interval); however, in order to add some testing flexibility the timing block is designed to generate signals that are suitable for high frequency operation (1 GHz) as well as low frequency operation (about 100 MHz). Additional testing flexibility has been added by using a controllable delay line to control the evaluation and precharge phase of the precharge signal (duty cycle) by controlling the local clock signal (clk) time delay. A dedicated test pin is assigned to switch between the high/low speed operation. Furthermore, the delay line time delay is further fine tuned using a variable DC control signal as shown in **Figure 5.20** where a control signal f_c is used to control the precharge signal duty cycle.

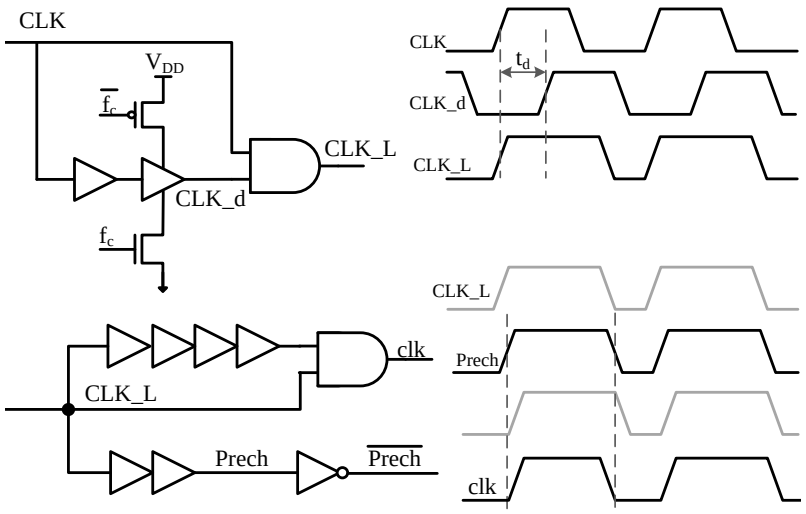


Figure 5.20: The Generation of The Timing Signals Used to Operate The Proposed 5T Array.

5.8 Chip Testing

Figure 5.21 shows the top-level layout of the fabricated test chip. A detailed top-level layout of the implemented 5T macro is shown in **Figure 5.22** along with a single cell layout and schematic superimposed. The test chip pins are assigned almost evenly among the four SRAM macros. However, due to some functional similarities, some of 5T macro test pins are shared with the 9T macro. A CLK-in/CLK-out test pin is used to verify the test chip input/output (IOs) pads functionality. The objective of that test pin is to make sure that the pad ring and the input/output pads (I/Os) are working properly. All SRAM macros implemented on the test chip are designed to be powered independently, *i.e.,* each macro has a separate supply voltage V_{DD} so that active and standby power consumption can be measured for each macro independently. Additionally, the I/Os are supplied with

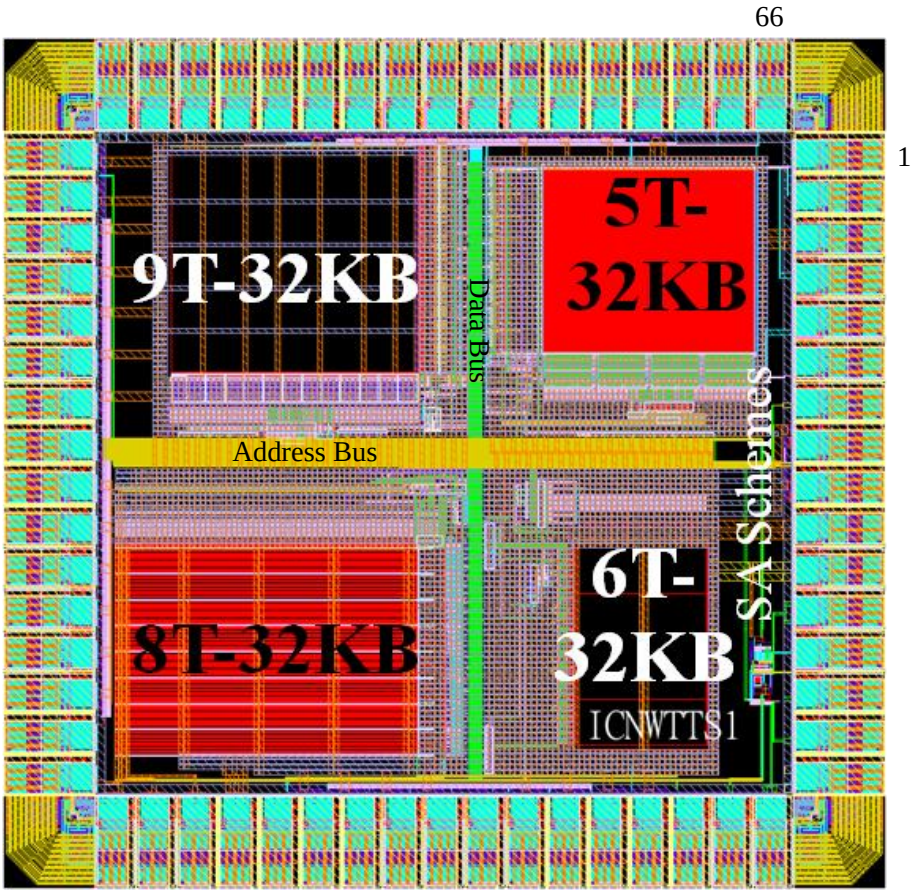


Figure 5.21: The Fabricated Test Chip Top-Level layout.

a dedicated supply voltage V_{DD} and ground terminal on two sides of the pad ring.

5.8.1 Testing Procedure

In order to measure the pad ring functionality, the measurement was initiated with all macros un-powered, *i.e.*, the supply voltage of the macro was not connected to V_{DD}. The measured voltage supply drop and the high current driven for the supply voltage source

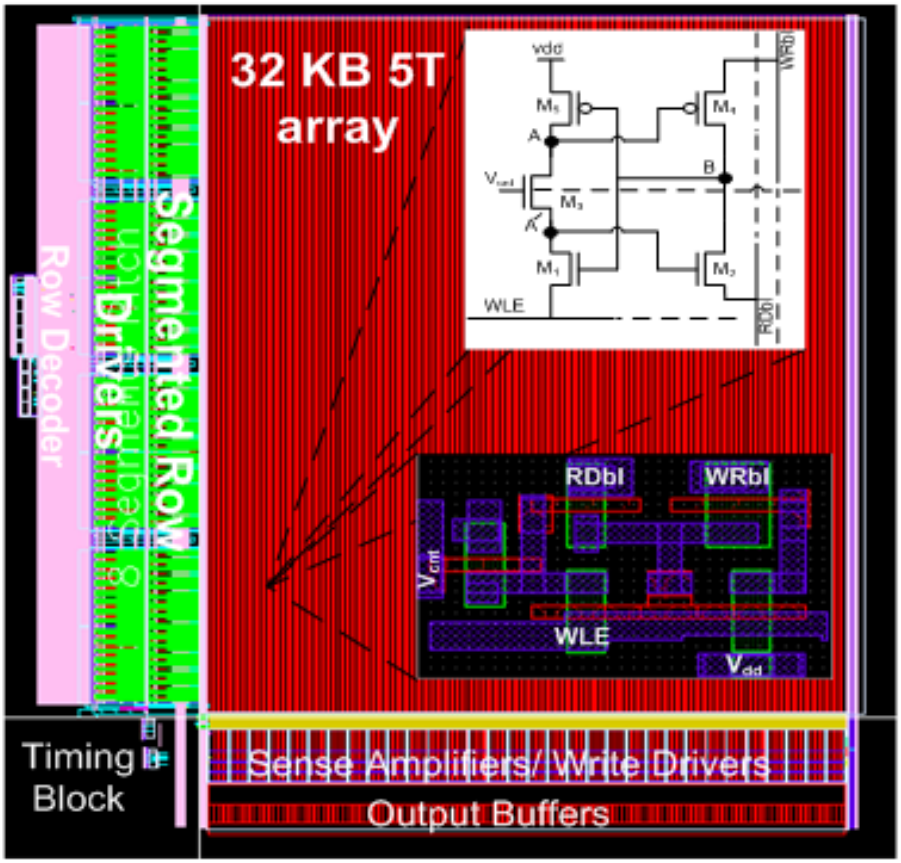


Figure 5.22: Top-Level Layout Implementation of a 32-Kbit SRAM Macro.

indicated a short circuit condition for supply voltage. Subsequently, the chip's top-level layout was investigated and a direct contact between the I/Os supply voltage V_{DD} and the ground was discovered in one location. A microscopic laser was used to cut the S/C point.

In the second attempt of the measurement, the supply voltage and the driven current measurement indicated that the physical S/C was repaired. However there was still a substantial amount of leakage current, above the anticipated value. Compared to another functional test chip implemented in our research group, the measured leakage current was

Table 5.4: Chip I/Os Leakage Current as a Function of the Supply Voltage V_{DD}.

Supply Voltage V_{DD} (V)	Leakage Current (mA)
1.0	20
0.9	18
0.8	15
0.7	13
0.6	10

comparable indicating that the higher leakage current was likely due to the fabrication process. As high leakage current could be attributed to a transistor latch up, we reduced the supply voltage V_{DD} and measured the corresponding leakage current. The linear relationship between the supply voltage V_{DD} and the leakage current indicated resistive voltage loss behavior. **Table 5.8.1** gives the measured leakage currents as a function of the supply voltage V_{DD} when none of the SRAM macros was powered, *i.e.,* the pads ring was unloaded. Furthermore, as individual arrays were powered up, the leakage current increased linearly, *i.e.,* each macro adds more leakage current. The high leakage current reduced the full swing of the measured logic levels. Under typical supply voltage (V_{DD}=1.0 V), the high level voltage swing was 800 mV which was 200 mV lower than the supply voltage V_{DD} (1.0 V).

(1) Functionality Test

In order to verify the chip pad ring (I/Os) functionality, the voltage level at the CLK-out pin when the CLK-in pin set to gnd and V_{DD}, alternately, was measured. The output CLK signal CLK-out showed that the chip was responding properly even though the logic "1" voltage level was below the full swing V_{DD}. In other words, when CLK-in pin set to "0",

the voltage level at CLK-out pin was "0" and 800 mV when CLK-in set to "1".

(a) Row Address Decoder and Column Multiplexer Functionality

Under read operation conditions (WR/RD set to gnd), the decoders functionality was verified by applying different data patterns (DC voltages) at the address bus bits A_0-A_7 and observing the 8 bits output data stream. The results obtained showed a high level data in some locations and low level in the others. By changing the address of the group decoder (A_{10}-A_{11}), different data out combinations were noticed. That indicated that the group select 2-4 decoder was responding properly. Similarly, by changing the address bits (A_8-A_9) (column multiplexer) different data out combinations were obtained, which indicated that the column multiplexer was working as well. It is worthwhile to mention that, the high logic level "1" measurement was below V_{DD} (800 mV) in all locations.

(b) Read/Write functionality Measurements

In order to verify the fabricated chip read/write functionality, AC signals using a data pattern generator available in our test laboratory were applied. Different combinations of address bus data, WR/RD, input data signals were used. The CLK signal frequency was set to 500 MHz. The test plan was to perform a successive read and write operations with different input data patterns and in different locations. The chip completely malfunctioned during the measurement procedure and no measurements could be taken. Even though we have spare test chips fabricated, no further measurements were performed since the spare test chips required a hardware repair using the laser facilities of another department that were inaccessible at the time.

5.9 Summary

A new five-transistor (5T) SRAM cell was presented in this chapter. The proposed cell operation and timing scheme is completely different from conventional SRAM schemes. The cell's stability during access and retention modes of operation was confirmed. Post-layout Monte Carlo simulation results were used to verify the proposed cell's stability under process and mismatch variations. Compared to the conventional 6T SRAM cell, the proposed cell demonstrates a 58% leakage reduction and 55% less energy consumption during a complete read-write cycle. Improvement in both cell drivability and SNM of 33% and 15.3%, respectively, were achieved at a cost of 7% area overhead. The proposed cell was incorporated in a 32-Kbit memory macro implemented in ST's standard 65-nm CMOS technology and a test chip is fabricated. A hardware laser repair was carried out to fix a mistakenly created layout short circuit. Leakage current and voltage level measurements were performed. Those measurements indicated a resistive IR voltage drop that caused about 200 mV voltage drop in logic "1" voltage level compared to the full swing supply voltage V_{DD} which was 1 V. The functionality of the row decoder and the column decoder were verified by using different address data pattern and monitor the output data voltage level. No transient measurements were conducted due to test chip failure during the transient measurements.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Conventional 6T SRAM bitcell integration and performance have been continuously benefiting from the aggressive scaling trend in CMOS technologies. However, in the nanometric CMOS regime cell reliability starts deteriorating due to increasing interconnect loading and process variations. Circuit-level techniques can be used to address the limited ability of the 6T SRAM cell to drive heavily loaded bitlines (cell drivability). Areas where circuit-level solutions are beneficial include:

1. **Cell supply voltage management**

   This includes the cell's power rails V_{DD} and gnd, and the cell's control signals: the bitline and WLE. While the cell power supply voltage has been widely studied and investigated, the WL control signal seems to be a valuable research topic that needs more research and investigation. Controllable WL boost technique can enhance the

SRAM cell reliability in many ways. The increase in cell voltage margins (RDM and WRM) is a direct consequence of using WL boost. Since WL boost reduces the on resistance of the access transistor, the access transistor channel length can be increased without losing the WL boost benefits. Increasing the access transistor's channel length contributes to a considerable reduction in cell leakage current.

2. **SRAM peripheral circuit design**

   The SRAM cell and the sense amplifier are two supplementary components in SRAM array operation. As such, the use of high-sensitivity sense amplifiers can significantly reduce the cell's soft read failures. In particular, current-mode sense amplifier is considered the most promising solution in this regards. Additionally, read-assist techniques are a great asset for SRAM cell reliability.

3. **SRAM bitcell topology**

   Conventional 6T SRAM bitcell design is based on delicate transistor sizing. One shortcoming of conventional 6T SRAM bitcell topology is the interdependency between read and write operations. For a given sizing scheme, fluctuations in transistor V_{TH} may result in destructive read operations. As such, other SRAM bitcell topologies are reported in which the read-write interdependency is eliminated by using separate read and write ports or by using additional transistors in order to, conditionally, break the cell's feedback mechanism. This approach allows the cell transistors to be sized independently.

6.2 Thesis Contributions

In this work we have analyzed and characterized conventional SRAM bitcells. In particular we focused on the conventional 6T bitcell as we have adopted it as a benchmark. We proposed a number of circuit-level techniques to enhance SRAM bitcell stability and reliability when operating in a high-density and low-voltage environment. The proposed techniques were simulated on a typical 256-cell column utilizing the conventional 6T SRAM cell. Column bitline loading has been extracted from post layout simulations to get realistic performance results. Furthermore, SPICE statistical simulations were conducted, whenever necessary, to validate the proposed scheme's functionality in the presence of process and mismatch variations.

The contributions of this research project can be divided into two categories: SRAM peripheral circuit designs and SRAM bitcell circuit design. In the category of peripheral circuits, we accomplished the following:

- design and implementation of a high-performance voltage sense amplifier featuring a controllable read-assist technique,
- analysis and design of a voltage sense amplifier featuring read-assist technique and write-back capabilities for low-voltage-operated SRAM applications,
- design and analysis of a current-mode sense amplifier featuring a read-assist technique, and
- design and analysis of a programmable wordline boost driver for low-voltage operated SRAM circuits.

In the category of SRAM bitcell circuit design, we accomplished the following:

- design, characterize and implementation of a novel five-transistor (5T) SRAM bitcell for low-power SRAM applications.

The following sections highlight the main achievements of this research.

1. **Read-Assist Voltage Sense Amplifiers Scheme I and II:**

   The proposed sense amplifier schemes were designed and laid-out in standard ST 90-nm CMOS technology. The proposed schemes' performance was compared to a reference (conventional) voltage-latch sense amplifier that features a read-assist. Simulation and post-layout simulation results show that, under the same operating conditions, proposed schemes I and II provide 3X and 2.5X faster response compared to the reference. Reductions in column power consumptions of 75% and 50% were achieved using Schemes I and II, respectively.

   Sense amplifier speed-bitline loading dependency is simulated using different bitline loading capacitance values. Simulation results show that the proposed scheme's speed dependence on bitline capacitance is lower than the conventional scheme. Proposed Scheme I, in particular, shows a marginal bitline loading dependence. Whereas the conventional scheme shows, an approximately, linear delay-bitline loading relationship, proposed Scheme I shows only a 5% delay increase when bitline loading is doubled.

   In order to verify the proposed sensing Scheme I with silicon measurements, a "Test Chip I" was fabricated in a standard ST 90-nm CMOS technology. Due to the limited availability of silicon area, only Scheme I was implemented. Even though Scheme I's functionality has been verified with post-layout simulation results, the test chip measurement outcomes unfortunately did not agree with the anticipated scheme operation, therefore we could not verify the proposed scheme's performance.

2. **Current-Mode Sense Amplifier:**

   The proposed current-mode sense amplifier was designed to operate with a 256-cell SRAM column. The performance of the proposed scheme was compared to a voltage-latch sense [9] amplifier because of common features they have, such as the number of transistors and the provision of a read-assist mechanism. In addition, the proposed scheme was compared to a conventional current-mode sense amplifier proposed in [25]. Monte Carlo simulation results show that the proposed scheme provides a 100% yield (0% read failures) with a 25-mV bitline voltage swing compared to a 120-mV swing required for the conventional scheme. Furthermore, the proposed scheme shows performance improvement as bitline voltage swing increases. A 41% speed improvement is achieved when bitline voltage swing is increased to 100 mV as opposed to a constant delay in the conventional scheme. Finally, at a reduced cell supply operating voltage (600 mV), the proposed scheme had no read failures at 45 mV bitline differential voltage swing compared to 115 mV required for the conventional scheme.

3. **Programmable Wordline Boost Driver:**

   A programmable wordline boost driver was designed and simulated in a 4-Kbit SRAM array. This scheme was used to enhance low voltage-operated SRAM drivability. The proposed scheme takes advantage of the DNM concept by applying a transitional wordline boost to increase the access transistor's drivability. According to the DNM concept, an SRAM cell can tolerate a high noise level for a short time interval. As such, the proposed wordline driver was designed with boost peak and pulse duration programmability. By boosting the WL signal, the proposed scheme enables a 3 times increase in the cell's access transistor channel length which contributes to a 39%

reduction in cell leakage current.

4. **Five-transistor 5T SRAM bitcell:**

   A new 5T SRAM bitcell was designed, characterized and implemented in a standard ST 65-nm CMOS technology. Simulation and post layout simulation were conducted to verify the proposed cell's functionality and performance. A 6T bitcell was used as a comparison reference. Compared to conventional 6T cell, post-layout simulation results showed 58% and 55% reductions in cell leakage current and read/write energy consumption, respectively. In addition, cell drivability and SNM improved by 15.3% and 33.3%, respectively. Although the transistor count and size of the proposed cell is smaller than that of the conventional 6T, the proposed cell's effective area is larger. This is due to the symmetry advantage that the 6T cell has which enables cell layout to overlap in four directions compared to the asymmetrical configuration of the proposed 5T cell which allows the cell layout to overlap in only three directions. As a result, the proposed cell exhibits 7% area overhead. Considering the 15.3% increase in cell drivability, 6T and 5T effective area will cross over each other if they are designed for equalized cell drivability.

   In order to verify the proposed bitcell's functionality and performance, a 32-kbit 5T memory macro was realized in the form of a test chip (Test Chip II) fabricated in standard ST 65-nm CMOS technology. Due to some fabrication procedure difficulties, the test chip top-level tape-out encountered a catastrophic physical defect that created a power-to-gnd short circuit. An attempt to fix the fabricated test chip was taken in which we performed a laser cut to the short circuit, however, measurement outcomes were not encouraging. We were able to measure the proposed array leakage current and test the CLK signal toggling, but the chip died and no further

measurements were possible.

6.3 Future Work

Since the core part of this work (Test Chip II) did not work properly, our next step will be to fabricate a new test chip. In fact, the last test chip has been designed to verify the proposed sense amplifier schemes and two novel SRAM bitcells. One of them is the 5T described above and the other one is a new 9T high-performance SRAM bitcell. Therefore, we intend to resubmit the proposed designs for fabrication in February 2012.

Moreover, we have been working to develop a new mathematical formulation to characterize the conventional 6T SRAM SNM and to apply that formula to the proposed 5T bitcell in order to analytically prove the proposed bitcell's superior stability. In addition, we have been working on new-array level techniques that can contribute to bitline loading reduction and thereby enhance the SRAM cell drivability.

APPENDICES

Appendix A

5T Read Inverter Design

The MOSFET transistor current can be defined in a generic current **Equation A.1** [1]:

$$I_{DS2,4} = K_{n,p} \times (W/L) \times \left[(V_{GS} - V_{TH})\, V_{min} - V_{min}^2/2 \right] \times (1 + \lambda V_{DS}) \qquad \text{(A.1)}$$

where, $V_{min} = min(V_{ov}, V_{DS}, V_{DSAT})$, $V_{ov} = (V_{GS} - V_{TH})$ and V_{DS} is the actual drain-source voltage.

According to the transistor operating voltages under read accessed mode, shown in **Figure 5.8**, $V_{min}4$ is: $V_{DD} - (V_{DD} - \Delta V) = \Delta V$; thus,

$$I_{Dp} = \frac{K_p (W/L)_p}{(1 + \Delta V/\xi_C L_e f f)} \left[(V_{DD} - V_{THp}) \Delta V - \Delta V^2/2 \right] \times (1 + \lambda \Delta_V) \qquad \text{(A.2)}$$

The high value of V_{DS2} makes M2 operate in the saturation region. However, short channel devices, operate at relatively high overdrive voltage V_{ov}, enter the velocity saturation regime before the drain-source voltage reaches the transistor overdrive voltage V_{ov}. As such, transistor M2 operates in the velocity saturation region and its drain current can be expressed by:

$$I_{Dn} = \frac{K_n(W/L)_n}{(1 + V_{DSAT}/\xi_C L_e f f)} \left[(V_{DD} - 2V_{THn})V_{DSAT} - V_{DSAT}^2/2\right] \times (1 + \lambda(V_{DD} - \Delta_V)) \tag{A.3}$$

where,

$$V_{DSAT} = \frac{(V_{DD} - 2V_{THn})}{(1 + (V_{DD} - 2V_{THp})/\xi_C L_e f f)} \tag{A.4}$$

In the velocity saturation region the average electric field in the channel (V_{ov}/L_{eff}) is higher than the critical field. As a result, V_{DSAT} is a fraction of V_{ov}, but is still higher than the accepted data-level degradation ΔV. For small values of ΔV and, assuming that V_{DSAT} is three times higher than ΔV, **Equations A.2** and **A.3** yield:

$$\begin{aligned} K_p(W/L)_p \left[(V_{DD} - V_{THp})\Delta V - \Delta V^2/2\right] &= \frac{K_n(W/L)_n}{(1 + 3\Delta V/\xi_C L_e f f)} \\ &\quad \left[(V_{DD} - 2V_{THn})3\Delta V - (3\Delta V)^2/2\right] \end{aligned} \tag{A.5}$$

If at the edge of the velocity saturation region the channel electric field is assumed equals to the critical electric field, *i.e.,* then **Equation A.5** can be reduced to:

$$R = \frac{W}{L} = \frac{K_n}{2K_p}\left[\frac{(V_{DD} - 2V_{THn})3\Delta V - (3\Delta V)^2/2}{(V_{DD} - V_{THp})\Delta V - (\Delta V)^2/2}\right] \tag{A.6}$$

Figure A.1 shows the cell ratio R as a function of the desired data level degradation ΔV. A carrier mobility ratio ($\mu n/\mu p$) of 1.4 is used and V_{THp} greater than V_{THn} according to the technology design rules manual.

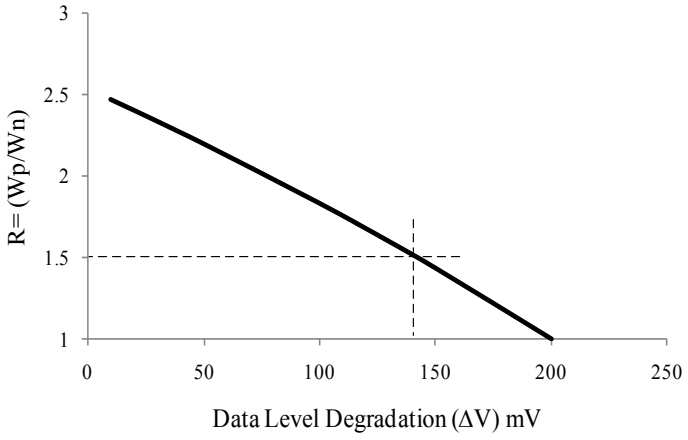


Figure A.1: The Relationship Between Targeted Data Level Degradation and Cell Ratio.

Appendix B

Publications

1. **T. Shakir**, D. Rennie, M. Sachdev, "High Stability, Low-leakage 5T Bitcell for Low-Power Embedded SRAM Applications," Submitted to IEEE Transaction on Very Large Scale Integration (**TVLSI**), April 2011.

2. **T. Shakir**, T. Charania, M. Sachdev, "A Programmable Worldline Boost Driver and Read-Assist, Write-Back Sense Amplifier for 400-mV Embedded SRAMs," Submitted to IEEE Transaction on Circuits and Systems Part I:Regular Paper (**TCAS I**), June 2011.

3. **T. Shakir**, D. Rennie, M. Sachdev, "Integrated Read-Assist Sense Amplifier Scheme for High Performance Embedded SRAMs,"IEEE International Midwest Symposium on Circuits and Systems (**MWSCAS**), pp. 137-140, May 2010.

4. D. Rennie, **T. Shakir**, and M. Sachdev, "Design Challenges in Nanometric Embedded Memories,"IEEE International Conference on Signals, Circuits and Systems (**SCS**), pp. 1-8, Nov 2009.

5. S. Jahinuzzaman, **T. Shakir**, S. Lubana, J. Shah, and M. Sachdev, "A Multiword Based High Speed ECC Scheme for Low-Voltage Embedded SRAMs," in European Solid-State Circuits Conference (**ESSCIRC**), pp. 226-229, Sept 2008.

References

[1] J. Rabaey, A. Chandrakasan, and B. Nikolic, *"Digital Integrated Circuits: A Design Perspective", 2nd Edition*. Upper Saddle River, NJ: Prentice-Hall, Inc, 2003. xiii, 4, 5, 47, 77, 164

[2] M. Bohr, "The New Era of Scaling in an SoC World," *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, pp. 23 –28, Feb 2009. xiii, xiv, 2, 7, 8, 13, 37

[3] H. Yamauchi, "Variation Tolerant SRAM Circuite Designs," *ISSCC Memory Toturial*, Feb 2009. xiv, 6, 52, 53

[4] D. Kanter, "Process Technology at IEDM 2008." `http://www.realworldtech.com/page.cfm?ArticleID=RWT072109003617&p=2`, 2009. [Online; accessed 5-June-2011]. xiv, 9, 10

[5] I. J. Chang and et.al, "A 32-KB 10T Sub-Threshold SRAM Array With Bit-Interleaving and Differential Read Scheme in 90-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 650 –658, Feb 2009. xiv, 17, 18, 38, 92, 111

[6] L. Chang and et.al, "A 5.3GHz 8T-SRAM with Operation Down to 0.41V in 65nm CMOS," *IEEE Symposium on VLSI Circuits*, pp. 252 –253, June 2007. xiv, 17, 18, 92, 111

[7] B. Calhoun and A. Chandrakasan, "A 256-KB 65-nm Sub-Threshold SRAM Design for Ultra-Low-Voltage Operation," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 680 –688, March 2007. xiv, 17, 18, 92, 111

[8] T. H. Kim, J. Liu, J. Keane, and C. Kim, "A 0.2 V, 480 KB Subthreshold SRAM With 1 K Cells Per Bitline for Ultra-Low-Voltage Computing," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 518 –529, Feb 2008. xiv, 17, 18, 92, 111

[9] H. Pilo and et.al, "An SRAM Design in 65-nm and 45-nm Technology Nodes Featuring Read and Write-Assist Circuits to Expand Operating Voltage," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 813 –819, April 2007. xvi, xvii, 35, 56, 57, 58, 69, 71, 72, 79, 82, 83, 103, 104, 111, 160

[10] K. Itoh, *"VLSI Memory Chip Design," Springer Series in Advanced Microelectronics*. Upper Saddle River, NJ: Springer-Verlag Berlin Heidlberg, 2001. 1, 35

[11] R. H. Dennard and et.al, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, pp. 256–268, October 1974. 4

[12] Bhavnagarwala and et.al, "The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Stability," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 658–665, April 2001. 7, 50, 75

[13] Tschanz and et.al, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage," in

IEEE International Solid-State Circuits Conference Digest of Technical Papers, vol. 2, pp. 344–539, Feb 2002. 9

[14] K. Goser and M. Pomper, "Five-Transistor Memory Cells in ESFI MOS Technology," *IEEE Journal of Solid-State Circuits*, vol. 8, pp. 324 – 326, Oct 1973. 10, 35, 49

[15] K. Zhang and et.al, "A 3-GHz 70-MB SRAM in 65-nm CMOS Technology with Integrated Column-Based Dynamic Power Supply," *IEEE Journal of Solid-State Circuits*, vol. 41, pp. 146 – 151, Jan 2006. 16, 56, 91

[16] Y. H. Chen and et.al, "A 0.6 V Dual-Rail Compiler SRAM Design on 45-nm CMOS Technology With Adaptive SRAM Power for Lower VDD_min VLSIs," *IEEE Journal of Solid-State Circuits*, pp. 55 –58, June 2008. 16, 91

[17] J. Pille and et.al, "Implementation of the CELL Broadband Engine in a 65-nm SOI Technology Featuring Dual-Supply SRAM Arrays Supporting 6GHz at 1.3V," *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, pp. 322 –606, Feb 2007. 16, 91

[18] M. Yabuuchi and et.al, "A 45nm Low-Standby-Power Embedded SRAM with Improved Immunity Against Process and Temperature Variations," *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, pp. 326 –606, Feb 2007. 16, 91

[19] S. Ohbayashi and et.al, "A 65-nm SoC Embedded 6T-SRAM Designed for Manufacturability With Read and Write Operation Stabilizing Circuits," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 820 –829, April 2007. 16

[20] H. Morimura and N. Shibata, "A Step-Down Boosted-Wordline Scheme for 1-V Battery-Operated Fast SRAM's," *IEEE Journal of Solid-State Circuits*, vol. 33, pp. 1220 –1227, Aug 1998. 16, 91

[21] K. Ishibashi and et.al, "A 1-V TFT-Load SRAM Using a Two-Step Word-Voltage Method," *IEEE Journal of Solid-State Circuits*, vol. 27, pp. 1519 –1524, Nov 1992. 16, 91

[22] E. Seevinck, P. 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. 26, pp. 525 –536, April 1991. 19, 34, 57, 59

[23] K. Ishibashi and et.al, "A 6-ns 4-MB CMOS SRAM With Offset-Voltage-Insensitive Current Sense Amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 30, pp. 480 –486, April 1995. 34

[24] A. Conte and et.al, "A High-Performance Very Low-Voltage Current Sense Amplifier for Nonvolatile Memories," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 507 – 514, Feb 2005. 34

[25] M. Hiraki and et.al, *"Current Sense Amplfier"*. No. 5534800, US Patent, 1996. 34, 57, 59, 79, 82, 83, 160

[26] J. Meyer and E. Boleky, "High Performance, Low Power CMOS Memories Using Silicon-on-Sapphire Technology," *International Electron Devices Meeting*, vol. 17, p. 44, 1971. 35

[27] K. Noda and et.al, "A Loadless CMOS Four-Transistor SRAM Cell in a 0.18- μm Logic Technology," *IEEE Transactions on Electron Devices*, vol. 48, pp. 2851 –2855, Dec 2001. 36

[28] H. Noguchi and et.al, "Which is the Best Dual-Port SRAM in 45-nm Process Technology? 8T, 10T Single End, and 10T differential," *IEEE International Conference on Integrated Circuit Design and Technology and Tutorial,ICICDT*, pp. 55 –58, June 2008. 37

[29] Y. Morita and et.al, "An Area-Conscious Low-Voltage-Oriented 8T-SRAM Design Under DVS Environment," *IEEE Symposium on VLSI Circuits*, pp. 256 –257, June 2007. 37, 38, 56

[30] H. Yamauchi, "Embedded SRAM Circuit Design Technologies for a 45-nm and Beyond," *7th International Conference on ASIC*, pp. 1028 –1033, Oct 2007. 37, 91

[31] R. Joshi and et.al, "6.6+ GHz Low Vmin, Read and Half Select Disturb-Free 1.2 MB SRAM," *IEEE Symposium on VLSI Circuits*, pp. 55 –58, June 2008. 38

[32] C.-T. Chuang and et.al, "High-Performance SRAM in Nanoscale CMOS: Design Challenges and Techniques," *IEEE International Workshop on Memory Technology, Design and Testing, MTDT*, pp. 4 –12, Dec 2007. 38

[33] E. Seevinck, F. List, and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," *IEEE Journal of Solid-State Circuits*, vol. 22, pp. 748 – 754, Oct 1995. 49, 117

[34] A. Kawasumi and et.al, "A Single-Power-Supply 0.7V 1GHz 45-nm SRAM with An Asymmetrical Unit-β -ratio Memory Cell," *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, pp. 382 –622, Feb 2008. 49, 113

[35] M. Sharifkhani and M. Sachdev, "SRAM Cell Stability: A Dynamic Perspective," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 609 –619, Feb 2009. 51, 95, 127

[36] Z. Guo and et.al, "Large-Scale SRAM Variability Characterization in 45-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 3174 –3192, Nov 2009. 54, 93

[37] A.-T. Do and et.al, "Design and Sensitivity Analysis of a New Current-Mode Sense Amplifier for Low-Power SRAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, pp. 196 –204, Feb 2011. 59

[38] T. Matsuura and et.al, "1.2 V Mixed Analog/Digital Circuits Using 0.3 μm CMOS LSI Technology," *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, pp. 250–251, Feb 1994. 59

[39] A. Hokazono and et.al, "MOSFET Design for Forward Body Biasing Scheme," *IEEE Electron Device Letters*, vol. 27, pp. 387 – 389, May 2006. 77

[40] M. Khellah and et.al, "A 4.2-GHz 0.3 mm^2 256-Kb Dual-V/Sub cc/ SRAM Building Block in 65nm CMOS," *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, pp. 2572 –2581, Feb 2006. 90

[41] K. Itoh, "Adaptive Circuits For the 0.5-V Nanoscale CMOS Era," *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, pp. 14 –20, Feb 2009. 90

[42] K. Itoh and et.al, "A Deep Sub-V, Single Power-Supply SRAM Cell With Multi-VT, Boosted Storage Node and Dynamic Load," *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, pp. 132 –133, June 1996. 90

[43] M. Khellah and et.al, "Wordline and Bitline Pulsing Schemes for Improving SRAM Cell Stability in Low-Vcc 65nm CMOS Designs," *Symposium on VLSI Circuits, Digest of Technical Papers*, pp. 9 –10, Dec 2006. 91

[44] K. Takeda and et.al, "A Read-Static-Noise-Margin-Free SRAM Cell for Low-V_{DD} and High-Speed Applications," *IEEE Journal of Solid-State Circuits*, vol. 41, pp. 113 – 121, Jan 2006. 92, 113

[45] J. Wang, S. Nalam, and B. Calhoun, "Analyzing Static and Dynamic Write Margin for Nanometer SRAMs," *IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 129 –134, Aug 2008. 95

[46] Y. Zhang, P. Li, and G. Huang, "Separatrices in High-Dimensional State Space: System-Theoretical Tangent Computation and Application to SRAM Dynamic Stability Analysis," *47th IEEE Design Automation Conference (DAC)*, pp. 567 –572, June 2010. 95

[47] K. Kushida and et.al, "A 0.7V Single-Supply SRAM With 0.495 μ m^2 Cell in 65-nm Technology Utilizing Self-Write-Back Sense Amplifier and Cascaded Bitline Scheme," *IEEE Symposium on VLSI Circuits*, pp. 46 –47, June 2008. 103

[48] T.-H. Joubert, E. Seevinck, and M. du Plessis, "A CMOS Reduced-Area SRAM Cell," *The 2000 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 3, pp. 335 –338, 2000. 112

[49] M. Wieckowski, S. Patil, and M. Margala, "Portless SRAM: A High-Performance Alternative to the 6T Methodology," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 2600 –2610, Nov 2008. 112, 113

[50] B. Geuskens and et.al, "Opportunities for PMOS Read and Write ports in Low Voltage Dual-Port 8T Bitcell Arrays," *IEEE Custom Integrated Circuits Conference*, pp. 1 –4, Sep 2010. 129

[51] K. Roy and et.al, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, pp. 305 – 327, Feb 2003. 136

[52]